

# Computational Linguistics: Language and Vision II

RAFFAELLA BERNARDI

# Contents

# 1. Recall: Language

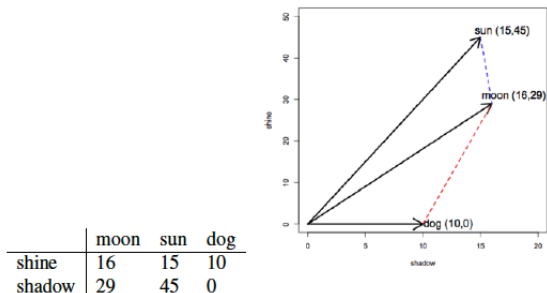
We found a cute, hairy wampimuk sleeping behind the tree.

what is a “wampimuk”?

We can understand the meaning of a word by its context.

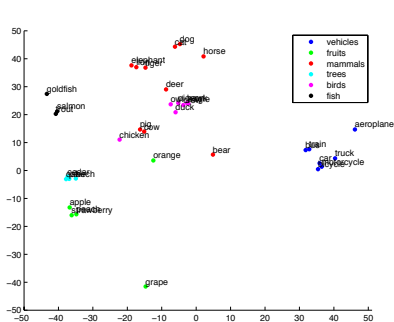
More generally, the meaning representation of a word is given by the words it occurs with. This info can be encoded into a vector.

$B = \{shadow, shine, \}$ ;  $A =$  frequency;  $S$ : angle measure (or Euclidean distance.)

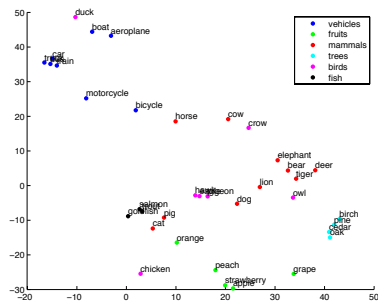


Smaller is the angle, more similar are the terms. (Cosine Similarity)

## 2. Language and Vision Spaces



Language



Vision

The two spaces are similar but different. We exploit both their similarity and their difference.

### 3. Similar: Exploit space similarity

Assumption: The two spaces encode similar information.

- ▶ Cross-Modal mappings provide semantic information about (unseen) concepts via the neighbour vectors of the vector projection.
- ▶ Images can be treated as visual phrases.
- ▶ Language Models can be used as prior knowledge for CV recognizers.

Deals with things not in the training data (“unseen”) by transferring in one modality knowledge acquired in the other (**generalization**).

## 3.1. Cross-modal mapping: Generalization

Angeliki Lazaridou, Elia Bruni and Marco Baroni. (ACL 2014)

**Generalization:** transferring knowledge acquired in one modality to the other one.

Learn to project one space into the other, from the visual space onto the language space.

- ▶ Learning: they use a set of  $N$ s seen concepts for which we have both image-based visual representations and linguistics vectors.
- ▶ The projection function is subject to an objective that aims at minimizing some cost function between the induced text-based representations.
- ▶ Testing: The induced function is then applied to the image-based representations of unseen objects to transform them into text-based representations.

## 3.2. Cross-modal mappings: Two tasks

- ▶ **Zero-Shot Learning:**
- ▶ **Fast Mapping:**

In both tasks, the projected vector of the unseen concept is labeled with the word associated to its cosine-based nearest neighbor vector in the corresponding semantic space.

### 3.3. Zero-Shot Learning

Learn a classifier  $X \rightarrow Y$ , s.t.  $X$  are images,  $Y$  are language vectors. Label an image of an unseen concept with the word associated to its cosine-based nearest neighbor vector in the language space.

For a subset of concepts (e.g., a set of animals, a set of vehicles), we possess information related to both their linguistic and visual representations.

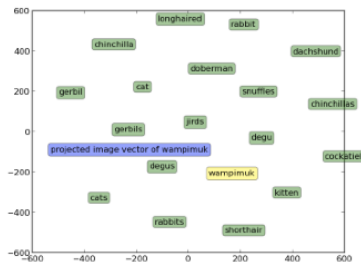
During training, this cross-modal vocabulary is used to induce a projection function, which intuitively represents a mapping between visual and linguistic dimensions.

