Why the pond is not outside the frog? Grounding in contextual representations by neural language models

Mehdi Ghanimifard

Assessor: Desmond Elliott, University of Copenhagen

University of Gothenburg

February 21 2020

Locative Expressions

Bring me *the big red book on the table.*

Photo: Adam C (CC-BY-2.0)

- Framework: <Target, Relation, Landmark>

- Also known as referent / relatum *(Miller and Johnson-Laird, 1976)*; figure / ground *(Talmy 1983)*; located object / reference object *(Herskovits 1986, Gapp 1994, Dobnik 2009)*

Geometric Expressions of Meaning

<frog, *next to*, pond>

- The **frog** *next to* the **pond**.
- The **frog** is *next to* the **pond**.
- There is a **frog** *next to* the **pond**.
- The **frog** *next to* the **pond** is watching us.

Figure: Ghanimifard 2020

Usage in Context

- Core issue in the usage of the functional / geometric form:

> "*functional sense* of relationships refers to the object-specific relationship between entities that is *not dependent* on the location or spatial configuration"
>
> Ghanimifard, 2020

RQ1: What spatial knowledge is learned in generative neural language models?

Study 3

Exploring the Functional and Geometric Bias of Spatial Relations Using Neural Language Models

Simon Dobnik* **Mehdi Ghanimifard*** **John D. Kelleher†**

- Hypothesis: it is possible to distinguish between functionally biased and geometrically biased spatial relations by examining the diversity of the contexts in which they occur.
- Estimate using a neural language model (*Hochreiter and Schmidhuber, 1997*) trained at the word level:

$$P(w_{1:T}) = \sum_{t=1}^{T} P(w_{t+1}|w_{1:t})$$

- Train the model with the Visual Genome Dataset (*Krishna et al. 2017*) of <target, relation, landmark> sequences

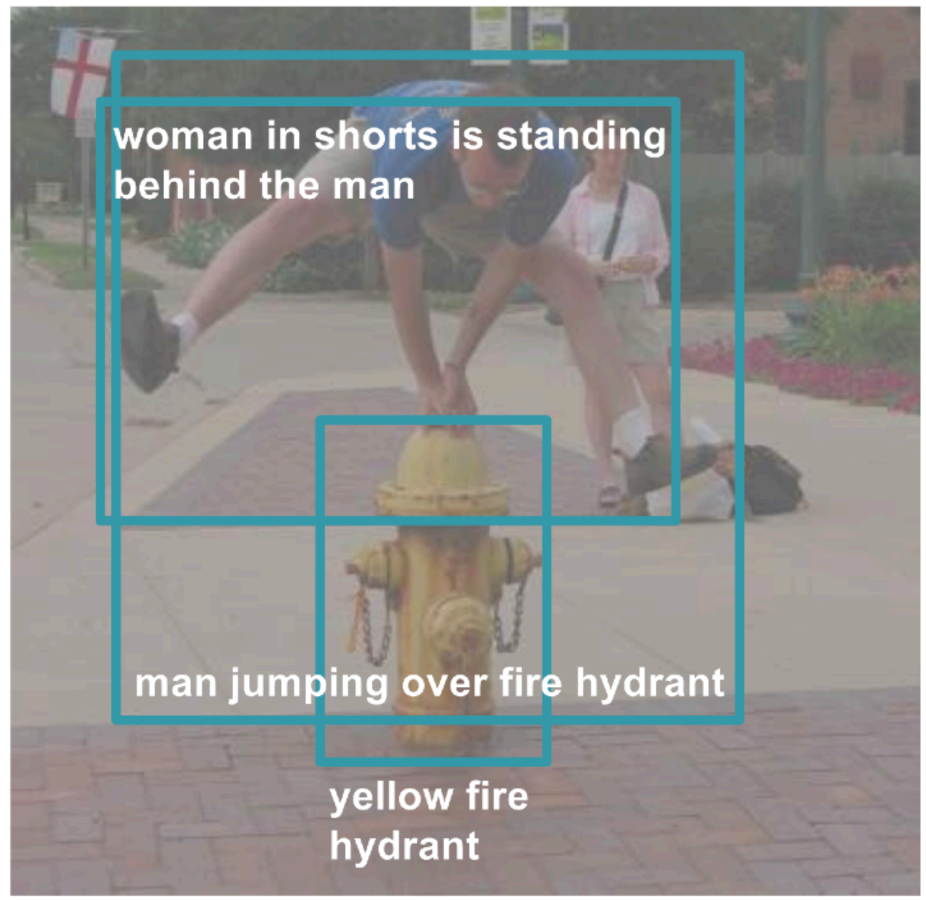


Figure: Krishna et al. 2017

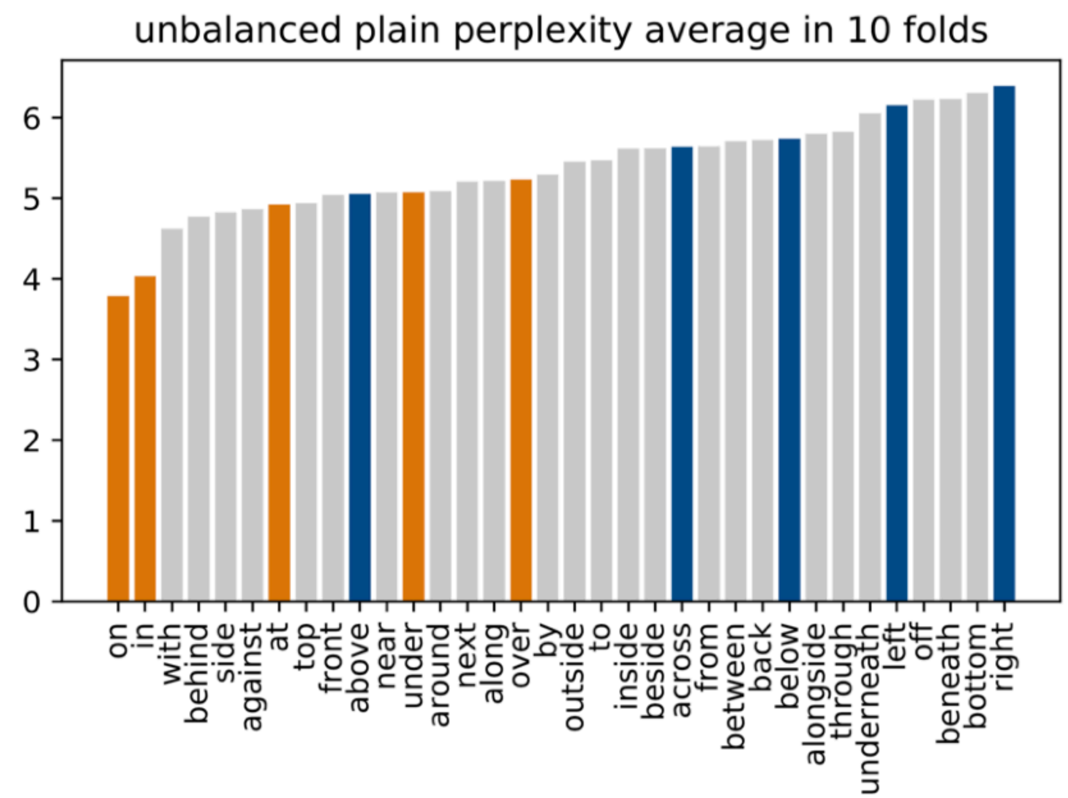
- Measure the perplexity of held-out sequences

