Attention as Grounding: Exploring Textual and Cross-Modal Attention on Entities and Relations in Language-and-Vision Transformer

Nikolai Ilinykh Simon Dobnik
Department of Philosophy, Linguistics and Theory of Science
Centre for Linguistic Theory and Studies in Probability (CLASP)
University of Gothenburg, Sweden
{name.surname}@gu.se

ACL 2022

Introduction

GÖTEBORGS UNIVERSITET

- In language-and-vision tasks, each word type should be grounded in both vision and language, but to a different degree (Ghanimifard and Dobnik, 2019).
- **In this paper** we inspect (i) cross-modal attention on objects and, (ii) masked self-attention on text of the multi-modal transformer and see how **two types** of words are learned in the multi-modal setting:
  - words denoting objects in the scene (e.g. "a red chair")
  - words depicting spatial relations between objects (e.g. "a chair *next to* the table")
- We train a simple multi-modal transformer for the task of image description sequence generation and inspect its attention patterns.
- Our results show that the model learns both *syntactic* and *semantic* knowledge about objects and relations, but this knowledge is not *linearly aligned* between two modules of the transformer.

Attention proportion

GÖTEBORGS UNIVERSITET

We compute attention proportion $\boldsymbol{P}$: the amount of attention on specific parts of the input when particular parts of the target sequence are generated:

$$\boldsymbol{P}_{\ell,h}(\alpha \mid S, T) = \frac{\sum_{u \in \mathbf{U}} \sum_{i=1}^{|S|} \sum_{t=1}^{|T|} \alpha(s_i, S | t_j, T)}{\sum_{u \in \mathbf{U}} \sum_{i=1}^{|S|} \sum_{t=1}^{|T|} \alpha(s_i, t_j, T)}, \tag{1}$$

where $\mathbf{S}$ are source units (words, object bounding boxes) with specific conditions imposed on them and $\mathbf{T}$ are target units (words), $\mathbf{U}$ is the set of texts.

Experiment I: masked self-attention on words

Question:

What type of semantic knowledge about objects and relations is captured by *attention on text* in the text decoder?

Conditions:

Restrict **T** to a token t_i.

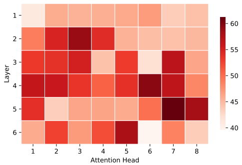
Restrict **S** to previously generated token(s) t_{i-n} which are either ⟨DET, ADJ, NOUN⟩ or ⟨VERB, ADP⟩ POS tags.

Examples:

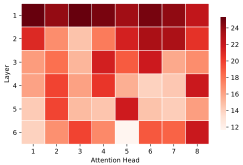
(1) There are **two gold framed pictures** on the walls.

(2) There are two gold framed pictures **on** the walls.

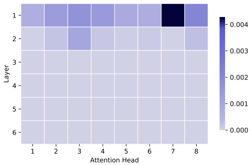
Experiment I: results



P on `DET,ADJ,NOUN`



P on `VERB,ADP`



Normalised entropy

- There is a clear separation between **when** different types of words are attended.
- The multi-modal text decoder captures both syntactic **and** semantic dependencies.
- Ilinykh and Dobnik (2021) compares multi-modal decoder with uni-modal decoder; uni-modal decoder captures more syntactic sequential knowledge whereas multi-modal decoder also captures some semantic despendencies.

Experiment II: cross-modal attention on objects

GÖTEBORGS UNIVERSITET

Question:

Is *grounding of noun phrases* expressed in the cross-modal attention to object bounding boxes?

Conditions:

Restrict **T** to word spans of noun phrases $(t_i, \ldots, t_j)$.
Restrict **S** to the ground truth objects v_n that the noun phrases depict.

Experiment II: cross-modal attention on objects

GÖTEBORGS UNIVERSITET

Question:

Is *grounding of noun phrases* expressed in the cross-modal attention to object bounding boxes?

Conditions:

Restrict **T** to word spans of noun phrases $(t_i, \ldots, t_j)$.
Restrict **S** to the ground truth objects v_n that the noun phrases depict.

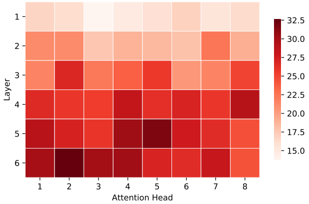
How do we achieve object-NP correspondence?