Thus, this function, given a visual vector, returns its corresponding linguistic representation.

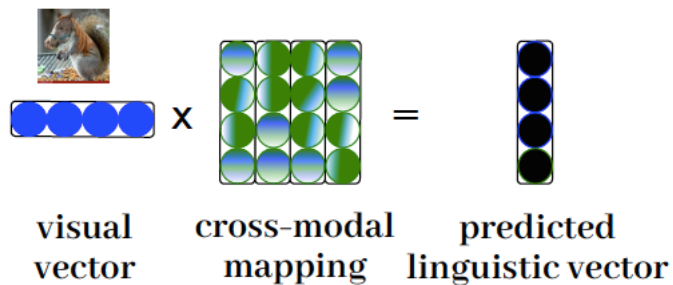
At test time, the system is presented with a previously unseen object (e.g., wampimuk). This object is projected onto the linguistic space and associated with the word label of the nearest neighbor in that space (containing all the unseen and seen concepts).



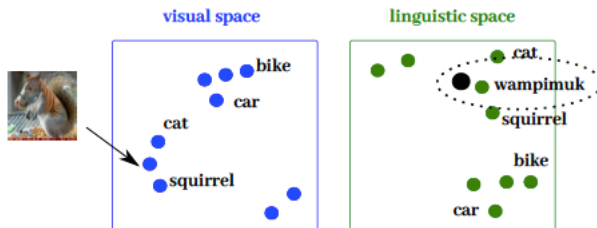
## 3.4. Zero-Shot Learning: the task



### 3.5. Zero-shot learning: linear mapping



### 3.6. Zero-shot learning: example



Step 1 Obtain “parallel data” of **linguistic** and **visual** vectors of concepts.

Step 2 Learn a cross-modal mapping between the two semantic spaces

Step 3 Map the **unknown** concept onto the **linguistic/visual** space

Step 4 Obtain a label through **nearest neighbor search**



### 3.7. Dataset

| Category                       | Seen Concepts                         | Unseen (Test) Concepts |
|--------------------------------|---------------------------------------|------------------------|
| aquatic mammals                | beaver, otter, seal, whale            | dolphin                |
| fish                           | ray, trout                            | shark                  |
| flowers                        | orchid, poppy, sunflower, tulip       | rose                   |
| food containers                | bottle, bowl, can, plate              | cup                    |
| fruit vegetable                | apple, mushroom, pear                 | orange                 |
| household electrical devices   | keyboard, lamp, telephone, television | clock                  |
| household furniture            | chair, couch, table, wardrobe         | bed                    |
| insects                        | bee, beetle, caterpillar, cockroach   | butterfly              |
| large carnivores               | bear, leopard, lion, wolf             | tiger                  |
| large man-made outdoor things  | bridge, castle, house, road           | skyscraper             |
| large natural outdoor scenes   | cloud, mountain, plain, sea           | forest                 |
| large omnivores and herbivores | camel, cattle, chimpanzee, kangaroo   | elephant               |
| medium-sized mammals           | fox, porcupine, possum, skunk         | raccoon                |
| non-insect invertebrates       | crab, snail, spider, worm             | lobster                |
| people                         | baby, girl, man, woman                | boy                    |
| reptiles                       | crocodile, dinosaur, snake, turtle    | lizard                 |
| small mammals                  | hamster, mouse, rabbit, shrew         | squirrel               |
| vehicles 1                     | bicycle, motorcycle, train            | bus                    |
| vehicles 2                     | rocket, tank, tractor                 | streetcar              |

Table 1: Concepts in our version of the CIFAR-100 data set

## 3.8. Fast Mapping

Learn a word vector from a few sentences, associate it to the referring image exploiting cosine-based neighbor vector in the visual space.

