Visual Features for Linguists

Basic image analysis techniques for multimodally-curious NLPers

Elia Bruni    Marco Baroni

Center for Mind/Brain Sciences
University of Trento

ACL Tutorial 2013
Acknowledgments

• Partial financial support from a Google Research Award

• Thanks to the authors of all the online materials we liberally sampled from (see slide credits)

• Special thanks to Jasper Uijlings, our in-house image analysis tutor
Outline of the tutorial

1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

4 Visual words for higher-level image representation

5 Example applications in computer vision

6 Going multimodal: Visual features in NLP
Suggested reading
Outline of the tutorial

1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

4 Visual words for higher-level image representation

5 Example applications in computer vision

6 Going multimodal: Visual features in NLP
There is plenty of text out there
and we have the tools to process it

Select all the guidelines you should follow when changing the...
www.weegy.com/ConversationId=56E5F47D
Make important text stand out with underlining or bold text. Use many different fonts and styles. Make sure there is plenty of white space or open space around text.

get out of there cat, my neck is plenty warm... - Cats. Where they...
getoutoftherecat.tumblr.com/post/45100104385/get-out-of-there-cat...
get out of there cat, my neck is plenty warm without you. ... own ask box in order to avoid cluttering everyone’s dash with text...

For all the Bronco haters out there (there are plenty) - Page 47 ...
forum.grasscity.com/bronco-haters-out-there-there-are-plenty-47.html
For all the Bronco haters out there (there are plenty) Originally Posted by fourty47 Just have to play smart. Anything can happen, but at the

Policing The Redding, Andersen Police, Shasta Co.Sheriff & Court...
polingthepoliceofredding.ca.blogspot.com/2008/11/i-know-there-is...
17/11/2008 - Policing the Police of Redding Ca. I know there is plenty of people there in the Redding area who have been victims of the R.P.D. Redding Police Dept.

Plenty of jobs out there? - Electricians Forums
www.electriciansforums.co.uk/...53412-plenty-jobs-out-there.html
Plenty of jobs out there? Discuss Plenty of jobs out there? in the Electrical Forum, General Electrical Forum at Electricians Forums Discussion Boards; Hi all, I'm ...

There - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/There
there, a deictic adverb in English; there, an English pronoun used in phrases such as ...
Text is available under the Creative Commons Attribution-ShareAlike License ...
There are also plenty of (labeled!) images out there!
And increasingly accurate tools to process them

... enabling sophisticated applications such as image stitching
And increasingly accurate tools to process them

ImageNet Large Scale Visual Recognition Challenge

2012 winner performed 1K-object classification with 16% 5-guess error
Why should WE care?

- Besides all the new applications we can pursue by bringing together language and vision...
- Image analysis might help us to deal with the very concrete problem of lack of grounding of linguistic symbols
Lack of grounding

To summarize, the semantic similarity structure of the MCSM representations is governed by different types of features depending on the semantic class. For example, animals are characterized by behaviors (<flies>) and taxonomic features (<is_animal>),

Fig. 2. Multidimensional scaling solution for an example 5-cluster solution from MCSM (top panel) and COALS (bottom panel) for nouns from McRae et al. (2005). Peaks are labeled with cluster numbers and the three most descriptive features for each cluster.
Lack of grounding

A sheep...

- According to the subject descriptions in the McRae et al.’s 2005 norms: **is white, has wool, has 4 legs, bahs**, . . .
- According to the text-generated descriptions of Baroni et al. 2010: **needs a shepherd, might suffer of scrapie, grazes, in a farm**, . . .

- Kelly et al. 2010, using large corpora, weak supervision, lexico-syntactic patterns, achieve max 24% precision, 48% recall at guessing McRae-subject-generated properties
Lack of grounding

The psychedelic world of corpus-determined color

- clover is blue
- coffee is green
- crows are white
- flour is black
- fog is green
- gold is purple
- mud is red
- the sky is green
- violins are blue
Lack of grounding

Andrews et al., Psych. Review 2009, Fig. 4
The image analysis pipeline

- Low-level feature detection and description
- Mapping low-level features to visual words
- Histogram representation
Outline of the tutorial

1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

4 Visual words for higher-level image representation

5 Example applications in computer vision

6 Going multimodal: Visual features in NLP
The PASCAL VOC dataset

- Around 10,000 images, with around 25,000 target objects
  - Objects from 20 categories (*person, car, bicycle, cow, table...*)
  - Objects are annotated with labeled bounding boxes

Slide credit: Pedro Felzenswalb
ImageNet at a glance

ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. Currently we have an average of over five hundred images per node. We hope ImageNet will become a useful resource for researchers, educators, students and all of you who share our passion for pictures. Click here to learn more about ImageNet, Click here to join the ImageNet mailing list.

What do these images have in common? Find out!

The ImageNet Challenge 2013 is announced!

© 2013 Stanford Vision Lab, Stanford University, Princeton University support@image-net.org Copyright infringement
ImageNet at a glance

- Animals
  - Birds
  - Fish
  - Mammal
  - Invertebrate
- Scenes
  - Indoor
  - Geological formations
- Sport activities
- Materials and fabric
- Instrumentation
  - Tools
  - Appliances
  - ...
- Plants
  - ...

