The Promise and the PERILS

of Learning Grounding from Visual-Textual Web data

Jack Hessel
Cornell University

What is visual-textual grounding?

What is visual-textual grounding?

A collection of tasks requiring connection between visual and textual content.

What is visual-textual grounding?

A collection of tasks requiring connection between visual and textual content.

Alt-text Generation

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

What is visual-textual grounding?

A collection of tasks requiring connection between visual and textual content.

Alt-text Generation

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.

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

Human-Robot Interaction

"Here are the yellow ones"

[Matuszek et al. 2012]

What is visual-textual grounding?

A collection of tasks requiring connection between visual and textual content.

<u>Alt-text Generation</u>

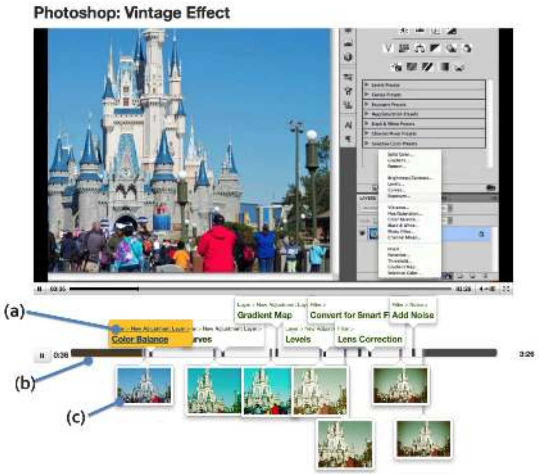
[Wu et al. 2017;
Sharma et al. 2019]

<u>Human-Robot Interaction</u>

"Here are the yellow ones"

[Matuszek et al. 2012]

<u>Web Video Parsing</u>



[Kim et al. 2014]

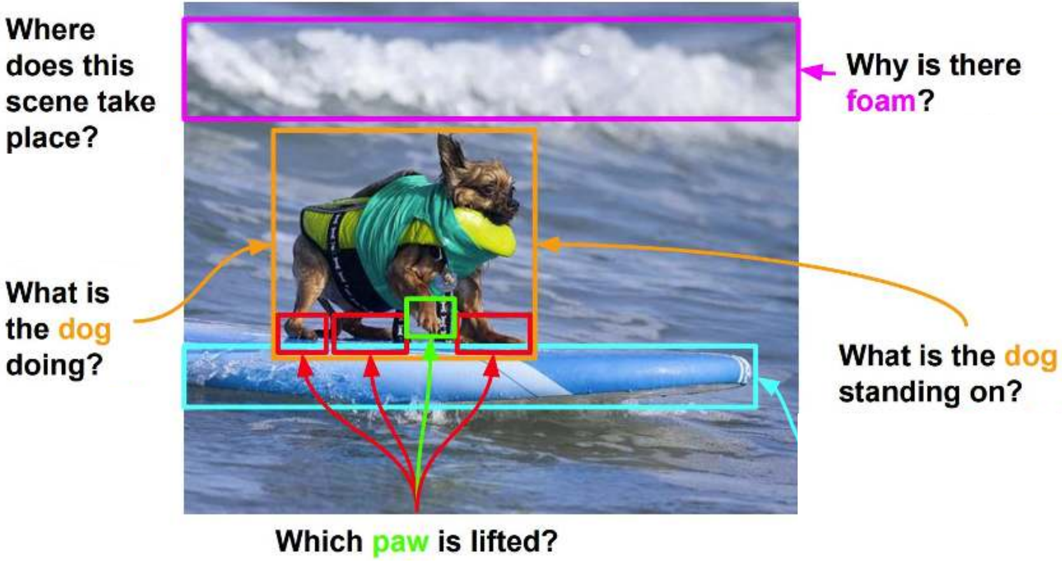
Why study visual-textual grounding?

Cross-modal reasoning is easy for humans, hard for computers

[Zhu et al. 2016;
Photo by Nathan Rupert]

Why study visual-textual grounding?

Cross-modal reasoning is easy for humans, hard for computers



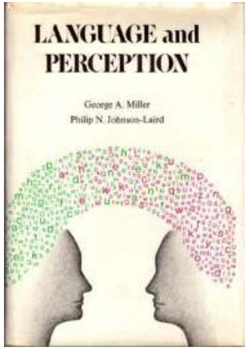
[Zhu et al. 2016;
Photo by Nathan Rupert]

Why study visual-textual grounding?

Cross-modal reasoning is important beyond AI

Cognitive psychology work since at least the 1970s.

[Miller and Johnson-Laird 1976]



"Symbol Grounding Problem"

[Harnad 1990]

"How are those symbols (e.g., the words in our heads) connected to the things they refer to?"

Why study multimodal *web data*?

Why study multimodal ***web data***?

Noisy web data is unreasonably effective

Why study multimodal *web data*?

Noisy web data is unreasonably effective

Web data is

"the best ally we have"

--- Halevy, Norvig, and Pereira, 2009

Why study multimodal *web data*?

Noisy web data is unreasonably effective

Why study multimodal *web data*?

Noisy web data is unreasonably effective

Unimodal Tasks

IMAGENET

GLUE

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

Image+Text Tasks

NLVR2

GQA

VQA

Flickr1K

[Goyal et al. 2017; Suhr et al. 2018; Hudson and Manning, 2019; Young et al. 2014]

Video+Text Tasks

Crosstask

YouCook2

[Zhukov et al. 2019; Zhou et al. 2018]

Why study multimodal *web data*?

Noisy web data is unreasonably effective

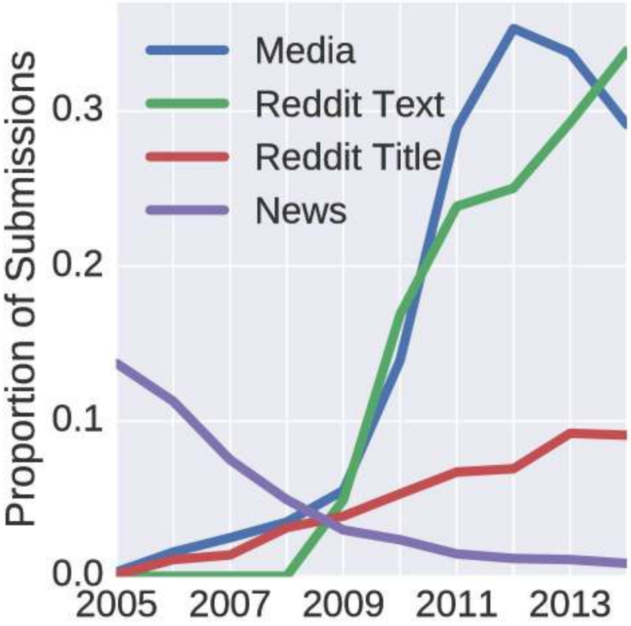
Unimodal Tasks	Image+Text Tasks	Video+Text Tasks
IMAGENET, GLUE	NLVR2, GQA, VQA, Flickr1K	Crosstask, YouCook2
[Deng et al. 2009; Wang et al. 2019]	[Goyal et al. 2017; Suhr et al. 2018; Hudson and Manning, 2019; Young et al. 2014]	[Zhukov et al. 2019; Zhou et al. 2018]
3.5B Tagged Instagram Images 34B Web Tokens	3M Webly Supervised Image-Caption Pairs	100M Web Video Clips + ASR
Common Crawl	Conceptual Captions	HowTo100M
[Mahajan et al. 2018; Raffel et al. 2019]	[Sharma et al. 2018]	[Miech et al. 2019]

Why study multimodal *web data*?

Important for understanding web communication

Why study multimodal *web data*?

Important for understanding web communication

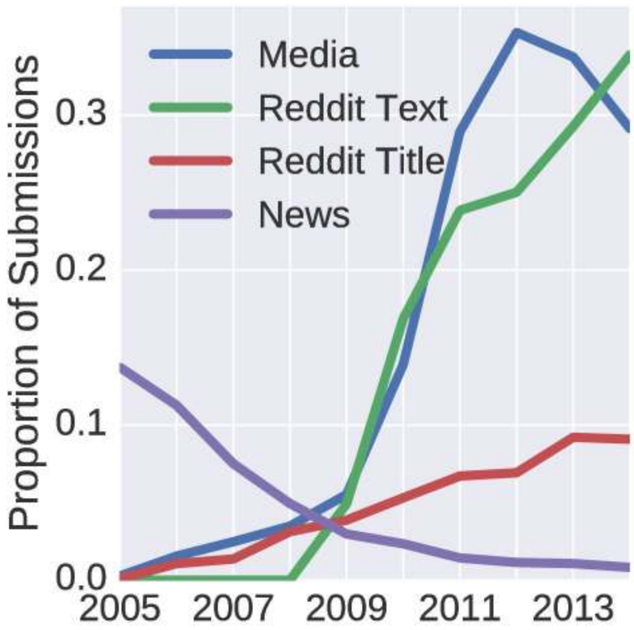


imgur.com,
youtube...

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

Why study multimodal *web data*?

Important for understanding web communication



imgur.com, youtube...

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

Semioticians have long argued multimodality is a fundamental part of communication

"The power of visual communication is multiplied when it is co-deployed with language in multimodal texts."
[Lemke 2002]

My Research Goals:

My Research Goals:

build better
grounding algorithms

understand web
communication

My Research Goals:

build better
grounding algorithms

need for
cross-modal
reasoning,
real-world
knowledge,
etc.

requires

understand web
communication

My Research Goals:

need for
cross-modal
reasoning,
real-world
knowledge,
etc.

