

LEARNING FROM MULTIMODAL WEB DATA

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

John Miles Hessel

August 2020

© 2020 John Miles Hessel
ALL RIGHTS RESERVED

LEARNING FROM MULTIMODAL WEB DATA

John Miles Hessel, Ph.D.

Cornell University 2020

Learning from large, unlabeled web corpora has proven effective for a variety of multimodal understanding tasks. But algorithms that leverage this type of data often assume literal visual-textual correspondences, ignoring the non-literal ways in which users actually communicate online. As user attention is increasingly dominated by multimedia content (e.g., combinations of text, images, videos, etc.), community moderators require tools capable of processing these complex forms of communication.

In this work, we detail our progress towards two related research goals. The first goal is to **leverage** multimodal web data in settings of weak (or “web”) supervision. The ultimate aim of this line of work is to build models capable of drawing connections between different modes of data, e.g., images+text. To this end, we present algorithms that discover grounded image-text relationships from noisy, long documents, e.g., Wikipedia articles and the images they contain. We also demonstrate that noisy web signals, such as speech recognition tokens from user-generated web videos, can be leveraged to improve performance in caption generation tasks.

While these results show that multimodal web data can be leveraged for building more powerful machine learning-based tools, the communicative intent of multimodal posts, which extend significantly beyond literal visual description, are not well understood. Thus, the second goal is to better **understand** communication in a non-textual web. We first conduct an *in-vivo* study

of several Reddit communities that focus on sharing and discussing image+text content; we train algorithms that are able to predict popularity in this setting, even after controlling for important, non-content factors like post timing. Finally, inspired by the fact that when text accompanies images online, rarely does the text serve as pure literal visual description (an assumption enforced by most curated image captioning datasets), we introduce algorithms capable of quantifying the visual concreteness of concepts in multimodal corpora. We find not only that our scoring method aligns with human judgements, but that concreteness is context specific: our method discovers that “London” is a consistent, identifiable visual concept in an image captioning dataset (because post-hoc annotators only mention “London” in captions if the image is iconically so), but not in a Flickr image tagging dataset (because users may tag any image that happens to be taken in London with the geotag “London”).

BIOGRAPHICAL SKETCH

Jack grew up in Portola Valley, CA, an area not notable for its ice hockey rinks. Fortunately, throughout his educational career, he managed to edge ever closer to the frozen expanses of Canada; prior to earning his PhD in Computer Science at Cornell, he spent his undergraduate years at Carleton College in Northfield, Minnesota, attempting¹ to play the sport. In his spare time, he studied Computer Science and Statistics, too.² After a decade in the cold, though, his inner-Californian is re-emerging, reminding him that, *no*, sun and Zambonis are not mutually exclusive. While he's not entirely sure computers can ever be intelligent, he certainly plans to keep on trying to make them so after graduation.

¹Poorly, according to some on-ice opponents.

²Though, there exist a number of alternate timelines wherein Jack majored in Physics. And Economics. And Ethnomusicology.

To my parents, Teri and Doug, and my sister, Sydney.

ACKNOWLEDGEMENTS

I have been *undeservedly fortunate* during my time at Cornell for my network of collaborators. My reverence for my advisor, Lillian Lee, started early: I recall walking out of Lillian’s office after one of our first meetings, thinking to myself: *“if I can make my brain work even 10% as well hers, my PhD will have been a success!”* I am so thankful for the opportunities I’ve had to learn from Lillian; her ability to appreciate (and even find wonder in!) complexity was alien to me as a beginning grad student, confused about why my regressions weren’t working. But, thanks to her patience, I’m more comfortable embracing situations where there is no “right answer” in research (and beyond). Years of support, kindness, discussions, and late-night deadline pushes cannot be meaningfully summarized in just a handful of words, but I will say: thank you, Lillian. My debt to you is unbounded.

Working with David Mimno has also been a blessing of my graduate career. Early on in my PhD, I was attending my first NLP conference (EMNLP 2015), nervously walking around the social event without much place to go. David saw me wandering, pulled me into his discussion, and introduced both me and the work I was presenting to his colleagues. *“David Mimno remembers my paper?!?”* a stunned younger me thought to himself. In the years since, David has played a formative role in my academic career; for his mentorship, collaboration, and warmth, I cannot thank him enough.

I’m similarly grateful for my committee members Steve Marschner and Karthik Sridharan. In addition to advising me in a number of capacities, they served as excellent supervisors for a greenhorn TA, more nervous about holding his first office hours than he might have let on.

Cornell CS+NLP are fantastic communities. If any potential PhD students hap-

pen to be reading this, I don't think you'll regret going there! I was particularly fortunate to have a number of collaborators, professors, mentors, and friends, including Rediet Abebe, Yoav Artzi, Maria Antoniak, Claire Cardie, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Molly Feldman, Dylan Foster, Liye Fu, Jake Gardner, Andrew Hirsch, Arzoo Katiyar, Balázs Kovács, Matt Kusner, Moontae Lee, Matthew Milano, Vlad Niculae, Geoff Pleiss, Maithra Raghu, Rahmtin Rotabi, Tobias Schnabel, Xanda Schofield, Tianze Shi, Ana Smith, Alane Suhr, Chenhao Tan, Laure Thompson, Andreas Veit, Gregory Yauney, and Justine Zhang.

Summer Internships: During my PhD, I was fortunate to do four research internships: one at Twitter in NYC, one at Facebook in Menlo Park, and two at Google Research in Mountain View. I was fortunate to have an amazing network of collaborators, mentors, and friends, including: Clément Farabet, Conrado Miranda, Bo Pang, Radu Soricut, Nikolai Yakovenko, Allan Zelener, and Zhenhai Zhu.

Carleton College: I briefly returned to Northfield during my fifth year of graduate school to be a visiting instructor of Computer Science at Carleton College. Seeing things on the "other side of the curtain" (as Eric Alexander once put it to me) was a growing experience. I thank the whole CS department not only for entrusting 50+(!) undergraduates across two classes to me (a relatively unpoven educator!), but also for the warmth of the community more broadly. My work there wouldn't have been possible without the particularly kind support of David Liben-Nowell, Dave Musicant, and Sneha Narayan.

Stewart Little Coop: I lived in the best cooperative house during the six years of my PhD. While there were many friends I unfortunately didn't get to live with for more than a few years, I'm thankful for my longer-term coop-mates,

without whom, I would have been far more lost on this journey, including: Gabriele Albertini, Martik Chatterjee, Jon Davenport, Jamie Freeman, Joshua Gancher, Neil Garson, Vanessa Kern, Samuel Leiboff, Kai Mast, Nandini Mehrotra, Annie Otwell, Zach Price, Yael Rhodes, Danny Rosenberg-Daneri, Marion Schelling, Nick Stepankiw, Kass Urban-Mead, April West, Natasha Zella, and Marty Ziech.

Hockey: I am of course thankful, too, for my teammates in the Ithaca Adult Ice Hockey Association. I played for four seasons, and while the championship alluded me, I made some great friends — far too many to name! But I am particularly thankful for Geoffrey Fatin, Ian Hewson, Justin Nicholatos, Tim Robinette, and my captians Merri Goodrow and Tom King. I also thank all members of the Leftovers D-6 SIAHL team in San Jose for keeping me fit and sane during my internships in CA.

Family: This dissertation is dedicated to my parents, Doug and Teri, and my sister Sydney. Without them, not only would I have not made it through the difficult times of my PhD, but I wouldn't have been in a PhD program in the first place. My parents are, and have always been, supportive of my computer science adventures (both even read of my arXiv submissions!) but in different and complementary ways. Sydney continues to be my lifelong role-model; having an older sibling as smart, open, and kind as her is one of my life's greatest blessings. And, of course, this journey wouldn't have been possible without Lisa Watkins; her support, humor, and energy have been (and continue to be!) a great source of happiness in my life.

TABLE OF CONTENTS

Biographical Sketch	iii
Dedication	iv
Acknowledgements	v
Table of Contents	viii
List of Tables	x
List of Figures	xii
1 Introduction and Organization	1
1.0.1 What is visual-textual grounding?	1
1.0.2 A brief theoretical case for studying grounding problems	3
1.1 A Practical Case for Considering Multimodal <u>Web</u> Data	4
1.2 Understanding Online Communication	6
1.2.1 Prior Theories of Image-Text Communication	7
1.3 The Connection Between Leveraging and Understanding	8
1.4 Organization and Contributions	9
2 Leveraging: Discovering Visual-Textual Connections in Noisy Web Documents	13
2.1 Brief Overview	13
2.2 Introduction	13
2.3 Task Formulation	16
2.4 Models	17
2.4.1 Alignment Model and Loss Function	18
2.4.2 Similarity Functions	19
2.4.3 Baselines	21
2.5 Experiments on Crowdlabeled Data	22
2.5.1 Crowdlabeled-Data Results	25
2.6 Experiments on RQA and DIY	28
2.7 Qualitative Exploration	31
2.8 Additional related work	33
2.9 Conclusion	34
2.10 Supplementary Material	35
2.10.1 Data preprocessing details	35
2.10.2 WIKI Fine-tuning Details	38
2.10.3 Additional Results	41
3 Leveraging: Generating Captions for Web Videos using Noisy ASR	44
3.1 Brief Overview	44
3.2 Introduction	44
3.3 Related Work	48
3.4 Dataset	49
3.4.1 A Closer Look at ASR tokens	50

3.4.2	A Closer Look at the Generation Task	51
3.5	Models	53
3.5.1	Transformer-based Neural Models	54
3.6	Experiments	56
3.6.1	Diversity of Generated Captions	58
3.7	Complementarity of Video and ASR	59
3.8	Oracle Object Detection	62
3.9	Conclusion	64
4	Understanding: Predicting Popularity in Multimodal Communities	65
4.1	Brief Overview	65
4.2	Introduction	66
4.3	Datasets	69
4.4	Time and Rich-getting-richer	72
4.5	Model Design	79
4.6	Results	86
4.6.1	Unimodal Experiments	87
4.6.2	Multimodal Experiments	88
4.7	Analysis of aww	92
4.8	Additional Related Work	94
4.9	Conclusion and Future Work	95
4.10	Brief Retrospective	97
5	Understanding: Quantifying the Visual Concreteness of Concepts	98
5.1	Brief Overview	98
5.2	Introduction	99
5.3	Related Work	101
5.4	Quantifying Visual Concreteness	103
5.4.1	Concreteness of discrete words	103
5.4.2	Extension to continuous topics	104
5.5	Datasets	105
5.6	Validation of Concreteness Scoring	107
5.6.1	Concreteness and human judgments	108
5.6.2	Concreteness within datasets	110
5.6.3	Concreteness varies across datasets	111
5.7	Learning Image/Text Correspondences	111
5.7.1	Concreteness scores and performance	116
5.8	Beyond Cross-Modal Retrieval	118
5.9	Future Directions	119
5.10	Brief Retrospective	120
6	Future Work	121
6.0.1	Ethical Considerations in Machine Learning	121
6.0.2	Future Multimodal Work	126

LIST OF TABLES

2.1	Dataset statistics: top half = crowd-labeled datasets; bottom half = organically-multimodal datasets. <i>Density</i> measures the sparsity of the ground truth graph as the number of ground-truth edges divided by the number of possible edges.	24
2.2	Results for crowd-labeled datasets (similar results for other settings are included in the supplementary material (§ 2.10)). Values are bolded if they are within 1% of the best-in-column performance.	25
2.3	Performance on the organically-multimodal data; values within 1% of best-in-column are bolded.	29
2.4	Results for crowd-labeled data with ground-truth annotation with $b = 20$ negative samples.	41
2.5	Results for crowd-labeled data with $b = 30$ negative samples.	42
2.6	Results for organically-multimodal data with ground-truth annotation with $b = 20$ negative samples.	42
2.7	Results for organically-multimodal data with $b = 30$ negative samples.	43
3.1	The performance of several state-of-the-art, video-only models, with lower (constant prediction) and upper (human estimate) bounds.	53
3.2	Caption generation performance: AT+Video is a multimodal model that adds visual frame features to AT. A bolded value in a column indicates a statistically-significant improvement, whereas an underline indicates a statistical tie for best ($p < .01$).	57
4.1	Number of unique users, number of Imgur submissions, and the average caption length for the communities used in this study. The number of unique users includes those who commented or submitted.	71
4.2	Statistics regarding the sampling used to generate ranking pairs. The maximum window is the maximum number of minutes that two submissions can be apart to be paired up, whereas the average window is the average time between all sampled pairs. The median and mean score differences between pairs is also given.	77
4.3	Human annotation accuracy results.	78
4.4	Unimodal accuracy results averaged over 15 cross-validation splits; higher accuracy is better. Bolded results are the best in the whole column and are underlined if differences are significant. Italicized results are tied for the best among their feature type. 95% CI are on average ± 0.5 and never exceed ± 1 for the non-timing features.	86

4.5	Multimodal accuracy results averaged over 15 cross-validation splits. Higher accuracy is better, and accenting follows Table 4.4. 95% CI are on average $\pm.5$ and never exceed $\pm.76$. The best unimodal model ResNet50 is generally outperformed by the multimodal model, Text + Image. User features alone (All User) generally perform better on their own than when they are combined with timing features.	88
4.6	Heldout, out-of-domain task accuracy results; bolded are best.	89
5.1	Dataset statistics: total number of images, average text length in words, and size of the train/test splits we use in §5.7.	106

LIST OF FIGURES

2.1	At training time, we assume we are given a set of multi-image/multi-sentence documents. At test-time, we predict links between individual images and individual sentences within single documents. Because no explicit multimodal annotation is available at training time, we refer to this task as unsupervised.	14
2.2	Sample documents from six of our datasets. Image sets and sentence sets may be truncated due to space constraints. The example from Story-DII is harder than is typical, but we include it to illustrate a point regarding image spread made in §2.5.1. *** denotes text-chunk delimiters present in the original data.	20
2.3	Inter-document objective (AP, $b = 10$) and intra-document AUC increase together during training.	24
2.4	Example test-time graph predictions from AP with $b = 10$. Each subfigure gives the top 5 image/sentence predictions per document, in decreasing order of confidence from left to right. Green edges indicate ground-truth pairs; edge widths show the magnitude of edges in \widehat{M}_i (only positive weights are shown). Examples are selected to be representative: per-document AUC (roughly) matches the average AUC achieved on the corresponding dataset.	30
2.5	Predicted sentences, with cosine similarities, for images in a 100-sentence ImageCLEF Wikipedia article on Mauritius. The first three predictions are reasonable, the last two are not. The third result is particularly good given that only two sentences mention <i>dodos</i> ; for comparison, the object-detection’s choice began “(Mauritian Creole people usually known as ‘Creoles’)”.	31
2.6	Inter-document objective (AP, $b = 10$, hard negative mining) and intra-document AUC during 50 epochs of training for all datasets we consider with ground-truth, intra-document annotations. While there are some interesting discontinuities, e.g., in DII-Stress’s training curves, in general, for a fixed neural architecture/similarity function, better retrieval performance, as measured by the negative-loss computed over the validation set, equates to better intra-document performance, as measured by AUC.	39
3.1	Illustration of a multimodal dense instructional video captioning task (the word “dense” refers to the fact that there are multiple captions per image). Models are given access to both video frames and ASR tokens, and must generate a recipe instruction step for each video segment. The speaker in the video <i>sometimes</i> (but not always) references literal objects and actions.	45

3.2	The AT+Video model. Both the encoder and decoder layers perform cross-modal attention.	55
3.3	Example generations from AT+Video in cases where it performs well, okay, and poorly.	58
3.4	The multimodal model AT+Video produces slightly more diverse captions than its unimodal counterparts.	59
3.5	Per-word classification results using ASR and/or Video features. Each point in the scatterplot represents a different word-type; x-coordinate values show how well a word is predicted by ASR-token features; y-coordinate values show how well a word is predicted by video features. Tables (a)-(d) show word types that are easy, universally difficult, better-predicted-by-ASR, and better-predicted-by-video, respectively.	59
3.6	The performance of the oracle methods increases as they are given access to an increasing number of object types.	64
4.1	Despite being submitted only 13 seconds apart to the subreddit <i>aww</i> , one of these submissions received over 1600 upvotes whereas the other received fewer than 20; the answer is in § 4.3. Images courtesy <i>imgur.com</i> , posted by Reddit users <i>mercurycloud</i> and <i>imsozzy</i>	67
4.2	Proportional popularity of types of Reddit posts over time across all subreddits.	72
4.3	Average score versus time of day (eastern) on <i>aww</i> with 95% CI (red) and activity levels.	72
4.4	Relationship between various measures of time and eventual submission score with 95% confidence intervals.	74
4.5	Relationship between various measures of time and eventual submission score for several subreddits, with 95% CI.	83
4.6	Examples from <i>FoodPorn</i> automatically scored by the ResNet50 model. The top, middle, and bottom rows are sampled from the 99th, 50th, and 1st percentiles of model scores respectively. While lighting effects likely relate to model scores, the underperformance of the color-only classifier and the performance jump when switching from VGG-19 to ResNet50 suggest that this is a rich computer vision task. Images courtesy <i>imgur.com</i>	91
4.7	Examples from one train/test split of <i>aww</i> scored by the ResNet50 model, the unigram model, and the text + image model. The top, middle, and bottom rows are sampled from the 99th, 50th, and 1st percentiles respectively. Images courtesy <i>imgur.com</i>	91

5.1	Demonstration of visual concreteness estimation on an example from the COCO dataset. The degree of visual clustering of textual concepts is measured using a nearest neighbor technique. The concreteness of “dogs” is greater than the concreteness of “beautiful” because images associated with “dogs” are packed tightly into two clusters, while images associated with “beautiful” are spread evenly. ³	99
5.2	Examples of text and images from our new Wiki/BL datasets. . .	105
5.3	Examples of the most and least concrete words/topics from Wiki, COCO, and Flickr, along with example images associated with each highlighted word/topic.	107
5.4	Spearman <u>correlations</u> between human judgment (USF) and our algorithm’s outputs, and dataset frequency. In the case of Flickr/COCO/WIKI our concreteness scores correlate with human judgement to a greater extent than frequency. For BL, neither frequency nor our concreteness measure is correlated with human judgement. ***/**/* := $p < .001/.01/.05$	109
5.5	Concreteness scores versus retrievability (plotted) for each dataset, along with Recall at 1% (in tables, higher is better) for each algorithm combination. Tables give average retrieval performance over 10-fold cross-validation for each combination of NLP/alignment algorithm; the top three performing combinations are bolded. The concreteness versus retrievability curves are plotted for the top-3 performing algorithms, though similar results hold for all algorithms. Our concreteness scores and performance are positively correlated, though the shape of the relationship between the two differs from dataset to dataset (note the differing scales of the y-axes). All results are for RN-ImageNet; the similar I3-OpenImages results are omitted for space reasons.	113
5.6	Wikipedia	117
5.7	British Library	117
5.8	COCO	117
5.9	Flickr	117
5.10	Correlation between word/topic frequency and retrievability for each of the four datasets. Compared to our concreteness measure (see Figure 5.5; note that the while x-axes are different, the y-axes are the same) frequency explains relatively little variance in retrievability.	117

CHAPTER 1

INTRODUCTION AND ORGANIZATION

Today's web plays host not only to an increasing diversity of communities, but also to an increasing diversity of *modalities of communication* as users mix text, images, videos, and audio. Content analysis tools like language and image processing algorithms have significant potential to offer insight into the dynamics of these communities and to enable the development of new tools for improving user experiences. But bridging the gap between content *recognition* (e.g., object detection in images or named-entity recognition in text) and *contextual understanding* remains a challenge.

The work in this dissertation details our efforts towards two related research goals. First, our goal is to **leverage** large, unlabeled web corpora to build computer systems capable of bridging the gap between different modalities; these tools enable a number of promising practical applications. Second, our goal is to **understand** the manner in which visual-textual content is used for communication in online contexts. As we will argue later in this introductory section, these two research goals are closely linked.

1.0.1 What is visual-textual grounding?

From the perspective of the machine learning researcher, visual-textual grounding can be defined as a set of tasks that require machines to understand connections between multiple modes of data, e.g., images and text. Many promising AI applications require understanding connections between multiple data modes. To list a few:

1. **Robot Navigation:** As robots play an increasingly important role in our day-to-day lives, users should ideally be able to interact with them via natural language commands, e.g., “go to the kitchen,” or “pick up the yellow ones” (Matuszek et al., 2012; Artzi and Zettlemoyer, 2013; Anderson et al., 2018). Because robots commonly interpret the physical world through cameras,¹ robot navigation thus frequently depends on a mapping of vision and language.
2. **Alt-text generation for the web:** A longstanding challenge in web design is how visual content can be made accessible to low vision and blind users (Lazar et al., 2007). The nonprofit organization WebAIM² says that “adding alternative text for images is the first principle of web accessibility,” but many web images do not provide this context. Automatic alt-text generation algorithms (Wu et al., 2017a; MacLeod et al., 2017), built to address this challenge at scale, must build a joint representation of images and generated text.
3. **Interactive accessibility tools:** Mobile computing devices, often equipped with cameras, offer a promising means for low vision and blind people to access visual information. Prior work has examined human-in-the-loop solutions for answering visual questions based on photos taken from these devices (Bigham et al., 2010; Brady, 2015), but automated methods might scale better. To this end, machine learning tasks like visual question answering/captioning for people who are blind have been framed (Gurari et al., 2018, 2019, 2020).
4. **Web Video Parsing:** Prior work in human-computer interaction suggests

¹Other modalities are also gaining popularity, e.g., LiDAR (in conjunction with visible spectrum camera data) in the case of self-driving cars (Sun et al., 2019c).

²<https://webaim.org/>

that users of instructional videos have improved experiences when presented with an annotated timeline of subgoals (Margulieux et al., 2012; Kim et al., 2014; Weir et al., 2015). To deploy timeline generation methods at web scale, automatic tools must be used. Recent work in video captioning and action localization e.g., Zhou et al. (2018b); Wang et al. (2019b), has taken strides towards this goal.

5. **Cross-modal Search+Retrieval:** Information needs for users are evolving: an ideal search engine should be able to support both multimodal queries and responses (Jeon et al., 2003; Rasiwasia et al., 2010). Similarly, the types of organizations with the need to organize and index multimedia content continues to expand: the Smithsonian recently released “nearly 3M 2D/3D digital items” from their collections,³ the Associated Press releases 1M images and 70K videos to accompany their news articles annually (The Associated Press, 2020), and the British Library extracted 400K images from historical volumes (mostly) from the 19th century (British Library Labs, 2016)

1.0.2 A brief theoretical case for studying grounding problems

The ability of humans to bridge gaps between different modes of perception is an intrinsic aspect of cognition. The study of capacity for multimodal comprehension has roots in cognitive science, and has been actively pursued since at least the 1970s (Miller and Johnson-Laird, 1976). In stating the well-known symbol grounding problem, Harnad (1990) asks: “How can the semantic interpretation of a formal symbol system be made intrinsic to the system, rather than

³<https://www.si.edu/openaccess>

just parasitic on the meanings in our heads?”

A grounded consideration of language, more broadly, is arguably required for systems capable of general language understanding, as well. Bender and Koller (2020) propose a thought experiment:⁴

Imagine that we were to train a [language model] on all of the well-formed Java code published on Github. The input is only the code. It is not paired with bytecode, nor a compiler, nor sample inputs and outputs for any specific program... We then ask the model to execute a sample program, and expect correct program output.

They argue that such a test is “patently unfair,” because one cannot expect to learn the semantics of a programming language, given access only to the surface code forms (and no sample executions). Extending that logic, some argue that, barring grounding to our physical environment, social contexts, and perceptions, some aspect of meaning will be “out-of-reach” for language processing systems (see Bisk et al. (2020) for a recent survey).

1.1 A Practical Case for Considering Multimodal Web Data

The history of artificial intelligence is fraught with boom-bust cycles of hype and disappointment.⁵ Over its 70+ year history, many predictions about AI have not

⁴One closely related to the “Chinese room” of Searle (1980).

⁵Indeed, an (arguably overly-)optimistic perspective has underlay AI since its founding. In addition to framing a number of important directions in AI (e.g., language processing, computational creativity, etc.) the organizers of the foundational Dartmouth Workshop of 1956 predicted that significant progress could be made on those directions “if a carefully selected group of scientists work on it together for a summer” McCarthy et al. (1955). And yet, even 60 years

come true, and specific methods and algorithms evolve quickly. One of the rare observations that has withstood the test of time (at least so far) is the *unreasonable effectiveness of data*: algorithms trained on large, unlabeled corpora scraped from the web tend to perform well on many benchmark AI tasks. Indeed, Halevy et al. (2009) called web data “the best ally we have” to address many important questions in AI.

Recent performance advances in language processing and computer vision continue to be largely driven by adapting existing algorithms to operate on unstructured, noisy, cheap-to-collect web datasets. To provide a brief (and incomplete) survey of some recent results across various data modalities:

- *Language processing*: Raffel et al. (2019) achieves high performance on a suite of 10 difficult benchmark language understanding tasks (Wang et al., 2019a),⁶ by pretraining a sequence-to-sequence neural network (Sutskever et al., 2014) on 35B tokens from common crawl⁷ web dumps.
- *Image classification*: Mahajan et al. (2018) collect approximately 3.5B Instagram images with user-generated hashtags. They pretrain a slightly older neural architecture (Xie et al., 2017) on these noisy images/tags, and achieve state-of-the-art performance on the ImageNet classification task (Russakovsky et al., 2015).
- *Video understanding*: Miech et al. (2019) pretrain a video/text model using 100M noisy web video clips from YouTube using a cross-modal retrieval

later, through multiple AI winters, many of these problems remain largely unsolved. A more complete history is given in Crevier (1993).

⁶SuperGLUE is a “meta benchmark” encompassing work from Levesque et al. (2011); Poliak et al. (2018); Rudinger et al. (2018); Pilehvar and Camacho-Collados (2019); Bentivogli et al. (2009); Giampiccolo et al. (2007); Bar Haim et al. (2006); Dagan et al. (2006); Zhang et al. (2018b); Khashabi et al. (2018); Roemmele et al. (2011); De Marneffe et al. (2019); Clark et al. (2019).

⁷<https://commoncrawl.org/>

model; they achieve high performance on several video understanding tasks (e.g., retrieval), and even exceeded the upper bound of training with supervision on an action localization task (Zhukov et al., 2019).

While it remains an open question as to whether or not consideration of even multimodal web data can truly lead to grounded language understanding,⁸ from a practical, tool-building perspective, it undoubtedly offers a promising method for training performant algorithms. Thus, one of the research goals discussed here is to **leverage** multimodal web data for constructing such tools.

1.2 Understanding Online Communication

As social interactions increasingly manifest online, the form and importance of digital communication continues to evolve. Natural language processing tools are promising for understanding web community dynamics; prior work has shown, for instance, that word-usage predicts user lifecycle (Danescu-Niculescu-Mizil et al., 2013), that message phrasing significantly impacts the spread of content (Tan et al., 2014) and that language correlates with how controversial a post will eventually become (Hessel and Lee, 2019).

The modern web, however, is not limited to text, as user attention is increasingly dominated by images, videos, etc. (Yu et al., 2011; Rainie et al., 2012; Singer et al., 2014). Indeed, traditional social media platforms like Facebook, Twitter, etc. universally support multimodal content; newer platforms like Snapchat and Instagram, solely focused on multimedia communication, continue to grow

⁸See § 6.0.2 for a fuller discussion.

in popularity. Thus, community moderators have a need to understand nontextual forms of communication from a computational perspective.

1.2.1 Prior Theories of Image-Text Communication

Fortunately, visual-textual communication has been previously studied: many early theories of image-text grounding derive from systemic-functional semiotics (see Barthes (1988); O’toole (1994); Lemke (1998); O’Halloran (2004)), a direction of work closely related to linguistics pursued with explaining language as a social semiotic system. Several taxonomies have been constructed to characterize image-text communications (Martinec and Salway, 2005; Marsh and Domas White, 2003) (e.g., images can “extend” text, images can “re-iterate” text, etc.). Recent efforts have been developed to operationalize some of these ideas into machine learning classifiers (Chen et al., 2015a; Alikhani et al., 2019; Vempala and Preoțiu-Pietro, 2019; Kruk et al., 2019).

Prior work in semiotics suggests that multimodal communication *can* be cross-modally compositional. One popular theory is multimodal *meaning multiplication* (Barthes, 1988) between images and text. Bateman (2014) summarizes:

The idea is that, under the right conditions, the value of a combination of different modes of meaning can be worth more than the information (whatever that might be) that we get from the modes when used alone. In other words, text ‘multiplied by’ images is more than text simply occurring with or alongside images.

Jones et al. (1979) provide experimental evidence of conditional, compositional

interactions between image and text in a humor setting, concluding that “it is the dynamic interplay between picture and caption that describes the multiplicative relationship” between modalities.

While “meaning multiplication” and related theories of communication are promising in theory, computational and statistical evidence of their manifestations are lacking, e.g., Chen et al. (2013b) note that many theories have not “been operationalized into [automated classifiers]” (an observation that largely remains true today).

1.3 The Connection Between Leveraging and Understanding

While **leveraging** data from the web and **understanding** how/why that data was created by users for communicative purposes may seem like disjoint research programs upon first examination, the two are closely connected.

As previously discussed, the manner in which users employ multimodal content for communicative purposes is potentially quite complex. Users may refer to dynamic real-world entities/events, communication may appear in difficult to represent social contexts, and unstated background “commonsense” information may be required for full comprehension. Complicating the setting compared to the text-only case: references, contexts, and commonsense references may require cross-modal reasoning to uncover; different modalities, in theory, may connect and interact via complex and difficult to specify mechanisms. Thus, to understand multimodal web communication, we undoubtedly need better image-text grounding tools capable of performing inferences between and across modalities.

Conversely, the tools that produce the best performance on different cross-modal reasoning benchmarks, in general, are dependent upon pretraining using a large, unlabeled web corpus. At present, pretraining is undertaken either via retrieval-style objectives that assume literal content overlap between visual and textual content (e.g., Miech et al. (2019)), or via curated annotated tasks (e.g., Tan and Bansal (2019)). Better understanding how multimodal content is used in practice will enable the design of more sophisticated unsupervised learning criteria, and, ultimately, more powerful representations for downstream tasks.

In summary, to better leverage multimodal web data, we must reason about its usage in online contexts. And, to better understand web communication, we need tools capable of making more sophisticated inferences compared to current models. Thus, there is a symbiotic relationship between these two research programs, with improvements in one likely leading to improvements in the other.

1.4 Organization and Contributions

The rest of this dissertation is organized into five chapters, detailing four published studies. Two of them are categorized under the **leveraging** heading and two of them are categorized under the **understanding** heading.

1. **Leveraging:** Discovering Visual-Textual Connections in Noisy Web Documents (Chapter 2). To effectively leverage large, visual-textual web corpora, we must develop algorithms that can learn connections between modalities. A majority of image-text grounding methods assume as in-

put a strongly aligned corpus, e.g., images paired with captions that literally describe those images. However, the data frequently found on the web does not have this form. In particular, images may appear in more diverse contexts, e.g., in longer documents containing many sentences that generally will not make literal reference to visual content. We develop algorithms that learn visual-textual grounding from such *multi-image, multi-sentence* documents. At training time, the algorithm must leverage document-level co-occurrence to learn a joint embedding of images and text. At test time, we task the same algorithms with a difficult within-document, single image-single sentence prediction task. We succeed in learning grounded information in this difficult setting, both from a quantitative and qualitative perspective. Our structured prediction methods outperform various baselines, including object detection, a version of our method without structured prediction, etc.