$$Perplexity(S,P) = 2^{E_S[-log_2(P(w_{1:T}))]}$$

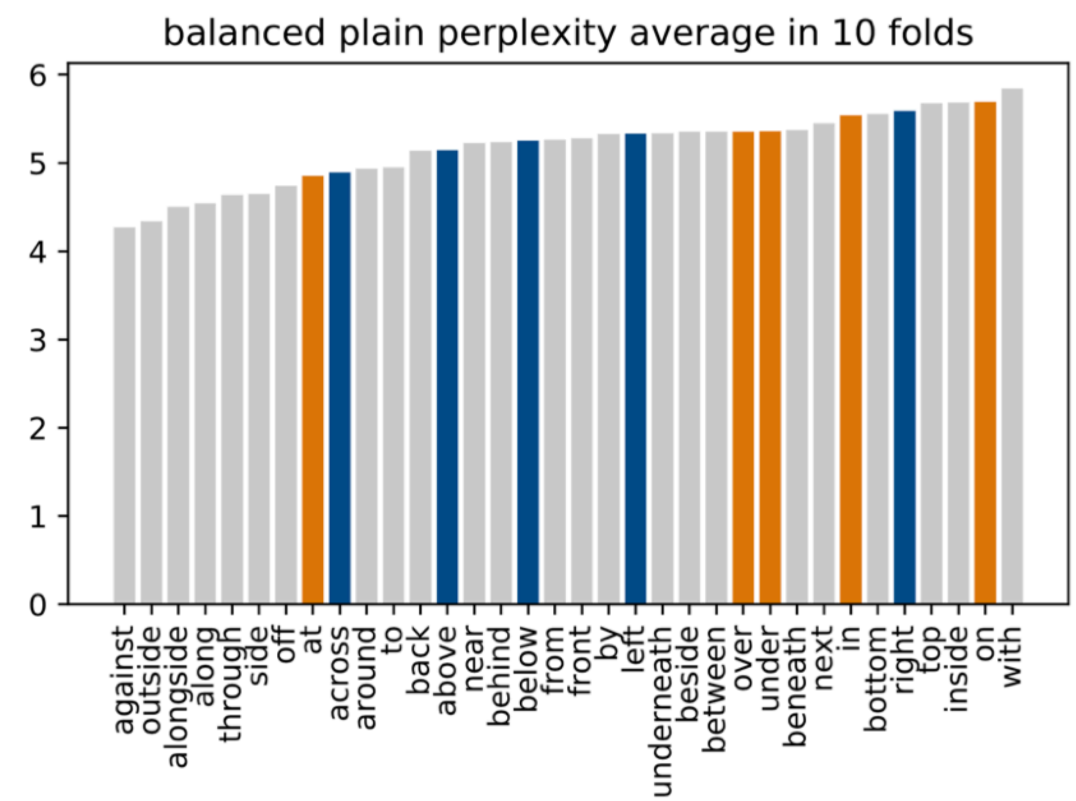
- The perplexity of functionally-biased relations is substantially affected by balancing the relations by downsampling the dataset.