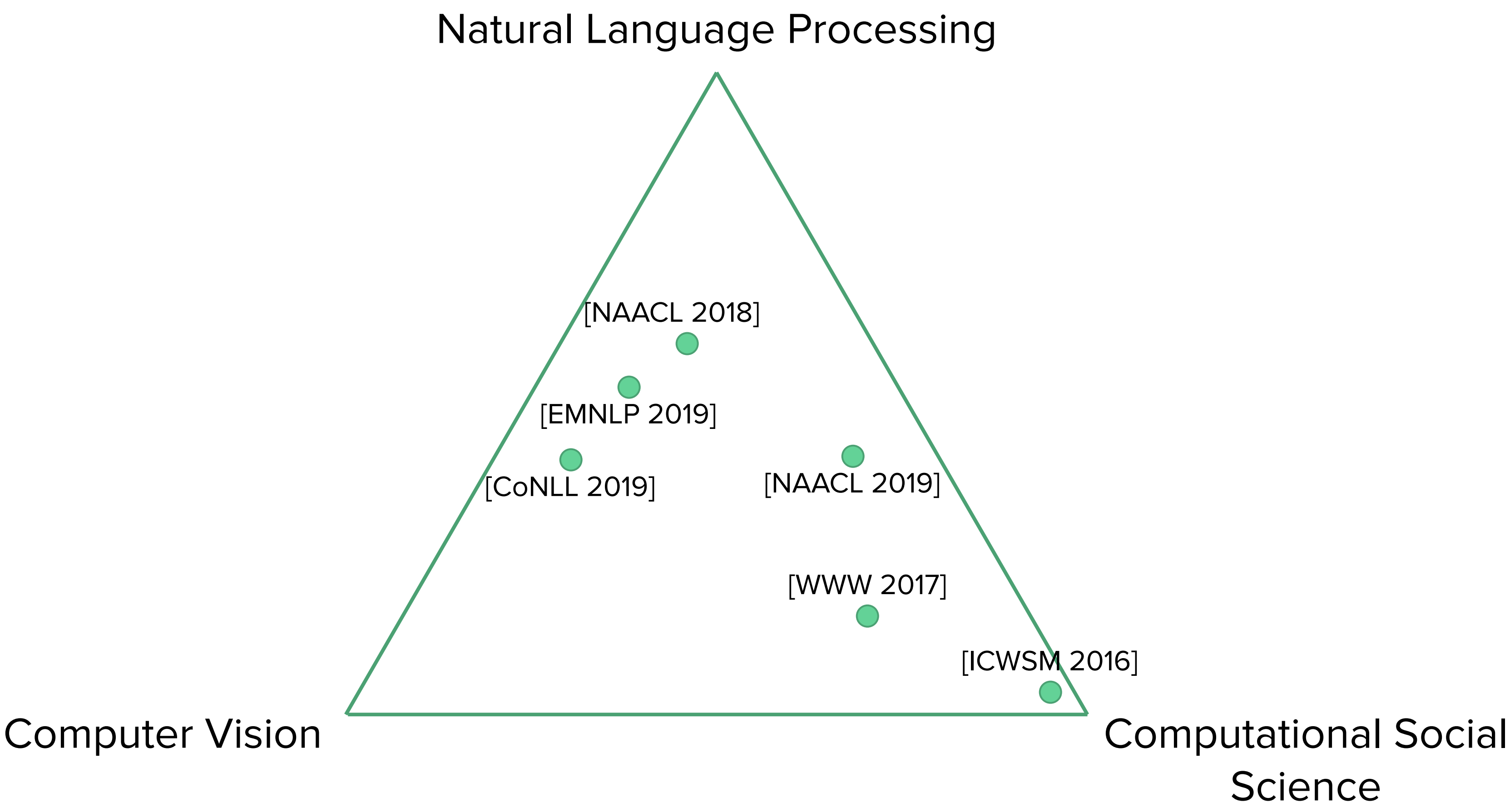
requires

build better
grounding algorithms

understand web
communication

requires

design
more effective
unsupervised
training
objectives
for web data


Natural Language Processing
[NAACL 2018]
[EMNLP 2019]
[CoNLL 2019]
[NAACL 2019]
[WWW 2017]
[ICWSM 2016]
Computer Vision
Computational Social Science

The Promise and the PERILS

The Promise and the PERILS

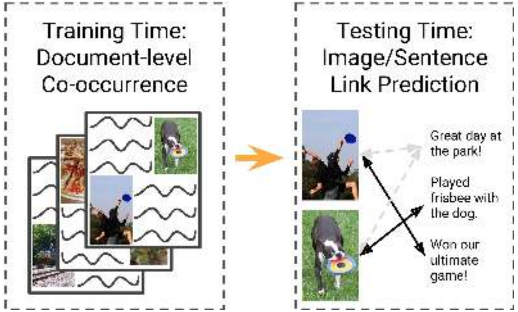
We can *do cool things* with multimodal webdata,

but web texts are *not literal image descriptions*

(even though most algorithms treat them that way)

The Promise and the PERILS

The Promise and the PERILS



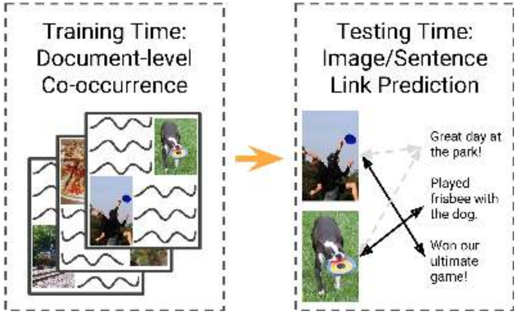
Learning from multi-sentence, multi-image web documents

[EMNLP 2019: Hessel, Lee, Mimno]

Learning from unlabelled web videos + ASR

[CoNLL 2019: Hessel, Pang, Zhu, Soricut; In Sub: Hessel, Zhu, Pang, Soricut]

The Promise and the PERILS



Learning from multi-sentence, multi-image web documents

[EMNLP 2019: Hessel, Lee, Mimno]

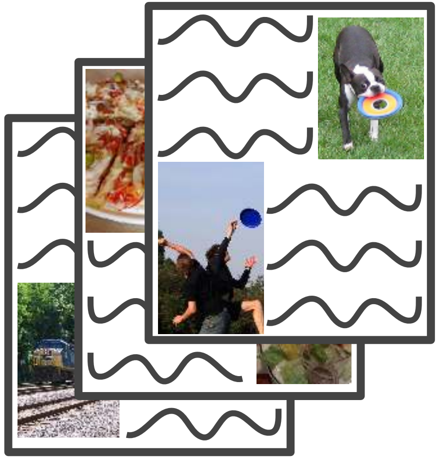
Learning from unlabelled web videos + ASR

[CoNLL 2019: Hessel, Pang, Zhu, Soricut;
In Sub: Hessel, Zhu, Pang, Soricut]

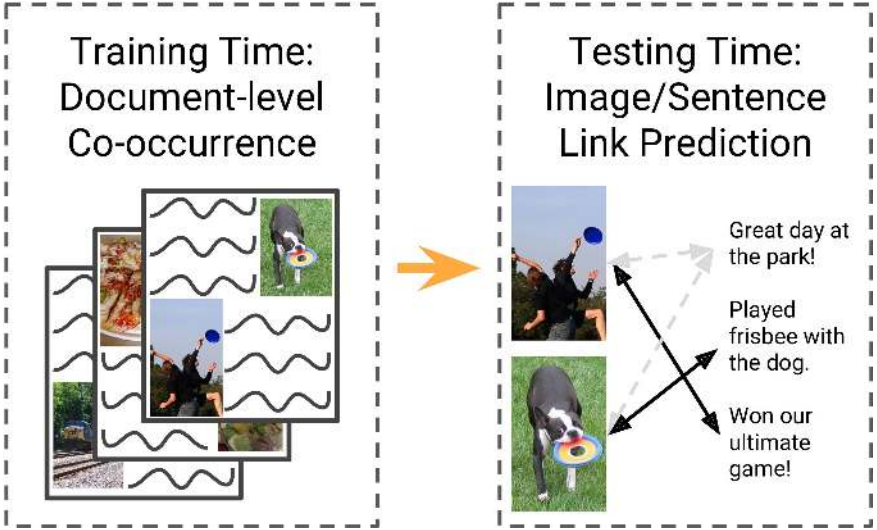
Multi-image, Multi-sentence documents?

Image captioning case:
one image, one sentence
explicit link by annotation

Our case:
multiple images, multiple sentences
no explicit links

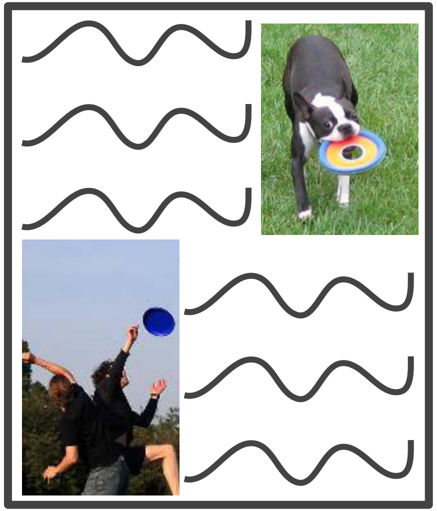


The Task: Unsupervised Link Prediction

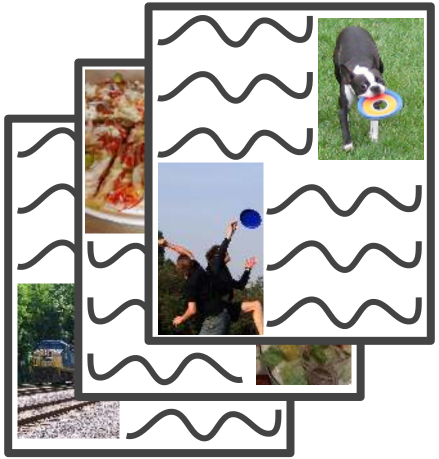


What's hard about this link prediction task?

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

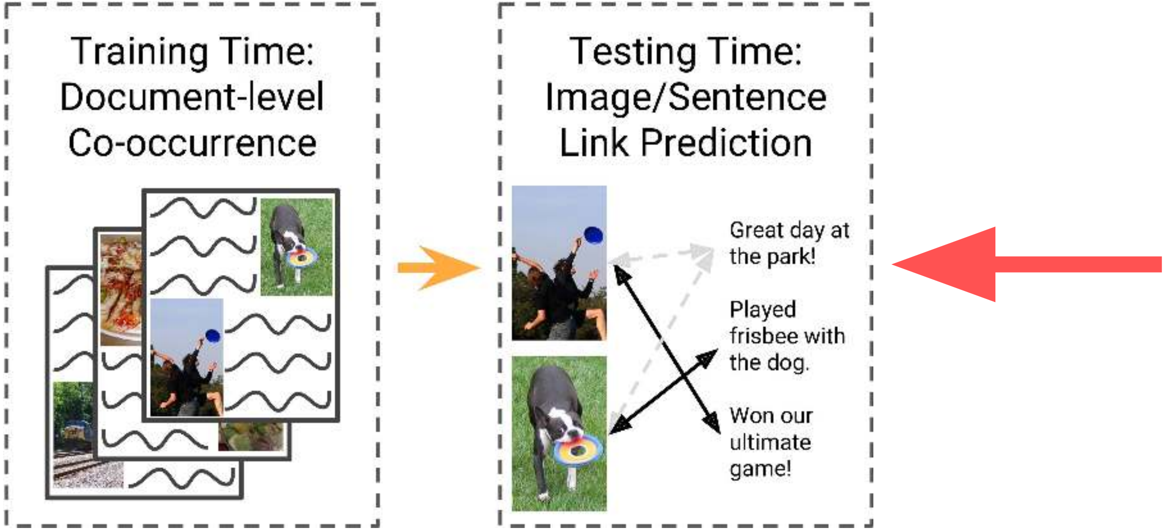


Multi-image/multi-sentence pretraining framework:



Web pages, product listings, books (current and historical), web comments on images, news articles...

The Task: Unsupervised Link Prediction



Why you might care about same document retrieval:

Why you might care about same document retrieval:

Microsoft

CaptionBot

"I think it's a group of people riding on the back of a boat."

Why you might care about same document retrieval:

Microsoft

CaptionBot

"I think it's a group of people riding on the back of a boat."

WIKIPEDIA

The Free Encyclopedia

"General Washington is emphasized by an unnaturally bright sky, while his face catches the upcoming sun. The colors consist of mostly dark tones, as is to be expected at dawn, but there are red highlights."

