Describing Images in Natural Language Part I

CVPR tutorial
Julia Hockenmaier
University of Illinois
juliahmr@illinois.edu

Overview

Part 1: High-Level Introduction to Sentence-Based Image Description

- What do we mean by image description?
- What kind of data sets are available?
- What kind of tasks have been proposed?
- How do we evaluate image description systems?
- A proposal for a shared task

Part 2: Digging deeper and going further

- A brief intro to NLP for image description
- Image description systems
- Image description and semantics

What is Sentence-based Image Description?

How would you describe this image?

A guy in a wetsuit petting a stingray

yes

Somebody kneeling down to touch a really flat fish

perhaps

Does that thing bite?

no

Why would you want to describe images?

Because you get to look at pictures of the cool things people do on their vacations?

Because you get to address some of the most fundamental problems in natural language understanding and artificial intelligence.

Because you get to work on a new, challenging task that could become really important.

Why would you want to describe images?

A test for grounded language understanding:
Image description requires the ability to associate sentences with images that depict the events, entities and scenes they describe.

A test for image understanding/vision:
Image description requires the ability to detect events, entities, and scenes in images.

Why would you want to describe images?

Traditional image retrieval maps text queries to text near the image. But the pictures you get from your camera/phone have no text associated with them.

We'd like to be able to associate text queries *directly* with images.

Sentence-based image description should improve:

... **Image search for everybody**

... **Accessibility to image collections for the visually impaired**

Word-based image annotation/search

Images and Words

Image tagging & search:

Annotate images with lists of keywords
Search for images by lists of keywords
(aka. Content-Based Image Retrieval)

True caption
market people

True caption
sky tree water

(Figures from Duygulu et al. 2002, Blei & Jordan 2003)

Region-based image annotation:

Annotate image regions with keywords

Content-based image retrieval

Image → Image (Query by visual example)

Given a query image, find images with similar content

Image ↔ Words (Semantic retrieval):

Induce a mapping between words and images from *weakly* labeled training data:

- Not all possible words are used to describe the image
- Words may not be associated with image regions
- Assumes a fixed vocabulary of words

Challenge:
The semantic gap

Mapping between images and words/concepts
is difficult because…

… different images of the same (kind of) object
may be visually very dissimilar
(due to different camera angles, lighting, pose, other attributes)

… images of different kinds of objects
may be visually very similar
(they may share textures, shapes, colors, etc.)

Data Sets

Corel5k (Duygulu et al. 2002), Corel30K (Vasconcelos)

5k or 30K tagged images

LabelMe data set (Russell et al., 2007)

Database of images with crowdsourced labeling of regions

http://labelme.csail.mit.edu

ImageNet (Jeng et al. 2009)

Augment WordNet synsets with images:

~22k synsets, 14M images (in 2010)

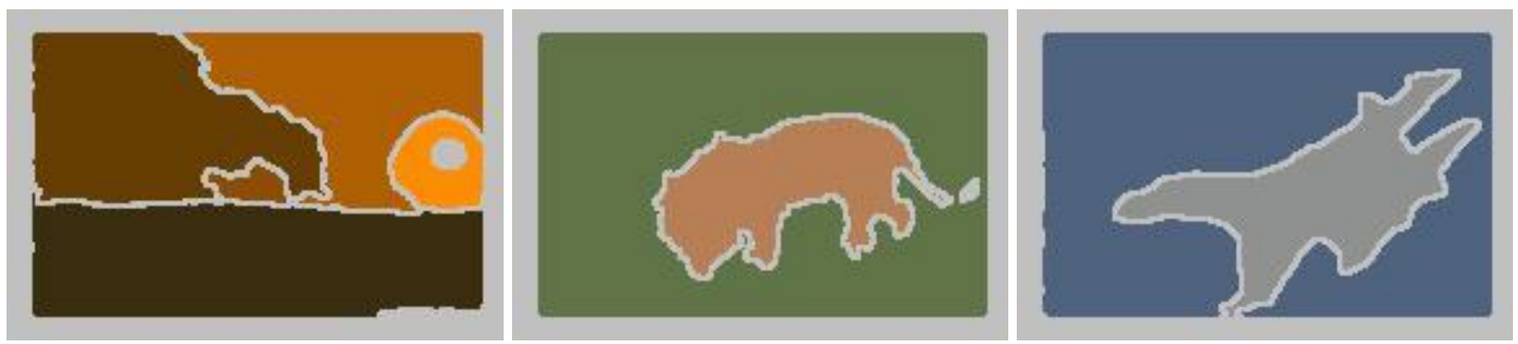
http://image-net.org

SUN (Scene UNderstanding) database (Xiao et al. 2010)

~100K images, ~900 scene classes

http://vision.princeton.edu/projects/2010/SUN/

Image annotation as machine translation Duygulu et al. 2002



sea sky sun waves | cat forest grass tiger | jet plane sky

Fig. 1. *Examples from the Corel data set. We have associated keywords and segments for each image, but we don't know which word corresponds to which segment. The number of words and segments can be different; even when they are same, we may have more than one segment for a single word, or more than one word for a single blob.*

Task: Annotate image regions with keywords (tags)

Model: IBM-style alignment

- map each image region to a visual vocabulary of 500 'blobs'
- train alignment model between blobs and tags

Data set: Corel5K

Bimodal topic models

Barnard et al. 2003, Blei & Jordan 2003

Basic idea:
Define a topic model in which topics generate image regions and keywords

Challenge:
Independence assumptions required by generative models may not be appropriate for this task

Deterministic annotation
Makadia et al. 2010

Input:
- a query image
- a pool of tagged images

Find k-Nearest Neighbor images to query image:
- Predefined image distance: Avg. over 7 basic distances (3 color histograms, 4 texture), each rescaled to lie between 0 and 1)

Transfer *n* of their labels to query image
- Use *n* most common labels of closest image
- If closest image has fewer labels: Remainder: based on remaining k-1 NN images.

Outperforms learning-based methods

Why are words alone not sufficient for description?

What do these words describe?

Object classes (*person*, *tree*) or instances (*Marilyn Monroe*)
Scenes (*market*, *crowd*)
'Stuff' (*grass*, *water*, *sky*)

But a bag of words cannot capture relations between objects or entities:

Spatial relations (*the book on the table*)
Actions (*the person is reading the book*)
Attributes (*the big book with the red cover*)

What do we want to say about an image?

How would you describe this image?

A boy in a yellow uniform carrying a football blocks another boy in a blue uniform.

yes

Two boys are playing rugby

perhaps

A dog is running on the beach.

no

How would you describe this image?

Jake tackled Kevin really hard.

perhaps

Last Sunday's game was really rough.

probably not

How would you describe this image?

There's a lot of vibrant blue and some pale green.

probably not

The image shows some shiny surfaces.

probably not

Image descriptions...

... should describe the depicted entities, events, scenes:
Who did what to whom, when and where?

... should *only* describe what is in the image:
No background information that cannot be seen.

... may differ in the *amount of detail* they provide
Each image has many correct descriptions.
Each sentence may describe many different images.

Image descriptions

[Shatford, Jaimes et al., Hollink et al.]

Perceptual image descriptions

- What kind of image?
  (photo vs. drawing, macro, panorama)
- Colors, textures, shapes

Non-visual image descriptions

- Additional context (*Last Sunday's game*)
- Metadata (*Nikon D90, f2.8, GPS coordinates*)

Image descriptions

[Shatford, Jaimes et al., Hollink et al.]

Conceptual image descriptions:
Who did what where to whom?

What events, scenes, entities are depicted?

- Generic: *Kids playing football.*
- Specific: *Jake tackling Kevin.*
- Abstract: *Childhood; Competition*

Most appropriate for image search etc., and for image description as a test for language understanding.

Summary:
What is image description?

Definition of sentence-based image description:

Sentence-based image description is the task of associating images with natural language sentences that describe what entities, events and scenes are depicted in them.

Applications of sentence-based image description:

- Searching online or personal image collections
- A testbed for image understanding
- A testbed for grounded language understanding

Comparing image description systems

Task definitions differ:

- Generate captions directly from image features
- Transfer captions from similar images
- Rank a pool of captions for each image

Models and **representations** differ:

- Image features: low-level features, detector responses
- Linguistic features: words, syntax, lexical semantics, roles
- 'Semantic' mapping between images and language

Data sets differ:

- UIUC Pascal: 1K images, 20 object types, crowdsourced captions
- UIUC 8K: 8k Flickr images, people/dogs, crowdsourced captions
- SBU data set: 1M Flickr images with Flickr captions

Evaluations differ:

- Human judgments or automated metrics

Image tagging vs. describing images with sentences

Image tagging is a **multi-label classification task**:
Given a large (but fixed, finite) set of tags,
predict which ones can be used for an image

Sentences are **compositional**:
We cannot assume we are dealing with a fixed,
finite set of labels.

Sentences have **ambiguous structure**:
Understanding a sentence requires disambiguation.

Data Sets for

Image Description

Data sets for sentence-based image description

To develop and evaluate sentence-based image description systems, we need corpora of images paired with appropriate captions.

- What data sets are available?
- What strengths and weaknesses do they have?
- What other data could be leveraged for this task?

Data sets for sentence-based image description

Using captioned images from the web (news, photo-sharing sites)

Advantage: Size, 'natural' captions

Disadvantage: Online captions may not describe images

SBU Captioned Photo data set; BBC data set

Using images with purposely created captions

Advantage: Sentences describe the images

Disadvantage: Smaller size, 'unnatural'

IAPR-TC; Illinois Pascal data set, Flickr 8K, Flickr30

News sites often use images just to embellish their stories

Drinking over recommended limit 'raises cancer risk'

COMMENTS (349)

Drinking more than a pint of beer a day can substantially increase the risk of some cancers, research suggests.

The Europe-wide study of 363,988 people reported in the British Medical Journal found one in 10 of all cancers in men and one in 33 in women were caused by past or current alcohol intake.

More than 18% of alcohol-related cancers in men and about 4% in women were linked to

Many people do not know that drinking alcohol can increase their cancer risk.

