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

Mehdi Ghanimifard Simon Dobnik
Centre for Linguistic Theory and Studies in Probability (CLASP)
Department of Philosophy, Linguistics and Theory of Science (FLoV)
University of Gothenburg, Sweden
{mehdi.ghanimifard,simon.dobnik}@gu.se

INLG 2019
31 October 2019

UNIVERSITY OF GOTHENBURG

CLASP centre for linguistic theory and studies in probability

Outline

UNIVERSITY OF GOTHENBURG

CLASP centre for linguistic theory and studies in probability

Motivations (1/3): Spatial Language In Image Descriptions

Figure: VisualGenome 2318741

UNIVERSITY OF GOTHENBURG

CLASP centre for linguistic theory and studies in probability

Motivations (1/3): Spatial Language In Image Descriptions

There is a teddy bear partially under a go cart.

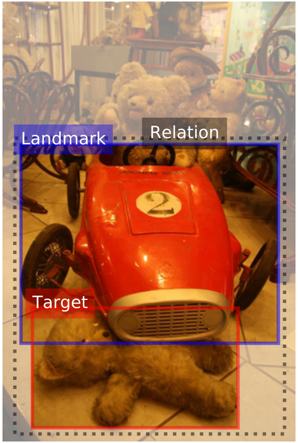
Figure: VisualGenome 2318741

Motivations (1/3): Spatial Language In Image Descriptions

There is $\underbrace{\textit{a teddy bear}}_{\textsc{Target}}$ $\underbrace{\textit{partially under}}_{\textsc{Relation}}$ $\underbrace{\textit{a go cart}}_{\textsc{Landmark}}$.

Figure: VisualGenome 2318741

Motivations (1/3): Spatial Language In Image Descriptions



TARGET-features
LANDMARK-features
Spatial Arrangements
Functional/Contextual Relations
Syntactic and linguistic features

⟨ teddy bear, partially under, go cart⟩

Figure: VisualGenome 2318741

CLASP centre for linguistic theory and studies in probability

Motivations (2/3): Processes of Grounding and Compositionality

UNIVERSITY OF GOTHENBURG

Two kinds of processes and representations:

- Bottom-up: data-driven / recognizing objects.
- Top-down: expectation-driven / recognizing relations.

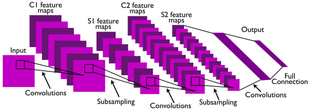
◇ How to integrate both in one system?

CLASP centre for linguistic theory and studies in probability

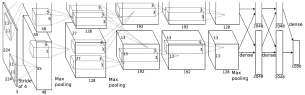
Motivation (3/3): Deep Neural Networks Paradigm

UNIVERSITY OF GOTHENBURG

Relaxing Spatial Transformation



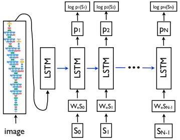
ConvNets (LeCun et al., 2010).



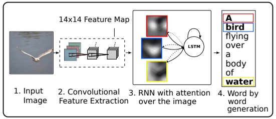
ImageNet: Object Recognition (Deng et al., 2009; Krizhevsky et al., 2012).

⇒

Generating Captions with Spatial Attention



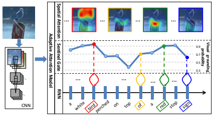
Conditional Recurrent LM (Vinyals et al., 2015).



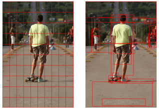
Spatial Attention (Xu et al., 2015).

⇒

Better Attention, Localization & Datasets!



Adaptive Attention (Lu et al., 2017)



Top-down localisation (Anderson et al., 2018).

CLASP centre for linguistic theory and studies in probability

Aims and Questions

- **Aims**:
  - To integrate top-down spatial knowledge in *recurrent language model*.
  - To investigate grounding of image descriptions in feature representations.

Aims and Questions

UNIVERSITY OF GOTHENBURG

- **Aims**:
  - To integrate top-down spatial knowledge in *recurrent language model*.
  - To investigate grounding of image descriptions in feature representations.
- **Questions**:
  - What kinds of top-down spatial knowledge improves generation?
  - How does each feature contribute to generating image descriptions?

