Describing Images in Natural Language Part II

EACL 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
- Visual features for image description
- Image description systems
- Image description and semantics

Visual Features for Image Description

Low-level features

Can be computed directly off of the image:

- Color
- Texture
- Key points/SIFT
- Scene descriptors (Gist)

"Bag-of-words" representation:
Real or vector-valued features are often quantized (e.g. by k-means clustering), and represented as a histogram of discrete values.

Color

Each pixel is represented as a vector in a color space:
- RGB: **r**ed-**g**reen-**b**lue
- HSV: **h**ue-**s**aturation-brightness **v**alue
- (CIE)LAB: designed to approximate human vision

Color features capture properties of the distribution of colors in an image or image region:
- Moments (mean, standard deviation, skewness)
- Histogram of quantized features

Texture

Texture is a property of an image patch.
Used to identify materials/stuff.

"Textons":
pass each image through a set (bank) of filters
cluster the responses into a vocabulary of texture words
represent image patches as histograms of filter responses
http://www.robots.ox.ac.uk/~vgg/research/texclass/

Gist descriptors

Capture dominant spatial structure of a scene along perceptual dimensions:
- Naturalness (natural or man-made environment?)
- Openness (coast, highway vs forest, city)
- Roughness (size of major components),
- Expansion (property of vanishing lines)
- Ruggedness (deviation of ground from horizon)

Yields low-dimensional descriptor (feature vector) of the entire image (*Spatial Envelope*)

A. Oliva & A. Torralba, 2001.
http://people.csail.mit.edu/torralba/code/spatialenvelope/

SIFT descriptors

Scale Invariant Feature Transform (Lowe 1999, 2004), developed as descriptor (feature vector) for key points:
- Invariant to translation, rotation, rescaling
- Robust to perspective and illumination changes
- (color: 3×)128-dimensions; computed over 4x4 patch

Sparse SIFT: applied to key points only
- useful for object matching across images

Dense SIFT: applied over a dense grid
- useful for object/scene classification
- vectors are clustered into a fixed number of 'words', represented as a histogram of discrete values

http://www.scholarpedia.org/article/SIFT

Region-based features

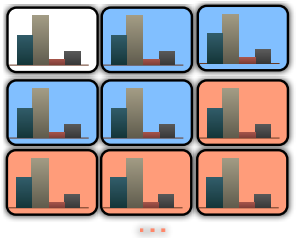
Fixed grid regions:
- easy to compute
- not scale and rotation invariant
- regions are not semantically or visually coherent

Regions based on image segmentation:
- variable number of regions/image
- objects may be over-segmented (or under-segmented)

Regions based on detector responses:
- should identify similar instances of an object class
- requires accurate detectors

Spatial Pyramids

Feature vector



Also used in kernels: Lazebnik et al. 2009

Objects

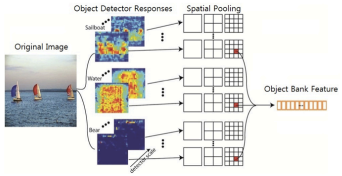
Object classification:
Does this image contain an object of category C?

Object detection (localization):
Find bounding boxes that depict objects of category C.

PASCAL Visual Object Classes (VOC) Challenges:
(2005-2012) 20 categories, 22.5K images:
Aeroplanes, Bicycles, Birds, Boats, Bottles, Buses, Cars, Cats, Chairs, Cows, Dining tables, Dogs, Horses, Motorbikes, People, Potted plants, Sheep, Sofas, Trains, TV/Monitors
ImageNet Large Scale Visual Recognition Challenge
(2010-...) 200-1000 object categories, 500K-1.5M images

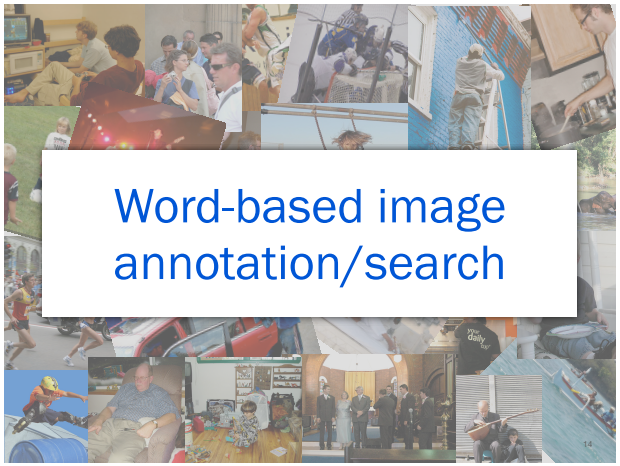
Detector responses

Use directly (to identify bounding boxes):
Provides localization, if detection is accurate
Use indirectly (as features):
Provides useful signal, even when not that accurate at detection
ObjectBank: http://vision.stanford.edu/projects/objectbank/



