Linguistic interpretability in neural models of grounded language learning

Grzegorz Chrupała

EMNLP Workshop on
Building Linguistically Generalizable NLP Systems

Tilburg University

In collaboration with

- Afra Alishahi

- Ákos Kádár

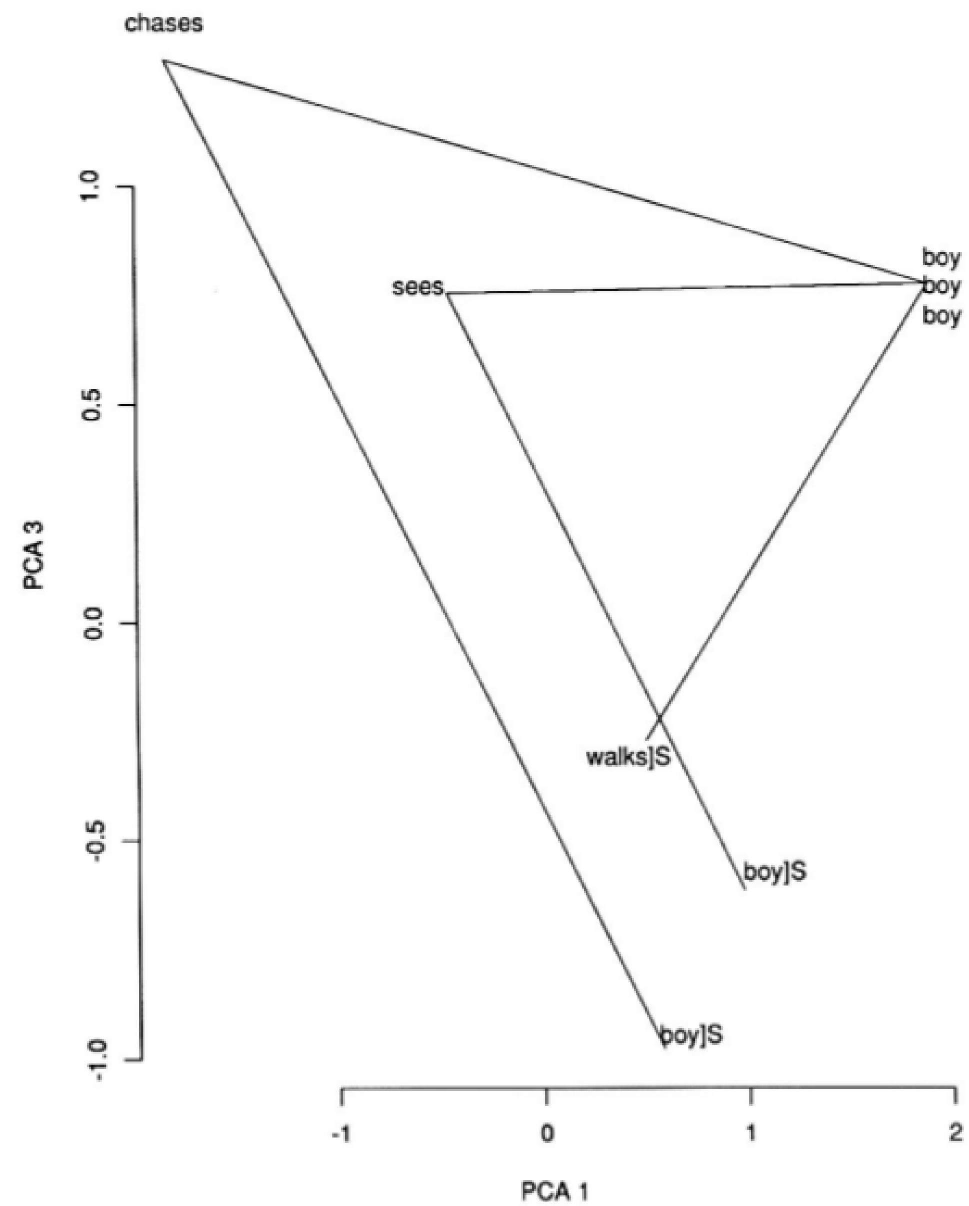
- Lieke Gelderloos

- Marie Barking

TILBURG UNIVERSITY

Understanding neural representations of language

- What representations emerge in neural nets?
- How much do they much linguistic analyses?
- Which parts of the architecture encode what?



Jeffrey L Elman. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning* 7(2-3):195–225.

Some modern work

Learning objectives

Language modeling

- Linzen et al. 2016

Sentiment classification

- Li et al. 2016a, 2016b

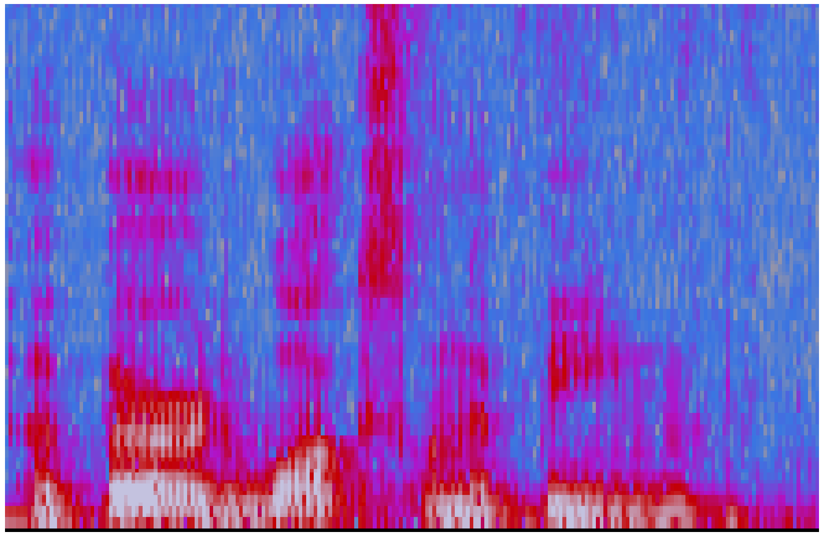
Autoencoding

- Adi et al. 2016

Translation

- Belinkov et al. 2017

Visually grounded language learning

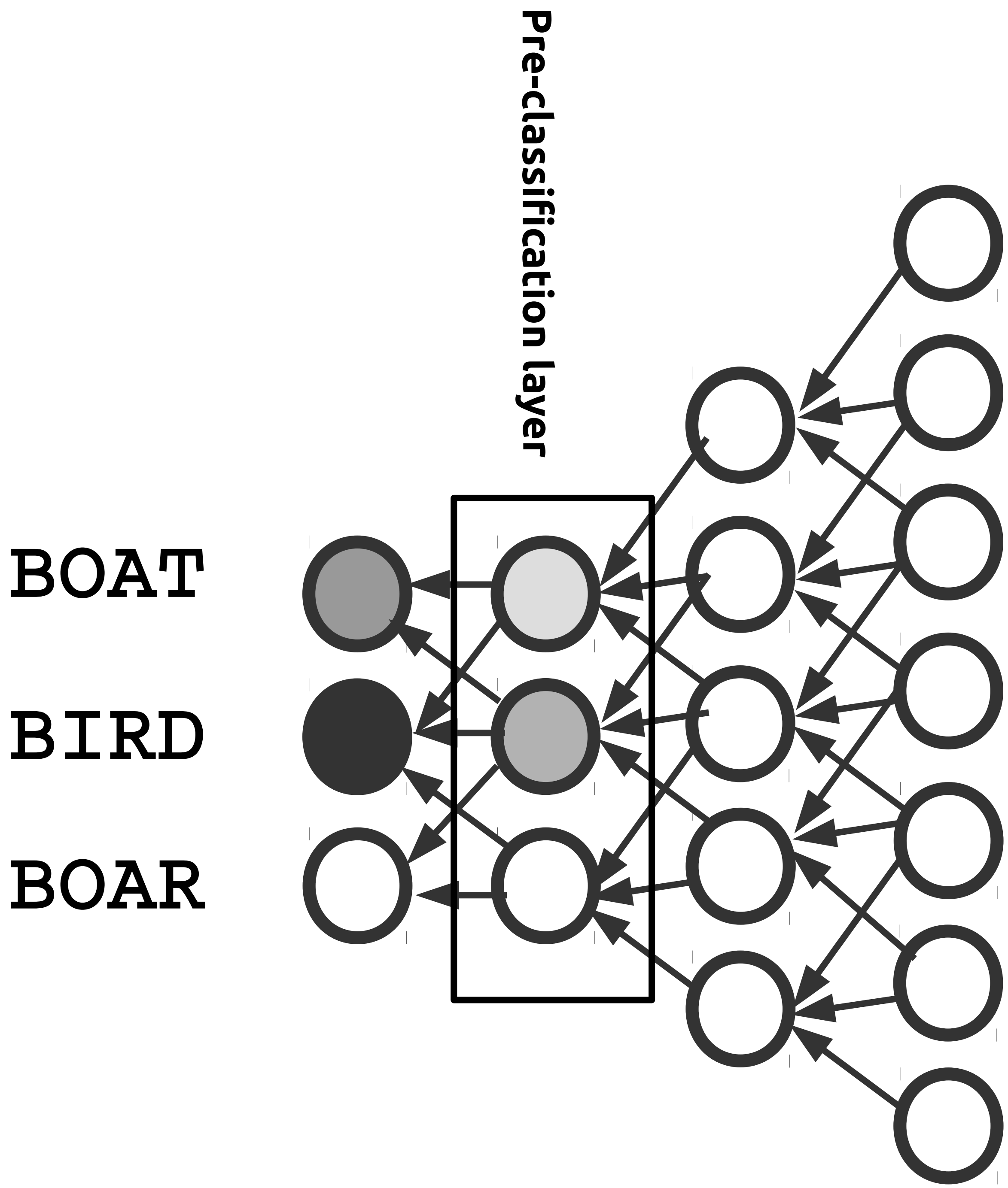


- Approximate human language acquisition
- Text / speech + visual perceptual input

Studies

Setting	Representations
Image + Text	Syntax
Image + Phonemes	Form vs Meaning
Image + Speech	Form vs Meaning
Image + Speech	Phonology

