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: red-green-blue
- HSV: hue-saturation-brightness value
- (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.

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)

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
1. Introduction

The fields of computer vision and cognitive science are closely related. Whereas the fields of computer vision and cognitive science are closely related, scene understanding research has lagged.

Scene recognition

SUN (Scene UNderstanding) database

Word-based image annotation/search

Challenges:
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 crowd-sourced labeling of regions http://labelme.csail.mit.edu


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

Task: Annotate image regions with keywords (tags)
Model: IBM-style alignment
map each image region to a visual vocabulary of 500 ‘blobs’
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

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.

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-t NN images.
Outperforms learning-based methods

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
label the image

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.

Naive image semantics

Represent an image as (Object, Action, Scene)
Assume a fixed type inventory for each slot:
- 23 objects
- 16 actions
- 9 scenes

Examples

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

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 images & sentences to an explicit semantic space

Farhadi et al. 2010

Semantics of images/sentences: \( \langle \text{Object}, \text{Action}, \text{Scene} \rangle \)

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

Mapping sentences to semantics

\( P((\text{Object}, \text{Action}, \text{Scene}) | \text{Sentence}) \)

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

Yang et al. 2011

Task: Generate sentences for UIUC Pascal images

Templates: \( \text{NP}_{\text{obj}} \text{ verb NP}_{\text{obj}}? \text{ prep NP}_{\text{scene}} \)

This is a NP

Image features to predict nouns (subj, obj, scene)
- 20 object types: Felzenszwalb detector responses
- 8 scenes: GIST descriptors

Language model to predict verbs and preposition:
Verb: based on \( \text{NP}_{\text{obj}} \) and \( \text{NP}_{\text{subj}} \)
Preposition: based on verb and \( \text{NP}_{\text{scene}} \) (or \( \text{NP}_{\text{obj}} \))
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

Elliott & Keller 2013

Visual Dependency Grammar

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.

Im2Text

Ordonez et al. 2011

Data set: SBU Captioned Photo Dataset
- 3M 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)
Kuznetsova et al. 2012

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 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 \( \mathbf{I}_{\text{test}} \) 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 \( f(I_{\text{query}}, S) \)

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 \( I \in \mathbf{I}_{\text{test}} \) by \( f(I, S_{\text{query}}) \)

Image search

Kernel CCA

KCCA learns projection weights that maximize the correlation between the kernel matrices \( \mathbf{K}_a, \mathbf{K}_s \)

Kernel matrix \( \mathbf{K}_d = k(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle \):
Contains implicit dot product of training items \( x_i, x_j \) in a high-dimensional space \( \varphi(x) \)

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

Image annotation

Canonical Correlation Analysis
Input: Pairs of items \( (A_i, B_i) \) drawn from different spaces \( A_i \in A, B_i \in B \)
Output: maximally correlated linear projections \( \mathbf{w}_A, \mathbf{w}_B \) that project items from \( A, B \) into an induced common space such that \( A \approx \) close to \( B \).

\[
\text{arg}\max_{\mathbf{w}_A, \mathbf{w}_B} \left\langle \mathbf{A}_w A, \mathbf{B}_w B \right\rangle \\
\left\| \mathbf{A}_w A \right\| \left\| \mathbf{B}_w B \right\|
\]
Kernel Canonical Correlation Analysis

Using KCCA for image description

Image Kernels

Text kernels

Lexical similarities

Our experiments

Our models

Examples
Image search examples

Two little girls practice martial arts

A dog in a grassy field, looking up.

Ranking-based evaluation

Recall of original item (automatic)

<table>
<thead>
<tr>
<th>Sentence NN</th>
<th>Image annotation</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td></td>
<td>2.5&quot;</td>
<td>7.6&quot;</td>
<td>9.7&quot;</td>
<td>2.5&quot;</td>
<td>4.7&quot;</td>
<td>7.2&quot;</td>
</tr>
<tr>
<td>BoW1</td>
<td></td>
<td>4.8&quot;</td>
<td>13.5&quot;</td>
<td>19.7&quot;</td>
<td>4.5&quot;</td>
<td>14.3&quot;</td>
<td>20.8&quot;</td>
</tr>
<tr>
<td>BoW5</td>
<td></td>
<td>6.2&quot;</td>
<td>17.1&quot;</td>
<td>24.3&quot;</td>
<td>5.8&quot;</td>
<td>16.7&quot;</td>
<td>23.6&quot;</td>
</tr>
<tr>
<td>Tri5Sem</td>
<td></td>
<td>8.3</td>
<td>21.6</td>
<td>30.3</td>
<td>7.6</td>
<td>20.7</td>
<td>30.1</td>
</tr>
</tbody>
</table>

Rate of Success (large-scale human evaluation)

<table>
<thead>
<tr>
<th>Sentence NN</th>
<th>Image annotation</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td></td>
<td>5.8&quot;</td>
<td>15.3&quot;</td>
<td>20.1&quot;</td>
<td>4.9&quot;</td>
<td>12.9&quot;</td>
<td>18.1&quot;</td>
</tr>
<tr>
<td>BoW1</td>
<td></td>
<td>12.2&quot;</td>
<td>30.3&quot;</td>
<td>39.7&quot;</td>
<td>11.4&quot;</td>
<td>30.5&quot;</td>
<td>40.2&quot;</td>
</tr>
<tr>
<td>BoW5</td>
<td></td>
<td>15.0&quot;</td>
<td>34.1&quot;</td>
<td>42.7&quot;</td>
<td>12.1&quot;</td>
<td>31.5&quot;</td>
<td>40.8&quot;</td>
</tr>
<tr>
<td>Tri5Sem</td>
<td></td>
<td>16.6</td>
<td>37.7</td>
<td>49.1</td>
<td>15.7</td>
<td>36.9</td>
<td>48.5</td>
</tr>
</tbody>
</table>

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 \( (i_1, s_1), (i_2, s_2) \) is greater than that of incorrect pairs \( (i_1, s_2), (i_2, s_1) \) by a margin \( \Delta \): \[ i_1s_1 > i_2s_2 + \Delta \quad \text{and} \quad i_2s_1 > i_1s_2 + \Delta \]
Denotational Semantics

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

\[ \lbrack s \rbrack = \{ w \in U : s \text{ is true in } w \} \]

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

\[ \lbrack s \rbrack = \{ i \in I : s \text{ describes (part of) } i \} \]

Young, Lai, Hodosh, Hockenmaier, TACL 2014.

Denotation Graph

Denotations induce a partial ordering over descriptions.

\[ \lbrack \text{a white dog runs on the beach} \rbrack \subset \lbrack \text{a dog runs} \rbrack \]

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 \[ \lbrack s \rbrack \geq 2 \]
~53K captions with \[ \lbrack s \rbrack \geq 5 \]
~22K captions with \[ \lbrack s \rbrack \geq 10 \]
~1.0K captions with \[ \lbrack s \rbrack \geq 100 \]
161 captions with \[ \lbrack s \rbrack \geq 1000 \]
   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

Summer-based image description


Note: The reference list contains a variety of sources, including academic papers, conference proceedings, and other scholarly publications, each with specific details such as authors, titles, publication years, and page numbers. This list covers a range of topics from computer vision to natural language processing, reflecting the interdisciplinary nature of research in these fields.