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...

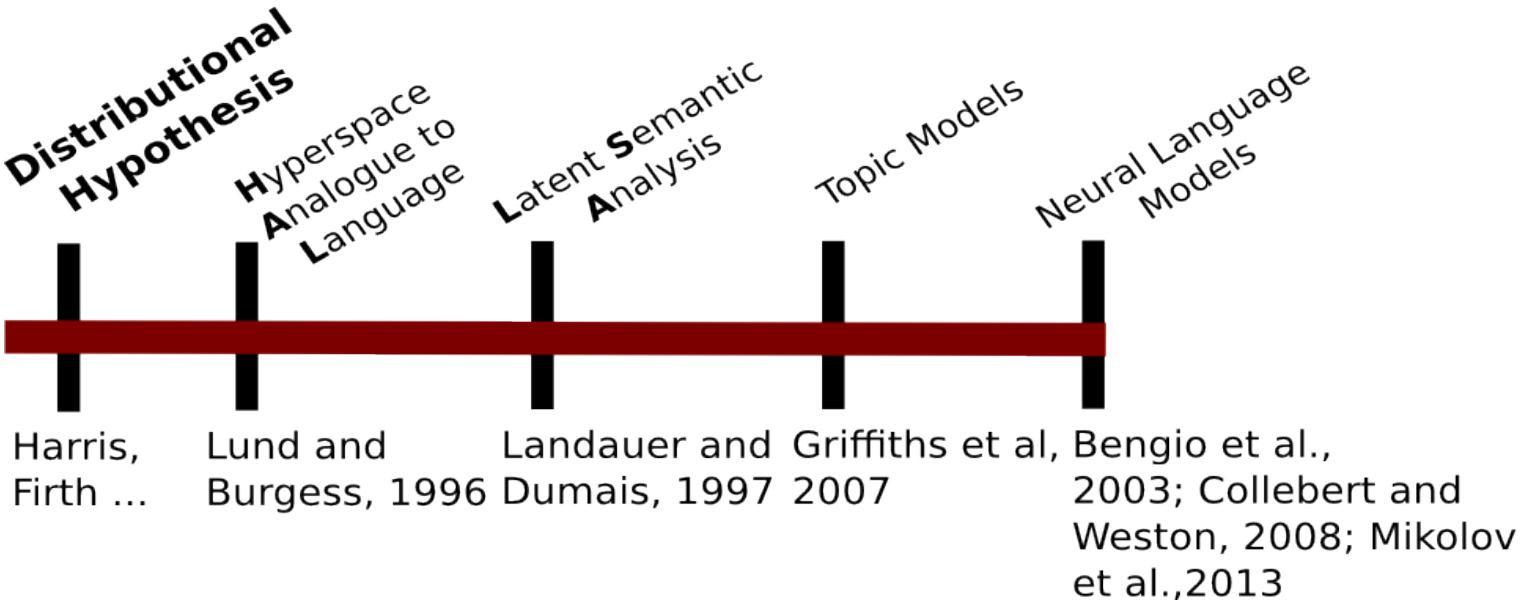
We found a cute, hairy wampimuk sleeping behind the tree

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



Are current models **cognitively plausible** mechanisms of language acquisition and usage?

