Every Picture Tells a Story: Generating Sentences from Images

M.Sc. Seminar: Recent Developments in Computational Semantics

Nikolina Koleva

Saarland University
Department of Computational Linguistics

December 9, 2013

Overview

Motivation

Humans are able to provide concise description of a picture that focuses on the most important depicted parts.
The descriptions are accurate and with good agreement.

Motivation

Humans are able to provide concise description of a picture that focuses on the most important depicted parts.
The descriptions are accurate and with good agreement.

A child on a pink stool milking a black and white cow.
A young boy is milking a cow.
A young boy milking a cow while sitting on a pink stool.
A young boy sitting on a pink stool milking a cow.
Child milking a cow outside.

Motivation

Hypothesis

Automatic methods can do so, too.

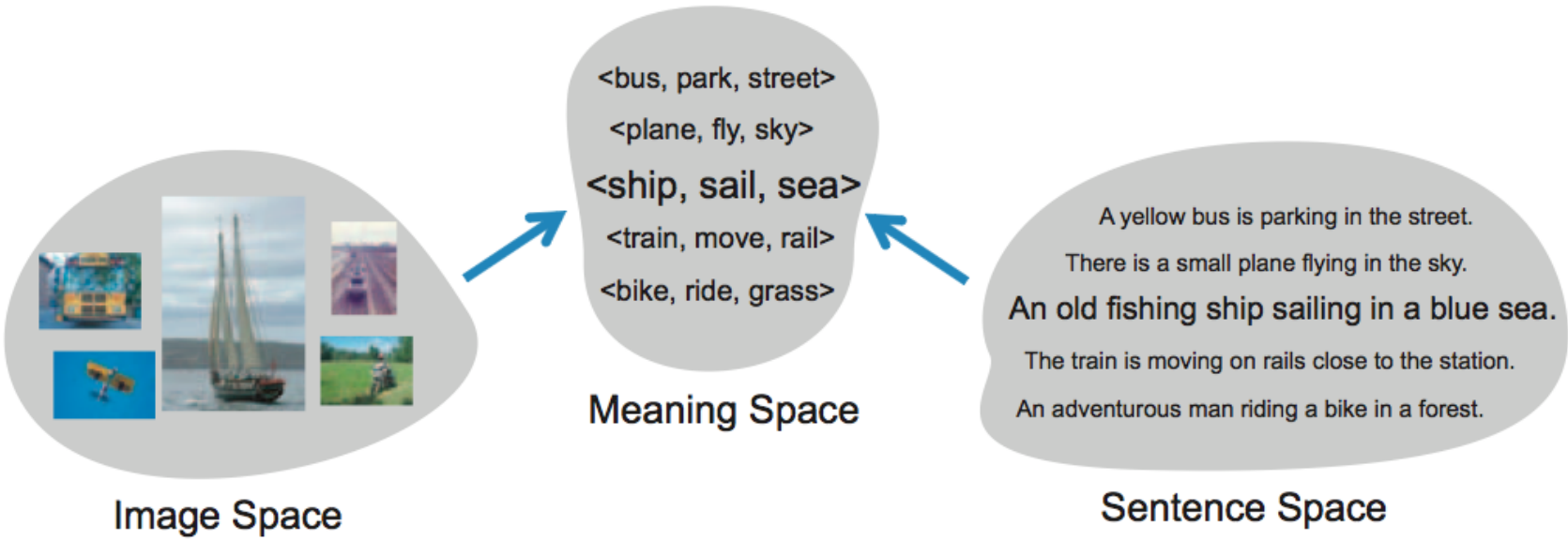
Goal

Demonstration of automatic correlation of a description to a given image and vice versa.

Motivation

Given	Retrieve
	A baby secured in a chair.
A man on a motorbike jumping with the sky behind him.	