2. **Leveraging:** Generating Captions for Web Videos using Noisy ASR (Chapter 3). Leveraging noisy web data can also improve performance in supervised learning tasks. Here, we examine a generation task, where the goal is to provide captions for web videos. Prior work on this dataset leveraged only the visual content of video frames as input signal. We show significant performance gains when also incorporating the noisy, automatically-generated speech recognition tokens, which aim to capture the literal utterances of speakers in web videos. What’s more, our results suggest that visual content and ASR tokens are complementary for the cooking instructional video corpus we consider: ASR tokens appear to capture fine-grained information that may be difficult to visually distinguish (e.g., “vegetable oil” vs. “olive oil”) whereas visual content captures

unstated background information (e.g., to mix ingredients together, you need a bowl).

3. **Understanding:** Predicting Popularity in Multimodal Communities (Chapter 4). Perhaps the best way to study multimodal communication online is by exploring the *in vivo* dynamics of real communities. In this work, we explore six communities from Reddit, and attempt to understand what types of image-text posts draw community-level attention (or not). However, when studying popularity dynamics, many non-content factors can obfuscate the relationship between content quality and ultimate reception. For example, because of diurnal patterns, *when* one posts significantly influences popularity: posts made at 9AM are at a significant advantage relative to posts made at 11AM. We carefully designed pairing experiments to control for timing (and other important factors), tasking algorithms to predict relative popularity between posts made in quick succession, e.g., within 30 seconds. The resulting models, on average, perform well — generally, they achieve higher accuracy than our estimate of human performance. Later work demonstrated that our models also generalize well to other domains.
4. **Understanding:** Quantifying the Visual Concreteness of Concepts (Chapter 5). When people use images and text to communicate on the web, they are unlikely to use literal descriptions: for example, when posting an image of a cat onto social media, who is likely to say “a cat is sitting on a blue bed”? But what does it mean for a language to be referentially concrete (e.g., describing literally pictured objects/actions) vs. abstract? Our intuition is that some concepts, e.g., “dog,” may be more visually concrete than others, e.g., “beauty.” We propose a quantification of this

intuition based on the feature geometry of the visual-textual input corpus. We not only demonstrate that our operationalization correlates with human judgement, but that our concreteness scores correlate with how well retrieval-style algorithms can learn — more concrete concepts are much easier for automated methods.

In Chapter 6, we present thoughts on two promising directions of future work, and perspectives on ethical implications of studying web data.

CHAPTER 2

LEVERAGING: DISCOVERING VISUAL-TEXTUAL CONNECTIONS IN NOISY WEB DOCUMENTS

2.1 Brief Overview

Images and text co-occur constantly on the web, but explicit links between images and sentences (or other intra-document textual units) are often not present. We present algorithms that discover image-sentence relationships *without* relying on explicit multimodal annotation in training. We experiment on seven datasets of varying difficulty, ranging from documents consisting of groups of images captioned post hoc by crowdworkers to naturally-occurring user-generated multimodal documents. We find that a structured training objective based on identifying whether *collections* of images and sentences co-occur in documents can suffice to predict links between specific sentences and specific images within the *same document* at test time.

The work in this chapter is joint with Lillian Lee and David Mimno, and was published in (Hessel et al., 2019a).

2.2 Introduction

Images and text act as natural complements on the modern web. News stories include photographs, product listings show multiple images providing detail for online shoppers, and Wikipedia pages include maps, diagrams, and pictures. But the exact matching between words and images is often left implicit.

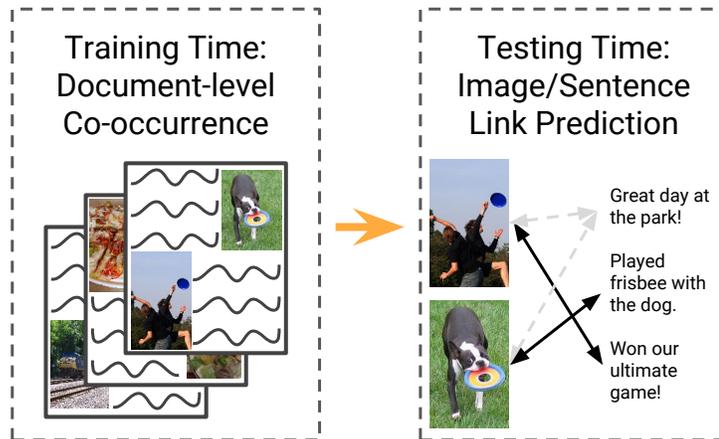


Figure 2.1: At training time, we assume we are given a set of multi-image/multi-sentence documents. At test-time, we predict links between individual images and individual sentences within single documents. Because no explicit multimodal annotation is available at training time, we refer to this task as unsupervised.

Algorithms that identify document-internal connections between specific images and specific passages of text could have both immediate and long-term promise. On the user-experience front, alt-text for vision-impaired users could be produced automatically (Wu et al., 2017b) via intra-document retrieval, and user interfaces could explicitly link images to descriptive sentences, potentially improving the reading experience of sighted users. Also, in terms of improving other applications, the text in multimodal documents can be viewed as a noisy form of image annotation: inferred image-sentence associations can serve as training pairs for vision models, particularly in domains lacking readily-available labeled data.

In this work, we develop *unsupervised* models that learn to identify multimodal within-document links *despite not having access to supervision at the individual image/sentence level during training*. Rather, the training documents contain multiple images and multiple sentences¹ that are not aligned, as illustrated in

¹Any discrete textual unit could be used, such as n-grams or paragraphs. We focus on sen-

Figure 2.1.

Our intra-document setting poses challenges beyond those encountered in the usual cross-modal retrieval framework, wherein “documents” generally consist of a single image associated with a single piece of text, e.g., an image caption. For the longer documents we consider, a sentence may have many corresponding images or no corresponding images, and vice versa. Furthermore, we expect that images *within* documents will be, on average, more similar than images *across* documents, thus making disambiguation more difficult than in the usual one-image/one-sentence case.

Our approach for this difficult setting is ranking-based: we train algorithms to score image collections and sentence collections that truly co-occur more highly than image collections and sentence collections that do not co-occur. The matching functions we consider predict a latent similarity-weighted bipartite graph over a document’s images and sentences; at test time, we evaluate this internal bipartite graph representation learned by our models for the task of intra-document link prediction.

We work with a variety of datasets (one of which we introduce), ranging from concatenations of individually-captioned images to organically-multimodal documents scraped from noisy, user-generated web content.² Despite having no supervision at the individual image-sentence level, our algorithms perform well on the same-document link prediction task. For example, on a visual storytelling dataset, we achieve 90+ AUC, even in the presence of a large number of sentences that do not correspond to any images in the

tences because there exist public sentence-level datasets that we can use for evaluation.

²Data and code: www.cs.cornell.edu/~jhessel/multiretrieval/multiretrieval.html

document. Similarly, for organically-multimodal web data, we are able to surpass object-detection baselines by a wide margin, e.g., for a step-by-step recipe dataset, we improve precision by 20 points on link prediction within documents by leveraging document-level co-occurrence during training.

We conclude by using our algorithm to *discover* links within a Wikipedia image/text dataset that lacks ground-truth image-sentence links. While the predictions are imperfect, the algorithm qualitatively identifies meaningful patterns, such as matching an image of a dodo bird to one of two sentences (out of 100) in the corresponding article that mention “dodo”.

2.3 Task Formulation

We assume as given a set of documents where each document $d_i = \langle S_i, V_i \rangle$ consists of a set S_i of $n_i = |S_i|$ sentences and a set V_i of $m_i = |V_i|$ images.³ For example, d_i could be an article about Paris with $n_i = 100$ sentences and $m_i = 3$ images of, respectively, the Eiffel Tower, the Arc de Triomphe, and a map of Paris. For each d_i , we are to predict an alignment — where some sentences or images may not be aligned to anything — represented by a (potentially sparse) bipartite graph on n_i sentence nodes and m_i image nodes. During training, we are *given no access* to ground-truth image-sentence association graphs, i.e., we do not know *a priori* which images correspond to which sentences, only that all images/sentences in a document co-occur together; this is why we refer to our task as *unsupervised*.

We produce a dense sentence-to-image association matrix $\widehat{M}_i \in \mathbb{R}^{n_i \times m_i}$, in

³Sentences and images can be considered as sequences rather than sets in our framework, but unordered sets are more appropriate for modeling some of the crowd-sourced corpora we used in our experiments.

which each entry is the confidence that there is an (undirected) edge between the corresponding nodes. Applying different thresholding strategies to \widehat{M}_i 's values yields different alignment graphs.

Evaluation. When we have ground-truth alignment graphs for test documents, we evaluate the correctness of the association matrix \widehat{M}_i predicted by our algorithms according to two metrics: AUROC (henceforth AUC) and precision-at- C ($p@C$). AUC , commonly used in evaluating link prediction (see Menon and Elkan (2011)) is the area under the curve of the true-positive/false-positive rate produced by sweeping over possible confidence thresholds; random is 50, perfect is 100. $p@C$ measures the accuracy of the algorithm's most confident C predicted edges (in our case, the most confident edges correspond to the largest entries in \widehat{M}_i). This metric models cases where only a small number of high-confidence predictions need be made per document. We evaluate using $C \in \{1, 5\}$.

2.4 Models

Our algorithm is inspired by work in cross-modal retrieval (Rasiwasia et al., 2010; Hodosh et al., 2013; Costa Pereira et al., 2014a; Kiros et al., 2015). Instead of operating at the level of individual images/sentences, however, our training objective encourages image *sets* and sentence *sets* appearing in the same document to be more similar than non-co-occurring sets.

2.4.1 Alignment Model and Loss Function

We assume that the dimensionality d_{multi} of the multimodal text-image space is predetermined.

Extracting sentence representations We pass the words in each sentence through a 300D word-embedding layer initialized with GoogleNews-pretrained word2vec embeddings (Mikolov et al., 2013). We then pass the sequence of word vectors to a GRU (Cho et al., 2014) and extract and L2-normalize a d_{multi} -dimensional sentence representation from the final hidden state.

Extracting image representations We first compute a representation for each image using a convolutional neural network (CNN).⁴ The network’s output is then mapped via affine projection to $\mathbb{R}^{d_{\text{multi}}}$ and L2-normalized.

Correspondence prediction The result of running the two steps above on an image-set/text-set pair $\langle S, V \rangle$ is $|S| + |V|$ vectors, all in $\mathbb{R}^{d_{\text{multi}}}$. From these, we compute the similarity matrix $\widehat{M} \in \mathbb{R}^{|S| \times |V|}$, where the $(j, k)^{\text{th}}$ entry is the cosine similarity between the j^{th} sentence vector and the k^{th} image vector.

Training Objective We train under the assumption that co-occurring image-set/sentence-set pairs should be more similar than non-co-occurring image-set/sentence-set pairs. We hope that use of this *document-level* objective will produce an \widehat{M}_i offering reasonable *intra-document* information at test time, even though such information is not available at training time.

The training process is modulated by a similarity function $\text{sim}(S, V)$ that mea-

⁴In some experiments, we use pre-computed image features from a pre-trained CNN (Sharif Razavian et al., 2014). In other cases, we fine-tune the full image network. We specify which representation we choose in a later section.

asures the similarity between a set of sentences and a set of images by examining the entries of the individual image/sentence similarity matrix \widehat{M}_i (specific definitions of $\text{sim}(S, V)$ are proposed in §2.4.2). We use a max-margin loss with negative sampling: we iterate through true documents $d_i = \langle S_i, V_i \rangle$, and negatively sample at the document level a set of b sets of images that did not co-occur with S_i , $\mathbb{V}' = \{V'_1, \dots, V'_b\}$, and a set of b sets of sentences that did not co-occur with V_i , $\mathbb{S}' = \{S'_1, \dots, S'_b\}$.

We then compute a loss for $\langle S_i, V_i \rangle$ by comparing the true similarities to the negative-sample similarities. We find that hard-negative mining (Dalal and Triggs, 2005; Schroff et al., 2015; Faghri et al., 2018), the technique of selecting the negative cases that maximally violate the margin within the minibatch, performs better than simple averaging. The loss for a single positive example is:

$$\mathcal{L}(S_i, V_i) = \max_{V' \in \mathbb{V}'} h(\text{sim}(S_i, V_i), \text{sim}(S_i, V')) + \max_{S' \in \mathbb{S}'} h(\text{sim}(S_i, V_i), \text{sim}(S', V_i)) \quad (2.1)$$

for hinge loss $h_\alpha(p, n) = \max(0, \alpha - p + n)$, where we set margin $\alpha = 0.2$ (Kiros et al., 2014; Faghri et al., 2018).

2.4.2 Similarity Functions

We explore several functions for measuring how similar a set of n sentences S is to a set of m images V . All similarity functions convert the matrix $\widehat{M} \in \mathbb{R}^{n \times m}$ corresponding to $\langle S, V \rangle$ into a bipartite graph based on the magnitude of the entries. The functions differ in how they determine which entries \widehat{M}_{ij} correspond to edges and edge weights.

Dense Correspondence (DC). The DC function assumes a dense correspondence between images and sentences; each sentence must be aligned to its most



Figure 2.2: Sample documents from six of our datasets. Image sets and sentence sets may be truncated due to space constraints. The example from Story-DII is harder than is typical, but we include it to illustrate a point regarding image spread made in §2.5.1. *** denotes text-chunk delimiters present in the original data.

similar image, and vice versa, regardless of how small the similarity might be:

$$\text{sim}(S, V) = \frac{1}{n} \sum_{i=0}^n \max_j \widehat{M}_{i,j} + \frac{1}{m} \sum_{j=0}^m \max_i \widehat{M}_{i,j}. \quad (2.2)$$

The underlying assumption of this function can clearly be violated in practice:⁵ sentences can have no image, and images no sentence.

Top-K (TK). Instead of assuming that every sentence has a corresponding image and vice versa, in this function only the top k most likely sentence \Rightarrow image (and image \Rightarrow sentence) edges are aligned. This process mitigates the effect of non-visual sentences by allowing algorithms to align them to no image. We discuss choices of k for particular experimental settings in §2.5.1.

Assignment Problem (AP). We may wish to consider the image-sentence alignment task as a *bipartite linear assignment problem* (Kuhn, 1955), such that each image/sentence in a document has at most one association. Each time we compute $\text{sim}(S, V)$ in the forward pass of our models, we solve the integer programming

⁵Karpathy et al. (2014) §3.3.1 discuss violations in the image fragment/single-word case.

problem of maximizing $\sum_{i,j} \widehat{M}_{ij} x_{ij}$ subject to the constraints:

$$\forall i, \sum_j x_{ij} \leq 1; \forall j, \sum_i x_{ij} \leq 1; \forall i, j, x_{ij} \in \{0, 1\}. \quad (2.3)$$

Despite involving a discrete optimization step, the model remains fully differentiable. Our forward pass uses tensorflow’s python interface, `tf.py_func`, and the `lapjv` implementation of the JV algorithm (Jonker and Volgenant, 1987) to solve the integer program itself. Given the solution x_{ij}^* , we compute (and backpropagate gradients through) the similarity function $\text{sim}(S, V) = (\sum_{i,j} M_{ij} x_{ij}^*) / r$ where r is the number of non-zero x_{ij}^* . Should we want to impose an upper bound k on the number of links, we can add the following additional constraint:⁶ $\sum_{i,j} x_{ij} \leq k(S, V)$. For example, one could set $k(S, V) = \frac{1}{2} \min(|S|, |V|)$.

The JV algorithm’s runtime is $O(\max(n, m)^3)$, and each positive example requires computing similarities for the positive case and the $2b$ negative samples from Eq. 2.1, for a per-example runtime of $O(b \cdot \max(n, m)^3)$. Fortunately, `lapjv` is highly optimized, so despite solving many integer programs, AP often runs *faster* than DC.

2.4.3 Baselines

We construct two baseline similarity functions, as we are not aware of existing models that directly address our task in an unsupervised fashion.

Object Detection. For each image in the document, we use DenseNet169 (Huang et al., 2017b) to find its K most probable ImageNet classes (e.g., “stingray”), and represent the image as the average of the word2vec embeddings of those K labels. We represent each sentence in a document as the mean

⁶Applying (Volgenant, 2004) polynomial-time algorithm.

word2vec embedding of its words. To form the strongest possible baseline, we compute the cosine similarity between all sentence-image pairs to form \widehat{M} for $K \in \{1 \dots 20\}$ and report the variant with the *best post-hoc performance on the test set*.

NoStruct. The similarity functions described in §2.4.2 rely on document-level, structural information, i.e., for a single image in a document, the *other images* in a document affect the overall similarity (and vice versa for sentences). However, this structural information may not be worth incorporating. Thus, we train a baseline that solely relies on single image/single sentence co-occurrence statistics. At training time, we randomly sample a single image and a single sentence from a document, compute the cosine similarity of their vector representations, and treat that value as the document similarity. While the randomly sampled image/sentence will not truly correspond for every sample, we still expect this baseline to produce above-random results when averaged over many iterations, as true correspondences have some (low) probability of being sampled.⁷

2.5 Experiments on Crowdlabeled Data

Our first set of experiments uses four pre-existing datasets created by asking crowdworkers to add sentence-long textual descriptions to images in a collection. Image-sentence alignments are therefore known by construction. We do not use these labels at training time: gold-standard alignments are only used at evaluation time to compare performance between algorithms.⁸ Statistics of these datasets are given in the top half of Table 2.1, and example documents

⁷This probability is equal to the density of the ground-truth, underlying image-sentence association graph.

⁸The supplementary material (§ 2.10) gives more details.

are given in Figure 2.2. Each crowd-labeled dataset is constructed to address a different question about our learning setting.

Q: Is this task even possible? Test: MSCOCO. MSCOCO (Lin et al., 2014) was created by crowdsourced manual captioning of single images. We construct “documents” from this data by first randomly aggregating five image-caption pairs. We then add five “distractor” images with no captions and five “distractor” captions with no images. Thus, a non-distractor image truly corresponds to the single caption that was written about it, and not to the other 9 captions in the document. There are a total of 10 images/sentences per document, and 5 ground-truth image-sentence links. *A priori*, we expect this to be the easiest setting for within-document disambiguation because mismatched images and sentences are completely independent.

Q: What if the images/sentences within a document are similar? Test: Story-DII. Huang et al. (2016) asked crowdworkers to collect subsets of images contained in the same Flickr album (Thomee et al., 2016) that could be arranged into a visual story. In the Story-DII (= “descriptions in isolation”) case, (possibly different) crowdworkers subsequently captioned the images, but only saw each image in isolation. We construct a set of documents from Story-DII so that each contains five images and five sentences. Because images come from the same album, images and captions in our Story-DII “documents” are more similar to each other than those in our MSCOCO “documents.”

Q: What if the sentences are cohesive and refer to each other? Test: Story-SIS. Huang et al. (2016) also presented all the images in a subset from the same Flickr album to crowdworkers simultaneously and asked them to caption the image subsets collectively to form a story (SIS = “story in sequence”). In contrast to

	train/val/test	n_i/m_i (median)	# imgs (unique)	density
MSCOCO	25K/2K/2K	10/10	83K	5%
Story-DII	22K/3K/3K	5/5	47K	20%
Story-SIS	37K/5K/5K	5/5	76K	20%
DII-Stress	22K/3K/3K	50/5	47K	2%
DIY	7K/1K/1K	15/16	154K	8%
RQA	7K/1K/1K	6/8	88K	17%
WIKI	14K/1K/1K	86/5	92K	N/A

Table 2.1: Dataset statistics: top half = crowd-labeled datasets; bottom half = organically-multimodal datasets. *Density* measures the sparsity of the ground truth graph as the number of ground-truth edges divided by the number of possible edges.

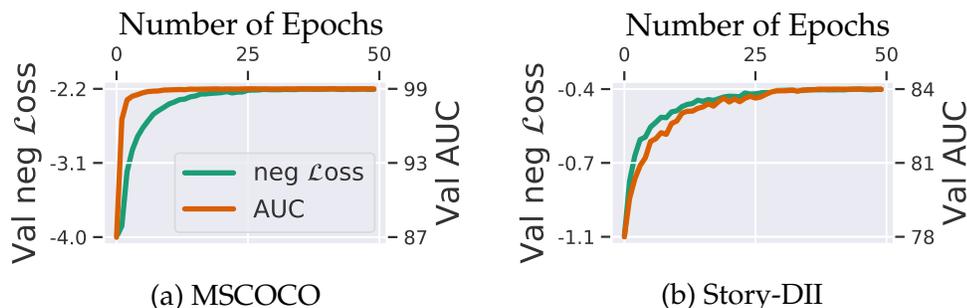


Figure 2.3: Inter-document objective (AP, $b = 10$) and intra-document AUC increase together during training.

Story-DII, the generated sentences are generally not stand-alone descriptions of the corresponding image’s contents, and may, for example, use pronouns to refer to elements from neighboring sentences and images.

Q: What if there are many sentences with no corresponding images? Test: DII-Stress. Because documents often have many sentences that do not directly refer to visual content, we constructed a setting with many more sentences than images. We augment documents from Story-DII with 45 randomly negatively sampled distractor captions. The resulting documents have five images and fifty sentences, where only five sentences truly describe images in the document.

	MSCOCO		Story-DII		Story-SIS		DII-Stress	
	AUC	p@1/p@5	AUC	p@1/p@5	AUC	p@1/p@5	AUC	p@1/p@5
Random	49.7	5.0/4.6	49.4	19.5/19.2	50.0	19.4/19.7	50.0	2.0/2.0
Obj Detect	89.5	67.7/45.9	65.3	50.2/35.2	58.4	40.8/28.6	76.9	25.7/17.5
NoStruct	87.5	50.6/34.6	76.6	60.1/46.2	64.9	43.2/33.7	84.2	21.4/15.6
DC	98.9	93.6/80.1	82.8	71.5/55.5	68.8	51.8/38.6	94.9	64.6/44.8
TK	98.9	93.9/80.1	82.9	71.4/55.5	68.8	50.9/38.7	95.2	65.6/45.3
↳ + $\frac{1}{2}k$	99.0	95.0/81.1	82.0	72.6/54.9	67.6	51.9/38.0	94.7	64.0/43.7
AP	98.7	91.0/78.0	82.6	70.5/55.0	68.5	50.5/38.3	95.3	65.5/45.7
↳ + $\frac{1}{2}k$	98.9	93.9/80.4	81.6	72.4/54.4	67.4	52.1/37.7	94.5	65.0/43.4

Table 2.2: Results for crowd-labeled datasets (similar results for other settings are included in the supplementary material (§ 2.10)). Values are bolded if they are within 1% of the best-in-column performance.

Experiment Protocols. We conduct our evaluations over a single randomly sampled train/dev/test split. For image features, we extract the pre-classification layer of DenseNet169 (Huang et al., 2017b) pretrained on the ImageNet (Russakovsky et al., 2015) classification task, unless otherwise specified. We train with Adam (Kingma and Ba, 2015) using a starting learning rate of .0001 for 50 epochs. We decrease the learning rate by a factor of 5 each time the loss in Eq. 2.1 over the dev set plateaus for more than 3 epochs. We set⁹ $d_{\text{multi}} = 1024$, and apply dropout with $p = .4$. At test time, we use the model checkpoint with the lowest dev error.

2.5.1 Crowdlabeled-Data Results

We tried all combinations of $b \in \{10, 20, 30\}$, $\text{sim} \in \{\text{DC}, \text{TK}, \text{AP}\}$. For TK and AP we set the maximum link threshold k to $\min(S_i, V_i)$ or $\lceil \frac{1}{2} \min(S_i, V_i) \rceil$ (denoted $\frac{1}{2}k$

⁹Anecdotally, we found that values of 256 and 512 produced similar performance in early testing.

in the results table).¹⁰

Table 2.2 shows test-set prediction results for $b = 10$ (results for $b \in \{20, 30\}$ are similar). The retrieval-style objectives we consider encourage algorithms to learn useful within-document representations, and incorporating a structured similarity is beneficial. All our algorithms outperform the strongest baseline (NoStruct) in all cases, e.g., by at least 10 absolute percentage points in p@1 on Story-DII.

We next show, as a sanity check, that our inter-document training objective function (Eq. 2.1) corresponds to intra-document prediction performance (the actual function of interest). Figure 2.3 plots how both functions vary with number of epochs, for two different validation datasets. In general, inter-document performance and intra-document performance rise together during training;¹¹ for a fixed neural architecture, models better at optimizing the inter-document loss in Eq. 2.1 also generally produce better intra-document representations.

In addition, we found that i) DC, despite assuming every sentence corresponds to an image, achieves high performance on DII-Stress, even though 90% of its sentences do not correspond to an image; ii) Allowing AP/TK to make fewer connections (i.e., setting $\frac{1}{2}k$) did not result in significant performance changes, even in the MSCOCO case, where the true number of links (5) was the same as the number of links accounted for by AP/TK+ $\frac{1}{2}k$; and iii) adding topical cohesion (MSCOCO \rightarrow Story-DII) makes the task more difficult, as does adding textual cohesion (Story-DII \rightarrow Story-SIS).

¹⁰For datasets where $m_i = n_i$ and the first choice of definition for k is used, DC and TK are the same. But running the duplicate algorithms anyway provides us with a rough sense of run-to-run variability.

¹¹See the supplementary material (§ 2.10) for plots for all datasets; while the general pattern is the same, some of the training curves exhibit additional interesting patterns.

Models have trouble with the same documents. We calculated AUC for each test document individually. The Spearman correlation between these individual-instance AUC values is very high: of all pairs in DC/TK/AP, over all crowd-labeled datasets at $b=10$, DC vs. AP on MSCOCO had the lowest correlation with $\rho = .89$.

Error analysis: content vs. spread. Why are some instances more difficult to solve for all of our algorithms? We consider two hypotheses. The “content” hypothesis is that some concepts are more difficult for algorithms to find multimodal relationships between: “beauty” may be hard to visualize, whereas “dog” is a concrete concept (Lu et al., 2008; Berg et al., 2010; Parikh and Grauman, 2011; Hessel et al., 2018; Mahajan et al., 2018). The “spread” hypothesis, which we introduce, is that documents with lower diversity among images/sentences may be harder to disambiguate at test time. For example, a document in which all images and all sentences are about horses requires finer-grained distinctions than a document with a horse, a barn, and a tractor. The Story-DII vs. Story-SIS example in Fig. 2.2 illustrates this contrast.

To quantify the spread of a document, we first extract vector representations of each test image/sentence.¹² We then L2-normalize the vectors and compute the mean squared distance to their centroid; higher “spread” values indicate that a document’s sentences/images are more diverse. To quantify the content of a document, for simplicity, we mean-pool the image/sentence representations and reduce to 20 dimensions with PCA.

We first compute an OLS regression of image spread + text spread on test AUC

¹²We use DenseNet169 features for images and mean word2vec for sentences. We don’t use internal model representations as we aim to quantify aspects of the dataset itself.

scores for Story-DII/Story-SIS/DII-Stress¹³ for AP with $b = 10$: 42/23/16% respectively (F-test $p \ll .01$) of the variance in AUC can be explained by the spread hypothesis alone. In general, documents with less diverse content are harder, with image spread explaining more variance than text spread. When adding in the image+text content features, the proportion of AUC variance explained increases to 52/35/38%; thus, for these datasets, both the “content” and “spread” hypotheses independently explain document difficulty, though the relative importance of each varies across datasets.

2.6 Experiments on RQA and DIY

The previous datasets had captions added by crowdworkers for the explicit purpose of aiding research on grounding: for MSCOCO, annotators providing image captions were explicitly instructed to provide literal descriptions and “not describe what a person might say” (Chen et al., 2015b). The manner in which users interact with multimodal content “in the wild” significantly differs from crowd-labeled data: (Marsh and Domas White, 2003) 49-element taxonomy of multimodal relationships (e.g., “decorate”, “reiterate”, “humanize”) observed in 45 web documents highlights the diversity of possible image-text relationships.

We thus consider two datasets (one of which we release ourselves) of organically-multimodal documents scraped from web data, where the original authors created or selected both images and sentences. Statistics of these datasets are given in the bottom half of Table 2.1.

¹³MSCOCO is omitted because the AUC scores are all large.

	RQA		DIY	
	AUC	p@1/p@5	AUC	p@1/p@5
Random	49.4	17.8/16.7	49.8	6.3/6.8
Obj Detect	58.7	25.1/21.5	53.4	17.9/11.8
NoStruct	60.5	33.8/27.0	57.0	13.3/11.8
DC	63.5	38.3/30.6	59.3	20.8/16.1
TK	67.9	44.0/35.8	60.5	21.2/16.0
$\downarrow + \frac{1}{2}k$	68.1	44.5/35.4	56.0	14.1/12.5
AP	69.3	47.3/37.3	61.8	22.5/17.2
$\downarrow + \frac{1}{2}k$	68.7	47.2/36.2	59.4	21.6/15.3

Table 2.3: Performance on the organically-multimodal data; values within 1% of best-in-column are bolded.

RQA. RecipeQA (Yagcioglu et al., 2018) is a question-answering dataset scraped from `instructibles.com` consisting of images/descriptions of food preparation steps; we construct documents by treating each recipe step as a sentence.¹⁴ Users of the Instructibles web interface put images and recipe steps in direct correspondence, which gives us a graph for test time evaluation.

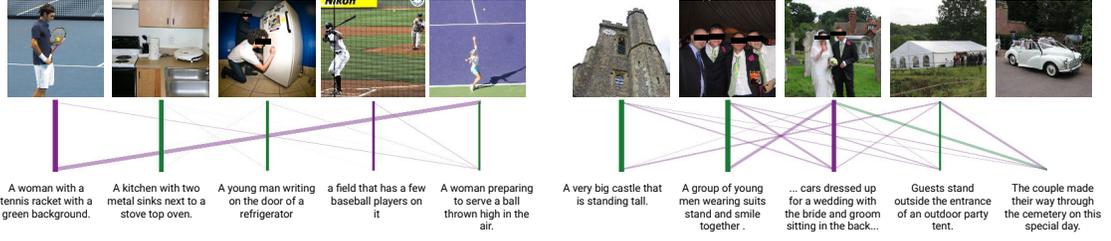
DIY (new). We downloaded a sample of 9K Reddit posts made to the community DIY (“do it yourself”). These posts¹⁵ consist of multiple images that users have taken of the progression of their construction projects, e.g., building a rock climbing wall (see Figure 2.2). Users are encouraged to explicitly annotate individual images with captions,¹⁶ and, for evaluation, we treat a caption written alongside a given image as corresponding to a true link.

We adopt the same experimental protocols as in §2.5, but increase the maximum sentence token-length from 20 to 50; Table 2.3 shows the test-set results.

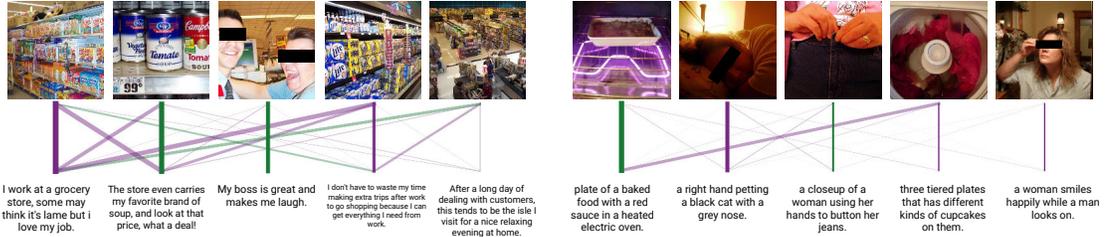
¹⁴Recipe steps have variable length, are often not strictly grammatical sentences, and can contain lists, linebreaks, etc.

¹⁵We required at least 25 upvotes per Reddit post to filter out spam and low-quality submissions.