BBC data set

Police confirmed Tony Blair was inside at the time of the incident

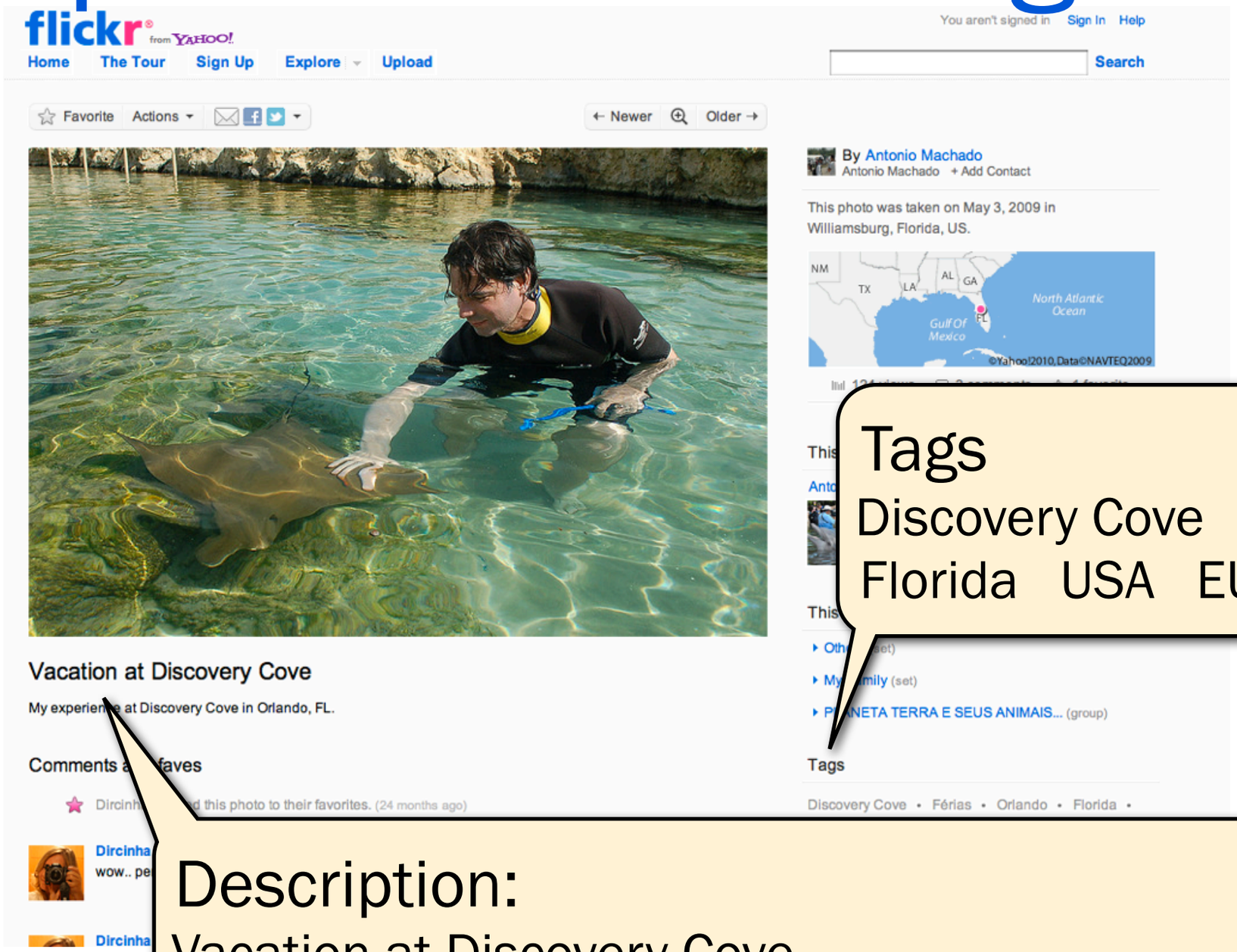
A man has been charged with possessing a knife and assaulting a police officer during an intrusion into the secure area of the prime minister's residence.
Byung Jin Lee, aged 32, was detained after scaling six-foot-high iron railings at Downing Street on Sunday.
Tony Blair was at home during the incident at the back of Number 10, but police said he was not "at risk".
Mr Lee was arrested after a brief struggle and will appear before magistrates on Tuesday.
Scotland Yard said: "We are satisfied that at no time was the prime minister at risk."
Mr Lee is due to appear at the City of Westminster Magistrates Court.
How the drama unfolded outside Downing Street

Crowds flocked to hear techno DJs play

Hundreds of thousands of revellers turned out for the return of Berlin's Love Parade to enjoy a sunshine-filled day of techno music.
The parade returned after a two-year gap caused by financial problems.
Organisers had hoped to attract one million people to the area around the Brandenburg Gate, which was recently home to the World Cup's Fan Mile.
About 40 decorated floats drove through the streets with DJs aboard to entertain the crowds.
The Love Parade started out as a small rave back in 1989 when the Berlin Wall fell.
It continued to grow to a peak of 1.5 million in 1999, but numbers then began falling with commercialism and growing costs blamed for its decline.
It was then scrapped for two years until a German fitness company stepped in with sponsorship in a bid to revive it.
This year, under the banner The Love is back, dance music fans from across the world flocked back to the German capital.
"This is great! I've been waiting three years for this," said Berliner Nicole Koehler, 25. "Hopefully it will be here every year from now on."
The festival went on all day, with parties continuing at the city's nightclubs.
Among those performing were international DJs Paul van Dyk, Westbam and Tiesto

Feng and Lapata 2010

On photo-sharing sites, people describe images...



Tags
Discovery Cove Férias Orlando Florida USA EUA Vacations

Description:
Vacation at Discovery Cove
My experience at Discovery Cove in Orlando, FL

… but they don't provide conceptual descriptions…

… because they write for (other) people—who can see what's in the picture.

Why bore them?

Gricean maxims:

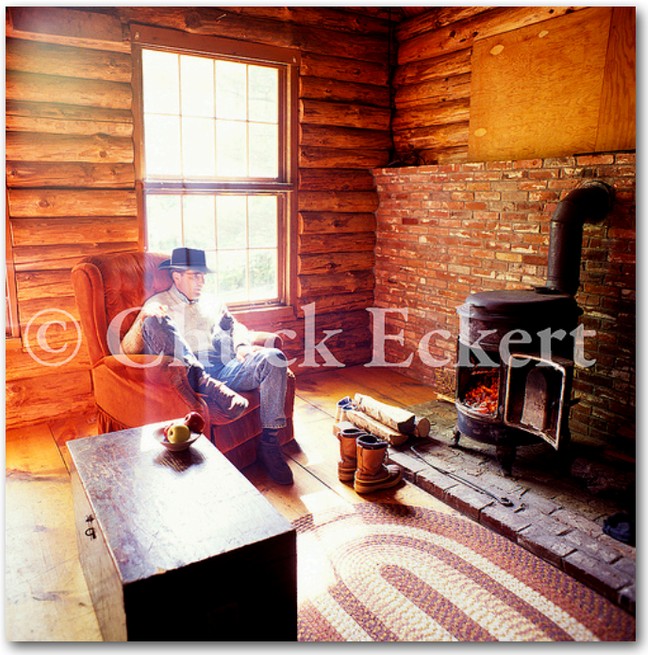
Be informative!

Be relevant!

SBU Captioned Photo Dataset

Not the best idea to roll around on my floor wearing white

Asbestos is used willy-nilly in ukraine as building material. Dad and Tanya have an asbestos roof and an asbestos fence.



Man in mountain cabin sitting by fireplace on a cold winters day

1M images and captions harvested from Flickr

Ordonez et al. 2011

IAPR-TC12 data set

Horse-Riding at the pampas

six people are riding on brown and white horses in a green, flat meadow in the foreground; cows behind them; white and grey clouds in a light blue sky in the background;

Buenos Aires, Argentina
9 December 2004

Panoramic View of the Iguazu Waterfalls

a cascading waterfall in the middle of the jungle; front view with pool of dirty water in the foreground;
this picture was taken from the Brazilian side;

Foz do Iguaçu, Brazil
March 2002

20,000 manually annotated and segmented images
Grubinger et al. 2006; Escalante et al. 2010

Illinois PASCAL data set

A grounded passage plane in a terminal.
An Air Pacific airplane sitting on the tarmac.
Large white commercial airliner parked on runway.
The back and right side of a parked passenger jet.
The passenger plane is sitting at the airport.

A hand holding bird seed and a small bird.
A person holding a small bluebird.
A person holds a bird and seeds.
A small bird is sitting on a person's hand that has bird seed in it.
A small black, white, and brown bird perched on and eating out of a man's hand.

1,000 images from the PASCAL VOC 2008 challenge (20 object categories) with 5 crowdsourced captions Rashtchian et al. 2010

Illinois Flickr8k/30k data sets

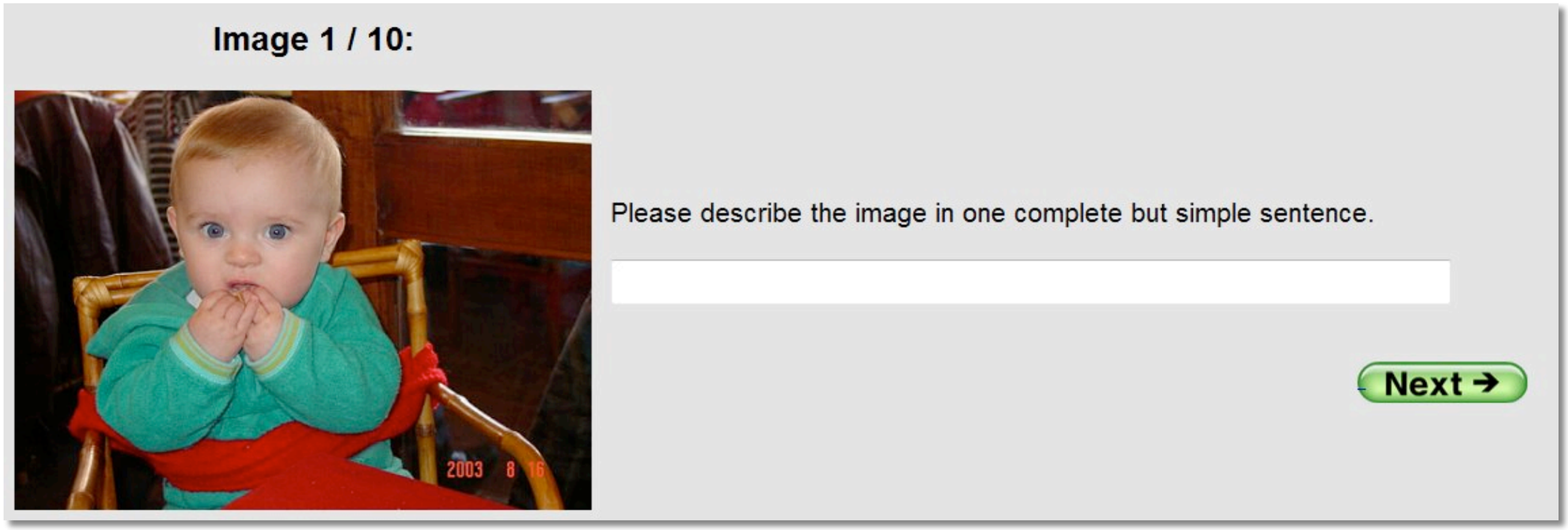
A goalie in a hockey game dives to catch a puck as the opposing team charges towards the goal.
The white team hits the puck, but the goalie from the purple team makes the save.
Picture of hockey team while goal is being scored.
Two teams of hockey players playing a game.
A hockey game is going on.

A group of people are getting fountain drinks at a convenience store.
Several adults are filling their cups and a drink machine.
Two guys getting a drink at a store counter.
Two boys in front of a soda machine.
People get their slushies.

32k images of people (and dogs) from Flickr with 5 crowdsourced captions

Rashtchian et al. 2010, Hodosh et al. 2013, Young et al. 2014

Image description with Amazon Mechanical Turk



Instructions:

Describe the objects and actions; Use adjectives; be brief

5 captions per image

Four basketball players in action.
Young men playing basketball in a competition.
Four men playing basketball, two from each team.
Two boys in green and white uniforms play basketball with two boys in blue and white uniforms.
A player from the white and green highschool team dribbles down court defended by a player from the other team.

A man crouched on a snowy peak.
A man in a green jacket stands in deep snow
at the base of a mountain.
A man kneels in the snow.
A man measures the depth of snow.
A mountain hiker is digging stakes
into the thick snow.

Image Description

Tasks

Comparing image description systems

Task definitions differ:

- Generate captions directly from image features
- Transfer captions from similar images
- Rank a pool of captions for each image

Models and **representations** differ:

- Image features: low-level features, detector responses
- Linguistic features: words, syntax, lexical semantics, roles
- 'Semantic' mapping between images and language

Data sets differ:

- UIUC Pascal: 1K images, 20 object types, crowdsourced captions
- UIUC 8K: 8k Flickr images, people/dogs, crowdsourced captions
- SBU data set: 1M Flickr images with Flickr captions

Evaluations differ:

- Human judgments or automated metrics

Defining f(I, S)

All image-description systems need a way to score image-sentence pairs (I, S).
This score may or may not be mediated by a (predefined or induced) semantic space.

f(I, S) can be:

- the score of a (discriminative) probabilistic model or classifier (e.g. CRF/MRF, RankSVM)
- the distance of I and S in an induced semantic space (Kernel Canonical Correlation Analysis, other joint embeddings)
- ...