General Idea

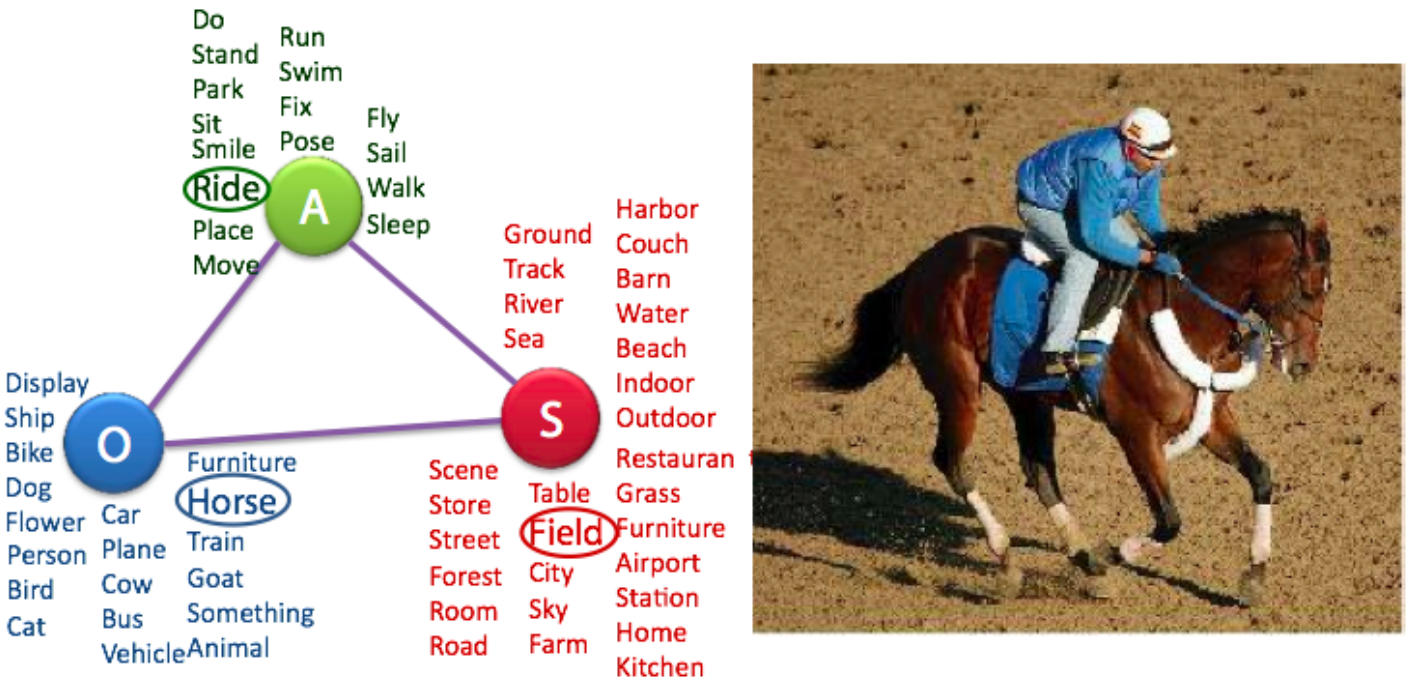


- Meaning represented as triplets: <object, action, scene>

☛ learn the projections from the image and sentence spaces to the meaning space

Mapping Image to Meaning

Solve a (small) multi-label Markov Random Field (MRF) to predict the triplet of an image



A, O and S have sets of discrete values

Image Node Potentials

1. *Image features* → linear combination of scores from different detector and classification responses
   - detector responses (Falzenwalb et al.)
   - classification responses (Hoiem et al.)
   - Gist-based scene classification responses (Oliva et al.)
2. *Node features* → predicted independently for each node by discriminative classifier (a linear `SVM`) given the *image features*
   - number-of-nodes-dimensional vector
   - each element gives score of for a node given an image

Image Similarity

1. obtaining the KNNs in the training set for a test image
   by matching (1) *image features* and (2) *node features* derived from classifiers and detectors
2. computing the average of the *node features* over those neighbours
   image side → what are the node features :
   sentence side → what does the sentence representation:
   (1) for similar images (2) for images that produce similar classifier and detector output

Sentence Similarity Measures

Compute the similarity of a sentence and the triplets.

1. compute dependency parses for each sentence
2. extract triplets of sentences
   object and action for a sentence: extract `subj`, direct `obj` and any `nmod` with a noun and a verb
   scene information: head nouns of the prepositional phrases (except "of" and "with")
3. Lin's similarity applied on **objects** and **scenes**
4. compute **action** co-occurrence scores
   detect similar verbs by checking if they appear in different captions for the same image
5. estimate sentence node potentials based on the measures 1-4

Sentence Node Potentials

- *sentence node feature*: similarity of each object, scene and action
- average of *sentence node features* for the other 4 captions
- `KNN` average of *sentence node features*
- average of the image node features for images of the neighbours
- average of the sentence node features of reference sentences for the neighbours

Edge Potentials

Defined as:

Linear combination of several estimates from node A to node B.

- the normalized frequency of the word A in the corpus, $f(A)$
- the normalized frequency of the word B in the corpus, $f(B)$
- the normalized frequency of A and B in the corpus, $f(A,B)$
- the ratio $\frac{f(A,B)}{f(A)\cdot f(B)}$

Learning and Inference

Triple prediction for pictures

discriminative learning based on a labeled training set

Learning the mapping from images to meaning:

find the set of weights of linear combinations of feature functions that maximize the ground truth triplets scores

Inference: search for a triple that gives the best score

additive: $\arg\max_y w^T \phi(x_i, y)$
multiplicative: $\arg\max_y \prod w^T \phi(x_i, y)$
ϕ the potential function
y is the triplet label
x_i the i-th image

Evaluation

- PASCAL Sentence Dataset:
  random selection of 50 images of 20 categories
  until set size 1000 images
- annotating each image with 5 sentences, resulting in 5000 sentences
- manual assignment of triplets
  173 different triplets in the training set (600 images)
  123 different triplets in the test set (400 images)
  overlap: 80 triplets
- 15 nearest neighbours for building the potentials for images and sentences
- 50 closest triples used by the matching

Evaluation

- scoring a match between an image and a sentence:
  ranking of k top triplets in the opposite space
  take the sum of ranks weighted by the inverse rank
  → low score = high similarity

Evaluation

- out of vocabulary words handled by distributional semantics methods unseen words are estimated during training by semantic similarity

Evaluation

- out of vocabulary words handled by distributional semantics methods
  unseen words are estimated during training by semantic similarity

From images to sentences

A red London United double-decker bus **drives** down a city street.