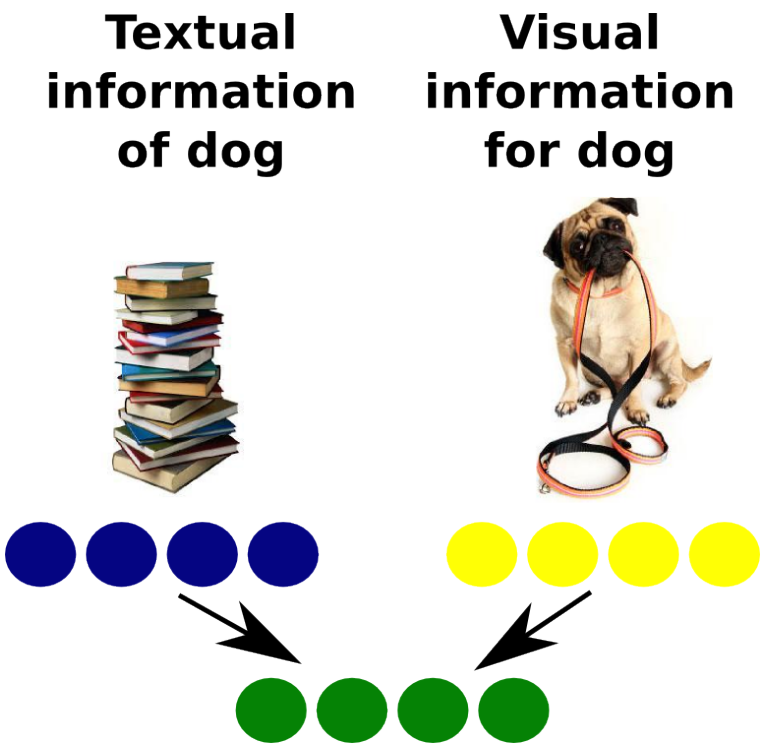
Landauer and Dumais, 1997; Lenci 2008

- 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.

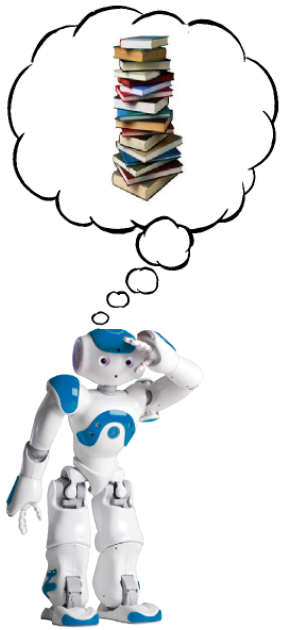
Why should we care?: Referent selection during language acquisition

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[1]

Using a powerful text-based vector model



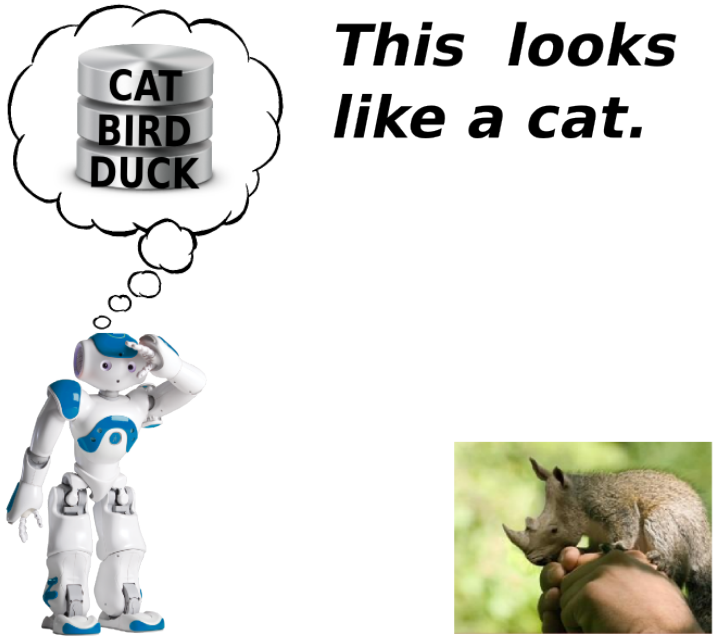
Wampimuk is semantically similar to a cat.

Is there a wampimuk in the room?