Image features

Low-level features to compare images/image regions
Color, texture, SIFT, HOG, GIST

Detector responses:
to identify regions that are likely to depict objects/stuff
to label the image
as (binary) features

Text features

Words and n-grams:

- possibly augmented with hypernyms
- possibly with lexical similarities

Grammatical roles and word-word dependencies to fill slots and to mediate between text and detectors

- NPs = actor/objects
- verb = activity
- PPs = scene (location) or ‘stuff’

Task definitions

Generate captions directly from image features:

- Requires an explicit mapping between image & text.
- Requires a surface realization model to guarantee fluency etc.
- Requires human evaluation of correctness & grammaticality.

Transfer (and combine) captions from similar images:

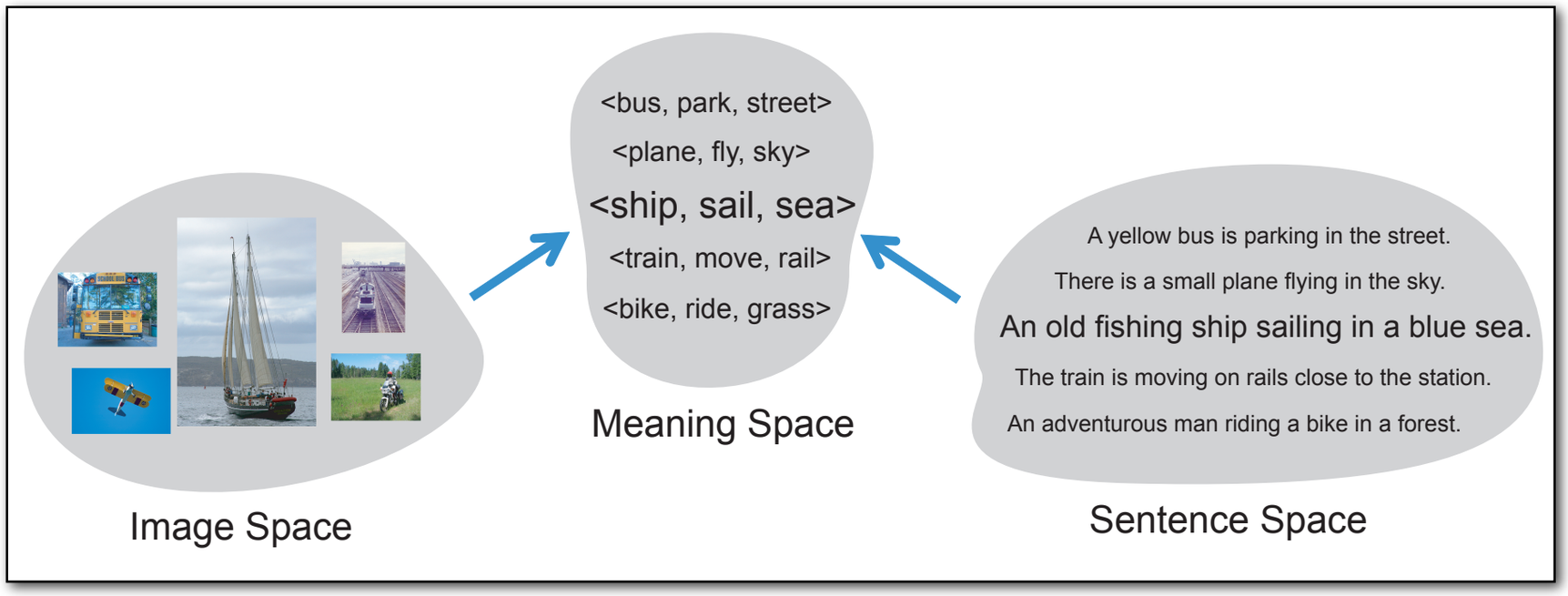
- Requires a unimodal (image) similarity metric
- May also require a surface realization model.
- Requires human evaluation of correctness & grammaticality.

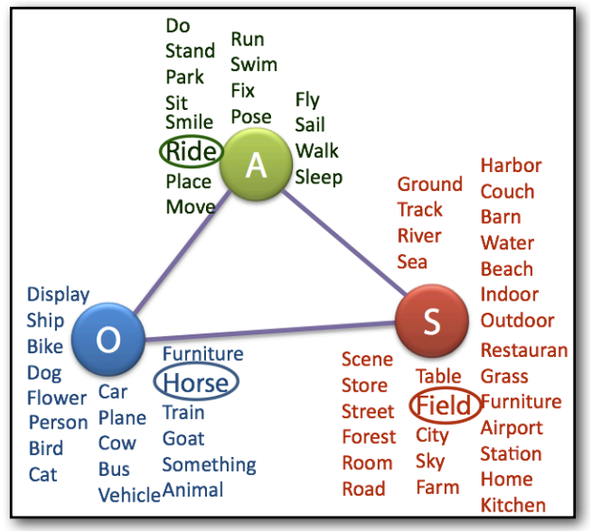
Score and rank a pool of captions for each image:

- Requires a cross-modal (image-sentence) similarity metric.
- Benefits from human relevance judgments,
- but may be evaluated automatically.

Image description as a *cross-modal ranking* task with explicit semantics

Mapping images & sentences to an explicit semantic space





Farhadi et al. 2010

Semantics of images/sentences:≈ ⟨Object, Action, Scene⟩

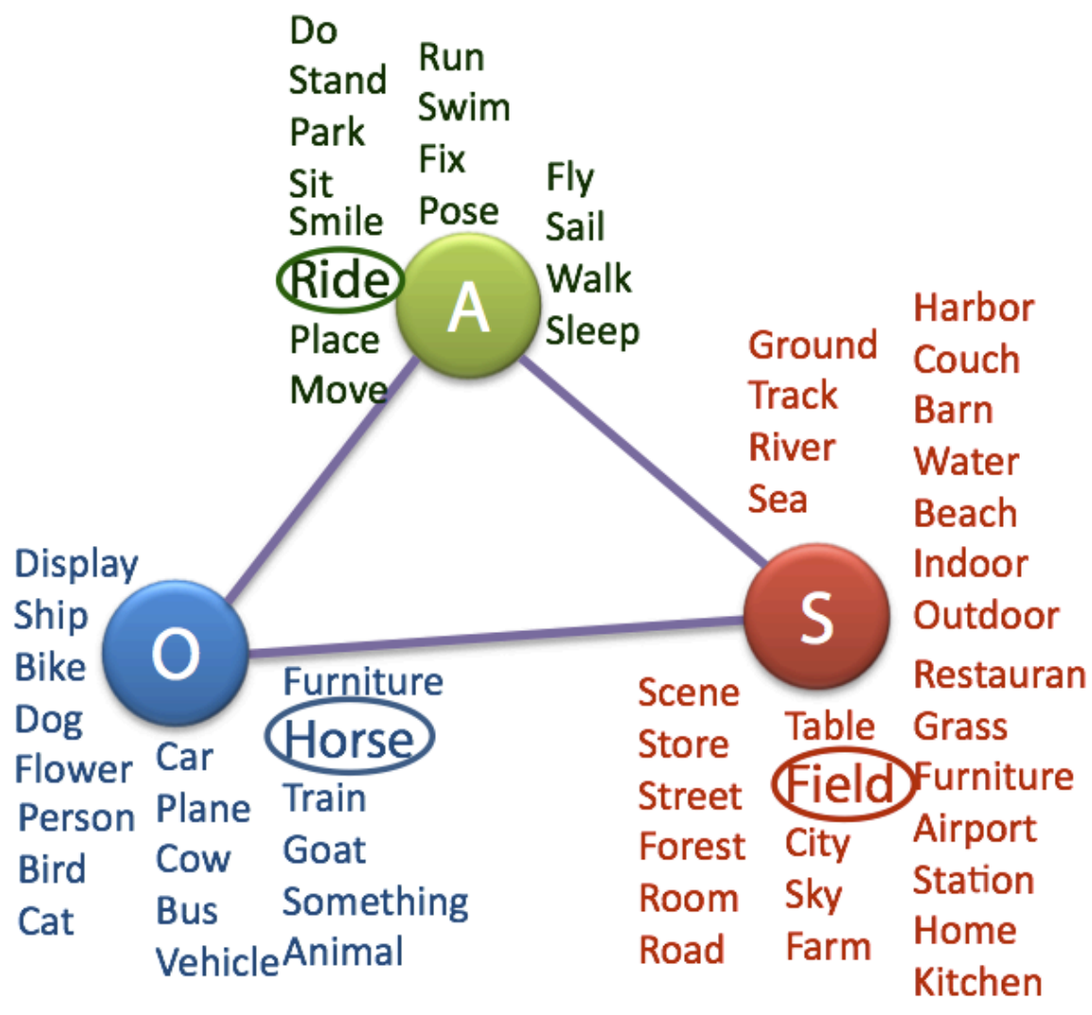
Use Markov Random Fields to predict most likely meaning triplet for images and sentences.

Naive image semantics

Represent an image as ⟨Object , Action, Scene⟩

Assume a fixed type inventory for each slot:

- 23 objects
- 16 actions
- 9 scenes

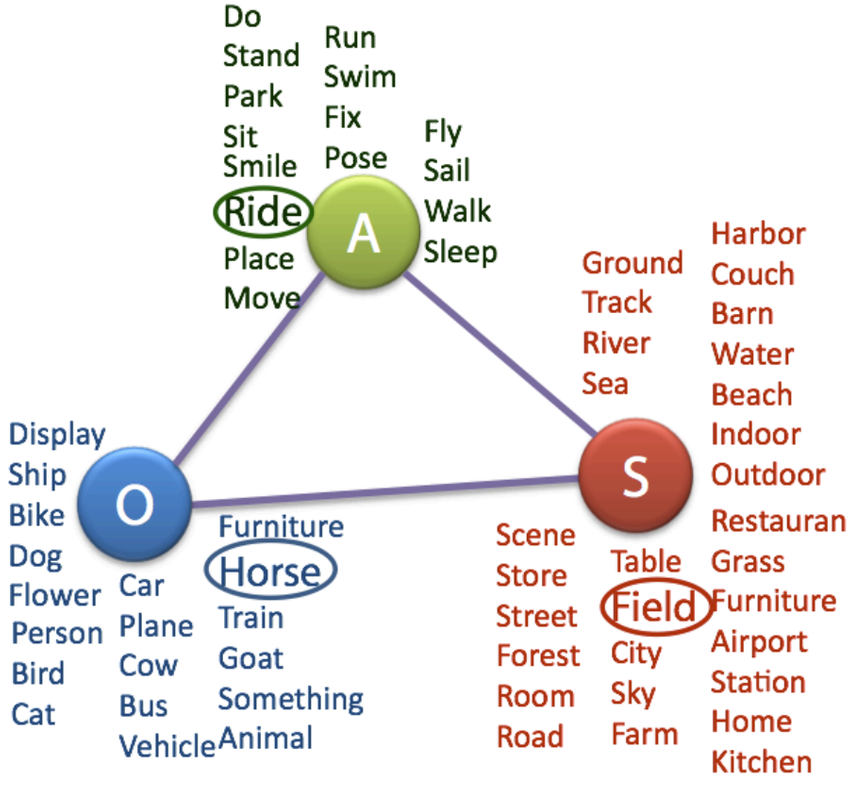


Mapping images to semantics

Discriminative probabilistic model:
P(⟨Object , Action, Scene⟩ | Image)

Markov Random Field:

- Node potentials:
  Based on image features
  (object & scene detectors)
- Edge potentials:
  How often do two labels
  co-occur?