Scene recognition

SUN (Scene UNderstanding) database
900 scene categories, 3800 object categories, 131K images
http://groups.csail.mit.edu/vision/SUN/



Word-based image annotation/search

Images ↔ Words

Image tagging & search:
Annotate images with lists of keywords
Search for images by lists of keywords
(aka. Content-Based Image Retrieval)
(Figures from Duygulu et al. 2002, Blei & Jordan 2003)

Region-based image annotation:
Annotate image regions with keywords

What do these words describe?
Object classes (*person, tree*) or instances (*Marilyn Monroe*)
Scenes (*market, crowd*)
'Stuff' (*grass, water, sky*)

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 MT
Duygulu et al. 2002

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

Sentence-based image description

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.

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

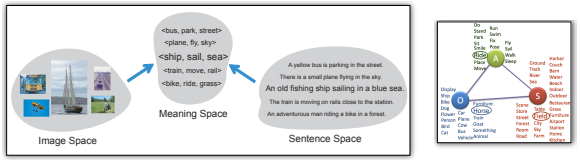
Generate captions directly from image features:
- Requires an explicit mapping between image & text.
- Requires a surface realization model to guarantee fluency etc.
- Requires human evaluation of correctness & grammaticality.

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



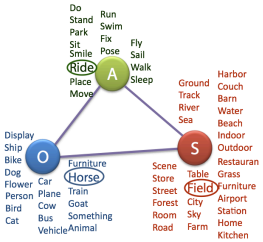
Farhadi et al. 2010
- Semantics of images/sentences:≈ ⟨Object, Action, Scene⟩
- Use Markov Random Fields to predict most likely meaning triplet for images and sentences.

Extended by Yang et al. 2011; Kulkarni et al, 2011, etc. for generation

Naive image semantics

Represent an image as ⟨Object , Action, Scene⟩

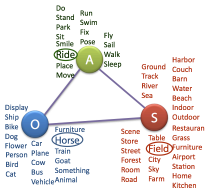
Assume a fixed type inventory for each slot:
- 23 objects
- 16 actions
- 9 scenes



Mapping images to semantics

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

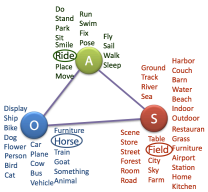
Markov Random Field:
- Node potentials:
  Based on image features (object & scene detectors)
- Edge potentials:
  How often do two labels co-occur?



Mapping sentences to semantics

P(⟨Object , Action, Scene⟩ | Sentence)

Node potentials based on similarity of subject/object, verb, and arguments of prepositions to node label



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

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

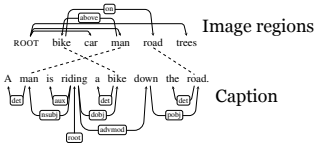
Description as Generation

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



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.

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.

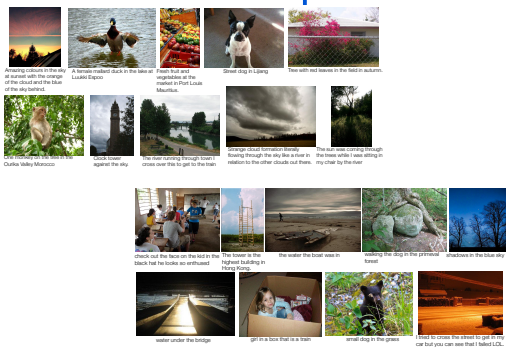
People and actions:
- Predict action and pose vector

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



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

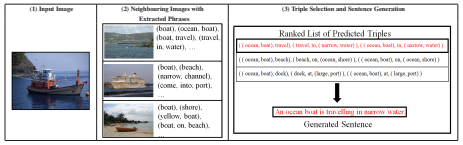
ILP: I like the way the clouds hanging down by the ground in Dupnitsa of Avikwalal.