The fast mapping setting can be seen as a special case of the zero-shot task. Whereas for the latter our system assumes that all concepts have rich linguistic representations (i.e., representations estimated from a large corpus), in the case of the former, new concepts are assumed to be encountered in a limited linguistic context and therefore lacking rich linguistic representations.

This is operationalized by constructing the text-based vector for these concepts from a context of just a few occurrences. In this way, we simulate the first encounter of a learner with a concept that is new in both visual and linguistic terms.

New paper: **Multimodal semantic learning from child-directed input** Angeliki Lazaridou, Grzegorz Chrupala, Raquel Fernandez and Marco Baroni NAACL 2016 Short <http://clic.cimec.unitn.it/marco/publications/lazaridou-etal-multimodal.pdf>

## 3.9. Images as Visual Phrases

- ▶ Given the visual representation of an object, can we “decompose” it into attribute and object?
- ▶ Can we learn the visual representation of attributes and learn to compose them with the visual representation of an object?

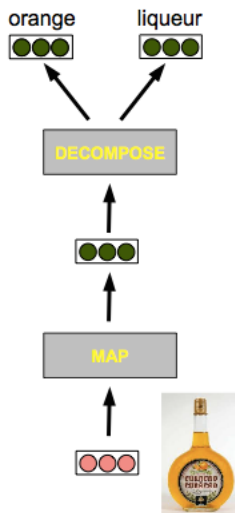
## 3.10. Visual Phrase: Decomposition

A. Lazaridou, G. Dinu, A. Liska, M. Baroni (TACL 2015)

- ▶ First intuition: vision and language space have similar structures (also w.r.t attribute/adjectives)
- ▶ Second intuition: Objects are bundles of attributes. Hence, attributes are implicitly learned together with objects.



### 3.11. Decomposition Model: attribute annotation



Evaluation: (unseen) object/noun and attribute/adjective retrieval.

## 3.12. Images as Visual Phrases: Composition

Coloring Objects: Adjective-Noun Visual Semantic Compositionality (VL'14)

D.T. Nguyen, A. Lazaridou and R. Bernardi

1. Assumption from linguistics: Adjectives are noun modifiers. They are functions from  $N$  into  $N$ .
2. From COMPOSES: adjectives can be learned from  $(ADJ\ N, N)$  inputs.
3. Applied to images: Compositional Visual Model?

### 3.13. Visual Composition

From the visual representation:

- ▶ Dense-Sift feature vectors as Noun vectors (e.g. car. light)
- ▶ Color-Sift feature vectors as Phrase vectors (e.g. red car. red light)

Learn the function (color) that maps the noun to the phrase. Apply that function to new (unseen) objects (e.g. red truck) and retrieve the image.

We compare the the composed visual vector (ATT OBJ) vs. composed linguistic vectors (ADJ N) vs. observed linguistic vectors.

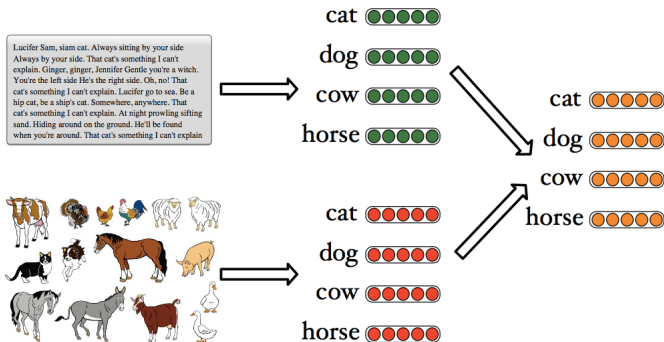
### 3.14. Coloring Objects: Results

|                                  | > 10 images | > 20 images | > 30 images |
|----------------------------------|-------------|-------------|-------------|
| $V_{phrase}^{comp} - V_{phrase}$ | 0.40        | 0.53        | <b>0.58</b> |
| $V_{phrase}^{comp} - W_{phrase}$ | 0.22        | 0.19        | <b>0.23</b> |

(Experiments: with Colors only).

