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

Nikolai Ilinykh and Simon Dobnik

CLASP, Department of Philosophy, Linguistics and Theory of Science,

University of Gothenburg, Sweden

`name.surname@gu.se`

UNIVERSITY OF GOTHENBURG

Why?

- Two types of words in image descriptions:

(1) object descriptions,

(2) spatial relations.

- Spatial relations are harder to ground.
- **Key challenge**: find an appropriate ratio between visual and linguistic information to generate each word type.

Questions

- **Q1**: What are the differences between **attention on text or objects** when generating relations?
- **Q2**: What knowledge is captured by **attention on text** in a multi-modal set-up?
- Q3: Does cross-modal **attention on objects** learn visual grounding of words?
  Answer: yes, check the paper for details.

How?

Dataset, task, models

- Image Description Sequence dataset.
- Generate paragraphs that describe images.
- Train a multi-modal transformer:
  $\mathbf{p}\,(t_i \mid \mathbf{V}; \mathbf{G}; (t_1, \dots, t_{i-1}); \theta)$.

1. There is a black and white fireplace on the left side of **the image**.
2. There is a green and maroon rug on **the floor**.
3. There are two gold framed pictures on **the walls**.
4. There are two clear flower vases on **the mantle**.
5. There is wooden chair and table in the middle of **the room**.

Attention proportion

Extract attention weights from **cross-modal attention** and **masked self-attention**:

$$\alpha_{\ell,h}(t_i \mid t_1, \dots, t_{i-1}) = \text{softmax}\left(\frac{Q_{MSA/CA} K^T_{MSA/CA}}{\sqrt{d_k}}\right), \qquad (1)$$

Compute **attention proportion** P:
the amount of attention on parts of the input (**S**) for specific parts of the output (**T**).

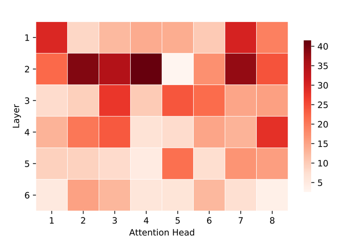
Experiments

Q1 : Attention on relations

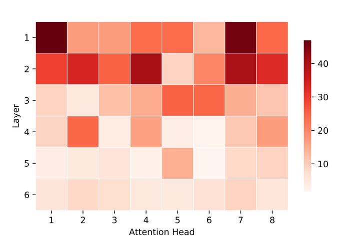
Conditions:

Restrict **T** to a token t_i of spatial relations.
Restrict **S** to ground truth objects $v_1, \dots, v_N$ (landmark, target) (cross-modal).
Restrict **S** to the previous token(s) t_{i-n} (text only).

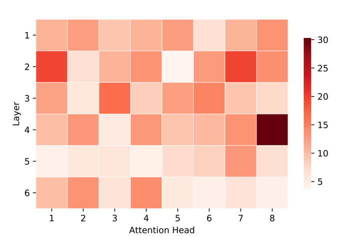
Results, text only:



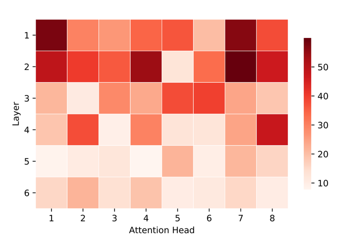
rel → *tar*



land → *rel*

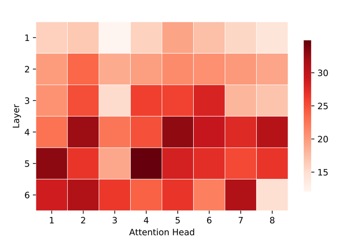


land → *tar*

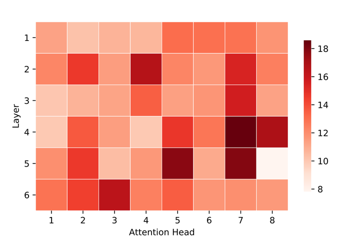


land → *rel* + *tar*

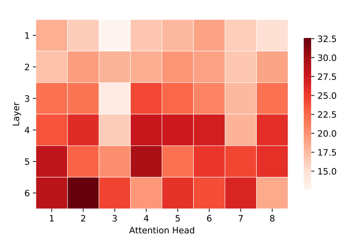
Results, cross-modal only:



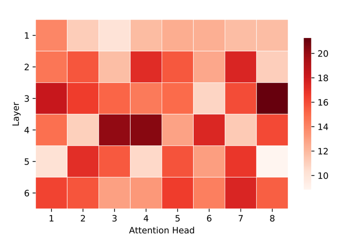
rel → *tar*



rel → *land*



land → *tar*



tar → *land*

Conclusions:

Text only:

- Target -> Relation -> Landmark, e.g. left-to-right "auto-regressive" attention pattern.
- Targets are important for relations, relations are important for targets.
- **Depth-wise**, the model learns relations and then exploits this knowledge to learn targets and landmarks.

Cross-modal only:

- Landmark -> Relation -> Target.
- Attention is **not aligned** with the sequential nature of the task.
- Landmarks are attended universally across many layers, including the surface ones.
- Targets are attended in deeper layers.

Q2 : Attention on words

Conditions:

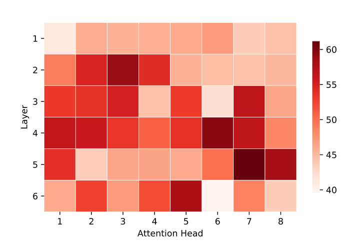
Restrict **T** to a token t_i.
Restrict **S** to previously generated token(s) t_{i-n} with either ⟨`DET, ADJ, NOUN`⟩ or ⟨`VERB, ADP`⟩ POS tags.

Examples:

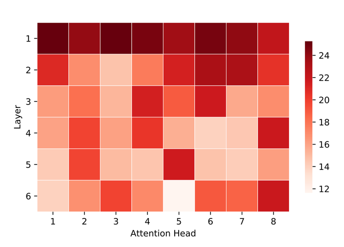
(1) There are **framed pictures** on the walls.

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

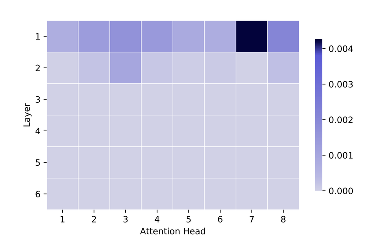
Results:



P on DET, ADJ, NOUN



P on VERB, ADP



Entropy on text

- Attention on text in a multi-modal set-up captures semantic differences between word types (object descriptions and spatial relations).

Summary

- Uni-modal and multi-modal components of the architecture capture **different alignments** of spatial relations.
- Linguistic representations capture semantic differences between word types **in the context of an image description task**.
- There is an impact of (i) the task and (ii) the structure of the model on what is learned.

Future work

- Other explainability methods (e.g., probing).
- Different feature representations: geometry, common sense knowledge, affordances.
- Check out our paper, code and data:

CLASP centre for linguistic theory and studies in probability