[1] For example, for executing natural language instructions (Branavan et al., 2009; Chen and Mooney, 2011)

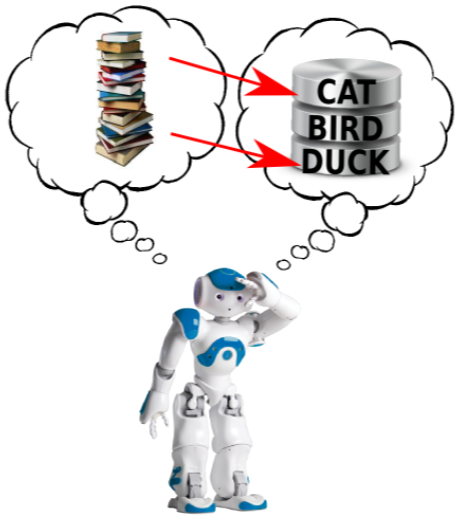
From fast mapping to zero-shot

Using a powerful object recognition component



From fast mapping to zero-shot

Knowledge tranfer from one modality to another

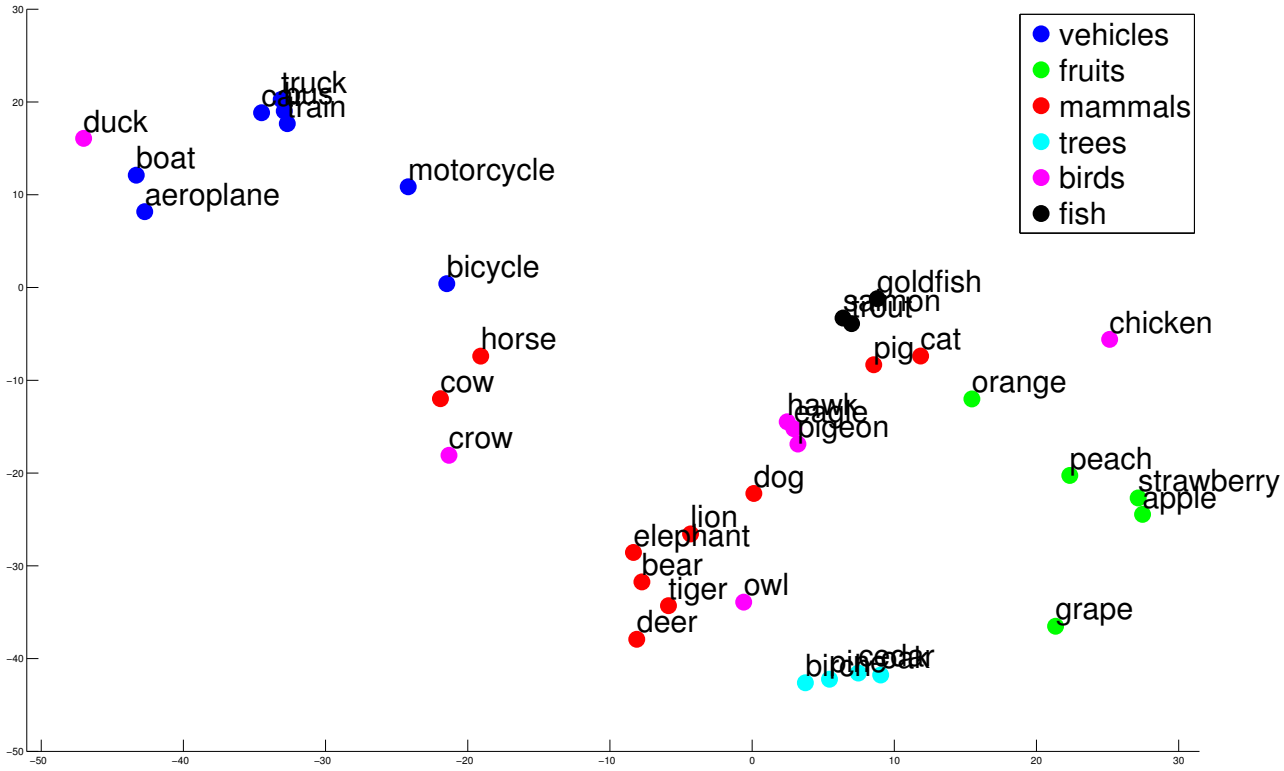


Wampimuk is semantically similar to a cat.
This looks like a cat.