Slide credit: Fei-Fei Li
ImageNet is a knowledge database

- Taxonomy on top of WordNet

Slide credit: Fei-Fei Li
ImageNet is large scale

**LabelMe**

Russell et al. 2005; statistics obtained in 2009

**IMAGENET**

Deng et al. 2009
statistics obtained in 2009

Slide credit: Fei-Fei Li
The ESP Game dataset

- 100K labeled images
- Labeled through a game:
  - two people are partnered together
  - both see the same image and have to agree on an appropriate word label
  - a word entered by both participants becomes a label for the image

- [http://www.cs.cmu.edu/~biglou/resources/](http://www.cs.cmu.edu/~biglou/resources/)
What do you see?

taboo words
- peace
- lay
guesses
- sheeps...
- sheep

Slide credit: Luis von Ahn
Grass
White
Sheep
Sheep's
Sheep

Slide credit: Luis von Ahn
New image tag: Sheep

Slide credit: Luis von Ahn
mirror, mud, white, person, stuck, car, jeep, door, tire, wheel

triangle, pink, building, tower, square, towers

band, sing, hair, arm, singer, man, guitar, mic, microphone

desert, soldier, army, man

coin, round, money, face, gold, old, man

imagine, in-depth, depth, uro, in, reports, more, euro
LabelMe

Label as many objects and regions as you can in this image

Instruction:
Use your mouse to click around the boundary of some objects in this image. You will then be asked to enter the name of the object (examples: car, window).

Labeling tools:
- Frase segment
- Zoom
- Fit Image

Polygons in this image:
- door
- door
- road
- stair
- window
- window
- sidewalk
- building region
- house
- window
- window
- window

Slide credit: Antonio Torralba
Polygon quality
Not all data is reliable

Most common labels:
- test
- adksdsasa
- woiieiiie
- ...

Slide credit: Antonio Torralba
Online hooligans

Slide credit: Antonio Torralba
More on annotated image datasets


Outline of the tutorial

1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

4 Visual words for higher-level image representation

5 Example applications in computer vision

6 Going multimodal: Visual features in NLP
Local features: main components

• Detection: Identify the interest points

• Description: Extract feature descriptor surrounding each interest point

\[ \mathbf{x}_1 = [x^{(1)}_1, \ldots, x^{(1)}_d] \]

\[ \mathbf{x}_2 = [x^{(2)}_1, \ldots, x^{(2)}_d] \]

Slide credit: Kristen Grauman
Chapter 1. Introduction

Local features: challenges

Figure 1.3: Images containing instances of the same object category can appear dramatically different; recognition methods must therefore be robust to a variety of challenging nuisance parameters.

Aside from these issues relating to robustness, today's recognition algorithms also face notable challenges in computational complexity and scalability. The fact that about half of the cerebral cortex in primates is devoted to processing visual information gives some indication of the computational load one can expect to invest for this complex task (Felleman & van Essen 1991). Highly efficient algorithms are necessary to accommodate rich high-dimensional image representations, to search large image databases, or to extend recognition to thousands of category types. In addition, scalability concerns also arise when designing a recognition system's training data: while unambiguously labeled image examples tend to be most informative, they are also most expensive to obtain. Thus, methods today must consider the trade-off between the extended cost algorithms require versus the advantages given to the learning process.

Slide credit: Grauman and Leibe
1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images
   Feature detectors
   Feature descriptors

4 Visual words for higher-level image representation

5 Example applications in computer vision

6 Going multimodal: Visual features in NLP
Local detectors

- Goals: Repeatable detection, precise localization, interesting content

Slide credit: Kristen Grauman
Corners as distinctive interest points

- **Design criteria**
  - We should easily recognize the point by looking through a small window *(locality)*
  - Shifting the window in any *direction* should give a *large change* in intensity *(good localization)*

  - **“flat” region:** no change in all directions
  - **“edge”:** no change along the edge direction
  - **“corner”:** significant change in all directions

Slide credit: Alyosha Efros
**Harris Detector: Responses**

*Effect:* A very precise corner detector.

*Slide credit:* Krystian Mikolajczyk
Harris Detector: Responses

Slide credit: Krystian Mikolajczyk
From keypoints to regions

Slide credit: Rick Szeliski
From keypoints (Harris) to regions (DoG)
Recent advances

Dense feature extraction [Nowak et al. 2006]
1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images
   Feature detectors
   Feature descriptors

4 Visual words for higher-level image representation

5 Example applications in computer vision

6 Going multimodal: Visual features in NLP
Local descriptors

- We know how to detect points
- Next question: How to describe them?

\[
\mathbf{x}_1 = [x_1^{(1)}, \ldots, x_d^{(1)}]
\]

\[
\mathbf{x}_2 = [x_1^{(2)}, \ldots, x_d^{(2)}]
\]

Slide credit: Kristen Grauman
SIFT descriptor [Lowe 1999]

- **Scale Invariant Feature Transform**
- **Descriptor computation:**
  - Divide regions into 4x4 cells
  - Compute histogram of gradient orientations (8 reference angles) for all pixels inside each sub-patch - Gradient = directional change in the intensity or color in an image
  - Resulting descriptor: 4x4x8 = 128 dimensions
SIFT descriptor - rotation invariance

- Estimation of the dominant direction
  - Extract gradient orientation
  - Histogram over gradient orientation
  - Peak in this histogram