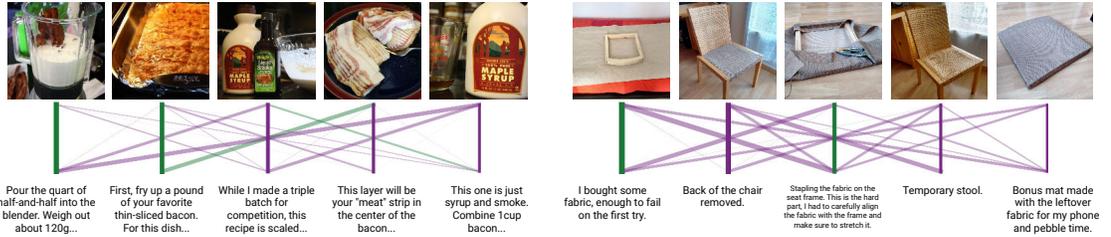
¹⁶As with RQA, DIY captions are not always grammatical.



(a) MSCOCO; 97 AUC, 10 sentences/10 images. (b) Story-DII; 83 AUC, 5 sentences/5 images.



(c) Story-SIS; 70 AUC, 5 sentences/5 images. (d) DII-Stress; 94 AUC, 50 sentences/5 images.



(e) RQA; 70 AUC, 9 sentences/18 images. (f) DIY; 62 AUC, 17 sentences/17 images.

Figure 2.4: Example test-time graph predictions from AP with $b = 10$. Each subfigure gives the top 5 image/sentence predictions per document, in decreasing order of confidence from left to right. Green edges indicate ground-truth pairs; edge widths show the magnitude of edges in \hat{M}_i (only positive weights are shown). Examples are selected to be representative: per-document AUC (roughly) matches the average AUC achieved on the corresponding dataset.

In general, the algorithms we introduce again outperform the NoStruct baseline. In contrast to the crowd-labeled experiments, AP (slightly) outperformed the other algorithms.¹⁷ DIY is the most difficult among the datasets we consider.

To see if the algorithms err on the same instances, we again compute the Spearman correlation ρ between test-instance AUC scores for DC/TK/AP, for

¹⁷This holds even when varying the number of negatively sampled documents; see the supplementary material (§ 2.10).

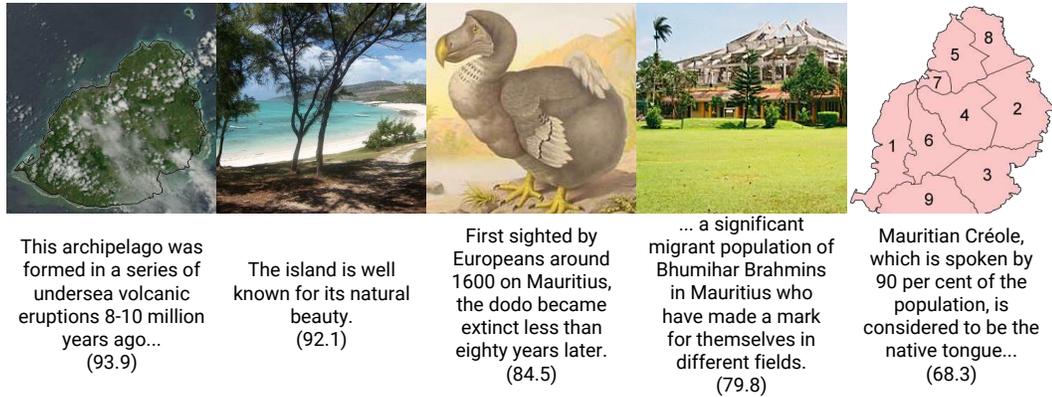


Figure 2.5: Predicted sentences, with cosine similarities, for images in a 100-sentence ImageCLEF Wikipedia article on Mauritius. The first three predictions are reasonable, the last two are not. The third result is particularly good given that only two sentences mention dodos; for comparison, the object-detection’s choice began “(Mauritian Creole people usually known as ‘Creoles’)”.

$b = 10$. We find greater variation in performance on organically-multimodal compared to crowd-labeled data. For example, on RQA, DC and AP have a ρ of only .64. We also repeat the regression on test-instance AUC scores introduced in §2.5.1 with different results; content generally explains more variance than spread, e.g., for AP, for RQA/DIY respectively, only 2/1% is explained by spread alone, but 18/13% is explained by spread+content.

2.7 Qualitative Exploration

To visualize the within-document prediction for document i , we compute \widehat{M}_i and solve the linear assignment problem described in §2.4.2, taking the edges with highest selected weights to be the most confident. Figure 2.4 contains example test predictions (along with \widehat{M}_i) from the datasets with ground-truth annotation. In an effort to provide representative cases, the selected examples have AUC scores close to average performance for their corresponding datasets.

The model mostly succeeds at associating literal objects and their descriptions: tennis players in MSCOCO, castles in Story-DII, a stapler in DIY, and bacon in a blender in RQA. Errors are often justifiable. For example, for the MSCOCO document, the chosen caption for a picture of two people playing baseball accurately describes the image, despite it having been written for a different image and thus counting as an error in our quantitative evaluation. Similarly, for RQA, a container of maple syrup is associated with a caption mentioning “syrup”, which seems reasonable even though the recipe’s author did not link that image/sentence.

In other cases, the algorithm struggles with what part of the image to “pay attention” to. In the Story-DII case (Figure 2.4b), the algorithm erroneously (but arguably justifiably) decides to assign a caption about a bride, groom, and a car to a picture of the couple, instead of to a picture of a vehicle.

For more difficult datasets like Story-SIS (Figure 2.4c), the algorithm struggles with ambiguity. For 2/5 sentences that refer to literal objects/actions (soup cans/laughter), the algorithm works well. The remaining 3 captions are general musings about working at a grocery store that could be matched to any of the three remaining images depicting grocery store aisles. DIY is similarly difficult, as many images/sentences could reasonably be assigned to each other.

WIKI. We also constructed a dataset from English sentence-tokenized Wikipedia articles (not including captions) and their associated images from ImageCLEF2010 (Popescu et al., 2010). In contrast to RQA and DIY, there are no explicit connections between individual images and individual sentences, so we cannot compute AUC or precision, but this corpus represents an important organically-multimodal setting. We follow the same experimental settings as in

§2.5 at training time, but instead of using pre-extracted features, we fine-tune the vision model’s parameters.¹⁸ Examining the predictions of the AP+fine-tuned CNN model trained on WIKI shows many of the model’s predictions to be reasonable. Figure 2.5 shows the model’s 5 most confident predictions on the 100-sentence Wikipedia article about Mauritius, chosen for its high image/text spread.

2.8 Additional related work

Our similarity functions are inspired by work in aligning image fragments, such as object bounding boxes, with portions of sentences without explicit labels (Karpathy et al., 2014; Karpathy and Fei-Fei, 2015; Jiang et al., 2015; Rohrbach et al., 2016; Datta et al., 2019); similar tasks have been addressed in supervised (Plummer et al., 2015) and semi-supervised (Rohrbach et al., 2016) settings. Our models operate at the larger granularity of entire images/sentences. Integer programs like AP have been used to align visual and textual content in videos, e.g., Bojanowski et al. (2015)

Prior work has addressed the task of identifying objects in single images that are referred to by natural language descriptions (Mitchell et al., 2010, 2013; Kazemzadeh et al., 2014; Karpathy et al., 2014; Plummer et al., 2015; Hu et al., 2016c; Rohrbach et al., 2016; Nagaraja et al., 2016; Hu et al., 2016b; Yu et al., 2016; Peyre et al., 2017; Margffoy-Tuay et al., 2018). In general, a supervised approach is taken (Mao et al., 2016; Krishna et al., 2017b; Johnson et al., 2017).

¹⁸In comparable settings, fine-tuning the vision CNN yields $\approx 20\%$ better performance in terms of the loss in Equation 2.1 computed over the validation/test sets. For memory reasons, we switched from DenseNet169 to NASNetSmall (Zoph et al., 2018); additional details are in the supplementary material (§ 2.10).

Related tasks involving multi-image/multi-sentence data include: generating captions/stories for image streams or videos (Park and Kim, 2015; Huang et al., 2016; Shin et al., 2016; Liu et al., 2017), sorting aligned (image, caption) pairs into stories (Agrawal et al., 2016), image/textual cloze tasks (Iyyer et al., 2017; Yagcioglu et al., 2018), augmentation of Wikipedia articles with 3D models (Russell et al., 2013), question-answering (Kembhavi et al., 2017), and aligning books with their film adaptations (Zhu et al., 2015); these tasks are usually supervised, or rely on a search engine.

2.9 Conclusion

We have demonstrated that a family of models for learning fine-grained image-sentence links *within documents* can produce good test-time results even if only given access to document-level co-occurrence at training time.

Future work could incorporate better models of sequence within document context (Kim et al., 2015; Alikhani and Stone, 2018). While using structured loss functions improved performance, image and sentence *representations* themselves have no awareness of neighboring images/sentences; this information should prove useful if modeled appropriately.¹⁹

¹⁹Attempts to incorporate document context information by passing the word-level RNN's output through a sentence-level RNN (Li et al., 2015; Yang et al., 2016) did not improve performance.

2.10 Supplementary Material

2.10.1 Data preprocessing details

MSCOCO. We downloaded the train/val 2017 images, and the train/val annotations from 2014 and 2017 from the MSCOCO website (but create our own training and validation splits). Then, we randomly designate half of the images as “true” images (which will eventually be paired with their true captions in documents) and half of the images as “fake” images, which will not be paired with their true captions in documents. Then, we randomly group all true images into groups of five, and all fake images into groups of five. Then, we pair each real-image set with a fake image set, and divide the resulting groups of 10 images into train/validation/test splits. Then, for each of the training/validation/testing document sets independently, for each document, we create (usually) 5 true versions of each document (for testing and validation, we only sample a single version of each document, and do not consider the alternate true captions provided by MSCOCO) because (in general) each MSCOCO image comes with 5 caption annotations. For each of these true versions, we randomly sample captions from a pool of all captions written on all images not in that document (but from the train/validation/test pools independently, so that there is no overlap between these sets, except in cases where captions happen to be identical). Then, we shuffle the sampled captions for each version. The result is 4968/1655/1655 train/validation/test documents, but each training “document” generally consists of 5 versions because MSCOCO images generally come with 5 captions each.

Story-DII/Story-SIS. We downloaded the Story-DII/Story-SIS train/val/test

splits along with all images from the Visual Storytelling Dataset website;²⁰ we preserve these splits for our train/validation/test sets. DII stories have multiple annotations per fixed image set, whereas SIS stories have multiple annotations per Flickr album, as human annotators were allowed to select images for their story from all the images within an album. We discard any story with any invalid or missing image (the FAQ page on the data download website mentions that images may be missing because users deleted them).

DII-Stress. We augmented the documents from Story-DII with 45 distractor captions (i.e., captions that were not written about any of the images in the document) selected uniformly at random. To preserve train/validation/test splits, we limit these uniform selections to within-split samples, i.e., training document distractor captions are sampled only from training documents.

RQA. We download the train and validation questions (29.6K/3.5K) and extract the “context” of each question, which consists of a list of recipe steps and their associated images; without filtering, there are 8.1K unique recipes in the training set, and 983 unique recipes in the validation data. We also download the training/validation images provided. We treat the provided validation split as the test data.

We concatenate the title and the body of the step (separating them with a space). We discard recipe steps that do not contain any tokens, and discard recipes for which there are no images that correspond to steps (e.g., if the only steps for which there were images contained empty text). Then, we reserve training recipes to act as our validation split. Then, we discard all recipes with fewer than 2 images/recipe steps. The result is 6502/946/878 train-

²⁰<http://visionandlanguage.net/VIST/>

ing/validation/test recipes, with 69K total images. The sizes of the documents are: mean/median/max number of images: 11/8/93; and mean/median/max number of sentences: 7/6/20.

DIY. We downloaded all the submissions on pushshift.io’s files page from Jan. 2013-Oct. 2018. We looped over all of them and found the ones available made to the subreddit “DIY,” for 241K posts. Then, we discard posts with score less than 25. While the semantics of the Reddit “score” field have changed over time,²¹ we intend for this filtration step to act as a basic spam filter. We only consider link submissions to imgur urls with “/a/” in the url, indicating that the imgur link is an album, rather than a single image. We then scrape the associated imgur album page and search for all “div” html fields that are “post-image-container,” and extract both the image associated with that field and its associated caption, if it’s not empty; users may leave image captions empty, but may not upload a caption without an associated image. We ignore imgur albums with no “post-image-container” fields. There are 13K documents after this step. We attempt to scrape all images for these documents, discarding gifs and invalid images for simplicity, resulting in 295K images.

Next, we search for any image duplicates using `findimagedupes` (<https://gitlab.com/opennota/findimagedupes>) with a neighbor threshold of 3. We discard any documents with any duplicate images. Then, we discard all documents without at least 2 image captions with at least 5 tokens, and discard documents without at least 2 valid images. Because a small number of documents are quite long, we discard documents with more than 40 images or more than 40 captions.²² We split the remaining documents into 6.8K/1K/1K

²¹Other confounding factors: Reddit has become more popular over time, DIY has likely changed in popularity, etc.

²²At this step, its possible for there to be more captions than images in a document, e.g.,

train/validation/test documents. Between these documents, there are 154K unique images. The sizes of the documents are: mean/median/max number of images: 17.4/16.0/40; mean/median/max number of sentences: 16.4/15.0/40.

WIKI. We downloaded the English-language subset of the ImageClef 2011 Wikipedia retrieval data as a starting point (<https://www.imageclef.org/wikidata>). This dataset contains the full text of Wikipedia articles, alongside a list of images in each article. We then stripped out wiki formatting, and used Spacy’s (<https://spacy.io/>) English-sentence tokenizer to split documents into sentences (the resulting sentence tokenization is imperfect, but sufficient). We keep only the first 100 identified sentences in a document. We discarded documents with fewer than 10 sentences, and documents with fewer than 3 images. The result is 16K articles, for which we used a 14K/1K/1K train/validation/test split. For the results discussed in the paper, we explore same-document predictions on training documents using a model checkpoint with low validation error. The sizes of the documents are: mean/median/max number of images: 6/5/108, mean/median/max number of sentences: 72/86/100.

Download. All datasets are available for download: www.cs.cornell.edu/~jhessel/multiretrieval/multiretrieval.html

2.10.2 WIKI Fine-tuning Details

We experiment with fine-tuning the parameters of our image model for the organically-multimodal data, as an alternative to extracting features from a because we discard animated gifs that may have been associated with captions.

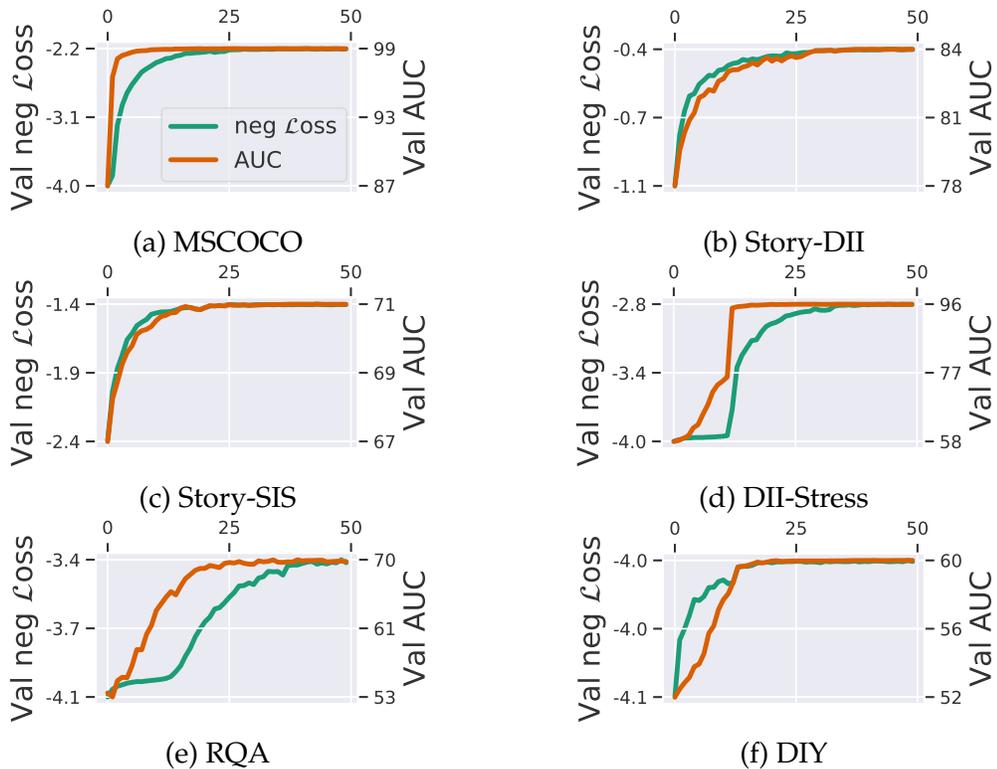


Figure 2.6: Inter-document objective (AP , $b = 10$, hard negative mining) and intra-document AUC during 50 epochs of training for all datasets we consider with ground-truth, intra-document annotations. While there are some interesting discontinuities, e.g., in DII-Stress’s training curves, in general, for a fixed neural architecture/similarity function, better retrieval performance, as measured by the negative-loss computed over the validation set, equates to better intra-document performance, as measured by AUC .

pretrained network. However, given that hundreds of images and sentences need to fit in GPU memory for each batch (we worked with a single GPU with 12GB of RAM), we needed to switch our CNN from DenseNet169 to one with a smaller memory footprint; we chose NASNetSmall. But even so, we still require a word-embedding matrix and a 1024-dimensional GRU in memory. Hence, additionally, at training time, for documents with more than 10 images/sentences, we randomly downsample images/sentences to a set of 10 (though at validation and test time, longer documents are kept intact). This subsampling process ensures that at most 110 images are in GPU memory at a time (for 10 negative

samples per positive sample). When training the CNN, we also perform random data augmentation to help regularize. We first resize images to 256 by 256, and, at training time, perform the following data augmentation: random horizontal flipping, up to 20 degree random image rotation, and a random crop to 224 by 224. At validation/test time, we use a center crop (with no rotations or flips).

We trained models with AP using fixed, NASNetSmall pre-extracted features, and compared those models to ones where we fine-tuned the additional 5M CNN parameters. The resulting *test* AUC/negative-loss ($-\mathcal{L}$) values are:

	RQA		DIY		WIKI	
	AUC	$-\mathcal{L}$	AUC	$-\mathcal{L}$	AUC	$-\mathcal{L}$
Fixed CNN	67.6	-.37	60.9	-.37	N/A	-.26
Finetuned CNN	65.7	-.40	57.9	-.39	N/A	-.21

Thus, we did not observe intra-document performance increases with fine-tuning for DIY and RQA for the experiment settings we consider. However, on WIKI, for negative-training-loss (the only metric we can compute on this no-ground-truth dataset), fine-tuning performed better.²³ Since Figure 2.6 demonstrates that, for a fixed architecture and for datasets where AUC can be computed, AUC and (the negative of) training loss rise together, we expect that fine-tuning is beneficial for WIKI.

²³Fine-tuning NASNetSmall also beat using DenseNet169 extracted features.

	MSCOCO		Story-DII		Story-SIS		DII-Stress	
	AUC	p@1/p@5	AUC	p@1/p@5	AUC	p@1/p@5	AUC	p@1/p@5
Random	49.7	5.0/4.6	49.4	19.5/19.2	50.0	19.4/19.7	50.0	2.0/2.0
Obj Detect	89.5	67.7/45.9	65.3	50.2/35.2	58.4	40.8/28.6	76.9	25.7/17.5
NoStruct	88.3	53.4/35.8	76.6	60.4/46.2	64.9	43.3/33.8	84.2	21.4/15.6
NoStruct+ hard neg	51.8	8.3/5.9	75.9	63.0/45.0	63.3	45.1/31.9	51.9	4.3/3.1
DC	98.8	92.0/78.6	81.8	69.1/53.7	68.0	49.7/37.6	93.8	58.3/40.1
DC+ hard neg	98.9	93.1/79.9	82.9	71.9/55.7	68.8	52.2/38.7	95.0	65.2/44.9
TK	98.8	92.1/78.6	81.8	69.6/53.8	68.0	49.7/37.6	94.4	60.2/42.2
TK+ hard neg	98.9	93.9/80.0	82.8	71.5/55.7	68.8	51.8/38.5	95.2	65.2/45.3
TK+ hard neg+ $\frac{1}{2}k$	99.0	95.0/81.4	81.9	71.4/54.5	67.6	51.5/37.8	94.7	64.5/43.4
AP	98.5	87.6/75.3	81.7	68.3/53.5	67.3	47.1/36.6	93.5	58.3/39.7
AP+ hard neg	98.7	91.1/77.9	82.6	70.7/55.0	68.6	50.6/38.3	95.4	65.4/45.5
AP+ hard neg+ $\frac{1}{2}k$	98.9	94.1/80.7	81.5	72.2/54.2	67.4	51.9/37.7	94.6	64.7/43.7

Table 2.4: Results for crowd-labeled data with ground-truth annotation with $b = 20$ negative samples.

2.10.3 Additional Results

Tables containing our full results are given in Tables 2.4, 2.5, 2.6, and 2.7. Compared to the results presented in the paper, here we explicitly compare additional hyperparameter configurations. Specifically: we show results for $b = 10, 20, 30$ negative samples (the main paper just shows $b = 10$) and compare using hard negative mining vs. not using hard negatives (the main paper just shows hard negative mining results, e.g., “AP+hard neg” in these tables is the same as the “AP” described in the main paper). In general, hard negative mining improves performance, and the number of negative samples doesn’t greatly affect performance in the range we examined.

	MSCOCO		Story-DII		Story-SIS		DII-Stress	
	AUC	p@1/p@5	AUC	p@1/p@5	AUC	p@1/p@5	AUC	p@1/p@5
Random	49.7	5.0/4.6	49.4	19.5/19.2	50.0	19.4/19.7	50.0	2.0/2.0
Obj Detect	89.5	67.7/45.9	65.3	50.2/35.2	58.4	40.8/28.6	76.9	25.7/17.5
NoStruct	87.5	50.8/34.7	76.6	59.9/46.2	64.9	43.4/33.7	84.1	21.3/15.6
NoStruct+ hard neg	52.0	10.3/6.0	75.9	63.0/45.0	63.0	44.5/31.5	51.8	4.0/2.9
DC	98.8	92.0/78.7	82.2	70.5/54.6	68.0	49.7/37.7	93.9	58.6/40.3
DC+ hard neg	98.9	93.4/79.9	82.8	71.3/55.5	68.8	52.1/38.6	95.0	63.8/44.5
TK	98.8	91.6/78.7	81.8	69.5/53.9	68.0	49.9/37.7	94.4	60.5/42.4
TK+ hard neg	98.9	93.3/80.0	82.8	71.4/55.7	68.8	51.0/38.6	95.2	65.3/45.7
TK+ hard neg+ $\frac{1}{2}k$	99.0	95.2/81.5	82.1	73.1/55.1	67.7	51.9/37.8	94.7	64.2/43.6
AP	98.5	87.3/75.4	81.7	67.7/53.4	67.3	47.1/36.6	93.4	57.2/39.8
AP+ hard neg	98.7	91.2/78.0	82.6	71.1/55.0	68.5	50.3/38.2	95.3	65.3/45.6
AP+ hard neg+ $\frac{1}{2}k$	98.9	94.1/80.5	81.6	72.8/54.4	67.4	51.8/37.8	94.4	64.3/43.2

Table 2.5: Results for crowd-labeled data with $b = 30$ negative samples.

	RQA		DIY	
	AUC	p@1/p@5	AUC	p@1/p@5
Random	49.4	17.8/16.7	49.8	6.3/6.8
Obj Detect	58.7	25.1/21.5	53.4	17.9/11.8
NoStruct	60.5	34.3/26.8	56.9	13.8/12.2
NoStruct+ hard neg	60.1	35.0/26.7	56.3	15.0/12.5
DC	67.1	43.8/34.9	59.5	19.3/15.2
DC+ hard neg	63.4	36.6/31.0	59.3	21.0/16.0
TK	65.2	41.6/33.1	60.0	20.4/15.5
TK+ hard neg	67.9	45.2/36.0	60.5	20.3/16.2
TK+ hard neg+ $\frac{1}{2}k$	67.7	44.4/35.0	56.1	14.8/12.0
AP	66.9	37.8/34.2	59.1	16.9/13.9
AP+ hard neg	69.4	45.9/37.8	61.9	23.3/17.9
AP+ hard neg+ $\frac{1}{2}k$	68.5	44.9/36.4	59.6	21.7/15.7

Table 2.6: Results for organically-multimodal data with ground-truth annotation with $b = 20$ negative samples.

	RQA		DIY	
	AUC	p@1/p@5	AUC	p@1/p@5
Random	49.4	17.8/16.7	49.8	6.3/6.8
Obj Detect	58.7	25.1/21.5	53.4	17.9/11.8
NoStruct	60.4	34.5/26.7	56.9	13.3/11.9
NoStruct+ hard neg	59.7	31.8/27.0	55.9	14.7/12.4
DC	66.7	42.7/34.1	59.5	18.9/14.7
DC+ hard neg	63.5	37.6/30.6	59.4	20.8/16.4
TK	65.3	41.2/32.8	60.1	20.0/15.9
TK+ hard neg	68.0	44.0/36.2	60.5	21.4/16.1
TK+ hard neg+ $\frac{1}{2}k$	67.8	43.2/35.1	57.3	19.1/13.5
AP	66.5	41.0/33.8	59.2	15.7/14.0
AP+ hard neg	69.3	47.5/37.4	61.9	24.4/17.8
AP+ hard neg+ $\frac{1}{2}k$	68.7	45.2/36.2	59.4	22.0/15.7

Table 2.7: Results for organically-multimodal data with $b = 30$ negative samples.

CHAPTER 3
LEVERAGING: GENERATING CAPTIONS FOR WEB VIDEOS USING
NOISY ASR

3.1 Brief Overview

Instructional videos get high traffic on video sharing platforms, and prior work suggests that providing time-stamped subtask annotations (e.g., “heat the oil in the pan”) improves user experiences. However, current automatic annotation methods based on visual features alone perform only slightly better than constant prediction. Taking cues from prior work, we show that we can improve performance significantly by considering automatic speech recognition (ASR) tokens as input. Furthermore, jointly modeling ASR tokens and visual features results in higher performance compared to training individually on either modality. We find that unstated background information is better explained by visual features, whereas fine-grained distinctions (e.g., “add oil” vs. “add olive oil”) are disambiguated more easily via ASR tokens.

The work in this chapter is joint with Bo Pang, Zhenhai Zhu, and Radu Soricut, and was published in Hessel et al. (2019b).

3.2 Introduction

Instructional videos increasingly dominate user attention on online video platforms. For example, 86% of YouTube users report using the platform often to learn new things, and 70% of users report using videos to solve problems related

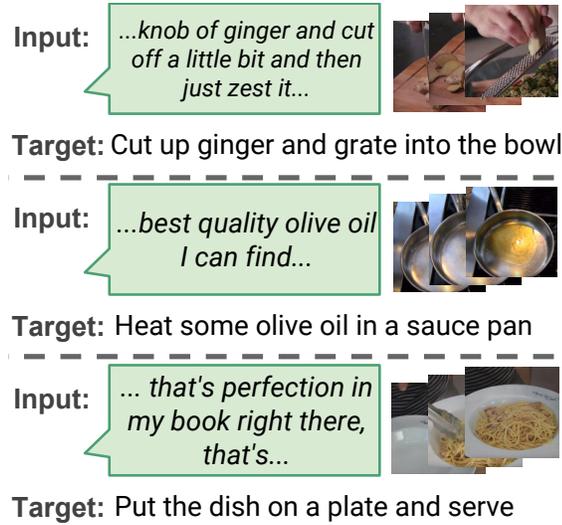


Figure 3.1: Illustration of a multimodal dense instructional video captioning task (the word “dense” refers to the fact that there are multiple captions per image). Models are given access to both video frames and ASR tokens, and must generate a recipe instruction step for each video segment. The speaker in the video *sometimes* (but not always) references literal objects and actions.

to work, school, or hobbies (O’Neil-Hart, 2018).

Prior work in user experience has investigated the best way of presenting instructional videos to users. Kim et al. (2014), for example, compare two options; first: presenting users with the video alone, and second: presenting the video with an additional *structured* representation, including a timeline populated with task subgoals. Users interacting with the structured video representation reported higher satisfaction, and external judges rated the work they completed using the videos as having higher quality. Margulieux et al. (2012) and Weir et al. (2015) similarly find that presenting explicit subgoals alongside how-to videos improves user experiences. Thus, presenting instructional videos with additional structured annotations is likely to benefit users.

These studies rely on human annotation of time-stamped subtask goals, e.g., time-stamped captions created through crowdsourcing. However, human-in-

the-loop annotation is infeasible to deploy for popular video sharing platforms like YouTube that receive hundreds of hours of uploads per minute. In this work, we address the task of *automatically* producing captions for instructional videos at the level of video segments. Ideally, generated captions provide a literal, imperative description of the procedural step occurring for a given video segment, e.g., in the cooking context we consider, “add the oil to the pan.”

Producing segment-level captions is a sub-task of dense video captioning, where prior work has mostly focused on visual-only models. Dense captioning is a difficult task, particularly in the instructional video domain, as fine-grained distinctions may be difficult or impossible to make with visual features alone. Visual information can be ambiguous (e.g., distinguishing between “olive oil” vs. “vegetable oil”) or incomplete (e.g., preparation steps may occur off-camera). In our study, a first important finding is that, for the dataset considered, current state-of-the-art, visual-features-only models only slightly outperform a constant prediction baseline, e.g., by 1.5 BLEU/METEOR points.

To improve performance in this difficult setting, we consider the *automatic speech recognition* (ASR) tokens generated by YouTube. These publicly available tokens are an ASR model’s attempts to map words spoken in videos into text. However, while a promising potential source for signal, it is not always trivial to transform even accurate ASR into the desired imperative target: while there are cases of clear correspondence between the literal actions in the video and the ASR tokens, in other cases, the mapping is imperfect (Fig. 3.1). For example, when finishing a dish, a user says “that’s perfection in my book right there” rather than “put the dish on a plate and serve.” There are also cases where no ASR tokens are available at all.

Despite these potential difficulties, previous work has demonstrated that ASR can be informative in a variety of instructional video understanding tasks (Naim et al., 2014, 2015; Malmaud et al., 2015; Sener et al., 2015; Alayrac et al., 2016; Huang et al., 2017a); though less work has focused on instructional caption *generation*, which is known to be difficult and sensitive to input perturbations (Chen et al., 2018).

We find that incorporating ASR-token-based features significantly improves performance over visual-features-only models (e.g., CIDEr improves $0.53 \Rightarrow 1.0$, BLEU-4 improves $4.3 \Rightarrow 8.5$). We also show that *combining* ASR tokens and visual features results in the highest performing models, suggesting that the modalities contain complementary information.

We conclude by asking: what information is captured by the visual features that *is not* captured by the ASR tokens (and vice versa)? Auxiliary experiments examining performance of models in predicting the presence/absence of individual word types suggest that visual signals are superior for identifying unspoken, implicit aspects of scenes; for instance, in order to mix ingredients, they must be placed in a bowl — and although bowls are often visually present in the scene, “bowl” is often not explicitly mentioned by the speaker. Conversely, ASR features readily disambiguate between fine-grained entities, e.g., “olive oil” vs. “vegetable oil”, a task that is difficult (and sometimes impossible) for visual features alone.