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

Image description as ranking

Hodosh, Young, Hockenmaier 2013

How well can we associate images with sentences? (*without detectors*)

Tasks:
- Sentence-based image search
- Sentence-based image annotation

Approach:
- Use Kernel Canonical Correlation Analysis (KCCA) to induce a joint semantic space of images and sentences.

In-depth study of evaluation metrics

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})

Image search

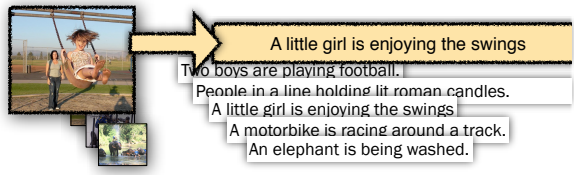
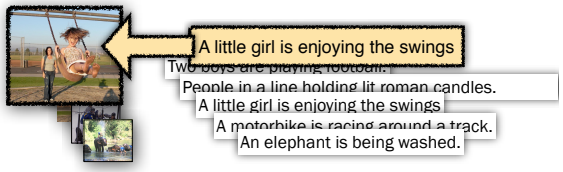


Image annotation



Canonical Correlation Analysis

Input: Pairs of items (A_i, B_i) drawn from different spaces $A_i \in \mathbf{A}$, $B_i \in \mathbf{B}$

Output: maximally correlated linear projections $\mathbf{w}_A$, $\mathbf{w}_B$ that project items from $\mathbf{A}$, $\mathbf{B}$ into an induced common space such that A_i is close to B_i

$$\text{argmax}_{\mathbf{w}_{\mathcal{A}}, \mathbf{w}_{\mathcal{B}}} \frac{\langle \mathbf{A}\mathbf{w}_{\mathcal{A}}, \mathbf{B}\mathbf{w}_{\mathcal{B}} \rangle}{\|\mathbf{A}\mathbf{w}_{\mathcal{A}}\| \|\mathbf{B}\mathbf{w}_{\mathcal{B}}\|}$$

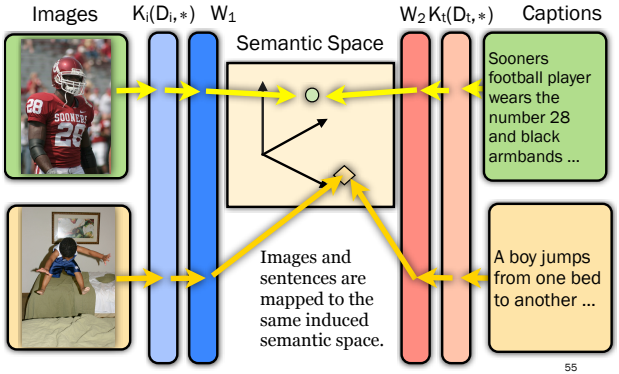
Kernel CCA

KCCA learns projection weights that maximize the correlation between the kernel matrices K_A, K_B

Kernel matrix $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = \langle \varphi(\mathbf{x}_i), \varphi(\mathbf{x}_j) \rangle$:
- Contains implicit dot product of training items $\mathbf{x}_i$ $\mathbf{x}_j$ in a high-dimensional space $\varphi(\mathbf{x}_i)$

Kernels:
- Primal representation of classifiers: learn feature weights
- Equivalent dual representation: weight vector is a linear combination of training examples
- Kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$ computes high-dimensional dot product $\langle \varphi(\mathbf{x}_i), \varphi(\mathbf{x}_j) \rangle$ efficiently

Kernel Canonical Correlation Analysis



Using KCCA for image description

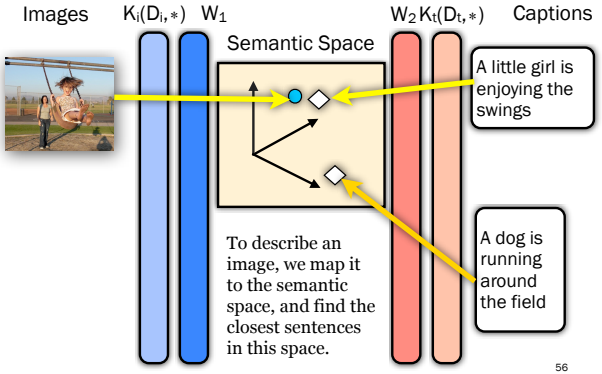
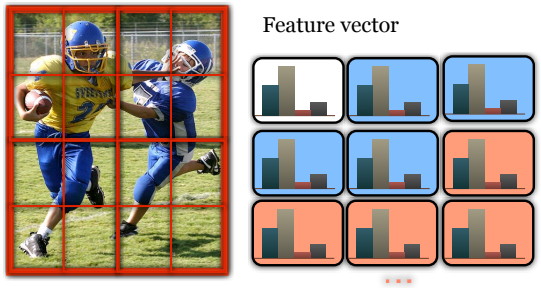


