Multimodal Grounding from User-generated Web Content

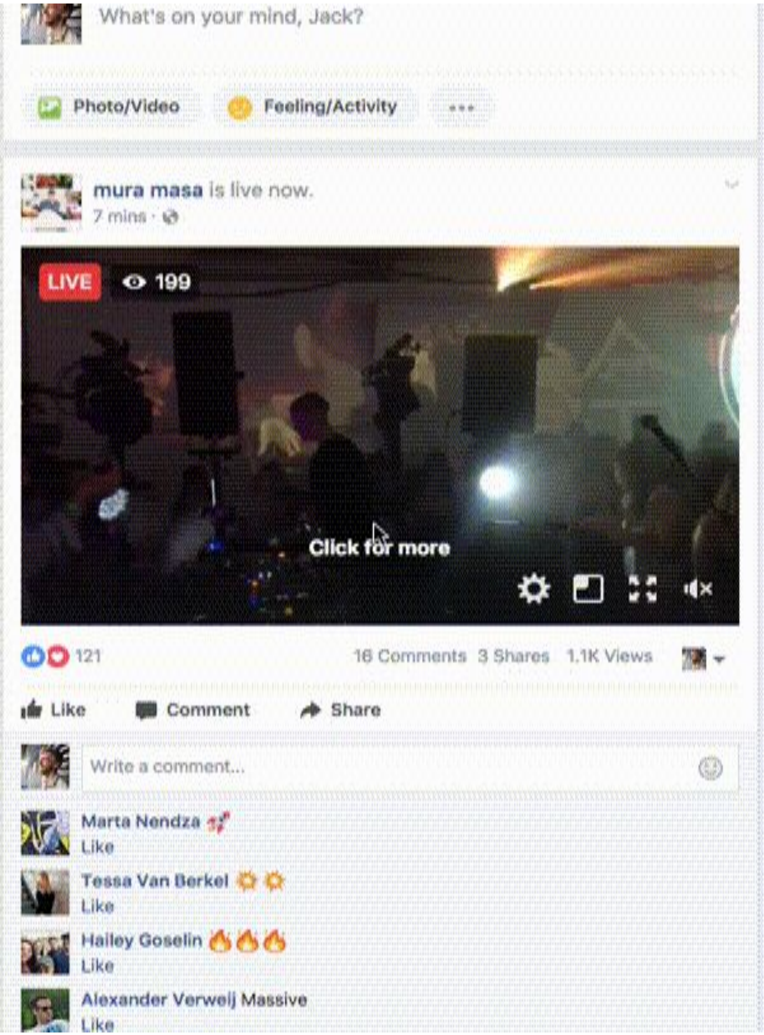
Jack Hessel
PhD Candidate
Cornell University

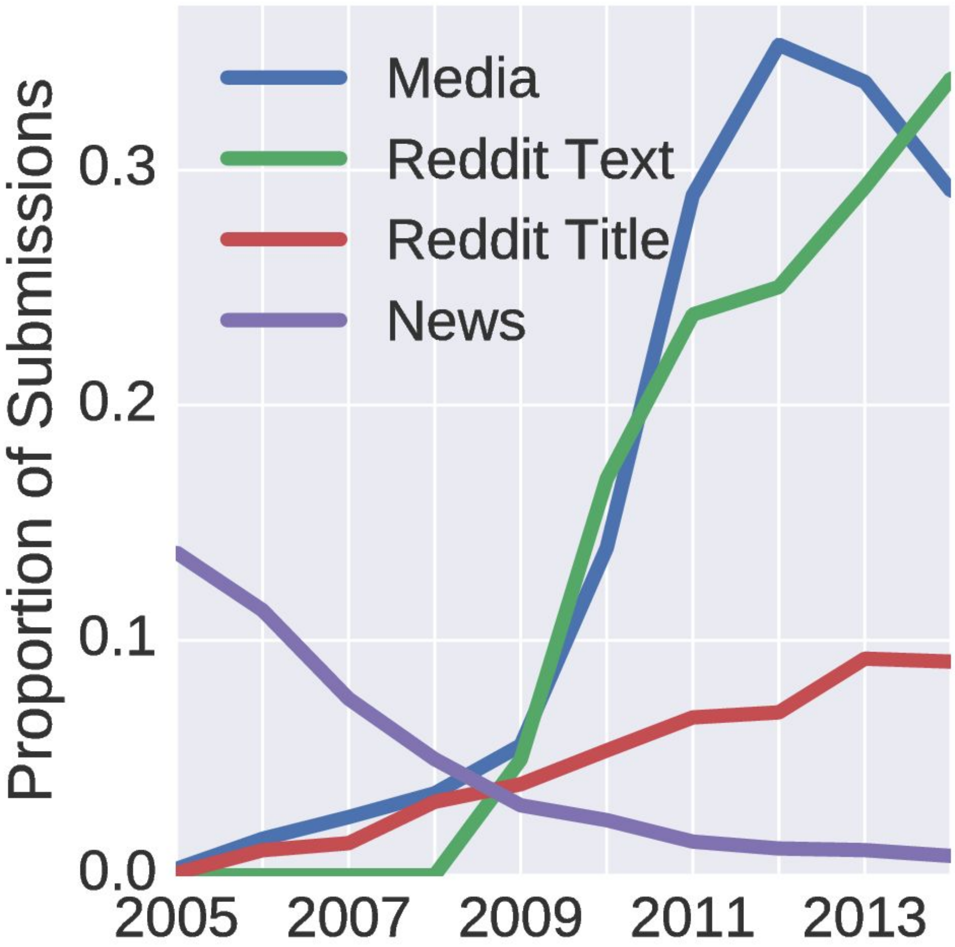
Why study multimodal web data?

Why study multimodal web data?

Increasingly, our social interactions manifest in *online*,

and online communities are increasingly multimodal


What's on your mind, Jack?
Photo/Video
Feeling/Activity
mura masa is live now.
7 mins
LIVE
199
Click for more
121
16 Comments 3 Shares 1.1K Views
Like
Comment
Share
Write a comment...
Marta Nendza
Like
Tessa Van Berkel
Like
Hailey Goselin
Like
Alexander Verwelj Massive
Like



imgur.com,
youtube...

cnn.com,
bbc.co.uk...

Perhaps the meteoric rise of multimodal content isn't so surprising...

Perhaps the meteoric rise of multimodal content isn't so surprising...

Semioticians have long argued multimodality is a fundamental part of communication

[Lemke 2002]

"The power of visual communication is multiplied when it is co-deployed with language in multimodal texts."

Perhaps the meteoric rise of multimodal content isn't so surprising...

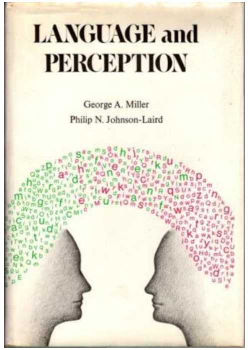
Semioticians have long argued multimodality is a fundamental part of communication

[Lemke 2002]

"The power of visual communication is multiplied when it is co-deployed with language in multimodal texts."

Cognitive psychologists have studied the connection between perceptions and language since at least the 1970s.

[Miller and Johnson-Laird 1976]



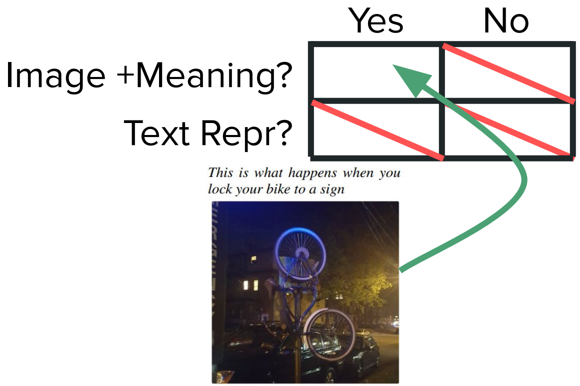
Why study multimodal web data?

Increasingly, our social interactions manifest in *online*,
and online communities are increasingly multimodal

Why study multimodal web data?

Increasingly, our social interactions manifest in *online*,
and online communities are increasingly multimodal

Study of web data gives an
in-vivo perspective
on communication and
communities!



[Vempala and Preoțiuc-Pietro 2019;
c.f. Chen et al., 2015,
Kruk and Lubin et al., 2019,
Alikhani et al., 2019]

Why study multimodal web data?

Why study multimodal web data?

if you don't care about communication in online communities

Why study multimodal web data?

if you don't care about communication in online communities

Unreasonable effectiveness
of (multimodal) web data

✓ IMAGENET

✓ GLUE

✓ Crosstask

[Deng et al. 2009;
Wang et al. 2019;
Zhukov et al. 2019]

Why study multimodal web data?

if you don't care about communication in online communities

Unreasonable effectiveness
of (multimodal) web data

IMAGENET

GLUE

Crosstask

[Deng et al. 2009;
Wang et al. 2019;
Zhukov et al. 2019]

Building tools that *require* grounding

10.09.19

Chrome's new AI feature solves one of the web's eternal problems

To help blind and low-vision users, Google is using machine learning to generate descriptions for millions of images.

[c.f. Wu et al. 2017;
Sharma et al. 2019]

Today

Today

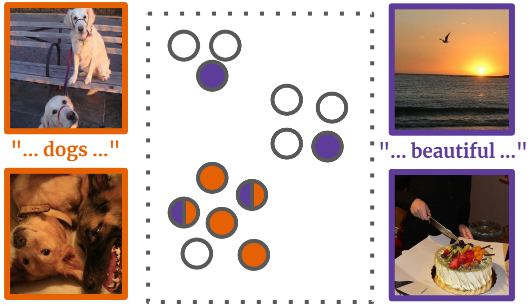
Does *multimodality* affect community reception of content?

[WWW 2017, H., Lee, Mimno]

Today

Does *multimodality* affect community reception of content?

[WWW 2017, H., Lee, Mimno]



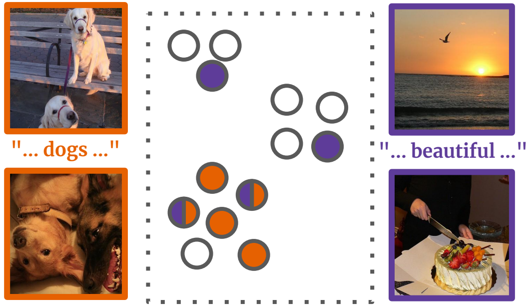
What concepts are "groundable," and in what context?

[NAACL 2018, H., Mimno, Lee]

Today

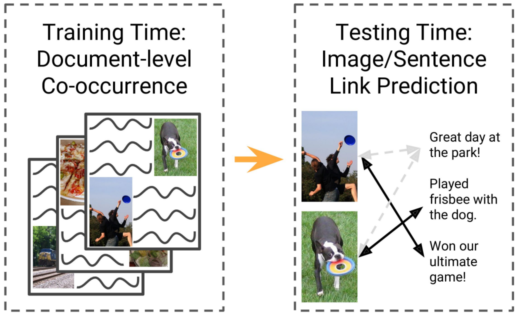
Does *multimodality* affect community reception of content?

[WWW 2017, H., Lee, Mimno]



What concepts are "groundable," and in what context?

[NAACL 2018, H., Mimno, Lee]



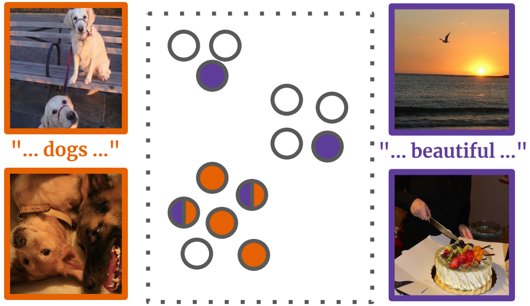
Can grounding be learned directly from multi-sentence, multi-image web documents?

[EMNLP 2019, H., Lee, Mimno]

Today

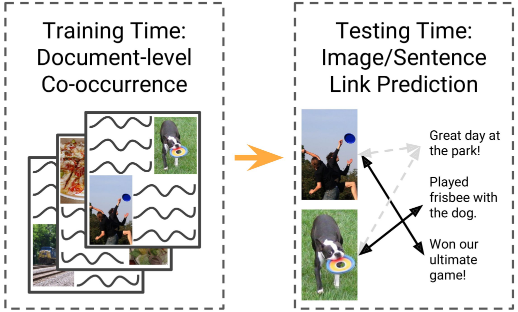
Does *multimodality* affect community reception of content?

[WWW 2017, H., Lee, Mimno]



What concepts are "groundable," and in what context?

[NAACL 2018, H., Mimno, Lee]



Can grounding be learned directly from multi-sentence, multi-image web documents?

[EMNLP 2019, H., Lee, Mimno]

Goal:
Recover Community Preferences
by predicting popularity of image/text posts

"Tonight, I carved a pumpkin. I also doused it in lighter fluid and lit it on fire." - /r/pics

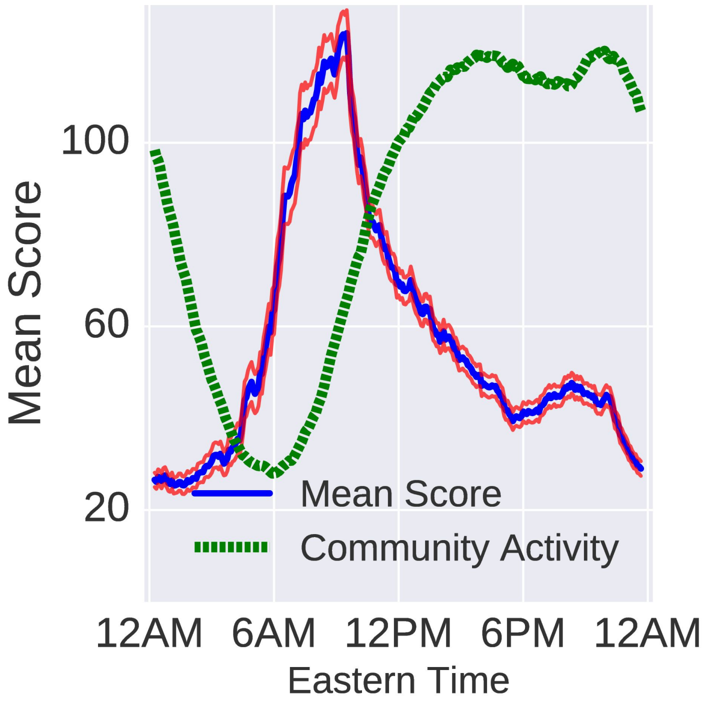
"You have to go to the border for food Fish Tacos [San Diego]" - /r/FoodPorn