CLASP centre for linguistic theory and studies in probability

Aims and Questions

- **Aims**:
  - To integrate top-down spatial knowledge in *recurrent language model*.
  - To investigate grounding of image descriptions in feature representations.
- **Questions**:
  - What kinds of top-down spatial knowledge improves generation?
  - How does each feature contribute to generating image descriptions?
- **Top-down spatial knowledge**:
  - Localisation
  - Semantic roles
  - Relational spatial features

Methodology

UNIVERSITY OF GOTHENBURG

Build comparable neural networks with spatial knowledge:

- Change spatial attention module.
- Enrich representations with spatial knowledge.

Methodology

UNIVERSITY OF GOTHENBURG

Build comparable neural networks with spatial knowledge:

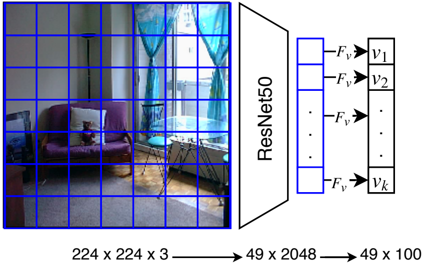
- Change spatial attention module.
- Enrich representations with spatial knowledge.

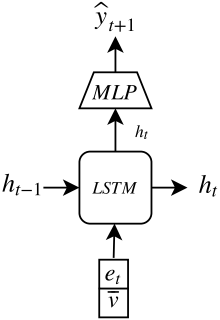
Experiments:

- Compare models' performance (loss / perplexity).
- Inspect contribution of features in word generation.

CLASP centre for linguistic theory and studies in probability

Baseline (1): Bottom-up Encoder-Decoder (*simple*)

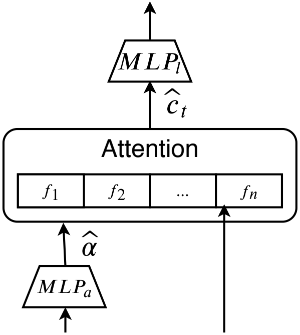




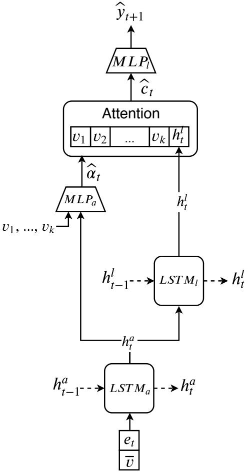
$$v_i = ReLU(W_v v_i' + b_v)$$

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

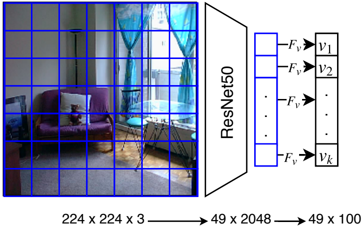
Baseline (2): Bottom-up Spatial Attention (*bu*49)



$$\hat{c} = \sum_{i=1}^{n} \alpha_i f_i$$

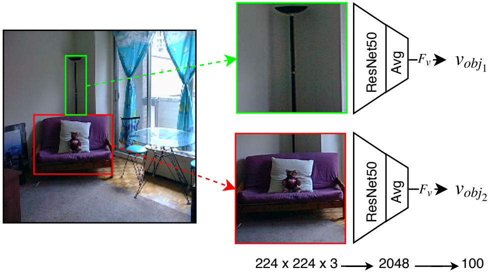


Method (1): Top-down localisation (1/2)



Bottom-up localisation ($bu49$)

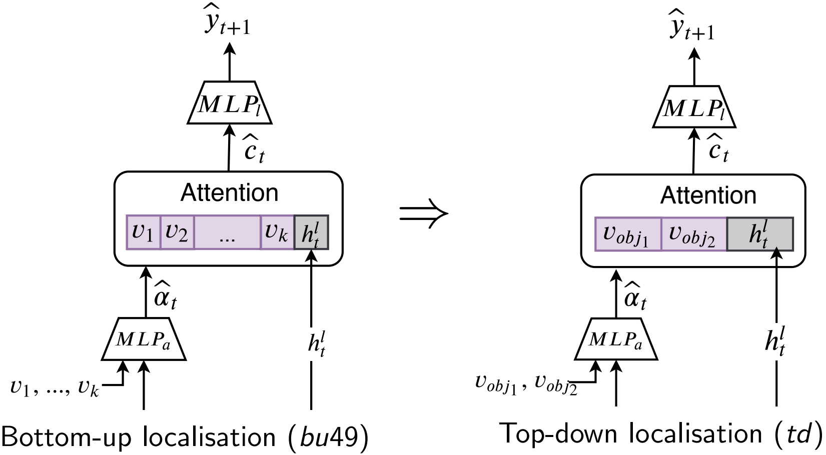
$\Rightarrow$



Top-down localisation (tu)