We link each noun phrase with object label(s) by searching for the most similar label comparing their embeddings with cosine similarity. We text different methods:

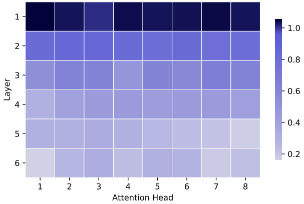
Combination Method	*Measure*	*mAP@K*	*Acc*
GloVe Multiply (Mitchell and Lapata, 2008)	cos	0.095	13.78
GloVe Add (Mitchell and Lapata, 2008)	cos	0.276	41.84
BERTScore (Zhang et al., 2020)	F_1	0.232	41.84
Sentence Transformer (Reimers and Gurevych, 2019)	cos	**0.313**	**44.39**

CLASP centre for linguistic theory and studies in probability

Experiment II: results



P on objects



Normalised entropy

- Visual grounding of noun phrases into objects happens in deeper layers.
- Low entropy in deeper layers indicates that the model becomes highly focused.

Experiment III: attention on relations (words or objects)

GÖTEBORGS UNIVERSITET

Question:
What are the differences in patterns between cross-modal attention
and attention on text in the decoder when generating descriptions of spatial relations?

Conditions:
Restrict **T** a token t_i that is either a target, relation or landmark.
Restrict **S** to the ground truth objects $v_1, \ldots, v_N$
which depict the related objects (landmark, target) (cross-modal).
Restrict **S** to previously generated token(s) t_{i-n}
which are parts of the specific description of a spatial relation (text only).

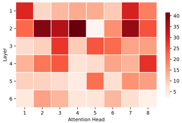
Example, text only:

There is **a green and maroon rug on** the floor.

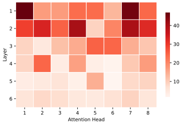
CLASP centre for linguistic theory and studies in probability

Experiment III: results, text only:

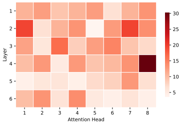
GÖTEBORGS UNIVERSITET



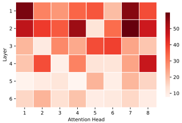
rel $\rightarrow$ *target*



land $\rightarrow$ *rel*



land $\rightarrow$ *target*

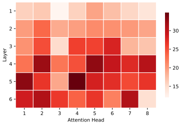


land $\rightarrow$ *rel* $+$ *target*

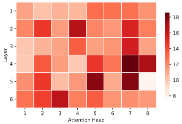
- The *sequential generation bias* is dominant: target is important for relation, relation is important for landmark.
- The model learns general semantic common-sense or functional knowledge about possible combinations between targets, relations and landmarks (in this order), e.g. “cup on table” vs “cup close to a table”.

Experiment III: results, cross-modal:

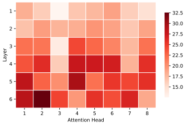
GÖTEBORGS UNIVERSITET



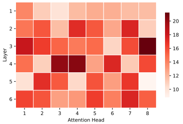
rel → *target*



rel → *land*



land → *target*



target → *land*

- Attention is **not aligned** with sequential nature of the task: this description refers to how we think the semantics works, not necessarily the model. As the lineraity is broken we get the pattern we see in the graphs above.
- The model learns semantic differences between different types of words in relation and attends to them differently.

Conclusions

GÖTEBORGS UNIVERSITET

- Attention in vision and language models captures more diverse linguistic knowledge, both **syntactic** and **semantic** which is **not linearly aligned**.
- Various factors affect how the model learns to attend across words and objects:
  - Bias of the task: auto-regressive nature of the generation task biases the attention to left-right direction.
  - Bias of the model: attention on text captures not only syntactic information, but also semantic one, coming from vision.
- Future work on grounding relations could explore the effects of different feature representations (common sense knowledge, objects' affordances) and the effect of the model architecture that could be built around them.

CLASP centre for linguistic theory and studies in probability

References I

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Mehdi Ghanimifard and Simon Dobnik. 2019. What goes into a word: generating image descriptions with top-down spatial knowledge. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 540–551, Tokyo, Japan. Association for Computational Linguistics.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image captioning: Transforming objects into words. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Nikolai Ilinykh and Simon Dobnik. 2021. How vision affects language: Comparing masked self-attention in uni-modal and multi-modal transformer. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 45–55, Groningen, Netherlands (Online). Association for Computational Linguistics.

References II

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Tell me more: A dataset of visual scene description sequences. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 152–157, Tokyo, Japan. Association for Computational Linguistics.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.