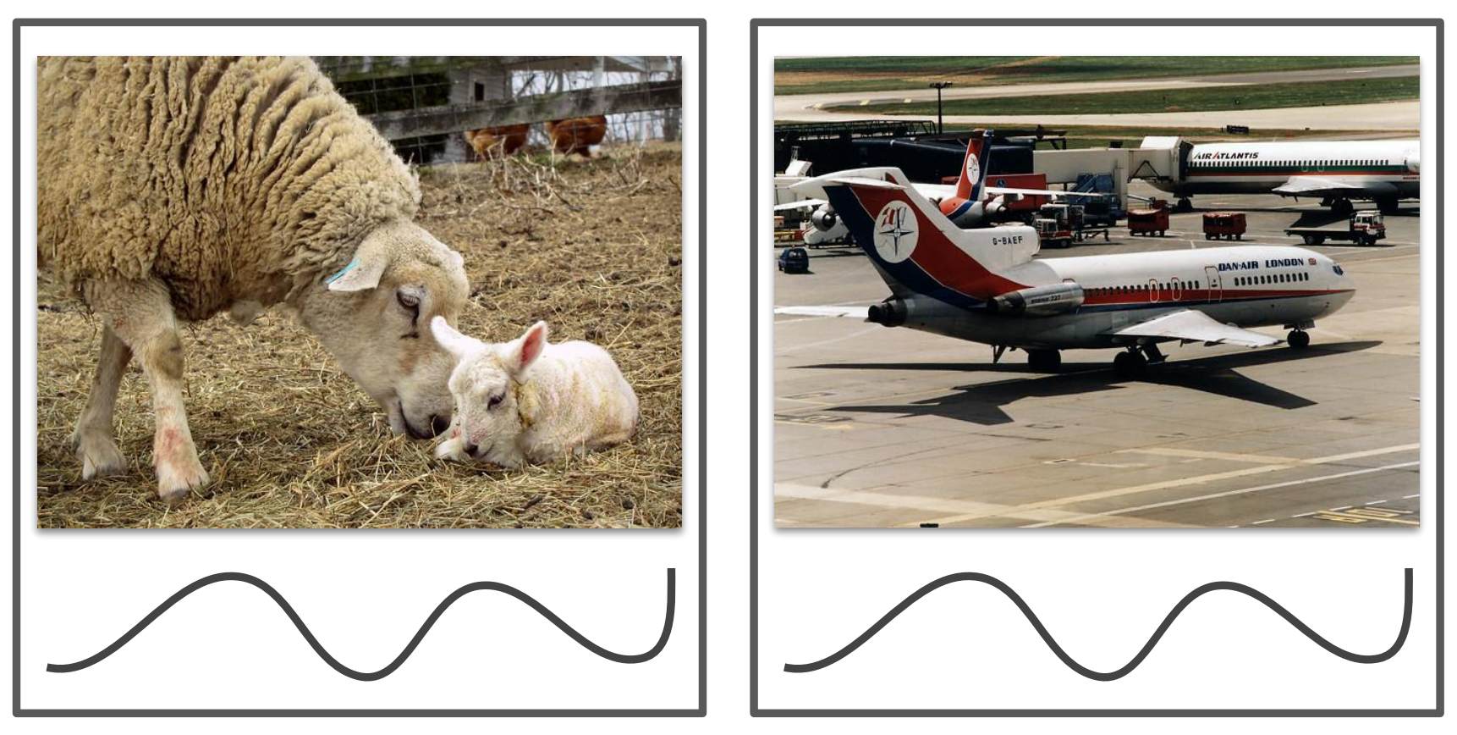
Unsupervised Discovery of Multimodal Links
in Multi-Image, Multi-Sentence Documents

Jack Hessel, Lillian Lee, David Mimno
Cornell University

What is a "multi-image, multi-sentence document"?