- Rotate patch in dominant direction

Slide credit: Cordelia Schmid
Many more descriptors than just SIFT

- SIFT [Lowe ’99]
  - Color SIFT [Sande et al. 2010]
  - Normalizing SIFT with square root transformation [Arandjelovic et Zisserman 2012]
- Textons [Leung and Malik ’01]
- HoG [Dalal and Triggs ’05]
- SURF [Bay et al. ’08]
- DAISY [Tola et al. ’08, Windler et al. ’09]
- Bag of (e.g., LAB) colors [Farhadi et al. ’09]
More on low-level features

Outline of the tutorial

1. Why image analysis?
2. Annotated image datasets
3. Extraction of low-level features from images
4. Visual words for higher-level image representation
5. Example applications in computer vision
6. Going multimodal: Visual features in NLP
1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

4 Visual words for higher-level image representation
   Constructing a vocabulary of visual words
   Classic Bags-of-visual-words representation
   Recent advances

5 Example applications in computer vision

6 Going multimodal: Visual features in NLP
BoVW: Feature extraction
BoVW: Feature extraction
BoVW: Feature extraction
BoVW: Feature extraction

SIFT 4x4

Descriptor Space
BoVW: Feature extraction
BoVW: Feature extraction

Descriptor Space
BoVW: Feature extraction

Descriptor Space
BoVW: Visual dictionary

Descriptor Space

The clusters partition the descriptor space. Each cluster is called a “Visual Word”.
BoVW: Visual dictionary

The clusters partition the descriptor space. Each cluster is called a “Visual Word”.
BoVW: Visual dictionary

Descriptor Space

The clusters partition the descriptor space. Each cluster is called a “Visual Word”.
BoVW: Visual dictionary

The clusters partition the descriptor space. Each cluster is called a “Visual Word”.

Descriptor Space
1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

4 Visual words for higher-level image representation
   Constructing a vocabulary of visual words
   Classic Bags-of-visual-words representation
   Recent advances

5 Example applications in computer vision

6 Going multimodal: Visual features in NLP
BoVW: Image representation

Descriptor Space

Global Representation

SIFT 4x4
BoVW: Image representation

SIFT 4x4

Descriptor Space

Global Representation
BoVW: Image representation
BoVW: Image representation

Descriptor Space

SIFT 4x4

Global Representation
BoVW: Image representation
1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

4 Visual words for higher-level image representation
   Constructing a vocabulary of visual words
   Classic Bags-of-visual-words representation
   Recent advances

5 Example applications in computer vision

6 Going multimodal: Visual features in NLP
Spatial pyramid representation [Lazebnik et al. 2006]

Locally orderless representation at several levels of spatial resolution

Slide credit: Svetlana Lazebnik
Spatial pyramid representation

Locally orderless representation at several levels of spatial resolution

Slide credit: Svetlana Lazebnik
Spatial pyramid representation

Locally orderless representation at several levels of spatial resolution

Slide credit: Svetlana Lazebnik
Fisher encoding [Perronnin et al. 2010]

- BoVW is only about **counting** the number of local descriptors assigned to each region
- Why not including other statistics?

Slide credit: Kristen Grauman
Fisher encoding

- BoVW is only about **counting** the number of local descriptors assigned to each region
- **Mean** of local descriptors

Slide credit: Kristen Grauman
Fisher encoding

- BoVW is only about **counting** the number of local descriptors assigned to each region
- **Variance** of local descriptors

Slide credit: Kristen Grauman
More on bag-of-visual-words


• F. Perronnin, J. Sanchez and T. Mensink. 2010. Improving the fisher kernel for large-scale image classification. *Proceedings of ECCV.*
Outline of the tutorial

1. Why image analysis?
2. Annotated image datasets
3. Extraction of low-level features from images
4. Visual words for higher-level image representation
5. Example applications in computer vision
6. Going multimodal: Visual features in NLP
1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

4 Visual words for higher-level image representation

5 Example applications in computer vision
   Emotion analysis
   Object recognition
   Visual attributes

6 Going multimodal: Visual features in NLP
Emotion analysis: Images evoke emotions

Slide credit: Yanulevskaya et al. 2012
Emotion analysis: The dataset

183 negative paintings

317 positive paintings

Slide credit: Yanulevskaya et al. 2012
Emotion analysis: The method

Proposed methodology

LAB
Visual Words
SIFT
Visual Words
Dictionary
Texture &
Shape
Feature 
Detection
Colours
Art Works

LAB descriptor

SIFT descriptor

Dictionary

Dictionary

Slide credit: Yanulevskaya et al. 2012
Emotion analysis: Results

- Task: Divide the dataset into train and test and use a classifier (SVM) to distinguish between positive and negative paintings

Accuracy:
- 0.63 Random
- 0.76 LAB visual words
- 0.73 SIFT visual words
- 0.78 LAB + SIFT

Slide credit: Yanulevskaya et al. 2012
More on emotion analysis


1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

4 Visual words for higher-level image representation

5 Example applications in computer vision
   - Emotion analysis
   - Object recognition
   - Visual attributes

6 Going multimodal: Visual features in NLP
Object recognition

- Image classification: assigning a class label to the image

- Object localization: define the location and the category

Slide credit: Cordelia Schmid
Object recognition: Typical pipeline

Slide credit: Chatfield et al. 2011
Pascal VOC 2012: Statistics

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Images</strong></td>
<td>11,540</td>
<td>10,994</td>
</tr>
<tr>
<td><strong>Objects</strong></td>
<td>27,450</td>
<td>27,078</td>
</tr>
</tbody>
</table>