3.3 Related Work

Narrated instructional videos. While several works have matched audio and video signals in an unconstrained setting (Arandjelovic and Zisserman, 2017; Tian et al., 2018), our work builds upon previous efforts to utilize accompanying speech signals to understand online *instructional* videos, specifically. Several projects focus on learning video-instruction alignments, and match a fixed set of instructions to temporal video segments (Regneri et al., 2013; Naim et al., 2015; Malmaud et al., 2015; Hendricks et al., 2017; Kuehne et al., 2017). Another line of previous work uses speech to extract and align language fragments, e.g., verb-noun pairs, with instructional videos (Gupta and Mooney, 2010; Motwani and Mooney, 2012; Alayrac et al., 2016; Huang et al., 2017a, 2018; Hahn et al., 2018). Sener et al. (2015), as part of their parsing pipeline, train a 3-gram language model on segmented ASR token inputs to produce recipe steps.

Dense Video Captioning. Recent work in computer vision addresses dense video captioning (Krishna et al., 2017a; Li et al., 2018; Wang et al., 2018), a supervised task that involves (i) segmenting the input video, and, (ii) generating a natural language description for each segment. Here, we focus on the second subtask of generating descriptions given a ground-truth segmentation; this setting isolates the language generation part of the modeling process.¹ Most related to the present work are several dense captioning approaches that have been applied to instructional videos (Zhou et al., 2018b,c). Zhou et al. (2018c) achieve state-of-the-art performance on the dataset we consider; their model is video-only, and combines a region proposal network (Ren et al., 2015) and a Transformer (Vaswani et al., 2017) decoder.

¹We find that state-of-the-art models perform poorly even for just this subtask (see § 3.4.2), so we reserve the full task for future work.

Multimodal Video Captioning. Several works have employed multimodal signals to caption the MSR-VTT dataset (Xu et al., 2016), which consists of 2K video clips from 20 general categories (e.g., “news”, “sports”) with an average duration of 10 seconds per clip. In particular, Ramanishka et al. (2016); Xu et al. (2017); Hori et al. (2017); Shen et al. (2017); Chuang et al. (2017); Hao et al. (2018) all report small performance gains when incorporating audio features on top of visual features. However, we suspect that the instructional video domain is significantly different than MSR-VTT (where the audio information does not necessarily correspond to human speech), as we find that ASR-only models significantly surpass the state-of-the-art video model in our case.

3.4 Dataset

We focus on YouCook2 (Zhou et al., 2018b), the largest human-captioned dataset of instructional videos publicly available.² It contains 2000 YouTube cooking videos, for a total of 176 hours, and spans 89 different recipes. Each video averages 5.26 minutes, and is annotated with an average of 7.7 temporal segments (i.e., start/end points) corresponding to semantically distinct recipe steps. Each segment is associated with an imperative caption, e.g., “add the oil to the pan”, for an average of 8.8 words per caption.

At the time of analysis (June 2018), over 25% of the YouCook2 videos had been removed from YouTube, and therefore we do not consider them. As a result, all our experiments operate on a *subset* of the YouCook2 data. While this makes direct comparison with previous and future work more difficult, our

²How2 (Sanabria et al., 2018) tackles the different task of predicting video uploader-provided descriptions/captions, which are not always appropriate summarizations.

performance metrics can be viewed as lower bounds, as they are trained on less data compared to, e.g., Zhou et al. (2018c). Unless noted otherwise, our analyses are conducted over 1.4K videos and the 10.6K annotated segments contained therein.

3.4.1 A Closer Look at ASR tokens

We collected the ASR tokens automatically generated by YouTube (available through the YouTube Data API³ with trackKind = ASR), which are then mapped to their temporally corresponding video segments. We start by asking the following questions: How much narration do users provide for instructional videos? And: can YouTube’s ASR system detect that speech?

Not surprisingly, speakers in videos tend to be more verbose than the annotated groundtruth captions: we find the length distribution of ASR tokens per segment to be roughly log-normal, with mean/median length being 42/28 tokens respectively (compared to a mean of 9 tokens/segment for captions). Over the 10.6K available segments, only 1.6% of them have zero associated tokens. Furthermore, based on automatic language identification provided by the YouTube API and some manual verification, we estimated that less than 1% of videos contain completely non-English speech (but we do not discard them from our experiments).

We also investigate the words-per-minute (WPM) ratio based on the video segment length. The mean value of 134 WPM is slightly lower than, but comparable to, previously reported figures of English speaking rates (Yuan et al., 2006),

³<https://developers.google.com/youtube/v3/docs/captions>

which indicates that, for this set of video segments, words are being detected at rates comparable to everyday English speech.

3.4.2 A Closer Look at the Generation Task

To better understand the generation task, we computed lower and upper bounds for generation performance using a constant-prediction baseline and human performance, respectively.

Lower bound: constant. For all segments at test time, we predict “heat some oil in a pan and add salt and pepper to the pan and stir.” This sentence is constructed by examining the most common n-grams in the corpus and pasting them together.

Upper bound: human estimate. We conducted a small-scale experiment to estimate human performance for the segment-level captioning task. Two of the authors of this paper, after being trained on segment-level captions from three videos, attempted to mirror that style of annotation for the segments of 20 randomly sampled videos, totalling over 140 segment annotations each.⁴ Both human annotators report low-confidence with the task; in particular, they found it difficult to maintain a consistent level of specificity in terms of how many factual details to include (e.g., “mix together” vs. “mix the peppers and mushrooms together.”)

Results: We compute corpus-level performance statistics using four standard generation evaluation metrics: ROUGE-L (Lin, 2004; Lin and Och,

⁴These preliminary experiments are not meant to provide a definitive, exact measure of inter-annotator agreement.

2004), CIDEr (Vedantam et al., 2015), BLEU-4 (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) (higher is better in all cases). ROUGE-L and BLEU-4 are based on n-gram overlaps. ROUGE-L automatically determines the longest common subsequence of tokens using a dynamic programming algorithm, and then computes an F-score after normalizing by the reference and candidate translation lengths. BLEU-4 is a modified form of precision that computes reference/prediction overlaps, and discounts the resulting score with a length penalty. METEOR and CIDEr are more complicated; the former computes similarity scores based on a predicted word-word alignment, while the later is motivated by efforts to explicitly compare against consensus in cases where there are multiple references (though the scoring method works for single-reference cases, too). The appendix of Vedantam et al. (2015) gives an excellent description of the detailed computational process of each of these scoring methods.

Note that our evaluation is micro-averaged at the segment level, and differs slightly from prior work on this dataset, which has mostly reported metrics macro-averaged at the video level. We switched the evaluation because some metrics like BLEU-4 exhibit undesirable sparsity artifacts when macro-averaging, e.g., any video without a correct 4-gram gets a zero BLEU score, even if there are many 1/2/3-grams correct. Segment-level averaging, the standard evaluation practice in fields like machine translation, is insensitive to this sparsity concern, and (we believe) provides a more robust perspective on performance.

This comparison highlights the gap that remains between the simplest possible baseline, several computer vision based models, and (roughly) how well humans perform at this task. Given that Sun et al. (2019a) is a highly tuned computer vision model transfer learned from a corpus of over 300K cooking videos,

	BLEU-4	METEOR	ROUGE-L	CIDEr
Constant Prediction	2.70	10.3	21.7	.15
Zhou et al. (2018c)	3.84	11.6	27.4	.38
Sun et al. (2019b)	4.07	11.0	27.5	.50
Sun et al. (2019a)	4.31	11.9	29.5	.53
Human Estimate	15.2	25.9	45.1	3.8

Table 3.1: The performance of several state-of-the-art, video-only models, with lower (constant prediction) and upper (human estimate) bounds.

from the perspective of building video captioning systems in practice, we suspect that incorporating additional modalities like ASR is more likely to result in performance gains versus building better computer vision models.

3.5 Models

In addition to the constant prediction baseline, we explore a series of ASR-based baseline methods:

ASR as the Caption (ASC) This baseline returns the test-time ASR token sequence as the caption. While the result is not a coherent, imperative step, performance of this method offers insight into the extent of word overlap between the ASR sequence and the target groundtruth, as measured by the captioning metrics.

Filtered ASR (FASC) Given that the ASR token sequences are much longer than groundtruth captions (§ 3.4.1), the performance of ASC incurs a length (or precision-based) penalty for several metrics. The FASC baseline strengthens ASC by removing word types that are less likely to appear in groundtruth captions, e.g., “ah”, “he”, “hello,” or “wish”. Specifically, we only keep words with

high $\frac{P(w|GT)}{P(w|ASR)}$ values, i.e., words that would be indicative of the groundtruth class if we were to build a Naive-Bayes classifier with add-one smoothing; probabilities are computed only over the training set to reduce the risk of overfitting. This baseline produces outputs that are shorter compared to ASC, but it is unlikely to yield fluent, readable text.

ASR-based Retrieval (RET) This retrieval baseline memorizes the recipe steps in the training set, and represents them each as tf-idf vectors. At test-time, the ASR sequence is converted into a tf-idf vector and compared to each training-set caption via cosine similarity.⁵ The training caption that is most similar to the test-time ASR according to this metric is returned as the “generated” caption. Note that, although a memorization-based technique, this baseline method produces de-facto captions as outputs.

3.5.1 Transformer-based Neural Models

We explore neural encoder-decoder models based on Transformer networks (Vaswani et al., 2017). In contrast to RNNs, Transformers abandon recurrence in favor of a mix of different types of feed-forward layers, e.g., in the case of the Transformer decoder, self-attention layers, cross-attention layers (attending to the encoder outputs), and fully connected feed-forward layers. We explore two variants of the Transformer, corresponding to different hypotheses about what information might be useful for captioning instructional videos.

ASR Transformer (AT) This model learns to map ASR-token sequences directly to captions using a standard sequence-to-sequence Transformer architec-

⁵We tried several variants of this method, e.g., comparing test ASR to train ASR, but found that comparing test ASR to train captions performed the best.

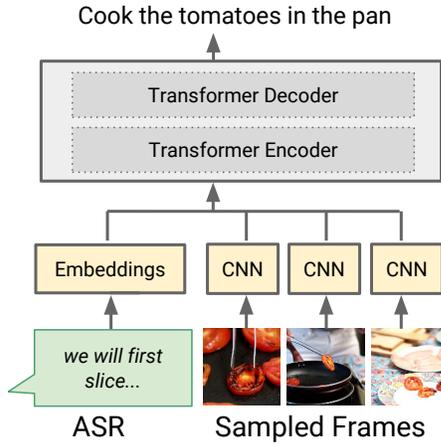


Figure 3.2: The AT+Video model. Both the encoder and decoder layers perform cross-modal attention.

ture. The model’s parameters are optimized to maximize the probability of the ground-truth instructions, conditioned on the input ASR sequences.

Multimodal model (AT+Video) We incorporate video features into the ASR transformer (Fig 3.2). For ease of comparison with prior and future work, we use features extracted from ResNet34 (He et al., 2016a) pretrained on the ImageNet classification task; these features are provided in the YouCook2 data release. Each video is initially uniformly sampled at 512 frames, with an average of 30 frames per captioned-segment.

To represent each video segment, first, k frames are randomly sampled with replacement. The sampled frames are temporally sorted to preserve ordering information, and their corresponding ResNet34 feature vectors are projected to the Transformer encoder hidden dimension via a width-1 1D convolution. We use $k = 10$ for all our experiments. The encoder self-attention layers perform *cross-modal attention* operations between the visual features and the ASR-token-based features. For each output token, the decoder attends to previously predicted tokens, and encoder outputs for all input frames / ASR tokens.

3.6 Experiments

We perform 10-fold cross-validation with randomly sampled 80/10/10 train/dev/test splits (split at the video-level), using the same splits for all models. After discarding the videos that were deleted at the time of data collection, each split contains roughly 1.1K training videos (averaging 8.3K training segments). We report mean performance over these splits according to four standard captioning accuracy metrics, introduced in §3.4.2. ROUGE-L, CIDEr, BLEU-4, and METEOR. We perform both Wilcoxon signed-rank tests (Demšar, 2006) and two-sided corrected resampled t-tests (Nadeau and Bengio, 2000) to estimate statistical significance. To be conservative and reduce the chance of Type I error, we take whichever p -value is larger between these two tests.

Transformer-based model details. For each cross-validation split, we use a batch size of 128, tie the Transformer model’s feed forward and model dimensions $d_{ffn} = d_{model}$, and optimize regularized cross-entropy loss using Adam (Kingma and Ba, 2015) with $lr = .001$. We train models for 100K steps, storing checkpoint files periodically. For each split, we train 8 model variants, conducting a grid search over model dimension, number of encoder/decoder layers, and L2 regularization: we consider all model parameter settings in $(d_{model}, N_{layer}, \lambda_{reg}) \in \{128, 256\} \times \{2, 3\} \times \{.0005, .001\}$ for each cross-validation split independently, and select the highest performing, checkpointed model according to ROUGE-L over the development set for that fold. Transformer models are implemented using `tensor2tensor` (Vaswani et al., 2018) and `Tensorflow` (Abadi et al., 2015). The vocabulary (average size 800) is determined separately using the training data for each cross-validation split. Words are considered if

	BLEU-4	METEOR	ROUGE-L	CIDEr
CNST	2.70	10.03	21.69	0.15
Sun et al. (2019a)	4.31	11.91	29.47	0.53
ASC	1.68	14.86	19.24	0.20
FASC	4.32	<u>18.47</u>	30.07	0.59
RET	5.68	14.29	28.06	0.80
AT	<u>8.55</u>	16.93	35.54	1.06
AT+Video	<u>9.01</u>	<u>17.77</u>	36.65	1.12

Table 3.2: Caption generation performance: AT+Video is a multimodal model that adds visual frame features to AT. A bolded value in a column indicates a statistically-significant improvement, whereas an underline indicates a statistical tie for best ($p < .01$).

they occur at least 5 times in the ground-truth of the current training set.⁶ This leads to an OOV rate of ~60% in the input. We truncate inputs at 80 tokens (~10-15% of transcripts are truncated in this process). For simplicity, decoding is done greedily in all cases.

Generation Experiment Results. Table 3.2 reports the performance of each model. For unimodal models, simple baselines like FASC (filtered ASR) and RET (training-caption retrieval) outperform the state-of-the-art video-only model of Sun et al. (2019a), according to the four automatic evaluation metrics. Overall, AT yields the best unimodal performance. Combining ASR and visual signals into a multimodal representation performs even better: the AT+Video model tends to outperform AT (and Sun et al. (2019a)), according to ROUGE-L, CIDEr, and METEOR ($p < .01$). Since AT and AT+Video have identical architectures and differ only in the available inputs, this result provides strong evidence that it is indeed the *multimodality* of AT+Video that leads to the (statistically significant) performance gains over the strongest unimodal models. We present some output examples in Fig. 3.3.

⁶Different vocabulary creation schemes, e.g., sub-word tokenization, led to small performance decreases.

Video						
ASR	"so I just want to go ahead and remove all of this fat from our chicken... cut it into about one inch pieces so you want pieces"		"... color them and then shape them ... tongs so as not to burn yourself it goes with total tacos in a frying pan ..."		"fattoush salad but you can add in cilantro and some other herbs if you prefer to do that instead of the parsley and one"	
Target	cut the chicken into pieces	prepare the tortillas and roll them using rolling pin	add chopped parsley to the mixture too	cut the circle in half	add chile powder	place the chicken on the rice
Pred.	cut the chicken into pieces	place the tortilla on the pan and roll	add cilantro to the salad	cut the dough into UNK pieces	add the coriander powder coriander...	add the sauce to the pot

Figure 3.3: Example generations from AT+Video in cases where it performs well, okay, and poorly.

3.6.1 Diversity of Generated Captions

In addition to the automatic quality metrics, we measure how diverse the generated caption are for each model, using the following metrics: vocabulary coverage (the percent of vocabulary that was predicted at test-time by each algorithm at least once); proportion not copied (the percent of generated captions that do not appear in the training set verbatim); and output uniqueness (the percent of generated captions that are unique). These metrics are useful because they can highlight undesirable, degenerate behavior for models.⁷ As an upper bound, we compute these metrics for the ground-truth (GT) test-time targets. Note that even the ground-truth targets do not achieve 100% in these diversity metrics: for vocabulary coverage, not all vocabulary items appear in the ground-truth captions for a given cross-validation split. Similarly, for proportion not copied/output uniqueness there are repeated captions in the label set.

According to all metrics, AT+Video outputs are slightly more diverse compared to the AT outputs (Fig. 3.4). This observation suggests that the multi-

⁷For instance, the constant prediction baseline we consider would score low in both vocab coverage and uniqueness.

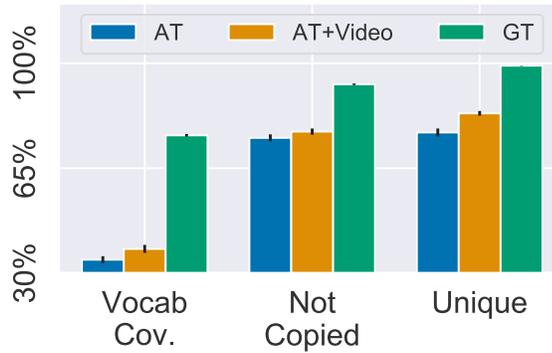


Figure 3.4: The multimodal model AT+Video produces slightly more diverse captions than its unimodal counterparts.

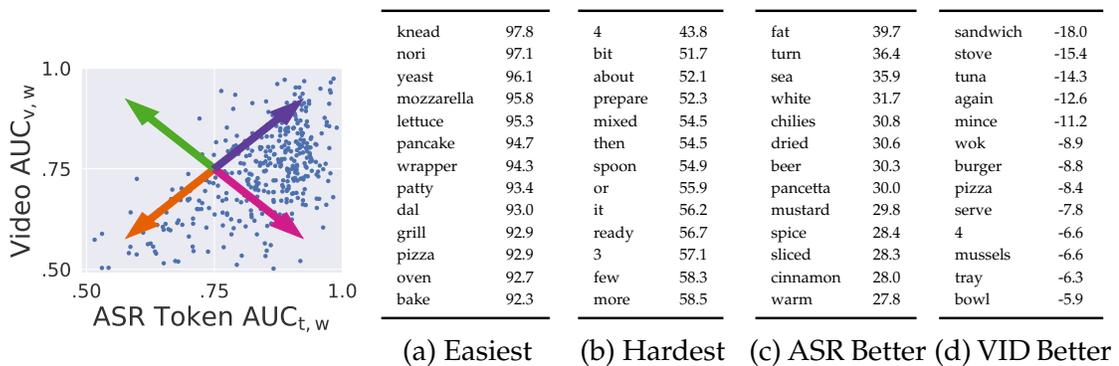


Figure 3.5: Per-word classification results using ASR and/or Video features. Each point in the scatterplot represents a different word-type; x-coordinate values show how well a word is predicted by ASR-token features; y-coordinate values show how well a word is predicted by video features. Tables (a)-(d) show word types that are easy, universally difficult, better-predicted-by-ASR, and better-predicted-by-video, respectively.

modal model is not simply exploiting a degeneracy to achieve its performance improvements.

3.7 Complementarity of Video and ASR

We now turn to the question of *why* multimodal models produce better captions: what type of signal does video contain that speech does not (and vice versa)?

Our initial idea was to quantitatively compare the captions generated by AT versus AT+Video; however, because the dataset is relatively small, we were unable to make observations about the generated captions that were statistically significant.⁸

Instead, we examine properties of the ASR-token-based and visual features directly. Following a procedure inspired from Lu et al. (2008); Berg et al. (2012); Dai et al. (2018); Mahajan et al. (2018), we consider the auxiliary task of predicting presence/absence of unigrams in the ground truth captions from features extracted from corresponding segments. We train two unimodal classifiers, one using ASR-token-based features and one using visual features, and measure their relative capacity to predict different word types; the goal is to measure which word types are most predictable from the ASR tokens and, conversely, which ones are most predictable from the visual features.

For each segment, we predict the unigram distribution of its corresponding caption using a unimodal softmax classifier: for simplicity, we use a 2-layer, residual deep averaging network (Iyyer et al., 2015) for both the visual and ASR-based classifier. We measure per-word-type performance using AUC, which is word-frequency independent. For a binary classification task with \mathcal{P} positive and \mathcal{N} negative instances, the AUC of a model m , which produces a score for inputs corresponding to the relative confidence the model has in the input belonging to the positive class is:

$$\text{AUC} = \frac{1}{|\mathcal{P}||\mathcal{N}|} \sum_{p \in \mathcal{P}} \sum_{n \in \mathcal{N}} \mathbb{I}[m(p) > m(n)]$$

In our case, for each word type w (e.g., $w = \text{beer}$) we measure how well w

⁸In general, making concrete statements about the causal link between inputs and outputs of sequence-to-sequence models is challenging, even in the text-to-text case, see Alvarez-Melis and Jaakkola (2017).

is predicted by the classifier based on ASR / spoken tokens $AUC_{t,w}$ (e.g., $AUC_{t,beer} = 98$) and, conversely, how well w is predicted by the visual classifier $AUC_{v,w}$ ($AUC_{v,beer} = 68$). For a given word type, we measure its overall difficulty by averaging $AUC_{t,w}$ and $AUC_{v,w}$; we call this $AUC_{\mu,w}$ ($AUC_{\mu,beer} = 83$). Similarly, we measure the difference in difficulty by subtracting $AUC_{t,w}$ and $AUC_{v,w}$ to give $AUC_{\Delta,w}$ ($AUC_{\Delta,beer} = 30$) with higher values indicating that a word type is predicted better by the spoken-token features compared to the visual features. We plot $AUC_{t,w}$ versus $AUC_{v,w}$ for 382 words in Fig. 3.5 (results are averaged over 10 cross-val splits).

Absolute Performance. Points in the upper-right quadrant of Fig. 3.5 represent words that are easy for both visual and ASR-token-based features to predict, whereas points in the lower-left represent words that are more difficult. Specific ingredients, e.g., “nori” and “mozzarella,” are often easy to detect, as are actions closely associated with particular objects (e.g., “dough” is almost always the object being “knead”-ed). Conversely, pronouns (e.g., “it”) and conjunctions (e.g., “or”) are universally difficult to predict.

Visual vs. ASR-token-based features. In general, ASR-token-based features carry greater predictive power, as evidenced by the skew towards the bottom right in the scatterplot in Fig. 3.5. One pattern in the cases where speech features perform better (Fig. 3.5c) is that words are often modifiers, e.g., *white* (pepper), *sea* (salt), *dried* (chilies), *olive* (oil), etc. Indeed, small, detailed distinctions may be often difficult to make from visual features, e.g., “vegetable oil” and “olive oil” may look identical in most YouTube videos.

Nonetheless, there are types better predicted by video features (Fig. 3.5d). Often, these are cases that require unstated, background knowledge, i.e., references to objects not explicitly stated by the speaker(s). To quantify this ob-

ervation, for each word type we compute the likelihood that it is *stated* by the speaker in the video, given that it appears in the ground-truth caption, i.e., $P(w \in \text{ASR} \mid w \in \text{GT})$. Aside from trivial cases (e.g., words misspelled in the GT never appear in the ASR), words that are often unstated include action words (e.g., “place”, “crush”) and cookware (e.g., “pan”, “wok”, “pot”). Words that are often stated include specific ingredients (e.g., “honey”, “coconut”, “ginger”). In contrast to word frequency (which is uncorrelated with $\text{AUC}_{\Delta,w}$, Spearman $\rho \approx 0$), stated rate *is* correlated with $\text{AUC}_{\Delta,w}$ ($\rho = 0.44, p < .01$).

3.8 Oracle Object Detection

The results in Table 3.2 indicate that, while adding visual information yields statistically significant improvements to the ASR-only model, the improvements are not large in magnitude. This leaves open the question of whether (a) any visual information simply does not provide much additional information on top of ASR, or (b) we need better visual modeling. We take a first step in addressing this question by experimenting with an “oracle” object detector that provides perfect-precision predictions.⁹ If even oracle object detection does not help, then the answer is more likely (a) rather than (b) above.

As part of a YouCook2 data release, bounding box annotations for selected objects in the recipe text (Zhou et al., 2018a) were provided. Unfortunately, while these could have served as an oracle, the actual annotations are only available for a small fraction of the data. Instead, we consider the set of 62 object labels made available. We simulate a high-precision, oracle object de-

⁹High-precision object detectors are gaining popularity in the computer vision community because the training data is easier to annotate, e.g., Krasin et al. (2017).

tor by identifying – per video segment – the overlap between (morphology-normalized) groundtruth caption mentions and the 62 object labels available.¹⁰ For instance, for the groundtruth caption “put the mushrooms in the pan”, the oracle object detector yields “mushroom” and “pan”. 89% of segments receive at least one oracle object. The oracle object detections are then fed into the Transformer encoder (in random order), either by themselves (Oracle) or along with the ASR token sequence (AT+Oracle). We perform the same cross-validation experiments as described in §3.6, and report the average ROUGE-L (we observe similar trends with other metrics):

	AT	AT+Video	Oracle	AT+Oracle
ROUGE-L	35.5	36.7	40.8	45.5

Because the AT+Oracle model achieves large improvements over AT+Video, we suspect that building higher-quality visual representations is a promising avenue for future work.

How weak of an oracle can still produce high performance? Fig. 3.6 shows performances of models using *subsets* of the 62 objects (most frequent 10% of objects through 90%) over one cross-validation fold. AT+Oracle gives better performance than AT+Video by detecting *just 6 object types*, and the oracle by itself (which is only given access to object sets) achieves comparable performance to AT+Video with 30 object types. These results suggest that, at least for this task, the Transformer decoder is likely not the main performance bottleneck, as it is able to paste together unordered object detections into captions effectively.

¹⁰This oracle is unlikely to be achievable, as it assumes 100% precision for the 62 objects considered (which also implies modeling *which* objects to talk about, a non-trivial task in itself (Berg et al., 2012)).

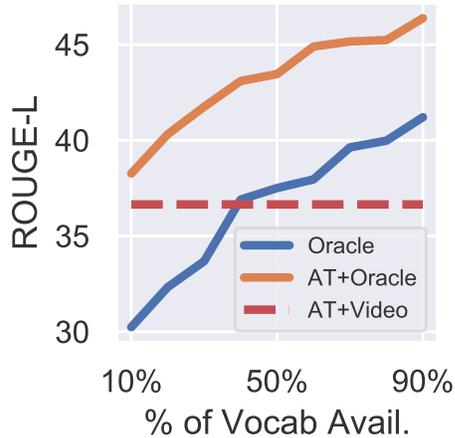


Figure 3.6: The performance of the oracle methods increases as they are given access to an increasing number of object types.

3.9 Conclusion

In this work, we demonstrate the impact of incorporating both visual and ASR-token-based features into instructional video captioning models. Additional experiments investigate the complementarity of the visual and speech signals. Our oracle experiments suggest that performance bottlenecks likely derive from the input encoding, as the decoder is able to paste together even simple sets of object detections into high-quality captions.

CHAPTER 4
UNDERSTANDING: PREDICTING POPULARITY IN MULTIMODAL
COMMUNITIES

4.1 Brief Overview

The content of today’s social media is becoming more and more rich, increasingly mixing text, images, videos, and audio. It is an intriguing research question to model the interplay between these different modes in attracting user attention and engagement. But in order to pursue this study of multimodal content, we must also account for context: timing effects, community preferences, and social factors (e.g., which authors are already popular) also affect the amount of feedback and reaction that social-media posts receive. In this work, we separate out the influence of these non-content factors in several ways. First, we focus on ranking pairs of submissions posted to the same community in quick succession, e.g., within 30 seconds; this framing encourages models to focus on time-agnostic and community-specific content features. Within that setting, we determine the relative performance of author vs. content features. We find that victory usually belongs to “cats and captions,” as visual and textual features together tend to outperform identity-based features. Moreover, our experiments show that when considered in isolation, simple unigram text features and deep neural network visual features yield the highest accuracy individually, and that the combination of the two modalities generally leads to the best accuracies overall.

The work in this chapter is joint with Lillian Lee and David Mimno, and was published in Hessel et al. (2017).

4.2 Introduction

Today's user-generated content is becoming more multimodal as users increasingly mix text, images, videos, and audio. Does one mode tend to be preferred over another — for example, on the Internet, is it indeed true that “a picture is worth a thousand words”? Or do the visual and the linguistic interact, sometimes reinforcing and sometimes counteracting each other's individual influence? Anecdotally, at least, it seems that there is interesting interplay between these different modes. For example, Figure 4.1 compares two posts made to the same forum on the same site, both containing captioned images of two cats. One could argue that the leftmost one has a more clever caption¹ but the second has a more attractive image. Which would more users prefer?

However, determining what multimodal content is most attractive is complicated by the fact that popularity can be strongly dependent on many non-content factors (Suh et al., 2010; Bakshy et al., 2011; Hong et al., 2011; Ma et al., 2012; Borghol et al., 2012; Romero et al., 2013; Lakkaraju et al., 2013). Posts by users that already have a large audience tend to enjoy an advantage over posts by relatively unknown people; posts that appear when users are most active are also more popular; and sometimes simply the fact that a post receives a few early clicks ensures that it gains even more popularity.

Yet to dismiss the importance of the *content* of a post would be wrong. From a user's perspective, if content matters less than identity and timing, why would they bother taking better pictures or writing wittier captions? Community moderators, who would ideally like to promote high quality content even if it was submitted at a less-than-optimal time or by a non-celebrity user, would also

¹One user comments in response: “A good title! Refreshing. Better than ‘this lil guy.’”



Figure 4.1: Despite being submitted only 13 seconds apart to the subreddit *aww*, one of these submissions received over 1600 upvotes whereas the other received fewer than 20; the answer is in § 4.3. Images courtesy *imgur.com*, posted by Reddit users *mercurycloud* and *imsozzy*.

appreciate a model of content alone. Researchers trying to understand community preferences/biases want to model users’ likes and dislikes, not the idiosyncrasies of ranking algorithms and random early upvoting patterns.

In this work, we seek to measure content preferences independent of confounding factors. We collect and analyze data from six sub-communities on *reddit.com* of varying size and focus. Each focuses on *multimodal* posts that include images and captions. Inspired by our prior work on wording effects (Danescu-Niculescu-Mizil et al., 2012; Tan et al., 2014), we select pairs of captioned images posted at *similar times* (e.g., 30 seconds) to the *same community* and then construct models to predict which of the two eventually becomes more popular. Comparing submissions in this time-controlled setting allows us to approach an “equal footing” assumption when modeling content, and to quantify the validity of that assumption.

We choose to explicitly control for time of posting because we find that it

is the most important contextual factor and because it is relatively easy to find comparable pairs. But there are other factors in play, some Reddit-specific and that are impossible for us to recover, e.g., the precise ordering of content displayed to users. However, we perform human annotation experiments that verify that these unrecoverable factors do not overwhelm the influence of content (see §4.4 for more details). For other factors, such as a user’s social status or experience, we take the approach of quantifying the predictive performance of such effects relative to content features, since explicitly controlling for both timing and user would leave us with too little data to work with.

When comparing “cats and captions” — that is, post content — to creator characteristics, we find that “cats and captions” are generally more important for the communities we examine. Also, while image features always outperform text features when both are considered independently (albeit only if deep learning is employed), in five of six Reddit communities, significant performance gains are observed when combining modalities.

The main contributions of this work are:

1. An exploration of time-sensitive content popularity across various communities on `reddit.com`, and an accompanying argument for framing these investigations in a time-controlled, ranking setting.
2. Several publicly available² datasets and ranking tasks involving the prediction of community response to multimodal content, plus estimates of human performance on these tasks.
3. A comparison of off-the-shelf image and language features against social and timing baselines, and a demonstration that multimodal features are

²www.cs.cornell.edu/~jhessel/cats/cats.html

worth incorporating. The models we consider can also be applied to submissions in isolation, enabling on-line scoring of novel content.