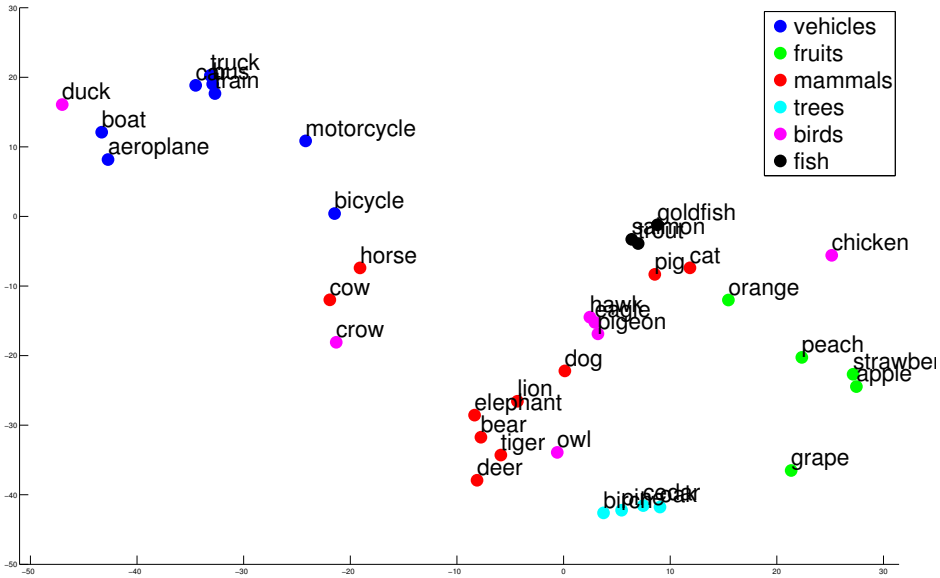
THUS, i know that this might be a wampimuk.

Is there a wampimuk in the room?

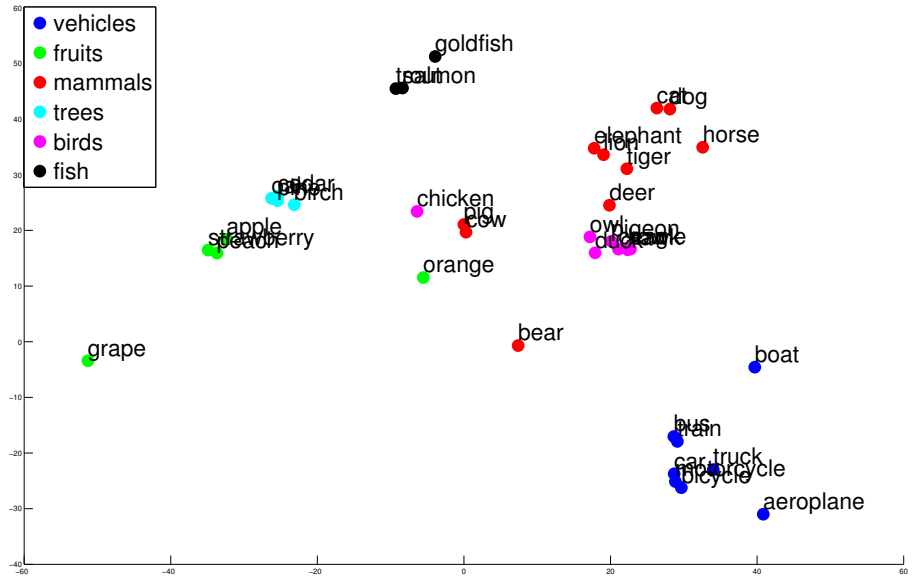
Visual or textual space?