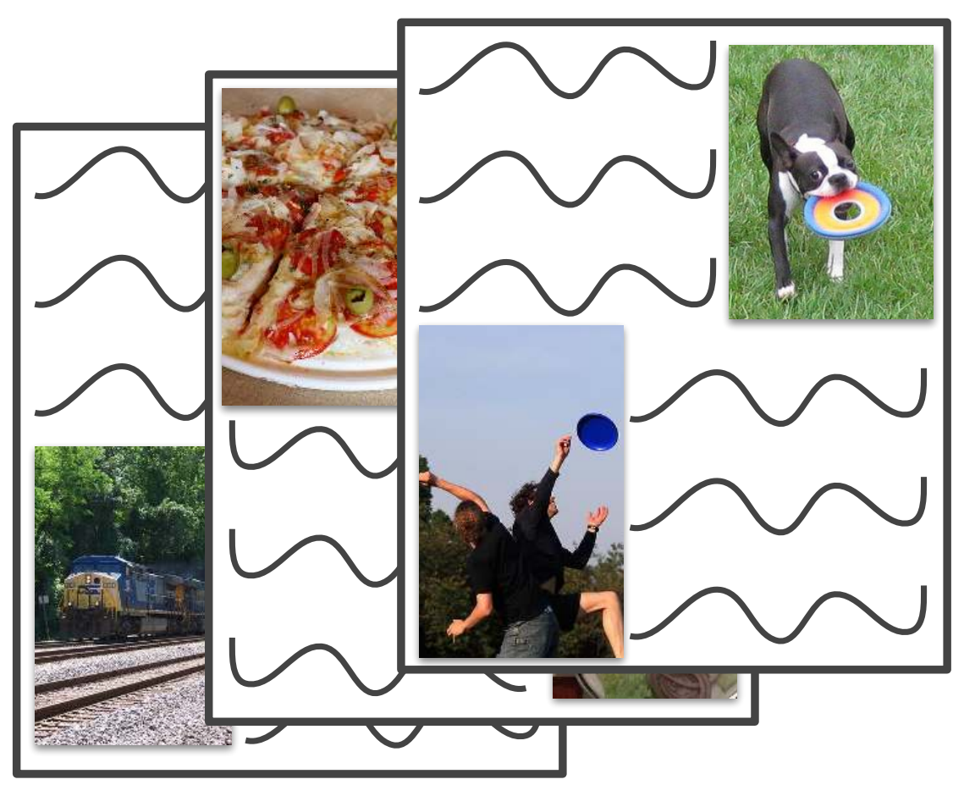
Image captioning/tagging case

single image,
explicit multimodal link by construction



Our case

Multiple images, multiple sentences,
no explicit multimodal links

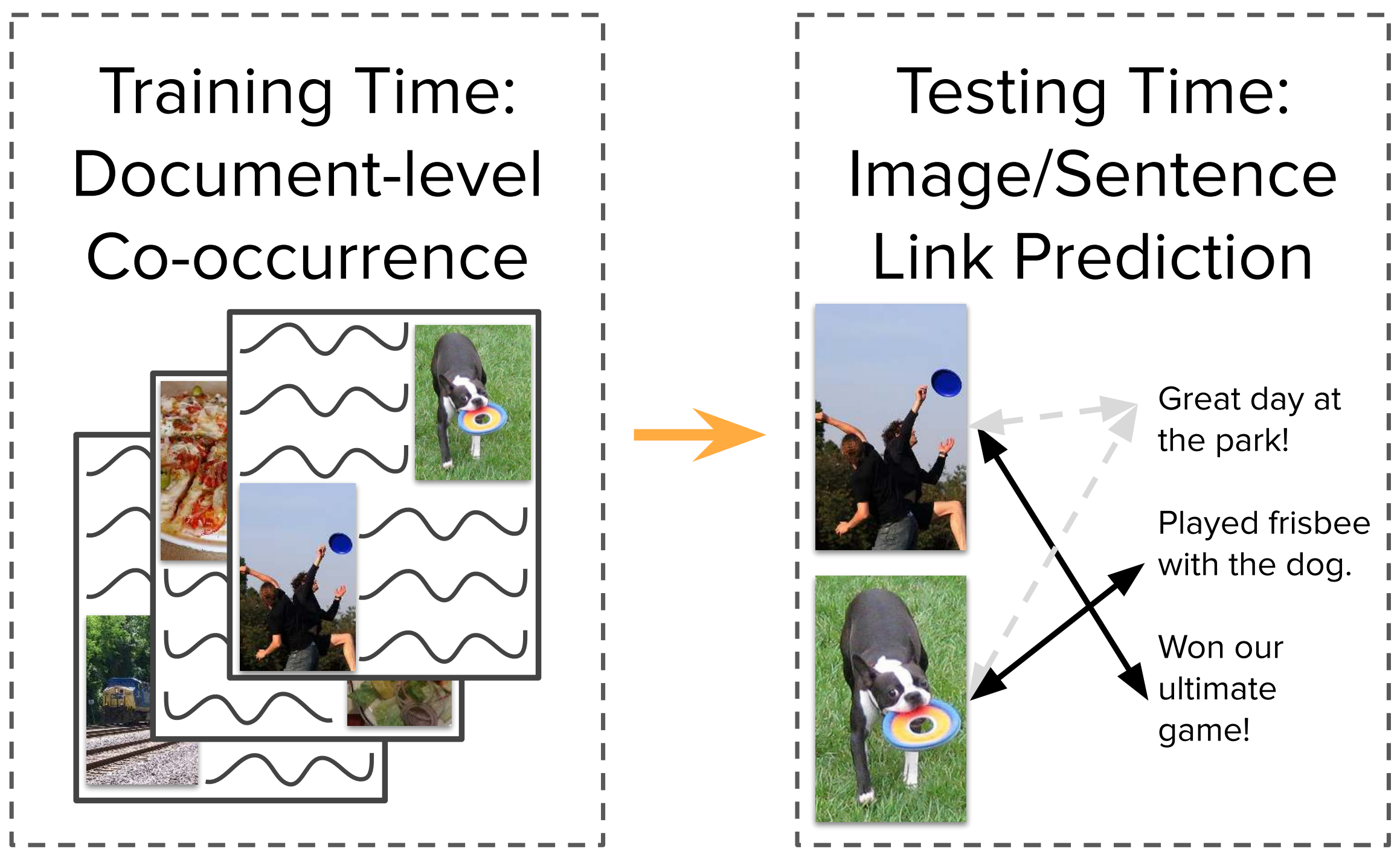


Web documents look less like **this** and more like **this!**

Multi-image, multi-sentence document use-cases:

1) provide **context-specific image captions** for low-vision and blind users
2) train image+text models **directly from unstructured web documents**

The Task: Unsupervised Link Prediction



Training Time: Document-level Co-occurrence

Testing Time: Image/Sentence Link Prediction

Datasets

Crowd-labeled Datasets:
Designed to address basic questions about this task

A run down street with grass growing in the middle it. A person's hand holding up a cell phone to a guinea pig in a cage. A man in a party hat sits at a table talking on a cell phone. A person doing a high jump on a skateboard. A keyboard sitting on a desk next to a large mouse pad. A man standing outside a building and practicing tennis. A person helping another person fix their skis. A photograph of sewing supplies including: scissors, a tape measure. Buttons and a needle & thread. A large white and blue bus driving down a street. Some people walking on the sand water and a kite surfer.

Q: Is this task even possible?
A: Microsoft COCO [1] "Documents"

The horses are small and in the pen. Two ponies are in a dirt covered field near a wire fence. Brown animals are standing up next to each other. Two horses are grazing on green grass outside. A brown horse with messy fur is staring at the camera.

Q: What if images/sentences are similar within a document?
A: Descriptions-in-Isolation [2]

[male] and [male] went to a fair on friday. There were lot of people there in the field. A big roller coaster was set up in the middle of the fair. There were also other ride to play on. Thankfully the last ride was the scariest ride that i refused to go on, was the one that went straight up and dropped down quickly.

Q: What if sentences are cohesive?
A: Stories-in-Sequence [2]

Q: What if many sentences do not refer to any image?
A: **DII-Stress**, a version of DII with 45 distractor sentences

Web-scraped Datasets:
Harder, more realistic cases

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... *** Finishing First Layer 1. Pour evenly into a pan... *** Onto the Mint! The Mint mixture can be changed ... Second Layer Is Finished! ...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 ...

RecipeQA [3]
9K documents, 88K images
6 sentences/8 images per doc

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.