4.3 Datasets

Our starting point for Reddit data is Tan and Lee (2015)'s dataset of all 106M submissions to Reddit from 2007 to 2014 and Hessel et al. (2016)'s extension of this dataset to include full Reddit comment trees. Reddit, which is the 25th most popular site on the Internet according to `Alexa.com` as of Fall 2016, consists of interest-centric subcommunities called subreddits. These datasets are based on the work of Jason Baumgartner of `pushshift.io` who scraped Reddit using their public API.

On Reddit, users are allowed to up/downvote content submitted by other users. While the exact counts of each of these votes are not made available,³ Reddit computes and displays a proprietary “engagement” metric based on the number of upvotes minus the number of downvotes. This quantity, called the *score* of a post, has been readily used in previous work, and is the measure of engagement we will be examining.

Content on Reddit is shared with topical subreddits (e.g., politics, Art); this allows us to control for the types of content by only comparing submissions within a given subreddit. In contrast to a majority of previous work that uses general-purpose image datasets from Flickr for popularity prediction, we examine a wide variety of *granularities* of content, ranging from highly general to very fine. Khosla et al. (2014), for example, find that objects like revolvers and

³The exact totals are obscured to prevent spam.

women’s bathing suits are predictive of popularity, whereas spatulas, plungers, and laptops have a negative impact. In other words, while previous work has addressed which types of objects tend to become popular, here we examine what objects of a *given* type become popular.

Many subreddits embody a larger growing trend towards images, video, and other media content. Nearly all major social media sites (e.g., Facebook, Twitter, Pinterest) support image and video, and some networks make multimodal content their focus (e.g., Instagram). We performed a similar analysis to Singer et al. (2014), and hand-categorize popular top-level domains on Reddit into “media” (e.g., *imgur*, *youtube*) “news” (e.g., *cnn*, *bbc*) and Reddit internal title-only and text posts. Figure 4.2 demonstrates the dramatic rise in multimedia content submitted to Reddit from 2005 to 2014. Note that this graph is proportional — the raw number of multimedia submissions to the site is still rising, even though the proportion has flattened. Roughly 30% of all submissions to Reddit are images, gifs, videos, and the like. In fact, more than 400 subreddits have each amassed more than 5,000 image submissions. If researchers frame problems carefully, these communities offer a diverse set of in-situ human and community reactions to multimedia content without the need for expensive annotations.

We focus on six image-centric subreddits of varying popularity, visual focus, and social structure. These communities range from *pics*,⁴ which has millions of subscribers and offers few guidelines about what types of images are permitted, to *RedditLaqueristas* [sic], where users submit photographs of artistically lacquered fingernails. Typical examples of image/text submissions made to *aww*

⁴According to the moderators: *pics* is “a place to share photographs and pictures.”

	# Users	#/% Imgur	Cap Len
pics	2108K	2472K/70%	9.84
aww	1010K	954K/81%	9.13
cats	109K	100K/73%	8.97
MakeupAddiction (MA)	77K	58K/57%	13.67
FoodPorn (FP)	74K	50K/77%	9.39
RedditLaqueristas (RL)	27K	39K/73%	11.12

Table 4.1: Number of unique users, number of Imgur submissions, and the average caption length for the communities used in this study. The number of unique users includes those who commented or submitted.

are shown in in Figure 4.1.⁵ General statistics about each of these datasets are presented in Table 4.1. Note that some community name abbreviations are also introduced here.

While users are able to submit links from any website on the Internet to any subreddit, the most common top-level domain is `imgur.com` (Alexa.com rank 48, Fall 2016), a site created to be an image hosting companion site to Reddit.⁶ Imgur allows users to upload content which can subsequently be shared to Reddit. All images in our datasets were fetched from Imgur.

We define a subreddit to be “active” if it receives more than 15 submissions on that day.⁷ We attempted to scrape all images from all active days from all six subreddits from Reddit’s inception until February 1st, 2014.

⁵The left submission was the more popular of the two, receiving at least 1K more upvotes than the right.

⁶<https://goo.gl/2fX34m>

⁷This is mostly done to filter out the unreliable feedback early in a community’s life. After the first active day, the proportion of active days thereafter varies from 96% in the case of `pics` to 55% in the case of `RL`, with an average of 83% over all datasets.

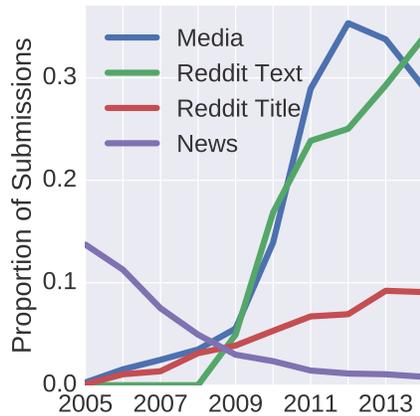


Figure 4.2: Proportional popularity of types of Reddit posts over time across all subreddits.

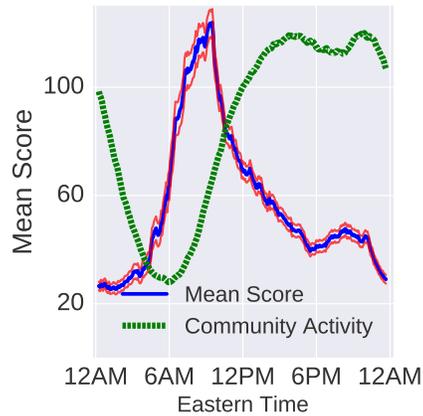


Figure 4.3: Average score versus time of day (eastern) on aww with 95% CI (red) and activity levels.

As preprocessing we remove any duplicate images,⁸ and any animated or corrupted images. Imgur albums consisting of multiple images are also discarded. All images are resized to 256 pixels by 256 pixels. All datasets, including train/test splits, are available at www.cs.cornell.edu/~jhessel/cats/cats.html.

4.4 Time and Rich-getting-richer

Our objective is to isolate content features and predict the relative popularity of two items posted at approximately the same time. This approach has the advantage that it is relatively insensitive to two factors: the time of posting and the absolute number of positive user votes. In this section we provide arguments that these factors are significant in our data set and that previous methods for

⁸We filter duplicate links by matching imgur ID and duplicate images by PHASH with a hand-picked hamming distance threshold of five. We attempt to discard *all* copies of repeat submissions to mitigate any effect of repeated submissions, though deleted posts and pathological cases prevent us from guaranteeing that there are no duplicates.

controlling for these factors are not sufficient.

Why Control for Time? Raw post scores are influenced by timing factors in complex and difficult-to-measure ways. Reddit is a dynamic, evolving platform, so expected popularity of submissions varies across many time scales. In Figure 4.3, for example, we show the mean score of submissions made at various times during the day averaged over a sliding 30 minute window in aww. The figure also shows the average activity level of the community as measured by number of submissions. There is a dramatic spike in average submission score for posts submitted at 9AM when compared to posts submitted at 6AM or 12PM.

Expected popularity also varies periodically between days of the week. Figure 4.4a shows posts binned by day of the week. The average score of submissions to aww, pics, and cats seems to be greater on weekends when compared to weekdays. These patterns are not always easily modeled; the number of upvotes in MakeupAddiction falls sharply on Tuesdays, potentially as a result of the community's "Text Tuesdays" tradition (when only text posts are allowed). We observed similar patterns in the other subreddits. Figure 4.4b illustrates binning by submission year. The average post score on Reddit seems to be increasing over time, but it is unclear whether this trend has continued in 2014, as vote totals might not have had the chance to stabilize at the time of scraping in early 2014.

Reddit communities more closely resemble time-sensitive "cultural markets" as described by Salganik et al. (2006) than any of the three image-sharing settings described by Khosla et al. (2014).

Mean Normalization. Lakkaraju et al. (2013)'s work offers a starting point for

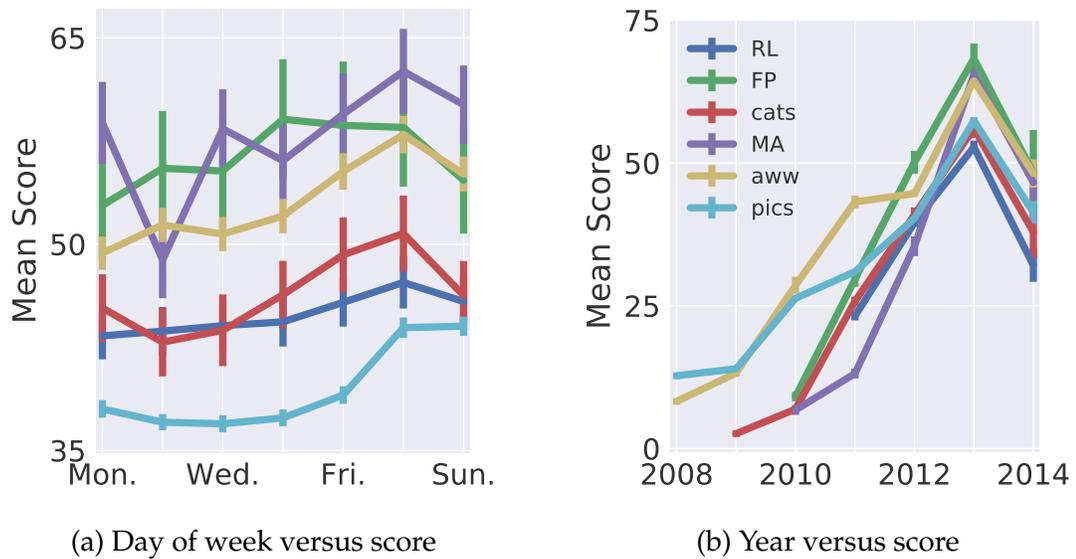


Figure 4.4: Relationship between various measures of time and eventual submission score with 95% confidence intervals.

designing a time-control mechanism. Their original goal was to control for popularity *between* subreddits, whereas we aim to control for time *within* a community.

We identified several problems when applying their time control method, which we call *mean normalization* (MN), to our setting. MN divides the score of a Reddit post by the average score of all posts surrounding it in an hour. Estimating a robust and accurate mean is difficult because of the dynamics of popularity. The submission distribution is skewed by rich-get-richer processes (for aww, average skew is 7.33, average kurtosis is 69.25), so *average* popularity as a statistic does not capture a fair notion of quality. Furthermore, submissions that are unlucky enough to be posted within the same hour as a popular post are unfairly downweighted by the rich-get-richer process.

For some subreddits, a one hour time window is probably too big. Figure 4.3 suggests that one hour can encompass large changes in expected popularity

in the fast-paced world of Reddit, e.g., the mean score for 6AM submissions is around 64, whereas the mean score for 7AM submissions is around 93.

Finally, in less popular communities, Lakkaraju et al. note that it is often difficult to get a stable estimate of the mean submission popularity within an hour. While their setting doesn't require estimating this mean, ours does. In FoodPorn, for example, just 44% of all submissions have at least 5 in-window submissions ($\mu = 4.57$) to take the mean over.

Raw Transformations on Reddit. Raw post scores, even normalized for temporal effects, may be too noisy to learn accurate models. Self-reinforcing “rich get richer” dynamics in online interfaces result in complex, non-linear relationships between quality and popularity (Salganik et al., 2006). Furthermore, recent work shows that these dynamics differ significantly from community to community; sometimes a small number of highly scored submissions is preferred, while in other cases, the scores are more evenly distributed (Lee et al., 2016). The complexity of this relationship is compounded by website interface changes, ranking algorithm modifications,⁹ and innumerable other subtle effects.

Transformations of raw votes are known to be more effective than highly-skewed raw values. Khosla et al. (2014), for example, successfully use a log-transformation on Flickr view counts. In the case of Reddit, however, heuristic transformations like these enforce complex biases that are not consistent between different subreddits. Also, it is not clear how to extend these to a time-controlled setting, in general.

⁹In fact, between the time of submission and publication, Reddit *did* entirely change their method for computing post scores: [g00.gl/zHcKzL](https://www.reddit.com/r/programming/comments/g00gl/zHcKzL)

Our Approach: Pairwise Sampling. Because only relative judgments need to be made, the comparison of submissions made in quick succession requires no assumptions about the skewness of the score distribution. We do not need to compute a stable estimate of average popularity, so sparse submission data can be handled. No ad-hoc transformation of raw scores is required, either. If the time difference between two posts is small we can train models using the assumption that posts start on roughly equal footing. We can then quantify the validity of those assumptions in terms of timing and user baselines, and directly compare cats and captions to creators and the clock.

While it would be ideal to design a pairing process that would control for other social effects, doing so would be substantially more difficult than accounting for time. For example, if we sample pairs of posts made by the same author in a short time window, we would lose—at the very minimum—the 75% of submissions made to pics by users who have deleted their accounts or who only submit a single time. Also, Reddit enforces a one-post-per-several-minutes submission rate on a majority of accounts, meaning our stringent time controls would need to be relaxed. We leave sophisticated user-identity controlling sampling procedures to future work, and focus on quantifying the performance of user features instead.

After scraping the images associated with each subreddit, our goal is to pair submissions to minimize differences in timing. The pairing process is controlled by several parameters. For each community we define a fixed, maximal allowed time-window so that pairs are not too far apart. We select pairs greedily to minimize this gap, so in practice the average time difference is smaller than the maximum window size. To mitigate the effects of noise, we force the score

	Max/Avg Win	Med/Avg Diff	# Pairs
pics	30/15 sec	117/478	44K
aww	30/15 sec	90/393	33K
cats	15/7 min	69/231	15K
MA	60/24 min	88/227	10K
FP	120/53 min	62/188	8K
RL	30/14 min	56/118	9K

Table 4.2: Statistics regarding the sampling used to generate ranking pairs. The maximum window is the maximum number of minutes that two submissions can be apart to be paired up, whereas the average window is the average time between all sampled pairs. The median and mean score differences between pairs is also given.

difference between members of a pair to be at least 20,¹⁰ and the eventually more popular submission must also be at least twice as popular as the other. Additionally, we ignore posts that received a score of less than two to avoid spam and other very low-quality submissions that received no upvotes.

Table 4.2 shows the maximum and average window sizes, along with the number of pairs that were sampled using a simple greedy algorithm. For aww and pics, the most popular communities we examine, sampled pairs are submitted 15 seconds apart on average.

Human Validation. We first consider the validity of this new task by conducting a small human study. Our goal with this study was to determine if the task of predicting relative engagement was even possible using these datasets, or whether there is no correlation between content and Reddit score. We asked annotators to predict which among two time-controlled submissions they thought would get more upvotes. For each of the six datasets we showed the same 20

¹⁰A majority of experiments were also conducted with the minimum difference parameter set to four; results were similar to those presented here.

	aww	pics	cats	MA	FP	RL
Humans	60.0	63.6	59.6	62.2	72.7	67.2

Table 4.3: Human annotation accuracy results.

pairs to annotators.¹¹ In total, we were able to gather 1400 human pairwise judgments. In addition, users were given the option of describing “why” they made the choice they did.

Annotators used a variety of techniques to make their decisions. Rationale ranged from basic aesthetic observations (“Much better photo;” submitted with a correct annotation. “Better photo;” correct annotation. “homemade + steak + picture resolution (so profesh);” correct annotation.) to comments about how unique images were (“Dude, it’s a cat with a pencil;” incorrect). Sometimes, the authors disqualified submissions based on the associated text, rather than on the images (“Less begging in the title;” incorrect). Many annotators used their perception of the communities when making judgements (“The Internet loves meat;” correct. “Easy. Desserts always win;” correct). Sometimes the annotators wished they were more familiar with the community, e.g., one user submitted an incorrect annotation, noting that “[they were not] sure whether FoodPorn is about the images or the food concept.” Some pairs were universally difficult. For example, 83% of annotators incorrectly selected a cute rabbit (“Dat bunny face;” +10 Reddit score) over an out-of-focus photo of a duck¹² with the caption “My brother got a duck yesterday..” [sic] (+115 Reddit score).

The resulting mean accuracy for each dataset is presented in Table 4.3. In

¹¹Due to a sampling bug, `pics` pairs in the human experiments were sampled from 2009-2012 instead of 2009-2014.

¹²One redditor comments regarding the misfocused image: “That trashcan [in the background] is in excellent quality.”

general, humans are able to guess pairwise rankings of submissions from images and captions, but the task is difficult.¹³ Having validated that the task is neither trivial nor impossible for humans, we now move on to our machine learning experiments.

4.5 Model Design

For relative popularity prediction, we use a pairwise learning-to-rank model (Herbrich et al., 1999; Joachims, 2002; Burges et al., 2005). Specifically, our data is of the form $\{x_{1i}, x_{2i}, y_i\}_{i=1}^n$ where $\langle x_{1i}, x_{2i} \rangle$ is a pair of Reddit submissions posted at similar times, and y_i is an indicator variable that encodes which submission became more popular. We train a linear classifier on top of the *vector difference* of two entities for predicting which of the two is more highly ranked (i.e. y_i). As such, we experiment with models of the form

$$\hat{y}_i = w^T(f(x_{1i}) - f(x_{2i})) \quad (4.1)$$

where w is a set of regression weights and f is one of a variety of Reddit submission representation functions. In all experiments, we use a hinge loss, which is minimized with respect to the coefficients of the regression itself and, if applicable, with respect to the trainable parameters of f .

Note that our model implicitly learns a scoring function that can assign a quality score to *unpaired* examples. Specifically $w^T f(x) \in \mathbb{R}$ is a value that correlates with the model’s ranking of that submission.¹⁴ This function could be

¹³Because the human study only considered a small subset of image pairs, the exact values reported are less precise than for the other results: the 95% confidence intervals for the human annotations are on average ± 6

¹⁴Rather than approximating the global raw Reddit score ranking, the model *induces* a ranking with desirable properties, e.g., it cannot be predicted from timing features.

used by moderators to compute model scores of novel, incoming submissions. We use this function in a later section to interpret our results.

Cats and Captions

The textual and visual characteristics of the six communities we examine are complex and varied. For example, most images in RL are of fingernails, which are out of domain for pretrained computer vision models. Similarly, complex social patterns and tags emerge within language e.g., “CCW” meaning “constructive criticism welcome.” As a result, a dataset-by-dataset examination of specific, higher-level processes like image (Khosla et al., 2012) or text (Danescu-Niculescu-Mizil et al., 2012) memorability transfer-learned from other domains is reserved for future work. The goal of this section is *not* to argue that these models are the best. Rather, we will use these generic feature extractors to demonstrate the importance of modeling content at all.

Image Models. We experiment with a combination of lower-level features and deep neural network models to represent image content. This mix of models is similar to those explored by Khosla et al. (2014).

The most basic building blocks of the human visual system are edges and colors, and the presentation of these features might effect how appealing an image is. Previous work has found that colors can play a role in human response to visual content (Bakhshi and Gilbert, 2015). As such, we examine a set of color features (**Color**), consisting of an l_1 normalized vector based on the RGB values of the colors in the image. We use Khan et al. (2013)’s 50 universal color descriptors and extraction code to compute this vector for each image.

The second feature set is histogram of oriented gradients (Dalal and Triggs, 2005). **HOG** features capture localized pixel gradients in an image. We use the HOG feature extractor in OpenCV (Bradski, 2000) with default parameters and use random projection to reduce the dimension of the resulting features to 2K from 34K.

Next, we examine the **GIST** image descriptors (Oliva and Torralba, 2001), which aim to capture “perceptual dimensions (naturalness, openness, roughness, expansion, ruggedness) that represent the dominant spatial structure of a scene.” We use the `pyleargist`¹⁵ library to extract these 960 features.

Recently, convolutional neural networks have been used to extract high-level concepts from image data. We use the popular **VGG-19** (Simonyan and Zisserman, 2014) and **ResNet50** (He et al., 2016b). Both of these networks are used as feature extractors¹⁶ by taking the final-layer activations from a set of weights trained for the ImageNet (Deng et al., 2009) ILSVRC-2014 classification task. Building a linear model over extracted features in this manner is known to offer an “astounding” baseline (Sharif Razavian et al., 2014).

Text Models. We first examine a set of **Structural** features of language. These include the message length (in tokens and characters), the token-to-type ratio, and a “punctuation proportion” feature to capture some signal about an author’s use of non-alpha-numeric characters (e.g., emoji).

We consider three models that do not use word order. The bag-of-words assumption is valuable both because of its relative simplicity and because of its high performance (see Hill et al. (2016) for some benchmarks). First we de-

¹⁵<https://bitbucket.org/ogrisel/pyleargist/>

¹⁶We found deeper residual networks and network fine-tuning to be unhelpful in early testing.

fine a set of **Unigram** features by mapping each caption to a vector of binary indicator variables. Second we extract topic distributions using a specialized biterm topic model (Yan et al., 2013) designed for short texts. We use 20 topics in all cases, and extract the resulting l_1 normalized **Topic** distributions. Third we use a variant¹⁷ of the deep averaging network (**DAN**) (Iyyer et al., 2015). This model averages a set of word embeddings and feeds the result through a simple multilayer perceptron. We consider a 3-layer DAN with 128 hidden units. The model’s word embeddings are tuned from a 100D GloVe (Pennington et al., 2014) pretrained set.

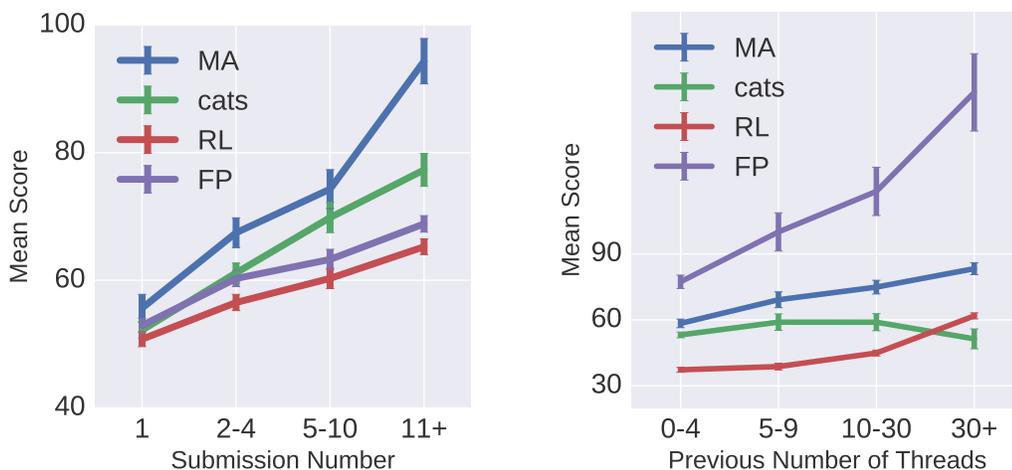
We also consider sequence-based features, specifically an order-sensitive recurrent neural network. We train an **LSTM** (Hochreiter and Schmidhuber, 1997) on the sequence of words in a caption. The parameters of the RNN are learned, and the word embeddings are tuned from the same 100-dimensional starting vectors as the DAN. For completeness, we also consider a bidirectional LSTM, **Bi-LSTM** (Graves and Schmidhuber, 2005).

Creators and the Clock

User Features. Can Reddit users get upvotes based on an attained status as on other social networks? An explicit and persistent user identity exists for some users on MakeupAddiction and RedditLaqueristas in the form of a *flair* that is displayed alongside a given user’s posts. Most often, the flair contains a link to a given user’s Instagram profile.

For other communities, however, a majority of users submit only a few

¹⁷We apply l_2 normalization after the averaging step, and don’t apply word-level dropout.



(a) Submission # vs score

(b) Conversation # vs score

Figure 4.5: Relationship between various measures of time and eventual submission score for several subreddits, with 95% CI.

times. Around 60% of submissions made to aww and cats, for example, are made by users who submit at most three times ever. Even if a celebrity status were earned and upvote counts were artificially inflated as a result, in these communities, this likely plays a lesser role.

Another hypothesis is that as a user gains familiarity with a community, they are better able to submit content of interest to that community. Indeed, a user who has a better sense of the types of content popular in a community might be more likely to submit high quality content than a newcomer.

Even though we cannot disentangle the effects of status and experience, we can still define features that capture aspects of a submitter's previous behavior within a community. Such features have previously been used in studies on Reddit (Althoff et al., 2014; Jaech et al., 2015) and Slashdot (Lampe and Resnick, 2004), among others.

Two easily measured quantities are how many times a user has previously

submitted and how many total threads a user has previously interacted with. Figure 4.5a and Figure 4.5b show that correlations exist between score and previous interactions. In *redditLaqueristas*, for example, if a submission is a user's fifth to tenth, it is more likely to receive upvotes than if a submission is a user's first. In *cats*, however, participating in more than 30 threads by commenting or posting seems to be associated with slightly lower average popularity.

The following set of user features are computed for each submitter at the time of their submission. When a statistic is not properly defined for a given user at a given time (e.g., average previous comment length when they have no previous comments; the submitter deleted their account, etc.) the mean value over the training set is substituted.¹⁸

Previous work (e.g., Dror et al. (2012)) has found information regarding *how much* a user participates in a community to be a useful predictor of their future behavior. The **Activity** feature set includes the number of previous posts/comments, how long the user has been a member of the community, the time since previous interaction, and the ratio of posts to posts plus comments for that author.

It is possible that *how* a user interacts with others in a community is more important than how much they interact with a community. The **Type** feature set includes average comment length, average comment token-to-type ratio, average conversation tree depth of comment, the proportion of previous comments with replies, the proportion of previous submissions wherein the user commented multiple times, and median time-to-response from thread start.

¹⁸We ran user experiments considering only pairs that consisted of no deleted users and no users without previous interaction data; the results were comparable.

Several variables are used to quantify the community-perceived **Quality** of a submitter’s previous interactions. Instead of using statistics based on raw scores, which can be skewed by a small number of very popular interactions, we use Jaech et al.’s Jaech et al. (2015) k -index, which counts the number of times a user has submitted either a post or a comment that received more than k upvotes. To normalize for a user’s total activity, we divide by the total number of posts/comments that user made to form a statistic we call k -rate. We compute k -rate for $k \in \{5, 10, 50, 100\}$ for both posts and comments. While the quality statistics might leak timing information, we would like the user baseline to be as strong as possible.

Timing Models. In the pairwise ranking setting, a **Random** guess is correct half of the time. Furthermore, it is possible that the post that was created **Earlier** has a tendency to get more upvotes because it has existed longer, so choosing the submission in the pair that was posted first makes for a good baseline.

Finally, we include a **Time** baseline to quantify how well the pairing process controls for time. Instead of attempting to hand-design a set of rules, because the effects of time are complicated we choose to simply learn a time feature classifier. We use a 1-hot encoding of the minute-in-hour, hour-in-day, day-in-week, and year of the post. Instead of subtracting the resulting encodings, we concatenate to give the model access to the absolute and relative timing. The classifier we use is a one-hidden-layer neural network with 100 hidden units to capture potential non-linear relationships.

		aww	pics	cats	MA	FP	RL
Timing	Random	50.0	50.0	50.0	50.0	50.0	50.0
	Earlier	<i>51.7</i>	<i>51.1</i>	<i>49.9</i>	<i>48.9</i>	<i>48.6</i>	<i>48.7</i>
	Time	50.2	50.2	50.7	50.4	49.7	50.6
User	Type	50.6	51.2	50.7	52.8	51.8	56.1
	Activity	51.1	53.6	52.8	55.0	53.9	60.6
	Quality	54.7	55.5	52.9	60.7	55.5	<u>67.3</u>
Textual	Struct	56.2	54.8	56.5	50.9	52.3	52.5
	Topic	55.2	55.8	56.8	60.4	55.2	55.5
	DAN	58.6	58.3	58.5	62.2	57.6	59.8
	LSTM	59.4	58.8	58.7	61.0	57.0	59.1
	Bi-LSTM	59.7	58.9	59.3	61.8	57.8	59.6
	Unigram	59.7	58.6	59.5	63.0	57.6	60.8
Visual	HOG	51.7	52.8	51.9	53.5	53.5	53.5
	GIST	52.7	53.0	53.5	55.9	56.5	56.3
	ColorHist	55.3	53.7	55.6	55.0	56.5	54.5
	VGG-19	63.4	58.9	61.1	62.4	62.8	62.1
	ResNet50	<u>64.8</u>	<u>60.0</u>	<u>62.6</u>	<u>64.9</u>	<u>65.2</u>	64.2

Table 4.4: Unimodal accuracy results averaged over 15 cross-validation splits; higher accuracy is better. Bolded results are the best in the whole column and are underlined if differences are significant. Italicized results are tied for the best among their feature type. 95% CI are on average ± 0.5 and never exceed ± 1 for the non-timing features.

4.6 Results

For all experiments, we compute 15-fold cross validation accuracy in an 80/20 train/test split. We withhold 10% of training data as a validation set, which is used to optimize regularization parameters and for early stopping. Models are trained using Keras (Chollet, 2015) with the Theano (Theano Development Team, 2016) backend.

4.6.1 Unimodal Experiments

Next we assess the individual ability of each modality to predict the eventual popularity of content. The results for each dataset and feature set are given in Table 4.4. Because the classification problem is a balanced two class task, we only report accuracy.

Pairwise Ranking Controls for Time. Our objective in using pairwise ranking is to reduce the effect of time-of-posting as a confounding factor. As shown in Section 4.4, time-based features are, in general, strongly predictive of average user engagement. But in the pairwise ranking setting, we were happy to see that neither the learned time classifier nor the “earlier” baseline were able to achieve meaningful performance above random. This suggests that we are effectively controlling for time of posting.

Previous Quality Predicts Current Quality. Among user features, quality of previous submitted content is the best predictor of future success. The particular types of interactions (e.g., posts vs. comments, comment length) also seem to be less important than the absolute volume of previous interactions.

For Words, Simpler is Better. Order-sensitive and deeper models rarely outperformed the shallower, order-unaware unigram models. Interestingly, structural features performed particularly well on cats and aww; we observed that longer, story-like titles worked well in both of those communities. For all datasets, the best text-only models performed worse than the best image-only models, suggesting that visual content is more predictive of relative popularity than textual content in these communities.

For Images, More Complicated is Better. For all datasets, the best performing

	aww	pics	cats	MA	FP	RL
Time + User	54.1	54.7	52.1	58.8	54.2	64.8
All User	56.3	55.3	54.6	60.9	56.0	68.4
ResNet50	64.8	60.0	62.6	64.9	65.2	64.2
Text + Image	67.1	62.7	65.9	67.7	65.8	66.4

Table 4.5: Multimodal accuracy results averaged over 15 cross-validation splits. Higher accuracy is better, and accenting follows Table 4.4. 95% CI are on average ± 0.5 and never exceed ± 0.76 . The best unimodal model ResNet50 is generally outperformed by the multimodal model, Text + Image. User features alone (All User) generally perform better on their own than when they are combined with timing features.

image algorithm was the deep neural network ResNet50. The fact that ResNet50 outperformed its shallower counterpart VGG-19 suggests that this task is well-formulated as a computer vision task. In general, the CNN approaches performed better than the lower-level image features, though all outperformed random.

4.6.2 Multimodal Experiments

We now directly compare “cats and captions” to “creators and the clock.” In particular, given the high performance of unigram and ResNet50 features, we use ‘s Lynch et al. (2016)’s elastic net regression method to jointly represent visual and textual content, and call the model **Text + Image**. Because timing features weren’t found to be helpful when concatenated with user features (**Time + User**), we also include a concatenation of all user features, **All User**. These results are presented in Table 4.5.

In five of six cases, content features outperform the user features for relative popularity prediction. In terms of relative improvement over random, the

	aww	pics	cats	MA	FP	RL
Time + User	55.5	51.7	52.6	56.9	52.8	60.5
All User	60.4	51.0	54.3	63.1	57.9	66.0
Text + Image	65.5	66.0	67.3	62.7	62.6	65.4

Table 4.6: Heldout, out-of-domain task accuracy results; bolded are best.

magnitude of this improvement is between 245% for cats and 62% for Make-upAddiction. In five of six cases performance significantly improves when we combine text and images, indicating that this task is well-formulated as a multi-modal task. In these cases, the relative improvement over random when adding text to the best image model varies between 27% for pics and 16% for aww.