"Snacks!" - /r/aww

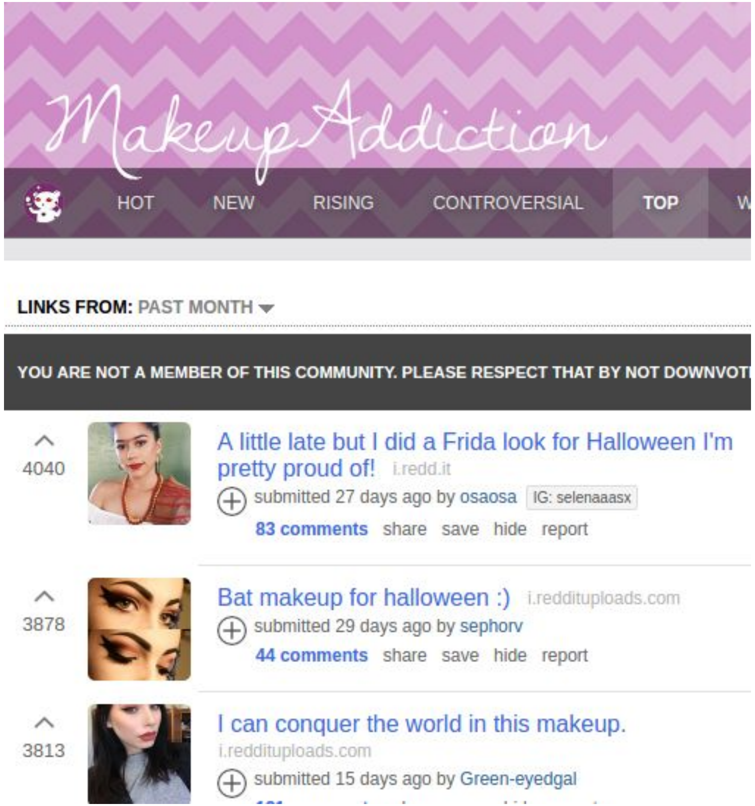
"Glamor Leaves" - /r/RedditLaquersitas

	# 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

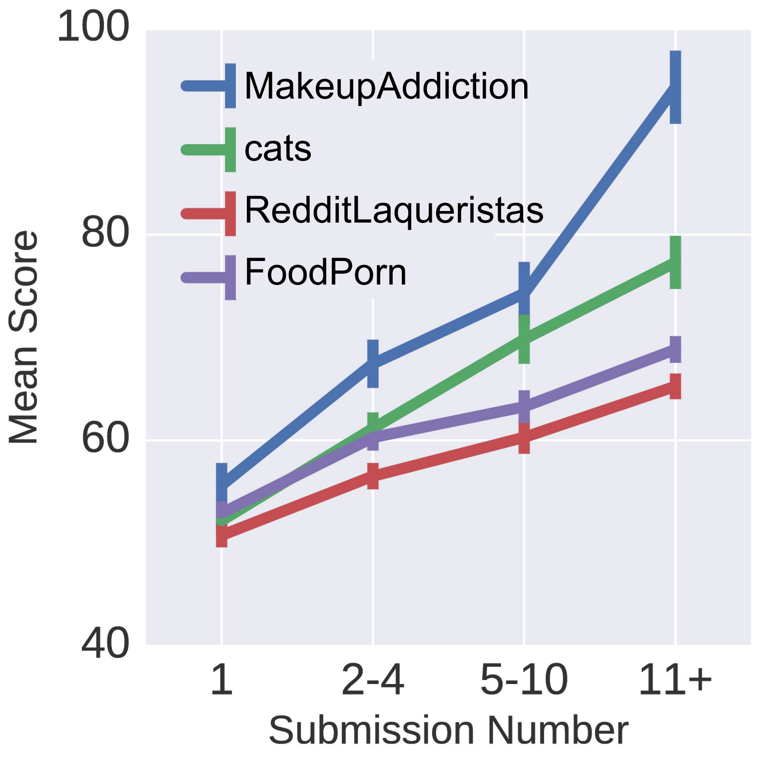
Complication: Minutes matter



Complication: Identity Matters



Complication: Identity Matters



Complication: Rich get richer

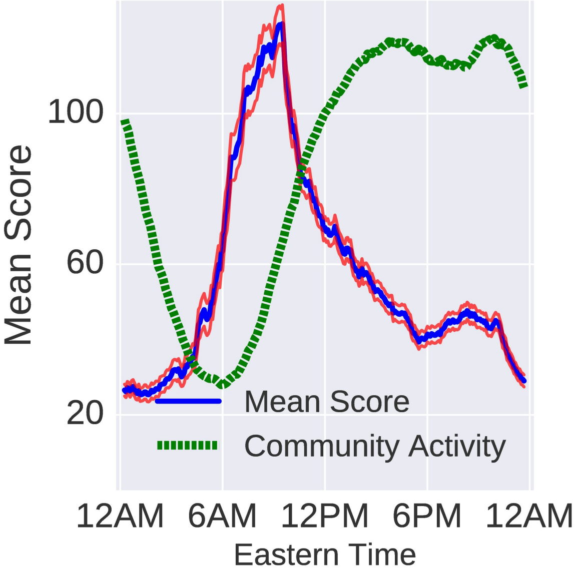
[Salganik et al. 2006]

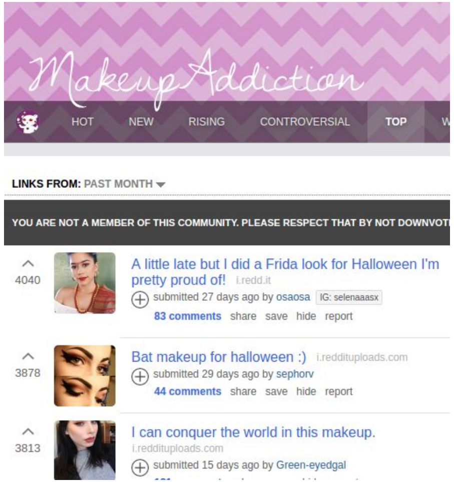
Complication: Rich get richer

"... small, random rating manipulations on social media submissions created significant changes in downstream ratings...
Positive treatment resulted in [an]
increased final rating [of] 11.02% on average."
-- Glenski et al. 2015

Reddit **overlooked 52% of the most popular links** the first time they were submitted.
-- Gilbert 2013


MUSIC LAB
COLUMBIA UNIVERSITY
CLICK HERE TO START
FREE MUSIC DOWNLOADS


Mean Score
100
60
20
Mean Score
Community Activity
12AM
6AM
12PM
6PM
12AM
Eastern Time


MakeupAddiction
HOT
NEW
RISING
CONTROVERSIAL
TOP
LINKS FROM: PAST MONTH
YOU ARE NOT A MEMBER OF THIS COMMUNITY. PLEASE RESPECT THAT BY NOT DOWNVOT
4040
A little late but I did a Frida look for Halloween I'm pretty proud of! i.redd.it
submitted 27 days ago by osaosa IG: selenaaasx
83 comments share save hide report
3878
Bat makeup for halloween :) i.reddituploads.com
submitted 29 days ago by sephorv
44 comments share save hide report
3813
I can conquer the world in this makeup.
i.reddituploads.com
submitted 15 days ago by Green-eyedgal

Can we isolate the effects of *content* rather than context?

Idea: impose strict timing controls

Idea: impose strict timing controls

Idea: impose strict timing controls

The grass is always greener

This is why you get two cats

13 Seconds Apart!

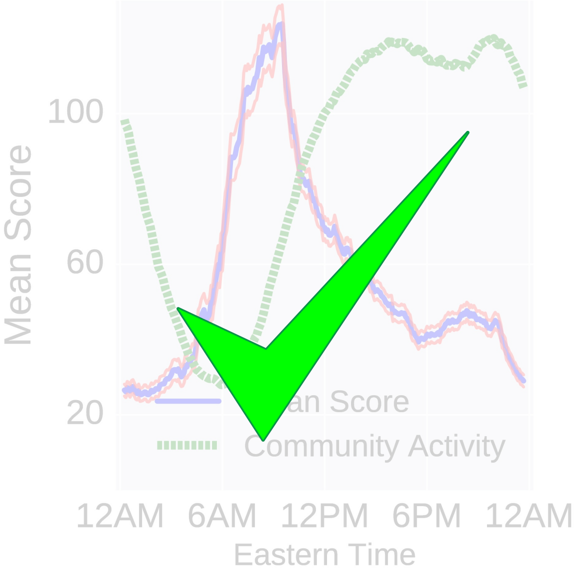
Idea: impose strict timing controls

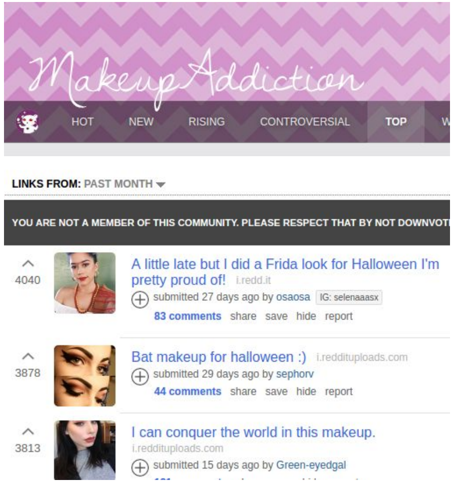
The grass is always greener

This is why you get two cats

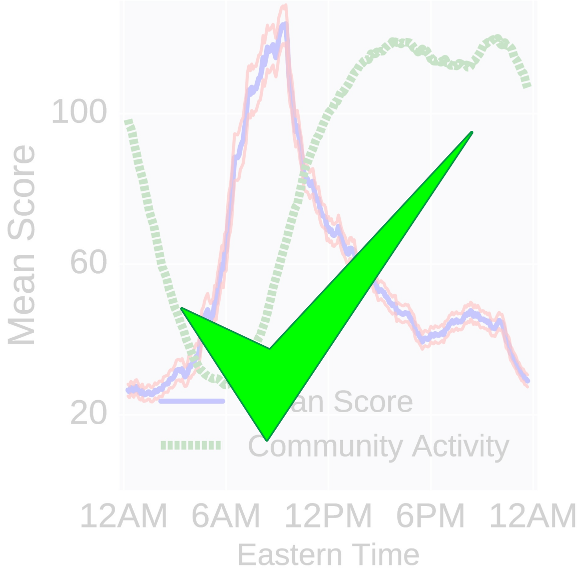
13 Seconds Apart!

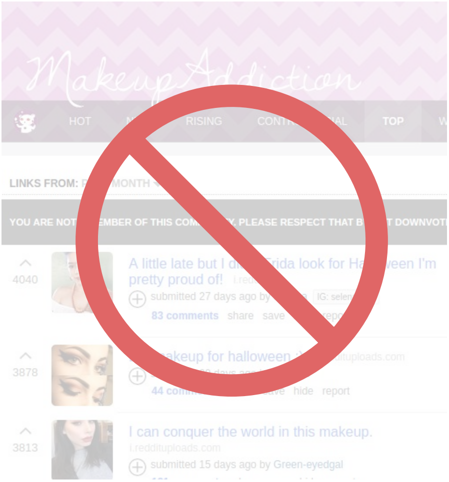

MUSIC LAB
COLUMBIA UNIVERSITY
CLICK HERE TO START
FREE MUSIC DOWNLOADS


Mean Score
100
60
20
Community Activity
12AM 6AM 12PM 6PM 12AM
Eastern Time

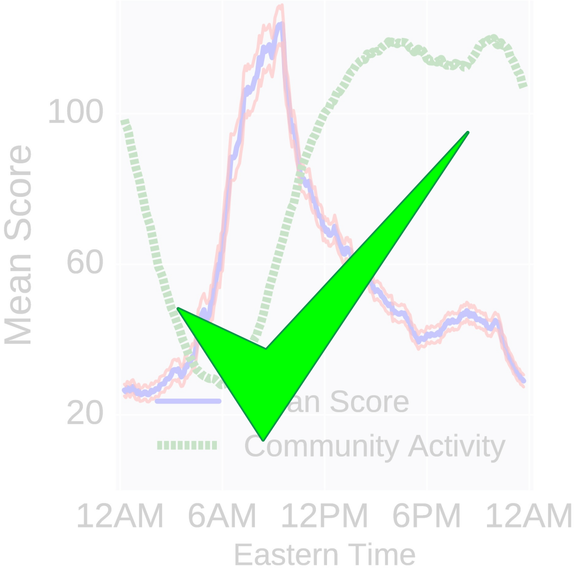

MakeupAddiction
HOT NEW RISING CONTROVERSIAL TOP
LINKS FROM: PAST MONTH
YOU ARE NOT A MEMBER OF THIS COMMUNITY. PLEASE RESPECT THAT BY NOT DOWNVOT
4040
A little late but I did a Frida look for Halloween I'm pretty proud of! i.redd.it
submitted 27 days ago by osaosa IG: selenaaasx
83 comments share save hide report
3878
Bat makeup for halloween :) i.reddituploads.com
submitted 29 days ago by sephorv
44 comments share save hide report
3813
I can conquer the world in this makeup.
i.reddituploads.com
submitted 15 days ago by Green-eyedgal

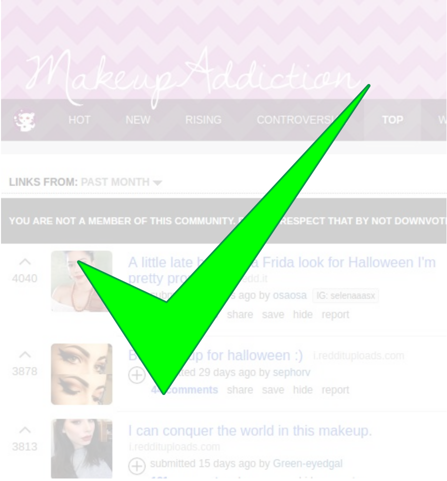

MUSIC LAB
COLUMBIA UNIVERSITY
CLICK HERE TO START
FREE MUSIC DOWNLOADS


Mean Score
100
60
20
Community Activity
12AM
6AM
12PM
6PM
12AM
Eastern Time

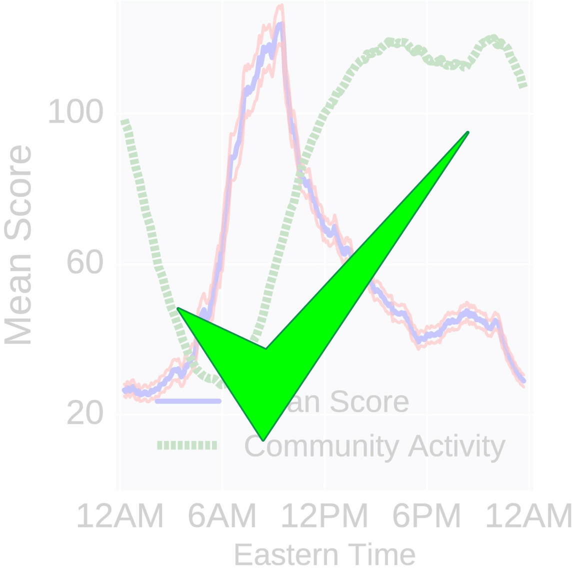

HOT
RISING
TOP
LINKS FROM:
4040
pretty proud of!
83 comments
share
save
3878
hide
report
3813
I can conquer the world in this makeup.
submitted 15 days ago by Green-eyedgal

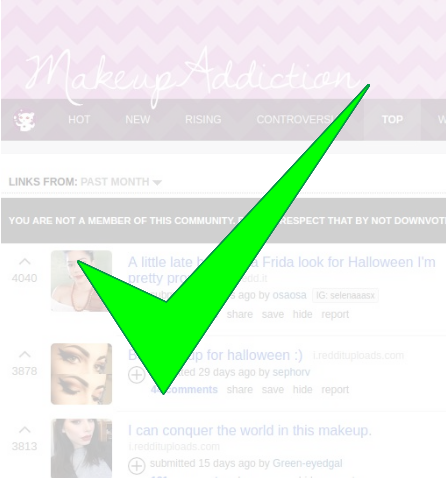

MUSIC LAB
COLUMBIA UNIVERSITY
CLICK HERE TO START
FREE MUSIC DOWNLOADS


Mean Score
100
60
20
Community Activity
12AM 6AM 12PM 6PM 12AM
Eastern Time


MakeupAddiction
HOT
NEW
RISING
TOP
LINKS FROM: PAST MONTH
4040
3878
3813
I can conquer the world in this makeup.
submitted 15 days ago by Green-eyedgal


MUSIC LAB
COLUMBIA UNIVERSITY
FREE MUSIC DOWNLOADS


Mean Score
100
60
20
Community Activity
12AM
6AM
12PM
6PM
12AM
Eastern Time


MakeupAddiction
HOT
NEW
RISING
TOP
LINKS FROM: PAST MONTH
4040
3878
3813
I can conquer the world in this makeup.
submitted 15 days ago by Green-eyedgal

The grass is always greener

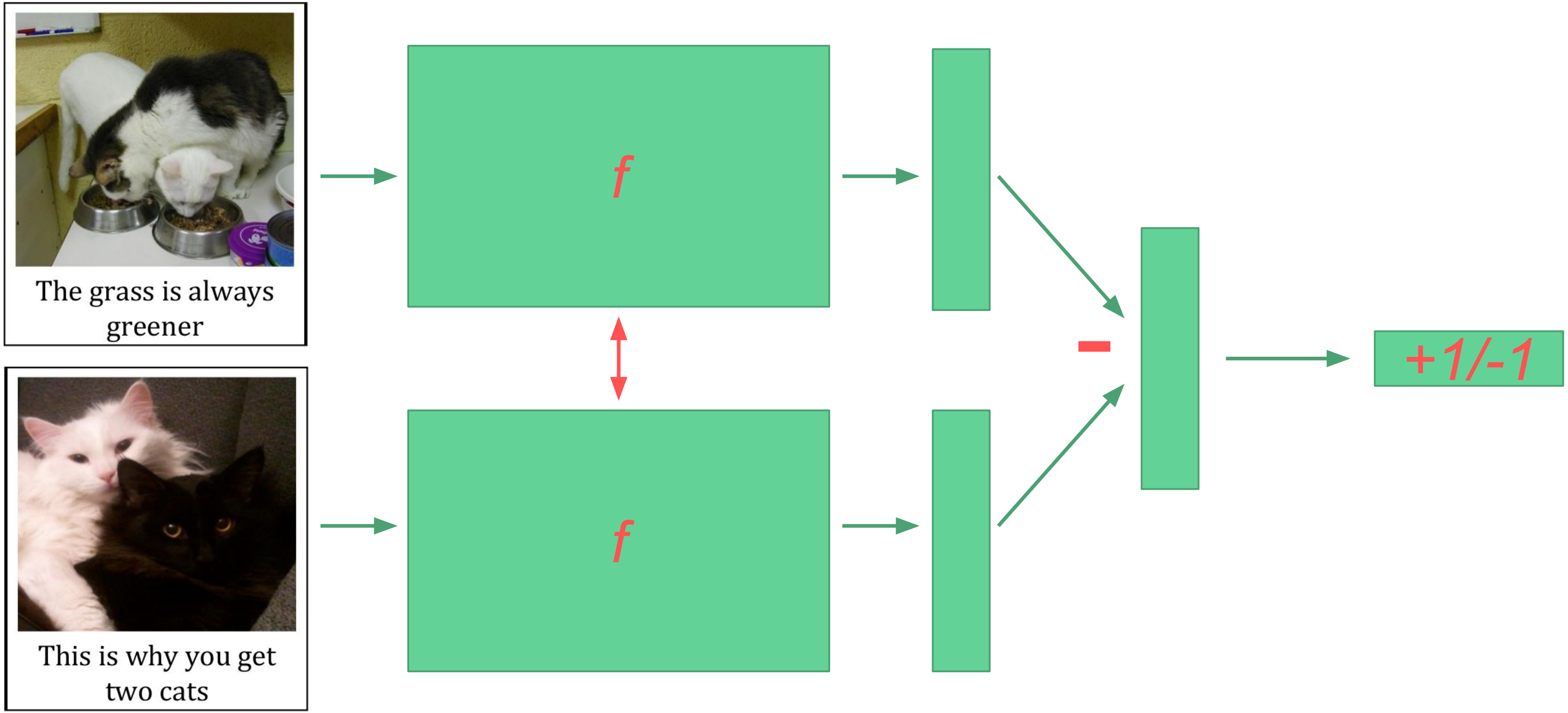
This is why you get two cats

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

	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

		aww	pics	cats	MA	FP	RL
Timing	Random	50.0	50.0	50.0	50.0	50.0	50.0
	Earlier	*51.7*	*51.1*	*49.9*	48.9	*48.6*	48.7
	Time	50.2	50.2	*50.7*	*50.4*	*49.7*	*50.6*

Machine learning experiments



Unimodal Results (crossval)

Unimodal Results (crossval)

	aww	pics	cats	MA	FP	RL
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*	**67.3**
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*
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	**64.8**	**60.0**	**62.6**	**64.9**	**65.2**	*64.2*

User Features

Text Features

Image Features

Unimodal Results (crossval)

	aww	pics	cats	MA	FP	RL
User Features						
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*	**67.3**
Text Features						
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*
Image Features						
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	**64.8**	**60.0**	**62.6**	**64.9**	**65.2**	*64.2*

Multimodal Results (crossval)

Multimodal Results (crossval)

	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

Multimodal Results (crossval)

Best unimodal

	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	**<u>68.4</u>**
ResNet50	64.8	60.0	62.6	64.9	*65.2*	64.2
Text + Image	**<u>67.1</u>**	**<u>62.7</u>**	**<u>65.9</u>**	**<u>67.7</u>**	**65.8**	66.4

Multimodal Results (crossval)

Best unimodal

	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

Multimodal beats unimodal!