Functionally biased

Geometrically biased



(a) test-set



(a) test-set

Study 4

What a neural language model tells us about spatial relations

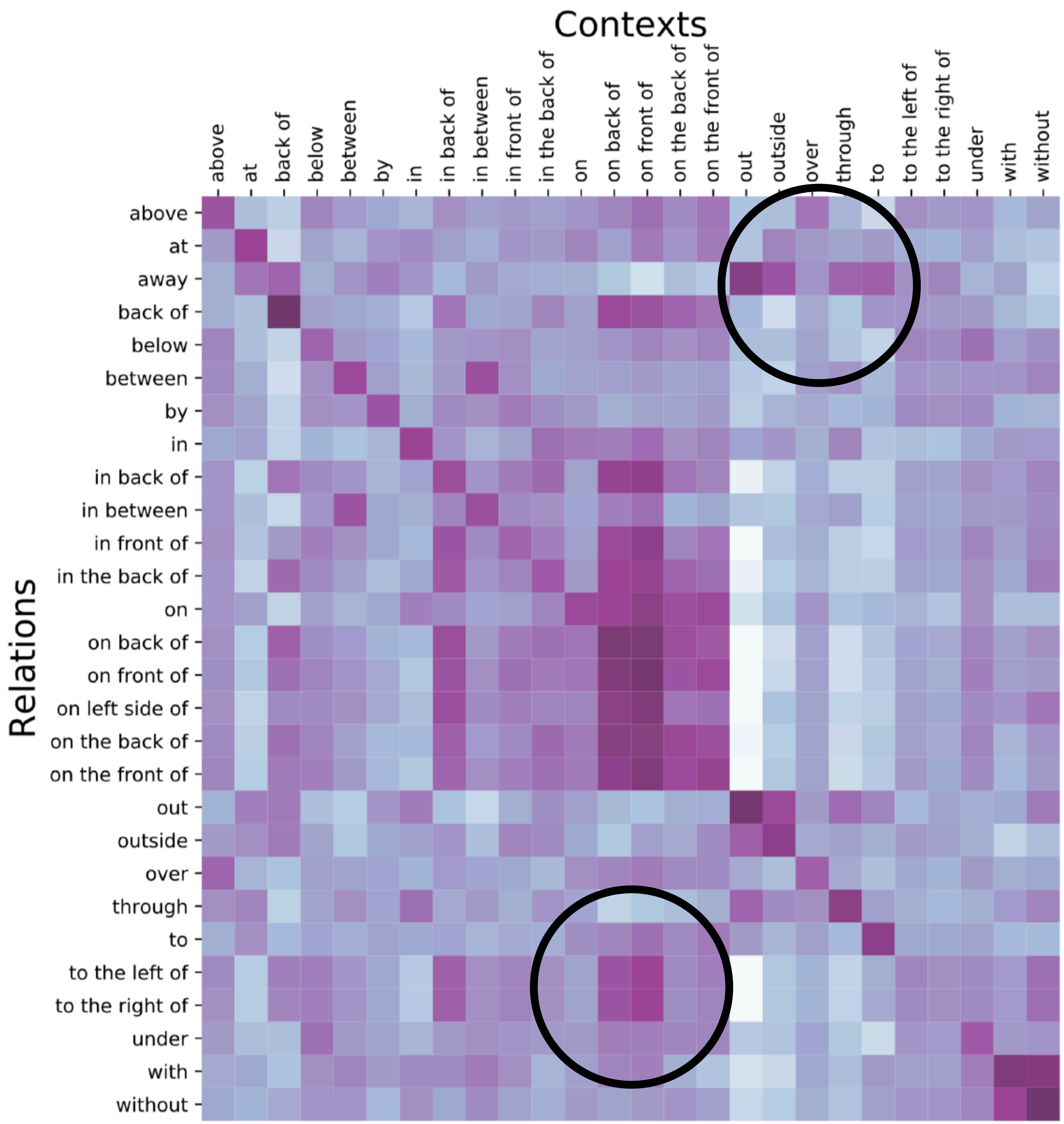
Mehdi Ghanimifard **Simon Dobnik**

- Investigate the knowledge about spatial relations learned from textual features in neural language models

- Estimate model perplexity on sentences in which the original relation *i* is replaced with an alternate relation *j*

Relation (rel_i)	Context bin (c_{rel_i})
above	scissors ______ the pen tall building ______ the bridge …
below	pen is ______ scissors bench ______ the green trees …
next to	a ball-pen ______ the scissors car ______ the water …

$$PP(S_{i\to j}) = PP_{i,j} = P(rel_i, c_{rel_j})^{\frac{1}{-N'}}$$



- Intuitive k-means clusters arise from the P-vectors

1. to	18. up; down; off
2. on	19. with; without
3. away	20. together; out
4. here	21. outside; inside
5. into	22. near; beside; by
6. from	23. top; front; bottom
7. during	24. in between; between
8. back of	25. along; at; across; around
9. through	26. beneath; below; under; behind
10. alongside	27. right; back; left; side; there
11. along side	28. to the left of; to the right of; next to
12. underneath	29. in back of; in the back of; on the back of; at the top of
13. in; against	30. on the top of; on side of; on the bottom of; on left side of; on top of; on the front of; on back of; on the side of; on front of; on bottom of
14. in front of	
15. above; over	
16. to the side	
17. onto; toward	

- Evidence that language models and the derived P-vectors capture spatial knowledge from only textual features.

RQ2: How is spatial knowledge learned in generative language models?

Study 5

Knowing When to Look For What and Where: Evaluating Generation of Spatial Descriptions with Adaptive Attention

Mehdi Ghanimifard[0000−0002−2598−5091] and Simon Dobnik[0000−0002−4019−7966]

- To what degree does an adaptive attention model attend to visual information when generating spatial relations?

$$\hat{\boldsymbol{c}}_t = \beta_t \boldsymbol{s}_t + (1 - \beta_t)\boldsymbol{c}_t$$