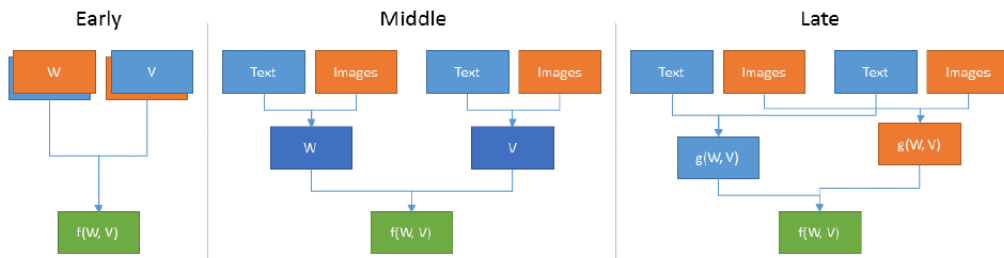
## 4. Different: Exploit differences

Assumption: The two spaces provide complementary information about concepts.



Multi-modal vectors are closer to human representations (**better quality**).



## 4.1. Multimodal fusion: approaches



- We need to perform **fusion** of textual and perceptual information.
  - Early: learn jointly, then compute function
  - Middle: learn separately, then combine, then compute function
  - Late: learn separately, compute function individually and combine function outputs

## 4.2. Multi-modal Semantics Models: Concatenation

E. Bruni, G.B. Tran and M. Baroni (GEMS 2011, ACL 2012, Journal of AI 2014)

|      | planet | night |  |  |
|------|--------|-------|---|---|
| moon | 10     | 22    | 22  | 0   |
| sun  | 14     | 10    | 15  | 0   |
| dog  | 0      | 4     | 0   | 20  |

- 1 Combine uni-modal representations



- 2 Compute function over multi-modal inputs, e.g. cosine

### 4.3. Multi-modal models: drawbacks

- ▶ First, they are generally constructed by first separately building linguistic and visual representations of the same concepts, and then merging them. This is obviously very different from how humans learn about concepts, by hearing words in a situated perceptual context.
- ▶ Second, MDSMs assume that both linguistic and visual information is available for all words, with no generalization of knowledge across modalities.
- ▶ Third, because of this latter assumption of full linguistic and visual coverage, current MDSMs, paradoxically, cannot be applied to computer vision tasks such as image labeling or retrieval, since they do not generalize to images or words beyond their training set.



## 5. Similar and Different

- ▶ **Cross-modal Mapping:** Generalization (transfer in one modality knowledge acquired in the other).
- ▶ **Multi-modal Models:** Grounded representation. Better quality.

Can we have both better quality and generalization?

## 5.1. Multimodal Skip-gram Model

Lazaridou, Pham, Baroni (NAACL 2015)

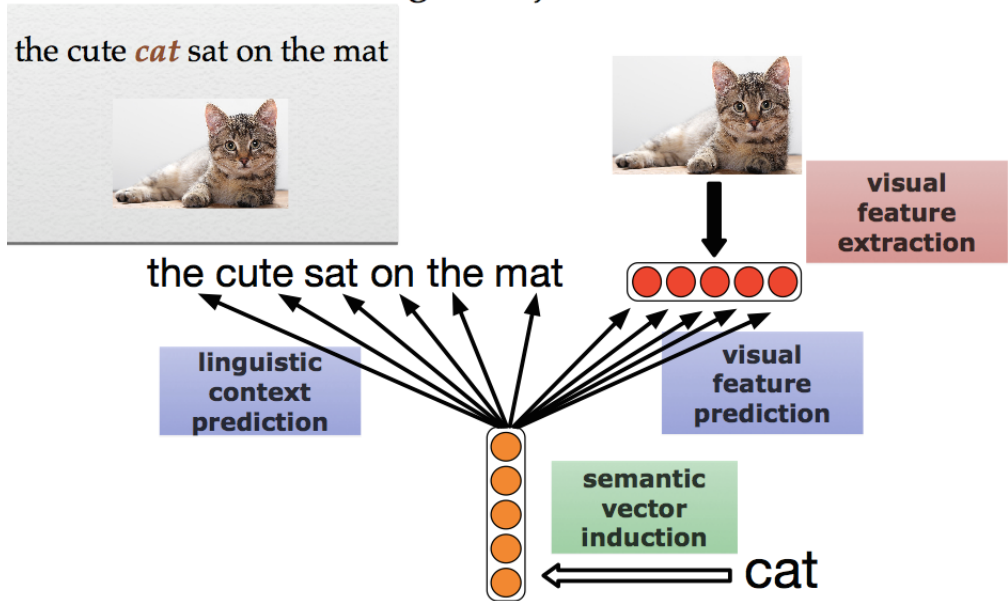
**Skip-Gram** Mikolov et al. (2013), constructs vector representations by learning, incrementally, to predict the linguistic contexts in which target words occur in a corpus.

