Distributional semantics from text and images

E. Bruni[1] G. B. Tran[2] M. Baroni[1]

[1]CIMeC, University of Trento

[2]L3S Research, Hannover

GEMS 31/07/2011

Aknowledgments

- This project is possible thanks to a Google Research Award

Outline

Outline

Distributional Semantic Models

- Use large text corpora to derive estimates of semantic similarities between words
- Distributional hypothesis: semantically similar words tend to appear in similar contexts (Harris, 1954)
- For example, the meaning of *spinach* (primarily) becomes the result of statistical computations based on the association between *spinach* and words like *plant*, *green*, *iron*, *Popeye*, *muscles*

Perceptual Models

- Humans also rely on non-verbal experience, and comprehension also involves the activation of non-linguistic representations (Barsalou et al., 2008; Glenberg, 1997; Zwaan, 2004)
- We need to ground words' meanings to bodily actions and perceptions in the environment (Harnad, 1990)
- Back to our example, the meaning of *spinach* should come (at least partially) from our experience with spinach, its colors, smell and the occasions in which we tend to encounter it

Two (apparently) mutual exclusive views

1. Meaning emerges from association between *linguistic units* reflected by statistical computations on large bodies of text
2. Meaning is still the result of an association process, but one that concerns the association between *words and perceptual information*

A unified model

- By combining the two models we could construct a richer and more human-like notion of meaning
- In particular, we concentrate on perceptual information coming from images, and we create a multimodal distributional semantic model extracted from texts and images
- Putting side by side techniques from NLP and computer vision

Outline

Bag of visual words

- As bag-of-words approach employed in information retrieval, the “bag of visual words” is a similar technique used mainly for scene classification (Yang et al., 2007):
  1. To represent an image using BoW model, an image can be treated as a document
  2. However, "words" in images do not come off-the-shelf like in text documents
  3. To achieve bag-of-words representation of image document, pipeline in next two slides is typically followed

Visual dictionary construction

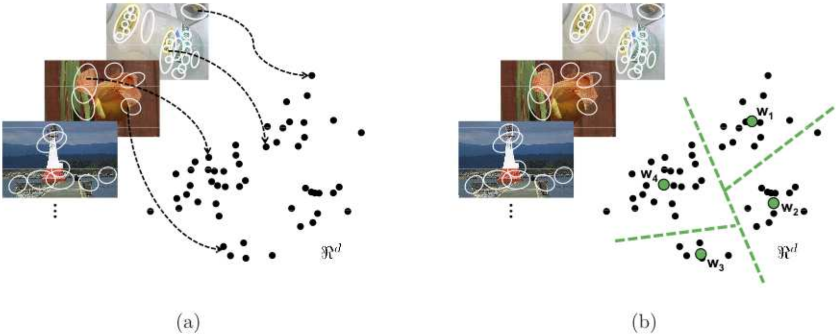
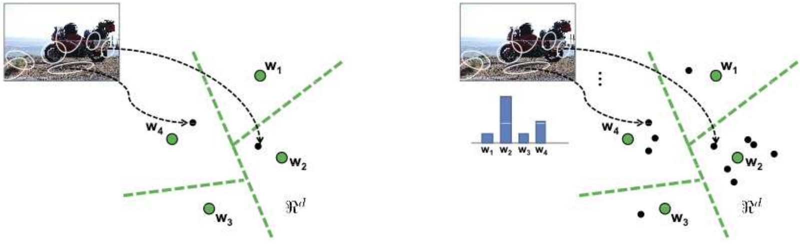


Figure: (a) A large corpus of representative images are used to populate the feature space with descriptor instances. (b) The sampled features are clustered in order to quantize the space into a discrete number of visual words (K. Grauman, B. Leibe)

Constructing histograms/vectors of visual words



(c) (d)

Figure: (c) Given a new image, the nearest visual word is identified for each of its features. This maps the image from a set of high-dimensional descriptors to a list of word numbers. (d) A bag-of-words histogram/vector can be used to summarize the entire image (K. Grauman, B. Leibe)

Outline

Bag of visual words

- Once we have represented images as bag-of-visual-words vectors, we can say that a word appearing in proximity of an image is co-occurring with the set of visual words present in the image

Words and visual words concatenation

- We can represent the document and its associated image as a mixture of textual and visual words

	leash	walk	run	owner	vw_1	vw_2	vw_3	vw_4
dog	3	5	2	5	7	3	0	4
cat	0	3	3	2	5	5	0	3
lion	0	3	2	0	5	5	3	6
light	0	0	0	0	0	0	4	0
bark	1	0	0	2	7	2	0	2
car	0	0	1	3	0	1	1	0