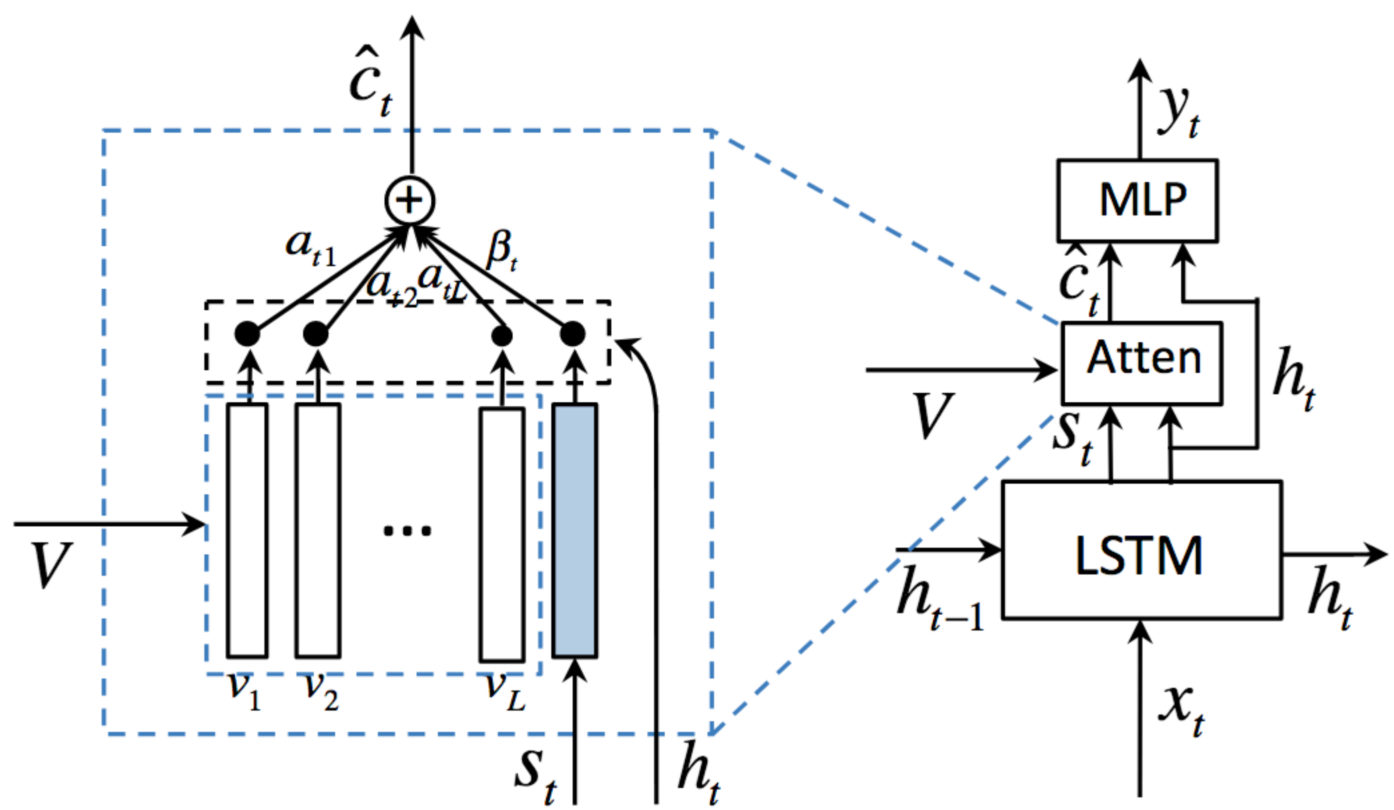


Figure: Lu et al. 2017

- Generate descriptions for 40K images in the MS COCO test set. Part-of-speech tag the generated sentences and determine the visual attention per type of word:

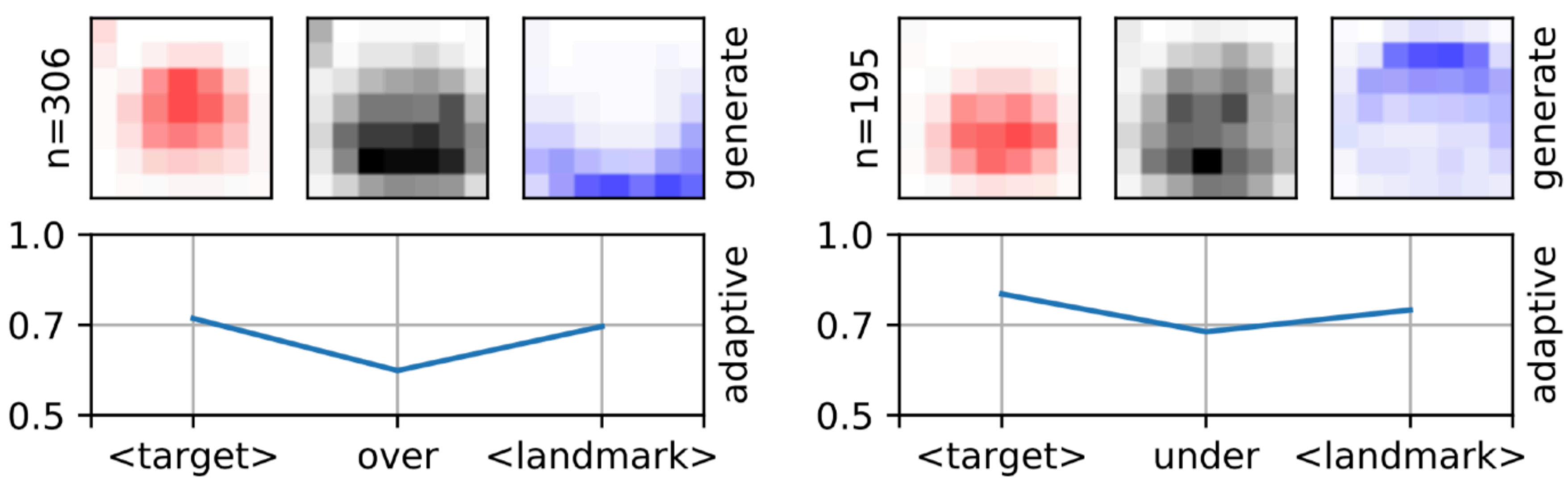
POS	Count	Mean $\pm$ std
NUM	1882	0.81 ± 0.08
NOUN	134332	0.78 ± 0.12
ADJ	23670	0.77 ± 0.14
DET	96641	0.73 ± 0.12
VERB	38381	0.70 ± 0.11
CONJ	6755	0.70 ± 0.13
ADV	184	0.69 ± 0.12
ADP	64332	0.62 ± 0.15
PRON	2347	0.53 ± 0.14
PRT	6462	0.52 ± 0.21

Spatial relations

Average visual attention $(1 - \beta_t)$

- Hypothesis: “When generating spatial relations, the visual attention is more spread over possible regions instead of focused on a specific object”

Descriptions	Average $(1 - \beta_t)$
Spatial Relations	**TRG, REL, LND**
under	0.84, **0.73**, 0.79
front	0.83, **0.70**, 0.82
next	0.82, **0.68**, 0.78
back	0.85, **0.68**, 0.84
in	0.82, **0.68**, 0.77
on	0.81, **0.68**, 0.75
near	0.80, **0.67**, 0.76
over	0.77, **0.62**, 0.75
above	0.73, **0.64**, 0.77



- Overall, adaptive attention focuses on visual objects

Study 6

What Goes Into A Word: Generating Image Descriptions With Top-Down Spatial Knowledge

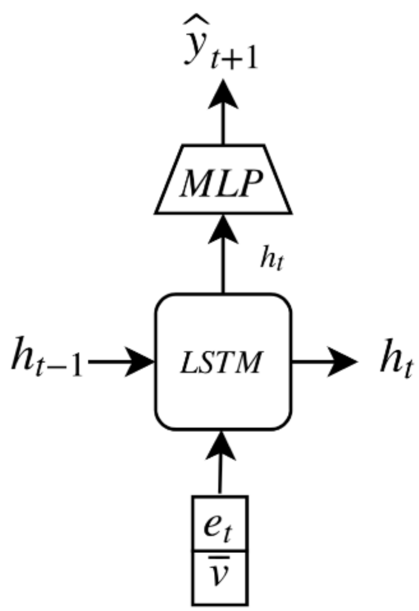
Mehdi Ghanimifard **Simon Dobnik**

- How much spatial information is needed to generate accurate descriptions of images?