Examples

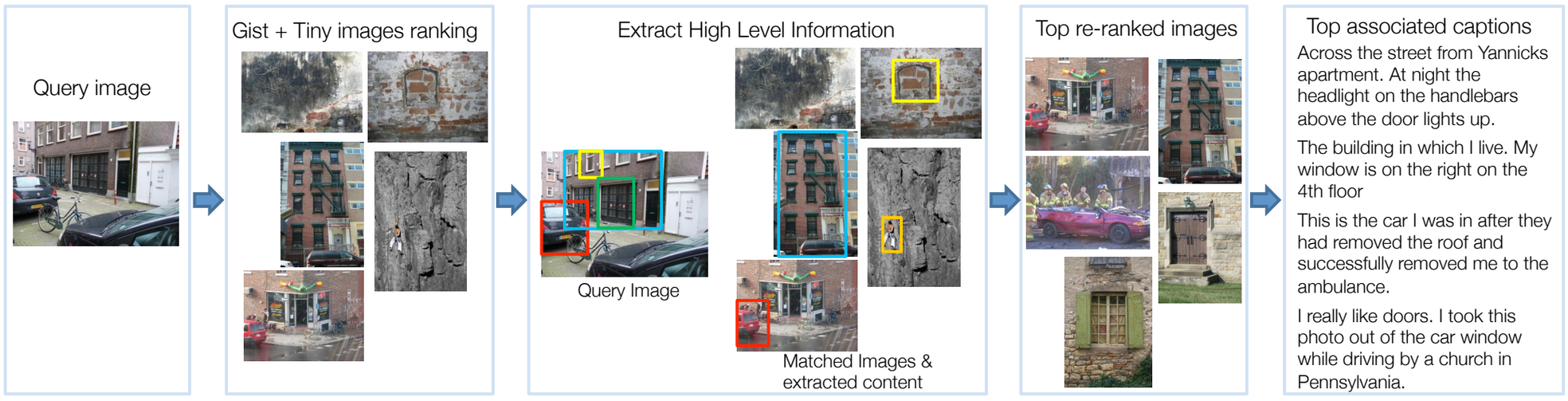
	(pet, sleep, ground) (dog, sleep, ground) (animal, sleep, ground) (animal, stand, ground) (goat, stand, ground)	see something unexpected. Cow in the grassfield. Beautiful scenery surrounds a fluffly sheep. Dog hearding sheep in open terrain. Cattle feeding at a trough.
	(furniture, place, furniture) (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

Challenges for explicit semantic mappings

Scalability:

- Restricted to predefined sets of semantic classes (objects, actions, scenes)
- Requires accurate detectors for each semantic class
  - Performance on Illinois PASCAL data may not be indicative of performance on other data (e.g. Flickr images)

Image description as unimodal retrieval



Ordonez et al. 2011:
Find most similar image among 1M Flickr images (SBU data set), and re-rank their captions.

Advantages of unimodal retrieval

Unimodal image-image similarities are much easier to define than cross-modal image-text similarities.

Can leverage large amounts of human-written captions.

Can be used to provide suggestions for generation-based approaches.

Challenges for retrieval

Ordonez et al.'s approach works because the SBU data set is very large (1 million images):

Query Image | 1k matches | 10k matches | 100k matches | 1million matches

This can only be done on harvested captions, which may not actually describe the images.

Image description as generation

Kulkarni et al.:
This is a picture of three persons, one bottle and one diningtable. The first rusty person is beside the second person. The second person is by the third rusty person. The colorful diningtable is near the first rusty person, and near the second person, and near the third rusty person.
Yang et al.:
Three people are showing th bottle on the street.
Mitchell et al.:
people with a bottle at the table

Kulkarni et al.:
This is a picture of two pottedplants, one dog and one person. The black dog is by the black person, and near the second feathered pottedplant.
Yang et al.:
The person is sitting in the chair in the room.
Mitchell et al.:
A person in black with a black dog by potted plants.

Comparison from Mitchell et al. (EACL 2012)

Image description as a *generation task* with explicit semantics

Yang et al. 2011

Task: Generate sentences for UIUC Pascal images

Templates: NP_{sbj} verb NP_{obj}? prep NP_{scene}

This is a NP

Image features to predict nouns (sbj, obj, scene)

20 object types: Felzenszwalb detector responses

8 scenes: GIST descriptors

Language model to predict verbs and preposition:

Verb: based on NP_{sbj} and NP_{obj}

Preposition: based on verb and NP_{scene} (or NP_{obj})

Kulkarni et al. 2011

Data set: UIUC Pascal images

For each query image:

detect 24 object classes and 6 ‘stuff’ categories
identify 21 attributes of candidate regions (adjectives)
process pairs of candidate regions to get spatial relations (PPs)
Use CRF to predict words for each object, attribute, stuff detection and for each pairwise relation
Use predicted words in a template-based generation system.

Midge (Mitchell et al. 2012)

For each query image:

- detect regions corresponding to objects/stuff with attributes
- detect actions/poses for each region
- detect spatial relations between regions

Each image caption contains:

- nouns + modifiers that refer to objects/stuff + attributes
- verbs that refer to poses/actions
- prepositions that refer to spatial relations between entities

Generation task:

- filter incorrect detections
- augment with syntax-based language model
- impose discourse constraints
- produce fluent caption

Description as Generation

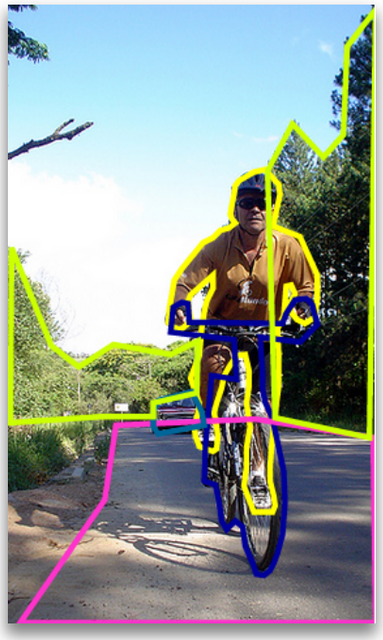
Kulkarni et al.:
This is a picture of three persons, one bottle and one diningtable. The first rusty person is beside the second person. The second person is by the third rusty person. The colorful diningtable is near the first rusty person, and near the second person, and near the third rusty person.
Yang et al.:
Three people are showing the bottle on the street.
Mitchell et al.:
people with a bottle at the table

Kulkarni et al.:
This is a picture of two pottedplants, one dog and one person. The black dog is by the black person, and near the second feathered pottedplant.
Yang et al.:
The person is sitting in the chair in the room.
Mitchell et al.:
A person in black with a black dog by potted plants.

Comparison from Mitchell et al. (EACL 2012)

Visual Dependency Grammar

Elliott & Keller 2013



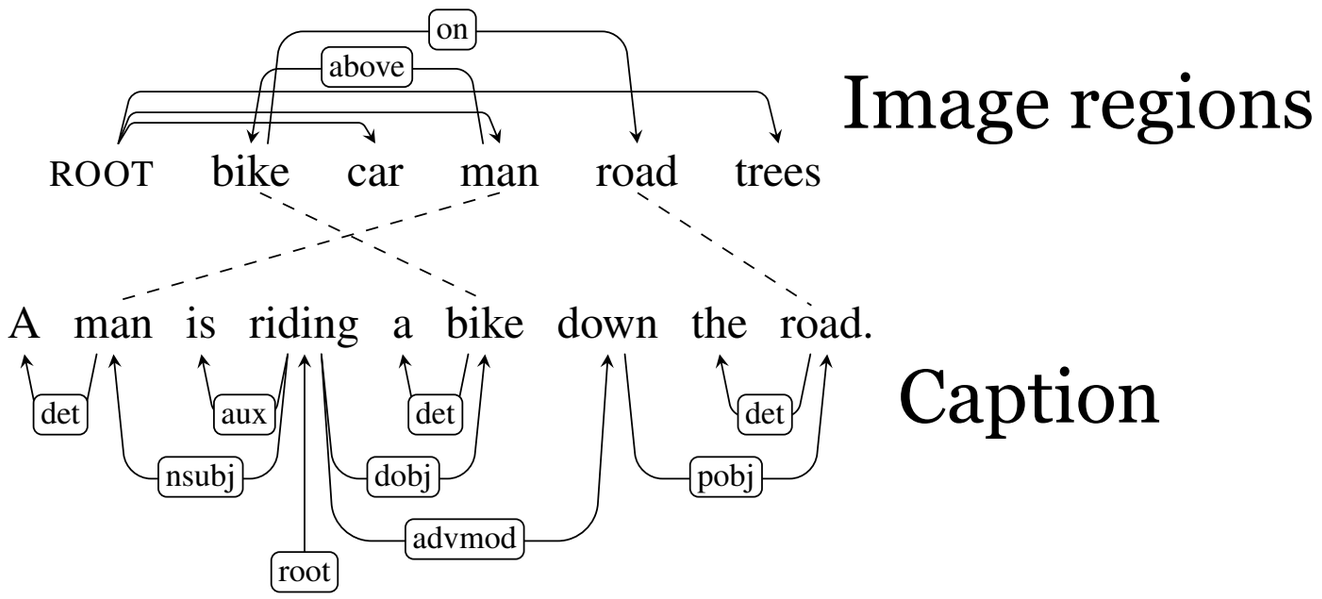


Image regions

Caption

Visual Dependency graph: DAG over image regions

Root = main actor

Edges = spatial relations (*on, surrounds, beside, opposite, above, below, in front of, behind*)

Generated from, and aligned with, image descriptions.

Shown to be beneficial for a template-based caption generation system that has access to gold regions.

Advantages of generation

Ultimately, this is the task we need to solve:
Given an image, produce a (possibly novel) sentence that describes the image well.

Easier to define the scope of what is described explicitly:

- Spatial relations between entities?
- Activities?
- Scenes?
- Attributes?

Can leverage other text resources.