Visual and Textual Semantic Spaces[2]



(a) Visual Semantic Space

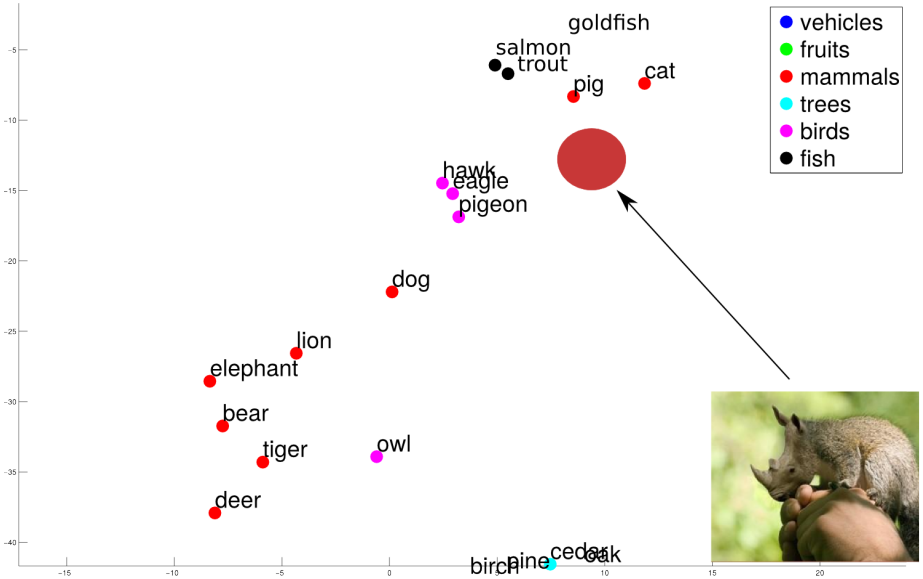


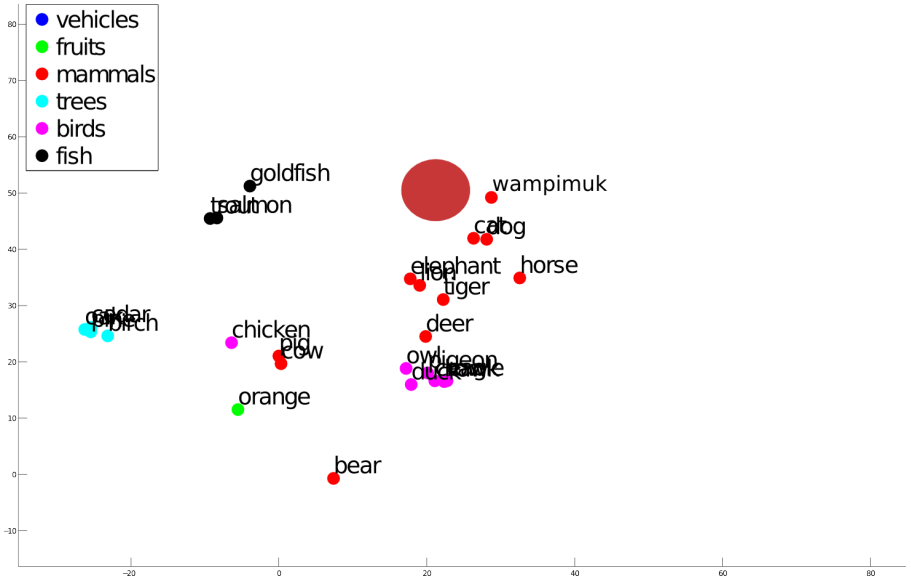
(b) Textual Semantic Space

[2] 0.5 correlation of pairwise distances in these spaces

Referent selection: Towards a solution

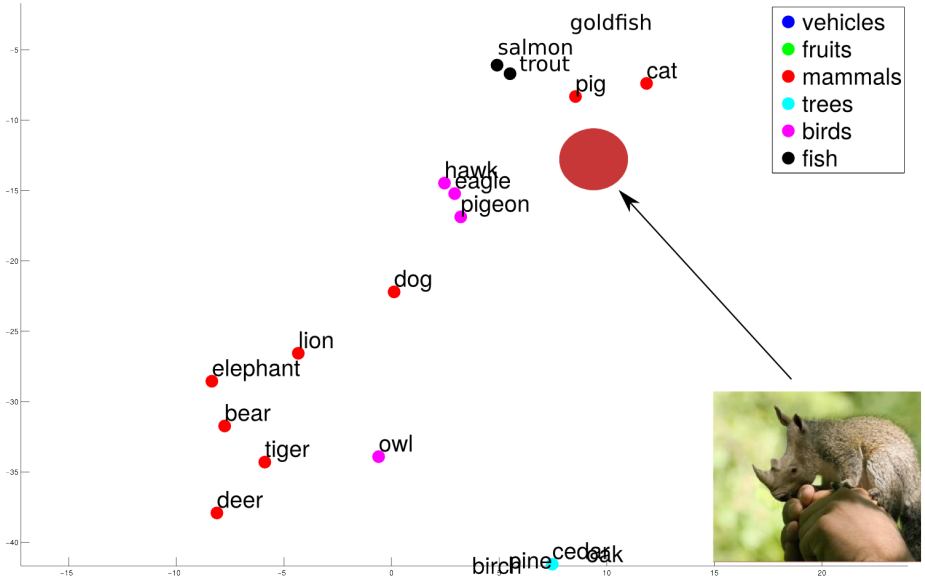
Cross-modal mapping (Frome et al., 2013; Socher et al., 2013)

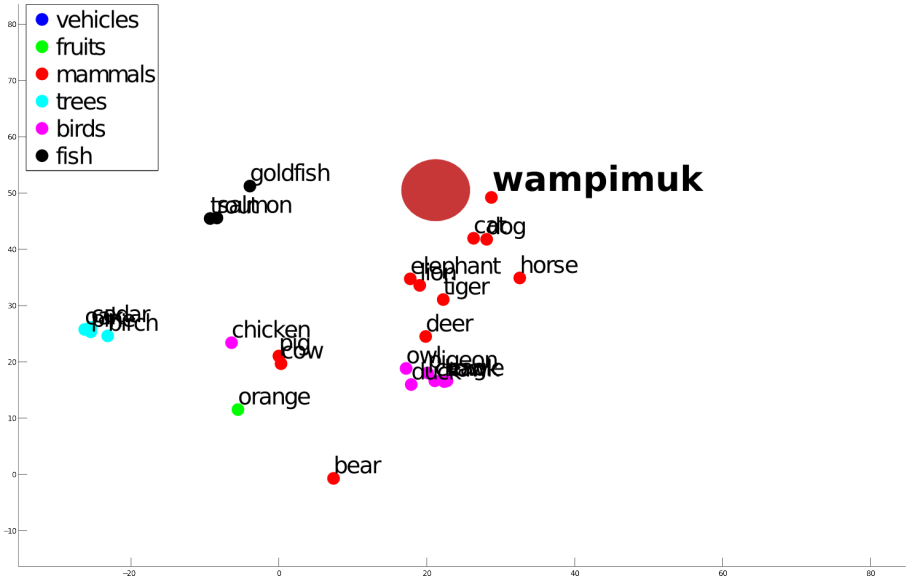




Referent selection: Towards a solution

Cross-modal mapping (Frome et al., 2013; Socher et al., 2013)





Cross-Modal Mapping function

Neural Network	$\mathrm{f}_{\mathrm{proj}_{\mathrm{v}\rightarrow\mathrm{w}}} = \mathbf{\Theta}_{\mathrm{v}\rightarrow\mathrm{w}}$
Linear Regression	$\mathrm{f}_{\mathrm{proj}_{\mathrm{v}\rightarrow\mathrm{w}}} = (\mathbf{V}_s^T\mathbf{V}_s)^{-1}\mathbf{V}_s^T\mathbf{W}_s$
CCA	$\mathrm{f}_{\mathrm{proj}_{\mathrm{v}\rightarrow\mathrm{w}}} = \mathbf{C}_V{\mathbf{C}_W}^{-1}$
SVD	$\mathrm{f}_{\mathrm{proj}_{\mathrm{v}\rightarrow\mathrm{w}}} = \mathbf{Z}_k\mathbf{Z}_k^T$