⟨ "bat", "over", "shoulder"⟩

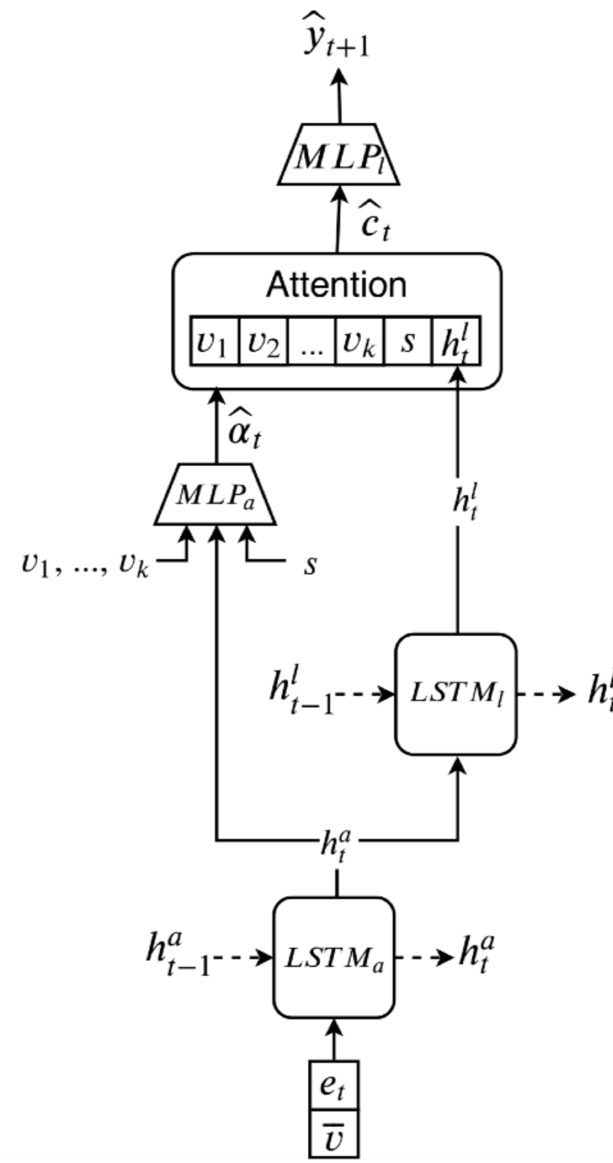
simple	player
bu49	man wearing shirt
td	bat in hand
td order	bat in hand
td order+VisKE	bat in hand

Global feature vector



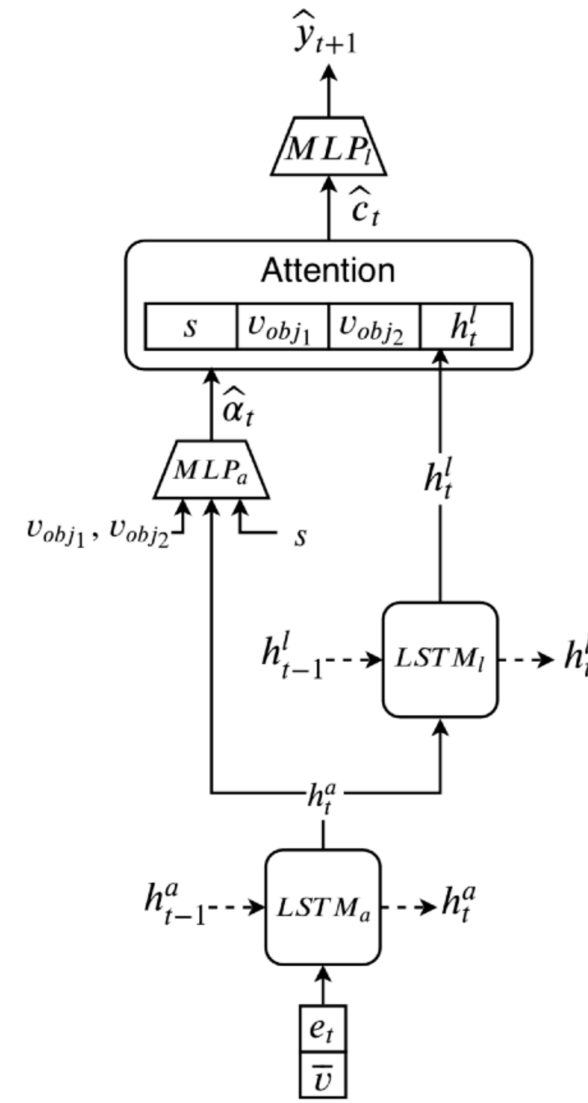
“simple”

7 x 7 grid of dense image features



“bu49”

