Is this a wampimuk?
Cross-modal mapping between distributional semantics and the visual world

Angeliki Lazaridou  Elia Bruni  Marco Baroni

University of Trento

ACL 2014
Computational Semantics Milestones

Distributional Hypothesis

Harris, Firth ...
Distributional Hypothesis

From theory...

We found a cute, hairy wampimuk sleeping behind the tree
## Distributional Hypothesis

... to today’s practise

|   | planet | night | full | shadow | shine | crescent |
|---|--------|-------|------|--------|-------|----------|
| moon | 10     | 22    | 43   | 16     | 29    | 12       |
| sun  | 14     | 10    | 4    | 15     | 45    | 0        |
| dog  | 0      | 4     | 2    | 10     | 0     | 0        |
Computational Semantics Milestones

**Distributional Hypothesis**
- Harris, Firth ...

**Hyperspace Analogue to Language**
- Lund and Burgess, 1996

**Latent Semantic Analysis**
- Landauer and Dumais, 1997

**Topic Models**
- Griffiths et al, 2007

**Neural Language Models**
- Bengio et al., 2003; Collebert and Weston, 2008; Mikolov et al., 2013
Are current models **cognitively plausible** mechanisms of language acquisition and usage?
Landauer and Dumais, 1997; Lenci 2008

- **Grounding Problem**
  - Limited in capturing the **holistic knowledge** about concepts
Grounding problem: Towards a solution
Feng and Lapata, 2010; Siblerer et al, 2013; Bruni et al, 2014; inter alia

- Enrichment of pure textual vectors with complementary information coming from perceptual visual features.
Are current models **cognitive plausible** mechanisms of language acquisition and usage?
Landauer and Dumais, 1997; Lenci 2008

- **Grounding Problem**
  - Limited in capturing the **holistic knowledge** about concepts
- **Lack of Reference**
  - Provide **no links** to the external world.
Why should we care?: Referent selection during language acquisition

Fast Mapping (Carey, 1978; Bloom, 2000; Alishahi et al. 2008)

- **Young learners** are able to select the correct referent of an *unfamiliar* word even from the very *first exposure* to it.
From fast mapping to zero-shot\textsuperscript{1}

Using a powerful text-based vector model

\textit{Wampimuk is semantically similar to a cat.}

\textbf{Is there a wampimuk in the room?}

\textsuperscript{1}For example, for executing natural language instructions (Branavan et al., 2009; Chen and Mooney, 2011)
From fast mapping to zero-shot
Using a powerful object recognition component

This looks like a cat.

Is there a wampimuk in the room?
From fast mapping to zero-shot
Knowledge transfer from one modality to another

Wampimuk is semantically similar to a cat. This looks like a cat. 

**THUS, I know that this might be a wampimuk.**

Is there a wampimuk in the room?
Visual or textual space?
Visual and Textual Semantic Spaces

(a) Visual Semantic Space

(b) Textual Semantic Space

\[ r_{corr} = 0.5 \]

0.5 correlation of pairwise distances in these spaces
Referent selection: Towards a solution

Cross-modal mapping (Frome et al., 2013; Socher et al., 2013)
Referent selection: Towards a solution
Cross-modal mapping (Frome et al., 2013; Socher et al., 2013)
Cross-Modal Mapping function

Neural Network
\[ f_{\text{proj}_{v \rightarrow w}} = \Theta_{v \rightarrow w} \]

Linear Regression
\[ f_{\text{proj}_{v \rightarrow w}} = (V_s^T V_s)^{-1} V_s^T W_s \]

CCA
\[ f_{\text{proj}_{v \rightarrow w}} = CV CW^{-1} \]

SVD
\[ f_{\text{proj}_{v \rightarrow w}} = Z_k Z_k^T \]
Visual Datasets

- CIFAR
  - Evaluation of various cross-modal mapping functions on an object recognition benchmark dataset
  - Search space: 90 classes

- ESP
  - Assess robustness of cross-modal mapping
  - Non-iconic images, where objects appear at their natural context
  - 100 times larger search space than CIFAR.
Visual Datasets

A chair...

CIFAR

ESP
Evaluation Setup

Given the visual representation $v_i$ for a wampimuk:

- project it with $f_{\text{proj}_{v \to w}}$ onto the text-based semantic space
- obtain $w'_i$
- rank its semantic neighbors of $w'_i$ through some metric, e.g. cosine similarity
- $squirrel, kitten, \textbf{wampimuk} \rightarrow \text{rank}=3$
Experiment 1: Referent selection in Distributional Semantics
Zero-shot in CIFAR

| Model | k | 1 | 5 | 10 | 20 |
|-------|---|---|---|----|----|
| Chance| 1 | 6 | 11| 22 |
| SVD   | 2 | 15| 29| 49 |
| CCA   | 3 | 18| 32| 52 |
| lin   | 2 | 19| 33| 55 |
| NN    | 4 | 22| 38| 58 |

**Table**: Percentage accuracy of retrieving the correct image label among the k nearest neighbors.
Interpretability of Hidden Layer of NN