Fully-held Out, Different Distribution Test. One useful property of the models we consider is the unpaired scoring function implicitly learned in the ranking process. While this scoring function could be used to process novel submissions made to a community, it’s unclear how well patterns learned across training data would generalize to testing data. Changing linguistic (Danescu-Niculescu-Mizil et al., 2013) and visual (Wu et al., 2016) preferences of communities complicate this task considerably.

We selected 1000 pairs from each community sampled outside of the training data’s time span, and therefore out of the exact distribution of the training data. These pairs were *fully held out* meaning that we evaluated them *exactly once* for each model. The accuracy of the content model and the user/timing model in the fully-held-out settings are given in Table 4.6.

While it is difficult to extrapolate from point estimates, the fully-held out results display interesting changes in performance. In particular, while differences in performance are relatively minor (indicating that we likely didn’t over-

fit) we see a roughly 28% decrease in performance in MakeupAddiction. We find some evidence suggesting that the community has evolved during the 10 month heldout period. In particular, for the image + text models, the average posting time of the correctly-classified pairs is 11 days earlier (and closer in time to the training data) than the average posting time of incorrectly-classified pairs. Because only 1K held-out pairs are considered, the statistical significance of this potential difference cannot be established for *all* models. However, this pattern was observed across several models we considered. Collectively, these observations suggest a potentially complex relationship between training set generalizability and time.

Model Score vs. Raw Score. Using traditional ranking metrics in this pairwise setting is difficult because, as we have argued above, there is no appropriate “gold standard” ranking to compare against. The scores received on Reddit would indeed provide *a* ranking, but not an *appropriate* ranking, because those scores are biased by precisely factors like timing we have discussed and constructed our pairwise task to mitigate. As a result, applying evaluation metrics like mean reciprocal rank (MRR) or precision-at-K (p@K) that assume a ground-truth ranking is not possible.

However, we understand that readers may still be curious to know whether the ranking induced by our method has any correlation with the scores that appear on Reddit, since other work (e.g., Khosla et al. (2014), who worked with a Flickr dataset) computes similar correlations. To satisfy the curious reader, we did go ahead and compare the Spearman correlation between raw popularity and our model’s scores. For the text + image models, the observed values averaged over cross-validation splits range between $\rho = .25$ for pics and $\rho = .37$ for



Figure 4.6: Examples from FoodPorn automatically scored by the ResNet50 model. The top, middle, and bottom rows are sampled from the 99th, 50th, and 1st percentiles of model scores respectively. While lighting effects likely relate to model scores, the underperformance of the color-only classifier and the performance jump when switching from VGG-19 to ResNet50 suggest that this is a rich computer vision task. Images courtesy imgur.com.

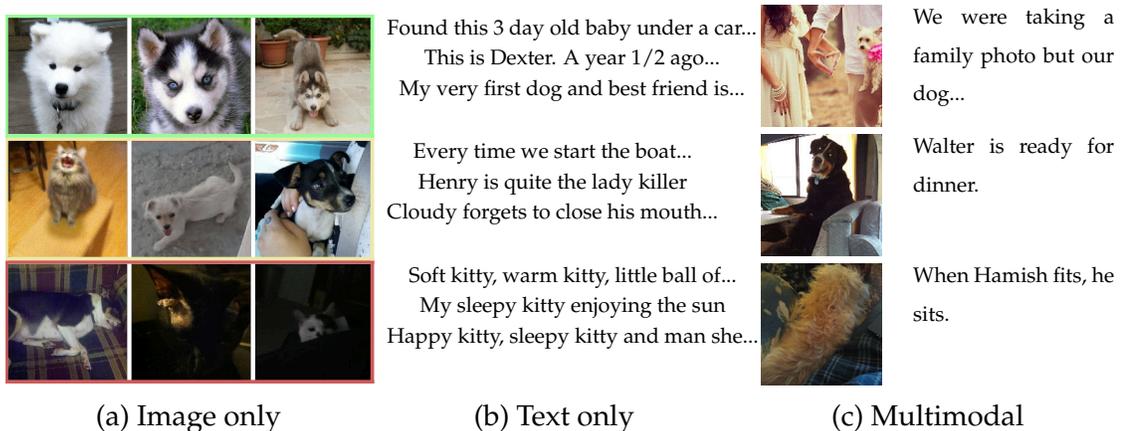


Figure 4.7: Examples from one train/test split of aww scored by the ResNet50 model, the unigram model, and the text + image model. The top, middle, and bottom rows are sampled from the 99th, 50th, and 1st percentiles respectively. Images courtesy imgur.com.

MakeupAddiction. In general, the correlations we observe are somewhat lower than those of Khosla et al. (2014)’s image-based model; whether the differences are due to the models or to the different domains is an open question.

4.7 Analysis of aww

We now qualitatively analyze the models’ performance on aww, though a similar analysis could be performed on any community (e.g., Figure 4.6 shows image examples from FoodPorn). Figure 4.7 shows several test examples scored by the image-only, text-only, and multimodal models from one of the aww cross-validation splits.

Figure 4.7a, which displays good, okay, and bad images as scored by ResNet50, illustrates that lighting is important. The model tends to assign lower scores in cases where an animal’s face isn’t visible. Having the animal taking up a majority of the image also seems to be important, though this could be an artifact of our resizing procedure. Also, we noticed that a disproportionate number of highly scored images were of dogs; among the cross-validation split we considered, in fact, the top ten images were all dogs. The model, and potentially the community, might be favoring particular types of animals.

To examine this possibility, we turn our focus to more interpretable object detections. Specifically, we turn to the canonical 1K ImageNet classes, which consist of a surprisingly high number of types of animals, e.g., 120/1000 classes are different types of dogs. As such, these classes are well-suited to analyzing aww. We extracted the pre-softmax input for each ImageNet class according to

ResNet50 for each image¹⁹ These features are the un-normalized log probabilities for each of the 1K ImageNet classes. For each of the 15 cross-validation splits, we computed the average Pearson correlation between our model’s score and the object detection features.

After applying Bonferroni-correction to our confidence intervals to account for the fact that there are 1K possible correlations, we observed many significant results. Among the 250 most common detections, the object-like features most correlated with success were “golden retriever,” “dingo,” and “labrador retriever” ($R = .23, .21, .19$, respectively, $p \ll .01$ that there is a true correlation). There were also dog breed features associated with failure, including “miniature schnauzer,” “maltese dog,” and “affenpinscher” ($R = -.23, -.21, -.21$, $p \ll .01$). Interestingly, non-bulldog terriers fared poorly; all 15 were negatively correlated with model score, though only 12/15 were significantly so. In contrast, 5/5 retriever classes were significantly correlated with higher scores. For cats, “cheetah” and “lion” features positively correlated ($R = .18, .09$) while “tabby,” “egyptian cat,” and “persian cat” features were all negatively correlated ($R = -.1, -.11, -.17$).

The story with text on aww is a simpler; Figure 4.7b shows that longer captions generally do better, and it also helps to have a story. Unigrams like saved ($\beta = .50$), wife ($\beta = .43$), roommate ($\beta = .42$), and cancer ($\beta = .41$), and are among the most predictive of success. Interestingly, sleeping animals seem to be predictive of failure, with unigrams like sleepy ($\beta = -.58$), sleeping ($\beta = -.47$), laying ($\beta = -.47$), and nap ($\beta = -.43$) being among the most predictive of failure.

When image and text features are combined, performance improves over

¹⁹Weights after the softmax transformation also produced some significant results, but the pre-softmax weights are known to contain more fine-grained information (Bucilua et al., 2006)

each by themselves, which suggests that the patterns discussed contain information orthogonal for predictive purposes. Because we simply concatenate image and text features rather than modeling interactions directly, the multimodal patterns likely mirror the unimodal patterns discussed here.

4.8 Additional Related Work

Content has been used to predict popularity in the past. Language (Petrovic et al., 2011; Hong et al., 2011; Guerini et al., 2011; Bandari et al., 2012; Danescu-Niculescu-Mizil et al., 2012; Artzi et al., 2012; Sun et al., 2013; Tan et al., 2014; Tsur and Rappoport, 2012), images (Khosla et al., 2014; Deza and Parikh, 2015; Wu et al., 2016), video (Shamma et al., 2011; Figueiredo, 2013; Pinto et al., 2013), or a combination of multiple modalities (Yamaguchi et al., 2014; McParlane et al., 2014; Gelli et al., 2015; Hu et al., 2016a; Chen, 2016) have been used for this task. Some previous work has controlled for, rather than modeled, multimodal content (Borghol et al., 2012; Lakkaraju et al., 2013). Our work builds upon previous studies that attempt to predict or analyze *crowd-level* preferences (Khosla et al., 2014; Figueiredo et al., 2014; Bakhshi et al., 2014; Bakhshi and Gilbert, 2015; Stoddard, 2015; Schifanella et al., 2015; Deza and Parikh, 2015; Mazloom et al., 2016; Almgren et al., 2016; Zakrewsky et al., 2016), as opposed to *user-level* preferences (Zhong et al., 2015). Glenski and Stoddard²⁰ describe human experiments similar to ours. While the setting we examine is different (e.g., we apply more stringent time controls), it was interesting to see that their human trial results were similar to ours.

²⁰<https://goo.gl/9M6Ioh>

Noisy Rich-get-richer Processes. Timing (Borghol et al., 2012; Lakkaraju et al., 2013), and even early random positive or negative treatments (Weninger et al., 2015) can affect the popularity of social media content. Salganik et al. (Salganik et al., 2006) show that while content does matter to an extent, presenting different orderings of songs to users results in wildly different most and least popular music. These effects likely underpin the widespread underprovision on Reddit (Gilbert, 2013), which causes “Reddit [to overlook] 52% of the most popular links the first time they were submitted.” Undoubtedly, content can never perfectly predict community response.

Social Features for Eventual Popularity. Social connections (Lerman and Hogg, 2010) and author identity (Suh et al., 2010) also effect the popularity of content. Solomon and Herman (Solomon and Herman, 1977) demonstrate that individuals with higher status are more likely to be recipients of prosocial behavior. In our case, this could mean higher status individuals in a community receive upvotes as a result of their celebrity status. Khosla et al. (2014) consider a simple set of social features of their Flickr dataset, and find that social features are significantly more predictive of popularity than image features when not controlling for user identity.

4.9 Conclusion and Future Work

In this work, we motivated the task of relative popularity prediction as a means of controlling for time. We also demonstrated that incorporating multimodal features generally resulted in improved performance. Future work in modeling could consider more sophisticated models of textual and visual interaction.

Also, it would be interesting to investigate visual trends within communities over time. Designing a model to identify “timely” or trend-setting image features is a promising avenue for future work.

Popularity prediction, too, is only one social factor of interest to moderators of multimodal communities. The text of comments, for example, offers a more fine-grained measure of community response than upvotes. Text features like sentiment could also be predicted from content in a similar time-controlled setting.

While we’ve provided evidence that there exist online communities wherein visual and textual content predict popularity more successfully than social features, it is important to point out the results presented here might not generalize to other communities, e.g., ones off of Reddit. We suspect that social connections are less salient on Reddit, which seems more centered on the content. Instagram, for example, is a social network based on image content wherein identity likely matters more. However, even on Reddit itself, we observed a case in *Reddit-Laqueristas* where our intuitions proved to be incorrect: celebrity-status/social features were more predictive than content in that subreddit.

Another caveat: while sampling pairs of posts made in quick succession provided good timing/ordering controls for us, in other settings there might not be enough posts to warrant such a sampling technique.

In the end, predicting what becomes popular in any given community requires accounting for timing, content, identity, social structure, and self-reinforcing rich-get-richer processes. While the relative predictive power of each varies on a case-by-case basis, we hope the results presented here encour-

age practitioners to investigate content-driven models in the face of complex confounding factors.

4.10 Brief Retrospective

Since the publication of this work in 2017, there have been two developments. First, upon learning about potential missing data, we replicated key results from these experiments.²¹ Second, Ding et al. (2019) applied our pretrained models to an out-of-domain popularity prediction task on Instagram. Notably, our model performed best among methods not specifically optimized for that domain. In particular, our models outperformed Khosla et al. (2014) by 10 accuracy points on a similar pairwise task (relative to a 50% random guessing baseline), despite being trained on an order of magnitude fewer images.

²¹<http://www.cs.cornell.edu/~jhessel/reddit/gaps.html>

CHAPTER 5

UNDERSTANDING: QUANTIFYING THE VISUAL CONCRETENESS OF CONCEPTS

5.1 Brief Overview

Multimodal machine learning algorithms aim to learn visual-textual correspondences. Previous work suggests that concepts with *concrete* visual manifestations may be easier to learn than concepts with abstract ones. We give an algorithm for automatically computing the visual concreteness of words and topics within multimodal datasets. We apply the approach in four settings, ranging from image captions to images/text scraped from historical books. In addition to enabling explorations of concepts in multimodal datasets, our concreteness scores predict the capacity of machine learning algorithms to learn textual/visual relationships. We find that 1) concrete concepts are indeed easier to learn; 2) the large number of algorithms we consider have similar failure cases; 3) the precise positive relationship between concreteness and performance varies between datasets. We conclude with recommendations for using concreteness scores to facilitate future multimodal research.

The work in this chapter is joint with David Mimno and Lillian Lee, and was published in Hessel et al. (2018).

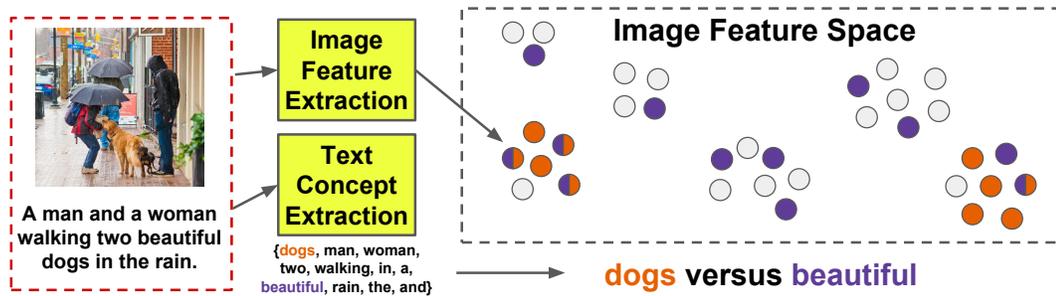


Figure 5.1: Demonstration of visual concreteness estimation on an example from the COCO dataset. The degree of visual clustering of textual concepts is measured using a nearest neighbor technique. The concreteness of “dogs” is greater than the concreteness of “beautiful” because images associated with “dogs” are packed tightly into two clusters, while images associated with “beautiful” are spread evenly.¹

5.2 Introduction

Text and images are often used together to serve as a richer form of content. For example, news articles may be accompanied by photographs or infographics; images shared on social media are often coupled with descriptions or tags; and textbooks include illustrations, photos, and other visual elements. The ubiquity and diversity of such “text+image” material (henceforth referred to as *multi-modal* content) suggest that, from the standpoint of sharing information, images and text are often natural complements.

Ideally, machine learning algorithms that incorporate information from both text and images should have a fuller perspective than those that consider either text or images in isolation. But Hill and Korhonen (2014b) observe that for their particular multimodal architecture, the level of *concreteness* of a concept being represented — intuitively, the idea of a *dog* is more concrete than that of *beauty* — affects whether multimodal or single-channel representations are more effective. In their case, concreteness was derived for 766 nouns and verbs from a

fixed psycholinguistic database of human ratings.

In contrast, we introduce an adaptive algorithm for characterizing the visual concreteness of all the concepts indexed textually (e.g., “dog”) in a given multi-modal dataset. Our approach is to leverage the geometry of image/text space. Intuitively, a visually concrete concept is one associated with more locally similar sets of images; for example, images associated with “dog” will likely contain dogs, whereas images associated with “beautiful” may contain flowers, sunsets, weddings, or an abundance of other possibilities — see Fig. 5.1.

Allowing concreteness to be dataset-specific is an important innovation because concreteness is contextual. For example, in one dataset we work with, our method scores “London” as highly concrete because of a preponderance of iconic London images in it, such as Big Ben and double-decker buses; whereas for a separate dataset, “London” is used as a geotag for diverse images, so the same word scores as highly non-concrete.

In addition to being dataset-specific, our method readily scales, does not depend on an external search engine, and is compatible with both discrete and continuous textual concepts (e.g., topic distributions).

Dataset-specific visual concreteness scores enable a variety of purposes. In this paper, we focus on using them to: 1) explore multimodal datasets; and 2) predict how easily concepts will be learned in a machine learning setting. We apply our method to four large multimodal datasets, ranging from image captions to image/text data scraped from Wikipedia,² to examine the relationship between concreteness scores and the performance of machine learning al-

²We release our Wikipedia and British Library data at <http://www.cs.cornell.edu/~jhessel/concreteness/concreteness.html>

gorithms. Specifically, we consider the cross-modal retrieval problem, and examine a number of NLP, vision, and retrieval algorithms. Across all 320 significantly different experimental settings (= 4 datasets \times 2 image-representation algorithms \times 5 textual-representation algorithms \times 4 text/image alignment algorithms \times 2 feature pre-processing schemes), we find that more concrete instances are easier to retrieve, and that different algorithms have similar failure cases. Interestingly, the relationship between concreteness and retrievability varies significantly based on dataset: some datasets appear to have a linear relationship between the two, whereas others exhibit a concreteness threshold beyond which retrieval becomes much easier.

We believe that our work can have a positive impact on future multimodal research. §5.9 gives more detail, but in brief, we see implications in (1) evaluation — more credit should perhaps be assigned to performance on non-concrete concepts; (2) creating or augmenting multimodal datasets, where one might *a priori* consider the desired relative proportion of concrete vs. non-concrete concepts; and (3) *curriculum learning* (Bengio et al., 2009), where ordering of training examples could take concreteness levels into account.

5.3 Related Work

Applying machine learning to understand visual-textual relationships has enabled a number of new applications, e.g., better accessibility via automatic generation of alt text (Garcia et al., 2016), cheaper training-data acquisition for computer vision (Joulin et al., 2016; Veit et al., 2017), and cross-modal retrieval systems, e.g., Rasiwasia et al. (2010); Costa Pereira et al. (2014b).

Multimodal datasets often have substantially differing characteristics, and are used for different tasks (Baltrušaitis et al., 2018). Some commonly used datasets couple images with a handful of unordered tags (Barnard et al., 2003; Cusano et al., 2004; Grangier and Bengio, 2008; Chen et al., 2013a) or short, literal natural language captions (Farhadi et al., 2010; Ordóñez et al., 2011; Kulka-rni et al., 2013; Fang et al., 2015). In other cross-modal retrieval settings, images are paired with long, only loosely thematically-related documents. (Khan et al., 2009; Socher and Fei-Fei, 2010; Jia et al., 2011; Zhuang et al., 2013). We provide experimental results on both types of data.

Concreteness in datasets has been previously studied in either text-only cases (Turney et al., 2011; Hill et al., 2013) or by incorporating human judgments of perception into models (Silberer and Lapata, 2012; Hill and Korhonen, 2014a). Other work has quantified characteristics of concreteness in multimodal datasets (Yatskar et al., 2013; Young et al., 2014; Hill et al., 2014; Hill and Korhonen, 2014b; Kiela and Bottou, 2014; Jas and Parikh, 2015; Lazaridou et al., 2015; Silberer et al., 2016; Lu et al., 2017; Bhaskar et al., 2017). Most related to our work is that of Kiela et al. (2014); the authors use Google image search to collect 50 images each for a variety of words and compute the average cosine similarity between vector representations of returned images. In contrast, our method can be tuned to specific datasets without reliance on an external search engine. Other algorithmic advantages of our method include that: it more readily scales than previous solutions, it makes relatively few assumptions regarding the distribution of images/text, it normalizes for word frequency in a principled fashion, and it can produce confidence intervals. Finally, the method we propose can be applied to both discrete and continuous concepts like topic distributions.

5.4 Quantifying Visual Concreteness

To compute visual concreteness scores, we adopt the same general approach as Kiela et al. (2014): for a fixed text concept (i.e., a word or topic), we measure the variance in the corresponding visual features. The method is summarized in Figure 5.1.

5.4.1 Concreteness of discrete words

We assume as input a multimodal dataset of n images represented in a space where nearest neighbors may be computed. Additionally, each image is associated with a set of discrete words/tags. We write w_v for the set of words/tags associated with image v , and V_w for the set of all images associated with a word w . For example, if the v^{th} image is of a dog playing frisbee, w_v might be {frisbee, dog, in, park}, and $v \in V_{\text{park}}$.

Our goal is to measure how “clustered” a word is in image feature space. Specifically, we ask: for each image $v \in V_w$, how often are v ’s nearest neighbors also associated with w ? We thus compute the expected value of MNI_w^k , the number of mutually neighboring images of word w :

$$\mathbb{E}_{P_{\text{data}}}[\text{MNI}_w^k] = \frac{1}{|V_w|} \sum_{v \in V_w} |\text{NN}^k(v) \cap V_w|, \quad (5.1)$$

where $\text{NN}^k(v)$ denotes the set of v ’s k nearest neighbors in image space.

While Equation 5.1 measures clusteredness, it does not properly normalize for frequency. Consider a word like “and”; we expect it to have low concreteness, but its associated images will share neighbors simply because “and” is a

frequent unigram. To correct for this, we compute the *concreteness* of a word as the ratio of $\mathbb{E}[\text{MNI}_w^k]$ under the true distribution of the image data to a random distribution of the image data:

$$\text{concreteness}(w) = \frac{\mathbb{E}_{P_{data}}[\text{MNI}_w^k]}{\mathbb{E}_{P_{random}}[\text{MNI}_w^k]} \quad (5.2)$$

While the denominator of this expression can be computed in closed form, we use $\mathbb{E}_{P_{random}}[\text{MNI}_w^k] \approx \frac{k|V_w|}{n}$; this approximation is faster to compute and is negligibly different from the true expectation in practice.

5.4.2 Extension to continuous topics

We extend the definition of concreteness to continuous concepts, so that our work applies also to topic model outputs; this extension is needed because the intersection in Equation 5.1 cannot be directly applied to real values. Assume we are given a set of topics T and an image-by-topic matrix $Y \in \mathbb{R}^{n \times |T|}$, where the v^{th} row³ is a topic distribution for the text associated with image v , i.e., $Y_{ij} = P(\text{topic } j | \text{image } i)$. For each topic t , we compute the average topic weight for each image v 's neighbors, and take a weighted average as:

$$\text{concreteness}(t) = \frac{n}{k} \cdot \frac{\sum_{v=1}^n [Y_{vt} \sum_{j \in \text{NN}^k(v)} Y_{jt}]}{(\sum_{v=1}^n Y_{vt})^2} \quad (5.3)$$

Note that Equations 5.1 and 5.3 are computations of means. Therefore, confidence intervals can be computed in both cases either using a normality assumption or bootstrapping.

³The construction is necessarily different for different types of datasets, as described in §5.5.

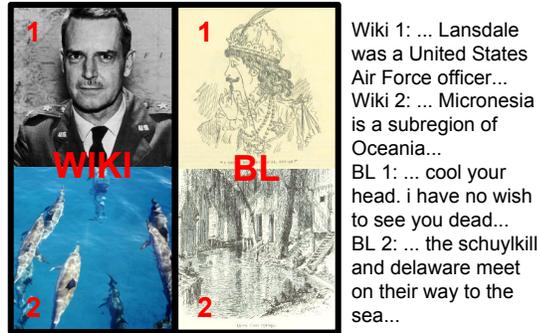


Figure 5.2: Examples of text and images from our new Wiki/BL datasets.

5.5 Datasets

We consider four datasets that span a variety of multimodal settings. Two are publicly available and widely used (COCO/Flickr); we collected and pre-processed the other two (Wiki/BL). The Wikipedia and British Library sets are available for download at <http://www.cs.cornell.edu/~jhessel/concreteness/concreteness.html>. Dataset statistics are given in Table 5.1, and summarized as follows:

Wikipedia (Wiki). We collected a dataset consisting of 192K articles from the English Wikipedia, along with the 549K images contained in those articles. Following Wilson’s popularity filtering technique,⁴ we selected this subset of Wikipedia by identifying articles that received at least 50 views on March 5th, 2016.⁵ To our knowledge, the previous largest publicly available multimodal Wikipedia dataset comes from ImageCLEF’s 2011 retrieval task (Popescu et al., 2010), which consists of 137K images associated with English articles.

Images often appear on multiple pages: an image of the Eiffel tower might appear on pages for Paris, for Gustave Eiffel, and for the tower itself.

⁴<https://goo.gl/B11yyO>

⁵The articles were extracted from an early March, 2016 data dump.

	# Images	Mean Len	Train	Test
Wiki	549K	1397.8	177K	10K
BL	405K	2269.6	69K	5K
COCO	123K	10.5	568K	10K
Flickr	754K	9.0	744K	10K

Table 5.1: Dataset statistics: total number of images, average text length in words, and size of the train/test splits we use in §5.7.

Historical Books from British Library (BL). The British Library has released a set of digitized books (British Library Labs, 2016) consisting of 25M pages of OCRred text, alongside 500K+ images scraped from those pages of text. The release splits images into four categories; we ignore “bound covers” and “embellishments” and use images identified as “plates” and “medium sized.” We associated images with all text within a 3-page window.

This raw data collection is noisy. Many books are not in English, some books contain far more images than others, and the images themselves are of varying size and rotation. To combat these issues we only keep books that have identifiably English text; for each cross-validation split in our machine-learning experiments (§5.7) we sample at most 10 images from each book; and we use *book-level* holdout so that no images/text in the test set are from books in the training set.

Captions and Tags. We also examine two popular existing datasets: Microsoft COCO (captions) (Lin et al., 2014) (**COCO**) and MIRFLICKR-1M (tags) (Huiskes et al., 2010) (**Flickr**). For COCO, we construct our own training/validation splits from the 123K images, each of which has 5 captions. For Flickr, as an initial preprocessing step we only consider the 7.3K tags that appear at least 200 times, and the 754K images that are associated with at least 3 of the 7.3K valid tags.

	Wiki (Topics)	MSCOCO (Unigrams)	Flickr (Unigrams)
Most Concrete	hockey 170.2	polar 296	écureuil 1647
	tennis 148.9	ben 247	cheetah 1629
	nintendo 86.3	stir 166	rongeur 1605
	guns 81.9	contents 160	pampaargentino 1600
	baseball 80.9	steam 157	scuridae 1588
	wrestling1 76.7	magnets 154	pampaargentina 1586
	wrestling2 71.4	sailboats 150	rodentia 1544
	software 70.4	wing 147	bodybuilding 1542
	auto racing 60.9	airlines 146	bodybuilder 1520
	currency 58.8	marina 137	sanantoniodeseaeco 1484
Least Concrete	australia 1.95	image 1.11	2007 2.16
	mexico 1.81	appears 1.09	activeasmnhtakly 2.11
	police 1.73	weird 1.06	artisticexpression 2.03
	law 1.71	appear 0.89	geotagged 1.92
	male names 1.65	photographed 0.75	2008 1.88
	community 1.58	thing 0.59	explored 1.86
	history 1.52	interesting 0.52	2009 1.75
	time 1.47	possibly 0.45	nikon 1.57
	months 1.46	somewhere 0.40	canon 1.57
	linguistics 1.22	caption 0.22	explore 1.55

Figure 5.3: Examples of the most and least concrete words/topics from Wiki, COCO, and Flickr, along with example images associated with each highlighted word/topic.

5.6 Validation of Concreteness Scoring

We apply our concreteness measure to the four datasets. For COCO and Flickr, we use unigrams as concepts, while for Wiki and BL, we extract 256-dimensional topic distributions using Latent Dirichlet Allocation (LDA) (Blei et al., 2003). For BL, topic distributions are derived from text in the aforementioned 3 page window; for Wiki, for each image, we compute the mean topic distribution of all articles that image appears in; for Flickr, we associate images with all of their tags; for COCO, we concatenate all captions for a given image. For computing concreteness scores for COCO/Flickr, we only consider unigrams associated with at least 100 images, so as to ensure the stability of MNI as defined in Equation 5.1.

We extract image features from the pre-softmax layer of a deep convolutional neural network, ResNet50 (He et al., 2016b), pretrained for the ImageNet classification task (Deng et al., 2009); this method is known to be a strong baseline (Sharif Razavian et al., 2014).⁶ For nearest neighbor search, we use the Annoy

⁶We explore different image/text representations in later sections.

library,⁷ which computes approximate kNN efficiently. We use $k = 50$ nearest neighbors, though the results presented are stable for reasonable choices of k , e.g., $k = 25, 100$.

5.6.1 Concreteness and human judgments

Following Kiela et al. (2014), we borrow a dataset of human judgments to validate our concreteness computation method.⁸ The concreteness of words is a topic of interest in psychology because concreteness relates to a variety of aspects of human behavior, e.g., language acquisition, memory, etc. (Schwanenflugel and Shoben, 1983; Paivio, 1991; Walker and Hulme, 1999; De Groot and Keijzer, 2000).

We compare against the human-gathered unigram concreteness judgments provided in the USF Norms dataset (USF) (Nelson et al., 2004); for each unigram, raters provided judgments of its concreteness on a 1-7 scale. For Flickr/COCO, we compute Spearman correlation using these per-unigram scores (the vocabulary overlap between USF and Flickr/COCO is 1.3K/1.6K), and for Wiki/BL, we compute topic-level human judgment scores via a simple average amongst the top 100 most probable words in the topic.

As a null hypothesis, we consider the possibility that our concreteness measure is simply mirroring frequency information.⁹ We measure frequency for each dataset by measuring how often a particular word/topic appears in it. A

⁷github.com/spotify/annoy

⁸Note that because concreteness of words/topics varies from dataset to dataset, we don't expect one set of human judgments to correlate perfectly with our concreteness scores. However, partial correlation with human judgment offers a common-sense "reality check."

⁹We return to this hypothesis in §5.7.1 as well; there, too, we find that concreteness and frequency capture different information.

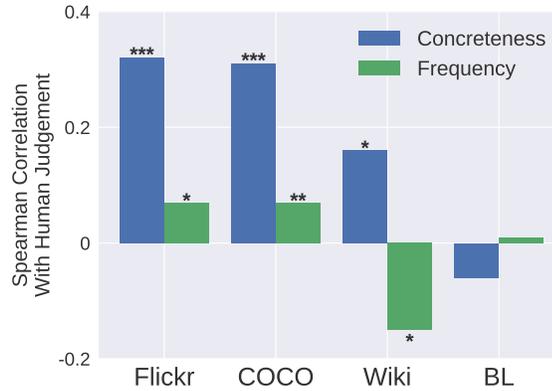


Figure 5.4: Spearman correlations between human judgment (USF) and our algorithm’s outputs, and dataset frequency. In the case of Flickr/COCO/WIKI our concreteness scores correlate with human judgement to a greater extent than frequency. For BL, neither frequency nor our concreteness measure is correlated with human judgement. ***/**/* := $p < .001/.01/.05$

useful concreteness measure should correlate with USF more than a simple frequency baseline does.

For COCO/Flickr/Wiki, concreteness scores output by our method positively correlate with human judgments of concreteness more than frequency does (see Figure 5.4). For COCO, this pattern holds even when controlling for part-of-speech (not shown), whereas Flickr adjectives are not correlated with USF. For BL, neither frequency nor our concreteness scores are significantly correlated with USF. Thus, in three of our four datasets, our measure tends to predict human concreteness judgments better than frequency.