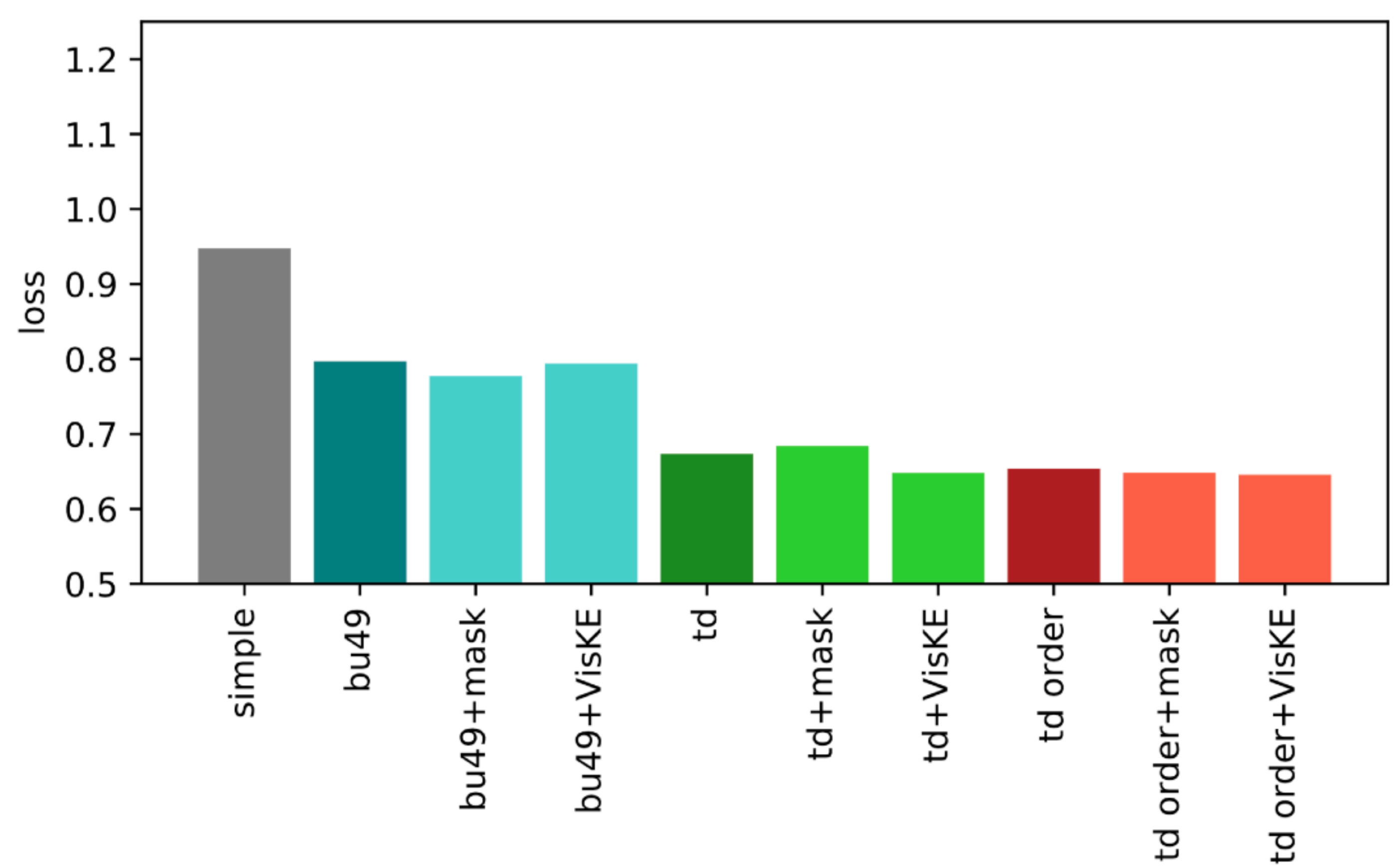
Feature vectors from bounding boxes



“td”

- Top-down features for TARGET-LANDMARK pairs is the most useful source of visual supervision.

- Geometric features do not have a significant effect



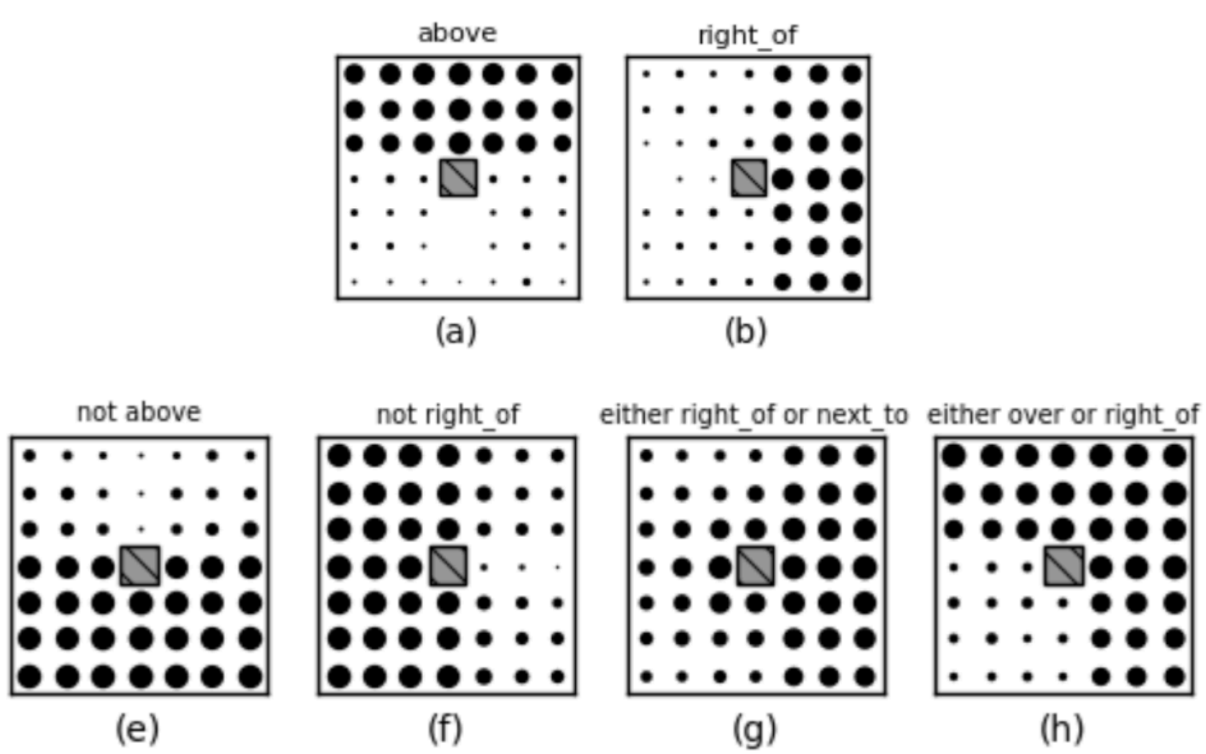
- Overall, top-down localisation is crucially important to generating accurate region descriptions.