"Do it Yourself"
9K documents, 154K images
15 sentences/16 images per doc

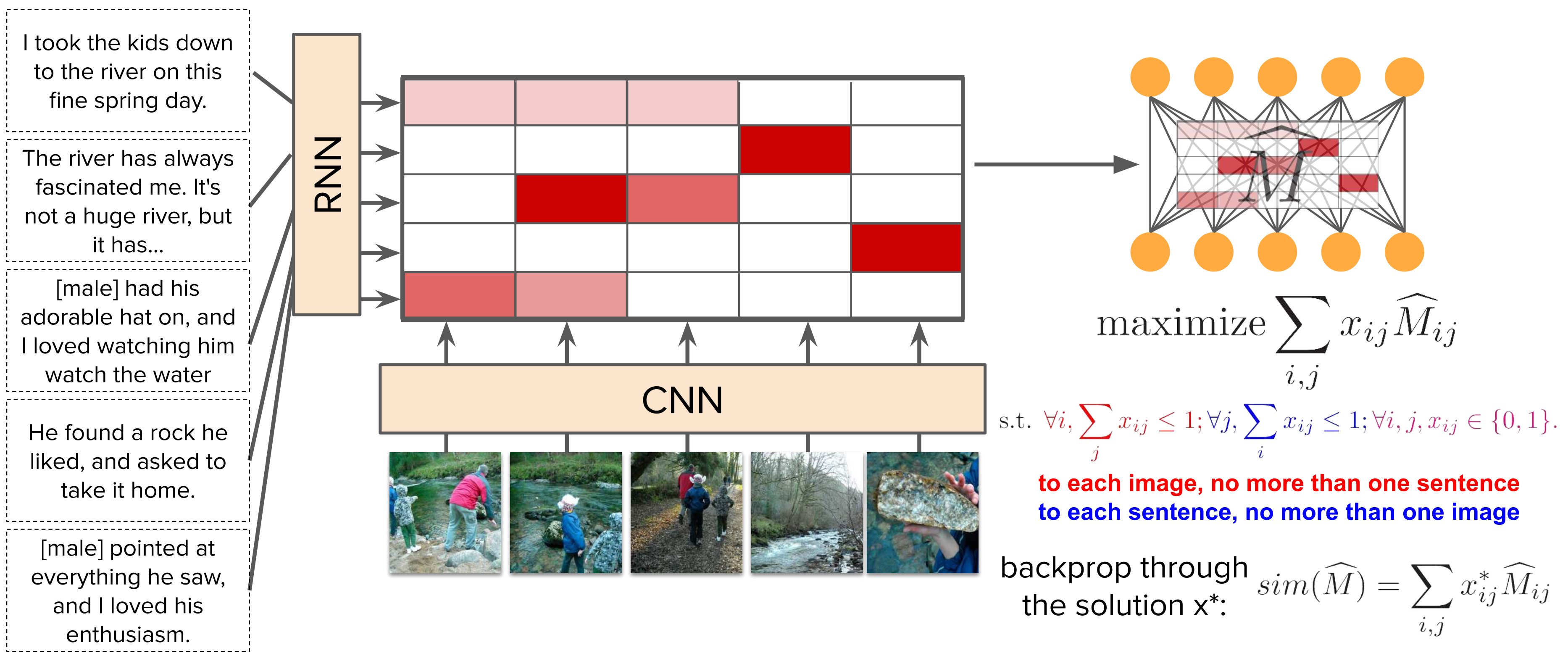
Rivet A rivet is a permanent mechanical fastener... Solid rivets consist simply of a shaft and head... Steel rivets can be found in static structures such as bridges, cranes, ... They are offered from 1/16-inch (1.6 mm) to 3/8-inch (9.5 mm) in diameter ... The most common machine is the impact riveter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

Wikipedia [4]
16K documents, 92K images
86 sentences/5 images per doc

[1] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In ECCV.
[2] 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.
[3] Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. In EMNLP.
[4] Adrian Popescu, Theodora Tsikrika, and Jana Kludas. 2010. Overview of the Wikipedia retrieval task at ImageCLEF 2010. In CLEF.

Our best-performing algorithm

CNNs and RNNs to extract features + solve *bipartite assignment* in the forward pass



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.

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

$$\text{s.t. } \forall i, \sum_j x_{ij} \le 1; \forall j, \sum_i x_{ij} \le 1; \forall i,j, x_{ij} \in \{0,1\}.$$

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

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

Training: maximize similarity between true (images, sentences), while minimizing similarity between negatively sampled (images, sentences)

Some baselines + quantitative results

Baseline 1: Object detection + word2vec

Baseline 2: NoStruct, a version of our algorithm with no structure

	MSCOCO		Story-DII		Story-SIS		DII-Stress		RQA		DIY	
	AUC	p@1/p@5	AUC	p@1/p@5	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	49.4	17.8/16.7	49.8	6.3/6.8
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	58.7	25.1/21.5	53.4	17.9/11.8
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	60.5	33.8/27.0	57.0	13.3/11.8
Proposed	**98.7**	**91.0/78.0**	**82.6**	**70.5/55.0**	**68.5**	**50.5/38.3**	**95.3**	**65.5/45.7**	**69.3**	**47.3/37.3**	**61.8**	**22.5/17.2**