Table: Textual and visual contexts

Outline

Feng and Lapata, 2010

- The model relies on the extraction of a single distributional model from the same mixed-media corpus, where text and visual words are same latent dimensions
- Text and visual words are represented in terms of the same shared latent dimensions (topics)

Cons

- The textual model has to be extracted from the same corpus images are taken from (no way to use state-of-the-art text-based model)
- Text context extraction methods must be compatible with the overall multimodal approach (e.g. difficult to integrate information from syntax)

Outline

The Distributional Memory

- Our off-the-shelf textual distributional semantic model
- Shown to be at the state of the art in many semantic tasks (Baroni and Lenci 2010)
- Available from: `http://clic.cimec.unitn.it/dm`

ESP game

- Invented by L. von Ahn (2003)
- 50k labeled images
- Labeled through a game:
  - two people are partnered
  - they both see the same image and the task is to agree on an appropriate word to label the image
  - once a word is entered by both partners, that word becomes a label for the image

Outline

Wordsim353

- (Approximately) continuous similarity judgment
- 203 noun pairs rated by 13 human subjects on a 0-10 similarity scale and averaged (our coverage 73%)
  - E.g.: *money-cash*, 9.08; *coast-hill*, 4.38; *stock-life*, 0.92
- (Spearman) correlation between cosine of angle between pair context vectors and the judgment averages

Performance

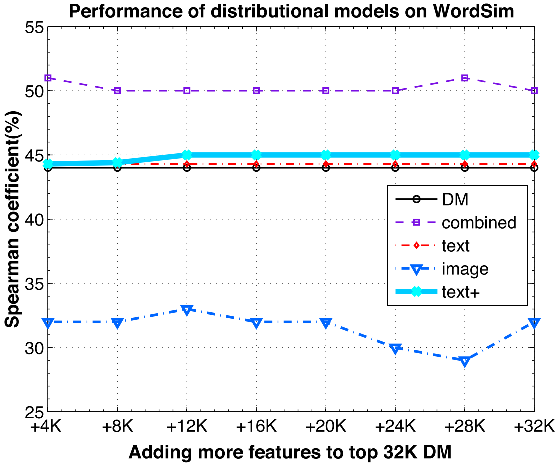


Figure: Wordsim performance

Concept categorization

- We used two **concept categorization** benchmarks, where the goal is to cluster a set of (nominal) concepts into broader categories
  - The **Almuhareb-Poesio** concept set (Almuhareb, 2006). In the version we cover, contains 230 concepts to be clustered into 21 classes such as vehicle (airplane, car...), time (aeon, future. . .) or social unit (brigade, nation)
  - The **Battig** set (Baroni et al., 2010), in the version we cover, contains 72 concepts to be clustered into 10 classes. Unlike AP, Battig only contains concrete basic-level concepts belonging to categories such as bird (eagle, owl...), kitchenware (bowl, spoon...) or vegetable (broccoli, potato...)

Performance

model	*AP*	*Battig*
DM	81	96
text	79	83
text+	80	86
image	25	36
combined	78	96

Table: Clustering performance

- Percentage AP and Battig purities of distributional models

BLESS

- Baroni-Lenci Evaluation of Semantic Similarity (BLESS) data set made available by the GEMS 2011 organizers
- In the version we cover, the data set contains 174 concrete nominal concepts
- Each concept paierd with a set of words that instantiate the following 5 relations: hypernymy (*spear*/*weapon*), coordination (*tiger*/*coyote*), meronymy (*castle*/*hall*), typical attribute (an adjective: *grapefruit*/*tart*) and typical event (a verb: *cat*/*hiss*)