- 20 categories
- Minimally, around 600 training objects per category
- Around 2,000 cars, 1,500 dogs and 8,500 people
- Approximately equal distribution across training and test datasets

Slide credit: Andrew Zisserman
Pascal VOC 2012: Submitted systems

• 7 systems, 5 groups

• Methods
  – Features: Dense SIFT, HoG, colour
  – Encodings: spatial pyramid, BoVW, Fisher vector
  – Classifier: SVM
Pascal VOC 2012: Results

Average precision by class

0
20
40
60
80
100

Average Precision

max
median
chance
Aeroplanes vs. bottles
Pascal VOC 2012: Results

Average Precision

NUSPSL_CTX_GPM_SCM
CVC_BOW_FK_COLORDET_SP
CVC_BOW_FK_COLORHOG
IT1_FK_FUSED_SIFT
IT1_FK_BS_GRAYSIFT
IMPERIAL_COMPLEX_LOGNORMAL

Median average precision by method
0
20
40
60
80
100
NUS
PSL
_C
TX
_G
PM
_S
CM
CVC_UV
A_UNITN_...
More on object recognition

- Results of the Pascal VOC Challenge 2012: [http://pascallin. ecs.soton.ac.uk/challenges/VOC/voc2012/index.html](http://pascallin. ecs.soton.ac.uk/challenges/VOC/voc2012/index.html)
1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

4 Visual words for higher-level image representation

5 Example applications in computer vision
   Emotion analysis
   Object recognition
   Visual attributes

6 Going multimodal: Visual features in NLP
Object-centric classification

Input Image

BoVW Representation

Object Classifier

Is it a DOG?
Attribute-centric classification

Input Image

BoVW Representation

Attribute Classifiers

Object Classifier

Is it **FURRY**?

Is it **BROWN**?

Is it a **DOG**?

Is it **SQUARE**?
The many uses of attributes

Farhadi et al., CVPR 2009, Fig. 1
Data sets

http://vision.cs.uiuc.edu/attributes/

**a-Pascal**  Pascal VOC 2008 data set, 20 categories (people, dog, bus, horse...), 150-to-1K images per category (5K for people), used for training and testing

**a-Yahoo**  12 categories similar to those in a-Pascal but different (statue, wolf, carriage, centaur...), used for testing
Base features

- BoVW spatial pyramid histograms from bounding boxes using:
  - HoG descriptors (good for parts)
  - Canny Edges (Canny, 1986; good for shapes)
  - Textons (good for materials)
  - LAB color features (good for materials)

- 9751-dimensional vectors (mostly HoG features)
Attributes

• 64 AmazonTurk-annotated “semantic” attributes:
  - **Shapes:** is 2D boxy, is 3D boxy, is cylindrical…
  - **Parts:** has head, has leg, has window…
  - **Materials:** is furry, has glass, is shiny…

• 1K automatically selected “discriminative” attributes
  - Selected to maximize discrimination between random subsets of classes or attributes
Learning

- Base feature selection by picking features that are best at within-category attribute learning
  - If you learn *has wheels* from cars, motorbikes, buses vs. horses, cats, bottles, you might learn *metallic* instead!
- Object classifiers trained on base features (standard approach) or on vectors of automatically assigned attributes
- Various learning algorithms used: linear SVMs, regularized logistic regression, nearest neighbour classification
- See the paper for details and more experiments
Identifying “new” objects
Farhadi et al., CVPR 2009, Fig. 9

Attribute classifiers trained on a-Pascal, predicted attributes used as features to train a-Yahoo object classifier (1NN)
Reporting missing attributes

68.2% accuracy

Aeroplane
No “wing”

Car
No “window”

Boat
No “sail”

Aeroplane
No “jet engine”

Motorbike
No “side mirror”

Car
No “door”

Bicycle
No “wheel”

Sheep
No “wool”

Train
No “window”

Sofa
No “wood”

Bird
No “tail”

Bird
No “leg”

Bus
No “door”
Reporting atypical attributes
47.3% accuracy
More on attributes


• Tutorial on attributes at CVPR, June 2013: [http://filebox.ece.vt.edu/~parikh/attributes/](http://filebox.ece.vt.edu/~parikh/attributes/)

• C. Silberer, V. Ferrari and M. Lapata. Models of semantic representation with visual attributes. *ACL 2013!*
Outline of the tutorial

1. Why image analysis?
2. Annotated image datasets
3. Extraction of low-level features from images
4. Visual words for higher-level image representation
5. Example applications in computer vision
6. Going multimodal: Visual features in NLP
1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

4 Visual words for higher-level image representation

5 Example applications in computer vision

6 Going multimodal: Visual features in NLP
   Generating image descriptions
   Multimodal distributional semantics
   Visual selectional preference
Generating image descriptions

We need image description \textit{generation} (not just \textit{retrieval})

Figure 1. Our system automatically generates the following descriptive text for this example image: “\textit{This picture shows one person, one grass, one chair, and one potted plant. The person is near the green grass, and in the chair. The green grass is by the chair, and near the potted plant}.”
This is a photograph of one person and one brown sofa and one dog. The person is against the brown sofa. And the dog is near the person, and beside the brown sofa.
Generating a labeled graph describing image contents