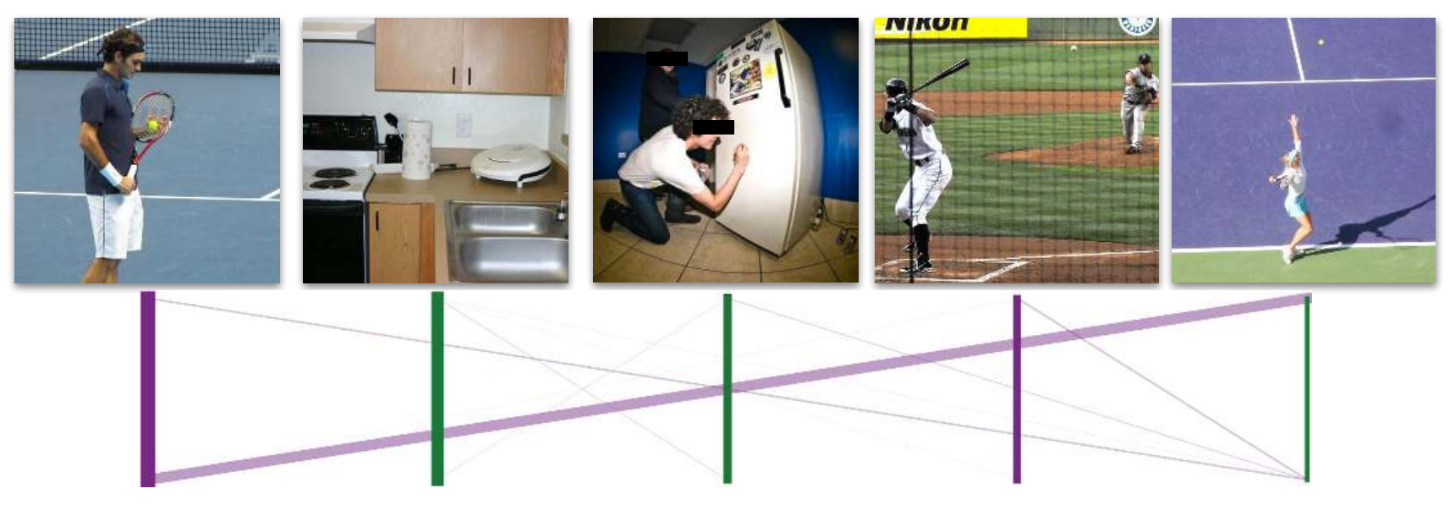
(higher = better)

Other variants/ablations are examined in the paper

Example Same-document Predictions

(Green is a ground-truth edge, purple is not)

Microsoft COCO



A woman with a tennis racket with a green background.

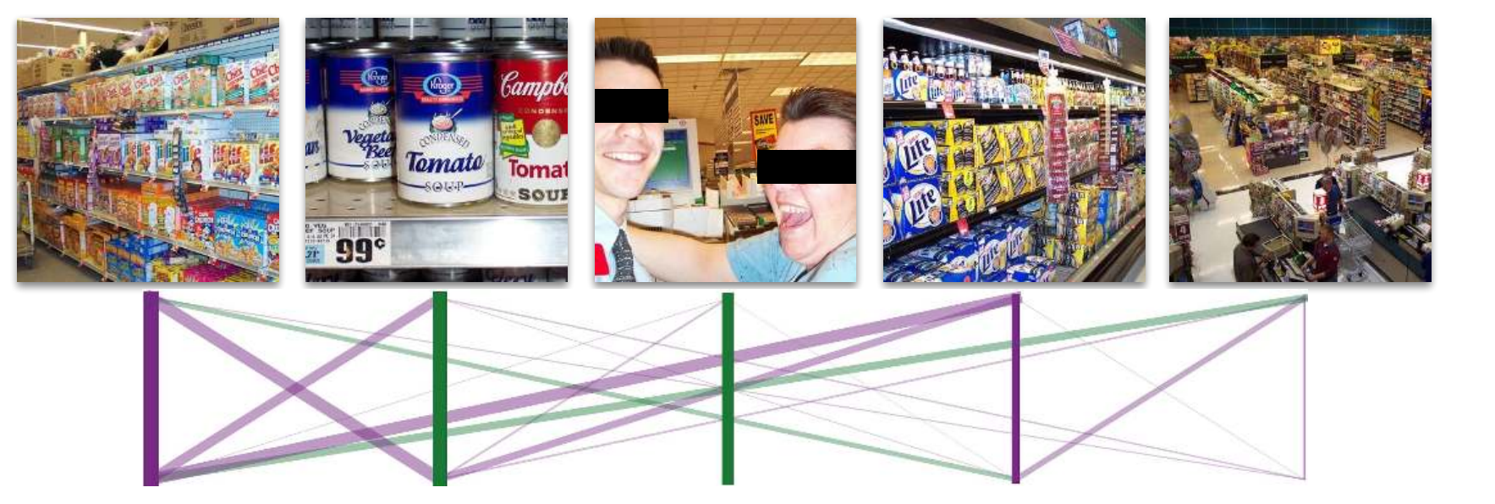
A kitchen with two metal sinks next to a stove top oven.

