# On the Interplay between Language and Vision in Transformers: How Much of a "Multi-Modal Learning" Do We Observe?

Nikolai Ilinykh
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

CLASP seminar, February 2, 2022

# Outline

# Introduction

GÖTEBORGS UNIVERSITET

What type of research questions do interpretability studies focus on?

- **NLP**: identification of linguistic phenomena captured in self-attention layers, attention heads and neurons (Raganato and Tiedemann, 2018; Belinkov and Glass, 2019; Vig and Belinkov, 2019; Voita et al., 2019; Rogers et al., 2020)
- **CV**: learning semantic information from pixels and image patches by tracking attention across layers (Dosovitskiy et al., 2021; Caron et al., 2021)
- **Language-and-Vision**:

  Cao et al. (2020) probe pre-trained architectures for different vision-and-language tasks

We need a better understanding of how structures within multi-modal transformers can be linked with findings from cognitive science (for "true-er" interpretability). In addition, we need more information about (i) inductive biases in transformers, (ii) the effects of features, tasks and model's architecture on what is learned by the model.

# Central Research Questions

GÖTEBORGS UNIVERSITET

1. What kind of knowledge is captured by different layers in image captioning transformer?
2. Is there a progression of attended representations from low-level local relations to high-level global dependencies? How does this relate to the hierarchical structure of language and perception observed in humans (Tenenbaum et al., 2011)?
3. Does language affect vision, e.g. is conceptual linguistic knowledge implicitly reflected in visual representations?

CLASP centre for linguistic theory and studies in probability

# Background: Model Architecture

GÖTEBORGS UNIVERSITET

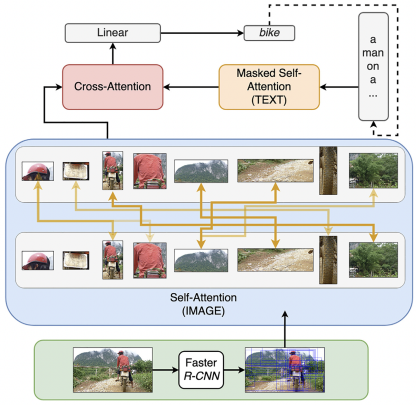



Figure: The model takes objects extracted with F-RCNN and sends them to the self-attention block; similarly, text is sent to its respective attention block. Later outputs from both modality-specific self-attentions are fused and processed by cross-attention. Once the next word is predicted, it is added to the textual input to keep generating until the END token is produced.

# Background: Incorporating Geometry



$$\lambda(m,n) = \left(\log\left(\frac{|x_m - x_n|}{w_m}\right), \log\left(\frac{|y_m - y_n|}{h_m}\right), \log\left(\frac{w_n}{w_m}\right), \log\left(\frac{h_n}{h_m}\right)\right)$$

$$\mathbf{E} = \mathrm{Emb}(\lambda)$$

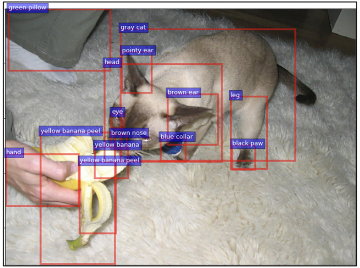
$$\mathbf{\Omega}^G = \mathbf{E}\mathbf{W}_g$$

$$\mathbf{\Omega}^V = \frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}$$

$$\mathbf{\Omega} = \log(\mathbf{\Omega}^{\mathbf{G}}) + \mathbf{\Omega}^{\mathbf{V}}$$

$$head_{h,\ell}(\mathbf{F}) = \mathrm{softmax}(\mathbf{\Omega})\mathbf{V}$$

# Background: Learning from Object Detections




The model is expected to benefit from the conceptual knowledge of the objects, as provided by the object extractor, pre-trained on human annotations of visual scenes.

# Experiment I: Thematic Relatedness

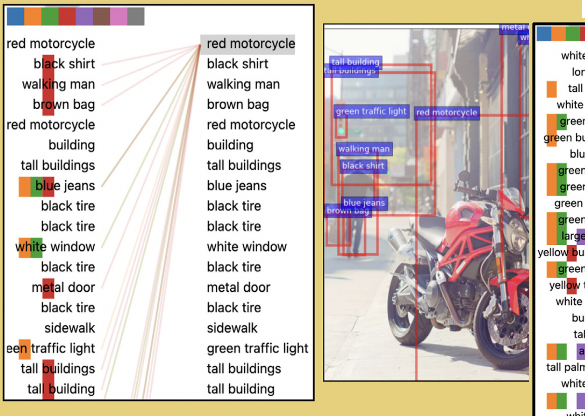
GÖTEBORGS UNIVERSITET

Does our model learn to attend between thematically related objects?
We inspect attention against ground-truth object clusters.

