UNIVERSITY OF GOTHENBURG

Spatial Knowledge in Neural Language Models

Mehdi Ghanimifard, Simon Dobnik
Centre for Linguistic Theory and Studies in Probability (CLASP)
University of Gothenburg

CLASP Seminar, Gothenburg, Sweden
10 October 2018

CLASP centre for linguistic theory and studies in probability

Generating Scene Description

Figure: *Flickr image in MSCOCO dataset id=330177*

Generating Scene Description

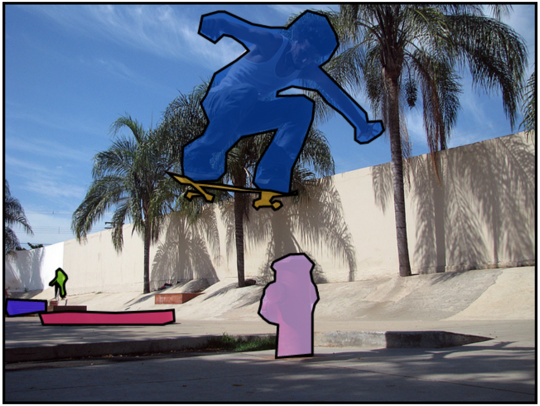


Figure: a **man** is *jumping on* a **skateboard** *over* a **fire hydrant**. [1]

[1] MSCOCO dataset id=330177

Generating Scene Description: Grounding and Compositionality

UNIVERSITY OF GOTHENBURG

Language Model

Words, Sentences, ...

["man", "jump", "hydrant", "a", "on", "over", …]

("a man is jumping on a skateboard over a fire hydrant." , )

Perceptual Models

***Objects**: man, skateboard, hydrant, …*
***Events**: jumping, walking, …*
***Spatial Relations**: over, on, in_front_of, …*

CLASP centre for linguistic theory and studies in probability

Generating Scene Description: Grounding and Compositionality

UNIVERSITY OF GOTHENBURG

- Extracting visual features.
- Connecting visual features with linguistic units.
- Generating acceptable word sequence.

CLASP centre for linguistic theory and studies in probability

Generating Scene Description: Grounding and Compositionality

- Extracting visual features. ($\rightarrow$ ConvNet)
- Connecting visual features with linguistic units. ($\rightarrow$ Conditional LM)
- Generating acceptable word sequence. ($\rightarrow$ Conditional RLM)

Generating Scene Description: Extracting visual features

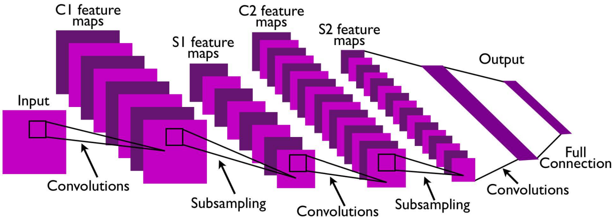


Figure: A ConvNet with two feature map layers (LeCun et al., 2010).

Generating Scene Description: Generating Description

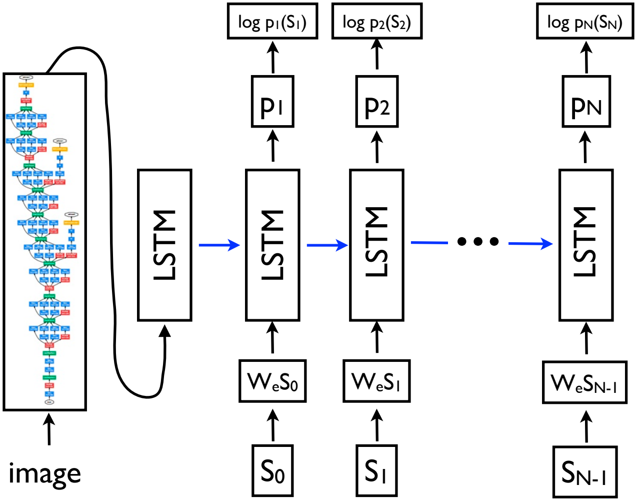


Figure: A Conditional Recurrent Language Model (Vinyals et al., 2015).

Question

How to ground "spatial relations" in visual clues?

Question

UNIVERSITY OF GOTHENBURG

How to ground "spatial relations" in visual clues?

- Not all aspects of meaning in spatial terms are visual.
- CNN features doesn't correspond to any explicit spatial representation.