Multimodal Results (crossval + fully heldout)

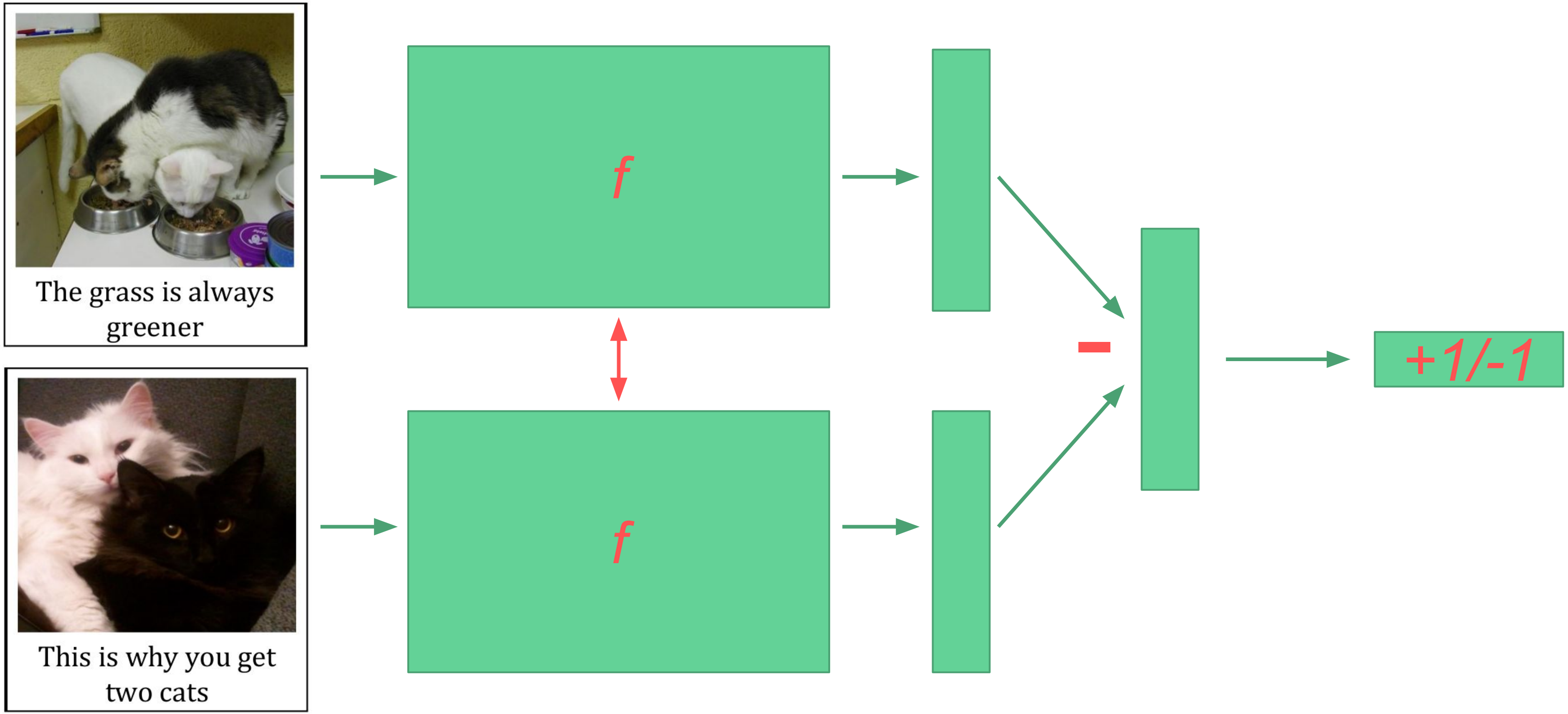
Best unimodal

	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

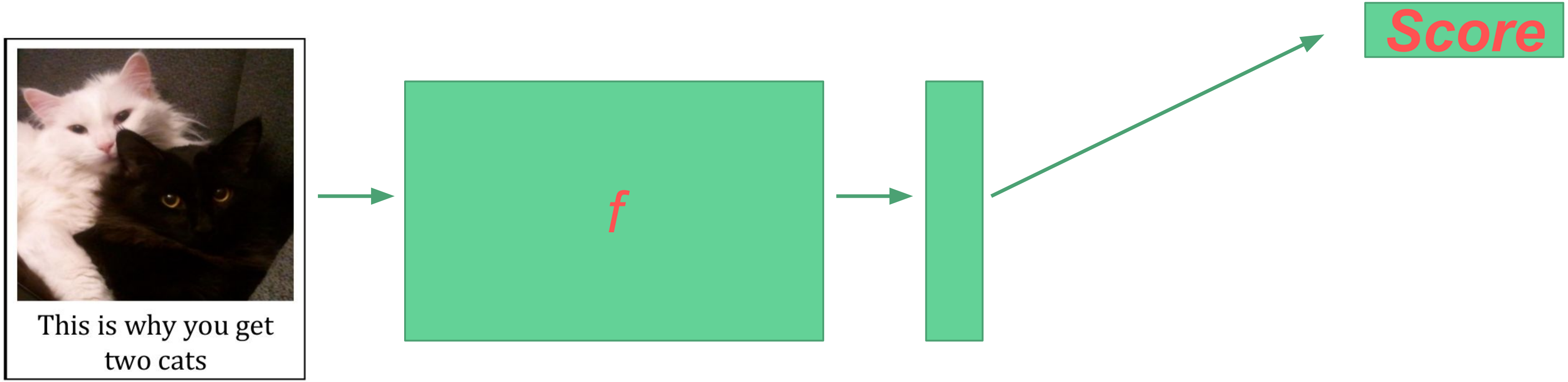
Multimodal beats unimodal!

	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

Machine learning experiments



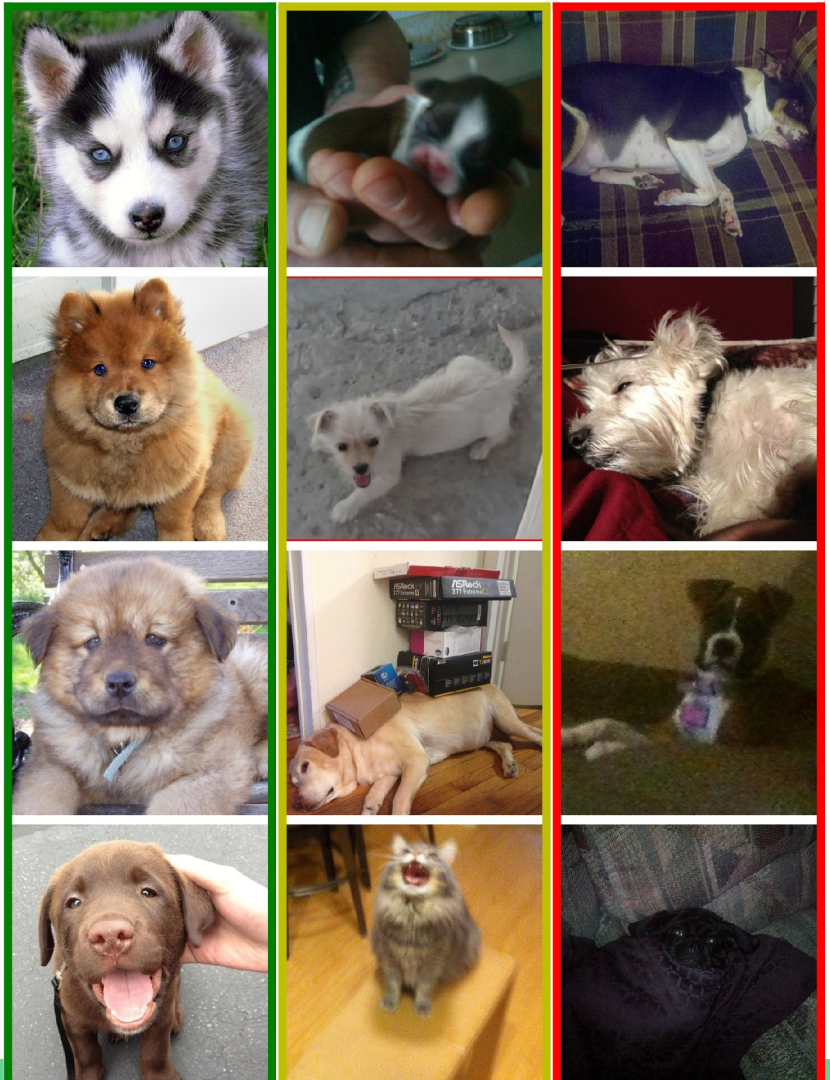
Machine learning experiments



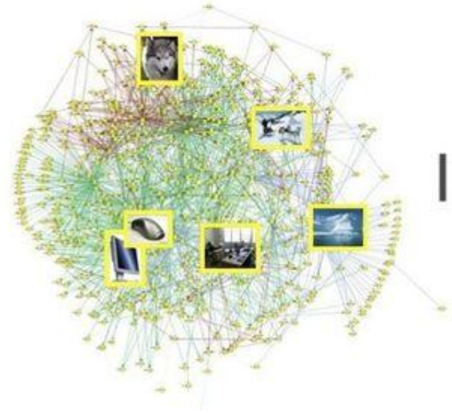
Highest Scores

Lowest Scores

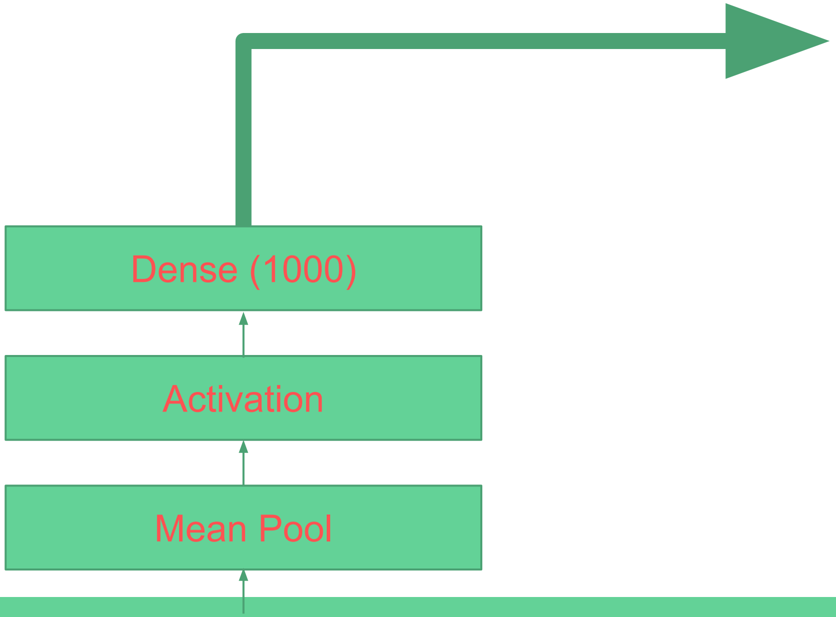
Highest Scores



Lowest Scores



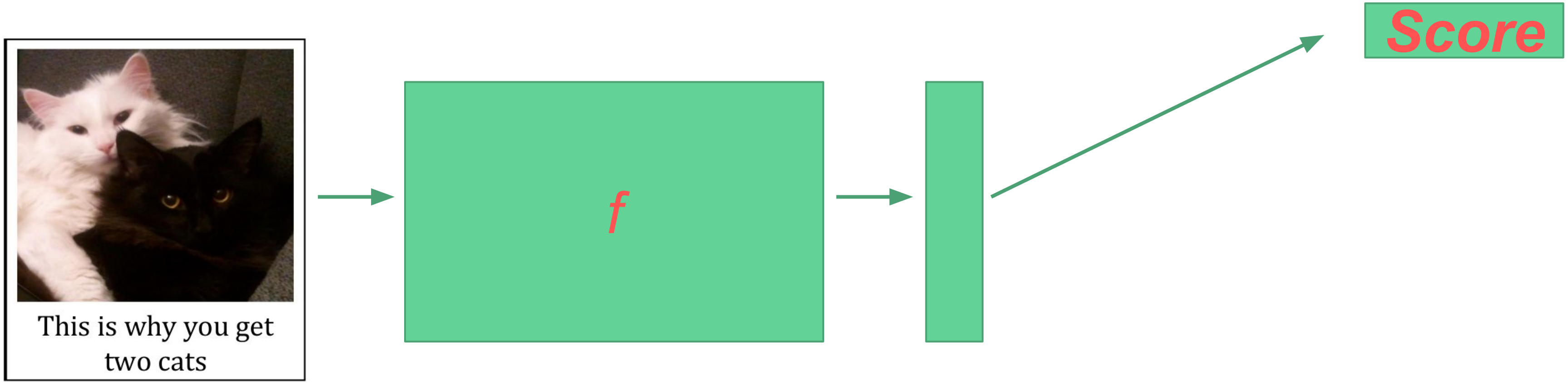
IMAGENET



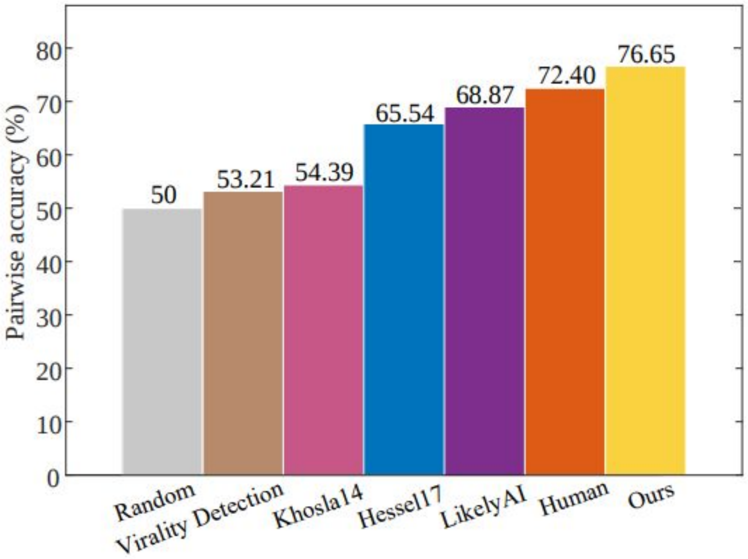

```
golden_retriever +0.2290 ***
dingo    +0.2126 ***
Labrador_retriever    +0.1960 ***
worm_fence    +0.1864 ***
cheetah    +0.1851 ***
Tibetan_mastiff    +0.1830 ***
...
Scotch_terrier    -0.2193 ***
bassinet    -0.2196 ***
wardrobe    -0.2231 ***
miniature_schnauzer    -0.2343 ***
four-poster    -0.2841 ***
mosquito_net    -0.2936 ***
```

(Significant after applying bonferroni correction)

Machine learning experiments



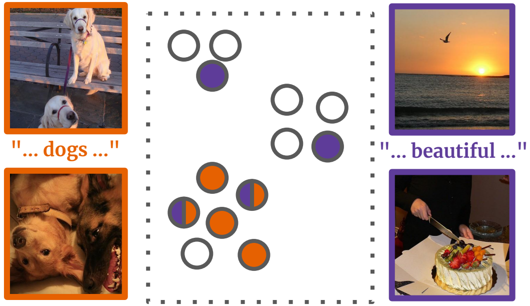
More evidence that controls are important: our models transfer well to other domains!