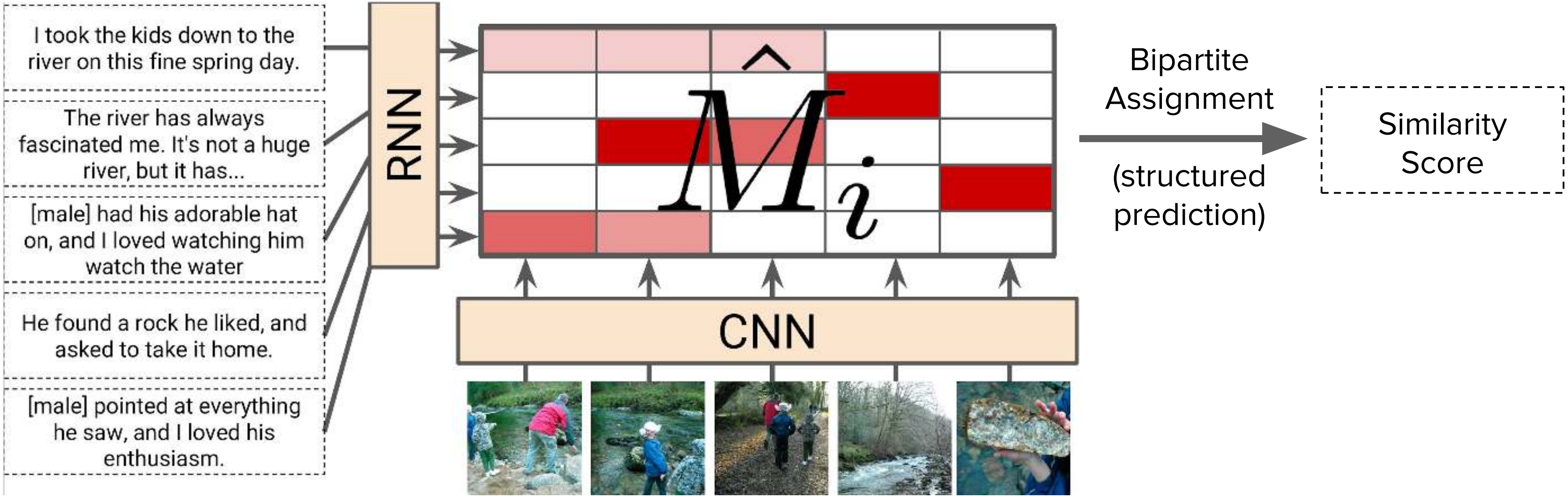
Stats for Web Datasets

	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

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
Bipartite Assignment
(structured prediction)
Similarity Score

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 ri

The river has always

[male] had his adorable hat

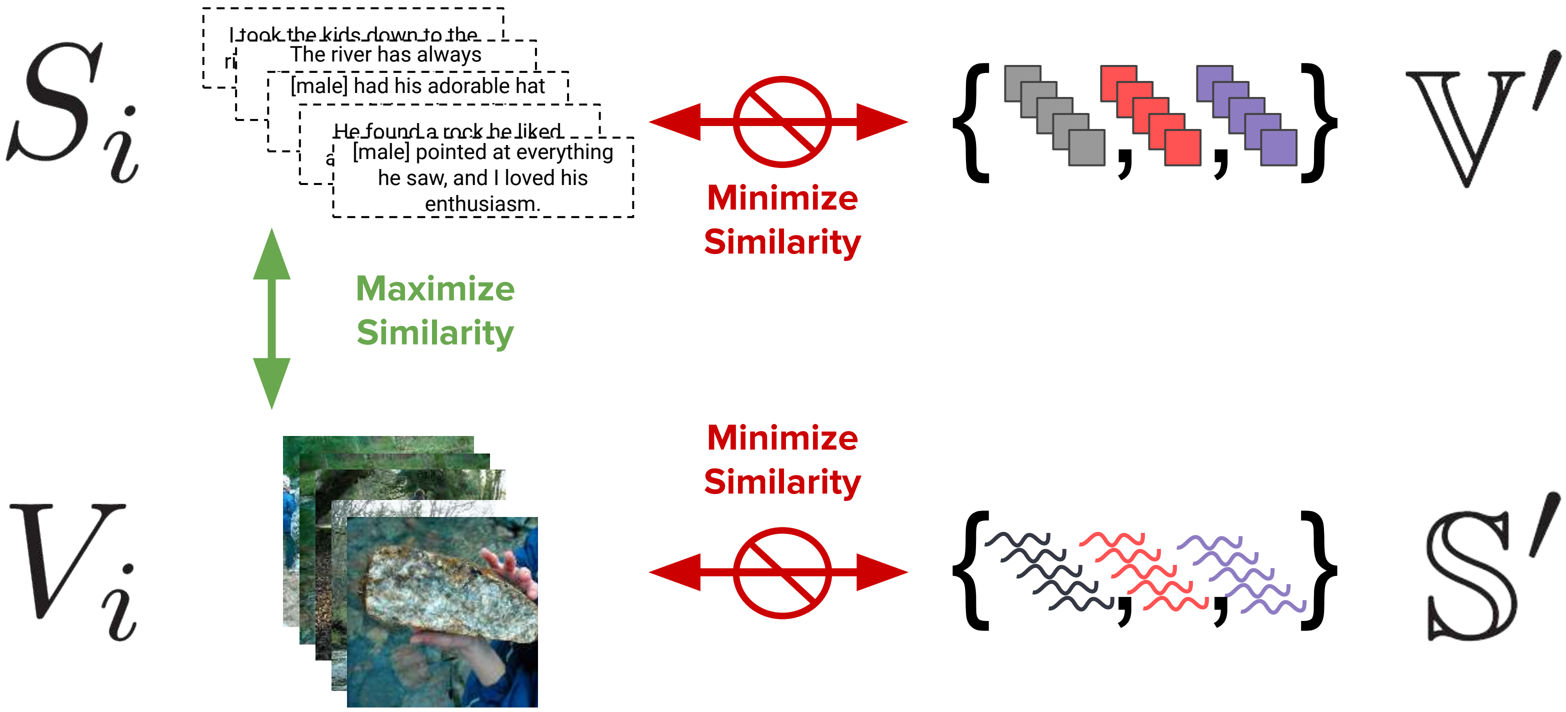
He found a rock he liked a

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

Maximize Similarity

V_i

Training: Max Margin loss with Negative Sampling



Quantitative Results

we have labels that are only used at test-time for evaluation for these datasets

(Higher = better)

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

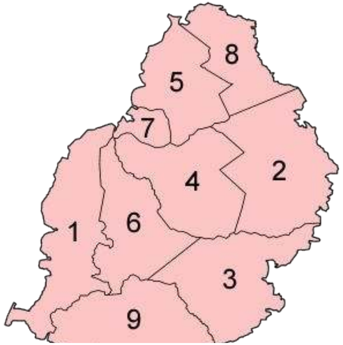
WIKI Prediction on 100-sentence Mauritius Article

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

WIKI Prediction on 100-sentence 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 an object detection baseline's selected sentence began with:
“(Mauritian Creole people usually known as ‘Creoles’)”

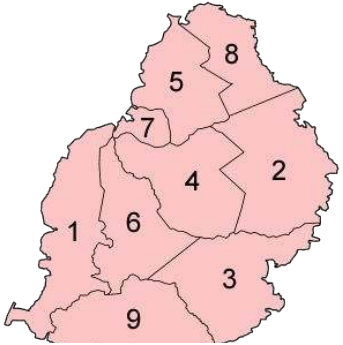
WIKI Prediction on 100-sentence Mauritius Article

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)

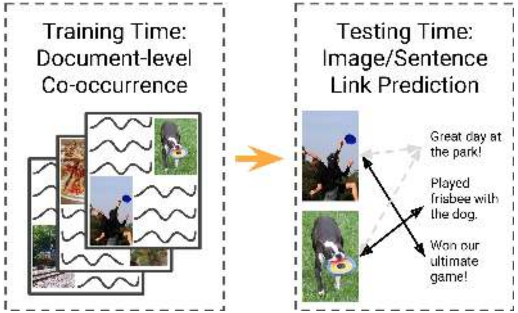
... a significant migrant population of Bhumihar Brahmins in Mauritius who have made a mark for themselves in different fields.
(79.8)



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

The Promise and the PERILS



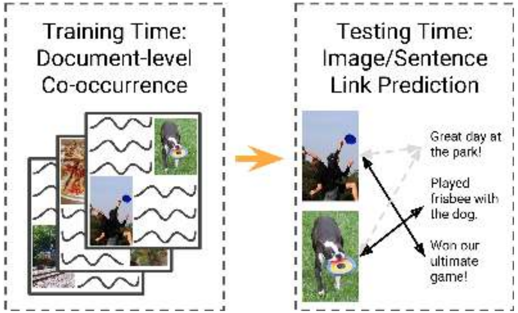
Learning from multi-sentence, multi-image web documents

[EMNLP 2019: Hessel, Lee, Mimno]

Learning from unlabelled web videos + ASR

[CoNLL 2019: Hessel, Pang, Zhu, Soricut;
In Sub: Hessel, Zhu, Pang, Soricut]

The Promise and the PERILS



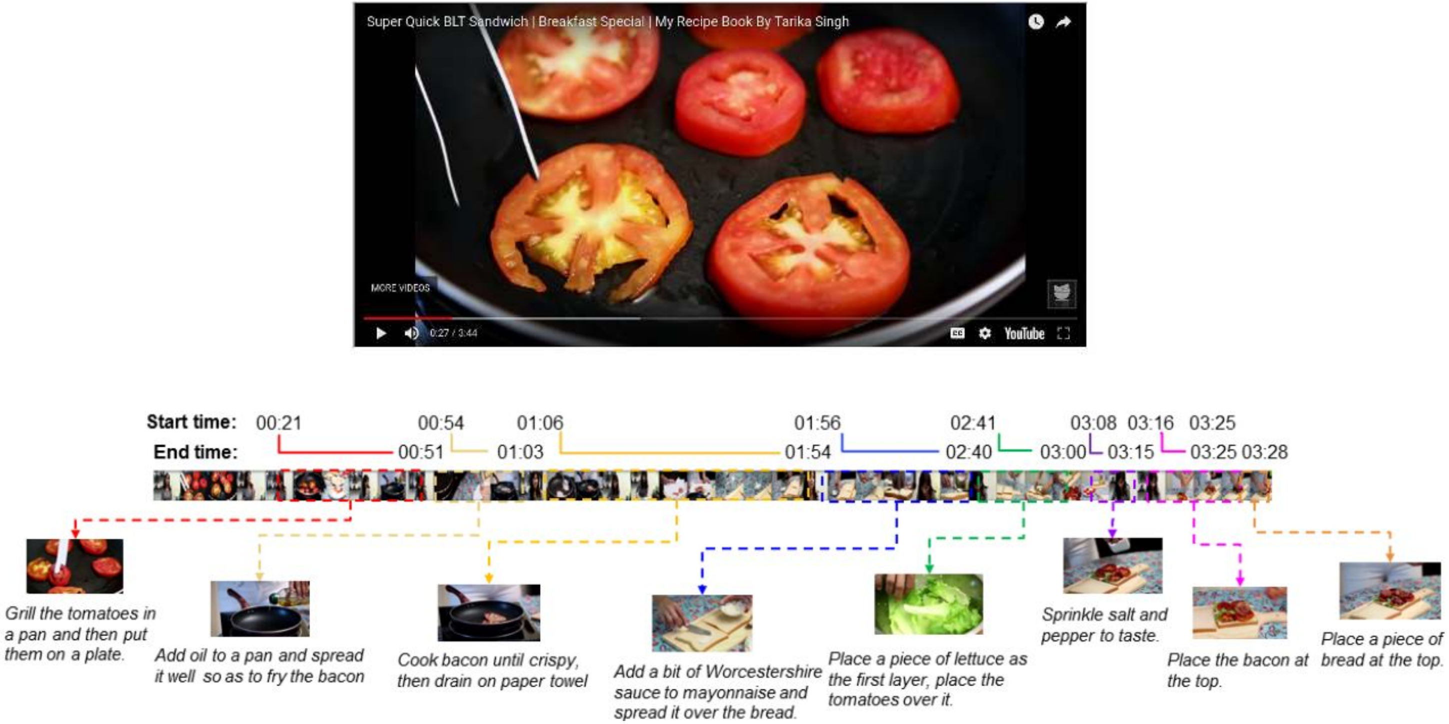
Learning from multi-sentence, multi-image web documents