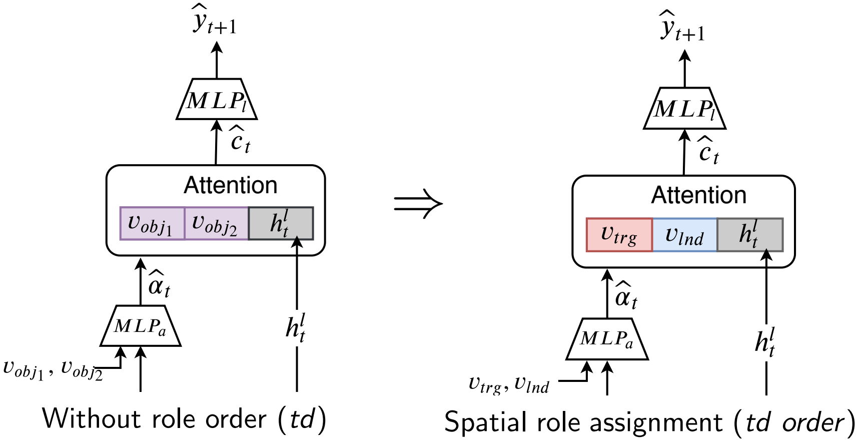
$$v_{obj_1} = ReLU(W_v v'_{obj1} + b_v)$$
$$v_{obj_2} = ReLU(W_v v'_{obj2} + b_v)$$

Method (1): Top-down localisation (2/2)

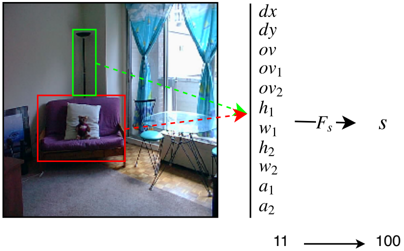


Method (2): Top-down role assignment

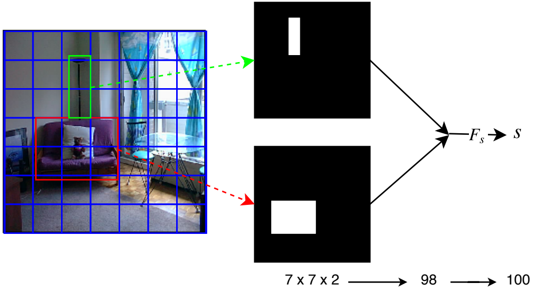
$$(object_1, object_2) \rightarrow (\textsc{Target}, \textsc{Landmark})$$



Method (3): Vectorizing Spatial Configurations (1/2)



VisKE (Sadeghi et al., 2015)

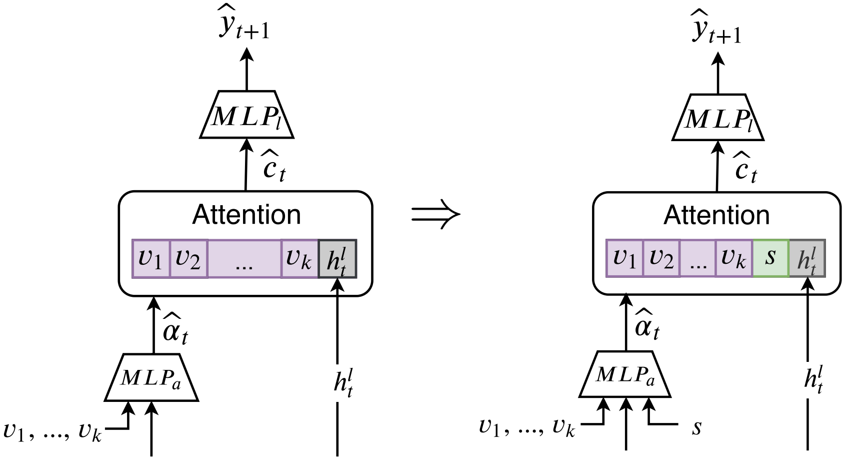


mask

Two strategies to represent s-features from bounding box information.