Visual Features via CNN



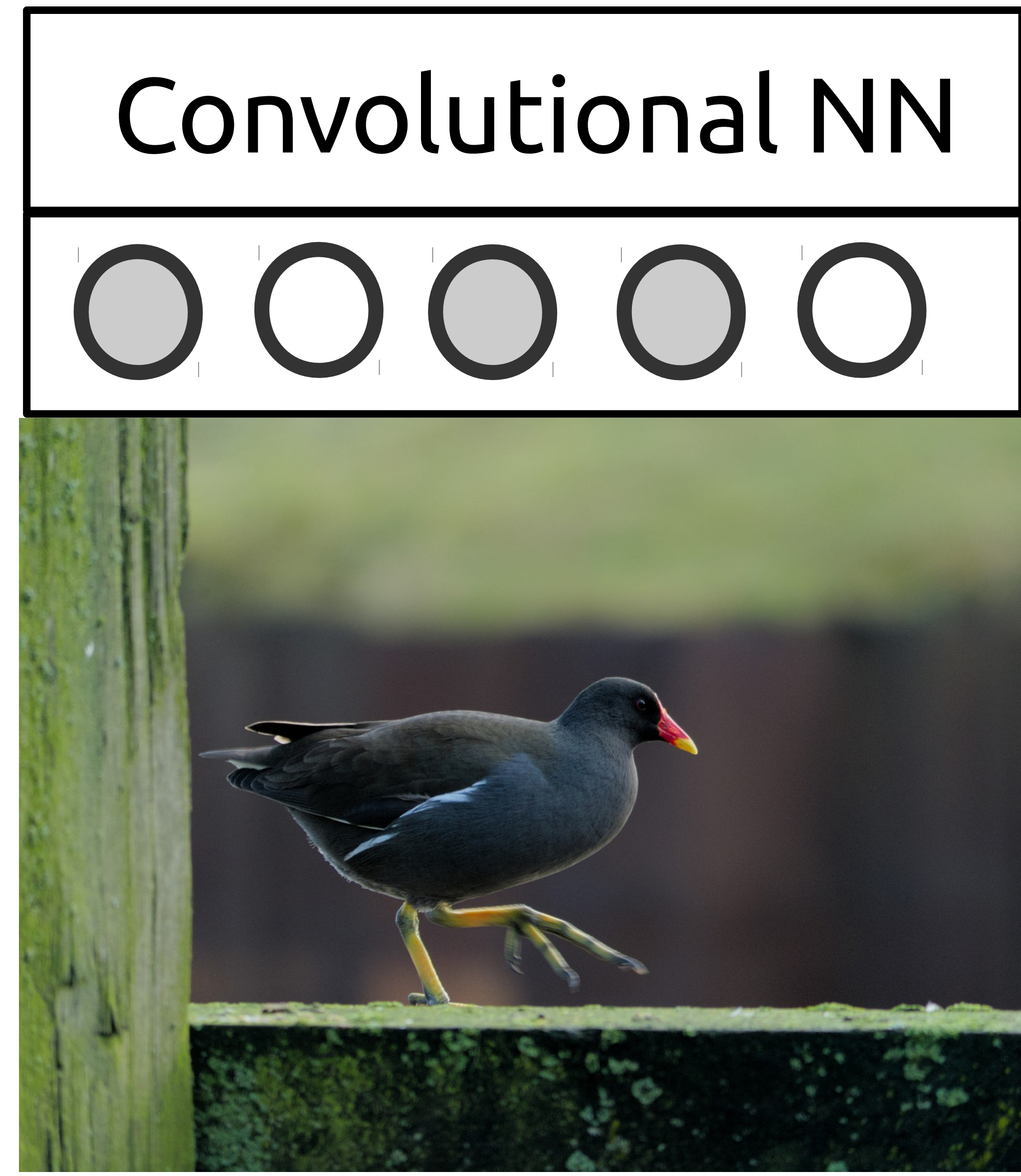
IMAGINET
Multi-task language/image model

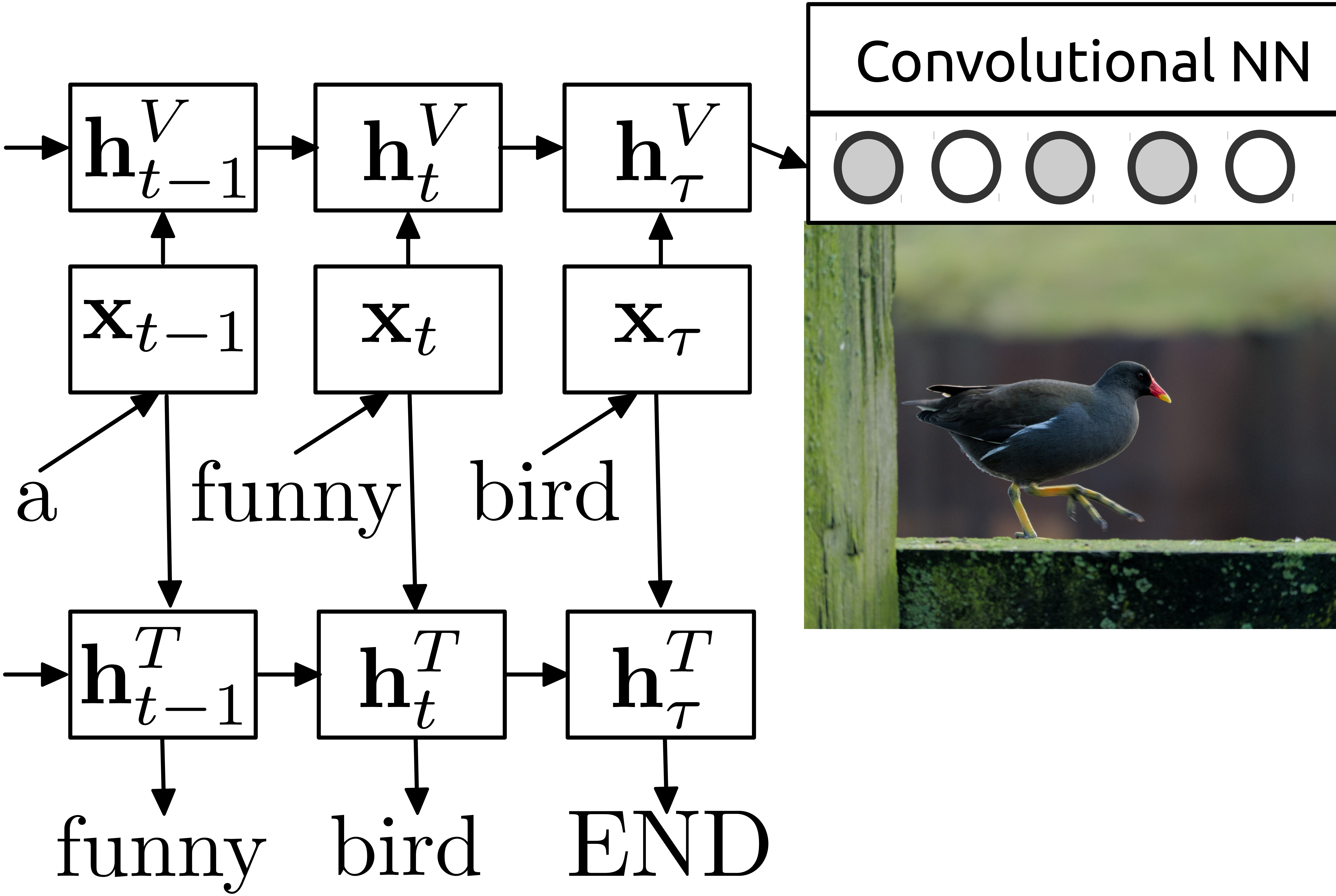
- Integrate distributional (textual) and perceptual (visual) clues
- Representations of phrases and complete sentences

Data

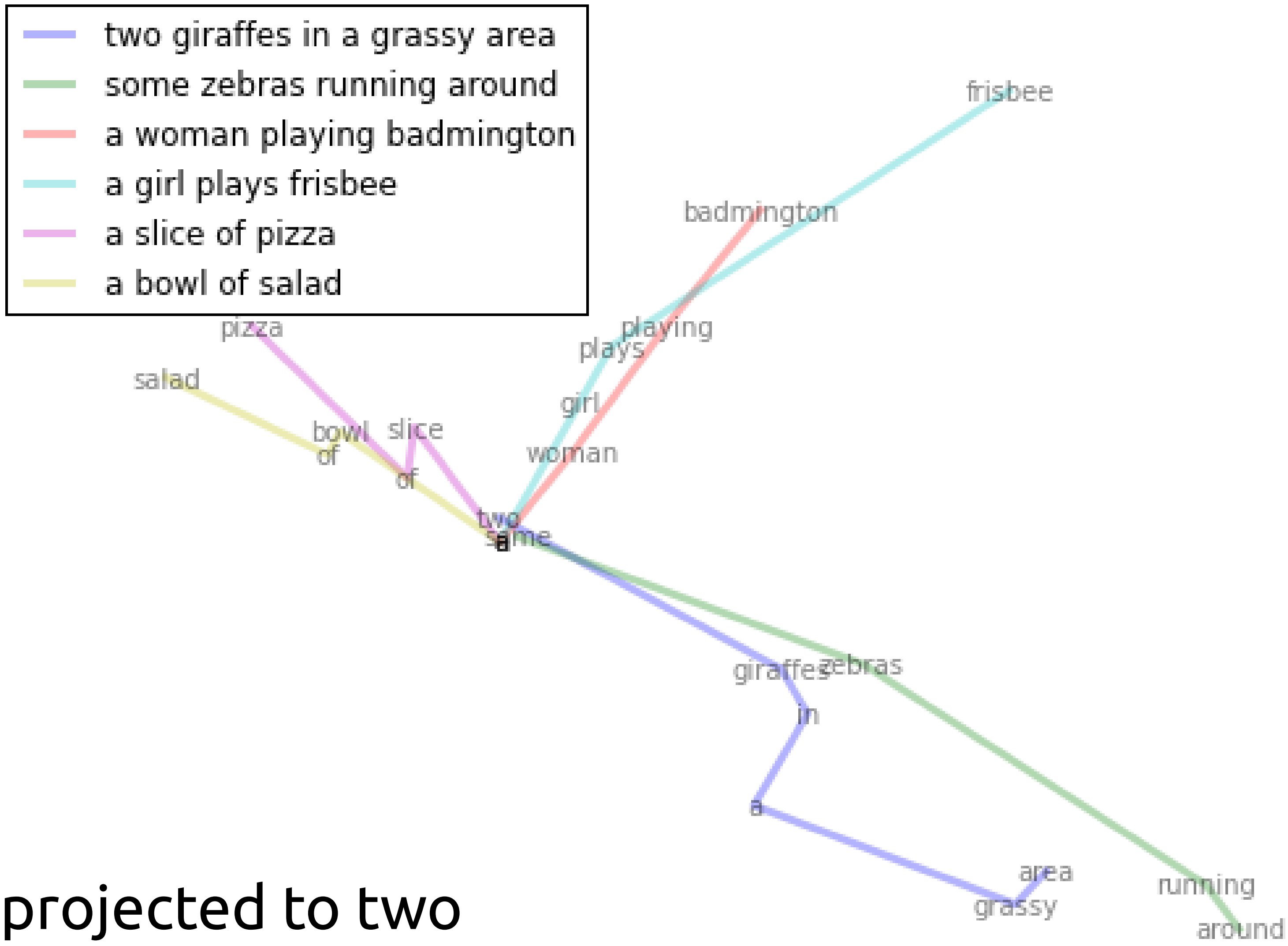
Microsoft COCO
Common Objects in Context