[EMNLP 2019: Hessel, Lee, Mimno]

Learning from unlabelled web videos + ASR

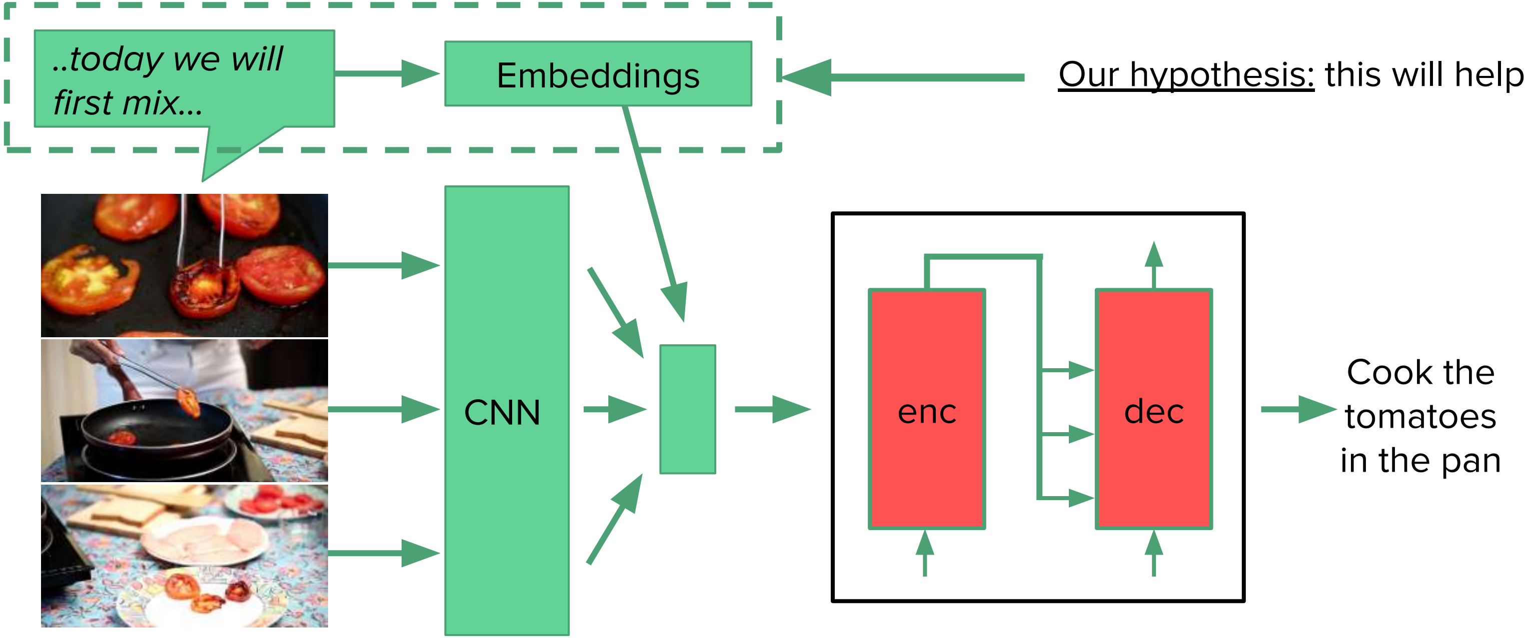
[CoNLL 2019: Hessel, Pang, Zhu, Soricut;
In Sub: Hessel, Zhu, Pang, Soricut]

Noisy ASR for Video Captioning



[Zhou et al. 2018]

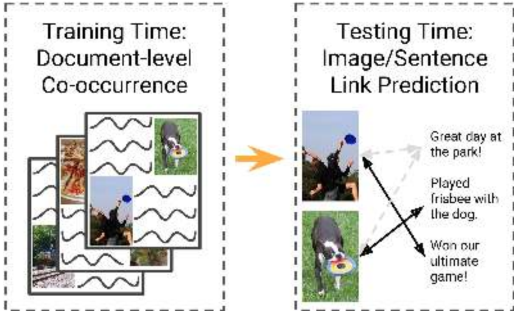
Noisy ASR for Video Captioning



Noisy ASR for Video Captioning

	BLEU-4	METEOR	ROUGE-L	CIDEr
Prev. SoTA (video only)	4.31	11.91	29.47	0.53
Noisy ASR (text only)	8.55	16.93	35.54	1.06
Video+ASR (multimodal)	9.01	17.77	**36.65**	**1.12**

The Promise and the PERILS



Learning from multi-sentence, multi-image web documents

[EMNLP 2019: Hessel, Lee, Mimno]

Learning from unlabelled web videos + ASR

[CoNLL 2019: Hessel, Pang, Zhu, Soricut; In Sub: Hessel, Zhu, Pang, Soricut]

The Promise and the PERILS

Many datasets/algorithms 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

Image-text relationships on the web

Q: "How does an illustration relate to the text with which it is associated, or, what are the functions of illustration?"

[Marsh and Domas White, 2003]

Image-text relationships on the web

Q: "How does an illustration relate to the text with which it is associated, or, what are the functions of illustration?"

A: It depends!

A Functions expressing little relation to the text	B Functions expressing close relation to the text	C Functions that go beyond the text
A1 Decorate	*B1 Reiterate*	*C1 Interpret*
A1.1 Change pace	B1.1 Concretize	C1.1 Emphasize
A1.2 Match style	B1.1.1 Sample	C1.2 Document
A2 Elicit emotion	B1.1.1.1 Author/Source	*C2 Develop*
A2.1 Alienate	B1.2 Humanize	C2.1 Compare
A2.2 Express poetically	B1.3 Common referent	C2.2 Contrast
A3 Control	B1.4 Describe	*C3 Transform*
A3.1 Engage	B1.5 Graph	C3.1 Alternate progress
A3.2 Motivate	B1.6 Exemplify	C3.2 Model
	B1.7 Translate	C3.2.1 Model cognitive process
	B2 Organize	C3.2.2 Model physical process
	B2.1 Isolate	C3.3 Inspire
	B2.2 Contain	
	B2.3 Locate	
	B2.4 Induce perspective	
	B3 Relate	
	B3.1 Compare	
	B3.2 Contrast	
	B3.3 Parallel	
	B4 Condense	
	B4.1 Concentrate	
	B4.2 Compact	
	B5 Explain	
	B5.1 Define	
	B5.2 Complement	

Table II.
Taxonomy of functions of images to the text

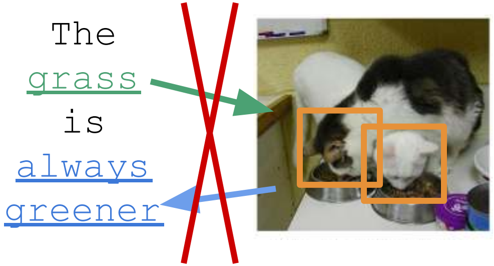
[Marsh and Domas White, 2003]

The Promise and the PERILS

The Promise and the PERILS

What concepts are "groundable"?

[NAACL 2018, Hessel, Mimno, Lee]



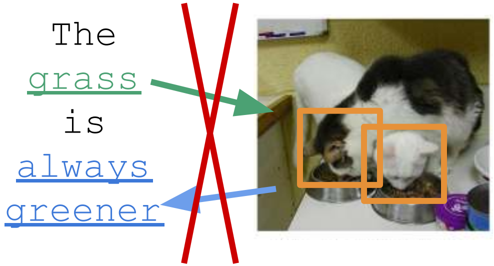
Does my model learn cross-modal interactions?

[In Sub to EMNLP 2020: Hessel, Lee; WWW 2017, Hessel, Lee, Mimno]

The Promise and the PERILS

What concepts are "groundable"?

[NAACL 2018, Hessel, Mimno, Lee]



Does my model learn cross-modal interactions?

[In Sub to EMNLP 2020: Hessel, Lee; WWW 2017, Hessel, 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**."*

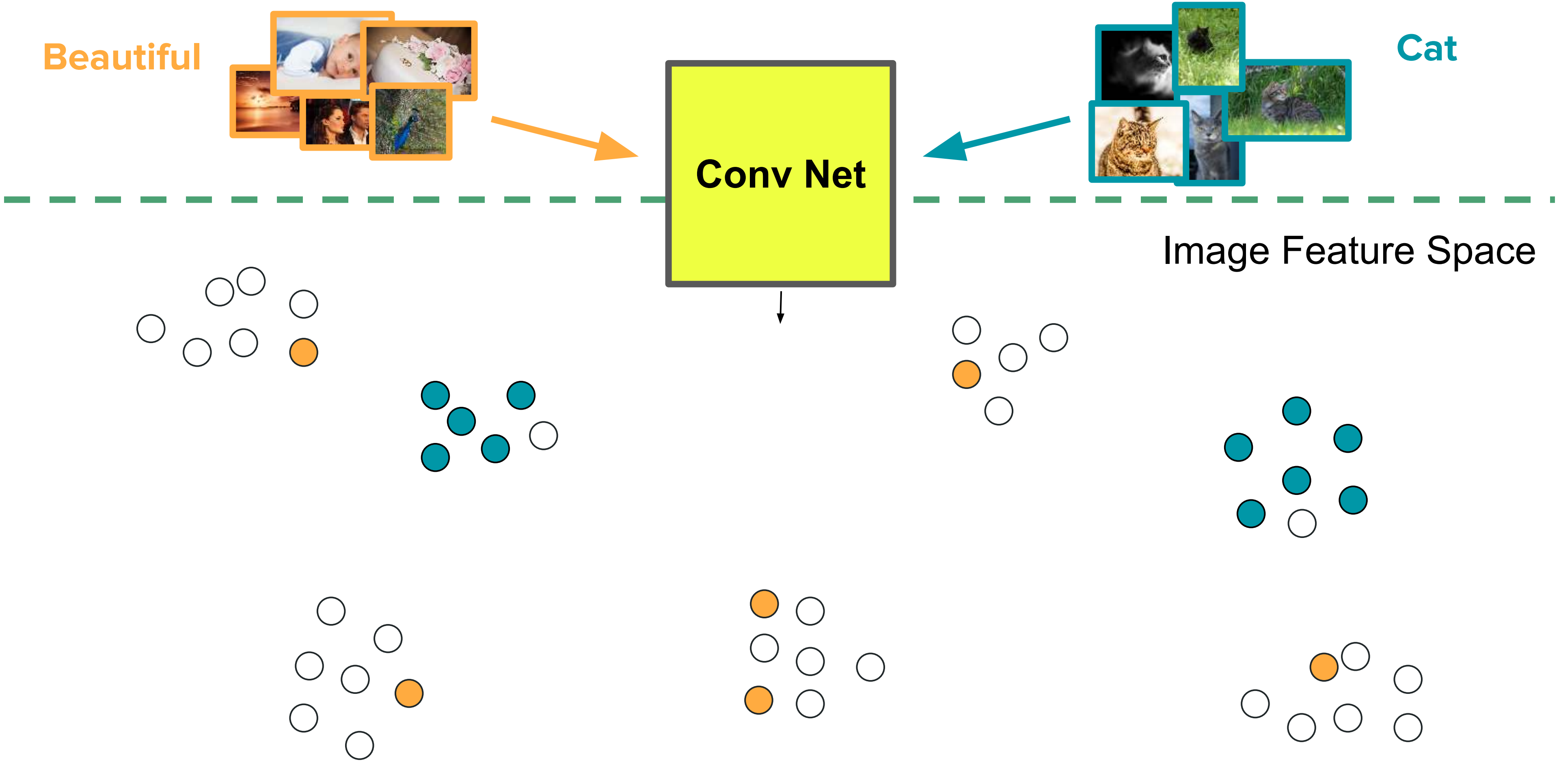
[Hill and Korhonen 2014]