Two young **women** with two little girl near them

From sentences to images

A very colorful **Volkswagen** Beetle.

Cattle feeding at a trough.

Quantitative Measures

- **Tree-F1** measure: reflects accuracy and specificity
  using taxonomy trees: Object → Animal → Cat
- standard **F1** measure
  precision: total # of matching edges with ground truth
  recall: total # of edges in the predicted path
- **BLUE** Measure: checks if a triplet is logically valid or not
  e.g. <bottle, walk, street> is not valid

Results for image → meaning space

	Obj	No Edge	FW(A)	SL(A)	FW(M)	SL(M)
Mean Tree-F1 for first 5	0.44	**0.52**	0.38	0.45	0.47	0.51
Mean BLUE for first 5	0.24	0.27	0.16	0.58	**0.76**	0.74
Mean Tree-F1 for first 5 objects	**0.59**	0.58	0.36	0.53	0.55	0.57
Mean Tree-F1 for first 5 actions	0.27	**0.52**	0.50	0.37	0.42	0.47
Mean Tree-F1 for first 5 scenes	0.28	0.48	0.28	0.44	0.46	**0.48**

Obj: consider only obj potentials

No Edge: uniform potentials over edges

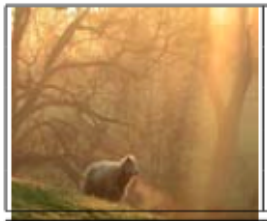
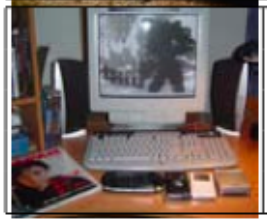
A: additive inference model

M: multiplicative inference model

FW: fixed weights

SL: structured learning

Results for sentence generation

	(pet, sleep, ground) (dog, sleep, ground) (animal, sleep, ground) (animal, stand, ground) (goat, stand, ground)	see something unexpected. Cow in the grassfield. Beautiful scenery surrounds a fluffly sheep. Dog hearding sheep in open terrain. Cattle feeding at a trough.
	(furniture, place, furniture) (furniture, place, room) (furniture, place, home) (bottle, place, table) (display, place, table)	Refrigerator almost empty. Foods and utensils. Eatables in the refrigerator. The inside of a refrigerator apples, cottage cheese, tupperwares and lunch bags. Squash apenny white store with a hand statue, picnic tables in front of the building.
	(transportation, move, track) (bike, ride, track) (transportation, move, road) (pet, sleep, ground) (bike, ride, road)	A man stands next to a train on a cloudy day A backpacker stands beside a green train This is a picture of a man standing next to a green train There are two men standing on a rocky beach, smiling at the camera. This is a person laying down in the grass next to their bike in front of a strange white building.
	(display, place, table) (furniture, place, furniture) (furniture, place, furniture) (bottle, place, table) (furniture, place, home)	This is a lot of technology. Somebody's screensaver of a pumpkin A black laptop is connected to a black Dell monitor This is a dual monitor setup Old school Computer monitor with way to many stickers on it

2 annotators for quality: 208 of 400 images have at least one of ten accurate sentence

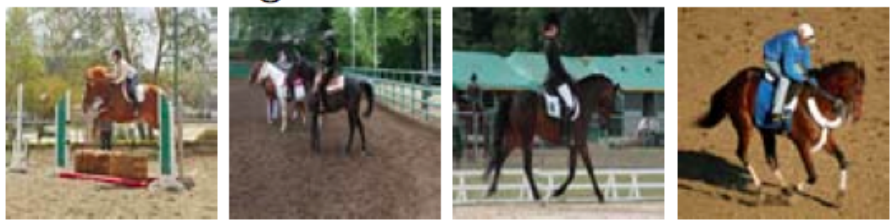
Retrieve images for sentences

A two girls in the store.

Yellow train on the tracks.

A small herd of animals with a calf in the grass.

A horse being ridden within a fenced area.



Failure examples

	A male and female giving pose for camera. A peaceful garden The food is ready on table.
	The two girls read to drive big bullet. Man with a goatee beard kneeling in front of a garden fence. Lone bicyclist sitting on a bench at a snowy beach.
	Black goat in a cage Horse behind a fence Wooly sheep standing next to a fence on a sunny day.

Summary

1. Image, Sentence and Meaning Space
2. leaning the projections of the image and sentence space to the triples in the meaning space
3. annotation of 1000 images each with five sentences as captions
4. meaning space used for generation of appropriate sentences given an image and retrieving images for a given sentence
5. out of vocabulary words handled with word similarity based on words distribution

Thank you for your attention!

Any questions?

References

Ali Farhadi, Seyyed Mohammad Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David A. Forsyth.

Every picture tells a story: Generating sentences from images.

In *ECCV (4)*, pages 15–29, 2010.

`http://vision.cs.uiuc.edu/pascal-sentences/`