- 300K images, five crowd-sourced captions each a funny bird



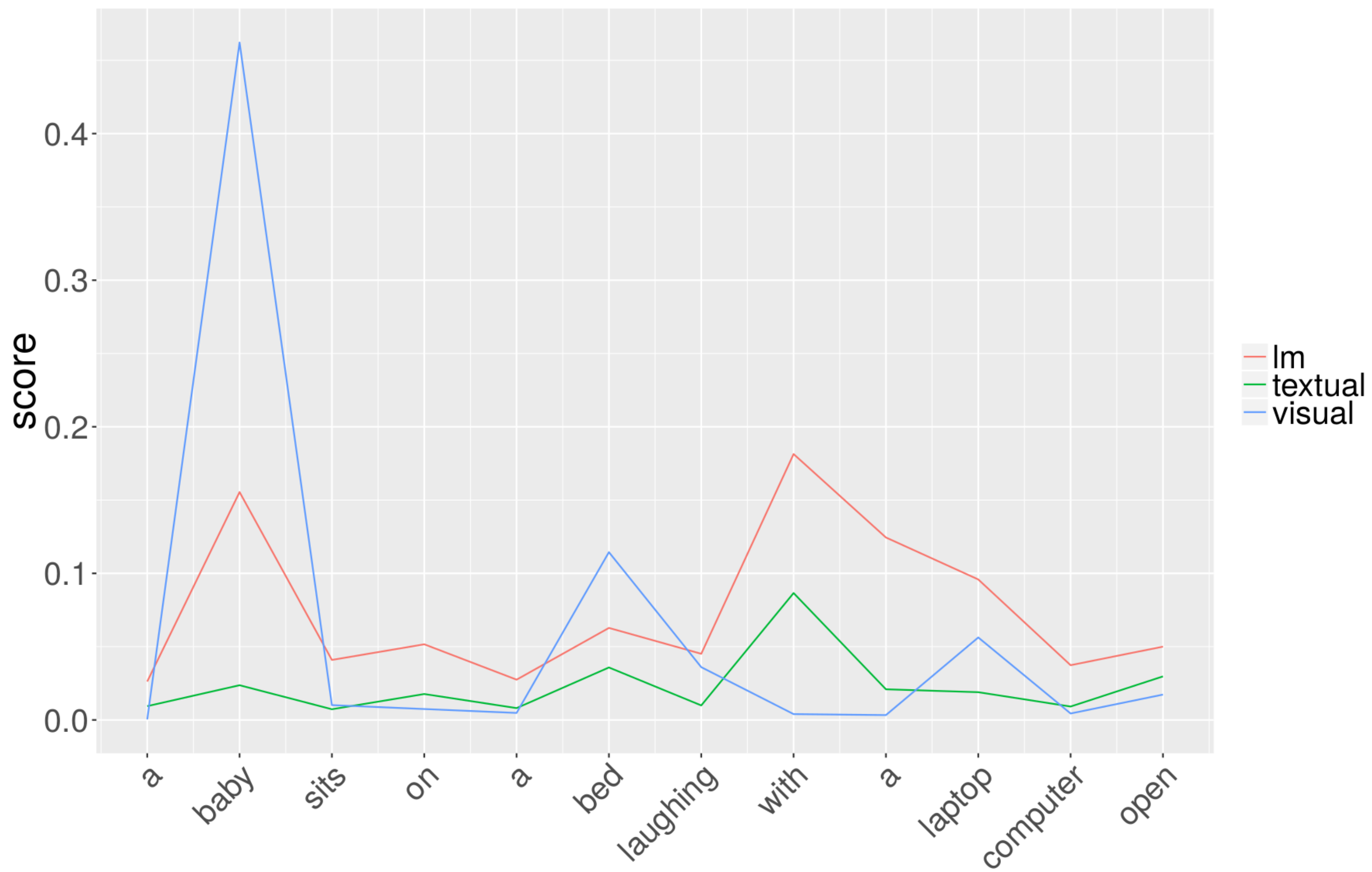

Convolutional NN
$\mathbf{h}^V_{t-1}$
$\mathbf{h}^V_t$
$\mathbf{h}^V_\tau$
$\mathbf{x}_{t-1}$
$\mathbf{x}_t$
$\mathbf{x}_\tau$
a
funny
bird
$\mathbf{h}^T_{t-1}$
$\mathbf{h}^T_t$
$\mathbf{h}^T_\tau$
funny
bird
END

Evolution of network state



Network states projected to two dimensions via PCA

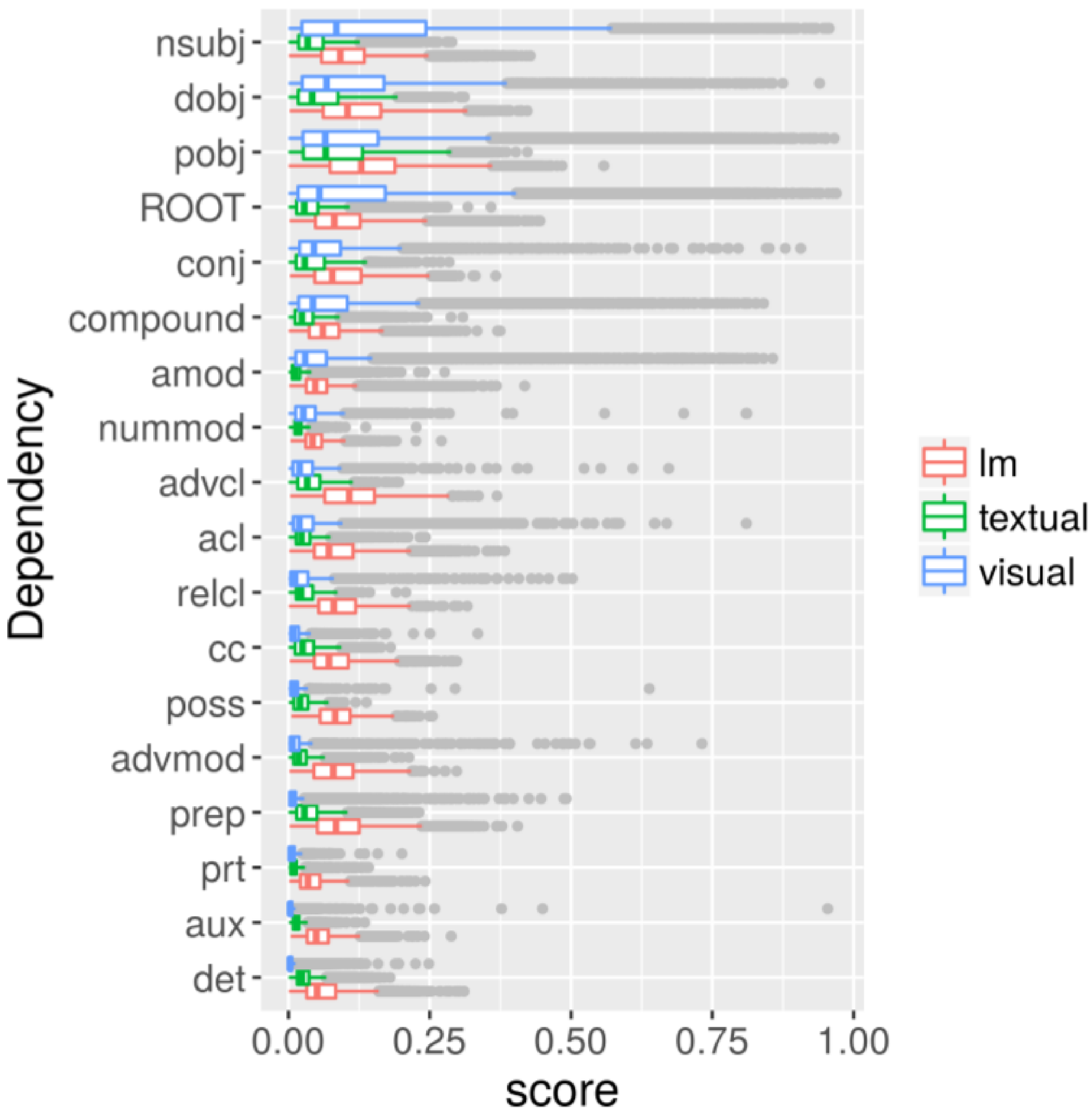
Quantifying importance



original sentence

omit **baby**

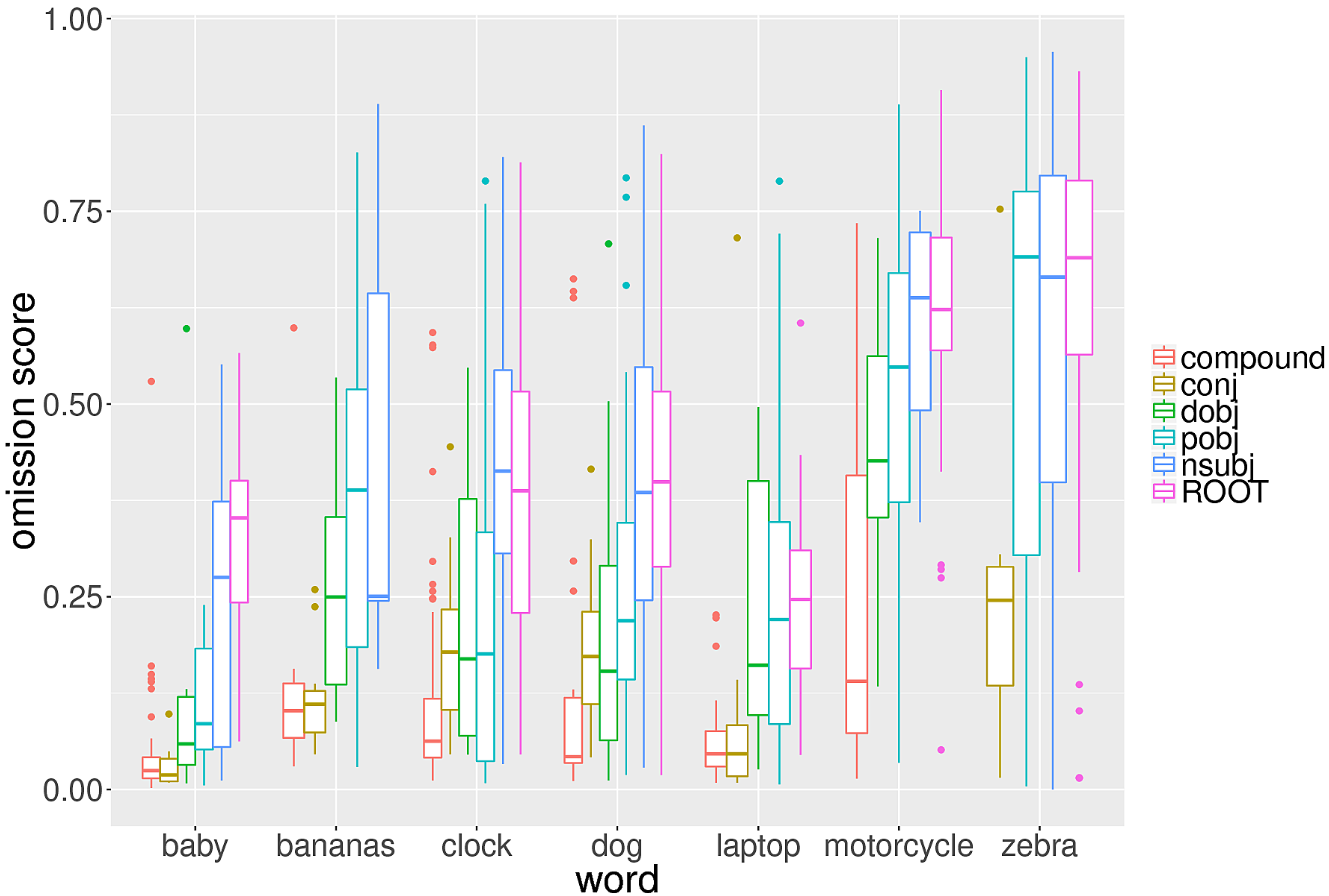
Grammatical functions



LM and Textual pays attention to all kinds of words

Visual pathway mostly focuses on content words like subjects, objects and main verbs