[Ding et al. 2019's instagram results]

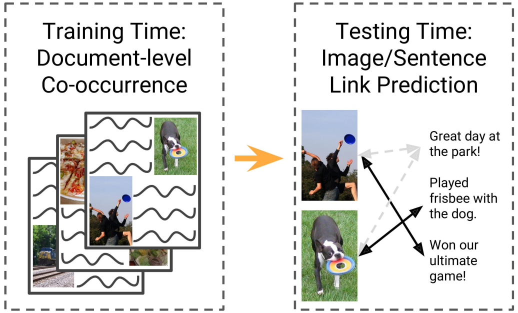
Does *multimodality* affect community reception of content?

[WWW 2017, H., Lee, Mimno]



What concepts are "groundable," and in what context?

[NAACL 2018, H., Mimno, Lee]

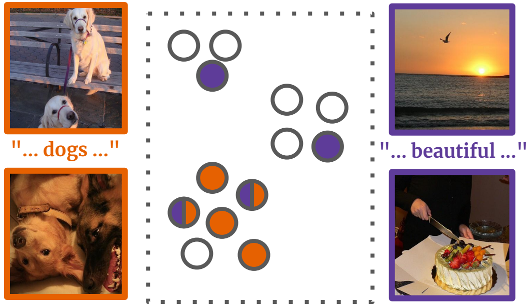


Can grounding be learned directly from multi-sentence, multi-image web documents?

[EMNLP 2019, H., Lee, Mimno]

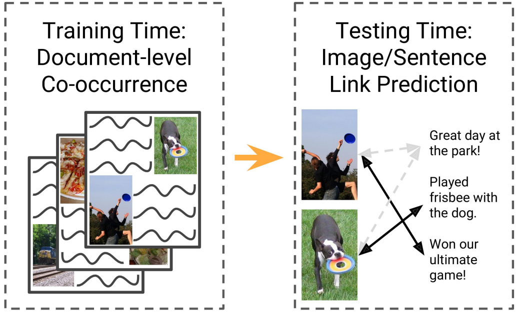
Does *multimodality* affect community reception of content?

[WWW 2017, H., Lee, Mimno]



What concepts are "groundable," and in what context?

[NAACL 2018, H., Mimno, Lee]



Can grounding be learned directly from multi-sentence, multi-image web documents?

[EMNLP 2019, H., Lee, Mimno]

"Performance advantages of [multi-modal approaches] over language-only models have been clearly established when models are required to learn ***concrete noun concepts****."*

[Hill and Korhonen 2014]

Many datasets focus only on literal objects/actions...

COCO
Common Objects in Context

[Lin et al 2014]

The man at bat readies to swing at the pitch while the umpire looks on.

"Do not describe what a person might say."
--- MSCOCO caption annotation guideline for mechanical turkers

... but we encounter lots of non-concrete language on the web!

Work on identifying hard/easy-to-ground concepts:

[Lu et al., 2008; Berg et al., 2010; Parikh and Grauman, 2011; Young et al., 2014; Kiela and Bottou, 2014; Jas and Parikh, 2015; Lazaridou et al., 2015; Silberer et al., 2016; Lu et al., 2017; Bhaskar et al., 2017; Mahajan et al., 2018; inter alia]

Our contributions:

- Fast algorithm for computing concreteness
- Extension from unigrams/bigrams to LDA topics
- Demonstration that concreteness is context specific

The **cat** is in the grass.

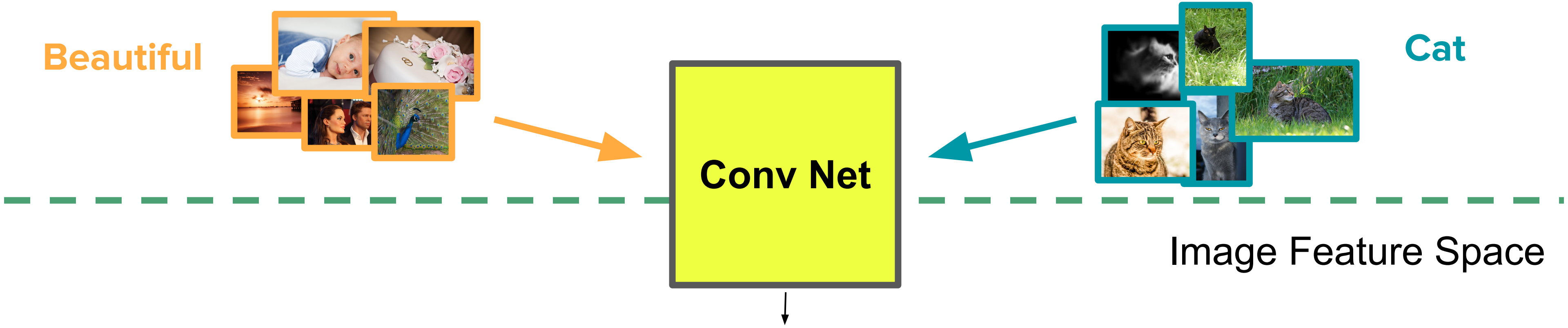
This **cat** is enjoying the sun.

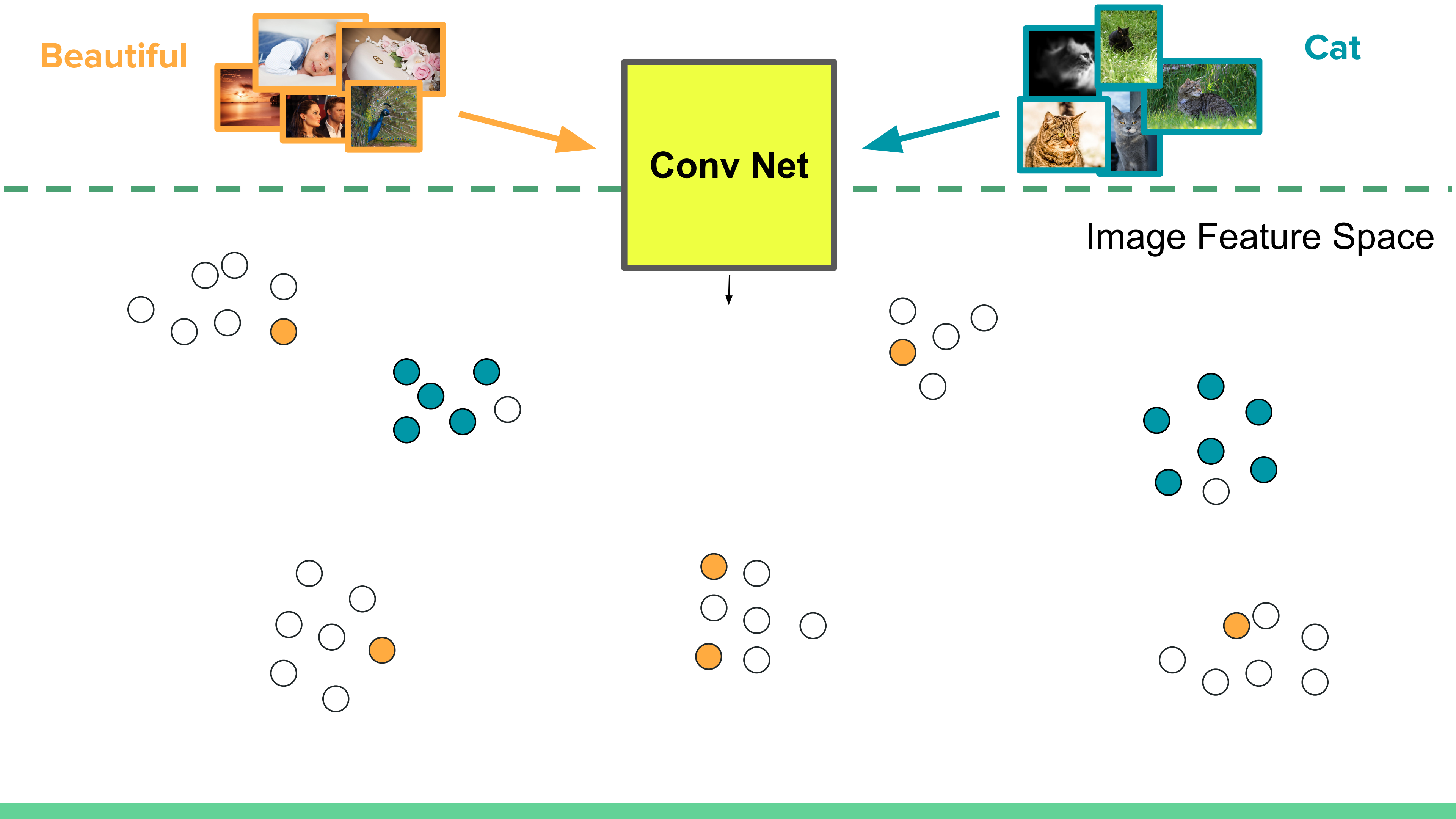
The **cat** is in the grass.

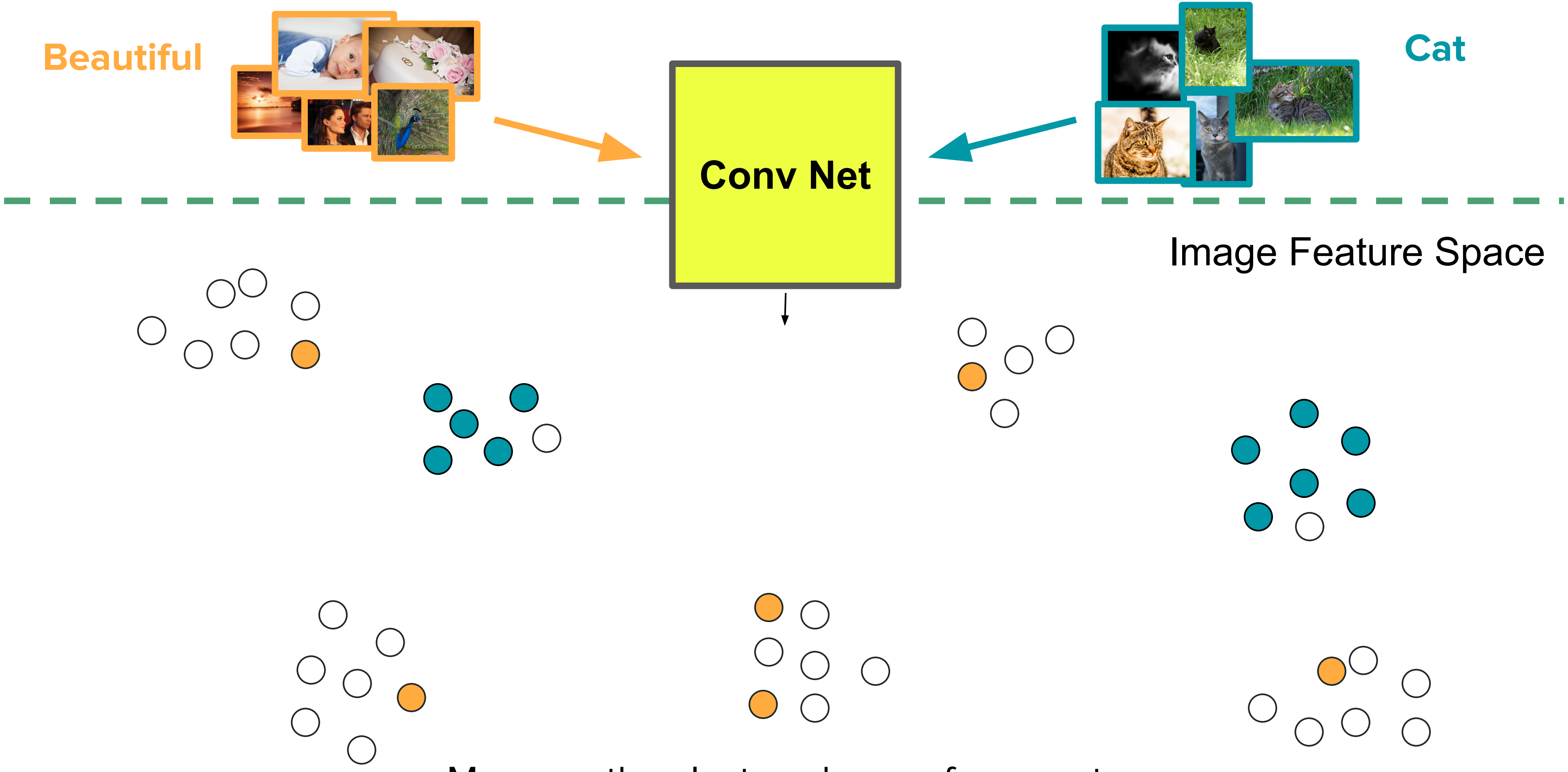
This **cat** is enjoying the sun.

This is a **beautiful** baby.

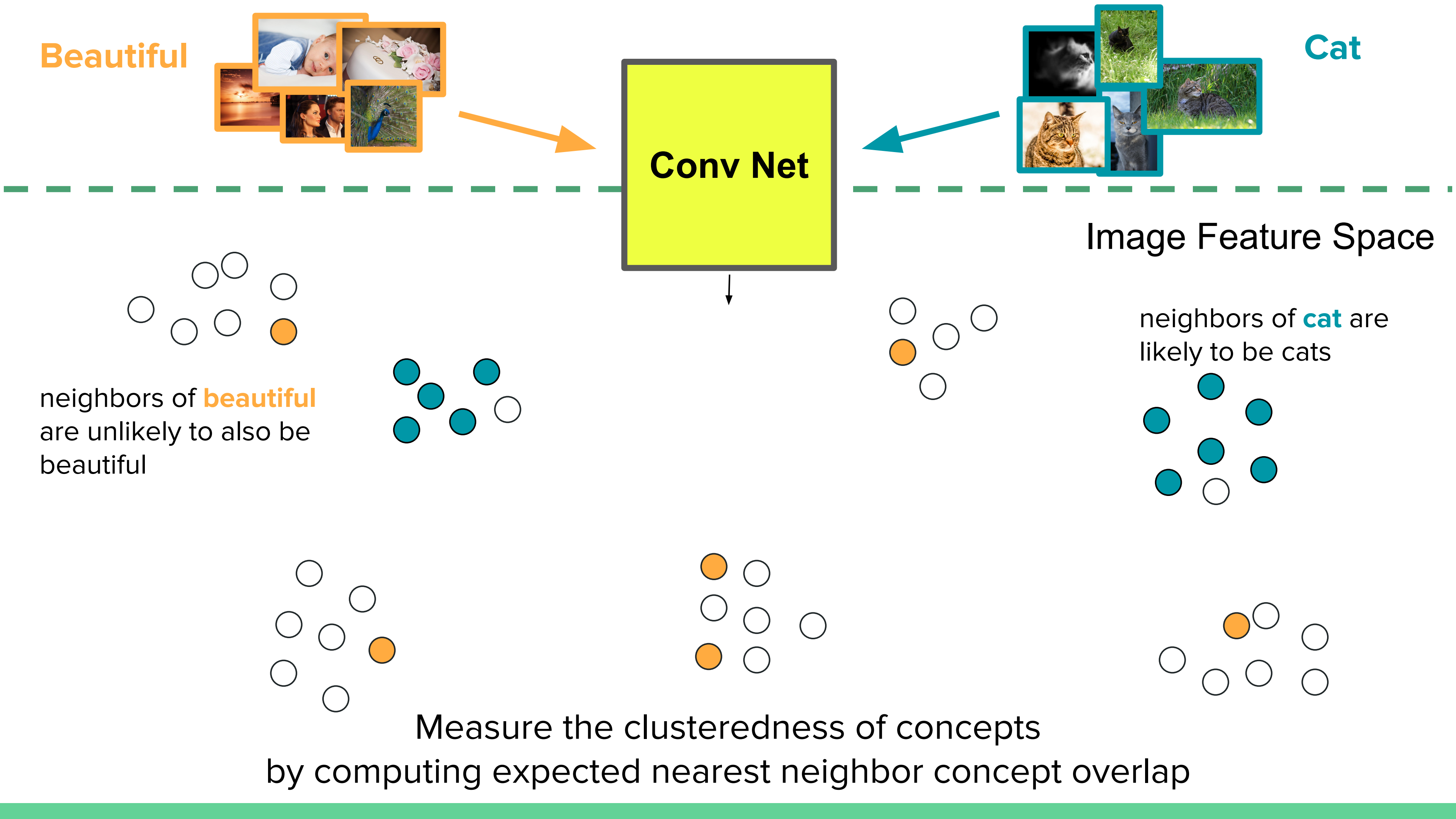
The sunset is **beautiful**.