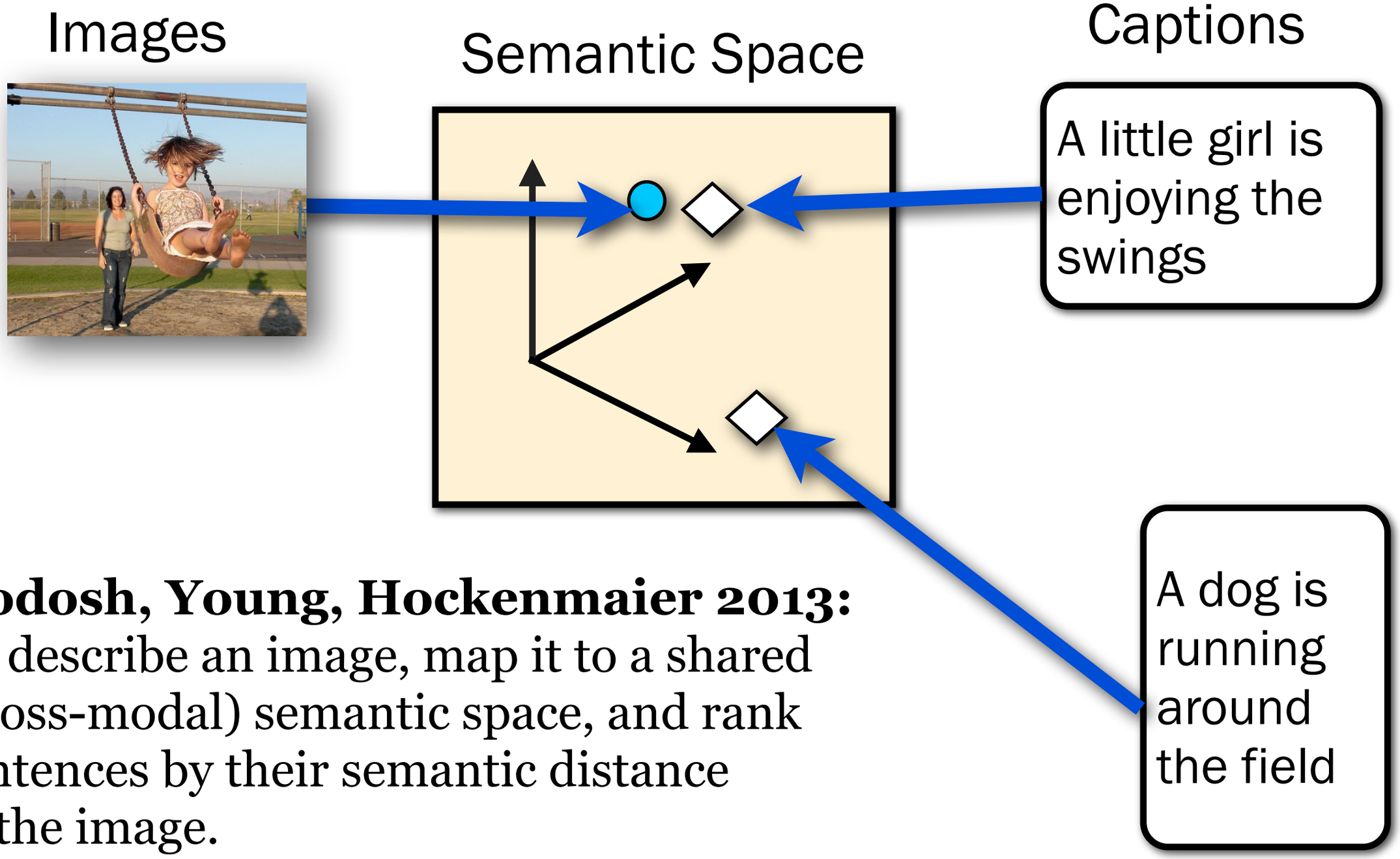
Issues for generation

For image and language understanding, the (truth-conditional) semantic question of whether a sentence describes an image or not is fundamental

Natural Language Generation has additional syntactic and pragmatic aspects that detract from the semantic question:
-Is the caption written in fluent/grammatical English?
-Is this a sentence people would use to describe images?

Natural Language Generation is difficult to evaluate

Image description as cross-modal ranking



Hodosh, Young, Hockenmaier 2013:
To describe an image, map it to a shared (cross-modal) semantic space, and rank sentences by their semantic distance to the image.

Image description as a *transfer* task

Im2Text
Ordonez et al. 2011

Data set: SBU Captioned Photo Dataset
- 1M images harvested from Flickr

Task: Transfer captions from visually similar images
1. Identify *k* visually similar images
2. Estimate image content: objects, stuff, people, scenes
3. Rerank captions of the *k* candidate images

Evaluation:
- Automatic: Bleu scores
- Human: Forced choice between 2 random images per caption

Im2Text: Find candidates

Represent each image as:
- Gist feature
- 'tiny image' (32 x 32 thumbnail)

Compute similarity between query image and each of the 1M images

Global matching:

Return the caption of most similar image

Content matching:

Return top 100 most similar images for further processing

Im2Text: Content matching

Objects (89 categories):

- If the caption mentions an object, run corresponding detector.
- Represent detected objects by shapes and visual attributes.

People and actions:

- Predict action and pose vector

Scenes (23 categories from SUN):

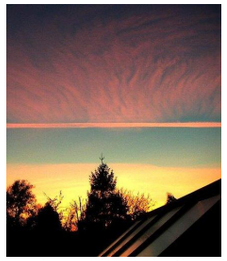
- Train 23 classifiers to predict a scene vector

Stuff (sky, road, building, tree, water)

Compare query image against each candidate image:

- Similarity of the regions corresponding to the detected objects, people, scenes, stuff
- Train classifier over these similarity vectors (to maximize Bleu)

Im2Text Examples



Amazing colours in the sky at sunset with the orange of the cloud and the blue of the sky behind.

A female mallard duck in the lake at Luukki Espoo

Fresh fruit and vegetables at the market in Port Louis Mauritius.

Street dog in Lijiang

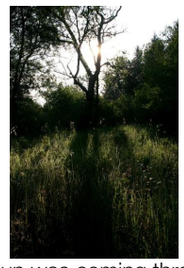
Tree with red leaves in the field in autumn.

One monkey on the tree in the Ourika Valley Morocco

Clock tower against the sky.

The river running through town I cross over this to get to the train

Strange cloud formation literally flowing through the sky like a river in relation to the other clouds out there.

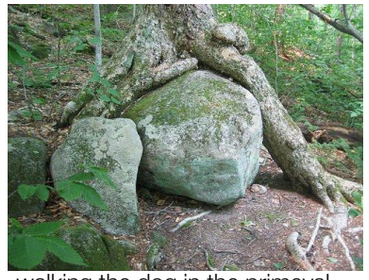


The sun was coming through the trees while I was sitting in my chair by the river

check out the face on the kid in the black hat he looks so enthused

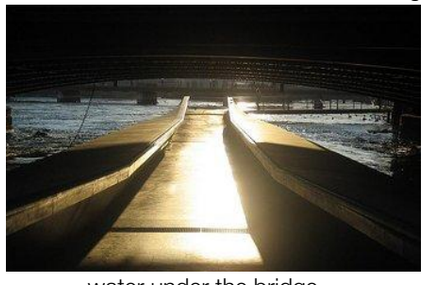
The tower is the highest building in Hong Kong.

the water the boat was in



walking the dog in the primeval forest

shadows in the blue sky



water under the bridge

girl in a box that is a train

small dog in the grass

I tried to cross the street to get in my car but you can see that I failed LOL.

Kuznetsova et al. 2012

ILP: This is a sporty little red convertible made for a great day in Key West FL. This car was in the 4th parade of the apartment buildings.

Human: Hard rock casino exotic car show in June

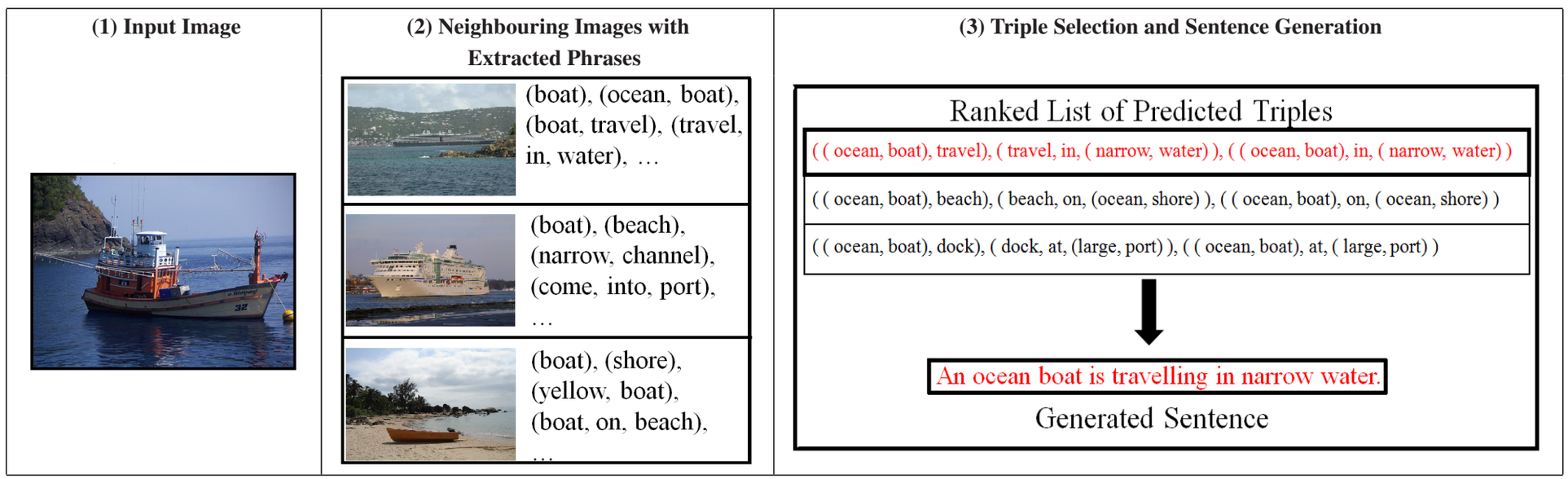
ILP: I like the way the clouds hanging down by the ground in Dupnitsa of Avikwalal.

Human: Car was raised on the wall over a bridge facing traffic..paramedics were attending the driver on the ground

Data: SBU data set, tested on 1,000 selected images with good detector responses

1. Process query image (Similar features to Im2Text)
2. For each detector response:
   - Retrieve images with visually similar responses
   - Transfer corresponding phrases from their captions
3. Generate one sentence per detected object
   - ILP formulation: word order, avoid redundancy, etc.

Gupta et al. 2012



Approach (on UIUC PASCAL data)

Generate caption from word-word dependencies that are transferred from k-nearest neighbor images.

Sentence features:

Word-word dependencies and Google n-gram counts

Image features:

color histograms (RGB, HSV)
texture: Gabor and Haar descriptors
scene (GIST), shape: SIFT

Image description as a *cross-modal ranking* task with implicit semantics

Hodosh, Young, Hockenmaier 2013

Cross-modal image annotation

A little girl is enjoying the swings

Two boys are playing football.

People in a line holding lit roman candles.

A little girl is enjoying the swings

A motorbike is racing around a track.

An elephant is being washed.

Cross-modal image search

A little girl is enjoying the swings

Two boys

People in a line holding lit roman candles.

A little girl is enjoying the swings

A motorbike is racing around a track.

An elephant is being washed.

Description as Ranking

Given a pool of unseen images $\boldsymbol{I}_{\text{test}}$ and unseen sentences $\mathbf{S}_{\text{test}}$, we can use an affinity function f(I, S) that is maximized when S describes I to define image description as two ranking tasks:

Sentence-based image annotation (over a pool of test sentences, $\mathbf{S}_{\text{test}}$):

For each $I_{\text{query}} \in \mathbf{I}_{\text{test}}$, rank all $S \in \mathbf{S}_{\text{test}}$ by f(I_{query}, S)

Sentence-based image search (over a pool of test images, $\mathbf{I}_{\text{test}}$):

For each $S_{\text{query}} \in \mathbf{S}_{\text{test}}$, rank all $I \in \boldsymbol{I}_{\text{test}}$ by f(I, S_{query})