Functions by word form



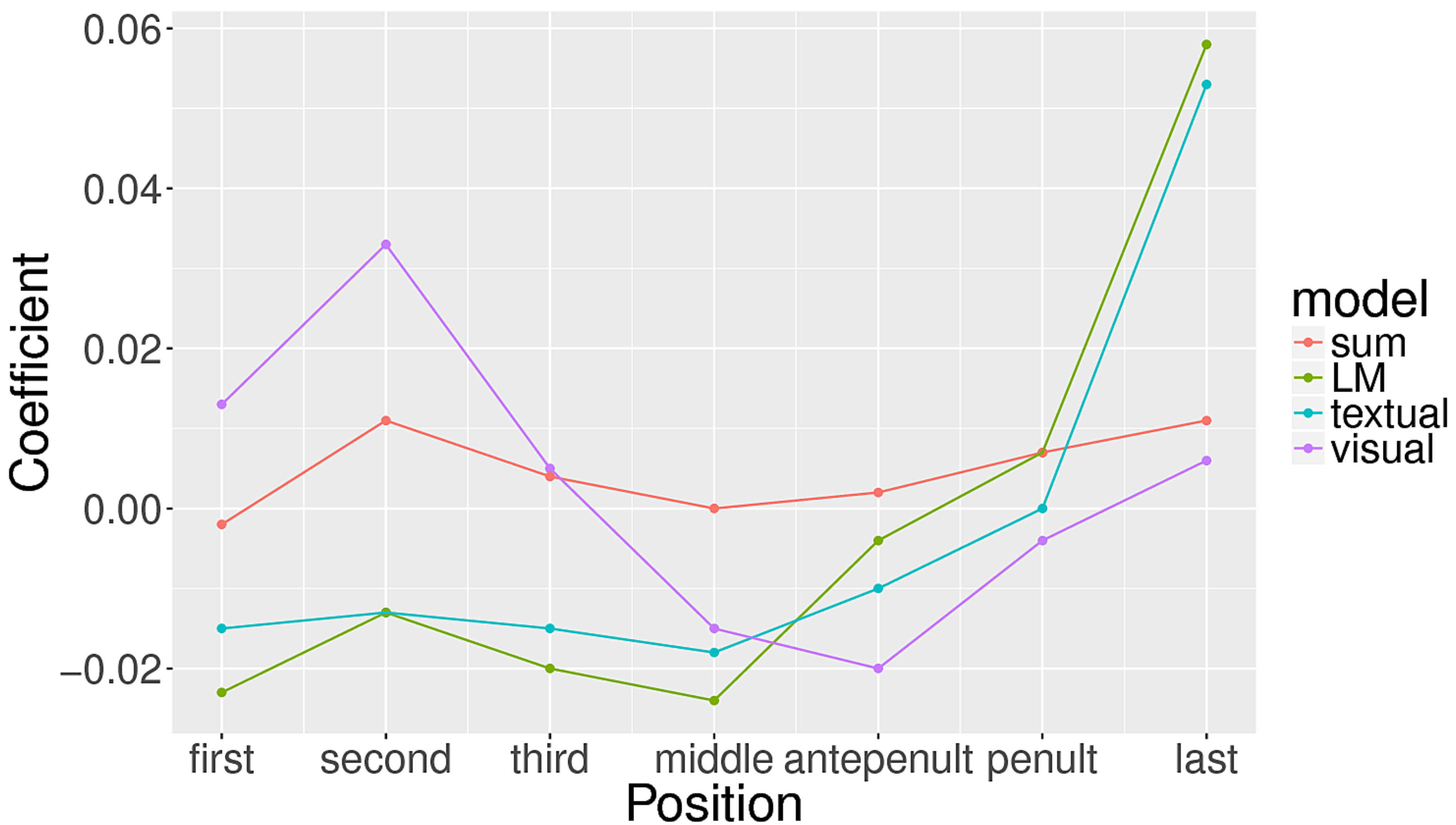
Omission score models

```
score ~ word + dep + pos + word:dep + word:pos
```