Performance

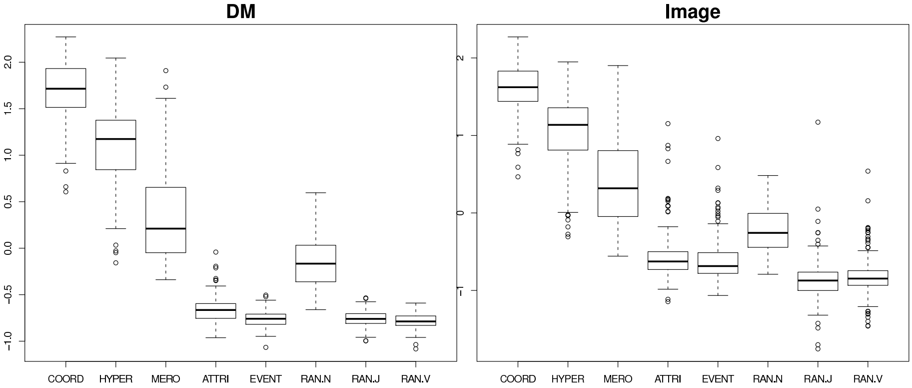


Figure: BLESS performance

Conclusion

- We proposed a simple method to augment a state-of-the-art text-based distributional semantic model with information extracted from image analysis
- Image-based models are more oriented towards capturing similarities between concrete concepts, and focus on their more imageable properties, whereas the text-based features are more geared towards abstract concepts and properties
- What next?
  - devising new benchmarks that address the special properties of image-enhanced models directly
  - combination techniques
  - new features

Thank you!

Questions?

Outline

Computing the angle

Example

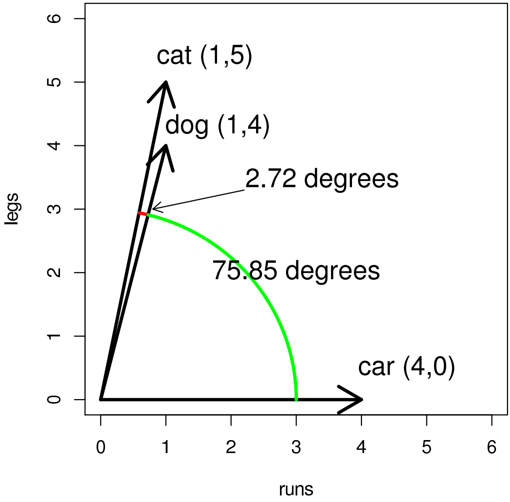


Figure: M. Baroni

SIFT feature vector

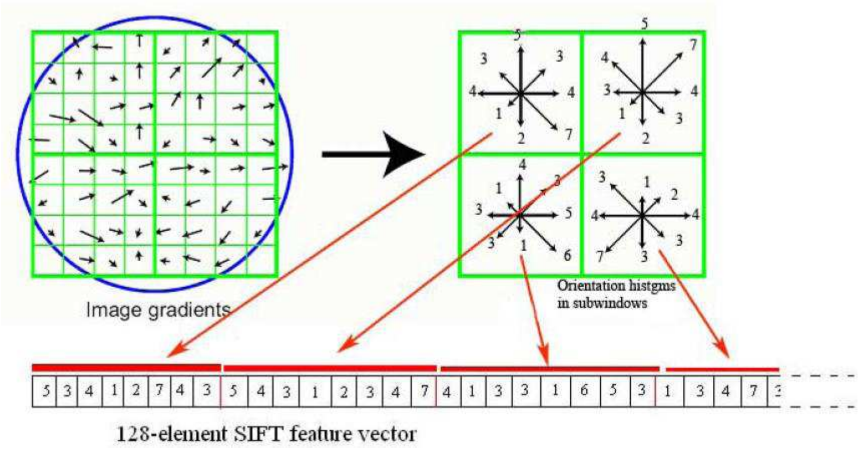


Figure: A visual feature

Challenges in basic features extraction

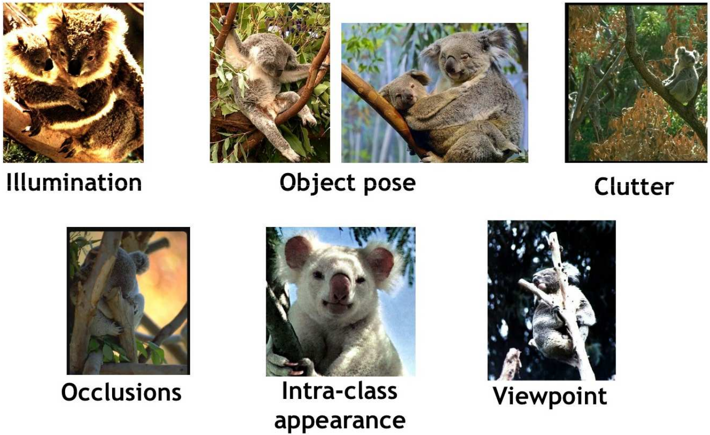


Figure: K. Grauman, B. Leibe

Nearest attributes

concept	*DM*	*image*
ant	small	black
cathedral	ancient	dark
pistol	dangerous	black
potato	edible	red
rifle	short	black
scooter	cheap	white
shirt	fancy	black
squirrel	fluffy	brown
truck	new	heavy
whale	large	gray

Table: Randomly selected cases where nearest attributes picked by DM and image differ