Kernel Canonical Correlation Analysis

Images

$K_i(D_i, *)$ W_1

Semantic Space

W_2 $K_t(D_t, *)$

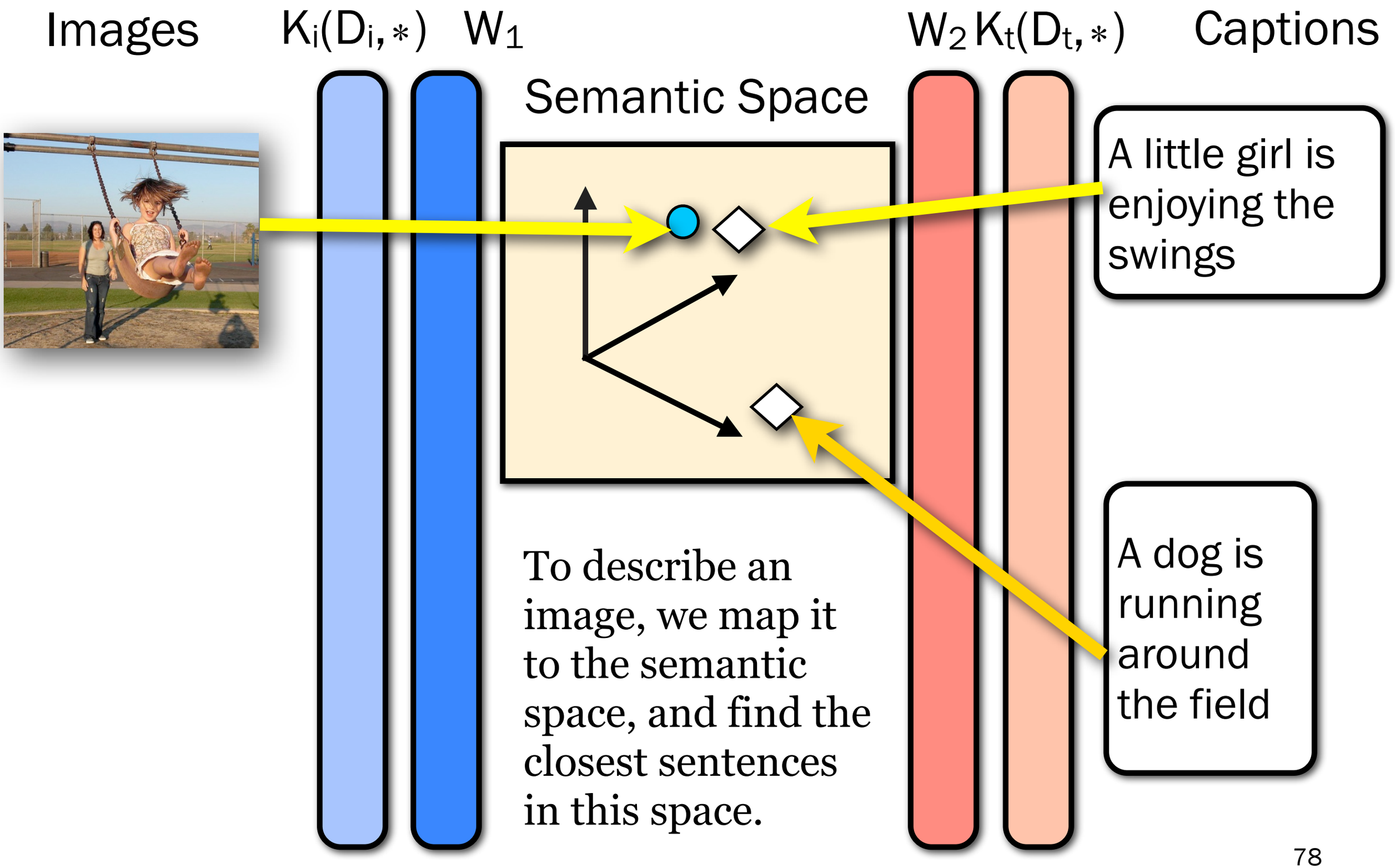
Captions

Sooners football player wears the number 28 and black armbands ...

A boy jumps from one bed to another ...

Images and sentences are mapped to the same induced semantic space.

Using KCCA for image description

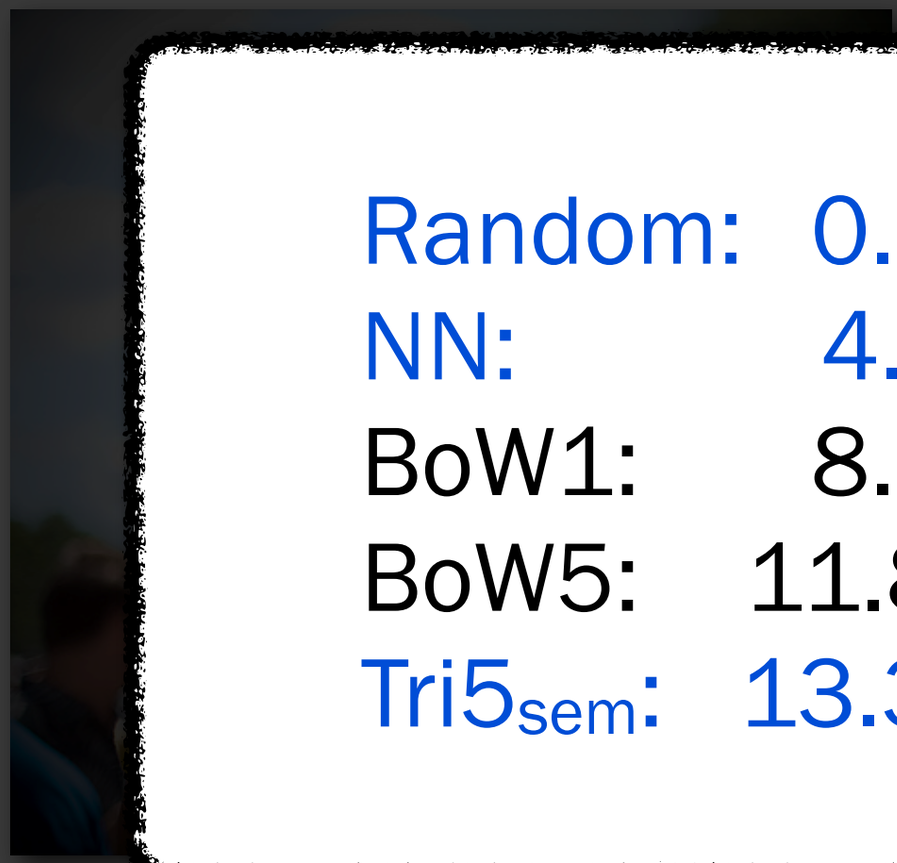


Examples

‘Expert’ human evaluation for image annotation:
Rate the highest ranked test caption for each test image on a scale from 1 to 4.

Search and annotation examples
Shown are the top 5 results per query image/sentence.
The response that belongs to the query is highlighted.

Score: 4 (Perfect description)

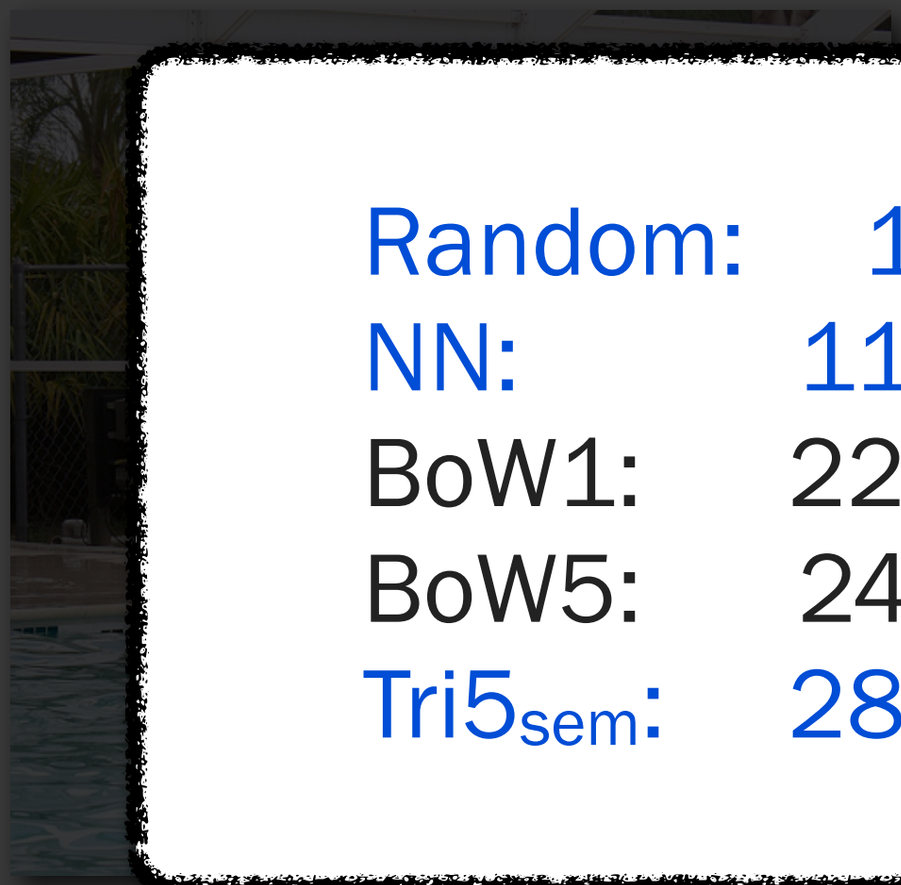


Random: 0.6%
NN: 4.1%
BoW1: 8.1%
BoW5: 11.8%
$Tri5_{sem}$: 13.3%

A girl wearing a yellow shirt and sunglasses smiles.

A man climbs up a sheer wall of ice.

Score: 3 (Minor errors)



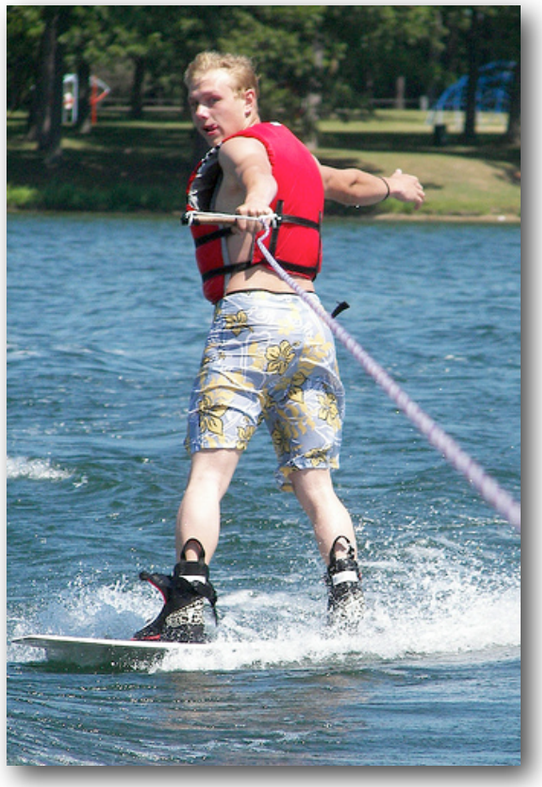
Random:	1.5%
NN:	11.4%
BoW1:	22.9%
BoW5:	24.7%
$Tri5_{sem}$:	28.1%

A boy jumps into the blue pool water.

A child jumping on a tennis court.

Score: 2 (Major errors)

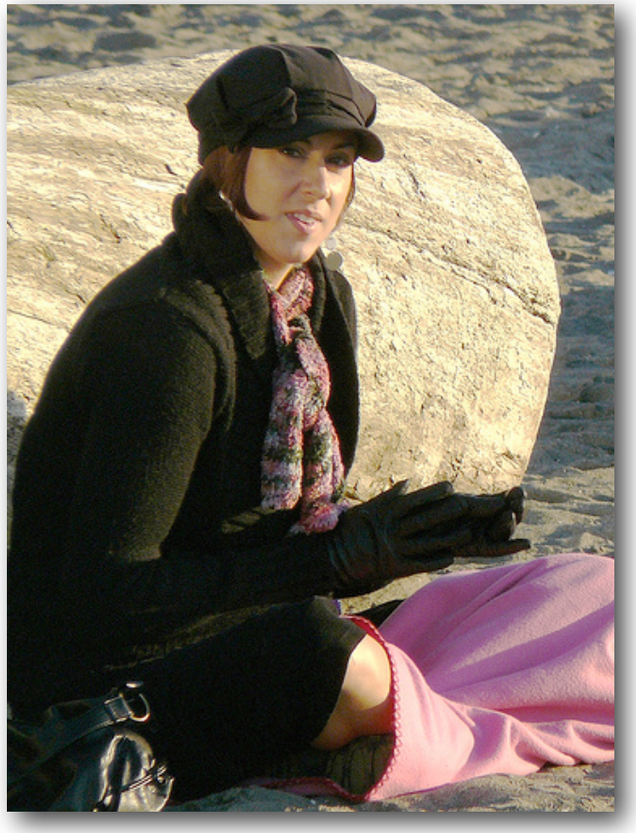
A dog in a grassy field, looking up .