- Images represented by graphs with nodes labeled with *objects* and *stuff*, *attributes* of objects/stuff and *prepositions* connecting objects/stuff
- Label assignment casted as Conditional Random Field energy minimization problem with both image- and text-based features
- CRF and image-based feature parameters trained on 153 images with (up-to-5) sentence descriptions from UIUC PASCAL data set

Graph from Kulkarni et al., 2011, Fig. 3
Image-derived features

Object detectors  Based on Felzenszwalb et al.’s *deformable part models*, trained for 24 categories on PASCAL 2010 and ImageNet data

Stuff detectors  Based on Farhadi et al.’s basic features, SVM trained to recognize sky, road, building, tree, water and grass on ImageNet

Attribute classifiers  SVM trained on Flickr, Google, ImageNet and Farhadi et al.’s images with attribute labels, 21 attributes (blue, furry, wooden, shiny…)  

Preposition functions  Hand-coded heuristics, e.g., $\text{near}(a, b)$ value is given by normalized minimum distance between the regions of $a$ and $b$
Text-derived features

- Text features provide *smoothing* of image-based hypotheses about objects, their attributes and preposition-expressed relations
- Co-occurrence counts for *attribute-object/stuff* and *object/stuff-preposition-object/stuff*
- Linearly combined counts from parsed Flickr image description corpus and Google
Generation

• From graph labels to sentences
  – E.g., from *(white cloud) in (blue sky)* to *There is a white cloud in the blue sky*

• Two simple generation strategies
  – N-gram language model used only to insert function words between nouns, adjectives and prepositions in the graph labels
  – Hand-coded templates
Language model decoding vs. templates

Kulkarni et al., 2011, Fig. 5
Quantitative performance

<table>
<thead>
<tr>
<th>Method</th>
<th>w/o</th>
<th>w/ synonym</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>0.50</td>
<td>0.51</td>
</tr>
<tr>
<td>Language model-based generation</td>
<td>0.25</td>
<td>0.30</td>
</tr>
<tr>
<td>Template-based generation</td>
<td>0.15</td>
<td>0.18</td>
</tr>
<tr>
<td>Meaning representation (triples)</td>
<td>0.20</td>
<td>0.30</td>
</tr>
</tbody>
</table>

Table 1. Automatic Evaluation: BLEU score measured at 1

<table>
<thead>
<tr>
<th>Method</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality of image parsing</td>
<td>2.85</td>
</tr>
<tr>
<td>Language model-based generation</td>
<td>2.77</td>
</tr>
<tr>
<td>Template-based generation</td>
<td>3.49</td>
</tr>
</tbody>
</table>

Table 2. Human Evaluation: possible scores are 4 (perfect without error), 3 (good with some errors), 2 (many errors), 1 (failure)

BLEU scores computed on 847 UIUC PASCAL images and associated sentences
Good description examples

Kulkarni et al., 2011, Fig. 4
Bad description examples 1

Missing detections:

Here we see one pottedplant.

Incorrect detections:

There are one road and one cat. The furry road is in the furry cat.

Incorrect attributes:

This is a photograph of two sheep and one grass. The first black sheep is by the green grass, and by the second black sheep. The second black sheep is by the green grass.

This is a picture of one tree, one road and one person. The rusty tree is under the red road. The colorful person is near the rusty tree, and under the red road.

This is a photograph of two horses and one grass. The first feathered horse is within the green grass, and by the second feathered horse. The second feathered horse is within the green grass.

This is a picture of one dog.
Bad description examples 2

Counting is hard!
There are two cows and one person. The first brown cow is against the brown person, and near the second cow. The brown person is beside the second cow.

Just all wrong!
There are one potted plant, one tree, one dog and one road. The gray potted plant is beneath the tree. The tree is near the black dog. The road is near the black dog. The black dog is near the gray potted plant.

This is a picture of four persons. The first colorful person is by the second pink person, and by the third colorful person. The second pink person is by the third colorful person, and by the fourth person.

This is a photograph of one person and one sky. The white person is by the blue sky.

Kulkarni et al., 2011, Fig. 6
More on image description generation/adaptation


1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

4 Visual words for higher-level image representation

5 Example applications in computer vision

6 Going multimodal: Visual features in NLP
   Generating image descriptions
   Multimodal distributional semantics
   Visual selectional preference
Multimodal distributional semantics

• Our work:
  – E. Bruni, N.K. Tran and M. Baroni. Submitted. Multimodal distributional semantics

• More:
Distributional semantics

The geometry of meaning (e.g., Turney and Pantel, 2010)

<table>
<thead>
<tr>
<th></th>
<th>shadow</th>
<th>shine</th>
</tr>
</thead>
<tbody>
<tr>
<td>moon</td>
<td>16</td>
<td>29</td>
</tr>
<tr>
<td>sun</td>
<td>15</td>
<td>45</td>
</tr>
<tr>
<td>dog</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

![Diagram showing the geometry of meaning with vectors for moon, sun, and dog]
Multimodal distributional semantics

<table>
<thead>
<tr>
<th></th>
<th>planet</th>
<th>night</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>moon</td>
<td>10</td>
<td>22</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>sun</td>
<td>14</td>
<td>10</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>dog</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>20</td>
</tr>
</tbody>
</table>
Distributional semantics from images

moon
Distributional semantics from images

moon

moon

+
Distributional semantics from images

Labeled images

moon

moon

moon

Instance counts
Distributional semantics from images

Labeled images

moon

Instance counts

= 

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Total counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>moon</td>
<td>31</td>
<td>65</td>
<td>56</td>
<td>28</td>
</tr>
</tbody>
</table>
Distributional semantics from images