Question: Distributional Bias In Language

UNIVERSITY OF GOTHENBURG

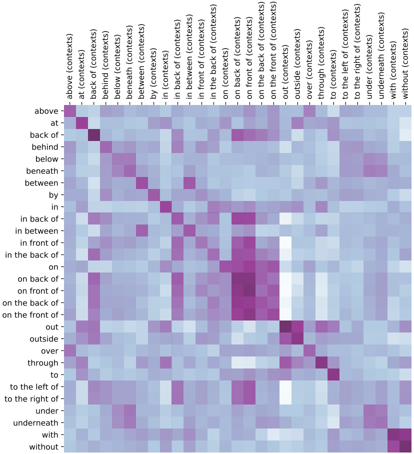


Figure: The distributional bias in language makes spatial relations predictable without looking at images. (Ghanimifard and Dobnik, in SLTC 2018)

CLASP centre for linguistic theory and studies in probability

Question: Spatial Attention

UNIVERSITY OF GOTHENBURG

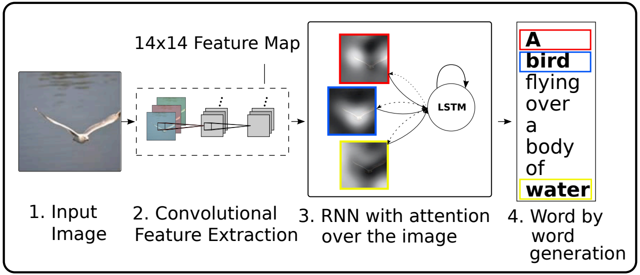


Figure: The spatial attention for caption generation in Xu et al. (2015).

- Weighted pool instead of average pool on last layer of ConvNet.
- Attention weights based on the hidden states of the language model.

CLASP centre for linguistic theory and studies in probability

Question: Spatial Attention

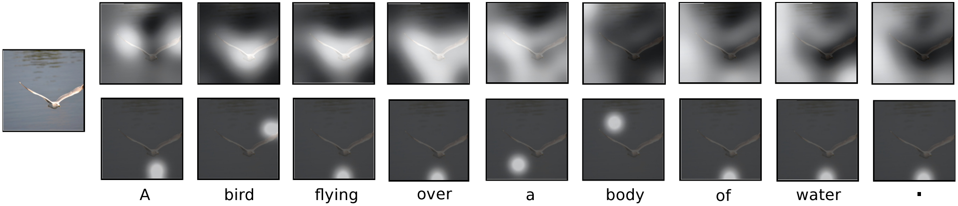


Figure: An example of spatial attention for caption generation in Xu et al. (2015).

Adaptive Attention

- Spatial attentions similar to Xu et al. (2015)
- \+ Attention on a representation from recurrent language model.

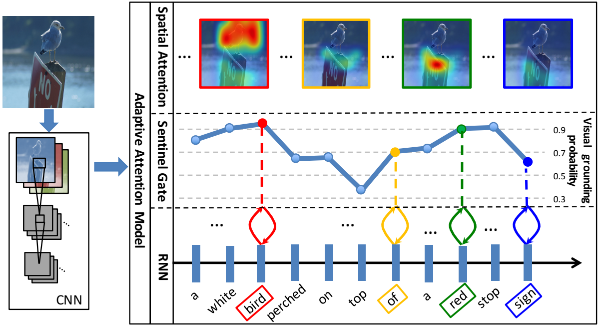


Figure: Adaptive attention on two types of input features Lu et al. (2017)

Adaptive Attention: When to Attend

POS	Mean $\pm$ std
NUM	0.81 ± 0.08
NOUN	0.78 ± 0.12
ADJ	0.77 ± 0.14
DET	0.73 ± 0.12
VERB	0.70 ± 0.11
CONJ	0.70 ± 0.13
ADV	0.69 ± 0.12
ADP	0.62 ± 0.15
PRON	0.53 ± 0.14
PRT	0.52 ± 0.21

Table: Visual attention is stronger on nominal phrase (NOUN, DET, ADJ) (Ghanimifard and Dobnik, 2018)

Adaptive Attention: When/Where to Attend

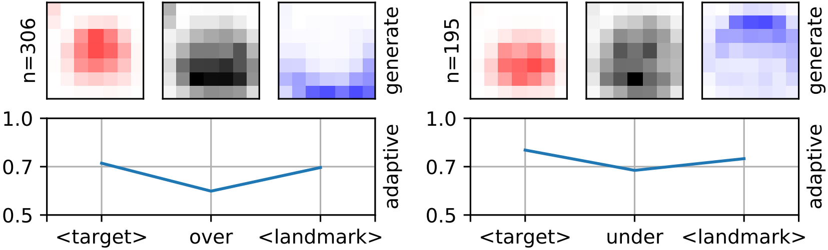


Figure: Visual attention on spatial relations are lower and more spread over 2D space (Ghanimifard and Dobnik, 2018).