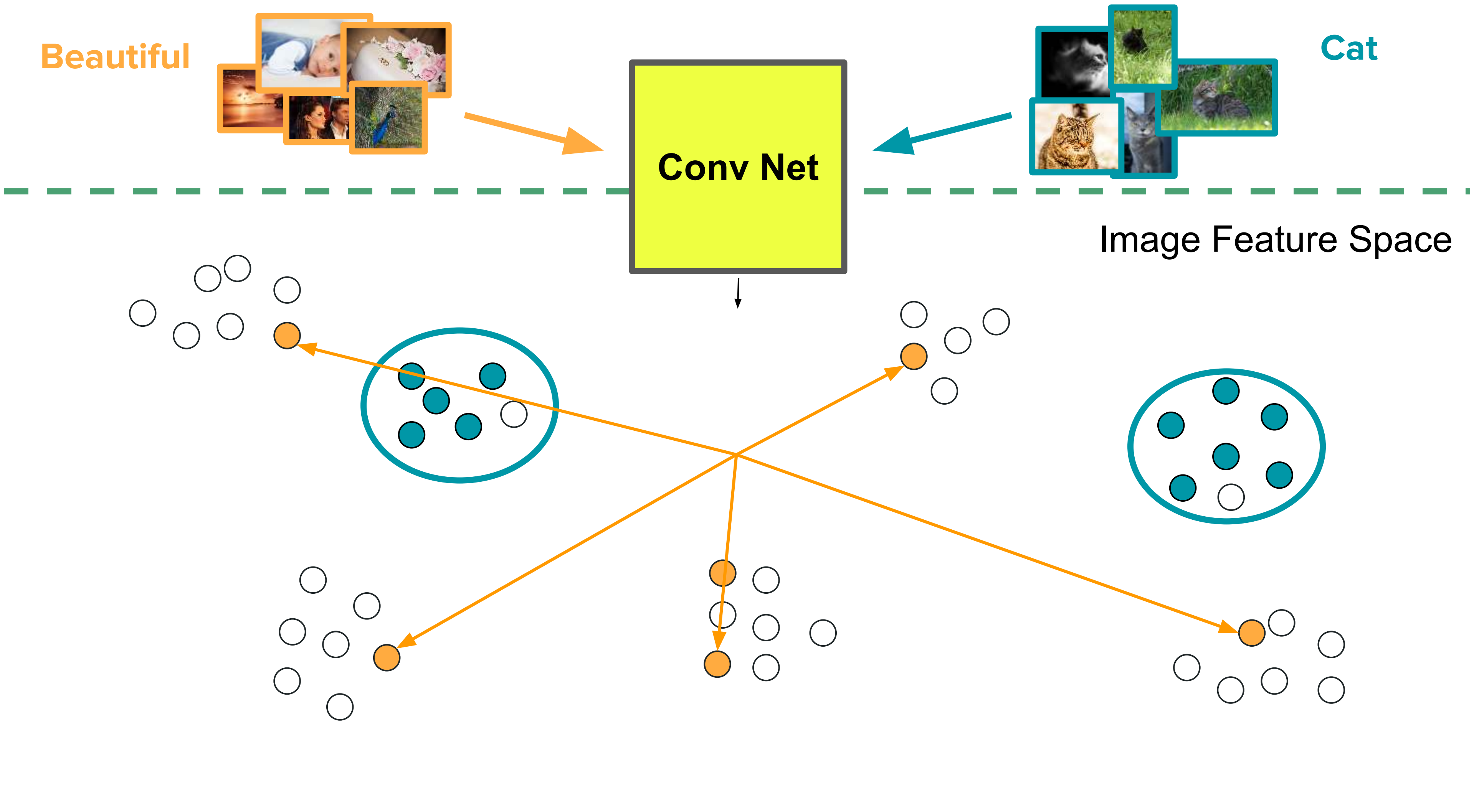
The **cat** is in the grass.

This **cat** is enjoying the sun.

This is a **beautiful** baby.

The sunset is **beautiful**.


Beautiful
Cat
Conv Net
Image Feature Space



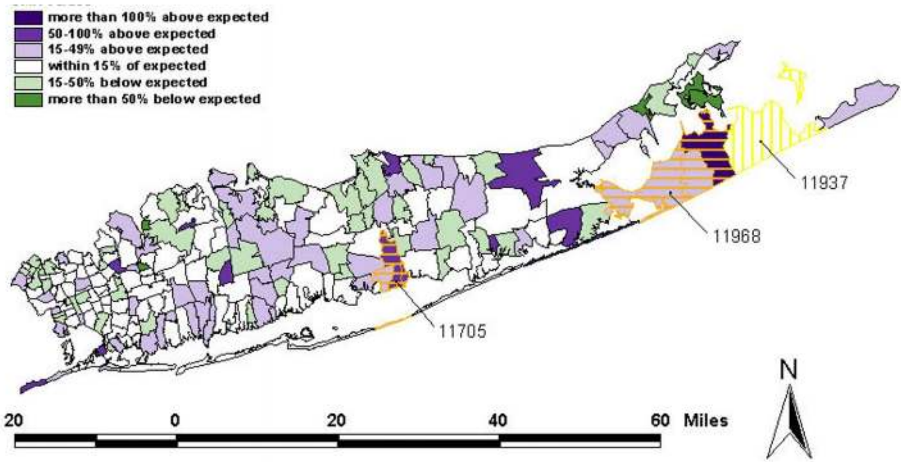
Beautiful
Cat
Conv Net
Image Feature Space

Connection to Geospatial Statistics

Local Indicators of Spatial Association—LISA

The capabilities for visualization, rapid data retrieval, and manipulation i graphic information systems (GIS) have created the need for new techniq exploratory data analysis that focus on the "spatial" aspects of the dat identification of local patterns of spatial association is an important conc

[Anselin 1995]



[Jacquez and Greiling 2003]