A boy in a blue life jacket jumps into the water .

Score: 1 (Caption unrelated)

Basketball players in action.



A black dog with a purple collar running.

Image search examples

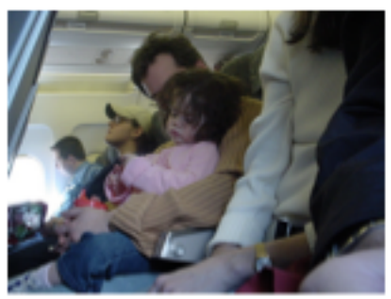
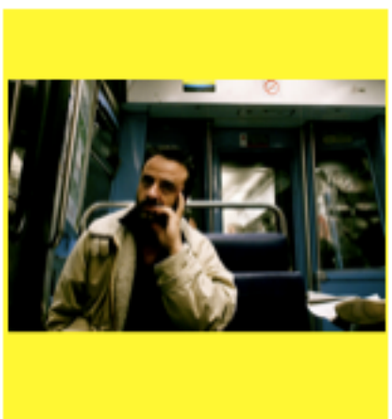
Two little girls practice martial arts

:






A man sitting on a subway



Advantages of cross-modal ranking

Aims to directly measure the quality of the semantic mapping between images and sentences.

- No need to worry about syntax or pragmatics, as in generation.

Makes it possible to define objective benchmarks that allow different systems to be compared directly.

- Everybody needs to rank the same set of sentences/images.

Can be evaluated automatically.

- Automatic evaluations correlate highly with human judgments.

Natural framework for annotation *and* search.

Issues for cross-modal ranking

Requires large(ish) pool of images paired with human-written sentences.

Benefits from additional human relevance judgments.

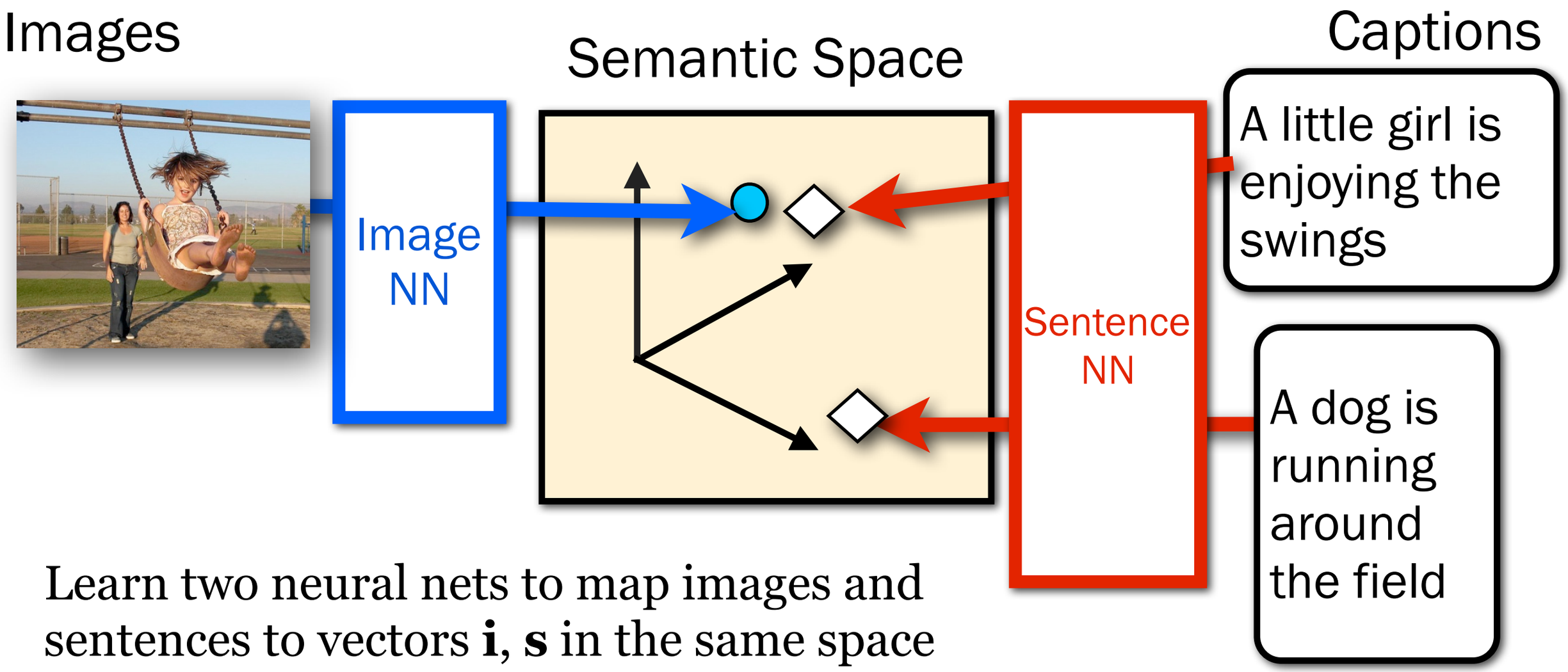
Is performance an artifact of what sentences are available in the pool?

Image annotation examples

Two girls with dark hair and white shirts.
A woman in a red shirt holding a cellphone.
The Asian girl wearing a pink and black striped top is walking next to the girl in the grey top .
A smiling woman embracing a young girl in a jacket with an apple print.
A woman in a headdress is holding a little boy wearing blue.

A person on a dirt bike is riding up a sandy hill.
A man riding a motorbike kicks up dirt .
Two motocross riders next to each other on a dirt track .
A person drives an ATV through mud.
A man wearing a white hat is on a red ATV driving on the dirt .

Using NNs for image description
(Socher et al., TACL 2014)



Learn two neural nets to map images and sentences to vectors **i**, **s** in the same space such that the dot product of correct image-sentence pairs ($\mathbf{i}_i, \mathbf{s}_i$) is greater than that of incorrect pairs ($\mathbf{i}_i, \mathbf{s}_j$), ($\mathbf{i}_j, \mathbf{s}_i$) by a margin Δ:
$\mathbf{i}_i\mathbf{s}_i > \mathbf{i}_i\mathbf{s}_j + \Delta$ and $\mathbf{i}_i\mathbf{s}_i > \mathbf{i}_j\mathbf{s}_i + \Delta$

Evaluating Image Description

Hodosh, Young, Hockenmaier, JAIR 2013

Desiderata for evaluation

Reliability:

- Higher-scoring systems produce better descriptions.
- Human evaluation: high inter-annotator agreement.

Reproducibility and objectivity:

- Repeated evaluation of the same system produces the same scores.
- Different systems can be evaluated independently, and their scores can still be compared directly.

Cost and efficiency:

- Evaluation should be as inexpensive as possible.
- Evaluation should be possible on a large scale.

Evaluating generated image descriptions

Human judgments

- Advantages: Accuracy
- Challenges: Reliability, reproducibility, cost

Automatic comparison to reference captions (BLEU, ROUGE)

- Advantages: Can be done on a large scale
- Challenges: Poor correlation with human judgments

Evaluating ranking-based description systems

Automatic evaluation is possible:
One image in the test pool is the image that the query sentence was written for.

- What is the median rank of that image?
- Recall@K: How often do we retrieve this image as the first result, among the first 5 results, among the first 10 results?

Challenges for evaluating ranking-based systems

The test pool may contain other images for which the test query is a good description.
- Automatic evaluation underestimates performance

We need to collect human judgments for all image-sentence pairs in the test pool.
- Prohibitively expensive to do this exhaustively, but:
- BLEU is a useful filter to identify implausible pairings
- Unlike for NLG, the same judgments can be re-used in evaluating different systems.

Collecting human relevance judgments

Binary task:

Which of these 10 captions describes the image, possibly with minor errors

Crowdsourcing:

On a test set of 1,000 image-caption pairs, we collected 3 judgments for the 10 highest results of 16 models on annotation and search

Preprocessing

Use BLEU (with stemming and stop word removal) to filter out implausible image-caption pairs.
Removes 86% of all pairs, but only 6.7% of good pairs

Using human relevance judgments

Rate of success (S@k):

What percentage of queries has at least one relevant response among the top k results?

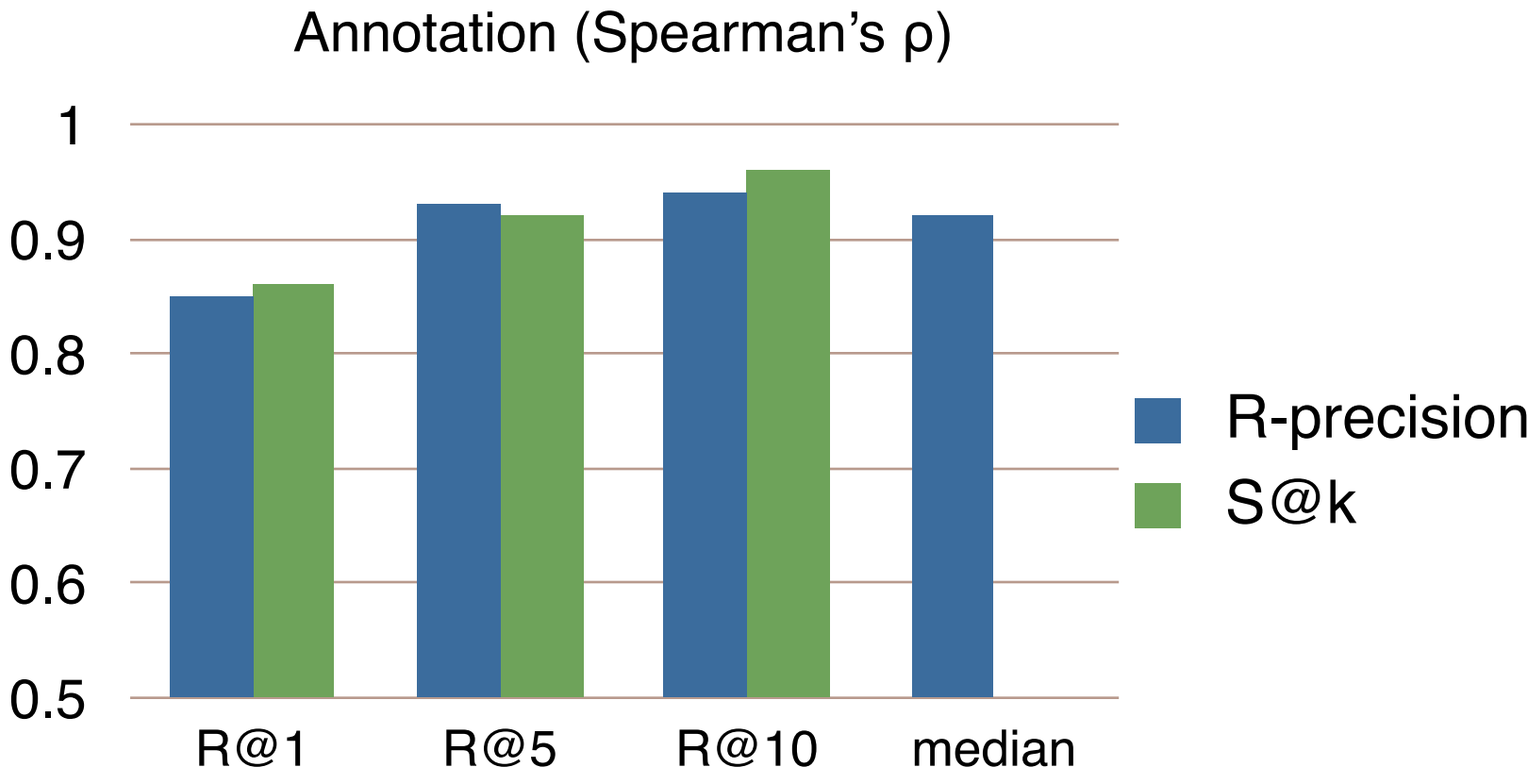
R-precision:

If query q_i has r_i relevant responses, report precision at rank r_i for q_i.