Beautiful
Cat
Conv Net
Image Feature Space


Beautiful
Conv Net
Cat
Image Feature Space



Measure the clusteredness of concepts
by computing expected nearest neighbor concept overlap


Beautiful
Cat
Conv Net
Image Feature Space
neighbors of cat are likely to be cats
neighbors of beautiful are unlikely to also be beautiful
Measure the clusteredness of concepts
by computing expected nearest neighbor concept overlap

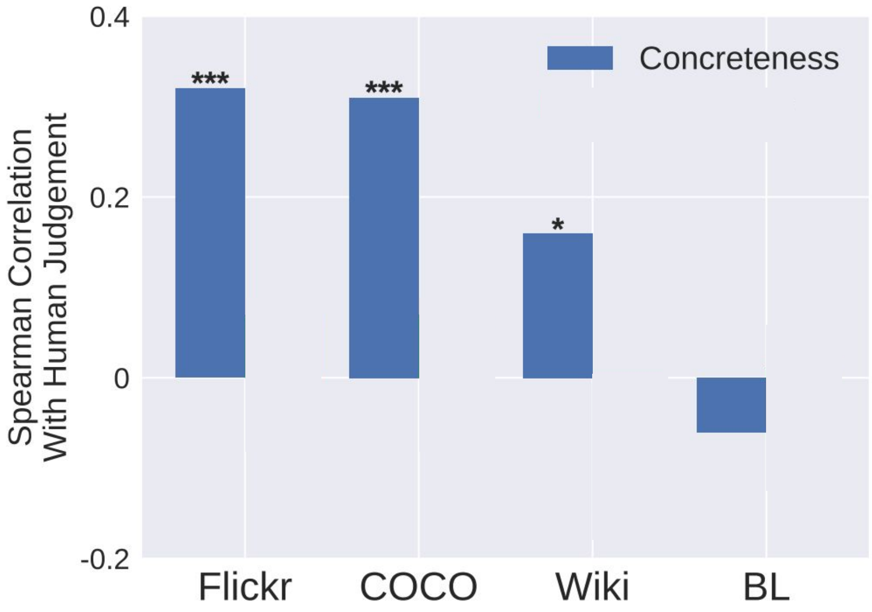
Connection to Geospatial Statistics

Local Indicators of Spatial Association—LISA

The capabilities for visualization, rapid data retrieval, and manipulation in geographic information systems (GIS) have created the need for new techniques of exploratory data analysis that focus on the "spatial" aspects of the data. The identification of local patterns of spatial association is an important concern in this respect. In this paper, I outline a new general class of local indicators of spatial association (LISA) and show how they allow for the decomposition of global indicators, such as Moran's I, into the contribution of each observation.

[Amelin 1995]

"Clusteredness" ≍ Concreteness



COCO Results

The man at bat readies to swing at the pitch while the umpire looks on.

COCO Results

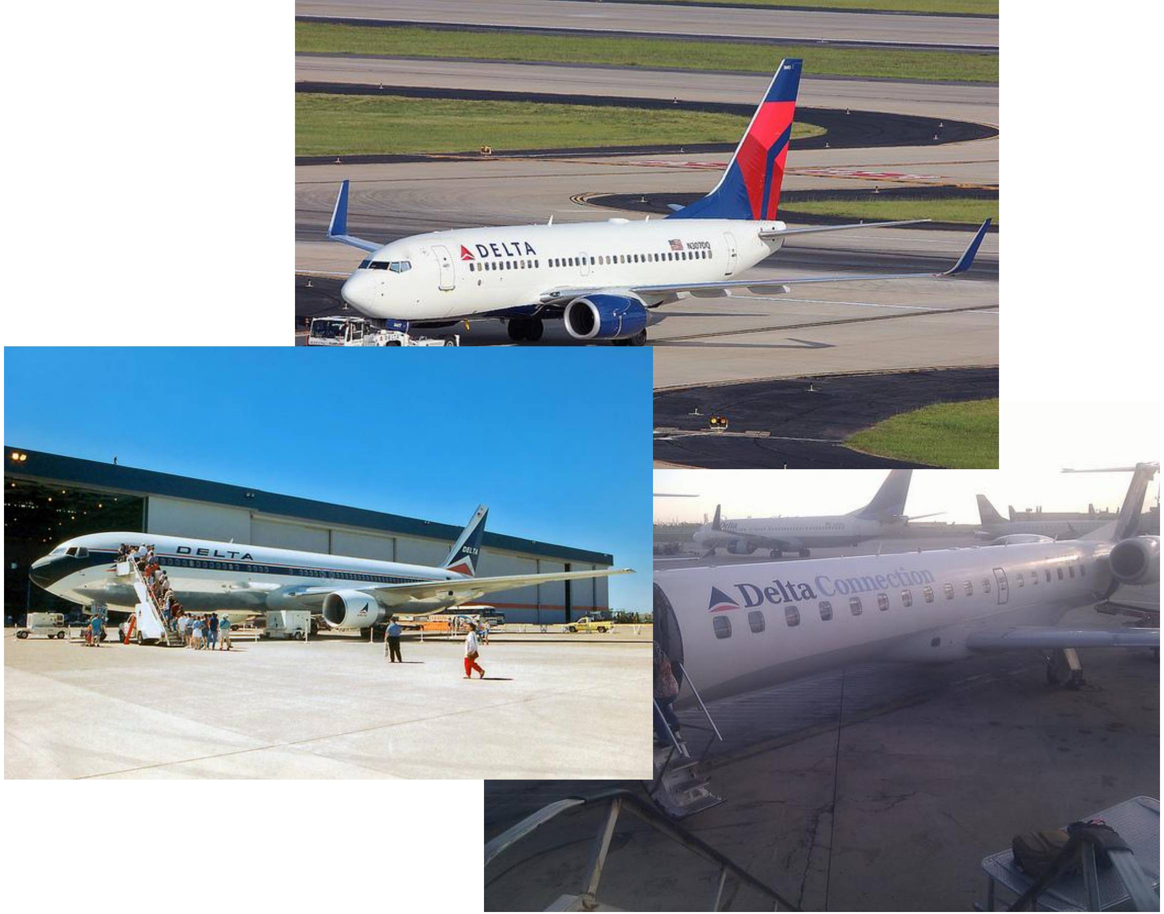
Most concrete

wok	315.595
hummingbird	291.804
vane	290.037
racer	269.043
grizzly	229.274
equestrian	219.894
taxiing	205.410
unripe	201.733
siamese	199.024
delta	195.618
kiteboarding	192.459
airways	183.971
compartments	182.015
burners	180.553
stocked	177.472
spire	177.396
tulips	173.850
ben	171.936

COCO Results

Most concrete

wok	315.595
hummingbird	291.804
vane	290.037
racer	269.043
grizzly	229.274
equestrian	219.894
taxiing	205.410
unripe	201.733
siamese	199.024
delta	**195.618**
kiteboarding	192.459
airways	183.971
compartments	182.015
burners	180.553
stocked	177.472
spire	177.396
tulips	173.850
ben	171.936



COCO Results

Most concrete

wok	315.595
hummingbird	291.804
vane	290.037
racer	269.043
grizzly	229.274
equestrian	219.894
taxiing	205.410
unripe	**201.733**
siamese	199.024
delta	195.618
kiteboarding	192.459
airways	183.971
compartments	182.015
burners	180.553
stocked	177.472
spire	177.396
tulips	173.850
ben	171.936

COCO Results

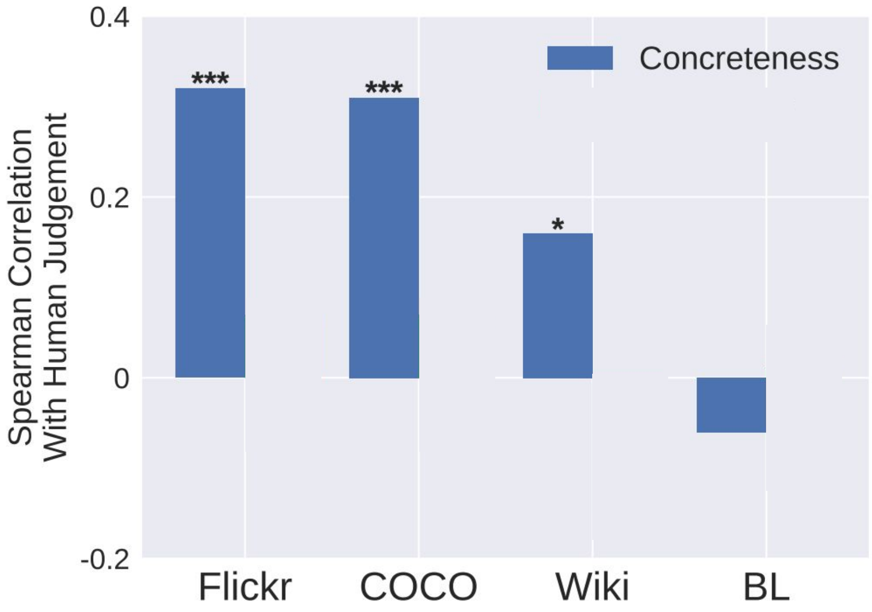
Most concrete

wok	315.595
hummingbird	291.804
vane	290.037
racer	269.043
grizzly	229.274
equestrian	219.894
taxiing	205.410
unripe	201.733
siamese	199.024
delta	195.618
kiteboarding	192.459
airways	183.971
compartments	182.015
burners	180.553
stocked	177.472
spire	177.396
tulips	173.850
ben	171.936

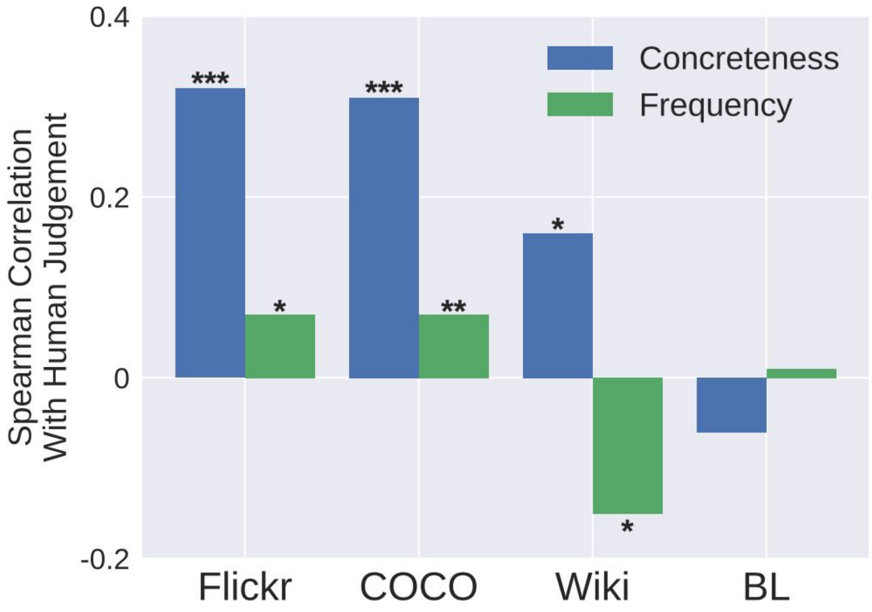
COCO Results

Most concrete		Somewhat concrete		Not concrete	
wok	315.595	motorcycle	10.291	side	1.770
hummingbird	291.804	fun	10.267	while	1.752
vane	290.037	including	10.262	other	1.745
racer	269.043	lays	10.232	sits	1.741
grizzly	229.274	fish	10.184	for	1.730
equestrian	219.894	goes	10.161	behind	1.709
taxiing	205.410	blurry	10.147	his	1.638
unripe	201.733	helmet	10.137	as	1.637
siamese	199.024	itself	10.128	image	1.620
delta	195.618	umbrellas	10.108	holding	1.619
kiteboarding	192.459	teddy	10.060	this	1.602
airways	183.971	bar	10.055	picture	1.589
compartments	182.015	fancy	10.053	couple	1.585
burners	180.553	sticks	10.050	from	1.569
stocked	177.472	himself	10.038	large	1.568
spire	177.396	take	10.016	person	1.561
tulips	173.850	steps	10.014	looking	1.502
ben	171.936	attempting	9.986	out	1.494

"Clusteredness" ≍ Concreteness

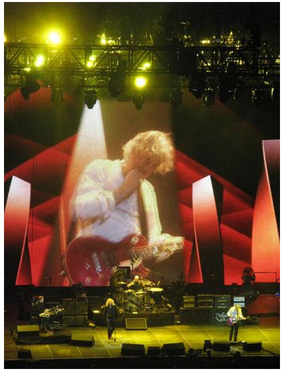


"Clusteredness" ≈ Concreteness



Context matters!

"London"
Top 1% Concrete
as a caption descriptor in
MSCOCO.

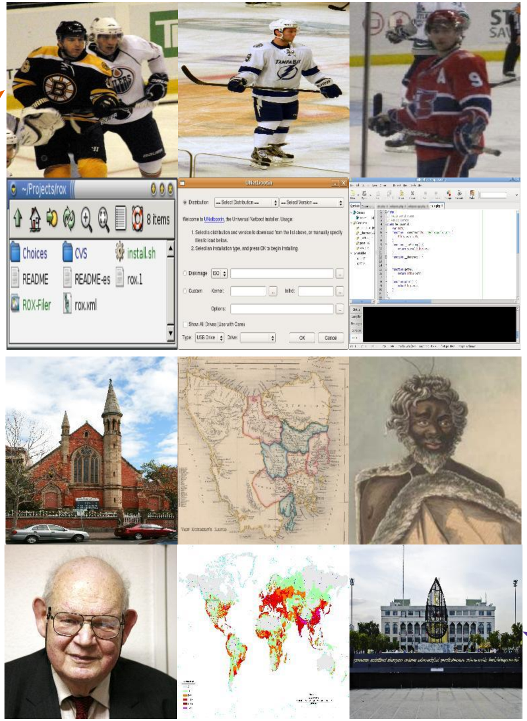


"#London"
Rank 1110/7K Concreteness
as a hashtag in a Flickr image
tagging dataset.

Experiments on Wikipedia with LDA topics:

Most Concrete

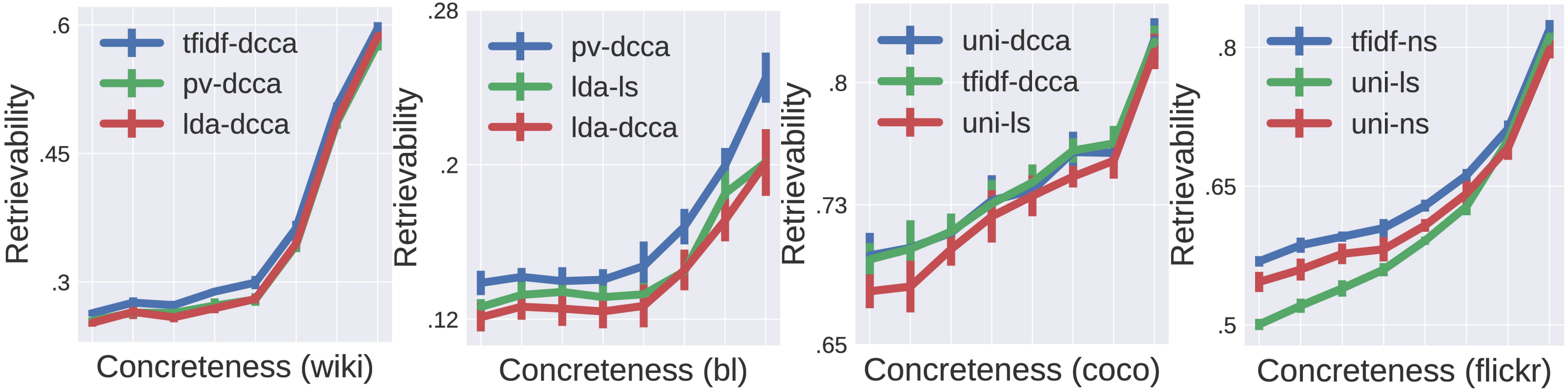
170.2	**hockey**
148.9	tennis
86.3	nintendo
81.9	guns
80.9	baseball
76.7	wrestling1
71.4	wrestling2
70.4	**software**
60.9	auto racing
58.8	currency