<table>
<thead>
<tr>
<th></th>
<th>□</th>
<th>★</th>
<th>△</th>
<th>☢</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>sun</strong></td>
<td>23</td>
<td>60</td>
<td>89</td>
<td>27</td>
</tr>
</tbody>
</table>

Total counts
Distributional semantics from images

<table>
<thead>
<tr>
<th></th>
<th>□</th>
<th>▫</th>
<th>▲</th>
<th>□</th>
<th>Total counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>23</td>
<td>60</td>
<td>89</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>dog</td>
<td>101</td>
<td>5</td>
<td>22</td>
<td>87</td>
<td></td>
</tr>
</tbody>
</table>
Distributional semantics from images

<table>
<thead>
<tr>
<th></th>
<th>:---:</th>
<th>:---:</th>
<th>:---:</th>
<th>:---:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><img src="image1.png" alt="blue square" /></td>
<td><img src="image2.png" alt="yellow star" /></td>
<td><img src="image3.png" alt="red triangle" /></td>
<td><img src="image4.png" alt="green circle" /></td>
</tr>
<tr>
<td><strong>moon</strong></td>
<td>31</td>
<td>65</td>
<td>56</td>
<td>28</td>
</tr>
<tr>
<td><strong>sun</strong></td>
<td>23</td>
<td>60</td>
<td>89</td>
<td>27</td>
</tr>
<tr>
<td><strong>dog</strong></td>
<td>101</td>
<td>5</td>
<td>22</td>
<td>87</td>
</tr>
</tbody>
</table>
What is this good for?

**Task 1** Predicting human *semantic relatedness* judgments

**Task 2** **Concept categorization**, i.e. grouping words into classes based on their semantic relatedness (*car* ISA *vehicle*; *banana* ISA *fruit*)

**Task 3** Find **typical color** of concrete objects (*cardboard* is *brown*, *tomato* is *red*)

**Task 4** Distinguish **literal vs. non-literal** usages of color adjectives (*blue uniform* vs. *blue note*)
Task 1: Semantic relatedness data sets

- **Data**
  - **WordSim353 dataset**
    - 353 word pairs (coverage: 252)
    - 16 subjects rate each pair on a 10-point scale, ratings averaged
    - *dollar/buck: 9.22, professor/cucumber: 0.31*
  - **MEN dataset (created by us)**
    - 3,000 word pairs, tags in image datasets
    - crowdsourcing: subjects see two word pairs and pick the pair containing most related words
    - each word pair is rated 50 times, score = selected / 50
    - *cold/frost: 0.9, eat/hair: 0.1*

- **Method**
  - for each model, compute cosine between word vectors
  - score: Spearman correlation against the human ratings
### Task 1: Semantic relatedness results

Bruni, N.K. Tran and Baroni (submitted)

| Model  | Window 2 |  | Window 20 |  |
|--------|----------|  |-----------|  |
|        | MEN      | WS | MEN       | WS |
| Text   | 0.73     | 0.70 | 0.68     | 0.70 |
| Image  | 0.43     | 0.36 | 0.43     | 0.36 |
| Fusion | **0.78** | **0.72** | **0.76** | **0.75** |

Spearman correlation of the models on MEN and WordSim
(all coefficients significant with $p < 0.001$).
Task 2: Concept categorization data sets

• Data
  – Battig (for training)
    • 77 concepts from 10 different classes
    • bird (eagle, owl...) vegetable (broccoli, potato...)
  – Almuhareb-Poesio (for testing)
    • 231 concepts from 21 different classes
    • vehicle (airplane, car...) time (aeon, future...)

• Method
  – cluster the words based on their pairwise cosines in the semantic space (using the CLUTO toolkit)
### Task 2: Concept categorization results

Bruni, N.K. Tran and Baroni (submitted)

<table>
<thead>
<tr>
<th></th>
<th>Window 2</th>
<th>Window 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td>0.73</td>
<td>0.65</td>
</tr>
<tr>
<td>Image</td>
<td>0.26</td>
<td>0.26</td>
</tr>
<tr>
<td>Fusion</td>
<td><strong>0.74</strong></td>
<td><strong>0.69</strong></td>
</tr>
</tbody>
</table>

Percentage purities of the models on Almuhareb-Poesio.
Task 3: Find typical color of concrete objects

• Data and task
  – spot typical color of 52 concrete objects: cardboard is brown, coal is black, forest is green
  – typical colors assigned by two judges by consensus
    • Berlin and Kay (1969)’s basic color adjectives: black, blue, brown, green, grey, orange, pink, purple, red, white, yellow

• Method
  – rank color adjective vectors by similarity to the noun vectors
  – good models will rank right color high
Task 3: Find typical color of concrete objects
Bruni, Boleda, Baroni and N.K. Tran 2012

<table>
<thead>
<tr>
<th>Model</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEXT\textsubscript{30K}</td>
<td>3</td>
</tr>
<tr>
<td>LAB\textsubscript{128}</td>
<td>1</td>
</tr>
<tr>
<td>SIFT\textsubscript{40K}</td>
<td>3</td>
</tr>
<tr>
<td>TEXT+LAB\textsubscript{128}</td>
<td>1</td>
</tr>
<tr>
<td>TEXT+SIFT\textsubscript{40K}</td>
<td>2</td>
</tr>
</tbody>
</table>