A young man writing on the door of a refrigerator

a field that has a few baseball players on it

A woman preparing to serve a ball thrown high in the air.

Stories-in-Sequence



I work at a grocery store, some may think it's lame...

The store even carries my favorite brand of soup...

My boss is great and makes me laugh.

I don't have to waste my time making extra trips after work...

...this tends to be the isle I visit for a nice relaxing...

Wikipedia

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

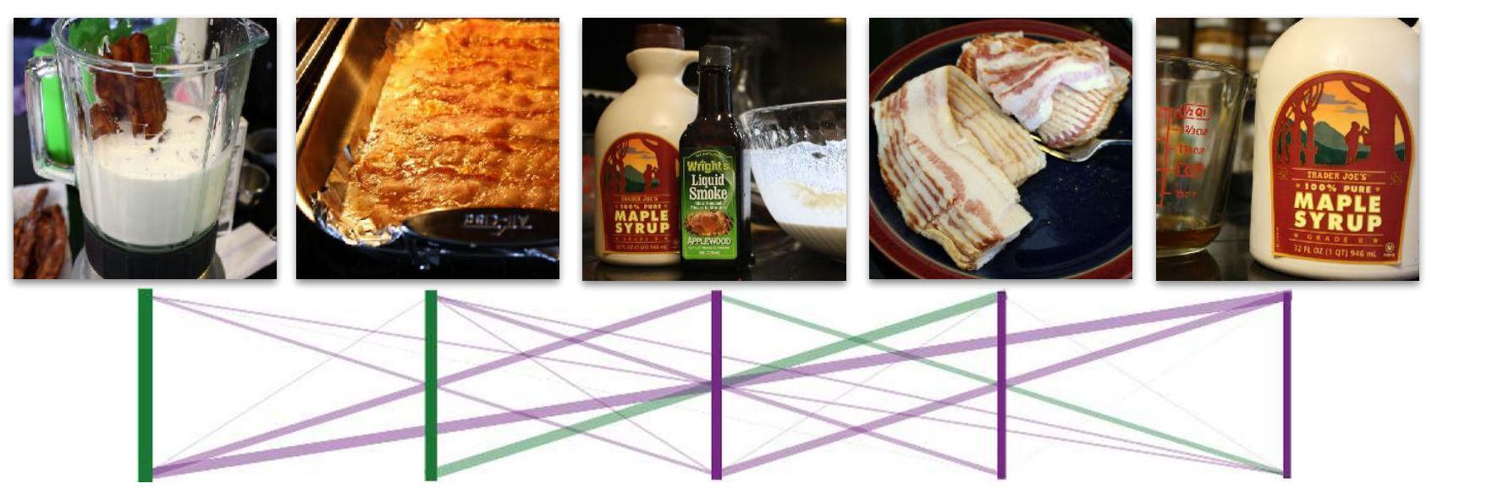
The island is well known for its natural beauty.

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

... population of Bhumihar Brahmins in Mauritius who have made a mark for themselves in different fields.

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

RecipeQA



Pour the quart of half-and-half into the blender.

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

... I made a triple batch for competition, this recipe...

... your "meat" strip in the center of the bacon...

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

What makes a document easier or harder?

Spread Hypothesis:

Documents with similar sentences/images will be harder to predict at test-time.

Content Hypothesis:

Some concepts are harder for image+text models to learn.

For crowd-labeled datasets, ***both the spread and content hypothesis*** explain document difficulty!

Data and Code Available!

http://www.cs.cornell.edu/~jhessel/multiretrieval/multiretrieval.html