Adaptive Attention: Question

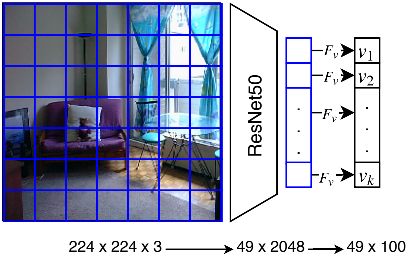
- Can we improve this?

Method

UNIVERSITY OF GOTHENBURG

- Gradually add modules to the neural network and compare the results.
  - Including different ways to use spatial features.
- Adaptive attention as the base architecture.
- Enhance the visual features with annotated information.
  - Annotations as feature extraction tool.
  - Annotations as explicit spatial features.

Method: Extracting visual features

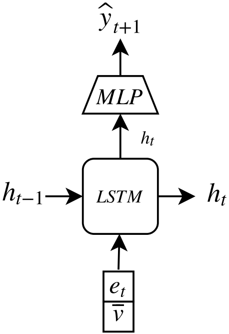


$$F_v : W_v \in \mathbb{R}^{100\times 2048}, b_v \in \mathbb{R}^{100}$$

$$v_i = ReLU(W_v v_i' + b_v)$$

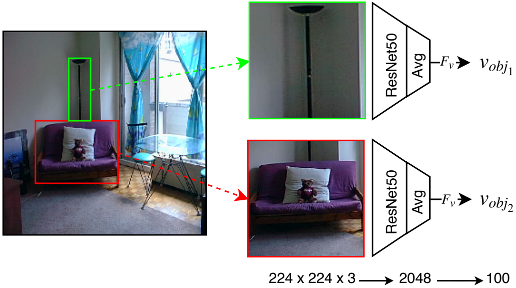
$$\bar{v} = \sum_{i=1}^{k} v_i$$

Method: The Simple Model



Method: Annotation for extracting features

UNIVERSITY OF GOTHENBURG



$$v_{obj_1} = ReLU(W_v v'_{obj1} + b_v)$$
$$v_{obj_2} = ReLU(W_v v'_{obj2} + b_v)$$

CLASP centre for linguistic theory and studies in probability

Method: Annotations as explicit spatial features

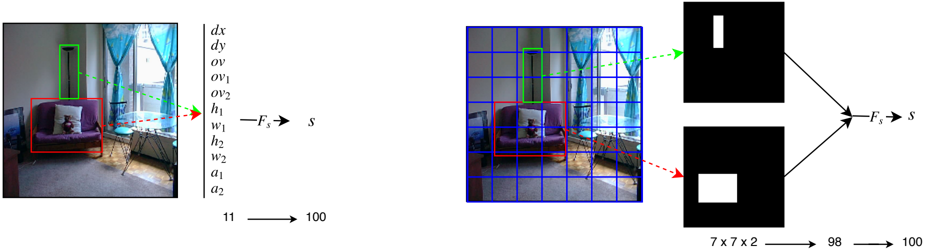
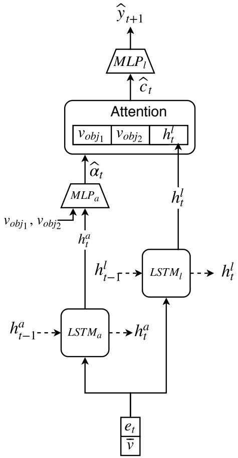


Figure: Two strategies to convert bounding box information into feature representation of their spatial relations.

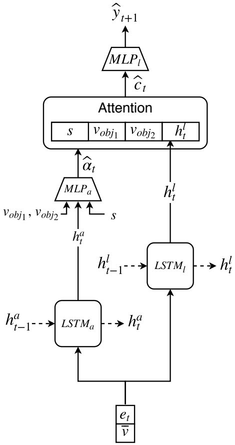
$$F_s : W_s^2 \in \mathbb{R}^{100\times100}, W_s^1 \in \mathbb{R}^{100\times11}(or\mathbb{R}^{100\times98})$$
$$s = W_s^2 tanh(W_s^1 s' + b_s^1)$$

CLASP centre for linguistic theory and studies in probability

Method: Models

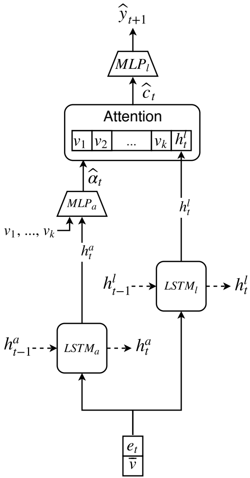


Use annotations to extract visual vectors



\+ Spatial features as an extra feature vector

Method: Models



(c)

Figure: (c) Inspired from adaptive attention in Lu et al. (2017)

Evaluation: Dataset

lamp behind couch
teddy bear on couch
chair next to table
table top near window
...