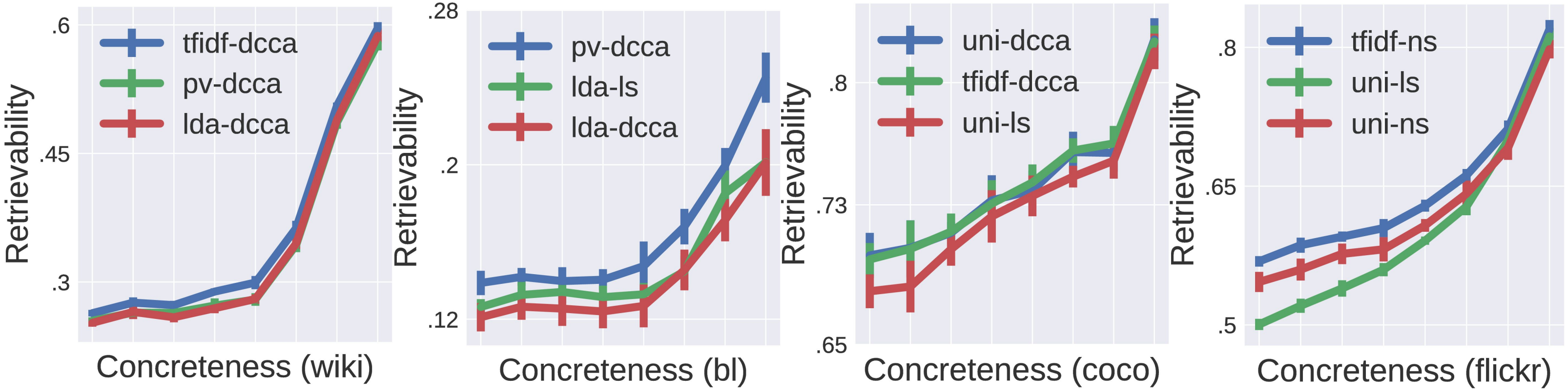
Least Concrete

australia	**1.95**
mexico	1.81
police	1.73
law	1.71
male names	1.65
community	1.58
history	1.52
time	1.47
months	1.43
linguistics	**1.29**

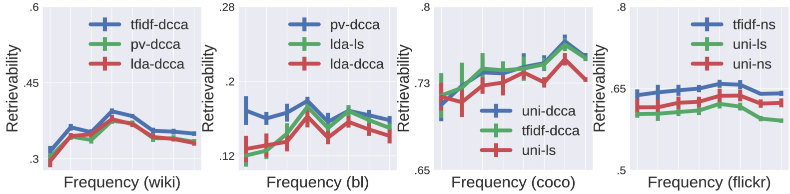
More concrete = easier to learn



More concrete = easier to learn

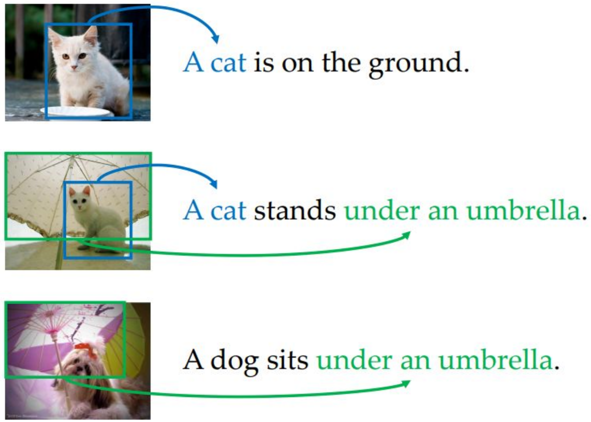


(more frequent =/= easier to learn)



Use Case from Shi et al. 2019 (ACL Best Paper Nom.)

Idea: unsupervised constituency parsing
based on the concreteness of spans in image captions

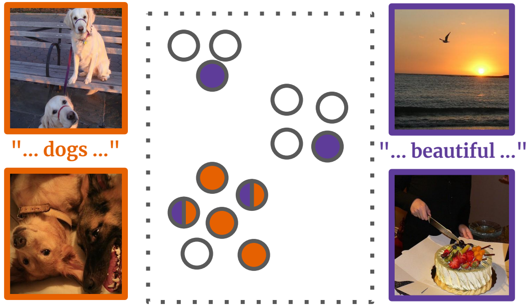


Model	NP	VP	PP	ADJP	Avg. F_1	Self F_1
Random	$47.3_{\pm 0.3}$	$10.5_{\pm 0.4}$	$17.3_{\pm 0.7}$	$33.5_{\pm 0.8}$	$27.1_{\pm 0.2}$	32.4
Left	51.4	1.8	0.2	16.0	23.3	N/A
Right	32.2	23.4	18.7	14.4	22.9	N/A
VG-NSL (ours)†	**79.6**$_{\pm 0.4}$	$26.2_{\pm 0.4}$	$42.0_{\pm 0.6}$	$22.0_{\pm 0.4}$	$50.4_{\pm 0.3}$	87.1
VG-NSL+HI (ours)†	$74.6_{\pm 0.5}$	$32.5_{\pm 1.5}$	**66.5**$_{\pm 1.2}$	$21.7_{\pm 1.1}$	**53.3**$_{\pm 0.2}$	**90.2**
VG-NSL+HI+FastText (ours)*†	$78.8_{\pm 0.5}$	$24.4_{\pm 0.9}$	$65.6_{\pm 1.1}$	$22.0_{\pm 0.7}$	**54.4**$_{\pm 0.4}$	89.8
Hessel et al. (2018)+HI†	72.5	**34.4**	65.8	26.2	52.9	N/A

(many more baselines in their paper)

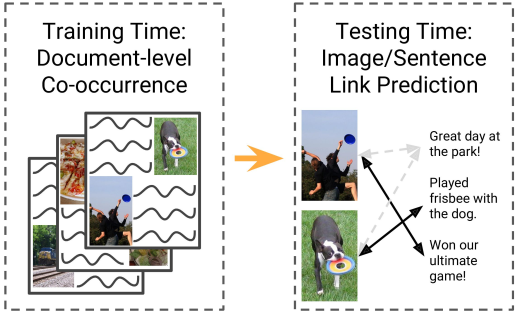
Does *multimodality* affect community reception of content?

[WWW 2017, H., Lee, Mimno]



What concepts are "groundable," and in what context?

[NAACL 2018, H., Mimno, Lee]

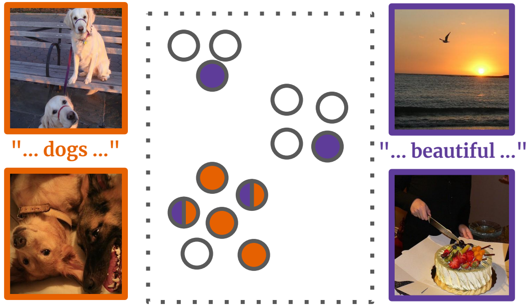


Can grounding be learned directly from multi-sentence, multi-image web documents?

[EMNLP 2019, H., Lee, Mimno]

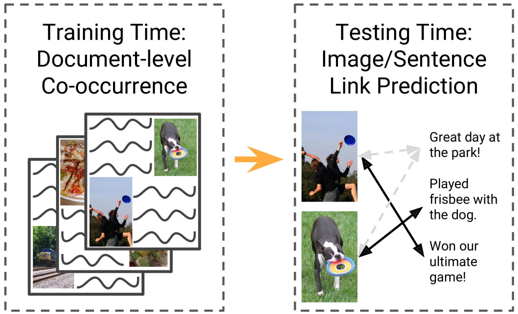
Does *multimodality* affect community reception of content?

[WWW 2017, H., Lee, Mimno]



What concepts are "groundable," and in what context?

[NAACL 2018, H., Mimno, Lee]



Can grounding be learned directly from multi-sentence, multi-image web documents?

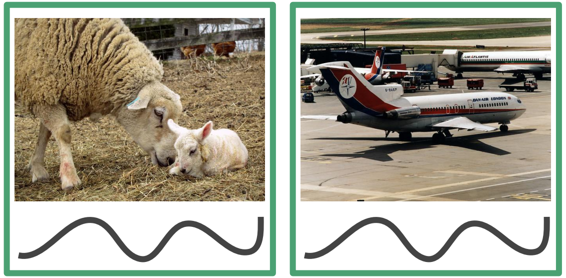
[EMNLP 2019, H., Lee, Mimno]

Multi-image, Multi-sentence documents?

Image captioning case:

one image, one sentence

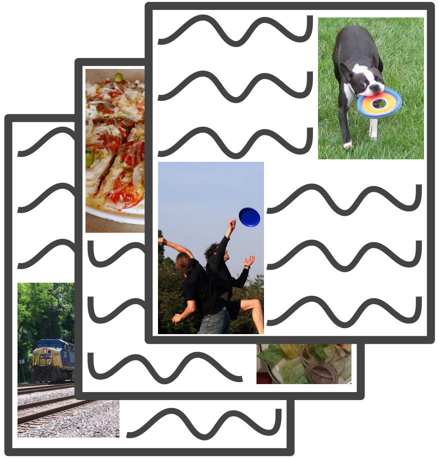
explicit link by annotation



Our case:

multiple images, multiple sentences

no explicit links

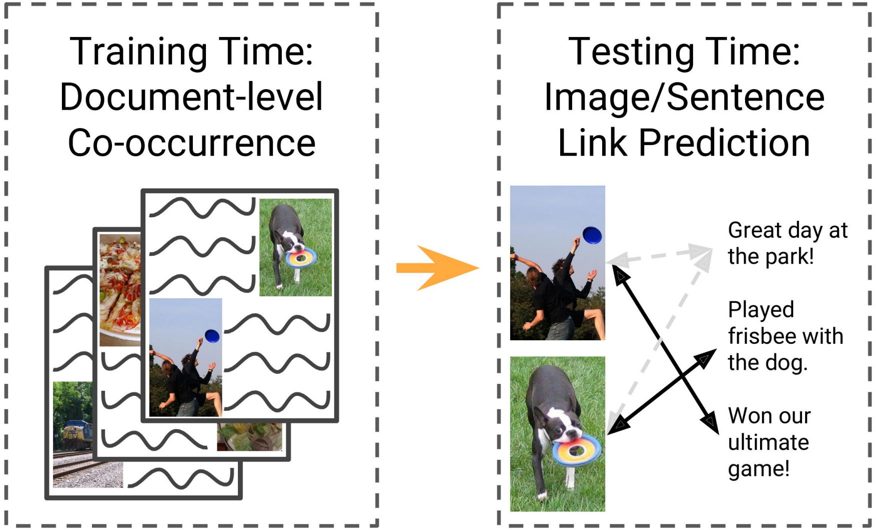


Why you might care about multi-image/multi-sentence documents

These types of documents are ubiquitous!

Web pages
Product listings
Books, current and historical
Web comments on images
News articles
...

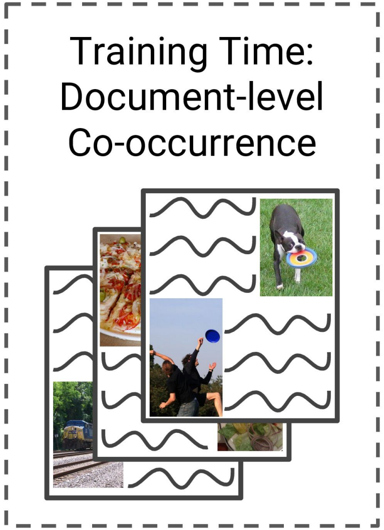
The Task: Unsupervised Link Prediction

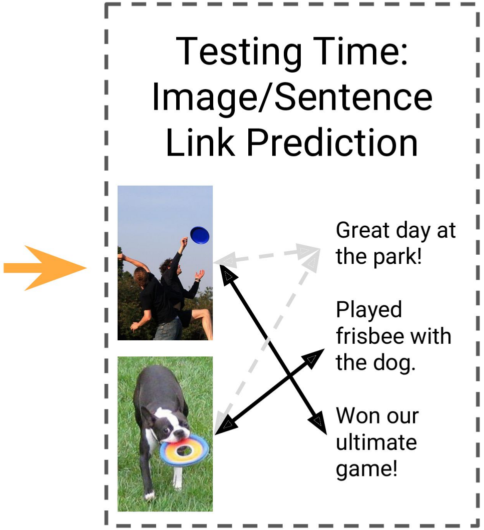


What's hard about this link prediction task?

- No explicit labels!
- Sentences may have no image
- Images may have no sentence
- Sentences may have multiple images
- Images may have multiple sentences

Evaluating Link Prediction





Metrics:

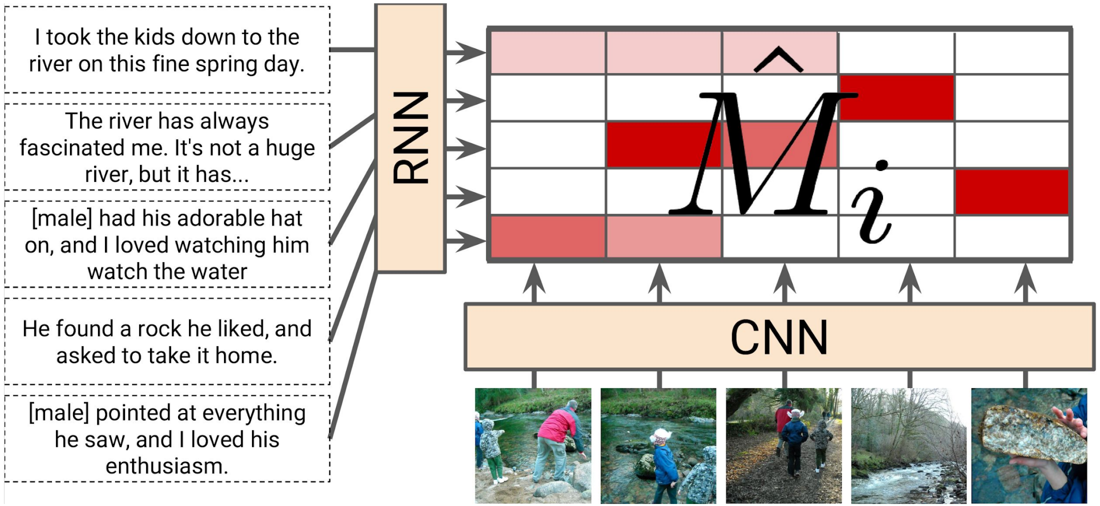
AUC: a standard link prediction metric

$$\propto \sum_{(i,j)\in G} \sum_{(i',j')\notin G} \mathbb{I}[s(i,j) > s(i',j')]$$

Precision-at-K (we use K=1,5):

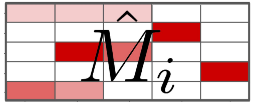
"If you had to make your K most confident predictions per-document, how accurate would you be?"

Model

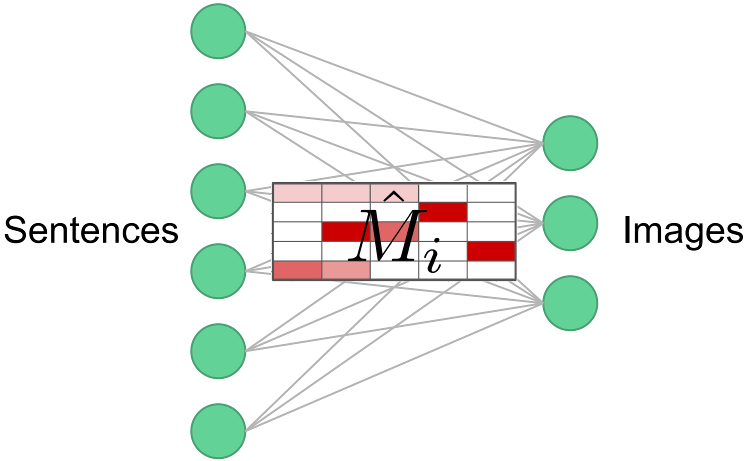


I took the kids down to the river on this fine spring day.
The river has always fascinated me. It's not a huge river, but it has...
[male] had his adorable hat on, and I loved watching him watch the water
He found a rock he liked, and asked to take it home.
[male] pointed at everything he saw, and I loved his enthusiasm.
RNN
$\hat{M}_i$
CNN