Median rank of correct color and # of top matches
## Task 3: Examples

<table>
<thead>
<tr>
<th>word</th>
<th>gold</th>
<th>LAB</th>
<th>SIFT</th>
<th>TEXT</th>
</tr>
</thead>
<tbody>
<tr>
<td>cauliflower</td>
<td>white</td>
<td>green</td>
<td>yellow</td>
<td>orange</td>
</tr>
<tr>
<td>cello</td>
<td>brown</td>
<td>brown</td>
<td>black</td>
<td>blue</td>
</tr>
<tr>
<td>deer</td>
<td>brown</td>
<td>green</td>
<td>blue</td>
<td>red</td>
</tr>
<tr>
<td>froth</td>
<td>white</td>
<td>brown</td>
<td>black</td>
<td>orange</td>
</tr>
<tr>
<td>gorilla</td>
<td>black</td>
<td>black</td>
<td>red</td>
<td>grey</td>
</tr>
<tr>
<td>grass</td>
<td>green</td>
<td>green</td>
<td>green</td>
<td>green</td>
</tr>
<tr>
<td>pig</td>
<td>pink</td>
<td>pink</td>
<td>brown</td>
<td>brown</td>
</tr>
<tr>
<td>sea</td>
<td>blue</td>
<td>blue</td>
<td>blue</td>
<td>grey</td>
</tr>
<tr>
<td>weed</td>
<td>green</td>
<td>green</td>
<td>yellow</td>
<td>purple</td>
</tr>
</tbody>
</table>
Task 4: Literal vs. non-literal

• Data and task
  – distinguish literal and non-literal usages of color adjectives: blue uniform, blue shark, blue note
  – 342 adjective-noun pairs, 227 literal, 115 non-literal, as decided by two judges by consensus

• Method
  – compute cosine between color adjective vector and noun vector
  – prediction: higher similarity of color and noun vectors for literal uses
### Task 4: Literal vs. non-literal

<table>
<thead>
<tr>
<th>Model</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEXT(_{30K})</td>
<td>0.53***</td>
</tr>
<tr>
<td>LAB(_{128})</td>
<td>0.25*</td>
</tr>
<tr>
<td>SIFT(_{40K})</td>
<td>0.57***</td>
</tr>
<tr>
<td>TEXT+LAB(_{128})</td>
<td>0.36***</td>
</tr>
<tr>
<td>TEXT+SIFT(_{40K})</td>
<td>0.73***</td>
</tr>
</tbody>
</table>

Average difference in normalized adj-noun cosines in literal vs. non-literal conditions with t-test significance.
The illustrated distributional hypothesis

Current development

The meaning of a visually depicted concept is (can be approximated by, derived from) the set of contexts in which it occurs in images
The illustrated distributional hypothesis

<table>
<thead>
<tr>
<th>Area</th>
<th>No</th>
<th>Manual</th>
<th>Automatic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept</td>
<td>NA</td>
<td>0.39</td>
<td>0.36</td>
</tr>
<tr>
<td>Context</td>
<td>NA</td>
<td>0.50</td>
<td>0.51</td>
</tr>
<tr>
<td>Concept+Context</td>
<td>0.47</td>
<td><strong>0.54</strong></td>
<td><strong>0.54</strong></td>
</tr>
</tbody>
</table>
1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

4 Visual words for higher-level image representation

5 Example applications in computer vision

6 Going multimodal: Visual features in NLP
   - Generating image descriptions
   - Multimodal distributional semantics
   - Visual selectional preference
Visual selectional preference

*Proceedings of RANLP*
What would you rather eat?

• migas?
• zeolite?
• carillons?
• a ficus?
• a mamey?
• manioc?
What would you rather eat?

Figure 1: Which out-of-vocabulary nouns are plausible direct objects for the verb eat? Examples include migas, zeolite, carillon, ficus, mamey, and manioc.

Our experiments evaluate the ability of these classifiers to correctly predict the selectional preferences of a small set of verbs. We evaluate two cases: 1) the case where the nouns are all assumed to be out-of-vocabulary, and the classifiers must make predictions without any corpus-based co-occurrence information, and 2) the case where we assume access to noun-verb co-occurrence information derived from web-scale N-gram data.

We show that visual features are useful for some verbs, but not for others. For verbs taking abstract arguments without definitive visual features, the classifier can often learn to disregard the visual data. On the other hand, for verbs taking physical arguments (such as food, animals, or people), the classifier can make accurate predictions using the nouns' visual properties. In these cases, visual information remains useful even after incorporating the web-scale statistics.

2. Visual Selection Preference

Consider determining whether the nouns carillon, migas, and mamey are plausible arguments for the verb eat. Existing systems are unlikely to have such words in their training data, let alone information about their edibility. However, after inspecting a few images returned by a Google search for these words (Figure 1), a human might reasonably predict which words are edible. Humans make this determination by observing both intrinsic visual properties (pits, skins, rounded shapes and fruity colors) and extrinsic visual context (circular plates, bowls, and other food-related tools) (Oliva and Torralba, 2007).