COCO Results

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

COCO Results

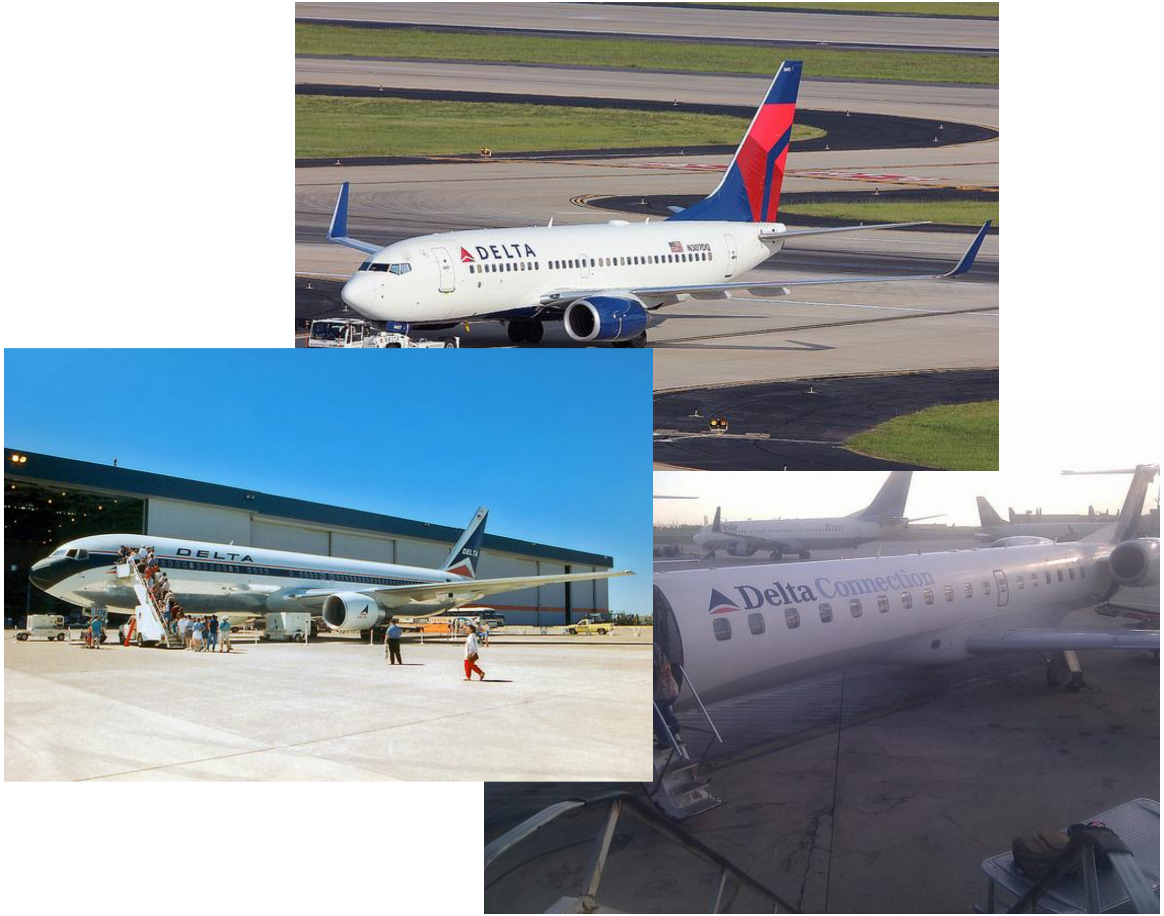
Most concrete

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

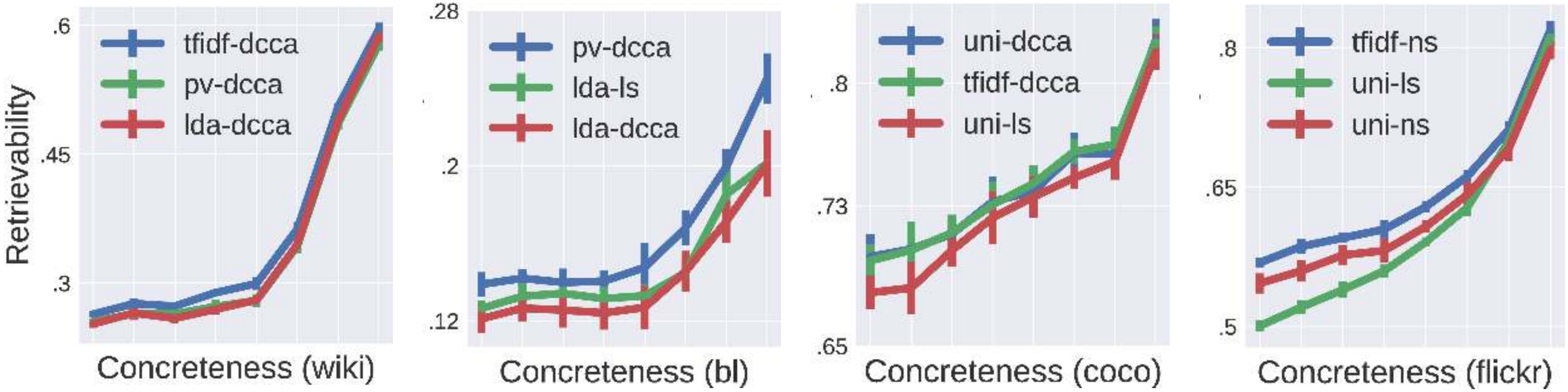
Most concrete

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

More concrete = easier to learn



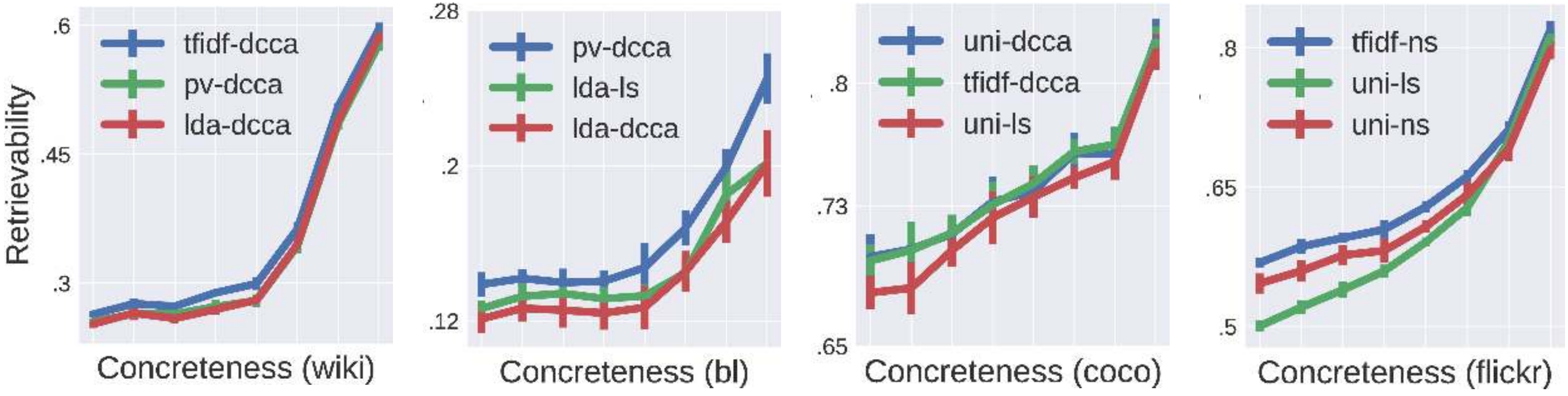
Bad news: success of retrieval objective largely determined by original feature geometry

Context matters!

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

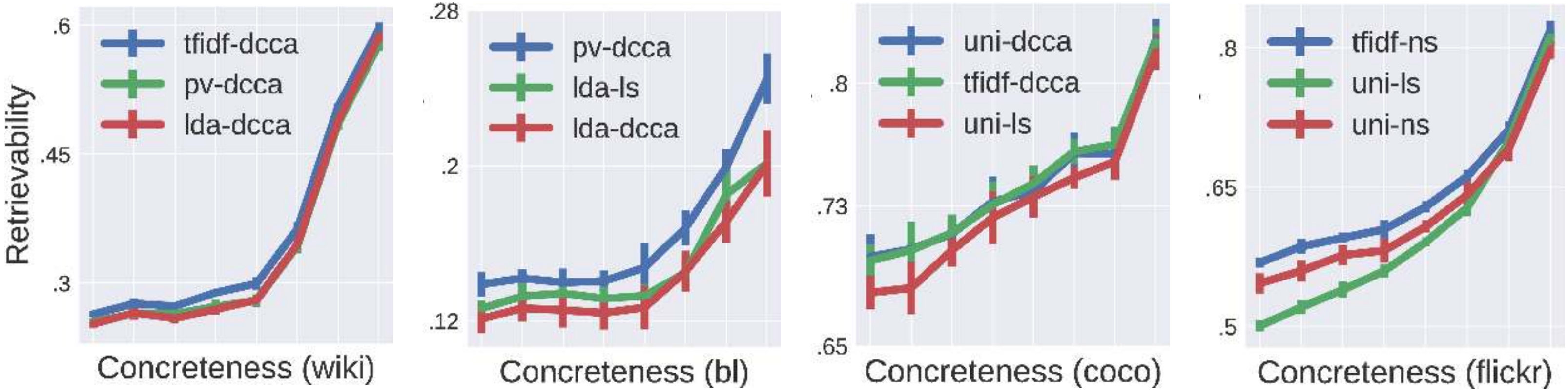
"#London"
Rank 1110/7K Concreteness
as a hashtag in a Flickr image
tagging dataset.

More concrete = easier to learn



Bad news: success of retrieval objective largely determined by original feature geometry

More concrete = easier to learn



Bad news: success of retrieval objective largely determined by original feature geometry

Open question: what are the limits of retrieval-style algorithms at scale?

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

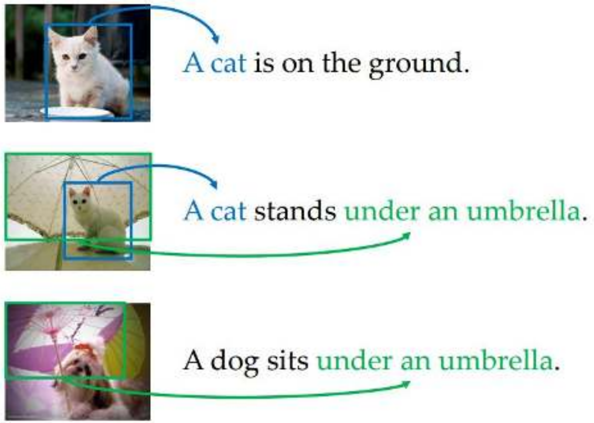
Least Concrete

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**

Use Case of Our Algorithm from Shi et al. 2019

(ACL Best Paper Nom.)

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



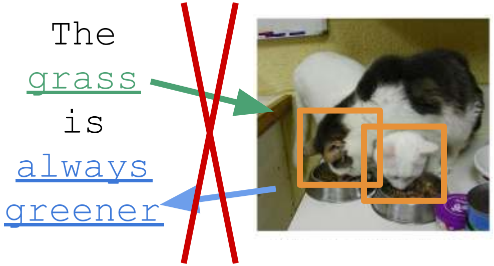
Model	NP	VP	PP	ADJP	Avg. F_1	Self F_1
Random	$47.3_{\pm 0.3}$	$10.5_{\pm 0.4}$	$17.3_{\pm 0.7}$	$33.5_{\pm 0.8}$	$27.1_{\pm 0.2}$	32.4
Left	51.4	1.8	0.2	16.0	23.3	N/A
Right	32.2	23.4	18.7	14.4	22.9	N/A
VG-NSL (ours)†	$\mathbf{79.6}_{\pm 0.4}$	$26.2_{\pm 0.4}$	$42.0_{\pm 0.6}$	$22.0_{\pm 0.4}$	$50.4_{\pm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

The Promise and the PERILS

What concepts are "groundable"?

[NAACL 2018, Hessel, Mimno, Lee]



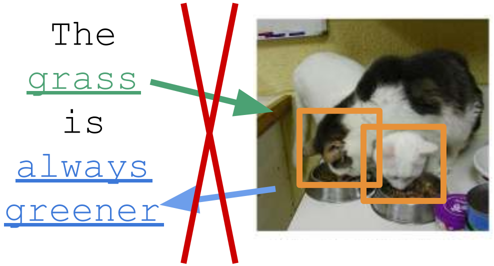
Does my model learn cross-modal interactions?

[In Sub to EMNLP 2020: Hessel, Lee; WWW 2017, Hessel, Lee, Mimno]

The Promise and the PERILS

What concepts are "groundable"?

[NAACL 2018, Hessel, Mimno, Lee]



Does my model learn cross-modal interactions?

[In Sub to EMNLP 2020: Hessel, Lee; WWW 2017, Hessel, Lee, Mimno]

Image-text relationships on the web

Q: "How does an illustration relate to the text with which it is associated, or, what are the functions of illustration?"

A: It depends!

A Functions expressing little relation to the text	B Functions expressing close relation to the text	C Functions that go beyond the text
A1 Decorate	*B1 Reiterate*	*C1 Interpret*
A1.1 Change pace	B1.1 Concretize	C1.1 Emphasize
A1.2 Match style	B1.1.1 Sample	C1.2 Document
A2 Elicit emotion	B1.1.1.1 Author/Source	*C2 Develop*
A2.1 Alienate	B1.2 Humanize	C2.1 Compare
A2.2 Express poetically	B1.3 Common referent	C2.2 Contrast
A3 Control	B1.4 Describe	*C3 Transform*
A3.1 Engage	B1.5 Graph	C3.1 Alternate progress
A3.2 Motivate	B1.6 Exemplify	C3.2 Model
	B1.7 Translate	C3.2.1 Model cognitive process
	B2 Organize	C3.2.2 Model physical process
	B2.1 Isolate	C3.3 Inspire
	B2.2 Contain	
	B2.3 Locate	
	B2.4 Induce perspective	
	B3 Relate	
	B3.1 Compare	
	B3.2 Contrast	
	B3.3 Parallel	
	B4 Condense	
	B4.1 Concentrate	
	B4.2 Compact	
	B5 Explain	
	B5.1 Define	
	B5.2 Complement	

Table II.
Taxonomy of functions of images to the text

[Marsh and Domas White, 2003]

Increasing number of multimodal, in-vivo studies

Proposing work	Task (structure)	Abbv.	# image+text
Kruk et al. (2019)	Instagram		
	↳ intent (7-way clf)	I-INT	1299
	↳ semiotic (7-way clf)	I-SEM	1299
	↳ contextual (7-way clf)	I-CTX	1299
Vempala and Preoţiuc-Pietro (2019)	Twitter visual-ness (4-way clf)	T-VIS	4471
Hessel et al. (2017)	Reddit popularity (Pairwise-ranking)	R-POP	88K
Borth et al. (2013)	Twitter sentiment (binary clf)	T-ST1	603
Niu et al. (2016)	Twitter sentiment (binary clf)	T-ST2	4511

Increasing number of multimodal, in-vivo studies

Proposing work	Task (structure)	Abbv.	# image+text
Kruk et al. (2019)	Instagram		
	↳ intent (7-way clf)	I-INT	1299
	↳ semiotic (7-way clf)	I-SEM	1299
	↳ contextual (7-way clf)	I-CTX	1299
Vempala and Preoţiuc-Pietro (2019)	Twitter visual-ness (4-way clf)	T-VIS	4471
Hessel et al. (2017)	Reddit popularity (Pairwise-ranking)	R-POP	88K
Borth et al. (2013)	Twitter sentiment (binary clf)	T-ST1	603
Niu et al. (2016)	Twitter sentiment (binary clf)	T-ST2	4511

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

"Snacks!" - /r/aww

	# 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

Our task:

popularity ranking

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

"Glamor Leaves" - /r/RedditLaquersitas

The grass is always greener

This is why you get two cats

The grass is always greener

This is why you get two cats

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

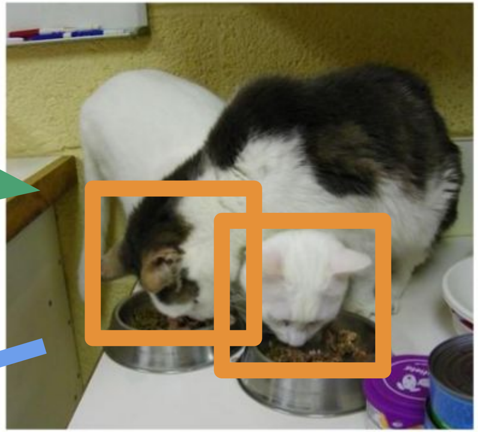
The

grass

is

always

greener



Visual-textual interactions: "meaning multiplication"

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.