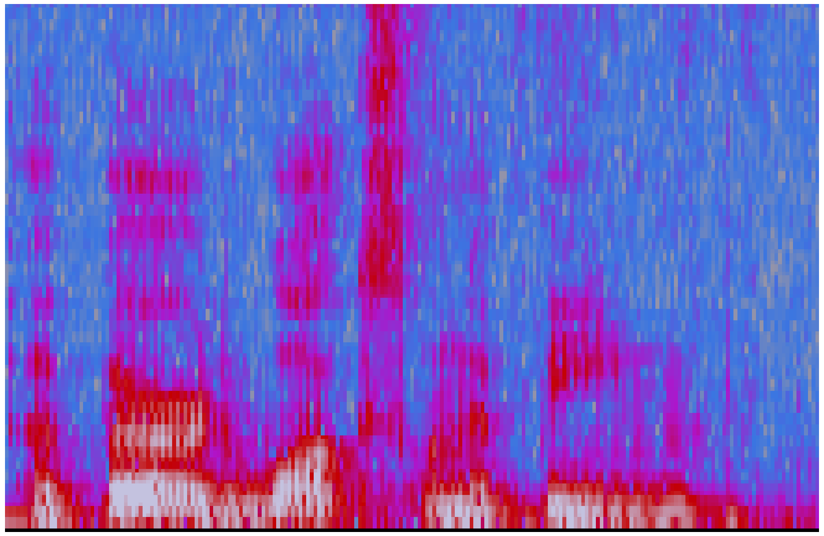
Visual pathway

Predictors	R^2
word	0.490
word+pos	0.506
word+dep	0.515
word+pos+dep	0.523

Information structure



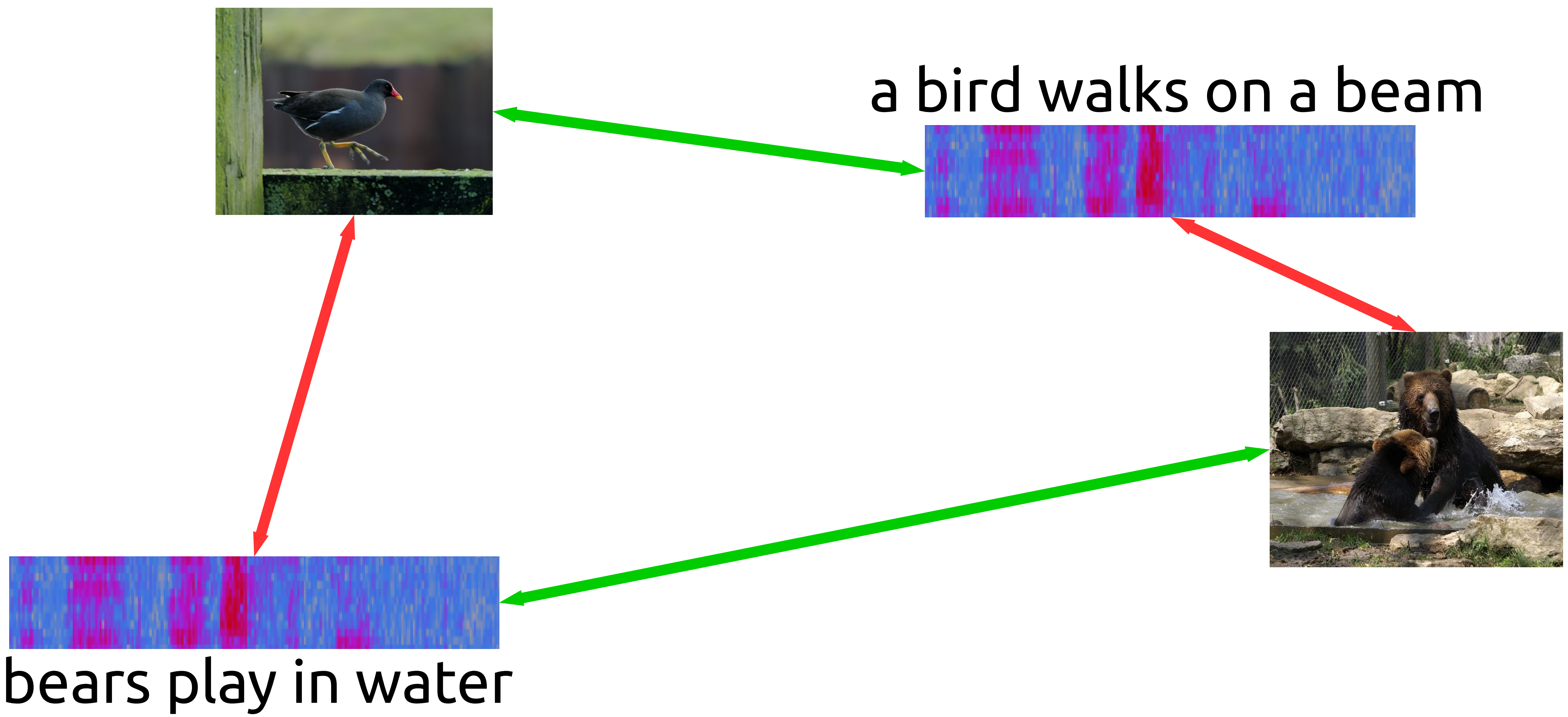
Speech + Image



Data

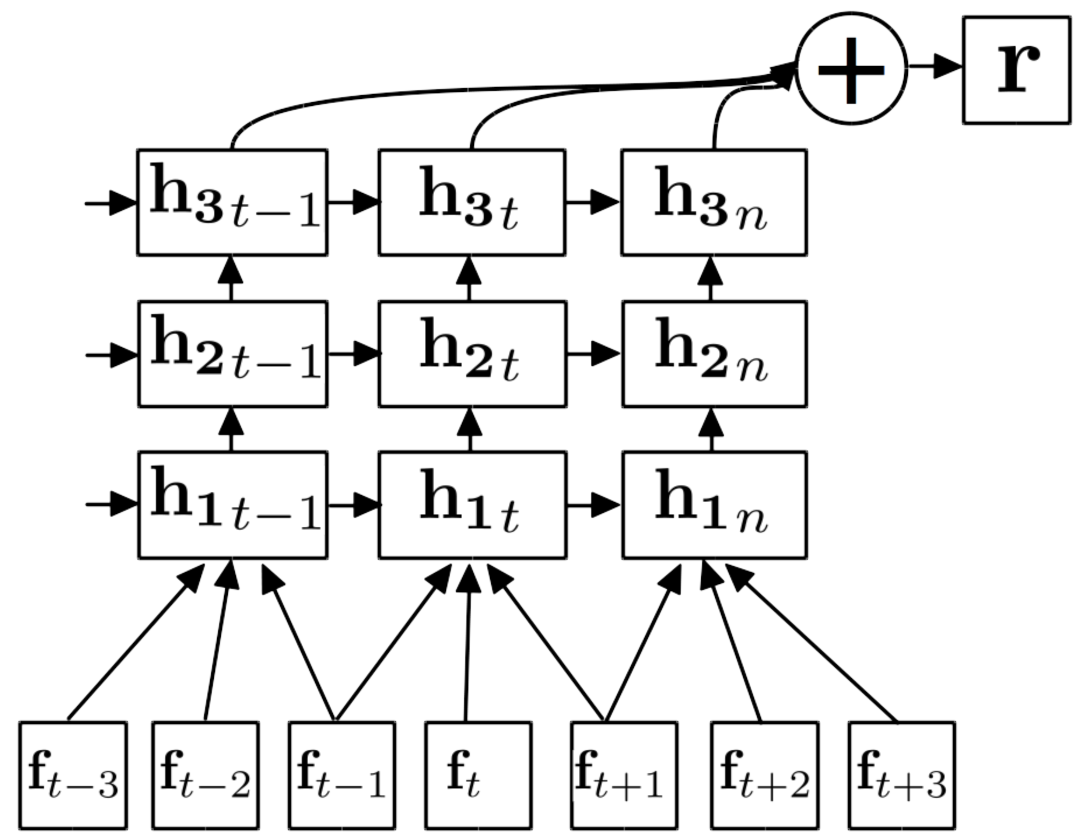
- Flickr8K Audio (Harwath & Glass 2015)
  - 8K images, five audio captions each
- MS COCO Synthetic Spoken Captions
  - 300K images, five synthetically spoken captions each

Project speech and image to joint space



Speech model

- Input: **MFCC**
- Subsampling CNN
- Recurrent Highway Network (Zilly et al 2016)
- Attention



Model settings

Flickr8K Speech
Attention 128
RHN depth 2, 1024
RHN depth 2, 1024
RHN depth 2, 1024
RHN depth 2, 1024
Conv 6x64, stride 2

COCO Speech
Attention 512
RHN depth 2, 512
RHN depth 2, 512
RHN depth 2, 512
RHN depth 2, 512
RHN depth 2, 512
Conv 6x64, stride 3

Flickr8K Text
RHN depth 1, 1024
Embedding 300

COCO Text
RHN depth 1, 1024
Embedding 300

Image retrieval

Flickr8K

Model	R@10	$\tilde{r}$
Speech $\mathrm{RHN}_{4,2}$	0.253	48
Harwath & Glass 2015	0.179	-
Text $\mathrm{RHN}_{1,1}$	0.494	11

MSCOCO

Model	R@10	$\tilde{r}$
Speech $\mathrm{RHN}_{5,2}$	0.444	13
Text $\mathrm{RHN}_{1,1}$	0.565	8

Newer CNN architecture: Harwath et al 2016 (NIPS), Harwath and Glass 2017 (ACL)

Levels of representation

- What aspects of sentences are encoded?
- Which layers encode form, which encode meaning?

Representational similarity

Utt 1	Utt 2	Sim 1	Sim 2
A slice of pizza	A bowl of salad	7.0	6.2
Two dogs run	A kitty running	8.0	9.0
A yellow and white bird	A kitty running	3.0	4.5

Correlation between similarity 1 and similarity 2

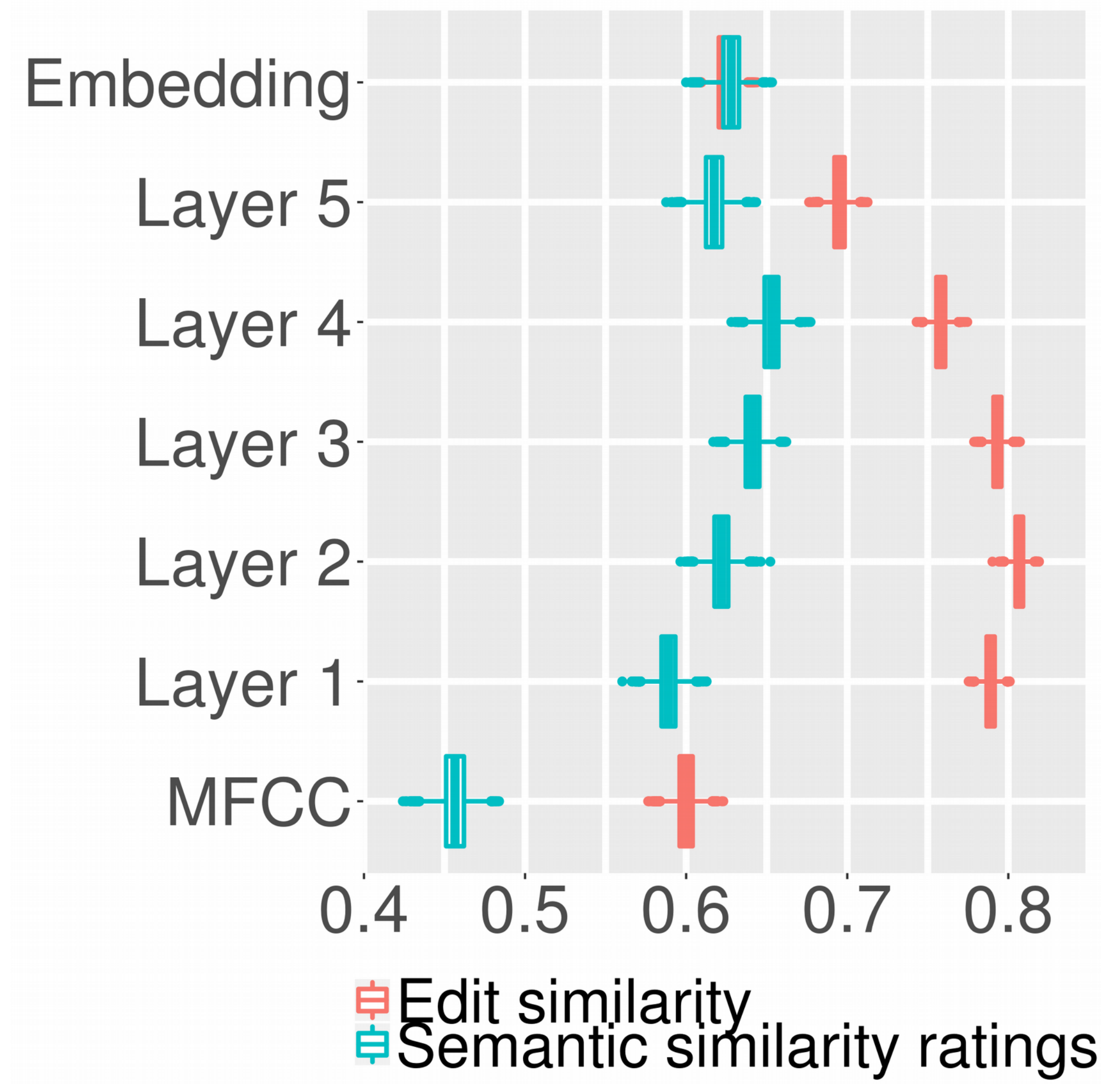
Representational Similarity

- Correlations between sets of pairwise similarities according to
  - Activations

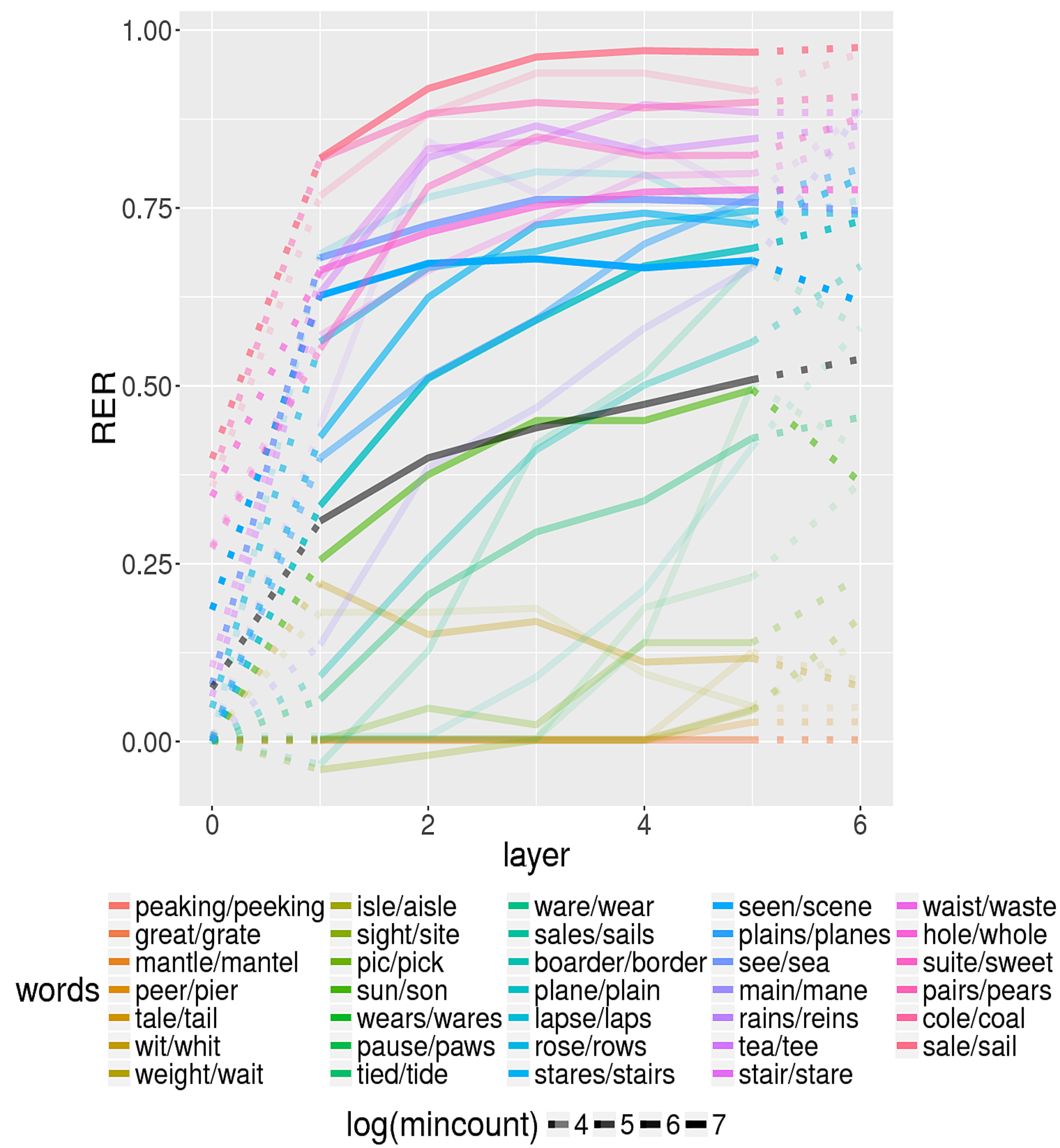
  vs

  - Edit ops on text
  - Human judgments

(SICK dataset)