Image Kernels



Spatial Pyramid Kernel (Lazebnik at al.)
Features: histogram of color, texture, and SIFT responses

Text kernels

Bag-of-Word (BoW):
How many words are shared?

Trigram (Tri_{sem}, truncated string kernel):
How many words, and bigram/trigram sequences (with gaps) are shared?
Augmented with lexical similarities: allow partial matches
IDF reweighting: downweight common words

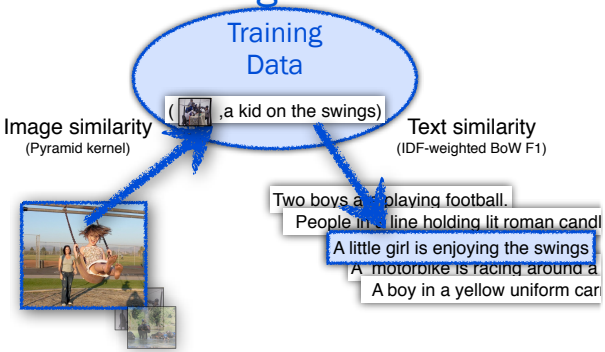
A child with red hair playing with a brown dog

A small child playing with a large dog on the carpet

Lexical similarities

	Alignment	$Dist_{captions}$	$Dist_{BNC}$	Lin
rider	biker	bike	ride	traveler
	bicyclist	dirt	horse	cyclist
	cyclist	motocross	race	bicyclist
	bmx	motocycle	bike	horseman
	bicycler	ride	jockey	jockey
swim	retrieve	pool	fish	bathe
	paddle	trunk	water	sport
	dive	water	sea	football
	come	dive	pool	activity
	wade	goggles	beach	soccer

Nearest Neighbor baseline



Our experiments

Training data: 6,000 image-caption pairs
('caption' = 1 caption or 5 captions)

Test data (disjoint from training data):
1,000 images and 1,000 captions

Tasks:
- **Image annotation:**
  Given a test image, rank all test captions
- **Image search:**
  Given a test caption, rank all test images

Our models

Baseline models
Random (Random)
Nearest Neighbor (NN; 5 captions/training image)

KCCA-based models
Bag of Words (BoW1; 1 captions/training image)
Bag of Words (BoW5; 5 captions/training image)
'Semantic' ($Tri5_{sem}$; 5 captions/training image)

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.

64

Score: 3 (Minor errors)

Random: 1.5%
NN: 11.4%
BoW1: 22.9%
BoW5: 24.7%
Tri5$_{sem}$: 28.1%

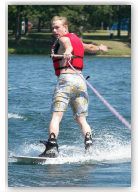
A boy jumps into the blue pool water.

A child jumping on a tennis court.

65

Score: 2 (Major errors)

A dog in a grassy field, looking up .



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

66

Score: 1 (Caption unrelated)

Basketball players in action.

A black dog with a purple collar running.

67

Image search examples

Two little girls practice martial arts

A man sitting on a subway

68

Ranking-based evaluation

Recall of original item (automatic)

	Image annotation			Image search		
	R@1	R@5	R@10	R@1	R@5	R@10
NN	2.5***	7.6***	9.7***	2.5***	4.7***	7.2***
BoW1	4.8***	13.5***	19.7***	4.5***	14.3***	20.8***
BoW5	6.2**	17.1***	24.3***	5.8**	16.7***	23.6***
Tri5$_{Sem}$	8.3	21.6	30.3	7.6	20.7	30.1

Rate of Sucsces (large-scale human evaluation)

	Image annotation			Image search		
	S@1	S@5	S@10	S@1	S@5	S@10
NN	5.8***	15.3***	20.1***	4.9***	12.9***	18.1***
BoW1	12.2***	30.3***	39.7***	11.4***	30.5***	40.2***
BoW5	15.0*	34.1**	42.7***	12.1***	31.5***	40.8***
Tri5$_{Sem}$	16.6	37.7	49.1	15.7	36.9	48.5