We propose using similar information to predict the plausibility of arbitrary verb-noun pairs. That is, we aim to learn the distinguishing visual features of all nouns that are plausible arguments for a given verb. This differs from work that has aimed to recognize, annotate and retrieve objects defined by a single phrase, such as tree or wrist watch (Feng and Lapata, 2010a). These approaches learn from labeled images during training in order to assign words to unlabeled images during testing. In contrast, we analyze labeled images (during training and testing) in order to determine their visual compatibility with a given predicate. Our approach does not need labeled training images for a specific noun in order to assess that noun during testing; e.g. we can make a reasonable prediction for the plausibility of eat mamey even if we've never encountered mamey before.

We now specify how we automatically 1) download a set of images for each noun argument, 2) extract visual features from each image, and 3) combine the visual features from multiple images into plausibility scores. Scripts, code and data are available at: www.clsp.jhu.edu/~sbergsma/ImageSP/.

2.1 Mining noun images from the web

To obtain a set of images for a particular noun argument, we submit the noun as a query to either the Flickr photo-sharing website (www.flickr.com), or Google's image search (www.google.com/imghp). In both cases, we download the thumbnails on the results page directly rather than downloading the source images. Flickr returns images by matching the query against user-provided tags and accompanying text. Google retrieves images based on the image caption, file-name, and surrounding text (Feng and Lapata, 2010a). Images obtained from Google are known to be competitive with “hand prepared datasets” for training object recognizers (Fergus et al., 2005).
Discriminative selectional preference

Bergsma et al. 2008

- Train a classifier for each verb to predict which nouns are acceptable objects:
  \[ y^v = \lambda^v \cdot \Phi^v(n) \]
- Requires large corpus to extract features such as co-occurrence of noun with other verbs
- For out-of-corpus-vocabulary nouns, use simple string-shape features
Enrich out-of-vocabulary noun representation with visual features

E.g. “migas” → Φᵥ(𝑛)

- Images retrieved from Flickr or Google Image Search, 6 images per noun
- Features: RGB-color and SIFT visual words (64- and 512-item color and 100- and 1000-item SIFT vocabularies, concatenated and fed to classifier)

Image from Shane Bergsma’s slides
Task

- Seven verbs: eat, inform, hit, kill, park, hunt and shoot down
- Training examples range from 500 to 10,000 per-verb, test instances from 50 to 1,000
  - Positive examples have AQUAINT corpus $PMI(v, n) > 0$, for negative examples $PMI(v, n) < 0$
- Baseline uses string features only
  - When using (co-)occurrence counts from the Web, visual features do not significantly improve over using textual data only!
### Accuracy across verbs

<table>
<thead>
<tr>
<th>Verb</th>
<th>Baseline</th>
<th>+Google features</th>
</tr>
</thead>
<tbody>
<tr>
<td>eat</td>
<td>68.3</td>
<td>79.5</td>
</tr>
<tr>
<td>inform</td>
<td>68.0</td>
<td>68.2</td>
</tr>
<tr>
<td>hit</td>
<td>68.7</td>
<td>68.7</td>
</tr>
<tr>
<td>kill</td>
<td>67.7</td>
<td>68.5</td>
</tr>
<tr>
<td>park</td>
<td>69.9</td>
<td>69.9</td>
</tr>
<tr>
<td>hunt</td>
<td>67.6</td>
<td>76.5</td>
</tr>
<tr>
<td>shoot down</td>
<td>70.0</td>
<td>72.0</td>
</tr>
</tbody>
</table>

From Shane Bergsma’s slides
Google Image better than Flickr

Average Accuracy Across All 7 Verbs

No visual features

From Shane Bergsma’s slides
The more images the better

From Shane Bergsma’s slides
Both types of visual features help

Accuracy On The Verb “Eat”

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Features</td>
<td>79.5</td>
</tr>
<tr>
<td>-Color Histogram</td>
<td>78.4</td>
</tr>
<tr>
<td>-SIFT Keypoint</td>
<td>78.1</td>
</tr>
<tr>
<td>-Color &amp; -SIFT</td>
<td>68.3</td>
</tr>
</tbody>
</table>

From Shane Bergsma’s slides
Welcome to the VSEM Website!

VSEM is a novel toolkit which allows the extraction of image-based representations of concepts in an easy fashion.

VSEM is equipped with state-of-the-art algorithms, from low-level feature detection and description up to the BoVW representation of images, together with a set of new routines necessary to move from an image-wise to a concept-wise representation of image content.

Download
- VSEM 0.1

Documentation
- MATLAB API
- Tutorials

Demos
- Pascal VOC demo

News

April 8, 2013
VSEM 0.1 released
The first version of VSEM has been released!

April 6, 2013
VSEM tutorials
The Bag of Visual Words, Concepts and Similarity Benchmark tutorials are now online.
SUPPLEMENTARY MATERIALS
Canny edges

Slide credit: Kristen Grauman
Canny edges

Slide credit: Kristen Grauman
Textons

- Texture is characterized by the repetition of basic elements or **Textons**


Slide credit: Cordelia Schmid
Textons


Slide credit: Cordelia Schmid
Color

Slide credit: Dieter Fox
LAB

Opponent color model of the L*a*b* color space

\( a^* \) channel opponent colors

\( b^* \) channel opponent colors
Emotion analysis: The method

Proposed methodology

LAB Visual Words SIFT Visual Words

Dictionary

Texture & Shape

Feature Detection

Colours

Art Works

LAB descriptor SIFT descriptor

Dictionary

Slide credit: Yanulevskaya et al. 2012

Slide credit: Victoria Yanuleskaya