--- Bateman, 2014
describing "Meaning Multiplication"
[Barthes 1988; Jones 1979]

Prediction Results

Best unimodal

(image only)

	aww	pics	cats	MA	FP	RL
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

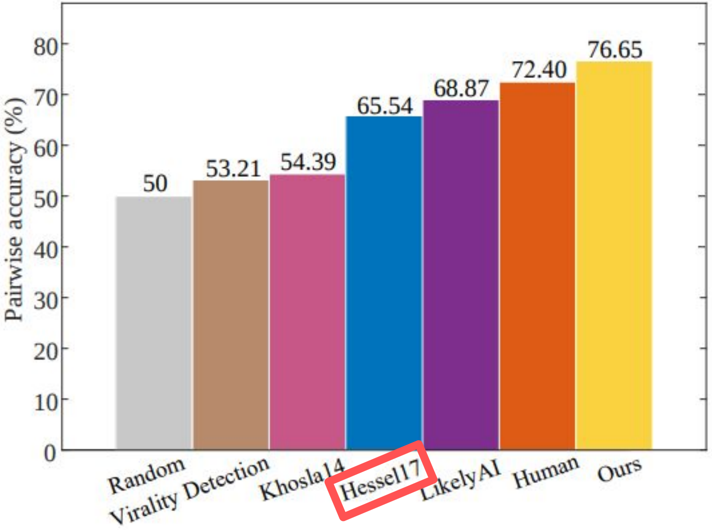
Multimodal
beats unimodal!

Prediction Results

Best unimodal
(image only)

	aww	pics	cats	MA	FP	RL
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!

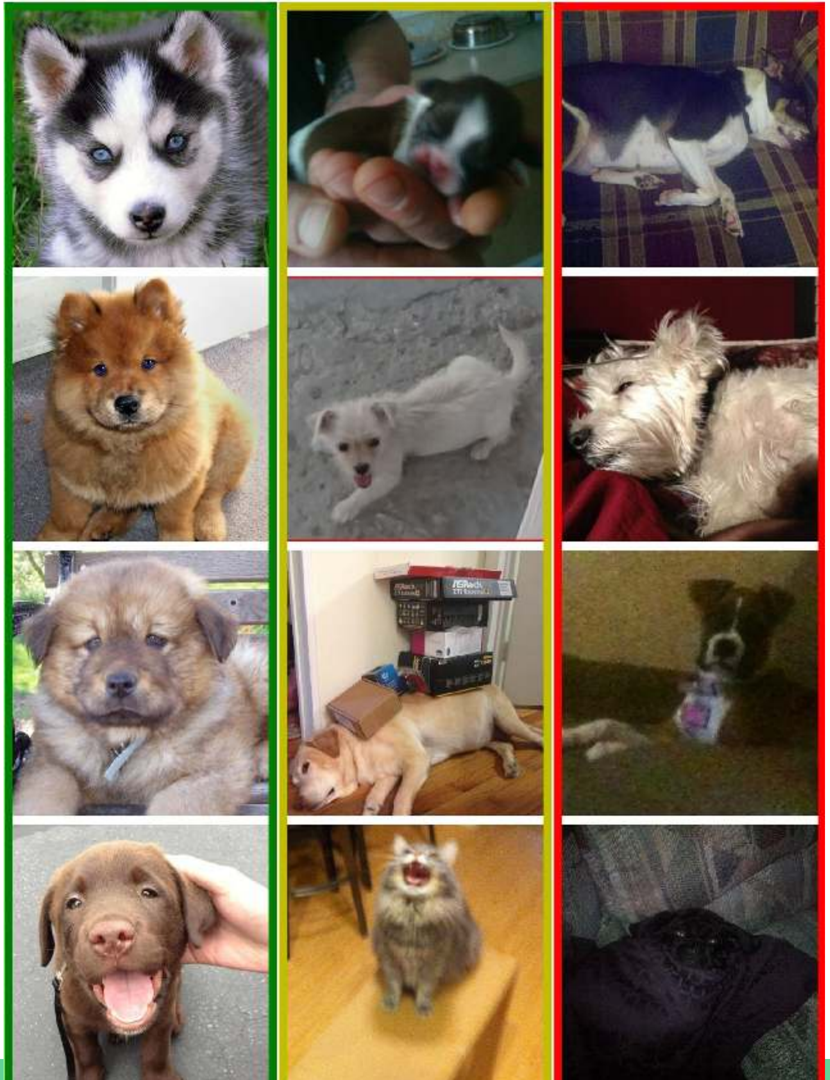


[Ding et al. 2019's instagram results]

Highest Scores

Lowest Scores

Highest Scores



Lowest Scores

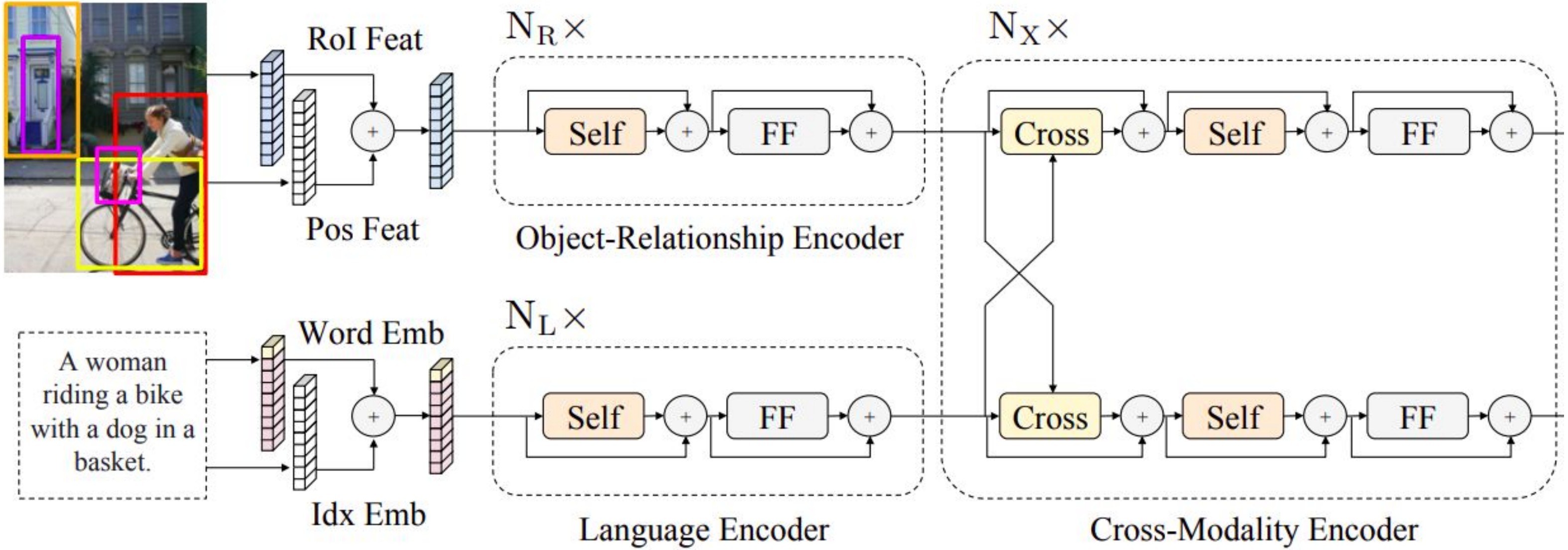
What is visual-textual grounding?

A collection of tasks ***requiring connection*** between visual and textual content.

In other words, text "multiplied by" images is more than text simply occurring with or alongside images.

--- Bateman, 2014
describing "Meaning Multiplication"
[Barthes 1988; Jones 1979]

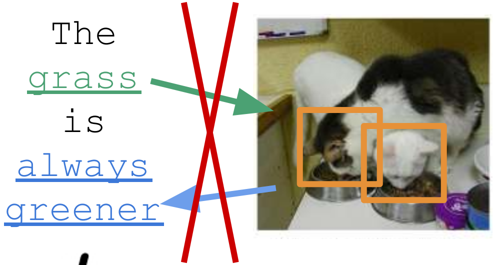
It can be difficult to tell what models learn...



[LXMERT: Tan and Bansal, 2019]

Can we formalize this a bit?