Standard IR metric when the number of relevant responses varies by query.

Can evaluation be automated?

Do automated metrics rank different (n = 30) systems the same way as human metrics?



Possibly, if you go beyond the first result

A proposal for a shared task

Why a shared task?

We need objective evaluation metrics and benchmarks to compare different systems directly.

(cf. CoNLL shared tasks, PASCAL VOC challenges, ImageNet)

We need reproducible experiments and evaluations.

No repeated human evaluation should be required.

We need data sets that both vision and language people can use.

We want to lower the barrier of entry for people from both fields.

Ranking-based image description

Test data:

A large pool of unseen images and their captions

Image annotation task:

For each test image, rank the test captions

Image search task:

For each test caption, rank the test images

Evaluation:

Without relevance judgments: median rank of gold item, R@k
With (exhaustive) relevance judgments: R-precision, S@k

Data sets

Hodosh et al. 2013 (JAIR):
Illinois Flickr 8k as benchmark for ranking-based image annotation & search

- Training set: 6,000 Flickr images with 5 crowdsourced captions
- Test set: 1,000 Flickr images with 1 crowdsourced caption each (+human relevance judgments)

Illinois Flickr 30k data set (Young et al., TACL, 2014):

- Mostly Creative-Commons licensed images

Alternative: IAPR-TC12

- (20,000 segmented images with captions)

REFERENCES

References

Image description in general
Jaimes, A., Jaimes, R., & Chang, S.-F. (2000). A conceptual framework for indexing visual information at multiple levels. In Internet Imaging 2000, Vol. 3964 of Proceedings of SPIE, pp. 2–15, San Jose, CA, USA.
Shatford, S. (1986). Analyzing the subject of a picture: A theoretical approach. Cataloging & Classification Quarterly, 6, 39–62.
Data sets (Images & sentences)
Everingham, M., Gool, L. V., Williams, C., Winn, J., & Zisserman, A. (2008). The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/.
Grubinger, M., Clough, P., Müller, H., & Deselaers, T. (2006). The IAPR benchmark: A new evaluation resource for visual information systems. In OntoImage 2006, LREC Workshop, pp. 13–23.
Hugo Jair Escalante, Carlos Hernandez, Jesus A. Gonzalez, Aurelio Lopez, Manuel Montes, Eduardo Morales, Enrique Sucar, Luis Villasenor, and Michael Grubinger (2010) The Segmented and Annotated IAPR-TC12 Benchmark. Computer Vision and Image Understanding Journal, 114(4):419–428, 2010.
Rashtchian, C., Young, P., Hodosh, M., & Hockenmaier, J. (2010). Collecting image annotations using Amazon's Mechanical Turk. In NAACL Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk, pp. 139–147.
Feng, Y., & Lapata, M. (2010). How many words is a picture worth? Automatic caption generation for news images. ACL, pp. 1239–1249.
Ordonez, V., Kulkarni, G., & Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. NIPS 24, pp. 1143–1151.
Hodosh, M., Young, P. & Hockenmaier J. (2013). Framing image description as a ranking task: data, models, and evaluation metrics, Journal of Artificial Intelligence Research, Volume 47, pages 853-899
Young, P., Lai, A., Hodosh M & Julia Hockenmaier J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL 2014.

References

Sentence-based image description

Feng, Y., & Lapata, M. (2008). Automatic image annotation using auxiliary text information ACL-08: HLT, pp. 272–280.

Feng, Y., & Lapata, M. (2010). How many words is a picture worth? Automatic caption generation for news images. ACL, pp. 1239–1249.

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. ECCV Part IV, pp. 15–29

Kuznetsova, P., Ordonez, V., Berg, A., Berg, T., & Choi, Y. (2012). Collective generation of natural image descriptions. 50th ACL (Volume 1: Long Papers), pp. 359–368.

Li, S., Kulkarni, G., Berg, T. L., Berg, A. C., & Choi, Y. (2011). Composing simple image descriptions using web-scale n-grams. CoNLL, pp. 220–228.

Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., & Berg, T. L. (2011). Baby talk: Understanding and generating simple image descriptions. CVPR, pp. 1601–1608.

Mitchell, M., Dodge, J., Goyal, A., Yamaguchi, K., Stratos, K., Han, X., Mensch, A., Berg, A., Berg, T., & Daume III, H. (2012). Midge: Generating image descriptions from computer vision detections. 13th EACL, pp. 747–756.

Yang, Y., Teo, C., Daume III, H., & Aloimonos, Y. (2011). Corpus-guided sentence generation of natural images. EMNLP, pp. 444–454

Gupta, A., Verma, Y., & Jawahar, C. (2012). Choosing linguistics over vision to describe images. AAAI.

Ordonez, V., Kulkarni, G., & Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. NIPS 24, pp. 1143–1151.

Elliott, D. and Keller F. 2013. Image Description Generation from Structured Image Representations. EMNLP, pp. 1292-1302.

Hodosh, M., Young, P. & Hockenmaier J. (2013). Framing image description as a ranking task: data, models, and evaluation metrics, Journal of Artificial Intelligence Research, Volume 47, pages 853-899

References

Sentence-based image description

Feng, Y., & Lapata, M. (2008). Automatic image annotation using auxiliary text information ACL-08: HLT, pp. 272–280.

Feng, Y., & Lapata, M. (2010). How many words is a picture worth? Automatic caption generation for news images. ACL, pp. 1239–1249.

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. ECCV Part IV, pp. 15–29

Kuznetsova, P., Ordonez, V., Berg, A., Berg, T., & Choi, Y. (2012). Collective generation of natural image descriptions. 50th ACL (Volume 1: Long Papers), pp. 359–368.

Li, S., Kulkarni, G., Berg, T. L., Berg, A. C., & Choi, Y. (2011). Composing simple image descriptions using web-scale n-grams. CoNLL, pp. 220–228.

Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., & Berg, T. L. (2011). Baby talk: Understanding and generating simple image descriptions. CVPR, pp. 1601–1608.

Mitchell, M., Dodge, J., Goyal, A., Yamaguchi, K., Stratos, K., Han, X., Mensch, A., Berg, A., Berg, T., & Daume III, H. (2012). Midge: Generating image descriptions from computer vision detections. 13th EACL, pp. 747–756.

Yang, Y., Teo, C., Daume III, H., & Aloimonos, Y. (2011). Corpus-guided sentence generation of natural images. EMNLP, pp. 444–454

Gupta, A., Verma, Y., & Jawahar, C. (2012). Choosing linguistics over vision to describe images. AAAI.

Ordonez, V., Kulkarni, G., & Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. NIPS 24, pp. 1143–1151.

Elliott, D. and Keller F. 2013. Image Description Generation from Structured Image Representations. EMNLP, pp. 1292-1302.

Hodosh, M., Young, P. & Hockenmaier J. (2013). Framing image description as a ranking task: data, models, and evaluation metrics, Journal of Artificial Intelligence Research, Volume 47, pages 853-899

References

Words and pictures
P. Duygulu, K. Barnard, N. de Freitas, D. Forsyth, Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary,_ECCV, pp IV:97-112, 2002 http://kobus.ca/research/data/eccv_2002/
Barnard, K., Duygulu, P., Forsyth, D., Freitas, N. D., Blei, D. M., & Jordan, M. I. (2003). Matching words and pictures. JMLR, 3, 1107–1135.
Blei, D. M., & Jordan, M. I. (2003). Modeling annotated data. In SIGIR 2003, pp. 127–134.
Grangier, D., & Bengio, S. (2008). A discriminative kernel-based approach to rank images from text queries. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30, 1371–1384.
Hardoon, D. R., Saunders, C., Szedmak, S., & Shawe-Taylor, J. (2006). A correlation approach for automatic image annotation. In Li, X., Zaïane, O. R., & Li, Z.-H. (Eds.), Advanced Data Mining and Applications, Vol. 4093 of Lecture Notes in Computer Science, pp. 681–692. Springer Berlin Heidelberg.
Socher, R., & Li, F.-F. (2010). Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. CVPR, pp. 966–973.
Weston, J., Bengio, S., & Usunier, N. (2010). Large scale image annotation: learning to rank with joint word-image embeddings. Machine Learning, 81(1), 21–35.
Hardoon, D. R., Szedmak, S. R., & Shawe-Taylor, J. R. (2004). Canonical correlation analysis: An overview with application to learning methods. Neural Computation, 16, 2639–2664.
Hotelling, H. (1936). Relations between two sets of variates. Biometrika, 28(3/4), 321–377.
Hwang, S., & Grauman, K. (2012). Learning the relative importance of objects from tagged images for retrieval and cross-modal search. International Journal of Computer Vision, 100(2), 134–153.
Makadia, A., Pavlovic, V., & Kumar, S. (2010). Baselines for image annotation. International Journal of Computer Vision, 90(1), 88–105.
Deschacht, K., & Moens, M.-F. (2007). Text analysis for automatic image annotation. 45th ACL, pp. 1000–1007

References

Data sets (Computer vision in general)
Everingham, M., Gool, L. V., Williams, C., Winn, J., & Zisserman, A. (2008). The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/.
J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database. IEEE Computer Vision and Pattern Recognition (CVPR), 2009
A. Berg, J. Deng, and L. Fei-Fei. ImageNet large scale visual recognition challenge 2010 http://www.image-net.org/challenges/lsvrc/2010/, 2010.
J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba (2010). SUN Database: Large-scale Scene Recognition from Abbey to Zoo. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
B. Russell, A. Torralba, K. Murphy, W. T. Freeman. LabelMe: a database and web-based tool for image annotation, IJCV 2007

Other image & sentence papers
P. Kordjamshidi, M. Van Otterlo, M. Moens (2010) Spatial Role Labeling: Task Definition and Annotation Scheme, LREC'10.
Deschacht, K., & Moens, M.-F. (2007). Text analysis for automatic image annotation. 45th ACL, pp. 1000–1007

References

Computer Vision

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2), 91–110

Varma, M., & Zisserman, A. (2005). A statistical approach to texture classification from single images. International Journal of Computer Vision, 62, 61–81.

Vedaldi, A., & Fulkerson, B. (2008). VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/.

Lazebnik, S., Schmid, C., & Ponce, J. (2009). Spatial pyramid matching. In S. Dickinson, A. Leonardis, B. S., & Tarr, M. (Eds.), Object Categorization: Computer and Human Vision Perspectives, chap. 21, pp. 401–415. Cambridge University Press.

Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. In Proceedings of the 2008 IEEE Conference on Com- puter Vision and Pattern Recognition (CVPR), pp. 1–8, Anchorage, AK, USA.

Li-Jia Li, Hao Su, Eric P. Xing and Li Fei-Fei (2010) Object Bank: A High-Level Image Representation for Scene Classification and Semantic Feature Sparsification. NIPS 2010.

Li-Jia Li, Hao Su* Yongwhan Lim and Li Fei-Fei (2010). Objects as Attributes for Scene Classification. ECCV Workshop on Parts and Attributes

A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, International Journal of Computer Vision, Vol. 42(3): 145-175, 2001

Image retrieval

Vasconcelos, N (2007). From Pixels to Semantic Spaces: Advances in Content-Based Image retrieval. IEEE Computer.

Popescu, A., Tsikrika, T., & Kludas, J. (2010). Overview of the Wikipedia retrieval task at ImageCLEF 2010. In CLEF (Notebook Papers/LABs/Workshops), Padua, Italy.