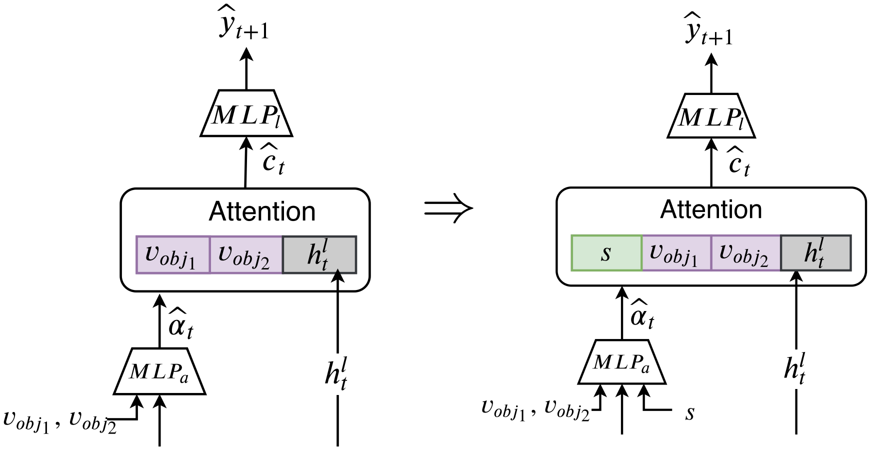
$$s = W_s^2 tanh(W_s^1 s' + b_s^1)$$

Method (3): Vectorizing Spatial Configurations (2/2)



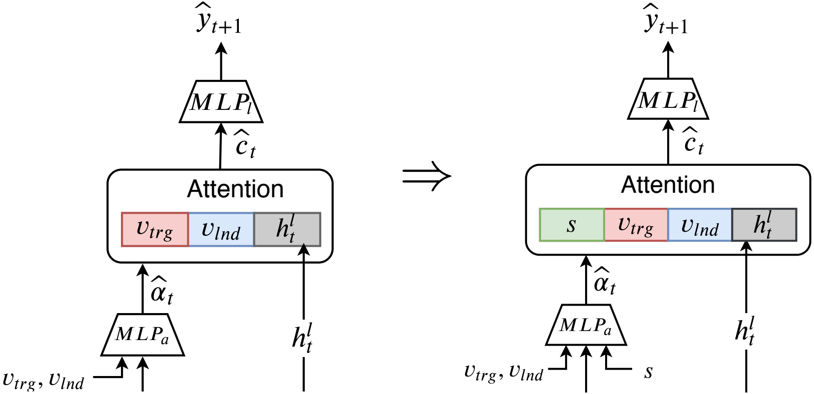
Method (3): Vectorizing Spatial Configurations (2/2)

UNIVERSITY OF GOTHENBURG



CLASP centre for linguistic theory and studies in probability

Method (3): Vectorizing Spatial Configurations (2/2)



Experiments: Dataset

Dataset:

- VisualGenome (Krishna et al., 2017)
- 108K Images.
- $\langle obj_1, rel, obj_2 \rangle \rightarrow$ token sequence (up to 15 tokens).
- 1.6 million examples (15 unique descriptions for each image)

Experiments: Dataset

Dataset:

- VisualGenome (Krishna et al., 2017)
- 108K Images.
- $\langle obj_1, rel, obj_2 \rangle \rightarrow$ token sequence (up to 15 tokens).
- 1.6 million examples (15 unique descriptions for each image)

Training:

- Training on 95% of images
- Experiment on 5% (80K descriptions)