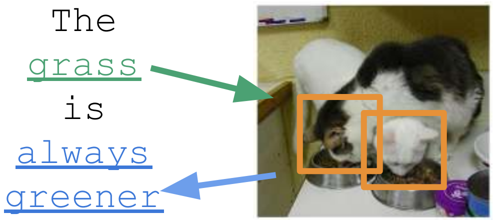
Multimodally additive model



$$t \qquad v$$

$$f(t, v) = f_t(t) + f_v(v)$$

Multimodally interactive model

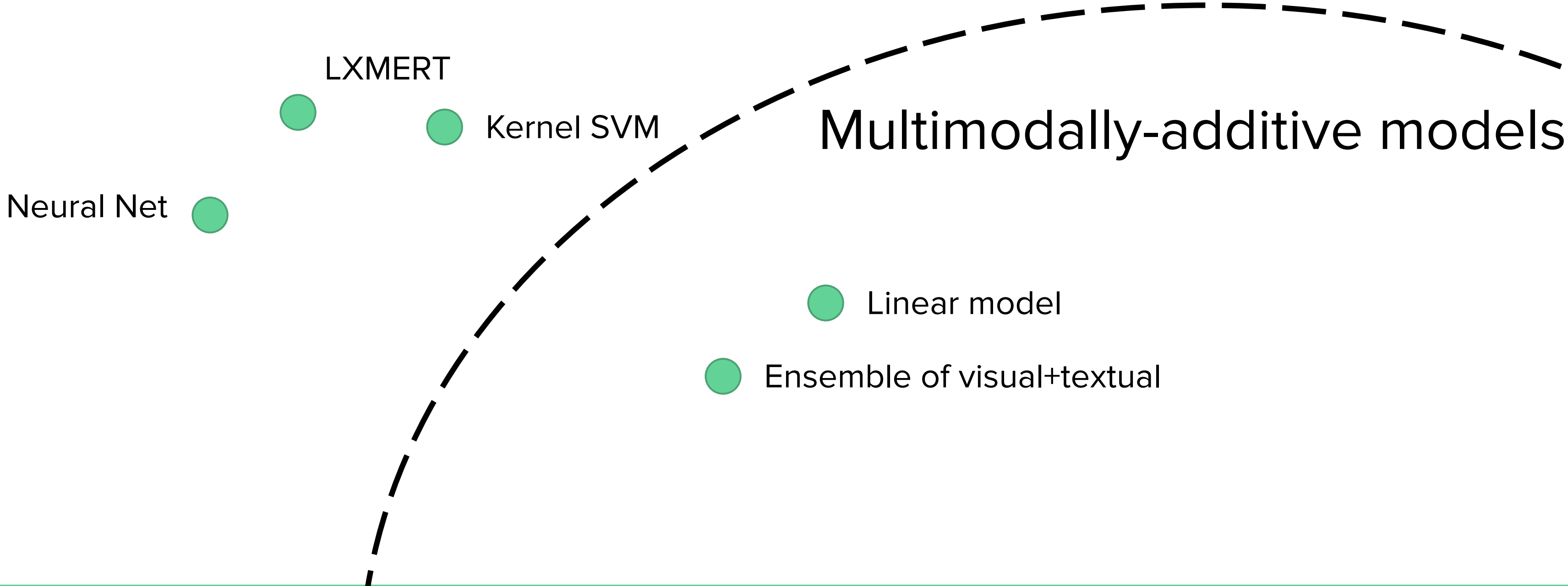


$$t \qquad v$$

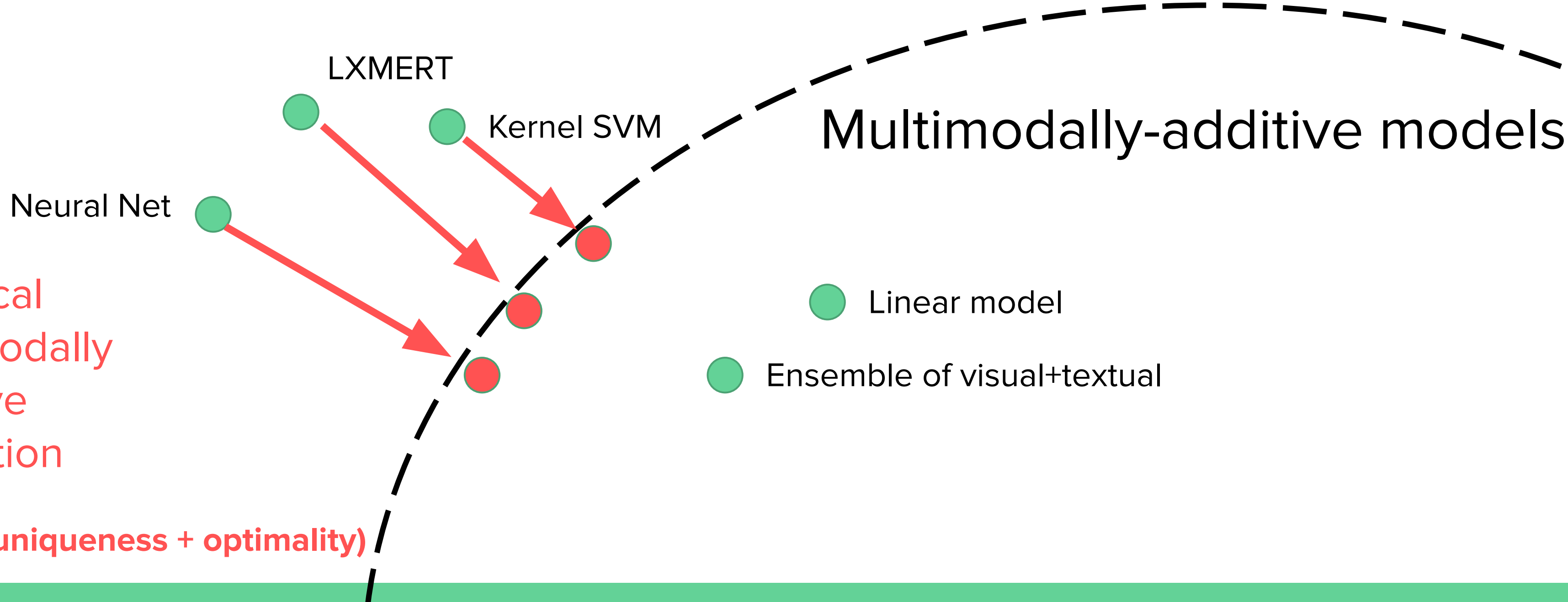
$$f(t, v) = f_{int}(t, v)$$

[Friedman 2001; Friedman et al. 2008; Hooker 2004]

Simplifying models with function projection



Simplifying models with function projection



Empirical
Multimodally
Additive
Projection

(We prove: uniqueness + optimality)

Proposing work	Task (structure)	Abbv.	# image+text
Kruk et al. (2019)	Instagram		
	↳ intent (7-way clf)	I-INT	1299
	↳ semiotic (7-way clf)	I-SEM	1299
	↳ contextual (7-way clf)	I-CTX	1299
Vempala and Preoţiuc-Pietro (2019)	Twitter visual-ness (4-way clf)	T-VIS	4471
Hessel et al. (2017)	Reddit popularity (Pairwise-ranking)	R-POP	88K
Borth et al. (2013)	Twitter sentiment (binary clf)	T-ST1	603
Niu et al. (2016)	Twitter sentiment (binary clf)	T-ST2	4511

	I-INT	I-SEM	I-CTX	T-VIS	R-POP	T-ST1	T-ST2
Metric	AUC	AUC	AUC	Weighted F1	ACC	AUC	ACC
Setup	5-fold	5-fold	5-fold	10-fold	15-fold	5-fold	5-fold
Prev. SoTA	85.3	69.1	78.8	44	62.7	N/A	70.5

	I-INT	I-SEM	I-CTX	T-VIS	R-POP	T-ST1	T-ST2
Metric	AUC	AUC	AUC	Weighted F1	ACC	AUC	ACC
Setup	5-fold	5-fold	5-fold	10-fold	15-fold	5-fold	5-fold
Prev. SoTA	85.3	69.1	78.8	44	62.7	N/A	70.5
Linear Model (A)	90.4	72.8	80.9	51.3	63.7	**75.6**	76.1

	I-INT	I-SEM	I-CTX	T-VIS	R-POP	T-ST1	T-ST2
Metric	AUC	AUC	AUC	Weighted F1	ACC	AUC	ACC
Setup	5-fold	5-fold	5-fold	10-fold	15-fold	5-fold	5-fold
Prev. SoTA	85.3	69.1	78.8	44	62.7	N/A	70.5
Linear Model (A)	90.4	72.8	80.9	51.3	63.7	**75.6**	76.1
Our Best Interactive (I)	**91.3**	**74.4**	**81.5**	**53.4**	**64.2***	**75.5**	**80.9**

	I-INT	I-SEM	I-CTX	T-VIS	R-POP	T-ST1	T-ST2
Metric	AUC	AUC	AUC	Weighted F1	ACC	AUC	ACC
Setup	5-fold	5-fold	5-fold	10-fold	15-fold	5-fold	5-fold
Prev. SoTA	85.3	69.1	78.8	44	62.7	N/A	70.5
Linear Model (A)	90.4	72.8	80.9	51.3	63.7	**75.6**	76.1
Our Best Interactive (I)	**91.3**	**74.4**	**81.5**	**53.4**	**64.2***	**75.5**	**80.9**
↳ + EMAP (A)	**91.1**	**74.2**	**81.3**	51.0	**64.1***	**75.9**	**80.7**

Well-balanced VQA datasets don't have this property

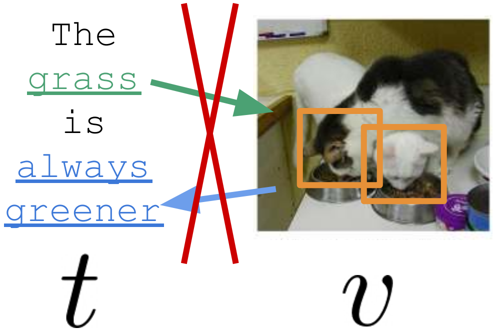
	LXMERT	+EMAP	Const.
VQA2	70.3	40.5	23.4
GQA	60.3	41.0	18.1

Accuracy results on dev set for LXMERT,
projected LXMERT, and constant prediction

Takeaway:

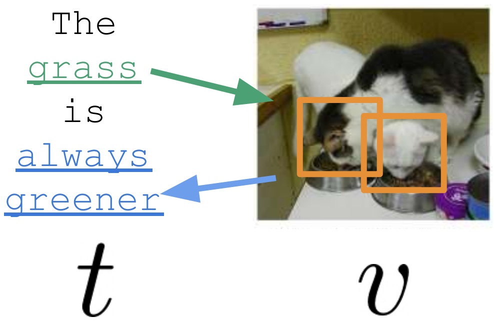
report the multimodally-additive projection performance!

<u>Multimodally additive model</u>



$$f(t, v) = f_t(t) + f_v(v)$$

<u>Multimodally interactive model</u>

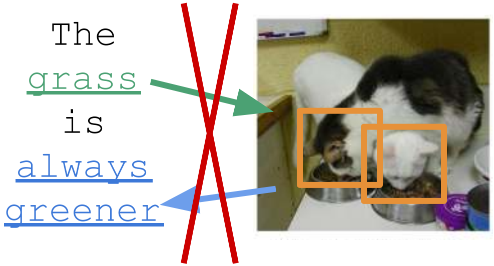


$$f(t, v) = f_{int}(t, v)$$

The Promise and the PERILS

What concepts are "groundable"?

[NAACL 2018, Hessel, Mimno, Lee]



Does my model learn cross-modal interactions?

[In Sub to EMNLP 2020: Hessel, Lee; WWW 2017, Hessel, Lee, Mimno]

The Promise and the PERILS

The Promise and the PERILS

We can *do cool things* with multimodal webdata,

but web texts are *not literal image descriptions*

(even though most algorithms treat them that way)

My Research Goals:

build better
grounding algorithms

requires

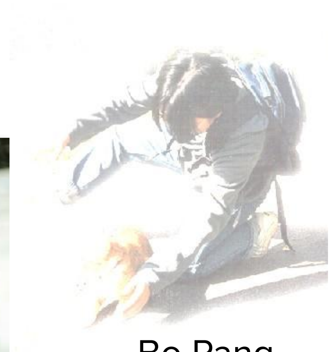
requires

understand web
communication

Thanks to my awesome collaborators!

Lillian Lee

David Mimno



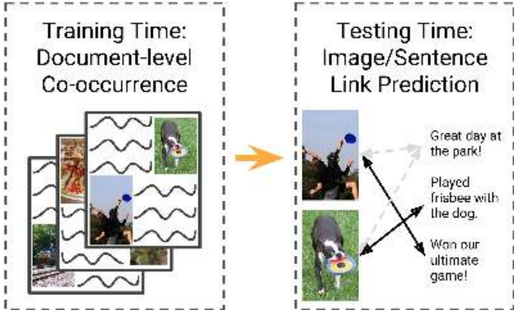
Bo Pang

Zhenhai Zhu

Radu Soricut

And thanks to you!!

The Promise and the PERILS



Contact:
jmhessel@gmail.com
@jmhessel on Twitter

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

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

[Lu et al., 2008; Berg et al., 2010; Parikh and Grauman, 2011; Yatskar et al. 2013; 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 empirical projection

Compute output for all image/text pairs, even mismatched ones not appearing in the data.

Return predictions with only additive structure that are minimally distant (according to squared error) from original predictions.

Algorithm 1 Multimodally-Additive Projection

Input: a trained model f that outputs logits; a set of text/visual pairs $\{(t_i, v_i)\}_{i=1}^{N}$

Output: the predictions of $\hat{f}$, the empirical $\mathcal{L}^2$ projection of f onto the set of multimodally-additive functions, on the input points.

$f_{cache} = \{\}, preds = \{\}$

for $i, j \in \{1, 2, \ldots, N\} \times \{1, 2, \ldots, N\}$ **do**

  $f_{cache}(i, j) = f(t_i, v_j)$

end for

$m = mean(f_{cache})$

for $i \in \{1, 2, \ldots, N\}$ **do**

  $proj_t = \frac{1}{N} \sum_{j=1}^{N} f_{cache}(i, j)$

  $proj_v = \frac{1}{N} \sum_{j=1}^{N} f_{cache}(j, i)$

  $preds[i] = proj_t + proj_v - m$

end for

return $preds$

(We prove: uniqueness + optimality)