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 $\mathrm{f}_{\mathrm{proj}_{\mathrm{v}\rightarrow\mathrm{w}}}$ 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*, **wampimuk** $\rightarrow$ rank=3

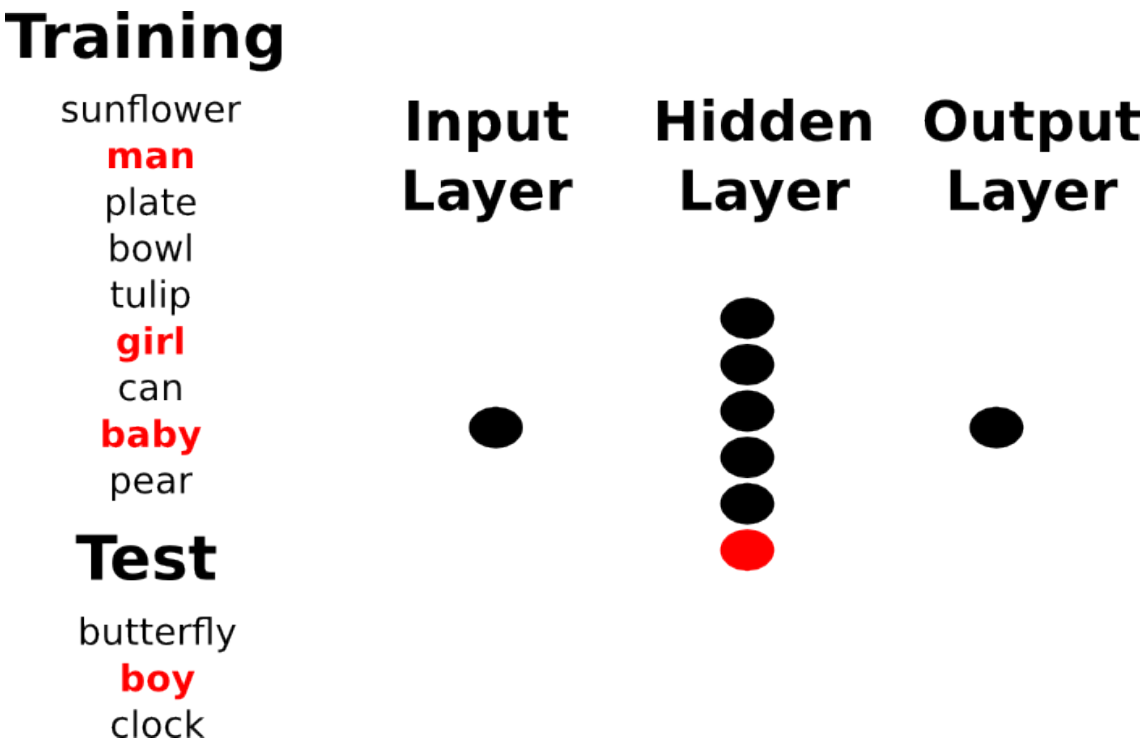
Experiment1: Referent selection in Distributional Semantics