Experiments: Overall Performance

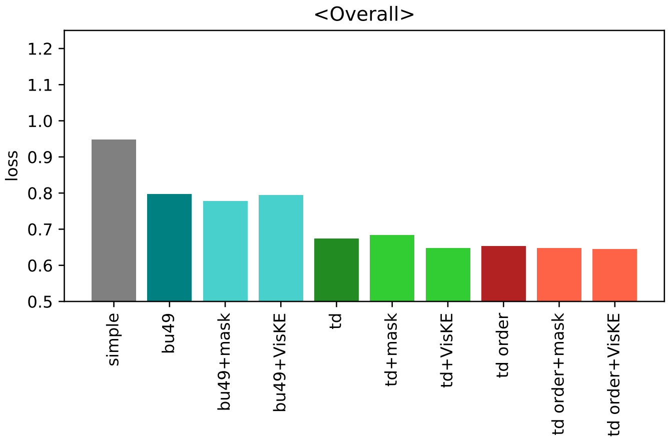


Figure: Cross-entropy loss of different model configurations on evaluation data.

Experiments: Qualitative Examples (Beam Search)

⟨ “bat”, “over”, “shoulder” ⟩

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

⟨ “hood”, “above”, “oven” ⟩

simple	window
bu49	pot on stove
td	oven has door
td order	vent above sink
td order + VisKE	cabinet has door

Figure: From VisualGenome: 2412051[1] 2413282[2]

[1] Herholz (2005): CC BY-SA 2.0.
[2] juanjogasp (2013): CC BY-NC-SA 2.0.

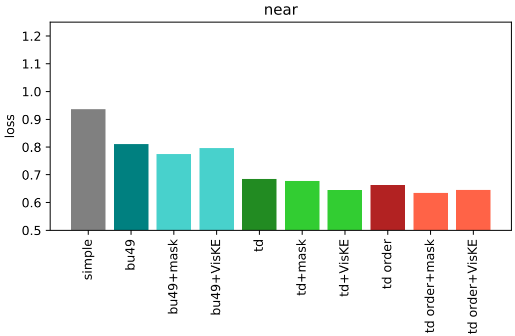
Experiments: *near*, *inside*

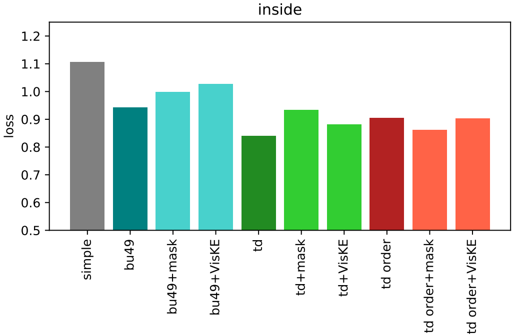
Role assignment effect:

- × roles are predictable. (objects predict context and their own roles)

s-features effect:

- × geometric features are not in 2D dimension.





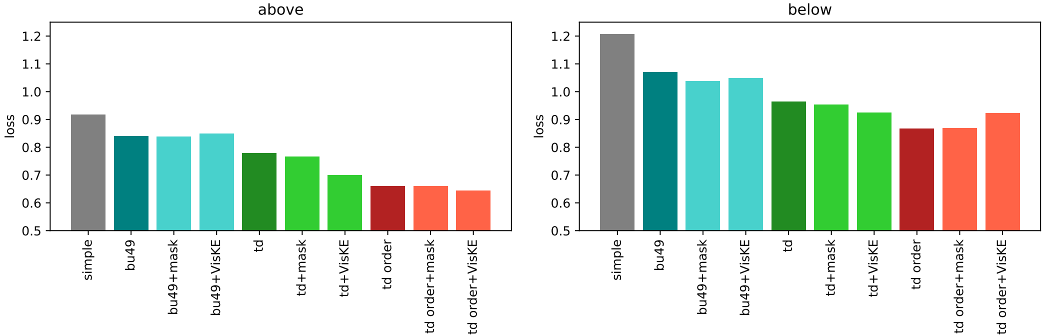
Experiments: *above*, *below*

Role assignment:

- ✓ *above* and *below* are more geometric (not predictable from objects alone).

s-features effect:

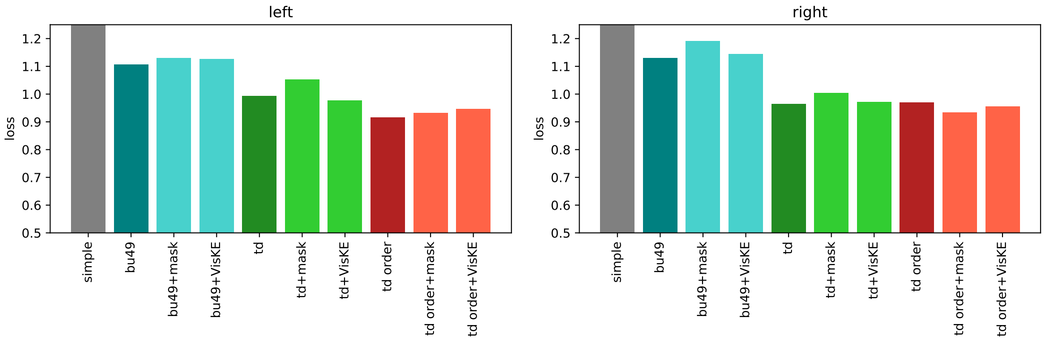
- × *below* is not frequent in training.



Experiments: *left*, *right*

Role assignment and *s*-features effect:

- × *left*, *right* are not frequent in training.

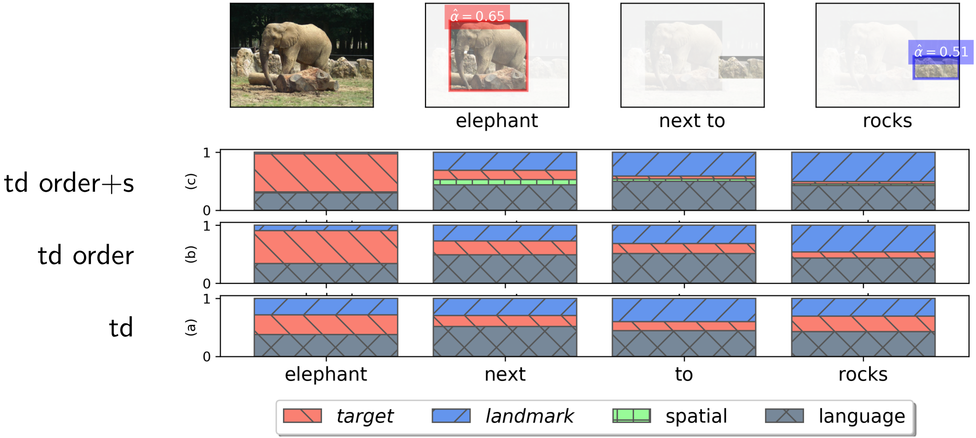


Experiments: Features Contributions

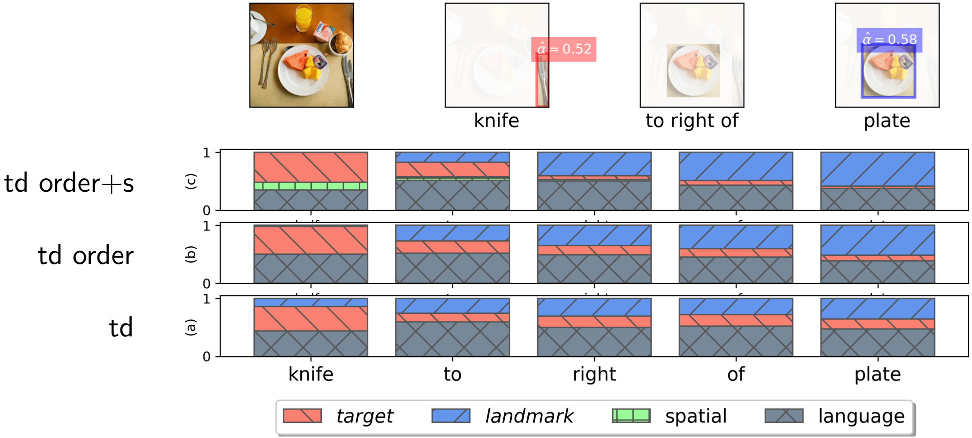
Magnitude of each feature after applying attentions:

$$\beta_{t,f_i} = \frac{\alpha_{t,f_i}||f_i||}{\sum_j \alpha_{t,f_j}||f_j||}$$

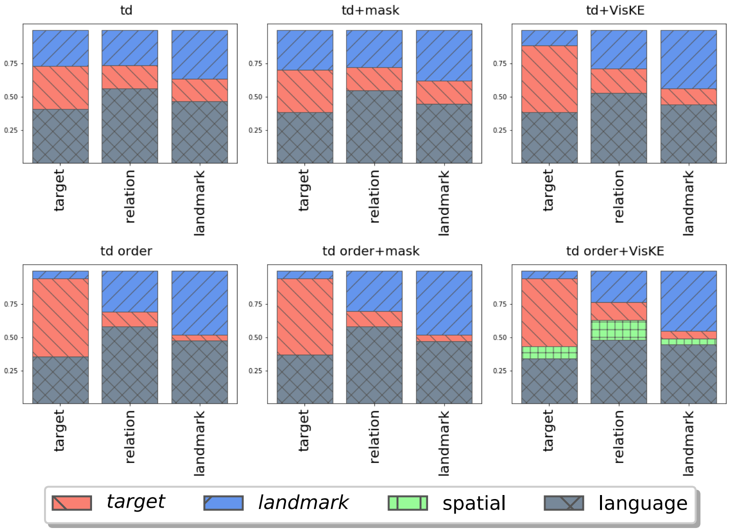
Experiments: Examples of features contributions



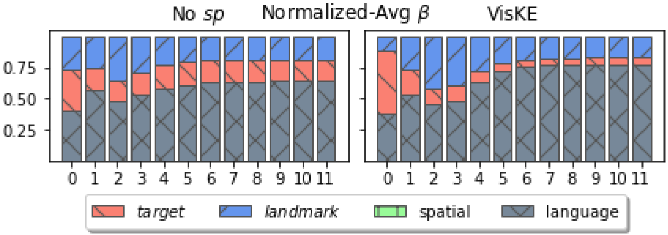
Experiments: Examples of features contributions



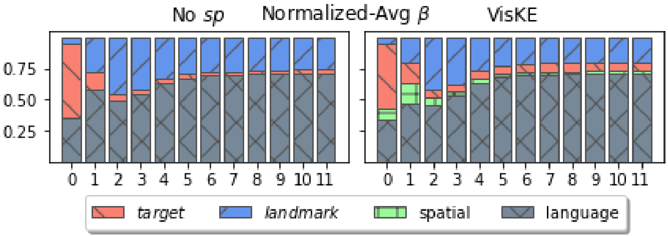
Experiments: Feature contribution based on spatial roles



Experiments: Feature contribution based on token's order



(a) *td* vs. *td+VisKE*



(b) *td order* vs. *td order+VisKE*

CLASP centre for linguistic theory and studies in probability

Summary

We

- ✓ integrated semantic structures as top-down knowledge in Recurrent LM.
- ✓ compared three groups of top-down spatial knowledge:
  - Localisation (bounding boxes)
  - Role Assignment (TRAGET-LANDMARK)
  - Spatial Configuration (*s*-features)
- ✓ measured their effect in model performance.
- ✓ inspected the feature contributions for different semantic roles.

Conclusions

- Overall top-down knowledge lead to better generation (perplexity measures).
- Localisation has the strongest effect.
- Effects of role assignment seems to be dependent on the relations:
  - × more functional / predictable roles (e.g. *inside*)
  - ✓ more geometric relations (e.g. *above*, *below*)
  - × rare relations (e.g. *left*, *right*)
- The effects of *s*-features are small.
  - It is depends on semantic roles assignments.
- Contextual embeddings are the most attended features.
  - Its contribution is increasing along the sequence.

◇ Corpus bias (image compositions)

◇ Task bias (image descriptions are not made to locate objects; i.e. *left*, *right*)

Thank you!

Source code and demo

http://bit.ly/36ixFfR

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. *CVPR 3*(5), 6.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255. IEEE.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena 42*(1-3), 335–346.

Herholz, D. (2005). Wide stance. VisualGenome image id 2412051.

juanjogasp (2013). Baltic trip. VisualGenome image id 2413282.

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

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105.

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.

Sadeghi, F., S. K. Kumar Divvala, and A. Farhadi (2015). Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1456–1464.

References III

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.