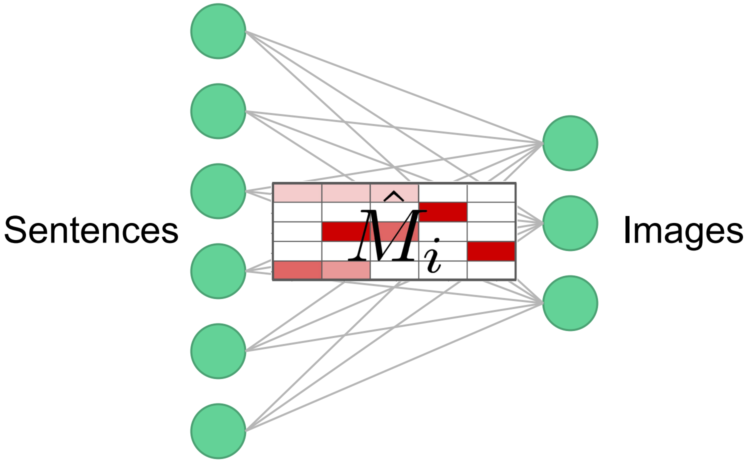
Model



Model

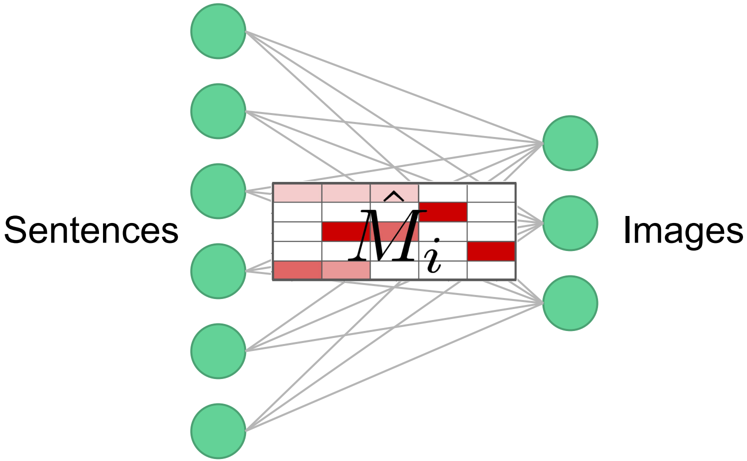


Model



maximize $\sum_{i,j} \widehat{M}_{ij} x_{ij}$

Model

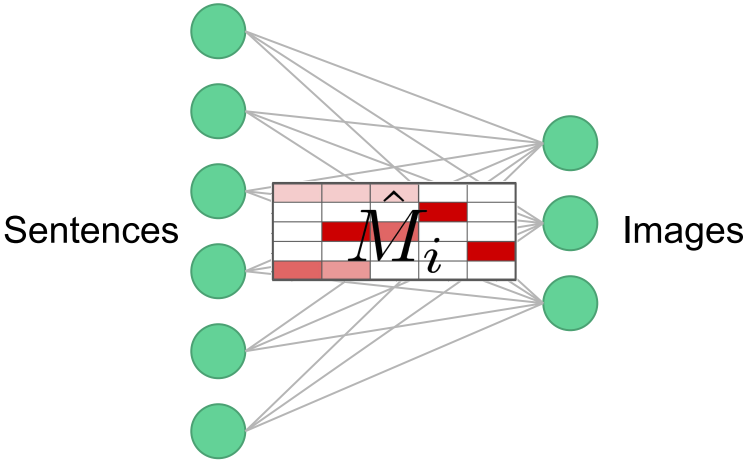


maximize $\sum_{i,j} \widehat{M}_{ij} x_{ij}$

$\forall i, \sum_j x_{ij} \leq 1; \forall j, \sum_i x_{ij} \leq 1; \forall i, j, x_{ij} \in \{0, 1\}$

to each image, no more than one sentence,
to each sentence, no more than one image

Model



maximize $$\sum_{i,j} \widehat{M}_{ij} x_{ij}$$

$$\forall i, \sum_j x_{ij} \leq 1; \forall j, \sum_i x_{ij} \leq 1; \forall i, j, x_{ij} \in \{0,1\}$$

to each image, no more than one sentence,
to each sentence, no more than one image

backprop through the solution x*: $$\mathrm{sim}(\widehat{M_i}) = \sum_{i,j} \widehat{M}_{ij} x^*_{ij}$$

Training: Max Margin loss with Negative Sampling

The river has always fascinated me. It's not a huge river, but it has...

I took the kids down to the river on this fine spring day.

He found a rock he liked, and asked to take it home.

[male] had his adorable hat on, and I loved watching him watch the water

[male] pointed at everything he saw, and I loved his enthusiasm.

Training: Max Margin loss with Negative Sampling

I took the kids down to the

The river has always

[male] had his adorable hat

He found a rock he liked

[male] pointed at everything he saw, and I loved his enthusiasm.

Training: Max Margin loss with Negative Sampling

S_i

I took the kids down to the

The river has always

[male] had his adorable hat

He found a rock he liked

[male] pointed at everything he saw, and I loved his enthusiasm.

V_i

Training: Max Margin loss with Negative Sampling

S_i

I took the kids down to the

The river has always

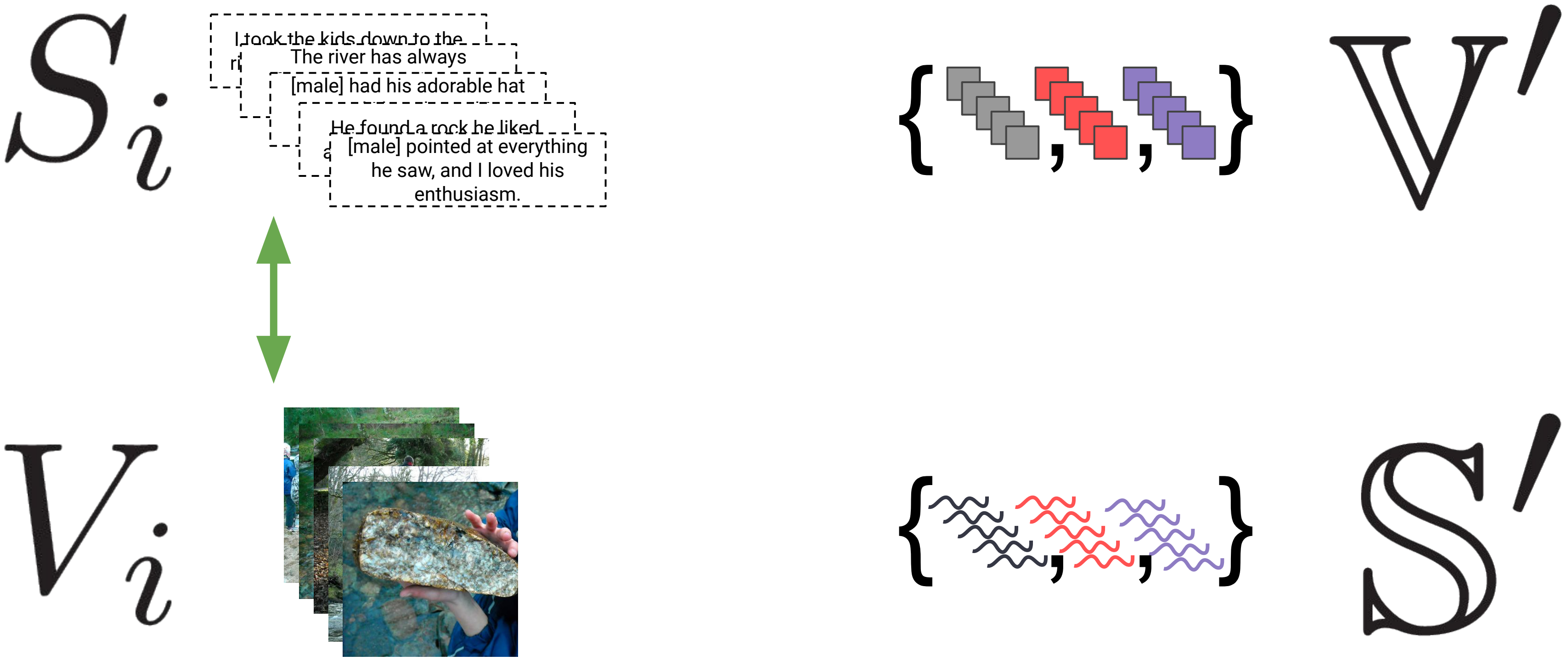
[male] had his adorable hat

He found a rock he liked

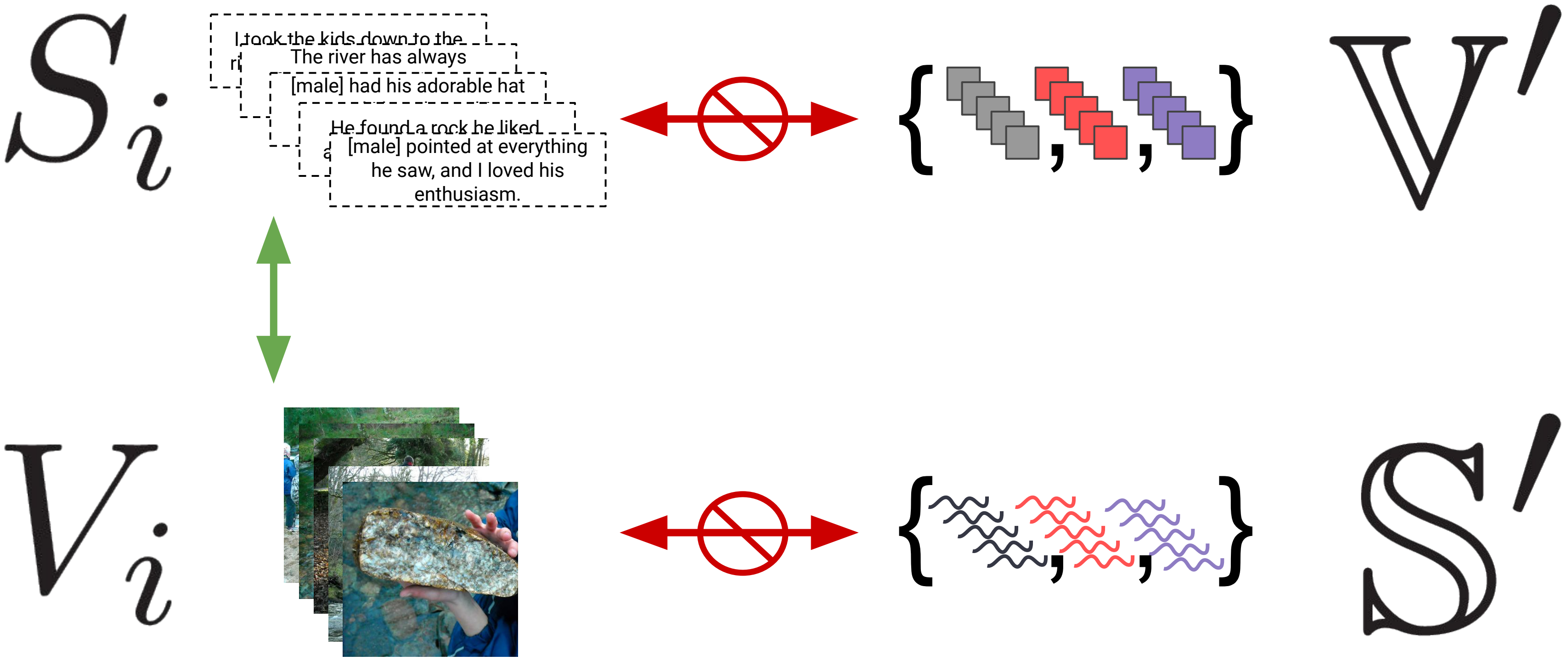
[male] pointed at everything he saw, and I loved his enthusiasm.