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



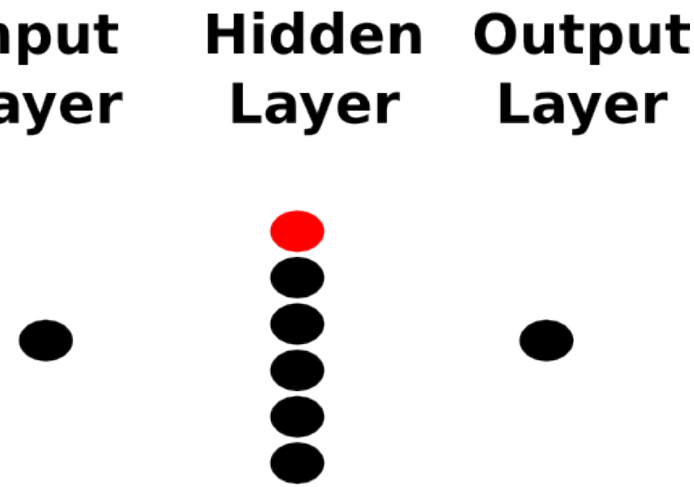
Interpretability of Hidden Layer of NN

Training

sunflower
man
plate
bowl
tulip
girl
can
baby
pear

Test

butterfly
boy
clock



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	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*	
jellyfish	**anemone, jellyfish, seashell, conch, hammerhead**	**cohyponymy**
cow	**bison, elephant, baboon, rhinoceros, giraffe**	**cohyponymy**
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

meronymy

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	

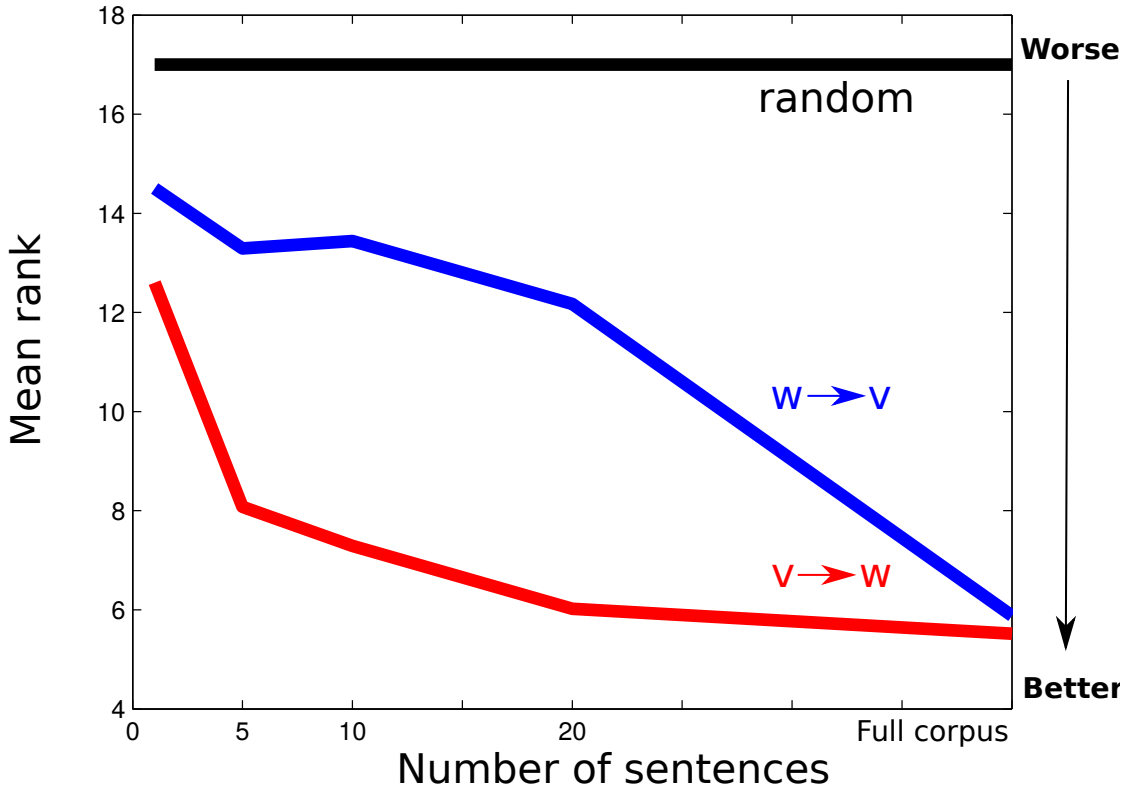
Examples

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



Discussion

- Tackle the referent selection problem by exploit common structure of modalities to learn a cross-modal mapping.
- Comparison of recenlty 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?? :-)