Concreteness and frequency. While concreteness measures correlate with human judgment better than frequency, we do expect *some* correlation between a word’s frequency and its concreteness (Gorman, 1961). In all cases, we observe a moderate-to-strong positive correlation between infrequency and concreteness ($\rho_{wiki}, \rho_{coco}, \rho_{flickr}, \rho_{bl} = .06, .35, .40, .71$) indicating that rarer words/topics are more concrete, in general. However, the correlation is not perfect, and con-

creteness and frequency measure different properties of words.

5.6.2 Concreteness within datasets

Figure 5.3 gives examples from Wiki, COCO, and Flickr illustrating the concepts associated with the smallest and largest concreteness scores according to our method.¹⁰ The scores often align with intuition, e.g., for Wiki, sports topics are often concrete, whereas country-based or abstract-idea-based topics are not.¹¹ For COCO, *polar* (because of polar bears) and *ben* (because of Big Ben) are concrete; whereas *somewhere* and *possibly* are associated with a wide variety of images.

Concreteness scores form a continuum, making explicit not only the extrema (as in Figure 5.3) but also the middle ground, e.g., in COCO, “wilderness” (rank 479) is more visually concrete than “outside” (rank 2012). Also, dataset-specific intricacies that are not obvious *a priori* are highlighted, e.g., in COCO, 150/151 references to “magnets” (rank 6) are in the visual context of a refrigerator (making “magnets” visually concrete) though the converse is not true, as both “refrigerator” (rank 329) and “fridge” (rank 272) often appear without magnets; 61 captions in COCO are exactly “There is no image here to provide a *caption* for,” and this dataset error is made explicit through concreteness score computations.

¹⁰The BL results are less interpretable and are omitted for space reasons.

¹¹Perhaps fittingly, the “linguistics” topic (top words: term, word, common, list, names, called, form, refer, meaning) is the least visually concrete of all 256 topics.

5.6.3 Concreteness varies across datasets

To what extent are the concreteness scores dataset-specific? To investigate this question, we compute the correlation between Flickr and COCO unigram concreteness scores for 1129 overlapping terms. While the two are positively correlated ($\rho = .48, p < .01$) there are many exceptions that highlight the utility of computing dataset-independent scores. For instance, “London” is extremely concrete in COCO (rank 9) as compared to in Flickr (rank 1110). In COCO, images of London tend to be iconic (i.e., Big Ben, double decker buses); in contrast, “London” often serves as a geotag for a wider variety of images in Flickr. Conversely, “watch” in Flickr is concrete (rank 196) as it tends to refer to the timepiece, whereas “watch” is not concrete in COCO (rank 958) as it tends to refer to the verb; while these relationships are not obvious *a priori*, our concreteness method has helped to highlight these usage differences between the image tagging and captioning datasets.

5.7 Learning Image/Text Correspondences

Previous work suggests that incorporating visual features for less concrete concepts can be harmful in word similarity tasks (Hill and Korhonen, 2014b; Kiela et al., 2014; Kiela and Bottou, 2014; Hill et al., 2014). However, it is less clear if this intuition applies to more practical tasks (e.g., retrieval), or if this problem can be overcome simply by applying the “right” machine learning algorithm. We aim to tackle these questions in this section.

The learning task. The task we consider is the construction of a joint embed-

ding of images and text into a shared vector space. Truly corresponding image/text pairs (e.g., if the text is a caption of that image) should be placed close together in the new space relative to image/text pairs that do not match. This task is a good representative of multimodal learning because computing a joint embedding of text and images is often a “first step” for downstream tasks, e.g., cross-modal retrieval (Rasiwasia et al., 2010), image tagging (Chen et al., 2013a), and caption generation (Kiros et al., 2015).

Evaluations. Following previous work in cross-modal retrieval, we measure performance using the top- $k\%$ hit rate (also called recall-at- k -percent, $R@k\%$; higher is better). Cross-modal retrieval can be applied in either direction, i.e., searching for an image given a body of text, or vice-versa. We examine both the image-search-text and text-search-image cases. For simplicity, we average retrieval performance from both directions, producing a single metric;¹² higher is better.

Visual Representations. Echoing Wei et al. (2016), we find that features extracted from convolutional neural networks (CNNs) outperform classical computer vision descriptors (e.g., color histograms) for multimodal retrieval. We consider two different CNNs pretrained on different datasets: ResNet50 features trained on the ImageNet classification task (**RN-Imagenet**), and InceptionV3 (Szegedy et al., 2015) trained on the OpenImages (Krasin et al., 2017) image tagging task (**I3-OpenImages**).

Text Representations. We consider sparse **unigram** and **tfidf** indicator vectors. In both cases, we limit the vocabulary size to 7.5K. We next consider latent-

¹²Averaging is done for ease of presentation; the performance in both directions is similar. Among the parametric approaches (LS/DCCA/NS) across all datasets/NLP algorithms, the mean difference in performance between the directions is 1.7% (std. dev=2%).

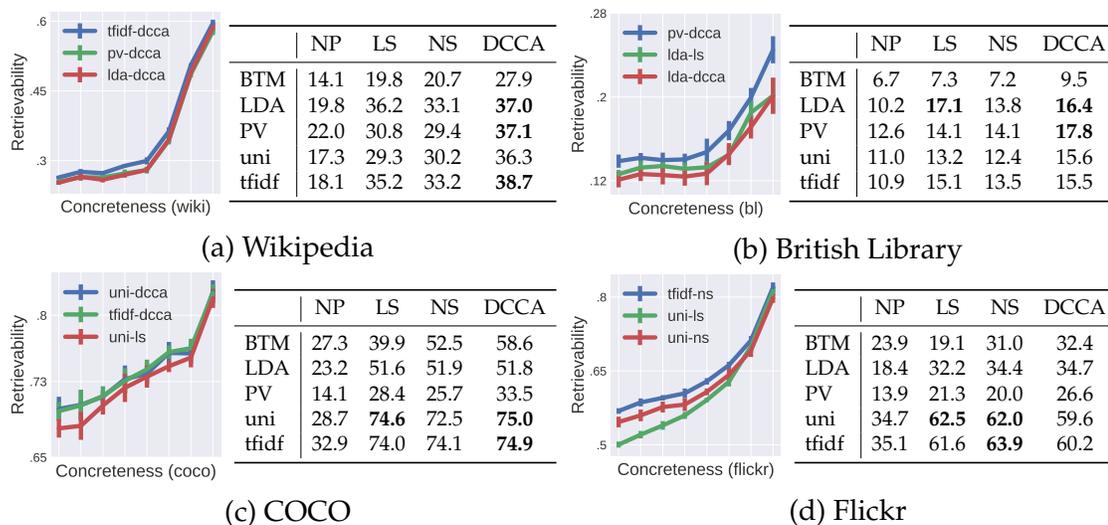


Figure 5.5: Concreteness scores versus retrievability (plotted) for each dataset, along with Recall at 1% (in tables, higher is better) for each algorithm combination. Tables give average retrieval performance over 10-fold cross-validation for each combination of NLP/alignment algorithm; the top three performing combinations are bolded. The concreteness versus retrievability curves are plotted for the top-3 performing algorithms, though similar results hold for all algorithms. Our concreteness scores and performance are positively correlated, though the shape of the relationship between the two differs from dataset to dataset (note the differing scales of the y-axes). All results are for RN-ImageNet; the similar I3-OpenImages results are omitted for space reasons.

variable bag-of-words models, including LDA (Blei et al., 2003) (256 topics, trained with Mallet (McCallum, 2002)) a specialized biterm topic model (**BTM**) (Yan et al., 2013) for short texts (30 topics), and paragraph vectors (**PV**) (Le and Mikolov, 2014) (PV-DBOW version, 256 dimensions, trained with Gensim (Řehůřek and Sojka, 2010)).¹³

Alignment of Text and Images. We explore four algorithms for learning correspondences between image and text vectors. We first compare against Hodosh et al. (2013)’s nonparametric baseline (**NP**), which is akin to a nearest-neighbor search. This algorithm is related to the concreteness score algorithm we previ-

¹³We also ran experiments encoding text using order-aware recurrent neural networks, but we did not observe significant performance differences.

ously introduced in that it exploits the geometry of the image/text spaces using nearest-neighbor techniques. In general, performance metrics for this algorithm provide an estimate of how “easy” a particular task is in terms of the initial image/text representations.

We next map image features to text features via a simple linear transformation. Let (t_i, v_i) be a text/image pair in the dataset. We learn a linear transformation W that minimizes

$$\sum_i \|W f_{\text{image}}(v_i) - f_{\text{text}}(t_i)\|_2^2 + \lambda \|W\|_F \quad (5.4)$$

for feature extraction functions f_{image} and f_{text} , e.g., RN-ImageNet/LDA. It is possible to map images onto text as in Equation 5.4, or map text onto images in an analogous fashion. We find that the directionality of the mapping is important. We train models in both directions, and combine their best-performing results into a single least-squares (**LS**) model.

Next we consider Negative Sampling (**NS**), which balances two objectives: true image/text pairs should be close in the shared latent space, while randomly combined image/text pairs should be far apart. For a text/image pair (t_i, v_i) , let $s(t_i, v_i)$ be the cosine similarity of the pair in the shared space. The loss for a single positive example (t_i, v_i) given a negative sample (t'_i, v'_i) is

$$h(s(t_i, v_i), s(t_i, v'_i)) + h(s(t_i, v_i), s(t'_i, v_i)) \quad (5.5)$$

for the hinge function $h(p, n) = \max\{0, \alpha - p + n\}$. Following Kiros et al. (2015) we set $\alpha = .2$.

Finally, we consider Canonical Correlation Analysis (**CCA**), which projects image and text representations down to independent dimensions of high multimodal correlation. CCA-based methods are popular within the IR community

for learning multimodal embeddings (Costa Pereira et al., 2014b; Gong et al., 2014). We use Wang et al. (2015b)’s stochastic method for training deep CCA (Andrew et al., 2013) (DCCA), a method that is competitive with traditional kernel CCA (Wang et al., 2015a) but less memory-intensive to train.

Training details. LS, NS, and DCCA were implemented using Keras (Chollet, 2015).¹⁴ In total, we examine all combinations of: four datasets, five NLP algorithms, two vision algorithms, four cross-modal alignment algorithms, and two feature preprocessing settings; each combination was run using 10-fold cross-validation.

Absolute retrieval quality. The tables in Figure 5.5 contain the retrieval results for RN-ImageNet image features across each dataset, alignment algorithm, and text representation scheme. We show results for $R@1\%$, but $R@5\%$ and $R@10\%$ are similar. I3-OpenImages image features underperform relative to RN-ImageNet and are omitted for space reasons, though the results are similar.

The BL corpus is the most difficult of the datasets we consider, yielding the lowest retrieval scores. The highly-curated COCO dataset appears to be the easiest, followed by Flickr and then Wikipedia. No single algorithm combination is “best” in all cases.

¹⁴We used Adam (Kingma and Ba, 2015), batch normalization (Ioffe and Szegedy, 2015), and ReLU activations. Regularization and architectures (e.g., number of layers in DCCA/NS, regularization parameter in LS) were chosen over a validation set separately for each cross-validation split. Training is stopped when retrieval metrics decline over the validation set. All models were trained twice, using both raw features and zero-mean/unit-variance features.

5.7.1 Concreteness scores and performance

We now examine the relationship between retrieval performance and concreteness scores. Because concreteness scores are on the word/topic level, we define a *retrievability* metric that summarizes an algorithm’s performance on a given concept; for example, we might expect that retrievability(dog) is greater than retrievability(beautiful).

Borrowing the $R@1\%$ metric from the previous section, we let $\mathbb{I}[r_i < 1\%]$ be an indicator variable indicating that test instance i was retrieved correctly, i.e., $\mathbb{I}[r_i < 1\%]$ is 1 if the the average rank r_i of the image-search-text/text-search-image directions is better than 1%, and 0 otherwise. Let s_{ic} be the affinity of test instance i to concept c . In the case of topic distributions, s_{ic} is the proportion of topic c in instance i ; in the case of unigrams, s_{ic} is the length-normalized count of unigram c on instance i . Retrievability is defined using a weighted average over test instances i as:

$$\text{retrievability}(c) = \frac{\sum_i s_{ic} \cdot \mathbb{I}[r_i < 1\%]}{\sum_i s_{ic}} \quad (5.6)$$

The retrievability of c will be higher if instances more associated with c are more easily retrieved by the algorithm.

Retrievability vs. Concreteness. The graphs in Figure 5.5 plot our concreteness scores versus retrievability of the top 3 performing NLP/alignment algorithm combinations for all 4 datasets. In all cases, there is a strong positive correlation between concreteness and retrievability, which provides evidence that more concrete concepts are easier to retrieve.

The shape of the concreteness-retrievability curve appears to vary between datasets more than between algorithms. In COCO, the relationship between the

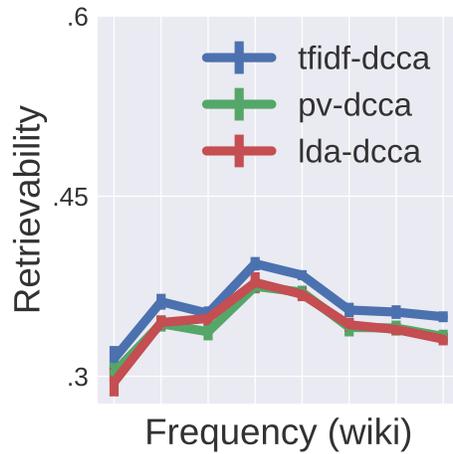


Figure 5.6: Wikipedia

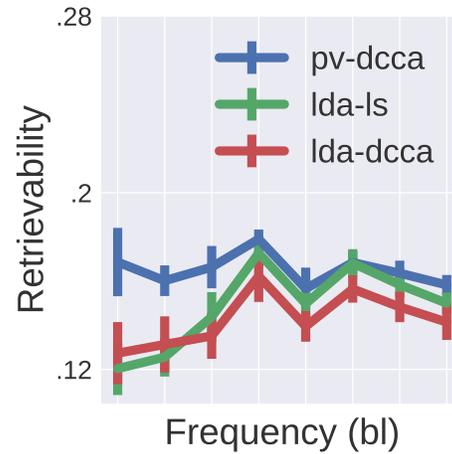


Figure 5.7: British Library

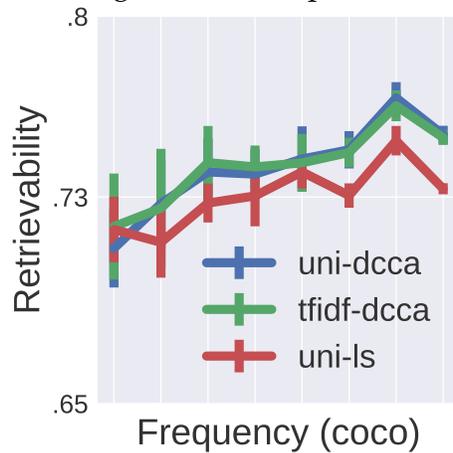


Figure 5.8: COCO

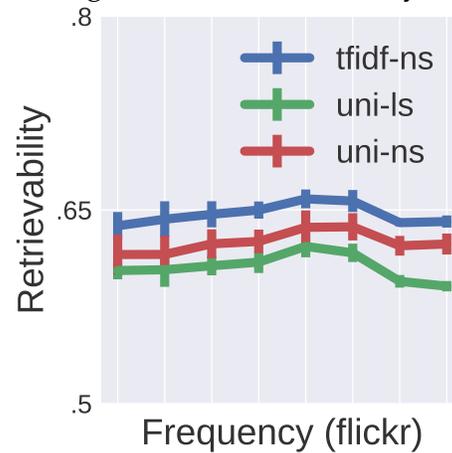


Figure 5.9: Flickr

Figure 5.10: Correlation between word/topic frequency and retrievability for each of the four datasets. Compared to our concreteness measure (see Figure 5.5; note that the while x-axes are different, the y-axes are the same) frequency explains relatively little variance in retrievability.

two appears to smoothly increase. In Wiki, on the other hand, there appears to be a concreteness threshold, beyond which retrieval becomes much easier.

There is little relationship between retrievability and frequency, further suggesting that our concreteness measure is not simply mirroring frequency. We re-made the plots in Figure 5.5, except we swapped the x-axis from concreteness to frequency; the resulting plots, given in Figure 5.10, are much flatter,

indicating that retrievability and frequency are mostly uncorrelated. Additional regression analyses reveal that for the top-3 performing algorithms on Flickr/Wiki/BL/COCO, concreteness explains 33%/64%/11%/15% of the variance in retrievability, respectively. In contrast, for all datasets, frequency explained less than 1% of the variance in retrievability.

5.8 Beyond Cross-Modal Retrieval

Concreteness scores do more than just predict retrieval performance; they also predict the difficulty of image classification. Two popular shared tasks from the ImageNet 2015 competition published class-level errors of all entered systems. We used the unigram concreteness scores from Flickr/COCO computed in §5.4 to derive concreteness scores for the ImageNet classes.¹⁵ We find that for both classification and localization, for all 10 top performing entries, and for both Flickr/COCO, there exists a moderate-to-strong Spearman correlation between concreteness and performance among the classes for which concreteness scores were available ($n_{\text{flickr}}, n_{\text{coco}} = 171, 288$; $.18 < \rho < .44$; $p < .003$ in all cases). This result suggests that concrete concepts may tend to be easier on tasks other than retrieval, as well.

¹⁵There are 1K classes in both ImageNet tasks, but we were only able to compute concreteness scores for a subset, due to vocabulary differences.

5.9 Future Directions

At present, it remains unclear if abstract concepts should be viewed as noise to be discarded (as in Kiela et al. (2014)), or more difficult, but learnable, signal. Because large datasets (e.g., social media) increasingly mix modalities using ambiguous, abstract language, researchers will need to tackle this question going forward. We hope that visual concreteness scores can guide investigations of the trickiest aspects of multimodal tasks. Our work suggests the following future directions:

Evaluating algorithms: Because concreteness scores are able to predict performance prior to training, evaluations could be reported over concrete and abstract instances separately, as opposed to aggregating into a single performance metric. A new algorithm that consistently performs well on non-concrete concepts, even at the expense of performance on concrete concepts, would represent a significant advance in multimodal learning.

Designing datasets: When constructing a new multimodal dataset, or augmenting an existing one, concreteness scores can offer insights regarding how resources should be allocated. Most directly, these scores enable focusing on “concrete visual concepts” (Huiskes et al., 2010; Chen et al., 2015b), by issuing image-search queries could be issued exclusively for concrete concepts during dataset construction. The opposite approach could also be employed, by prioritizing less concrete concepts.

Curriculum learning: During training, instances could be up/down-weighted in the training process in accordance with concreteness scores. It is not clear if placing more weight on the trickier cases (down-weighting concreteness), or

giving up on the harder instances (up-weighting concreteness) would lead to better performance, or differing algorithm behavior.

5.10 Brief Retrospective

While the original published version of this work pointed towards several potential applications of concreteness scores, we were excited to find that our scoring method has proven useful for a task we did not consider. Specifically: Shi et al. (2019) utilized concreteness scores as a baseline for unsupervised constituency parsing. By first computing the average visual concreteness of different spans of words and then applying a greedy clustering step, something quite close to a sentence parse tree can be constructed. Follow up work from Kojima et al. (2020) has confirmed the efficacy of this and related methods. Even for more sophisticated models with greater internal representational capacity, collapsing models to a single hidden dimension results in almost no performance degradation, and the single-dimension representation the model learns ends up closely correlating with the concreteness scores computed by our method.

CHAPTER 6

FUTURE WORK

The focus of this thesis was multimodal web data. Four primary projects were discussed. Two focused on leveraging this data to build better machine learning tools capable of drawing connections between different data modalities. The other two focused on building understanding of how people actually utilize images and text to communicate online. We argued that these two objectives, leveraging web data and understanding web data, are linked: improvements in one program drive improvements in the other.

6.0.1 Ethical Considerations in Machine Learning

In the era of “big data,” important conversations about ethical considerations of machine learning are happening in-step with algorithmic improvements. Machine learning datasets are generally gathered from the web, constructed via crowdsourcing, or some combination of the two (e.g., post-hoc captions generated by crowdworkers on images scraped from Flickr). While some arguments have been made regarding the potentially unethical implications of constructing a dataset purely via crowdsourcing (Fort et al., 2011; Williamson, 2016), our focus will be three major themes that apply to the the exploration of web data: consent, bias, and application. As a preface, we largely agree with the framing of Ess and Jones (2004): they argue 1) that “Internet research are ethical problems precisely because they evoke more than one ethically defensible response to a specific dilemma or problem. Ambiguity, uncertainty, and disagreement are inevitable.” but also 2) that “recognizing the possibility of a range of defensible ethical responses to a given dilemma does not commit us to ethical relativism

(‘anything goes’).”

Consent

In the case of web data, rarely are authors or subjects (in text, images, or videos) explicitly asked whether or not they are okay with their data being aggregated into a training set, and most are unaware that their data is even viewable by researchers (Fiesler and Proferes, 2018). While recent work highlights issues of consent in the ImageNet dataset (Prabhu and Birhane, 2020) (collected from Flickr) almost no machine learning dataset collected from web data meets, for example, the criteria of *informed consent*. Furthermore, legal standards of consent (e.g., a user agreeing to an EULA, or a user uploading an image under a copyright-permissive license) often do not accord with ethical ideals.

Is informed consent the “correct” framework for data collection in machine learning? Consider the case of the Europarl corpus (Koehn, 2005), which underlies many machine translation datasets/challenges, e.g., the WMT14 datasets (Macháček and Bojar, 2014). This data consists of transcripts of politicians speaking publicly, presumably with the understanding that their actions are being recorded. But even in this case, these people were not explicitly asked whether or not they would be okay with their data being used in a machine learning context, nor can they remove their utterances from now widely-distributed WMT corpora. The purpose of this example isn’t to single out Europarl, but to point out that, even if a corpus is not explicitly advertised as “web data” and is typically viewed as innocuous by the community, it can still violate a strict application of the informed consent standard.

boyd and Marwick (2011) explore additional axes of data consent, comparing being “in public” (as in: being observable) vs. “being public” (as in: actively seeking public attention) online. While a useful distinction, consideration of publicity expectations on the side of the author doesn’t fully solve the consent problem. For example, if a researcher constructs a dataset of hate speech in an online social network, should individuals posting that hate speech in the collection still be afforded the “right to be forgotten,” thereby hindering research on, e.g., extremism in online social networks?

Bias

One alluring aspect of training on a web corpus is that fewer preprocessing decisions must be explicitly made by a practitioner. On the surface, a “uniform sample” from the web may seem a promising path towards “objectivity,” especially when the alternative is bias arising from dataset curation or construction. But this reasoning is incomplete, especially when the end goal is to make a deployable tool. Given that statistical generalization is about recognizing patterns in training data (and reproducing those patterns come test-time), models certainly have the capacity to learn pernicious human biases. One simple but powerful example is due to Bolukbasi et al. (2016): they show that models based on word co-occurrences readily exhibit sexist correlations, e.g., “man” is to “doctor” as “woman” is to “nurse.” These word embedding models are frequently used as parameter initializations for training neural networks for downstream tasks.

Given that models often reflect the societal (or curation) biases present in

(pre)training data, the deployment¹ of machine learning-based tools should be undertaken with extreme care. Many cases of discrimination have been documented: Buolamwini and Gebru (2018) show racial discrepancies in commercial facial recognition software; Noble (2018) shows that several Google Search features, e.g., autocomplete, reflect societal biases; and Amazon was forced to abandon an automatic resume screening tool because it was sexist.²

Statistical models are optimized to reproduce trends present in the training data when making predictions. So, especially when deploying a tool trained on web data, achieving high accuracy is insufficient. We must also be vigilant to how and why algorithms make the predictions that they do, assess the cost/benefit of deploying/building tools, and be prepared to act when the impact of deployed tools is discriminatory, despite best efforts.

Application

Before auditing biased data or biased predictions, it's always worth reflecting on why are particular datasets and questions are being investigated in the first place. Tatman (2020) summarizes: *"Can I minimize differences in accuracy between subgroups" is less important than "should this be built at all."* Reasoning about the risks of unintended dual use (Jonas, 1979) is complex, personal, and ambiguous, but not unique to data science. However, this reflection is particularly important for machine learning researchers, not only because of the broad social impact

¹This is in contrast to the use case of exploring societal biases encoded in large-scale web data using statistical models. Abebe et al. (2020) refer to this case as "Computing as Diagnostic," and computational tools are deployed to "measure social problems and diagnose how they manifest in technical systems." Thus, "debiasing" methods may be inappropriate if the goal is diagnostic.

²<https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraped-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>

of technology, but also because non-technical practitioners³ (often unjustifiably) favor automatic predictions to human ones (Cummings, 2004). In a machine learning context, this “Automation Bias” problem may be exacerbated by “intelligent” behavior being ascribed to “AI” algorithms (Challen et al., 2019).

Steps to addressing ethical concerns

Minimally, as summarized by boyd and Crawford (2012): “Researchers must keep asking themselves — and their colleagues — about the ethics of their data collection, analysis, and publication.” Within machine learning, steps have been taken towards this end. Institutional review boards, including Cornell University’s,⁴ have recently posted new guidelines for working with social media data. Improved documentation (Bender and Friedman, 2018; Gebru et al., 2018; Mitchell et al., 2019) offers a step towards clarifying assumptions made during collection/modeling and specifying intended use cases. Legal regulation can offer practical requirements: while primarily focused on businesses rather than researchers, the General Data Protection Regulation (Council of European Union, 2014) (or GDPR, for short) aims to provide individuals with opportunities for affirmative consent, the right to be forgotten, etc. (though the translation between business requirements and research needs is imperfect). While *not* a panacea, in special cases, some mathematical notions of “fairness” can be optimized (Zafar et al., 2017). Major NLP conferences like ACL and EMNLP have recently updated their review form to explicitly ask reviewers to consider ethical concerns of submissions. Finally, guidelines for considering the ethics of internet research problems were proposed by Markham and Buchanan (2012):

³And undoubtedly some machine learning researchers.

⁴<https://researchservices.cornell.edu/policies/irb-policy-20-use-social-networking-sites-or-mobile-devices-human-participant-research>

they give an outline of questions to spark introspection, e.g., “What are the potential harms or risks associated with this study — for individuals, for online communities, for researchers, for research?” and “What are potential benefits associated with this study?”

Of the guidelines proposed by Markham and Buchanan (2012), one particularly resonates with us: “How are we recognizing the autonomy of others and acknowledging that they are of equal worth to ourselves and should be treated so?” It is our hope that with empathy, openness to critique, and willingness to listen, that tools built upon web data can continue to improve lives while minimizing harm.

6.0.2 Future Multimodal Work

Given the difficulty of many cross-modal reasoning tasks, there is a significant amount of future work to be undertaken towards these joint research agendas. I conclude by highlighting two high-level directions, one related to leveraging web data, and the other related to understanding web data.

1. As discussed in the introduction of this dissertation, a strong argument for utilizing large, unlabeled web corpora is pragmatic. Put simply: unsupervised pretraining “works” better than most other methods, at least in terms of accuracy (and related evaluations) for the most challenging datasets available today.

However, there are at least two possible shortcomings of web data. First, the *function vs. form* problem is highlighted by Bender and Koller (2020),

who argue that training on “surface forms” of language (e.g., those available in a web data dump) will never lead to true language understanding. Without grounding to some sort of communicative intent (or similar), statistical methods trained on web corpora may be doomed to shallow pattern recognition, rather than true “language understanding.” Even the difference between these two modes of comprehension is not yet fully defined, and, thus, it’s not clear whether or not we would even *know* if our systems were inching from recognition to understanding.

Second, the *reporting bias* (Van Durme, 2010) problem with web data is that not all useful information is stated in the “proper” frequency. A multimodal example is the “Black Sheep” problem.⁵ While most sheep in the world are white, the term “black sheep” appears with greater frequency in most web corpora vs. “white sheep.” Thus, a language-only model might be inclined to guess that most sheep are indeed black (entirely missing the point of the “black sheep” idiom). Even worse, some information may not be stated at all. Forbes et al. (2019) summarizes succinctly: “Any implication that can be trivially understood by a person is precisely the kind of information left unsaid.” Present exciting work (e.g., Sap et al. (2019)) is addressing this concern by building commonsense knowledge datasets containing usually-unstated background information. Even for tasks related to these types of corpora, however, pretraining on a large web corpus still results in significant performance improvement (Bosselut et al., 2019).

Thus, it largely remains an open question as to 1) whether or not web data

⁵The summary here is inspired by Daumé’s blogpost (<https://nlpers.blogspot.com/2016/06/language-bias-and-black-sheep.html>), which, in turn, was inspired by discussions with Meg Mitchell.

(coupled with the proper unsupervised training objective and sufficient scale) will be sufficient to build models capable of truly understanding text; and 2) if not, then what additional structured information should act as a supplement (or replacement).

2. Several prior works have proposed typologies of image-text communication (Zhang et al., 2018a; Alikhani and Stone, 2019; Vempala and Preotiu-Pietro, 2019; Alikhani et al., 2019), for example Marsh and Domas White (2003)'s categorization specifies "49 relationships and groups them in three categories according to the closeness of the conceptual relationship between image and text;" these range from "the image reiterates the text" to "the image transforms the text." These taxonomies are often inspired by studies of multimodal communication in various contexts, e.g., advertising, comic books, the web, etc. (Schwarcz, 1982; Hobbs, 1990; McCloud and Manning, 1998; Martinec and Salway, 2005; Cohn, 2013).

Statistical methods potentially offer a complementary means of corpus exploration (versus pre-specifying a particular taxonomy). Specifically, it would be ideal to develop unsupervised tools that cluster image-text messages automatically according to their communicative intent/strategy; this is in contrast to most multimodal clustering methods which focus on *content*, e.g., cat-related images are clustered with cat-related captions. In doing so, it may well be possible to automatically discover different context-specific typologies of image-text relations.

BIBLIOGRAPHY

- Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from [tensorflow.org](https://www.tensorflow.org).
- Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. 2020. Roles for computing in social change. In *ACM FAccT*.
- Harsh Agrawal, Arjun Chandrasekaran, Dhruv Batra, Devi Parikh, and Mohit Bansal. 2016. Sort story: Sorting jumbled images and captions into stories. In *EMNLP*.
- Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. 2016. Unsupervised learning from narrated instruction videos. In *CVPR*.
- Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. 2019. CITE: a corpus of image–text discourse relations. In *NAACL*.
- Malihe Alikhani and Matthew Stone. 2018. Exploring coherence in visual explanations. In *Multimedia Information Processing and Retrieval*.

- Malihe Alikhani and Matthew Stone. 2019. 'Caption' as a coherence relation: Evidence and implications. In *Second Workshop on Shortcomings in Vision and Language*.
- Khaled Almgren, Jeongkyu Lee, and Minkyu Kim. 2016. Predicting the future popularity of images on social networks. In *MISNC*.
- Tim Althoff, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2014. How to ask for a favor: A case study on the success of altruistic requests. In *ICWSM*.
- David Alvarez-Melis and Tommi S Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *EMNLP*.
- Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*.
- Galen Andrew, Raman Arora, Jeff A. Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *ICML*.
- Relja Arandjelovic and Andrew Zisserman. 2017. Look, listen and learn. In *ICCV*.
- Yoav Artzi, Patrick Pantel, and Michael Gamon. 2012. Predicting responses to microblog posts. In *NAACL*.
- Yoav Artzi and Luke Zettlemoyer. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *TACL*.