**MMSKI-Gram** builds vector-based word representations by learning to predict linguistic contexts in text corpora. However, for a restricted set of words,

- ▶ the models are also exposed to visual representations of the objects they denote (extracted from natural images), and
- ▶ must predict linguistic and visual features jointly.
- ▶ The joint objective encourages the propagation of visual information to representations of words for which no direct visual evidence was available in training. The resulting multimodally-enhanced vectors achieve remarkably good performance both on traditional semantic benchmarks, and in their new application to the zero-shot image labeling and retrieval scenario

## 5.2. Multi-modal Skip-gram Model

Learning from joint contexts



## 5.3. Multi-modal Skip-gram Model

- ▶ Better quality vector representation tested against:
  - ▷ Word similarity (MEN, SimLex-999, SemSim and VisSim)
- ▶ Generalization tested against:
  - ▷ Image Retrieval of (unseen) objects

## 5.4. MMSkip-gram

| <i>Target</i> | SKIP-GRAM                      | MMSKIP-GRAM-A                 | MMSKIP-GRAM-B                 |
|---------------|--------------------------------|-------------------------------|-------------------------------|
| donut         | fridge, diner, candy           | pizza, sushi, sandwich        | pizza, sushi, sandwich        |
| owl           | pheasant, woodpecker, squirrel | eagle, woodpecker, falcon     | eagle, falcon, hawk           |
| mural         | sculpture, painting, portrait  | painting, portrait, sculpture | painting, portrait, sculpture |
| tobacco       | coffee, cigarette, corn        | cigarette, cigar, corn        | cigarette, cigar, smoking     |
| depth         | size, bottom, meter            | sea, underwater, level        | sea, size, underwater         |
| chaos         | anarchy, despair, demon        | demon, anarchy, destruction   | demon, anarchy, shadow        |

Table 2: Ordered top 3 neighbours of example words in purely textual and multimodal spaces. Only *donut* and *owl* were trained with direct visual information.

## 5.5. Multimodal Models: Evaluation Tasks

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

Improved!

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

Improved!

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

Improved!

**Task 4** Distinguish **literal vs. non-literal** usages of color adjectives (**blue uniform** vs. **blue note**)

Improved!

## 5.6. Multi-modal models: predicting colors

E. Bruni, G. Boleda, M. Baroni and N. Tran (ACL 2012)

| <i>word</i> | <i>gold</i> | <i>LAB</i> | <i>SIFT</i> | <i>TEXT</i> |
|-------------|-------------|------------|-------------|-------------|
| banana      | yellow      | yellow     | blue        | orange      |
| cauliflower | white       | green      | yellow      | orange      |
| cello       | brown       | brown      | black       | blue        |
| deer        | brown       | green      | blue        | red         |
| froth       | white       | brown      | black       | orange      |
| gorilla     | black       | black      | red         | grey        |
| grass       | green       | green      | green       | green       |
| pig         | pink        | pink       | brown       | brown       |
| sea         | blue        | blue       | blue        | grey        |
| weed        | green       | green      | yellow      | purple      |

## 5.7. Application: predict concreteness

D. Kiela, F. Hill, A. Korhonen and S. Clark (2014) Improving multimodal representation using image dispersion: Why less is sometimes more. ALC 2014