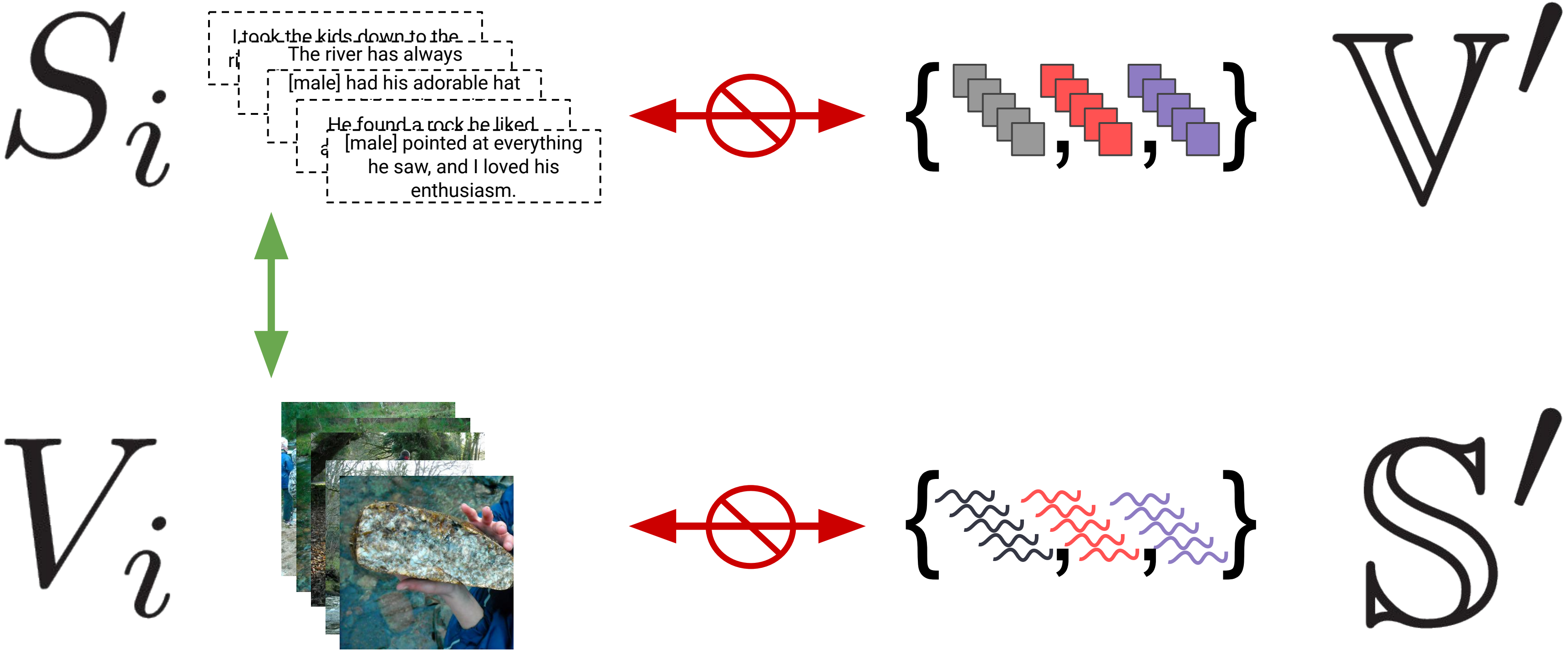
V_i

Training: Max Margin loss with Negative Sampling

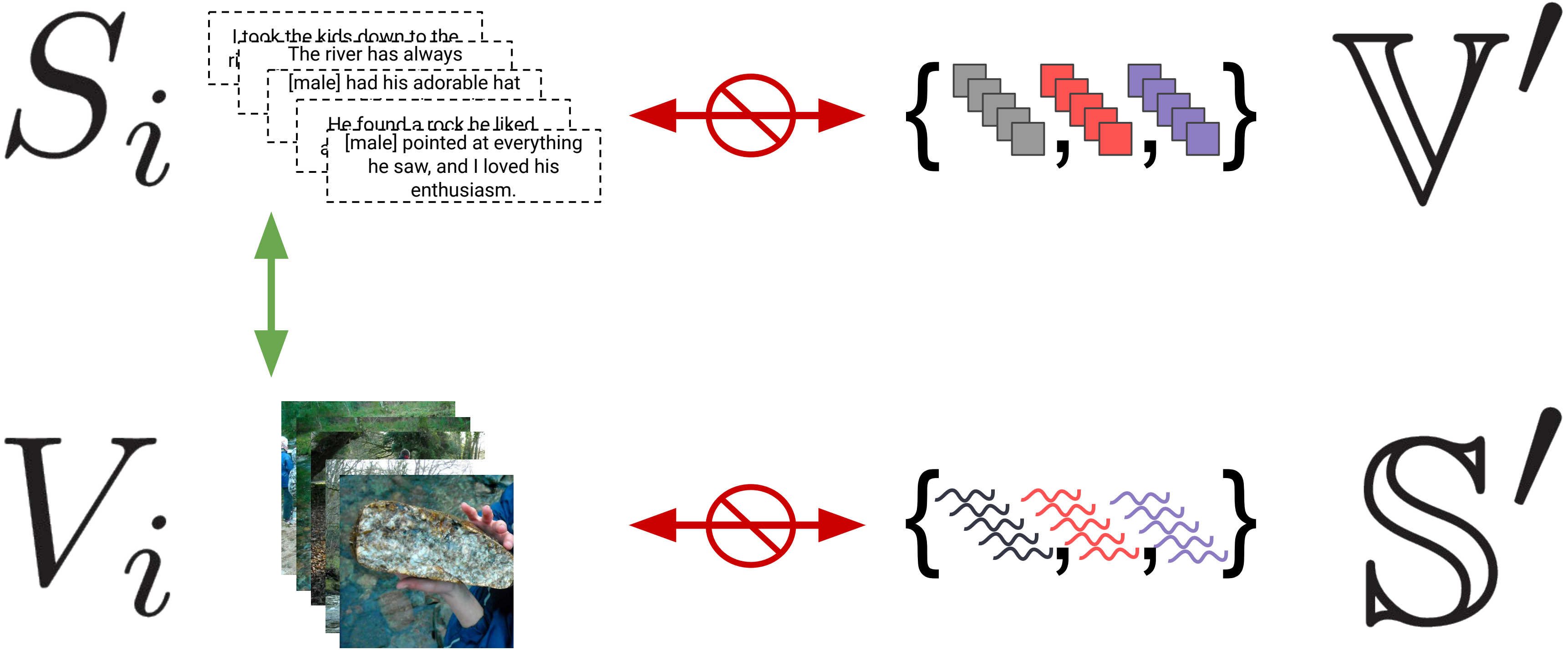


Training: Max Margin loss with Negative Sampling



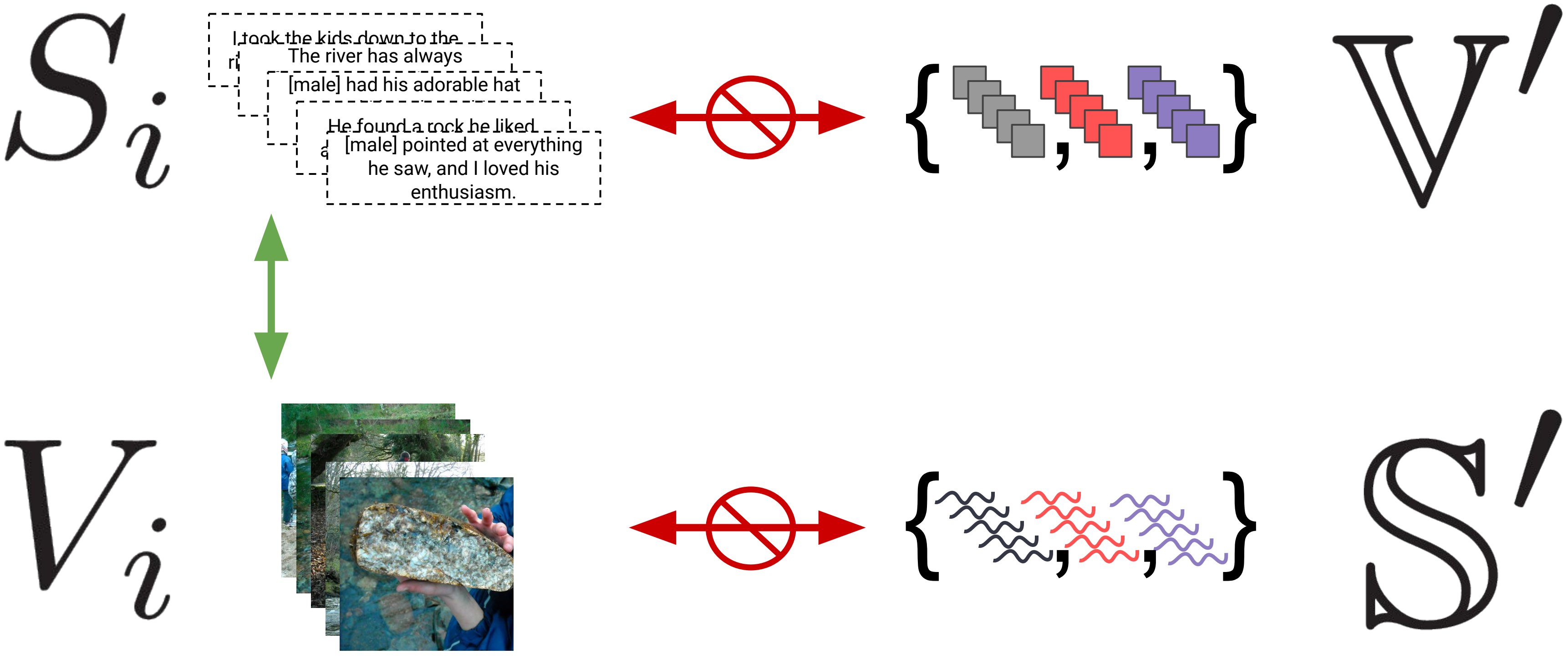


$$\mathcal{L}\left(S_i, V_i\right) = \max_{V' \in \mathbb{V}'} h\left(\mathrm{sim}(S_i, V_i), \mathrm{sim}(S_i, V')\right) + \max_{S' \in \mathbb{S}'} h\left(\mathrm{sim}(S_i, V_i), \mathrm{sim}(S', V_i)\right)$$



$$\mathcal{L}\left(S_i, V_i\right) = \max_{V' \in \mathbb{V}'} h\left(\boxed{\mathrm{sim}(S_i, V_i)}, \boxed{\mathrm{sim}(S_i, V')}\right)$$

$$+ \max_{S' \in \mathbb{S}'} h\left(\boxed{\mathrm{sim}(S_i, V_i)}, \boxed{\mathrm{sim}(S', V_i)}\right)$$



$$\mathcal{L}(S_i, V_i) = \max_{V' \in \mathbb{V}'} h\left(\boxed{\mathrm{sim}(S_i, V_i)}, \boxed{\mathrm{sim}(S_i, V')}\right) + \max_{S' \in \mathbb{S}'} h\left(\boxed{\mathrm{sim}(S_i, V_i)}, \boxed{\mathrm{sim}(S', V_i)}\right)$$

Hard-negative mining
(at the document level)

The paper has results on four crowdsourced datasets

but we'll focus on the web-scraped data for now...

Datasets Scraped from the Web

Datasets Scraped from the Web

Data scraped from instructables.com;

Via web interface, authors associate multiple images with recipe steps, which gives us a graph for evaluation

RecipeQA

Ingredients Mint Layer 1. 1 sticks butter 2. 1 cup powdered sugar 3. 1 table spoon milk ... *** Chocolate Layer #1 Although the chocolate layers are perhaps the simplest... until smooth *** Finishing First Layer 1. Pour evenly into a pan... *** Onto the Mint! The Mint mixture can be changed ... Second Layer Is Finished! Now comes a bit of a tricky part. ...The possibilities are endless :D *** Repeat Step #2 ... and final layer of your beautiful snack. *** Pulling It All Together! 1. Remove the dually layered bar ... *** Finishing Notes Allow the bar to acclimate...

Datasets Scraped from the Web

Data scraped from reddit's do-it-yourself (DIY) community

Via web interface, authors associate images with descriptive steps, which gives us a graph for evaluation

DIY

So my partner and I decided that we want to build our first In-Home rock climbing wall... *** We set aside a budget of $1200 and began a model to estimate... *** Each box represents one square foot of climbing space... *** After cutting a bit more plywood and lining it up... *** I insisted in putting a few cross braces into the angled section... *** I'm going to have fun with this.

Datasets Scraped from the Web

Data scraped from wikipedia

There are no ground-truth links between images and text (so we are limited to qualitative observation).

Imageclef-Wiki

Rivet A rivet is a permanent mechanical fastener... Solid rivets consist simply of a shaft and head... Steel rivets can be found in static structures such as bridges, cranes, ... They are offered from 1/16-inch (1.6 mm) to 3/8-inch (9.5 mm) in diameter ... The most common machine is the impact riveter and the most common use of semitubular rivets is in lighting, brakes ...

Stats for Web Datasets

	train/val/test	n_i/m_i (median)	# imgs (unique)	density
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

sentences/doc

images/doc

Quantitative Results on RQA/DIY

	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

Quantitative Results on RQA/DIY

	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
AP	**69.3**	**47.3/37.3**	**61.8**	**22.5/17.2**

Pour the quart of half-and-half into the blender. Weigh out about 120g...

First, fry up a pound of your favorite thin-sliced bacon. For this dish...

While I made a triple batch for competition, this recipe is scaled...

This layer will be your "meat" strip in the center of the bacon...

This one is just syrup and smoke. Combine 1cup bacon...

Example from RQA

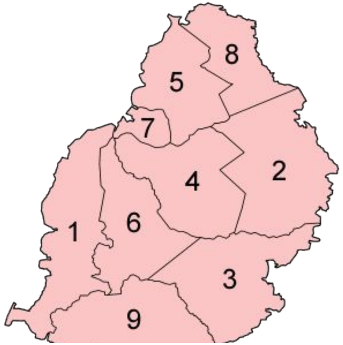
WIKI Prediction on Mauritius Article

This archipelago was formed in a series of undersea volcanic eruptions 8-10 million years ago...
(93.9)

The island is well known for its natural beauty.
(92.1)

First sighted by Europeans around 1600 on Mauritius, the dodo became extinct less than eighty years later.
(84.5)

... a significant migrant population of Bhumihar Brahmins in Mauritius who have made a mark for themselves in different fields.
(79.8)

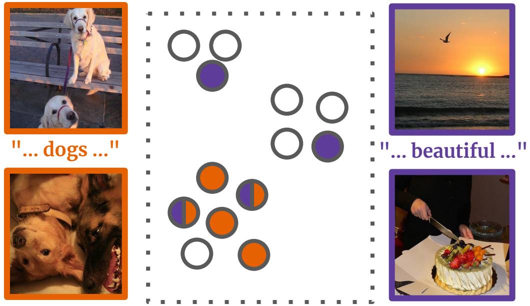


Mauritian Créole, which is spoken by 90 per cent of the population, is considered to be the native tongue...
(68.3)

For the dodo, the object detection baseline's selected sentence began with:
“(Mauritian Creole people usually known as ‘Creoles’)”

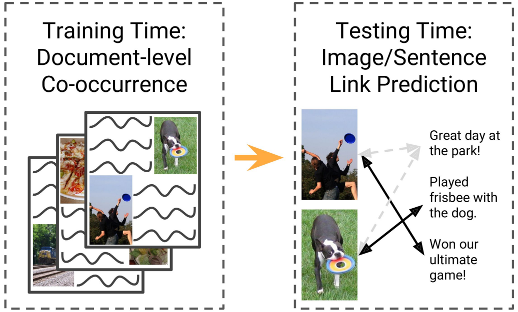
Does *multimodality* affect community reception of content?

[WWW 2017, H., Lee, Mimno]



What concepts are "groundable," and in what context?

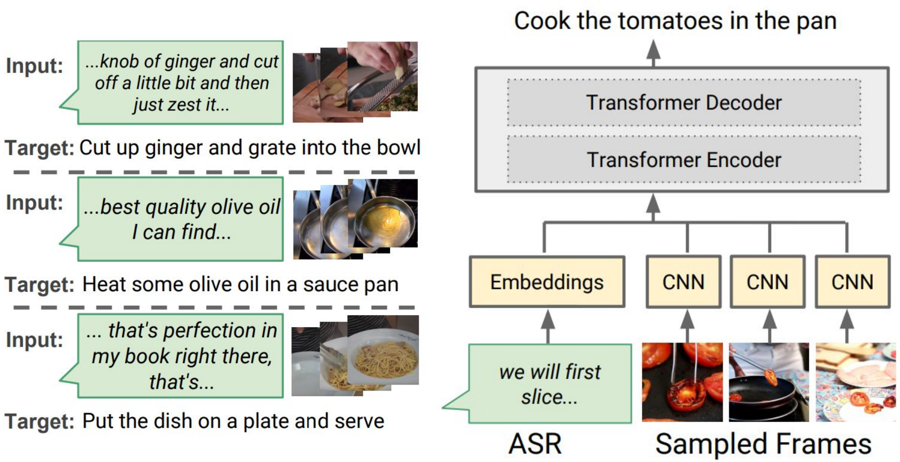
[NAACL 2018, H., Mimno, Lee]



Can grounding be learned directly from multi-sentence, multi-image web documents?

[EMNLP 2019, H., Lee, Mimno]

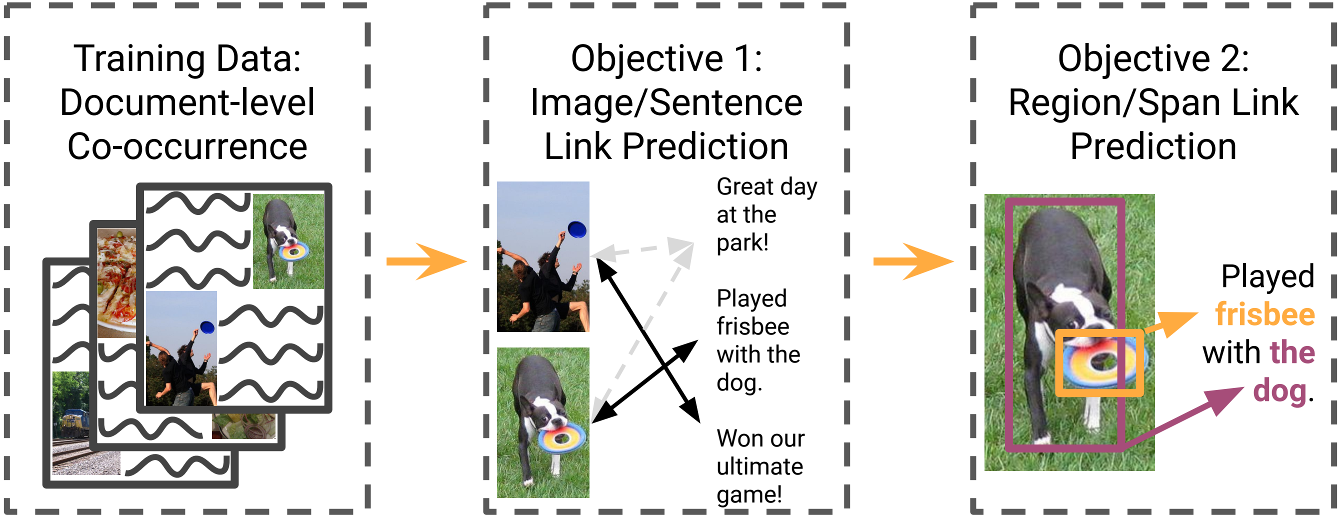
Ongoing work: exploring grounding in web videos



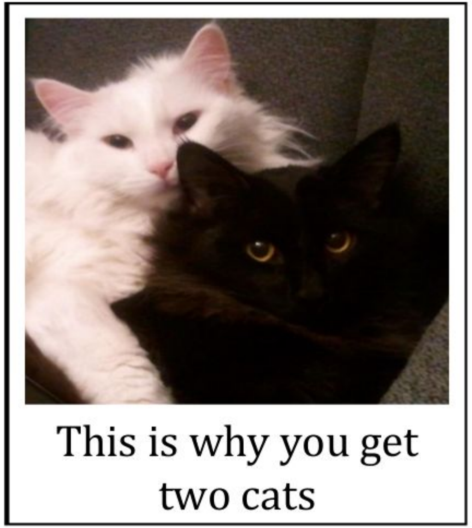
A Case Study on Combining ASR and Visual Features
for Generating Instructional Video Captions

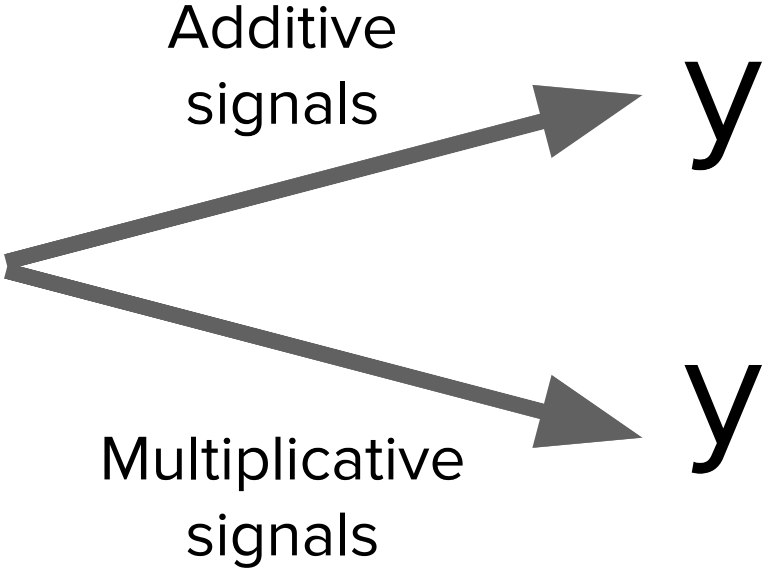
[CoNLL 2019 H., Pang, Zhu, Soricut;
H., Pang, Zhu are planning ACL submission :)]

Ongoing work: incorporating structure into multi-retrieval models



Ongoing work: decoupling additive vs. multiplicative interactions





Thanks to my awesome collaborators!

Lillian Lee

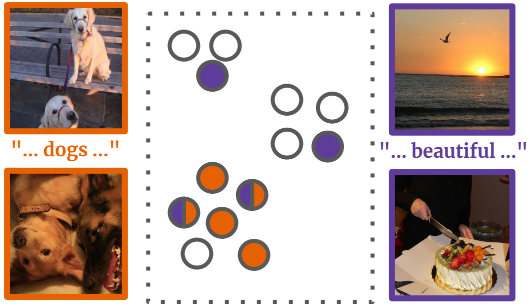
David Mimno

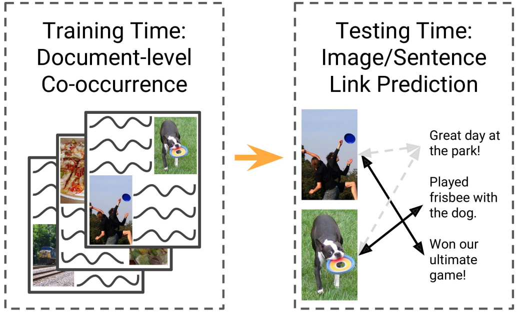
Bo Pang

Zhenhai Zhu

Radu Soricut

And thanks to you for having me!!





Contact:
jmhessel@gmail.com
@jmhessel on Twitter

Code, data, and papers are all available:
http://www.cs.cornell.edu/~jhessel/