- Saeideh Bakhshi and Eric Gilbert. 2015. Red, purple and pink: The colors of diffusion on Pinterest. *PloS One*.
- Saeideh Bakhshi, David A. Shamma, and Eric Gilbert. 2014. Faces engage us: Photos with faces attract more likes and comments on instagram. In *CHI*.
- Eitan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. Everyone’s an influencer: Quantifying influence on twitter. In *WSDM*.
- Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multi-modal machine learning: A survey and taxonomy. *TPAMI*.
- Roja Bandari, Sitaram Asur, and Bernardo A. Huberman. 2012. The pulse of news in social media: Forecasting popularity. In *ICWSM*.
- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop on Evaluation Measures for MT and Summarization*.
- Roy Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second PASCAL recognising textual entailment challenge.
- Kobus Barnard, Pinar Duygulu, David Forsyth, Nando De Freitas, David M. Blei, and Michael I. Jordan. 2003. Matching words and pictures. *JMLR*.
- Roland Barthes. 1988. *Image-music-text*. Macmillan.
- John Bateman. 2014. *Text and image: A critical introduction to the visual/verbal divide*. Routledge.
- Emily Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *ACL*.

- Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *TACL*.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *ICML*.
- Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The fifth PASCAL recognizing textual entailment challenge. In *TAC*.
- Alexander C. Berg, Tamara L. Berg, Hal Daumé III, Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Aneesh Sood, Karl Stratos, et al. 2012. Understanding and predicting importance in images. In *CVPR*.
- Tamara L. Berg, Alexander C. Berg, and Jonathan Shih. 2010. Automatic attribute discovery and characterization from noisy web data. In *ECCV*.
- Sai Abishek Bhaskar, Maximilian Köper, Sabine Schulte im Walde, and Diego Frassinelli. 2017. Exploring multi-modal text+image models to distinguish between abstract and concrete nouns. In *IWCS Workshop on Foundations of Situated and Multimodal Communication*.
- Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, et al. 2010. VizWiz: nearly real-time answers to visual questions. In *ACM symposium on User Interface Software and Technology*.
- Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Alek-

- sandr Nisnevich, et al. 2020. Experience grounds language. *arXiv preprint arXiv:2004.10151*.
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *JMLR*.
- Piotr Bojanowski, Rémi Lajugie, Edouard Grave, Francis Bach, Ivan Laptev, Jean Ponce, and Cordelia Schmid. 2015. Weakly-supervised alignment of video with text. In *ICCV*.
- Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to home-maker? debiasing word embeddings. In *NeurIPS*.
- Youmna Borghol, Sebastien Ardon, Niklas Carlsson, Derek Eager, and Anirban Mahanti. 2012. The untold story of the clones: Content-agnostic factors that impact YouTube video popularity. In *KDD*.
- Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *ACL*.
- danah boyd and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679.
- danah boyd and Alice E. Marwick. 2011. Social privacy in networked publics: Teens’ attitudes, practices, and strategies. In *A decade in internet time: Symposium on the dynamics of the internet and society*.
- G. Bradski. 2000. OpenCV library. *Dr. Dobb’s Journal of Software Tools*.

- Erin Brady. 2015. Getting fast, free, and anonymous answers to questions asked by people with visual impairments. *ACM SIGACCESS Accessibility and Computing*.
- British Library Labs. 2016. Digitised books. <https://data.bl.uk/digbks/>.
- Cristian Buciluă, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *KDD*, pages 535–541. ACM.
- Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *ACM FAccT*.
- Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *ICML*.
- Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. 2019. Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*, 28(3):231–237.
- Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh. 2018. Attacking visual language grounding with adversarial examples: A case study on neural image captioning. In *ACL*.
- Jingyuan Chen. 2016. Multi-modal learning: Study on a large-scale micro-video data collection. In *ACM MM*.
- Minmin Chen, Alice X. Zheng, and Kilian Q. Weinberger. 2013a. Fast image tagging. In *ICML*.
- Tao Chen, Dongyuan Lu, Min-Yen Kan, and Peng Cui. 2013b. Understanding and classifying image tweets. In *ACM MM*.

- Tao Chen, Hany M SalahEldeen, Xiangnan He, Min-Yen Kan, and Dongyuan Lu. 2015a. Velda: Relating an image tweet’s text and images. In *AAAI*.
- Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015b. Microsoft COCO captions: Data collection and evaluation server. *Computing Research Repository*, arXiv:1504.00325. Version 2.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *EMNLP*.
- François Chollet. 2015. Keras. <https://github.com/fchollet/keras>.
- Shun-Po Chuang, Chia-Hung Wan, Pang-Chi Huang, Chi-Yu Yang, and Hung-Yi Lee. 2017. Seeing and hearing too: Audio representation for video captioning. In *IEEE Automatic Speech Recognition and Understanding Workshop*.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*.
- Neil Cohn. 2013. Visual narrative structure. *Cognitive science*.
- Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. 2014a. On the role of correlation and abstraction in cross-modal multimedia retrieval. *TPAMI*.
- Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert R.G. Lanckriet, Roger Levy, and Nuno Vasconcelos. 2014b. On the role of correlation and abstraction in cross-modal multimedia retrieval. *TPAMI*.

- Council of European Union. 2014. Council regulation (EU) no 269/2014.
<http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1416170084502&uri=CELEX:32014R0269>.
- Daniel Crevier. 1993. *AI: the tumultuous history of the search for artificial intelligence*. Basic Books, Inc.
- Mary Cummings. 2004. Automation bias in intelligent time critical decision support systems. In *AIAA 1st Intelligent Systems Technical Conference*, page 6313.
- Claudio Cusano, Gianluigi Ciocca, and Raimondo Schettini. 2004. Image annotation using SVM. In *Electronic Imaging*.
- Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising textual entailment*, pages 177–190. Springer.
- Bo Dai, Sanja Fidler, and Dahua Lin. 2018. A neural compositional paradigm for image captioning. In *NeurIPS*.
- Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *CVPR*.
- Cristian Danescu-Niculescu-Mizil, Justin Cheng, Jon Kleinberg, and Lillian Lee. 2012. You had me at hello: How phrasing affects memorability. In *ACL*.
- Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *WWW*.

- Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. 2019. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *ICCV*.
- Annette De Groot and Rineke Keijzer. 2000. What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning*, 50(1):1–56.
- Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. In *Sinn und Bedeutung 23*. Data can be found at <https://github.com/mcdm/CommitmentBank/>.
- Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *JMLR*.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*.
- Arturo Deza and Devi Parikh. 2015. Understanding image virality. In *CVPR*.
- Keyan Ding, Kede Ma, and Shiqi Wang. 2019. Intrinsic image popularity assessment. In *ACM MM*.
- Gideon Dror, Dan Pelleg, Oleg Rokhlenko, and Idan Szpektor. 2012. Churn prediction in new users of yahoo! answers. In *WWW*.
- Charles Ess and Steven Jones. 2004. Ethical decision-making and internet research: Recommendations from the aoir ethics working committee. In *Readings in virtual research ethics: Issues and controversies*, pages 27–44. IGI Global.

- Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving visual-semantic embeddings with hard negatives. In *BMVC*.
- Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In *CVPR*.
- Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *ECCV*.
- Casey Fiesler and Nicholas Proferes. 2018. “participant” perceptions of twitter research ethics. *Social Media+ Society*, 4(1).
- Flavio Figueiredo. 2013. On the prediction of popularity of trends and hits for user generated videos. In *WSDM*.
- Flavio Figueiredo, Jussara M. Almeida, Fabrício Benevenuto, and Krishna P. Gummadi. 2014. Does content determine information popularity in social media?: A case study of YouTube videos’ content and their popularity. In *CHI*, pages 979–982. ACM.
- Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? *Conference of the Cognitive Science Society*.
- Karèn Fort, Gilles Adda, and K Bretonnel Cohen. 2011. Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.
- Dario Garcia Garcia, Manohar Paluri, and Shaomei Wu. 2016. Under the hood: Building accessibility tools for the visually impaired on facebook.

<https://code.facebook.com/posts/457605107772545/under-the-hood-building-accessibility-tools-for-the-visually-impaired-on-facebook/>.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets.

Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo, and Shih-Fu Chang. 2015. Image popularity prediction in social media using sentiment and context features. In *ACM MM*.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9. Association for Computational Linguistics.

Eric Gilbert. 2013. Widespread underprovision on Reddit. In *CSCW*.

Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*.

Aloysia M. Gorman. 1961. Recognition memory for nouns as a function of abstractness and frequency. *Journal of Experimental Psychology*, 61(1):23–29.

David Grangier and Samy Bengio. 2008. A discriminative kernel-based approach to rank images from text queries. *TPAMI*.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*.

- Marco Guerini, Carlo Strapparava, and Gözde Özbal. 2011. Exploring text virality in social networks. In *ICWSM*.
- Sonal Gupta and Raymond J Mooney. 2010. Using closed captions as supervision for video activity recognition. In *AAAI*.
- Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P Bigham. 2019. VizWiz-Priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *CVPR*.
- Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. VizWiz grand challenge: Answering visual questions from blind people. In *CVPR*.
- Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. *arXiv preprint arXiv:2002.08565*.
- Meera Hahn, Nataniel Ruiz, Jean-Baptiste Alayrac, Ivan Laptev, and James M Rehg. 2018. Learning to localize and align fine-grained actions to sparse instructions. *arXiv preprint arXiv:1809.08381*.
- Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems*.
- Wangli Hao, Zhaoxiang Zhang, and He Guan. 2018. Integrating both visual and audio cues for enhanced video caption. In *AAAI*.
- Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016a. Deep residual learning for image recognition. In *CVPR*.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016b. Deep residual learning for image recognition. In *CVPR*.
- Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *ICCV*.
- Ralf Herbrich, Thore Graepel, and Klaus Obermayer. 1999. Large margin rank boundaries for ordinal regression. In *NeurIPS*.
- Jack Hessel and Lillian Lee. 2019. Something’s Brewing! Early Prediction of Controversy-causing Posts from Discussion Features. In *NAACL*.
- Jack Hessel, Lillian Lee, and David Mimno. 2017. Cats and captions vs. creators and the clock: Comparing multimodal content to context in predicting relative popularity. In *WWW*.
- Jack Hessel, Lillian Lee, and David Mimno. 2019a. Unsupervised discovery of multimodal links in multi-image, multi-sentence documents. In *EMNLP*.
- Jack Hessel, David Mimno, and Lillian Lee. 2018. Quantifying the visual concreteness of words and topics in multimodal datasets. In *NAACL*.
- Jack Hessel, Bo Pang, Zhenhai Zhu, and Radu Soricut. 2019b. A case study on combining asr and visual features for generating instructional video captions. In *CoNLL*.
- Jack Hessel, Chenhao Tan, and Lillian Lee. 2016. Science, AskScience, and Bad-Science: On the coexistence of highly related communities. In *ICWSM*.

- Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *NAACL*.
- Felix Hill, Douwe Kiela, and Anna Korhonen. 2013. Concreteness and corpora: A theoretical and practical analysis. In *Workshop on Cognitive Modeling and Computational Linguistics*.
- Felix Hill and Anna Korhonen. 2014a. Concreteness and subjectivity as dimensions of lexical meaning. In *ACL*.
- Felix Hill and Anna Korhonen. 2014b. Learning abstract concept embeddings from multi-modal data: Since you probably can't see what I mean. In *EMNLP*.
- Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Multi-modal models for concrete and abstract concept meaning. *TACL*.
- Jerry R Hobbs. 1990. *Literature and cognition*. Center for the Study of Language (CSLI).
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8).
- Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR*.
- Liangjie Hong, Ovidiu Dan, and Brian D. Davison. 2011. Predicting popular messages in Twitter. In *WWW companion volume*.
- Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. 2017. Attention-based multi-modal fusion for video description. In *ICCV*.

- Jiani Hu, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2016a. Multimodal learning for image popularity prediction on social media. In *Consumer Electronics-Taiwan*.
- Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. 2016b. Segmentation from natural language expressions. In *ECCV*.
- Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016c. Natural language object retrieval. In *CVPR*.
- De-An Huang, Shyamal Buch, Lucio Dery, Animesh Garg, Li Fei-Fei, and Juan Carlos Niebles. 2018. Finding “it”: Weakly-supervised reference-aware visual grounding in instructional videos. In *CVPR*.
- De-An Huang, Joseph J Lim, Li Fei-Fei, and Juan Carlos Niebles. 2017a. Un-supervised visual-linguistic reference resolution in instructional videos. In *CVPR*.
- Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017b. Densely connected convolutional networks. In *CVPR*.
- Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *NAACL*.
- Mark J. Huiskes, Bart Thomee, and Michael S. Lew. 2010. New trends and ideas in visual concept detection: the MIR flickr retrieval evaluation initiative. In *ACM MIR*.

- Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *JMLR*.
- Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *ACL*.
- Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan Boyd-Graber, Hal Daumé III, and Larry S. Davis. 2017. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In *CVPR*.
- Aaron Jaech, Victoria Zayats, Hao Fang, Mari Ostendorf, and Hannaneh Hajishirzi. 2015. Talking to the crowd: What do people react to in online discussions? In *EMNLP*.
- Mainak Jas and Devi Parikh. 2015. Image specificity. In *CVPR*.
- Jiwoon Jeon, Victor Lavrenko, and Raghavan Manmatha. 2003. Automatic image annotation and retrieval using cross-media relevance models. In *SIGIR*.
- Yangqing Jia, Mathieu Salzmann, and Trevor Darrell. 2011. Learning cross-modality similarity for multinomial data. In *ICCV*.
- Xinyang Jiang, Fei Wu, Xi Li, Zhou Zhao, Weiming Lu, Siliang Tang, and Yuet-ing Zhuang. 2015. Deep compositional cross-modal learning to rank via local-global alignment. In *ACM MM*.
- Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *KDD*.

- Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*.
- Hans Jonas. 1979. *Das Prinzip Verantwortung. Versuch einer Ethik für die technologische Zivilisation*. Insel Verlag.
- James M Jones, Gary Alan Fine, and Robert G Brust. 1979. Interaction effects of picture and caption on humor ratings of cartoons. *The Journal of Social Psychology*, 108(2):193–198.
- Roy Jonker and Anton Volgenant. 1987. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*.
- Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. 2016. Learning visual features from large weakly supervised data. In *ECCV*.
- Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.
- Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *NeurIPS*.
- Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *EMNLP*.
- Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? Textbook question answering for multimodal machine comprehension. In *CVPR*.

- Inayatullah Khan, Amir Saffari, and Horst Bischof. 2009. TVGraz: Multi-modal learning of object categories by combining textual and visual features. In *Workshop of the Austrian Association for Pattern Recognition*.
- Rahat Khan, Joost Van de Weijer, Fahad Shahbaz Khan, Damien Muselet, Christophe Ducottet, and Cecile Barat. 2013. Discriminative color descriptors. In *CVPR*.
- Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *NAACL*.
- Aditya Khosla, Atish Das Sarma, and Raffay Hamid. 2014. What makes an image popular? In *WWW*.
- Aditya Khosla, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2012. Memorability of image regions. In *NeurIPS*.
- Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *EMNLP*.
- Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. 2014. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *ACL*.
- Gunhee Kim, Seungwhan Moon, and Leonid Sigal. 2015. Ranking and retrieval of image sequences from multiple paragraph queries. In *CVPR*.
- Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J Guo, Robert C Miller, and Krzysztof Z Gajos. 2014. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *CHI*.

- Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.
- Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In *ICML*.
- Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2015. Unifying visual-semantic embeddings with multimodal neural language models. *TACL*.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.
- Noriyuki Kojima, Hadar Averbuch-Elor, Alexander M Rush, and Yoav Artzi. 2020. What is learned in visually grounded neural syntax acquisition. In *ACL*.
- Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. 2017. Open-Images: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from <https://github.com/openimages>*.
- Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017a. Dense-captioning events in videos. In *ICCV*.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David Ayman Shamma, Michael Bernstein, and Li Fei-Fei. 2017b. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*.

- Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in instagram posts. In *EMNLP*.
- Hilde Kuehne, Alexander Richard, and Juergen Gall. 2017. Weakly supervised learning of actions from transcripts. *Computer Vision and Image Understanding*.
- Harold W. Kuhn. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*.
- Girish Kulkarni, Visruth Premraj, Vicente Ordóñez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *TPAMI*.
- Himabindu Lakkaraju, Julian J. McAuley, and Jure Leskovec. 2013. What’s in a name? understanding the interplay between titles, content, and communities in social media. In *ICWSM*.
- Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *CHI*.
- Jonathan Lazar, Aaron Allen, Jason Kleinman, and Chris Malarkey. 2007. What frustrates screen reader users on the web: A study of 100 blind users. *International Journal of HCI*.
- Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *NAACL*.
- Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*.

- Moontae Lee, Seok Hyun Jin, and David Mimno. 2016. Beyond exchangeability: The Chinese voting process. In *NeurIPS*.
- Jay Lemke. 1998. Multiplying meaning. *Reading science: Critical and functional perspectives on discourses of science*.
- Kristina Lerman and Tad Hogg. 2010. Using a model of social dynamics to predict popularity of news. In *WWW*.
- Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2011. The Winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, volume 46, page 47.
- Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. In *ACL*.
- Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. 2018. Jointly localizing and describing events for dense video captioning. In *CVPR*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.
- Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *ACL*.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *ECCV*.
- Yu Liu, Jianlong Fu, Tao Mei, and Chang Wen Chen. 2017. Let your photos talk:

- Generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks. In *AAAI*.
- Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*.
- Yijuan Lu, Lei Zhang, Qi Tian, and Wei-Ying Ma. 2008. What are the high-level concepts with small semantic gaps? In *CVPR*.
- Corey Lynch, Kamelia Aryafar, and Josh Attenberg. 2016. Images don't lie: Transferring deep visual semantic features to large-scale multimodal learning to rank. In *KDD*.
- Zongyang Ma, Aixin Sun, and Gao Cong. 2012. Will this #hashtag be popular tomorrow? In *SIGIR*.
- Matouš Macháček and Ondřej Bojar. 2014. Results of the WMT14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301.
- Haley MacLeod, Cynthia L Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding blind people's experiences with computer-generated captions of social media images. In *CHI*.
- Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. 2018. Exploring the limits of weakly supervised pretraining. In *ECCV*.
- Jonathan Malmaud, Jonathan Huang, Vivek Rathod, Nick Johnston, Andrew Rabinovich, and Kevin Murphy. 2015. What's cookin'? interpreting cooking videos using text, speech and vision. In *NAACL*.

- Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *CVPR*.
- Edgar Margffoy-Tuay, Juan C Pérez, Emilio Botero, and Pablo Arbeláez. 2018. Dynamic multimodal instance segmentation guided by natural language queries. In *ECCV*.
- Lauren E Margulieux, Mark Guzdial, and Richard Catrambone. 2012. Subgoal-labeled instructional material improves performance and transfer in learning to develop mobile applications. In *Conference on International Computing Education Research*.
- Annette Markham and Elizabeth Buchanan. 2012. Ethical decision-making and internet research: Version 2.0. recommendations from the aoir ethics working committee. Available online: aoir.org/reports/ethics2.pdf.
- Emily E Marsh and Marilyn Domas White. 2003. A taxonomy of relationships between images and text. *Journal of Documentation*, 59(6).
- Radan Martinec and Andrew Salway. 2005. A system for image–text relations in new (and old) media. *Visual communication*, 4(3).
- Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012. A Joint Model of Language and Perception for Grounded Attribute Learning. In *ICML*.
- Masoud Mazloom, Robert Rietveld, Stevan Rudinac, Marcel Worring, and Willemijn van Dolen. 2016. Multimodal popularity prediction of brand-related social media posts. In *ACM MM*.

- Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. <http://mallet.cs.umass.edu>.
- John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon. 1955. A proposal for the dartmouth summer research project on artificial intelligence. *Reprinted online at <http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html>*.
- Scott McCloud and AD Manning. 1998. Understanding comics: The invisible art. *IEEE Transactions on Professional Communications*.
- Philip J McParlane, Yashar Moshfeghi, and Joemon M Jose. 2014. Nobody comes here anymore, it's too crowded; Predicting image popularity on Flickr. In *Multimedia Retrieval*.
- Aditya Krishna Menon and Charles Elkan. 2011. Link prediction via matrix factorization. In *ECML PKDD*.
- Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NeurIPS*.
- George A Miller and Philip N Johnson-Laird. 1976. *Language and perception*. Belknap Press.
- Margaret Mitchell, Kees van Deemter, and Ehud Reiter. 2010. Natural reference to objects in a visual domain. In *International Natural Language Generation Conference*.

- Margaret Mitchell, Ehud Reiter, and Kees Van Deemter. 2013. Typicality and object reference. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *ACM FAccT*.
- Tanvi S Motwani and Raymond J Mooney. 2012. Improving video activity recognition using object recognition and text mining. In *ECAI*.
- Claude Nadeau and Yoshua Bengio. 2000. Inference for the generalization error. In *NeurIPS*.
- Varun K. Nagaraja, Vlad I. Morariu, and Larry S. Davis. 2016. Modeling context between objects for referring expression understanding. In *ECCV*.
- Iftekhar Naim, Young C Song, Qiguang Liu, Liang Huang, Henry Kautz, Jiebo Luo, and Daniel Gildea. 2015. Discriminative unsupervised alignment of natural language instructions with corresponding video segments. In *NAACL*.
- Iftekhar Naim, Young Chol Song, Qiguang Liu, Henry A Kautz, Jiebo Luo, and Daniel Gildea. 2014. Unsupervised alignment of natural language instructions with video segments. In *AAAI*.
- Douglas L. Nelson, Cathy L. McEvoy, and Thomas A. Schreiber. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.
- Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

- Kay O'Halloran. 2004. *Multimodal discourse analysis: Systemic functional perspectives*. A&C Black.
- Aude Oliva and Antonio Torralba. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*.
- Celie O'Neil-Hart. 2018. Why you should lean into how-to content in 2018. www.thinkwithgoogle.com/advertising-channels/video/self-directed-learning-youtube/. Accessed: 2019-09-03.
- Vicente Ordóñez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *NeurIPS*.
- Michael O'toole. 1994. *The language of displayed art*. Fairleigh Dickinson Univ Press.
- Allan Paivio. 1991. Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology*, 45(3):255–287.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.
- Devi Parikh and Kristen Grauman. 2011. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*.
- Cesc C. Park and Gunhee Kim. 2015. Expressing an image stream with a sequence of natural sentences. In *NeurIPS*.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP*.
- Sasa Petrovic, Miles Osborne, and Victor Lavrenko. 2011. RT to win! Predicting message propagation in Twitter. In *ICWSM*.

- Julia Peyre, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2017. Weakly-supervised learning of visual relations. In *ICCV*.
- Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: The word-in-context dataset for evaluating context-sensitive meaning representations. In *NAACL*.
- Henrique Pinto, Jussara M Almeida, and Marcos A Gonçalves. 2013. Using early view patterns to predict the popularity of YouTube videos. In *WSDM*.
- Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*.
- Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *EMNLP*.
- Adrian Popescu, Theodora Tsirikla, and Jana Kludas. 2010. Overview of the Wikipedia retrieval task at ImageCLEF 2010. In *CLEF*.
- Vinay Prabhu and Abeba Birhane. 2020. Large image datasets: A pyrrhic win for computer vision? Preprint Under Review.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

- Lee Rainie, Joanna Brenner, and Kristen Purcell. 2012. Photos and videos as social currency online. *Pew Internet & American Life Project*.
- Vasili Ramanishka, Abir Das, Dong Huk Park, Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, and Kate Saenko. 2016. Multimodal video description. In *ACM MM*.
- Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *ACM MM*.
- Michaela Regneri, Marcus Rohrbach, Dominikus Wetzels, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *TACL*.
- Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *LREC Workshop on NLP Frameworks*.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*.
- Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.
- Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. In *ECCV*.
- Daniel M. Romero, Chenhao Tan, and Johan Ugander. 2013. On the interplay between social and topical structure. In *ICWSM*.

- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *NAACL*.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252.
- Bryan C. Russell, Ricardo Martin-Brualla, Daniel J. Butler, Steven M. Seitz, and Luke Zettlemoyer. 2013. 3D Wikipedia: Using online text to automatically label and navigate reconstructed geometry. *ACM Transactions on Graphics*.
- Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*.
- Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS.
- Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI*.
- Rossano Schifanella, Miriam Redi, and Luca Aiello. 2015. An image is worth more than a thousand favorites: Surfacing the hidden beauty of Flickr pictures. In *ICWSM*.
- Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*.

- Paula J. Schwanenflugel and Edward J. Shoben. 1983. Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(1):82–102.
- Joseph H Schwarcz. 1982. *Ways of the illustrator: Visual communication in children's literature*. American Library Association Chicago.
- John R Searle. 1980. Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424.
- Ozan Sener, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. 2015. Unsupervised semantic parsing of video collections. In *ICCV*.
- David A. Shamma, Jude Yew, Lyndon Kennedy, and Elizabeth F. Churchill. 2011. Viral actions: Predicting video view counts using synchronous sharing behaviors. In *ICWSM*.
- Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: An astounding baseline for recognition. In *CVPR workshops*.
- Zhiqiang Shen, Jianguo Li, Zhou Su, Minjun Li, Yurong Chen, Yu-Gang Jiang, and Xiangyang Xue. 2017. Weakly supervised dense video captioning. In *CVPR*.
- Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. 2019. Visually grounded neural syntax acquisition. In *ACL*.
- Andrew Shin, Katsunori Ohnishi, and Tatsuya Harada. 2016. Beyond caption to narrative: Video captioning with multiple sentences. In *International Conference on Image Processing*.

- Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2016. Visually grounded meaning representations. *TPAMI*.
- Carina Silberer and Mirella Lapata. 2012. Grounded models of semantic representation. In *EMNLP*.
- Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. In *ICLR*.
- Philipp Singer, Fabian Flöck, Clemens Meinhart, Elias Zeitfogel, and Markus Strohmaier. 2014. Evolution of Reddit: From the front page of the internet to a self-referential community? In *WWW*.
- Richard Socher and Li Fei-Fei. 2010. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *CVPR*.
- Henry Solomon and Lorraine Herman. 1977. Status symbols and prosocial behavior: The effect of the victim's car on helping. *The Journal of Psychology*, 97(2).
- Greg Stoddard. 2015. Popularity dynamics and intrinsic quality in reddit and hacker news. In *ICWSM*.
- Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. 2010. Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In *SocialCom*.
- Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. 2019a. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743*.

- Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019b. Videobert: A joint model for video and language representation learning. In *ICCV*.
- Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. 2019c. Scalability in perception for autonomous driving: Waymo open dataset.
- Tao Sun, Ming Zhang, and Qiaozhu Mei. 2013. Unexpected relevance: An empirical study of serendipity in retweets. In *ICWSM*.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NeurIPS*.
- Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *CVPR*.
- Chenhao Tan and Lillian Lee. 2015. All who wander: On the prevalence and characteristics of multi-community engagement. In *WWW*.
- Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic-and author-controlled natural experiments on Twitter. In *ACL*.
- Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP*.
- Rachel Tatman. 2020. What I won't build. WiNLP @ ACL 2020 Keynote Talk.

- The Associated Press. 2020. AP information: <https://www.ap.org/en-us/>, accessed may 14, 2020.
- Theano Development Team. 2016. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688.
- Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: the new data in multimedia research. *Communications of the ACM*.
- Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. 2018. Audio-visual event localization in unconstrained videos. In *ECCV*.
- Oren Tsur and Ari Rappoport. 2012. What’s in a hashtag?: Content based prediction of the spread of ideas in microblogging communities. In *WSDM*.
- Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *EMNLP*.
- Benjamin Van Durme. 2010. *Extracting implicit knowledge from text*. Ph.D. thesis, University of Rochester.
- Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for neural machine translation. *CoRR*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.
- Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. 2017. Learning from noisy large-scale datasets with minimal supervision. In *CVPR*.
- Alakananda Vempala and Daniel Preoțiuc-Pietro. 2019. Categorizing and inferring the relationship between the text and image of twitter posts. In *ACL*.
- Anton Volgenant. 2004. Solving the k-cardinality assignment problem by transformation. *European Journal of Operational Research*.
- Ian Walker and Charles Hulme. 1999. Concrete words are easier to recall than abstract words: Evidence for a semantic contribution to short-term serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(5):1256–1271.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *NeurIPS*.
- Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. 2018. Bidirectional attentive fusion with context gating for dense video captioning. In *CVPR*.
- Weiran Wang, Raman Arora, Karen Livescu, and Jeff A. Bilmes. 2015a. On deep multi-view representation learning. In *ICML*.
- Weiran Wang, Raman Arora, Karen Livescu, and Nathan Srebro. 2015b. Stochastic optimization for deep cca via nonlinear orthogonal iterations. In *Communication, Control, and Computing*.

- Weiyang Wang, Yongcheng Wang, Shizhe Chen, and Qin Jin. 2019b. Youmakeup: A large-scale domain-specific multimodal dataset for fine-grained semantic comprehension. In *EMNLP*.
- Yunchao Wei, Yao Zhao, Canyi Lu, Shikui Wei, Luoqi Liu, Zhenfeng Zhu, and Shuicheng Yan. 2016. Cross-modal retrieval with cnn visual features: A new baseline. *Transactions on Cybernetics*.
- Sarah Weir, Juho Kim, Krzysztof Z Gajos, and Robert C Miller. 2015. Learner-sourcing subgoal labels for how-to videos. In *CSCW*.
- Tim Weninger, Thomas James Johnston, and Maria Glenski. 2015. Random voting effects in social-digital spaces: A case study of Reddit post submissions. In *Hypertext*.
- Vanessa Williamson. 2016. On the ethics of crowdsourced research. *PS: Political Science & Politics*, 49(1):77–81.
- Bo Wu, Wen-Huang Cheng, Yongdong Zhang, and Tao Mei. 2016. Time matters: Multi-scale temporalization of social media popularity. In *ACM MM*.
- Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017a. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *CSCW*.
- Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017b. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *CSCW*.
- Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *CVPR*.

- Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*.
- Jun Xu, Ting Yao, Yongdong Zhang, and Tao Mei. 2017. Learning multimodal attention lstm networks for video captioning. In *ACM MM*.
- Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. In *EMNLP*.
- Kota Yamaguchi, Tamara L Berg, and Luis E Ortiz. 2014. Chic or social: Visual popularity analysis in online fashion networks. In *ACM MM*.
- Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *WWW*.
- Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL*.
- Mark Yatskar, Svitlana Volkova, Asli Celikyilmaz, Bill Dolan, and Luke Zettlemoyer. 2013. Learning to relate literal and sentimental descriptions of visual properties. In *NAACL*.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*.
- Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. In *ECCV*.

- Louis Yu, Sitaram Asur, and Bernardo A Huberman. 2011. What trends in chinese social media. In *5th SNA-KDD Workshop*.
- Jiahong Yuan, Mark Liberman, and Christopher Cieri. 2006. Towards an integrated understanding of speaking rate in conversation. In *International Conference on Spoken Language Processing*.
- Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*.
- Stephen Zakrewsky, Kamelia Aryafar, and Ali Shokoufandeh. 2016. Item popularity prediction in e-commerce using image quality feature vectors. *arXiv preprint arXiv:1605.03663*.
- Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. 2018a. Equal but not the same: Understanding the implicit relationship between persuasive images and text. In *BMVC*.
- Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018b. ReCoRD: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint 1810.12885*.
- Changtao Zhong, Dmytro Karamshuk, and Nishanth Sastry. 2015. Predicting Pinterest: Automating a distributed human computation. In *WWW*.
- Luowei Zhou, Nathan Louis, and Jason J Corso. 2018a. Weakly-supervised video object grounding from text by loss weighting and object interaction. In *BMVC*.
- Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018b. Towards automatic learning of procedures from web instructional videos. In *AAAI*.

- Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. 2018c. End-to-end dense video captioning with masked transformer. In *CVPR*.
- Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*.
- Yue Ting Zhuang, Yan Fei Wang, Fei Wu, Yin Zhang, and Wei Ming Lu. 2013. Supervised coupled dictionary learning with group structures for multi-modal retrieval. In *AAAI*.
- Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. 2019. Cross-task weakly supervised learning from instructional videos. In *CVPR*.
- Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. 2018. Learning transferable architectures for scalable image recognition. In *CVPR*.