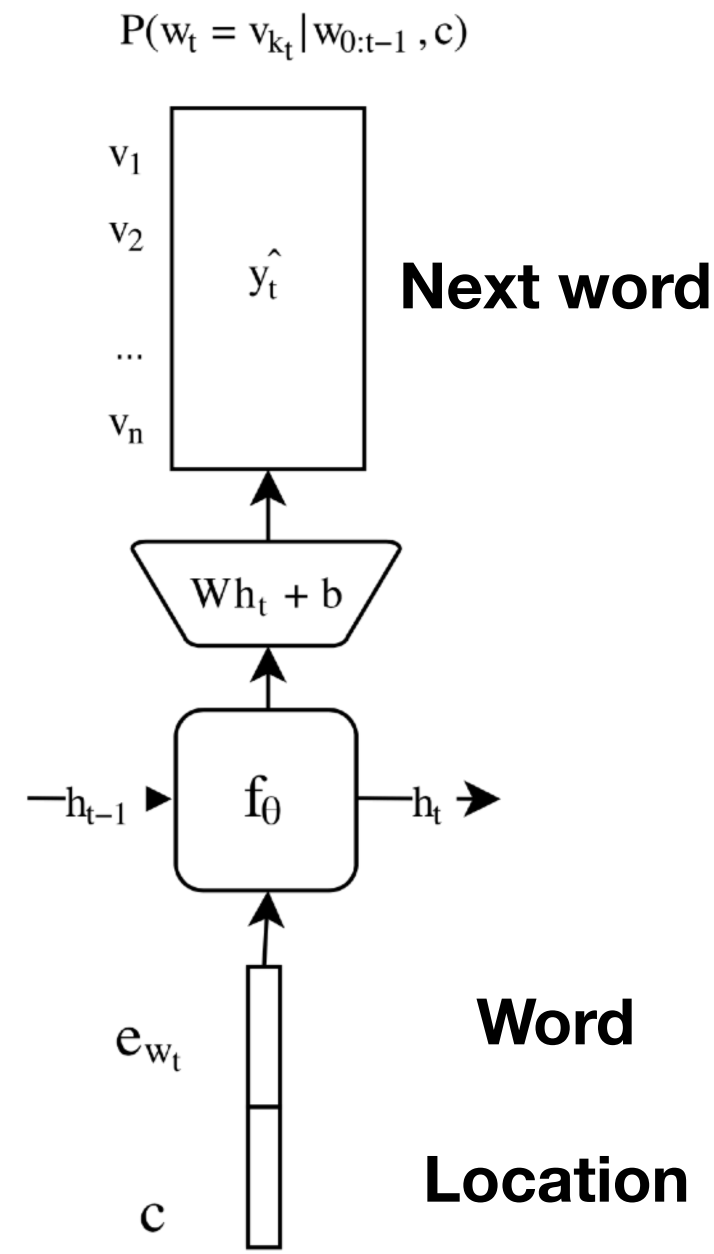
RQ3: Are neural language models capable of systematic generalisation?

Study 1

Learning to Compose Spatial Relations with Grounded Neural Language Models

- To what extent is the language model grounded in spatial representations?
- Work with **spatial templates** over 7 x 7 grids (Logan and Sandler, 1997)

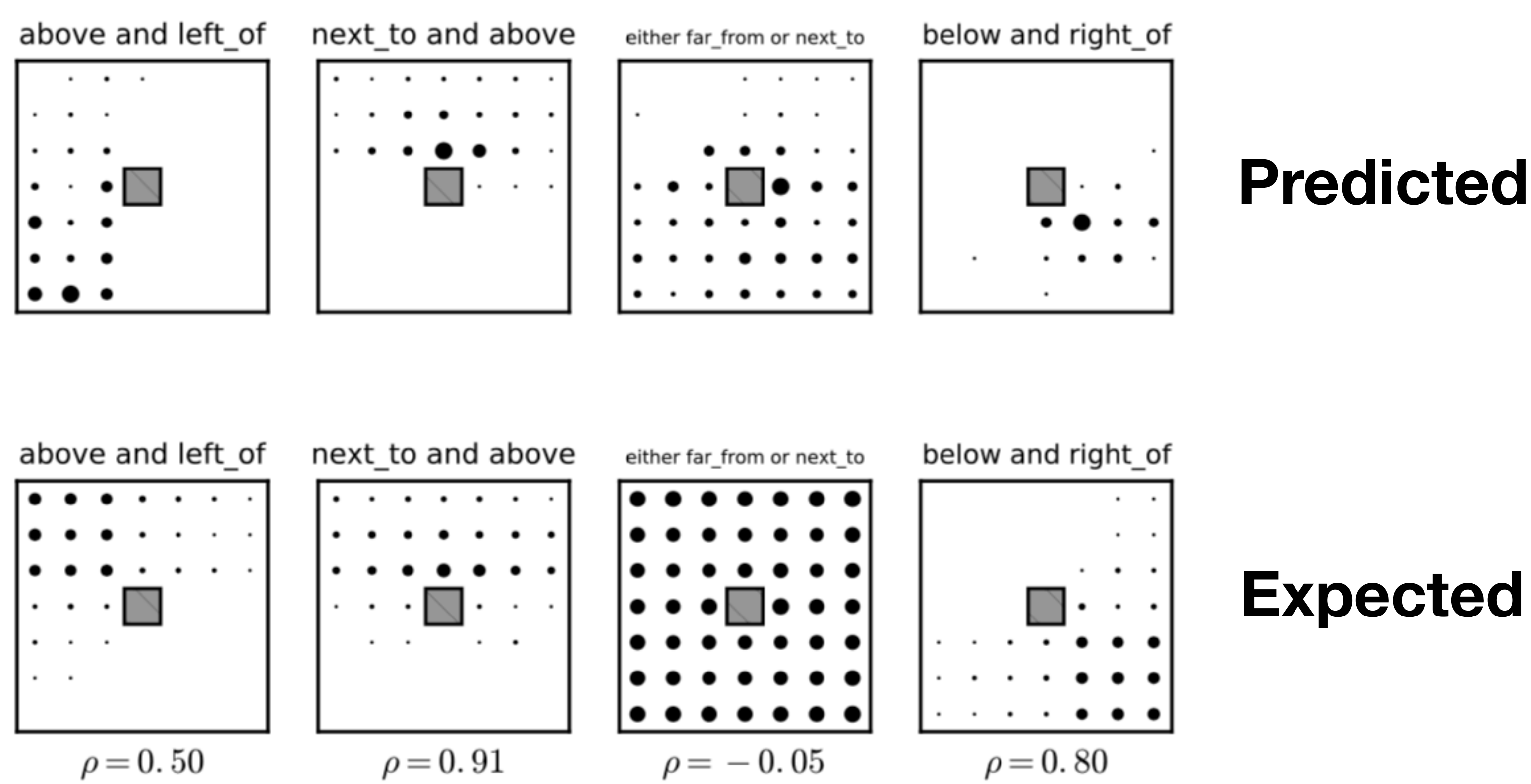




	Simple phrases	With distractors	Untrained
AND-phrases	0.87	0.85	-0.00
NEG-phrases	0.72	0.82	0.03
OR-phrases	0.79	0.80	-0.03
SINGLE-word	0.92	0.91	-0.05
All previous	0.83	0.83	-0.01
All previous + distractors	NaN	0.84	-0.03

- Models are sensitive to the amount of training data

Proportions of 90 combinations	10%	20%	30%	40%	50%
AND-phrases	0.84	0.8	0.78	0.76	0.71
OR-phrases	0.74	0.73	0.69	0.67	0.56



Study 2

"Deep" Learning: Detecting Metaphoricity in Adjective-Noun Pairs*

- Predicting the metaphoricity of adjective-noun pairs in 8,592 pairs in the Gutiérrez et al. (2016) dataset.