Figure: Annotated relations in VisualGenome (Krishna et al., 2017)[2].

[2] https://cs.stanford.edu/people/rak248/VG_100K/4.jpg

Evaluation: Dataset

- **Dataset**: VisualGenome with 108K Images.
- $\sim$ 2 million annotated triplets in relation dataset: [obj1, rel, obj2] (maximum 15 words)
- After pre-processing: 1.6 million phrases.
- **Training**: trained on 95% and test on 5% (80K phrases)

Evaluation: Results

Token level loss on validation data after 15 epochs training (cross-entropy error):

- 0.9490 no attention
- 0.7968 adaptive attention on ConvNet regions
- 0.6522 only object vectors
- 0.6484 object vectors + (98D) explicit spatial vector
- 0.6455 object vectors + (11D) explicit spatial vector

Evaluation: How it works?

UNIVERSITY OF GOTHENBURG

h_t	s	obj_1	obj_2	word
0.802	0.028	**0.165**	0.005	man
0.773	**0.170**	0.020	0.037	in
0.839	0.031	0.008	**0.121**	jacket
0.899	0.012	0.013	0.076	EOS

CLASP centre for linguistic theory and studies in probability

Evaluation: How it works?

UNIVERSITY OF GOTHENBURG

h_t	s	obj_1	obj_2	word
0.796	**0.054**	**0.111**	0.039	topping
0.876	**0.034**	0.030	**0.060**	on
0.846	0.002	0.018	**0.134**	a
0.849	0.003	0.009	**0.140**	pizza
0.900	0.004	0.011	0.085	EOS

CLASP centre for linguistic theory and studies in probability

Evaluation: Analysis

Observations:

- Language model gets the highest attention.
- Using spatial annotations improves the results.
- Spatial annotations as feature vector has potentials for deeper investigation.

Summary

We compared end-to-end language generation enriched with spatial knowledge:

- Spatial knowldge to extract visual features.
- Spatial knowldge as feature vectors.

Discussions

- Visual grounding of spatial terms can be grounded in:
  - (1) visual clues from locations (where)
  - (2) visual clues from objects (what)
- Spatial knowledge about "where" can be used for finding "what".
- Spatial relations are different from just location of two objects.
- Visual relations are rich concepts:
  non-spatial aspects in spatial relations "on", "in", etc.
  spatial aspects in non-spatial relations "wearing", "working on", etc.

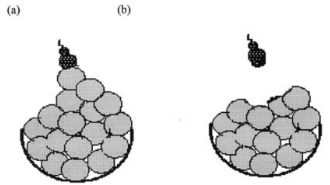


Figure: The meaning of "in" is an interplay between functional and geometric aspects (Coventry et al., 2001)

Future

- More in-depth evaluating CNNs for detecting spatial relations.
- Augment other model such as (Anderson et al., 2018) with spatial information.
- Explore other spatial representations (i.e. AVS).
- Test on other datasets.
- Report on different part of speech.

Thank you!

References I

Anderson, P., X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, Volume 3, pp. 6.

Coventry, K. R., M. Prat-Sala, and L. Richards (2001). The interplay between geometry and function in the comprehension of over, under, above, and below. *Journal of memory and language 44*(3), 376–398.

Ghanimifard, M. and S. Dobnik (2018). Knowing when to look for what and where: Evaluating generation of spatial descriptions with adaptive attention. In *Proceedings of the 1st Workshop on Shortcomings in Vision and Language (SiVL'18), ECCV, 2018*.

Krishna, R., Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision 123*(1), 32–73.

References II

LeCun, Y., K. Kavukcuoglu, C. Farabet, et al. (2010). Convolutional networks and applications in vision. In *ISCAS*, Volume 2010, pp. 253–256.

Lu, J., C. Xiong, D. Parikh, and R. Socher (2017). Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Volume 6.

Vinyals, O., A. Toshev, S. Bengio, and D. Erhan (2015). Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pp. 3156–3164. IEEE.

Xu, K., J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pp. 2048–2057.