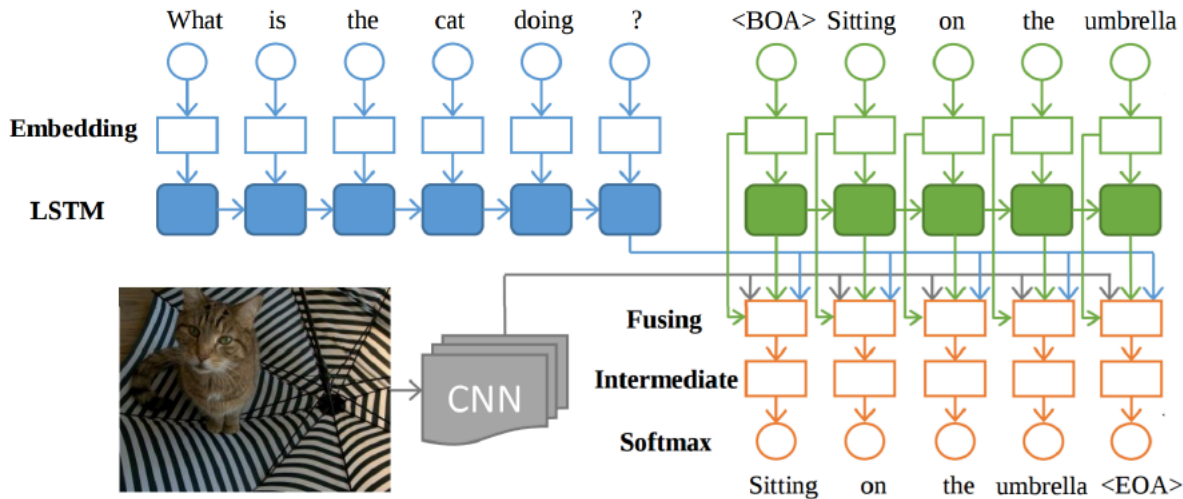
## 5.8. Application: methaphor detection

Shutova et al 2016

| [Mohammad et al., 2016] - SV/VO |              |
|---------------------------------|--------------|
| blister foot                    | literal      |
| blister administration          | metaphorical |
| blur vision                     | literal      |
| blur distinction                | metaphorical |
| [Tsvetkov et al., 2014] - AN    |              |
| cold beer                       | literal      |
| cold heart                      | metaphorical |
| foggy morning                   | literal      |
| foggy brain                     | metaphorical |

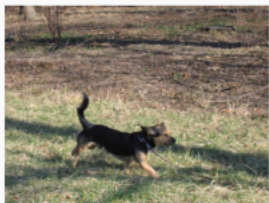


## 6. VQA



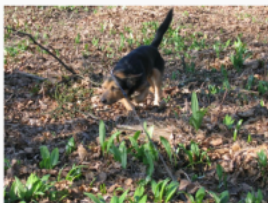
## 7. Visual Story Telling

1



**The dog was ready to go.**

2



**He had a great time on  
the hike.**

3



**And was very happy to be  
in the field.**

Photos by kameraschwein / CC BY-NC-ND 2.0



## 8. FOIL it!

task 1:  
classification



People riding bicycles down  
the road approaching a dog.

**FOIL**

task 2:  
foil word detection



People riding bicycles down  
the road approaching a **dog**.

task 3:  
foil word correction



People riding bicycles down  
the road approaching a **bird**.

## 9. Administrativa

- ▶ Next week (26th) last frontal class: on going work on Vision and quantities at CIMEC/clic.
- ▶ 11th of May 15:00-18:00 (aula 1): Project presentation
- ▶ 17th of May 10:30-12:30 (aula 1): written exercises

## 10. Open questions from last time

- ▶ L1 loss function is also known as least absolute deviations (LAD), least absolute errors (LAE). It is basically minimizing the sum of the absolute differences between the target value and the estimated values
- ▶ L2-norm loss function is also known as least squares error (LSE). It is basically minimizing the sum of the **square** of the differences between the target value and the estimated values.