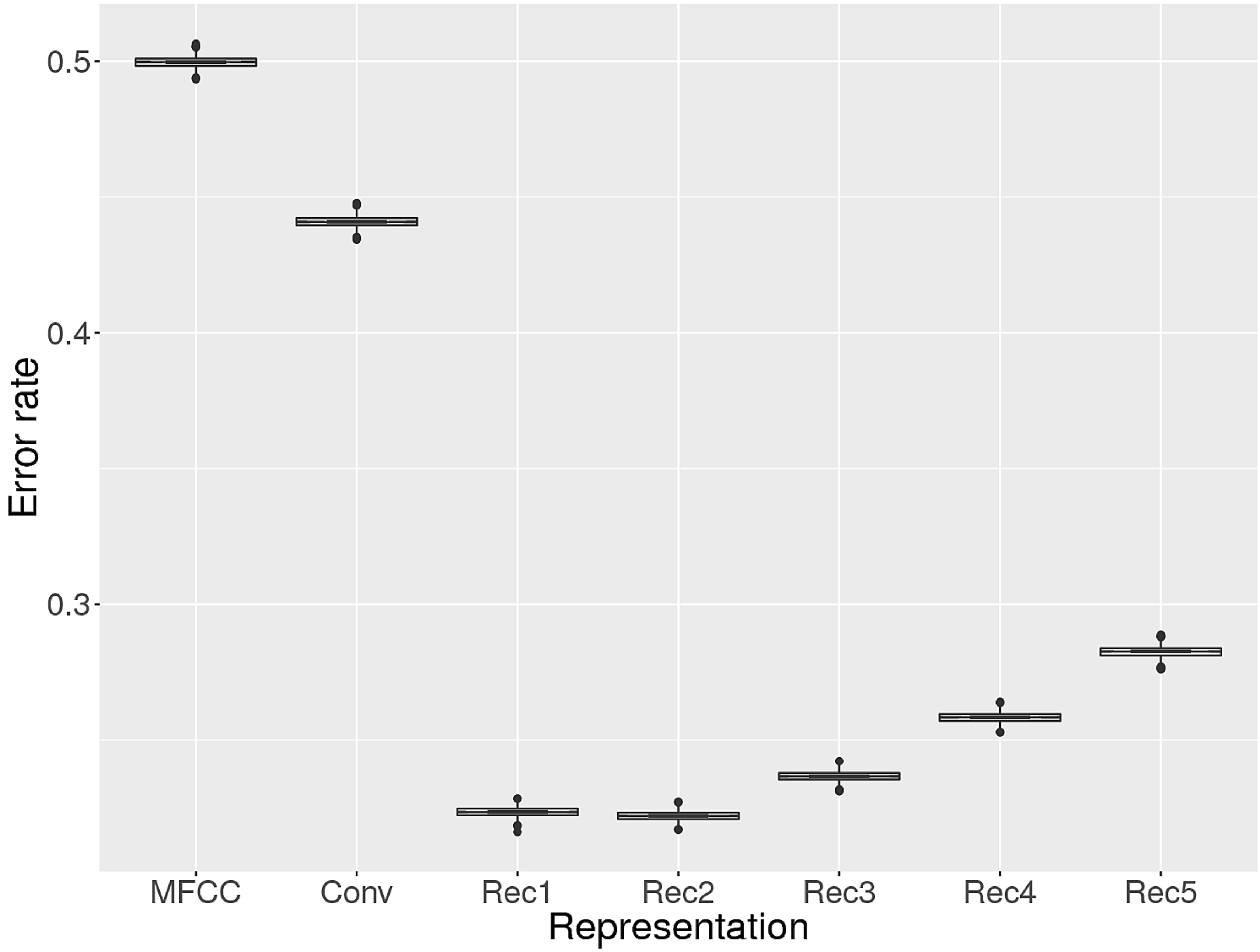
Homonym disambiguation



Phonological form

Phoneme decoding

- Classify representations of speech segments
- L2-penalized Logistic regression



Phoneme discrimination

ABX task (Schatz et al. 2013)

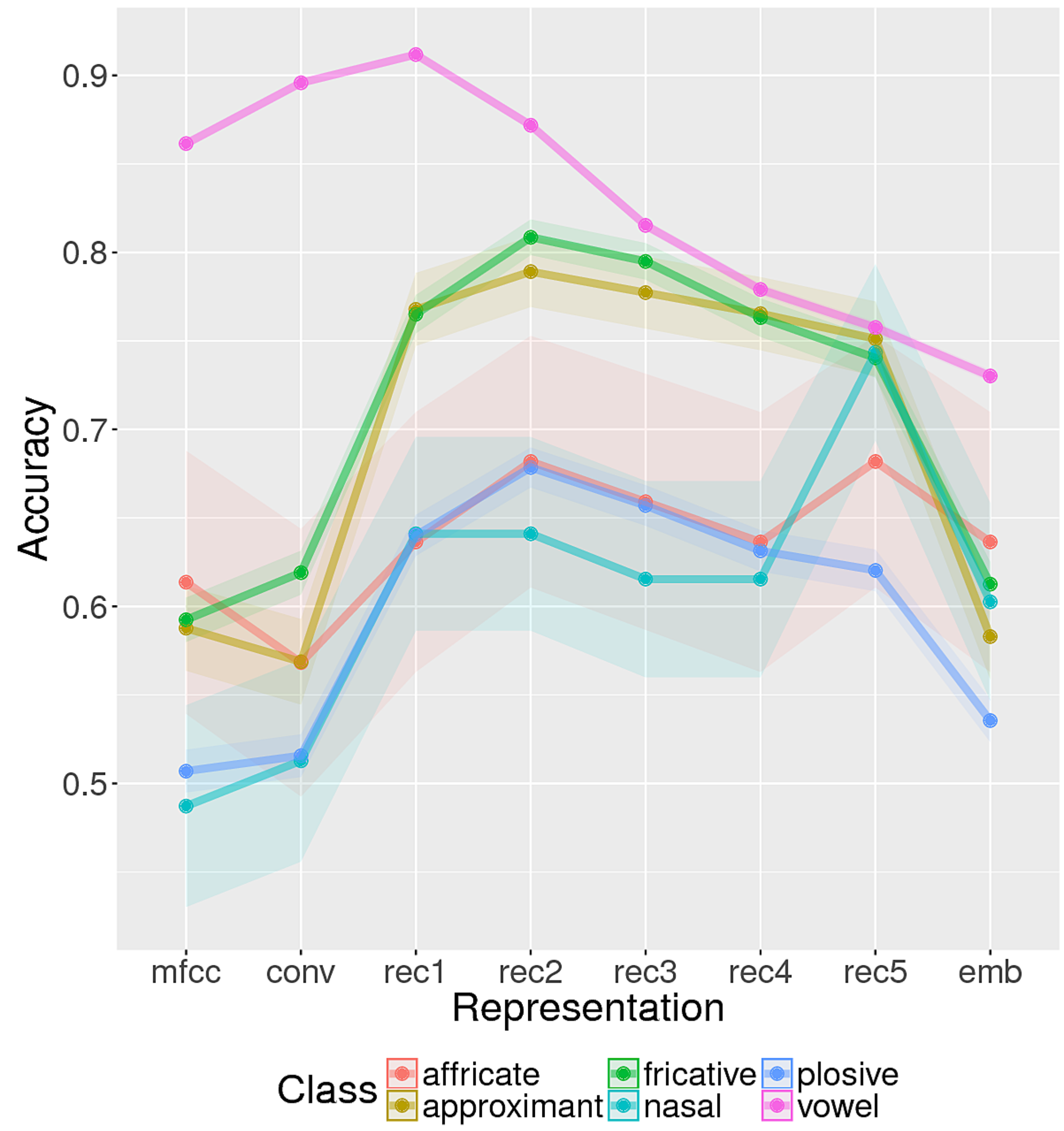
A: /bi/

B: /mi/

X: /maɪ/

ABX

Especially challenging when the target (B) and distractor (A) belong to same phoneme class.

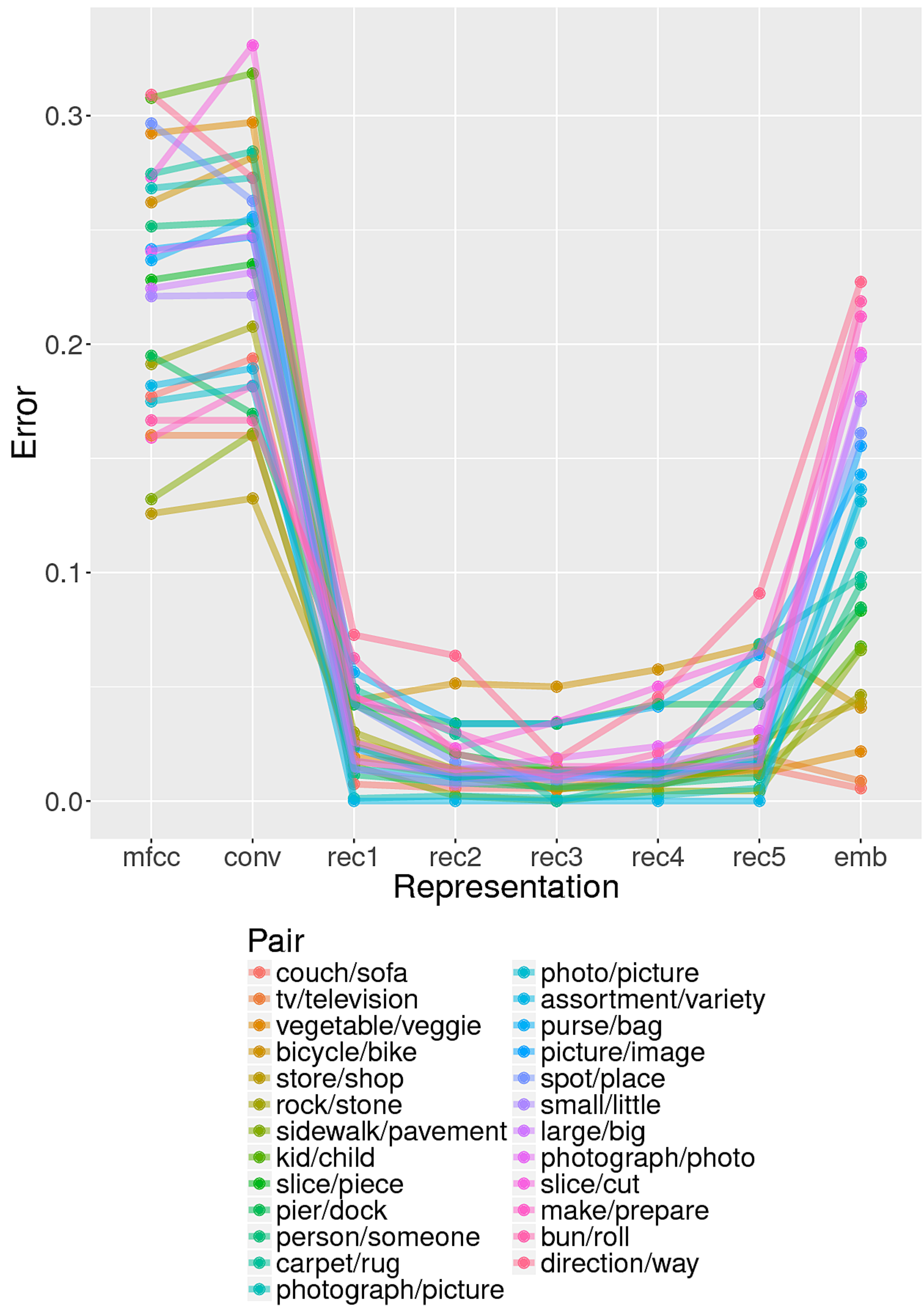


Synonym discrimination

- Disentangle phonological form and semantics.
- Discriminate between synonyms in identical context:

  A girl looking at a photo.