Training
- sunflower
- man
- plate
- bowl
- tulip
- girl
- can
- baby
- pear

Test
- butterfly
- boy
- clock

Input Layer  Hidden Layer  Output Layer
Interpretability of Hidden Layer of NN

**Training**
- sunflower
- man
- plate
- bowl
- tulip
- girl
- can
- baby
- pear

**Test**
- butterfly
- boy
- clock

**Diagram**
- Input Layer
- Hidden Layer
- Output Layer
Experiment 2: Cross-modal mapping on non-iconic images, where objects appear in their natural context

Zero-shot in ESP

| Model | k | 1   | 5   | 10  | 50  |
|-------|---|-----|-----|-----|-----|
| Chance| Chance | 0.01 | 0.05 | 0.10 | 0.5 |
| NN    | 1 | 6   | 10  | 31  |

**Table:** Percentage accuracy of retrieving the correct image label among the k nearest neighbors.
### Examples

| Target       | Nearest neighbors of mapped visual vector | Cohyponymy |
|--------------|------------------------------------------|------------|
| jellyfish    | anemone, jellyfish, seashell, conch, hammerhead |            |
| cow          | bison, elephant, baboon, rhinoceros, giraffe |            |
| phone        | headset, smartphone, microphone, earpiece, sony |            |
| instrument   | sitar, percussion, accordion, rhythm, xylophone |            |
| kiss         | happy, hate, dad, sweetheart, sad |            |
| participate  | cheese, sour, refrigerate, cooking, ketchup |            |
### Examples

| Target         | Nearest neighbors of mapped visual vector |
|----------------|------------------------------------------|
| jellyfish      | anemone, jellyfish, seashell, conch      |
|                | hammerhead                               |
| cow            | bison, elephant, baboon, rhinoceros, giraffe |
| phone          | headset, smartphone, microphone, earpiece, sony |
| instrument     | sitar, percussion, accordion, rhythm, xylophone |
| kiss           | happy, hate, dad, sweetheart, sad       |
| participate    | cheese, sour, refrigerate, cooking, ketchup |
### Examples

| Target          | Nearest neighbors of mapped visual vector |
|-----------------|------------------------------------------|
| jellyfish       | anemone, jellyfish, seashell, conch, hammerhead |
| cow             | bison, elephant, baboon, rhinoceros, giraffe |
| phone           | headset, smartphone, microphone, earpiece, sony |
| instrument      | sitar, percussion, accordion, rhythm, xylophone | **hyponymy** |
| kiss            | happy, hate, dad, sweetheart, sad |
| participate     | cheese, sour, refrigerate, cooking, ketchup |
## Examples

| **Target**     | Nearest neighbors of mapped visual vector |
|----------------|------------------------------------------|
| jellyfish      | anemone, jellyfish, seashell, conch, hammerhead |
| cow            | bison, elephant, baboon, rhinoceros, giraffe |
| phone          | headset, smartphone, microphone, earpiece, sony |
| instrument     | sitar, percussion, accordion, rhythm, xylophone |
| kiss           | happy, hate, dad, sweetheart, sad, adjectives, verbs |
| participate    | cheese, sour, refrigerate, cooking, ketchup |
### Examples

| Target         | Nearest neighbors of mapped visual vector |
|----------------|------------------------------------------|
| jellyfish      | anemone, jellyfish, seashell, conch      |
| hammerhead     |                                          |
| cow            | bison, elephant, baboon, rhinoceros, giraffe |
| phone          | headset, smartphone, microphone, earpiece, sony |
| instrument     | sitar, percussion, accordion, rhythm, xylophone |
| kiss           | happy, hate, dad, sweetheart, sad       |
| participate    | cheese, sour, refrigerate, cooking, ketchup | weird? events?
Experiment 3: Simulating a fast mapping scenario

- Is the model able to do referent selection with minimal exposure to the linguistic input just like children do?
- Regulate the amount of context we use to construct the text-based vectors with 1, 5, 10, 20 sentences used as well as the full corpus.
- $v \rightarrow w$: first visual encounter with the object, then search for its referent in the on-going spoken discourse.
- $w \rightarrow v$: first exposed to a new word, then search for its referent in the on-going visual discourse.
Fast mapping in ESP

![Graph showing mean rank vs. number of sentences for different mapping types.]
Discuss the referent selection problem by exploit common structure of modalities to learn a cross-modal mapping.

Comparison of recently proposed models on a visual recognition dataset.

Evaluation of cross-modal mapping on a larger dataset with non-iconic images

- Paves the way to applications of cross-modal mapping for more complex tasks, e.g. caption generation/retrieval

Preliminary experiments towards assessing viability of cross-modal mapping as a grounded word-meaning acquisition mechanism.
Future Work

- Exploit to a greater extent the common and hierarchical structure of modalities
  - Deep Boltzmann Machines, structured regularizers, unsupervised alignment
- More realistic simulations of fast mapping experiments
  - Designing of novel-word experiments
  - Use of corpora with child-directed-like speech, e.g. CHILDES, Simple Wikipedia
Thank you!

Questions?\(^3\)

\(^3\)Apart from what a wampimuk really is?? :-)

Lazaridou, Bruni, Baroni (University of Trento)  Cross-modal Mapping in Distributional Semantics  ACL 2014  34 / 34