Bright painting / bright idea

- Model with a sigmoidal function of the dot product between the adjective-noun phrase vector **p** and a learned *metaphoricity vector* **q**

$$\hat{y} = \sigma(\mathbf{p} \cdot \mathbf{q} + b_1) = \frac{1}{1 + e^{-\mathbf{p} \cdot \mathbf{q} + b_1}}$$

- Main ideas:
  - transfer learning from pre-trained embeddings
  - learned composition with neural networks

Models

- Concatenate

$$\mathbf{p} = f_\theta(\mathbf{u}, \mathbf{v}) = W^T \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} + b$$

- Additive with shared projection matrix

$$\mathbf{p} = f_\theta(\mathbf{u}, \mathbf{v}) = W^T\mathbf{u} + W^T\mathbf{v} + b$$

- Element-wise multiplicative interaction

$$\mathbf{p} = f_\theta(\mathbf{u}, \mathbf{v}) = (\mathbf{u} \times \mathbf{v})W + b$$

Accuracy

	Random W	**Trained W**
cat-linear	0.8973	0.9153
cat-relu	0.8763	0.9228
sum-linear	0.8815	0.9068
sum-relu	0.8597	0.9150
mul-linear	0.7858	0.8066
mul-relu	0.7795	0.8186

More abstract

Top ten	reluctance, reprisal, resignation, response, rivalry, satisfaction, storytelling, supporter, surveillance, vigilance
Bottom ten	saucepan, flour, skillet, chimney, jar, tub, fuselage, pellet, pouch, cupboard

More concrete

Summary

- This thesis offers a comprehensive study of the representation of spatial language in neural network language models.

- The experiments on the role of visual context are illuminating and demonstrate the utility of bounding box object representations.

- Raises important questions about what is needed from the visual component of a vision and language model.

Questions

- So, why is the pond not outside the frog? What evidence do the studies in this thesis bring to this question?
- What do distributional representations of language tell us about the substitutability of TARGETS and LANDMARKS?
- Do you think the results from your experiments would hold for different languages?
  - How would you go about testing this?

More Questions

- Study 1: why do distractors improve the correlation with the original spatial templates for NEG-phrases? (Table 1)
- Study 2: what exactly is **q**? How would we understand the learning process that generates a *metaphoricity* vector?
- Study 4: would you expect to find similar results if you worked with pre-trained language models? (Fewer tokens would fall below the 100 token threshold.)
- Study 5: do attention-based captioning models attend to objects “just-in-time” or in order to generate a sequence of tokens?
- Study 6, what would be the performance of the Top-down localisation approach if you did not have annotated bounding boxes?

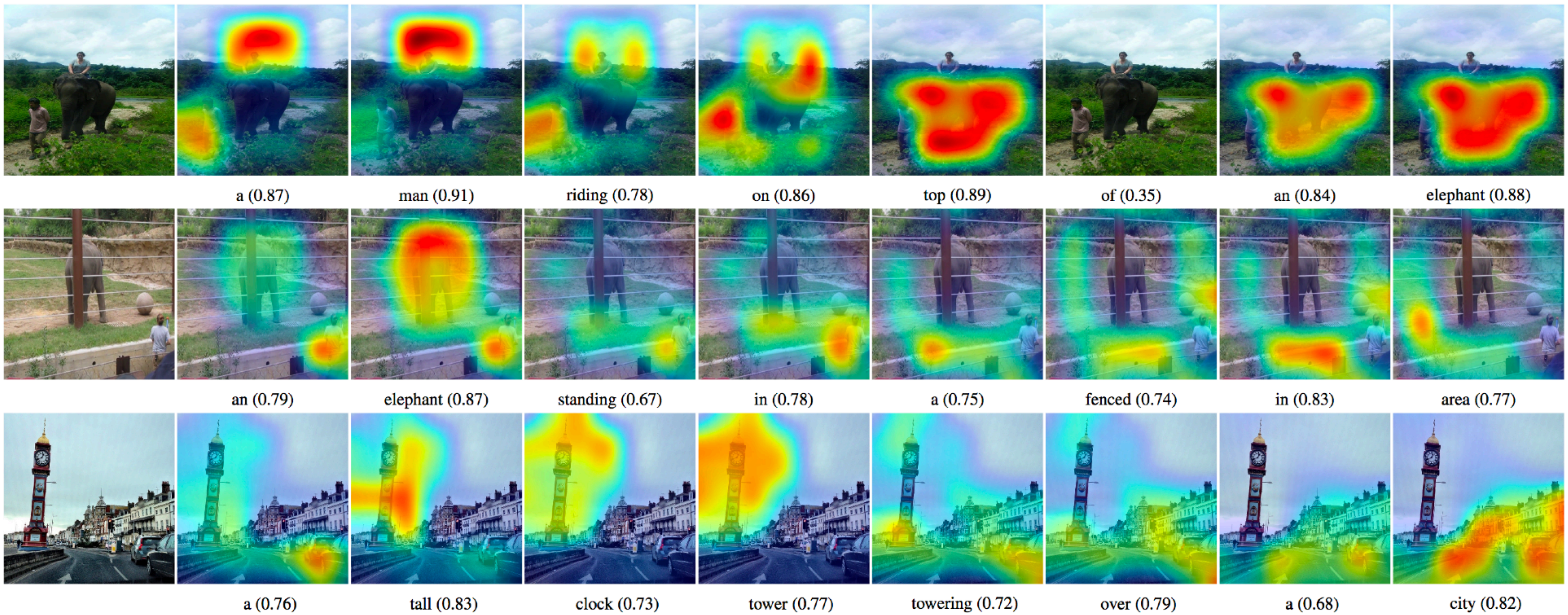


Figure: Lu et al. 2017