  A girl looking at a picture.
- **How invariant to phonological form is a representation?**


0.3
0.2
0.1
0.0
Error
mfcc
conv
rec1
rec2
rec3
rec4
rec5
emb
Representation
Pair
couch/sofa
tv/television
vegetable/veggie
bicycle/bike
store/shop
rock/stone
sidewalk/pavement
kid/child
slice/piece
pier/dock
person/someone
carpet/rug
photograph/picture
photo/picture
assortment/variety
purse/bag
picture/image
spot/place
small/little
large/big
photograph/photo
slice/cut
make/prepare
bun/roll
direction/way

Conclusion

- Visually grounded RNNs implicitly learn approximations of (some) linguistic concepts
  - Grammatical functions
  - Phonemes
- Bottom layers encode form, top layers meaning
- Even top layers are far from form-invariant

Some open questions

- RNNs' biases are weak and not motivated by structure of language
- Inject stronger, more specific bias?
  - Hard-wire them?
  - Learn them from massive data?
- Triangulate using cross-language setting?

References

- Grzegorz Chrupała, Ákos Kádár, Afra Alishahi. 2015. Learning language through pictures. In ACL.
- Lieke Gelderloos and Grzegorz Chrupała. 2016. From phonemes to images: levels of representation in a recurrent neural model of visually-grounded language learning. In Coling.
- Ákos Kádár, Grzegorz Chrupała and Afra Alishahi. 2017. Representation of linguistic form and function in recurrent neural networks. Computational Linguistics (in press).
- Grzegorz Chrupała, Lieke Gelderloos and Afra Alishahi. 2017. Representations of language in a model of visually grounded speech signal. In ACL.
- Afra Alishahi, Marie Barking and Grzegorz Chrupała. 2017. Encoding of phonology in a recurrent neural model of grounded speech. In CoNLL.

Code/data

- github.com/gchrupala/visually-grounded-speech
- github.com/gchrupala/encoding-of-phonology
- zenodo.org/record/400926