69

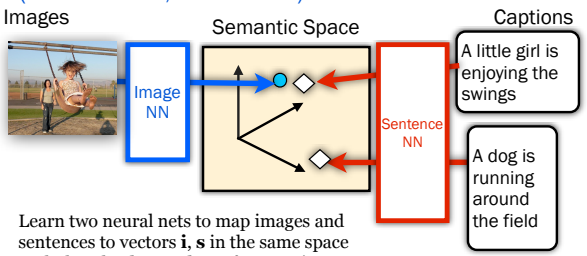
Image annotation examples

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 .

70

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 $\mathbf{i}_i\mathbf{s}_i > \mathbf{i}_j\mathbf{s}_i + \Delta$

71

72

KCCA: Semantics

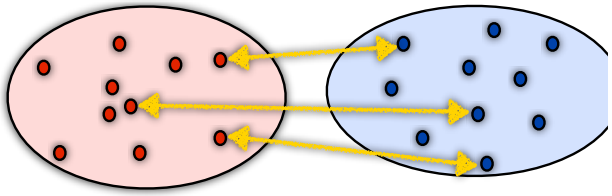
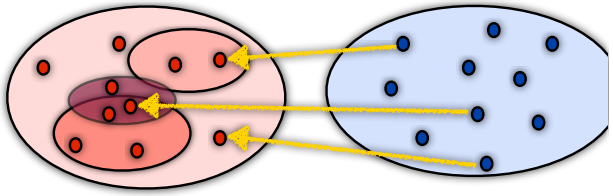
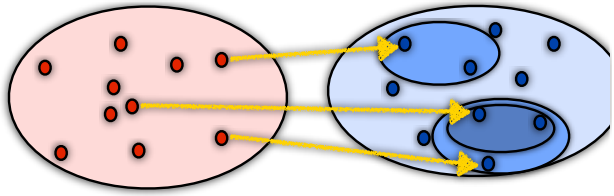


Image semantics



Denotational Semantics



Denotational Semantics

The denotation of a (declarative) sentence is the set of all possible worlds/situations in which it is true:

⟦s⟧ = {w ∈ U: s is true in w }

The *visual* denotation of a (descriptive) sentence is the set of all images for which it is a correct description:

⟦s⟧ = {i ∈ I: s describes (part of) i }

Young, Lai, Hodosh, Hockenmaier, TACL 2014.

Denotation Graph

Denotations induce a partial ordering over descriptions.

⟦a white dog runs on the beach ⟧ ⊂ ⟦a dog runs⟧

This yields a subsumption hierarchy/lattice over image descriptions

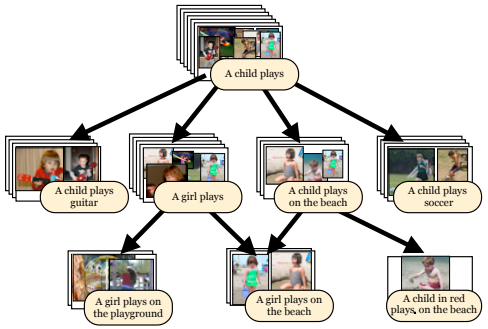
Constructing the graph

1. Normalize captions:
Spelling; capitalization
Lemmatization
Normalize determiners

2. Make captions more generic:
Replace nouns by hypernyms
Drop modifiers (adjectives, adverbs, PPs)

3. Extract simpler constituents

The denotation graph



Statistics

Original data (~32,000 images)
~160K distinct captions

Denotation graph:
~1750K distinct captions:
~230K captions with ⟦s⟧ ≥ 2
~53K captions with ⟦s⟧ ≥ 5
~22K captions with ⟦s⟧ ≥ 10
~1.9K captions with ⟦s⟧ ≥ 100
161 captions with ⟦s⟧ ≥ 1000
e.g. *person play instrument*, *woman standing*, ...

Explicit semantic representations?

Should image description be mediated by explicit semantic representations?
Linguists are developing semantic representations for spatial information, e.g.:
Spatial role labeling (Kordjamshidi et al. 2010)
ISO-Space annotation scheme (Pustejovsky & Yochum 2014)

REFERENCES

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.