Extras

Dependency and position

- Omission ~

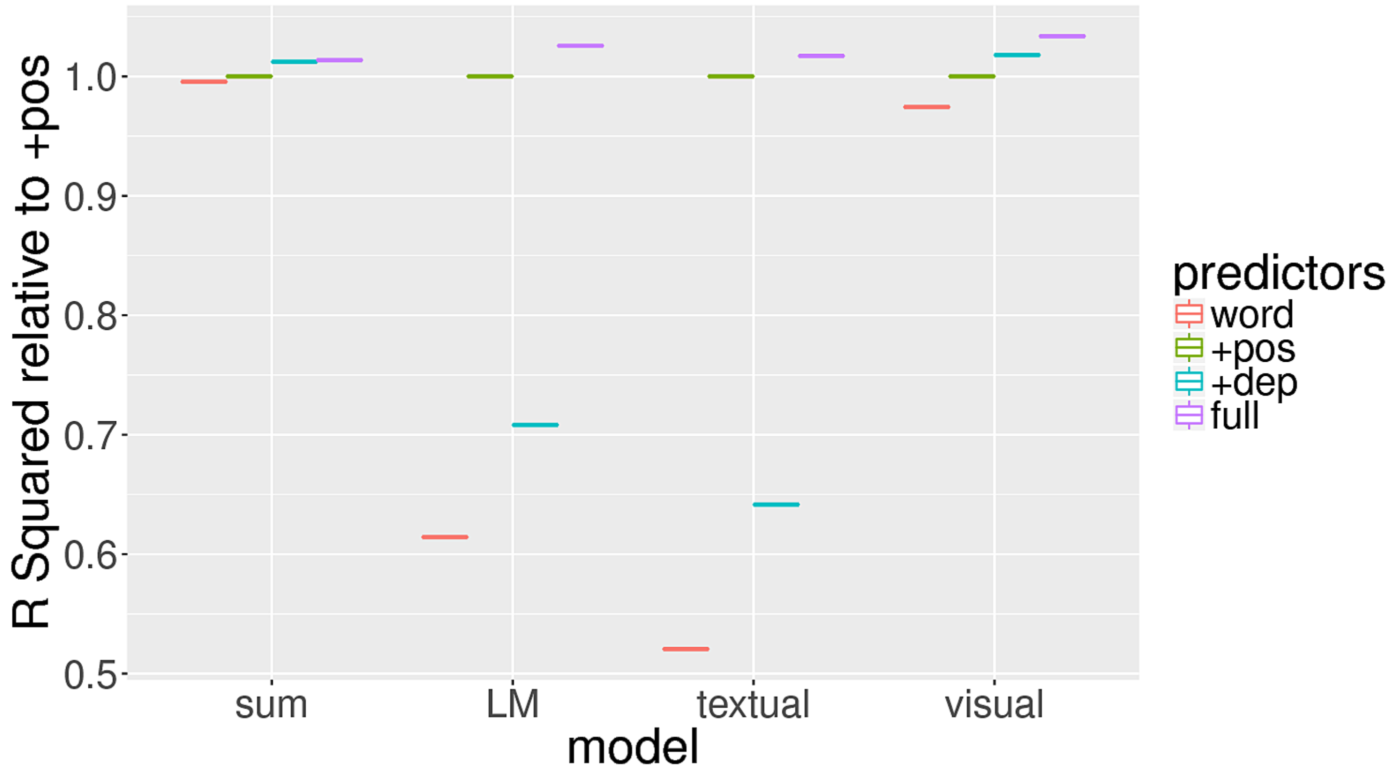
  Word +

  Pos +

  Dep +

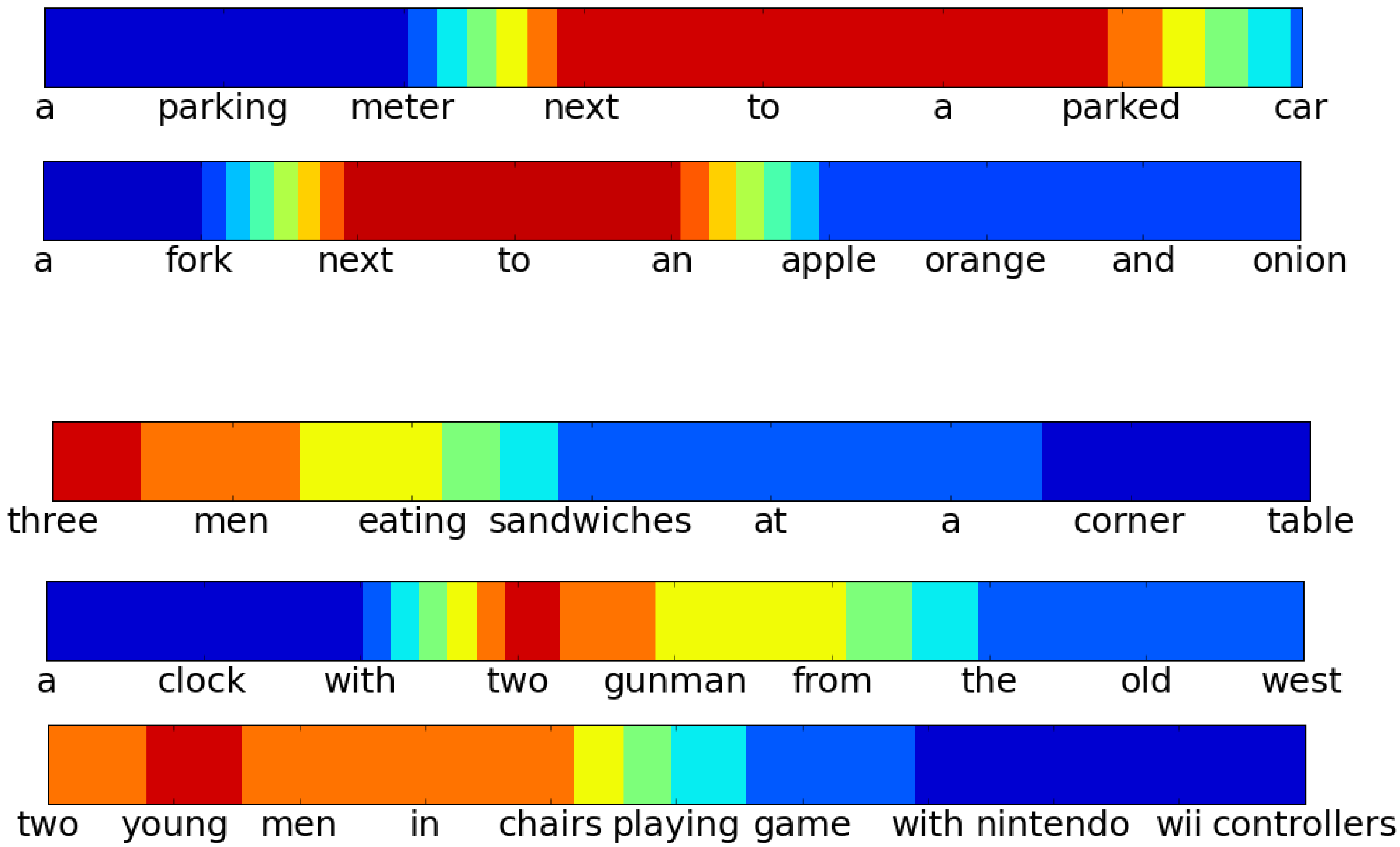
  Word:Pos +

  Word:Dep



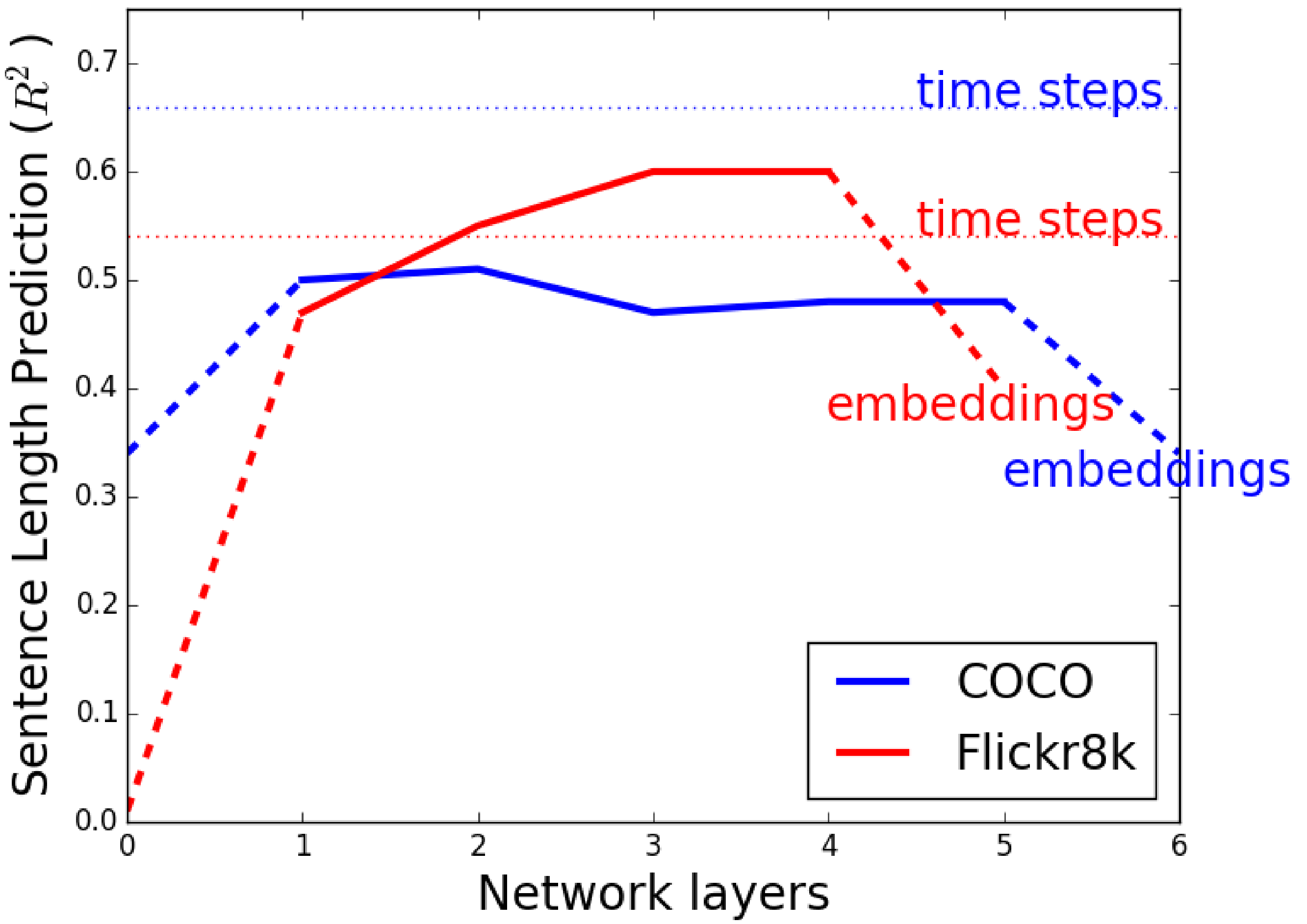
	word	+pos	+dep	full
SUM	0.654	0.661	0.670	0.670
LM	0.358	0.586	0.415	0.601
TEXTUAL	0.364	0.703	0.451	0.715
VISUAL	0.490	0.506	0.515	0.523

Specificity of neurons



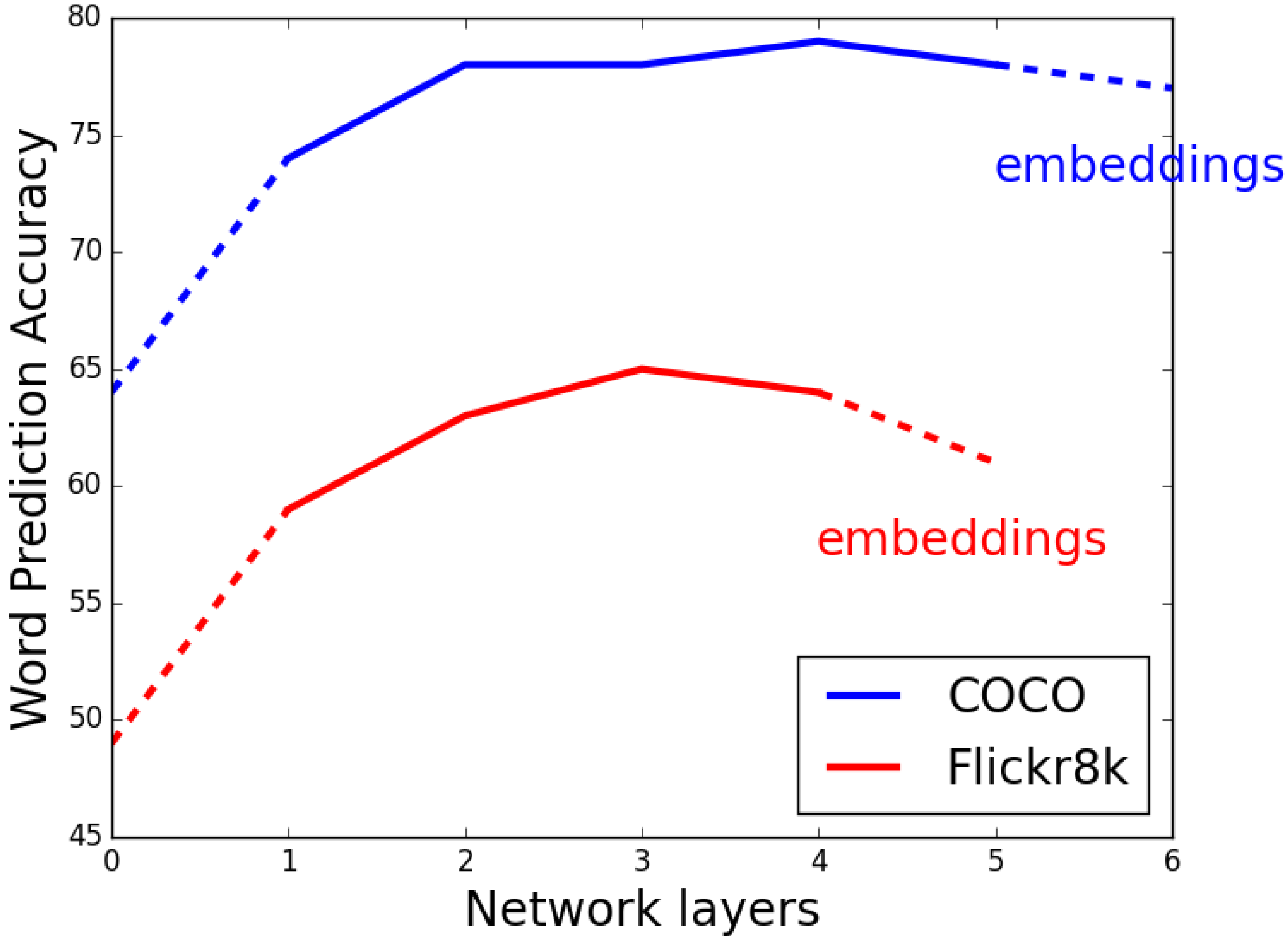
Number of words

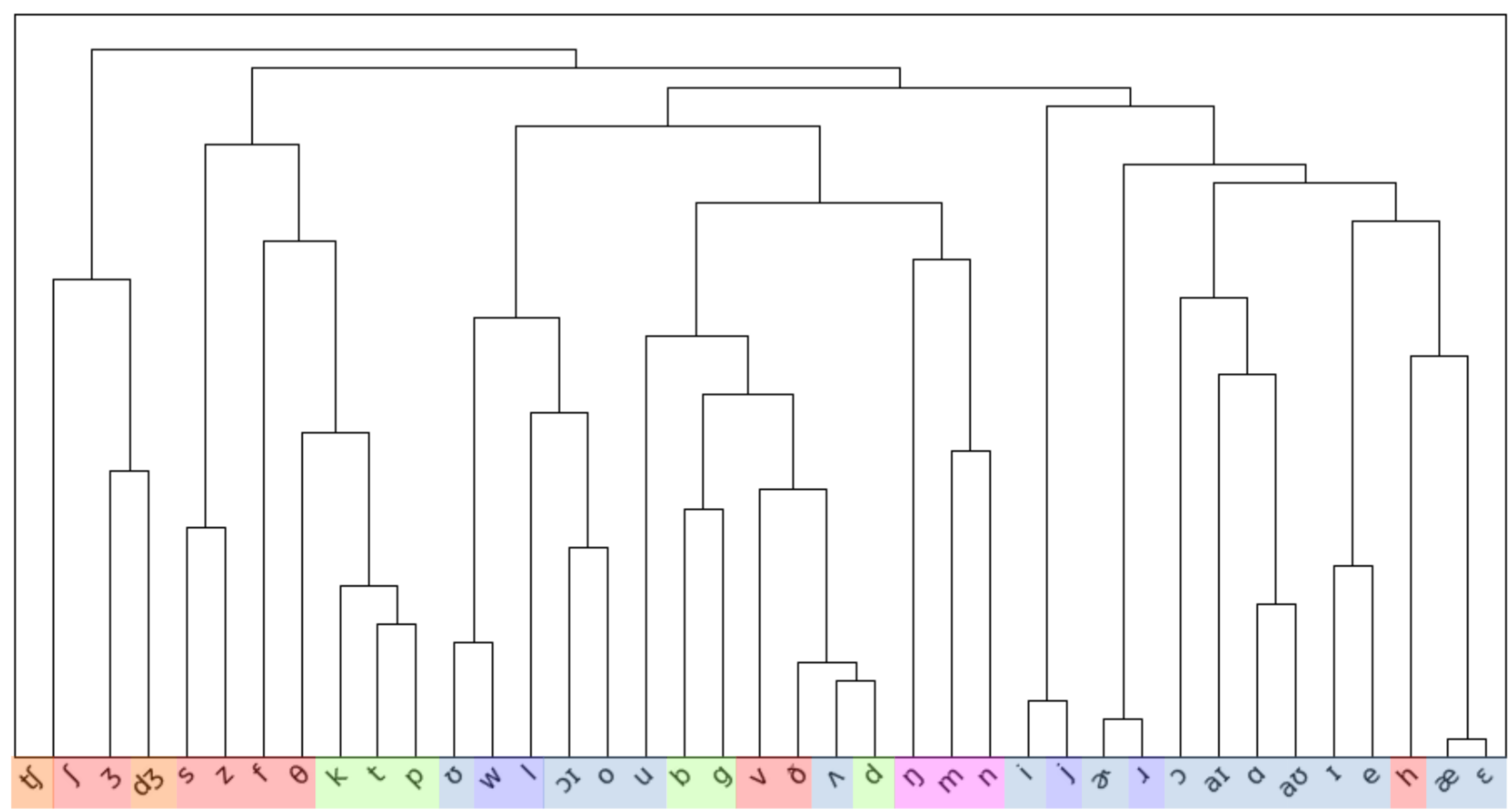
- Input
  - Activations for utterance
- Model
  - Linear regression



Word presence

- Input
  - Activations for utterance
  - MFCC for word
- Model
  - MLP




tʃ ʃ ʒ dʒ s z f θ k t p ʊ w l ɔɪ o u b g v ð ʌ d ŋ m n i j ɚ ɹ ɔ aɪ ɑ aʊ ɪ e h æ ɛ