# Analysis I: Thematic Relatedness



Does our model learn to attend between thematically related objects?
We inspect attention against ground-truth object clusters.




Layer VI


Layer VI

# Analysis I: Result

GÖTEBORGS UNIVERSITET

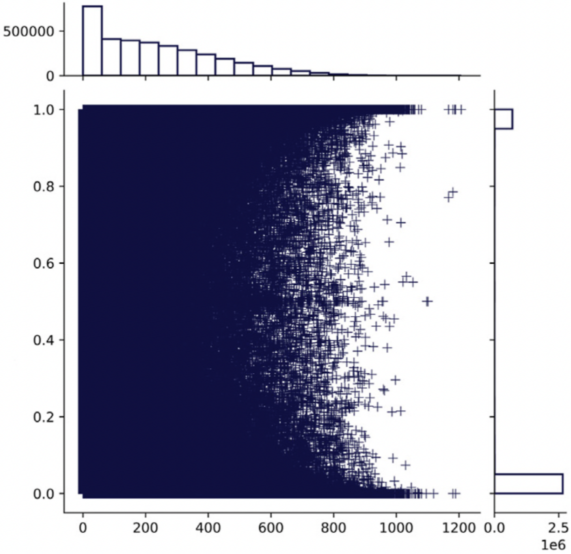
Earlier layers encode visual properties within thematic categories, whereas deeper layers look beyond automatically identified thematic categories.
Why?

Cosine similarity on objects in different clusters: *0.50* for the same cluster, *0.31* for two different clusters. Our model learns that semantic similarity entails visual similarity, as observed in both humans (Rosch, 1975) and machines (Deselaers and Ferrari, 2011).

Deselaers and Ferrari (2011): distant elements are semantically less similar.
**Q**: Does our model associate distances between elements with their semantic similarity? Is it able to infer that parts of the objects have to be geometrically close to each other?

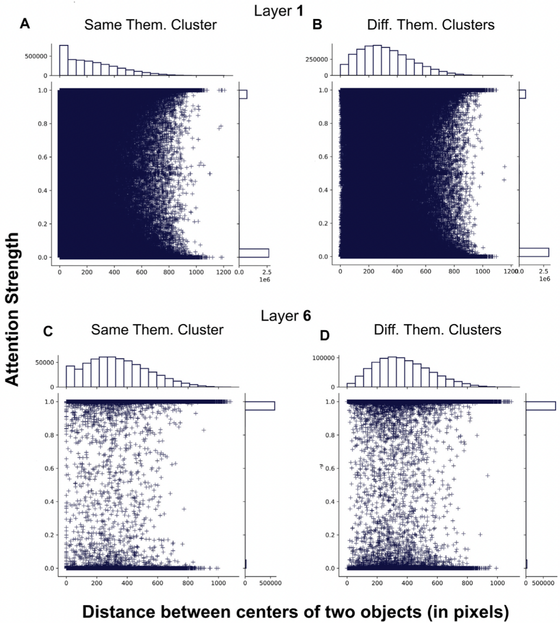
# Analysis II: Visual Proximity and Semantic Similarity

GÖTEBORGS UNIVERSITET




CLASP centre for linguistic theory and studies in probability

# Analysis II: Visual Proximity and Semantic Similarity



# Analysis II: Result

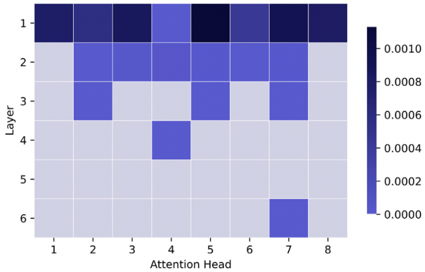


- In general, earlier layers and deeper layers capture different sources of information.
- Local relations between semantically similar objects are captured in the first layer, less so in the last layer. These attention weights differ in terms of their strength.
- The model seems to understand that if two objects are close to each other, they must be semantically similar, e.g. part-whole relations.
- At the same time, the model builds on local knowledge and learns to attend between more distant objects, perhaps, *whole* objects, not their parts.

**Q**: How does attention entropy change between layers? Why would deeper layers be much more confident (high attention)?

# Analysis III: Attention Entropy

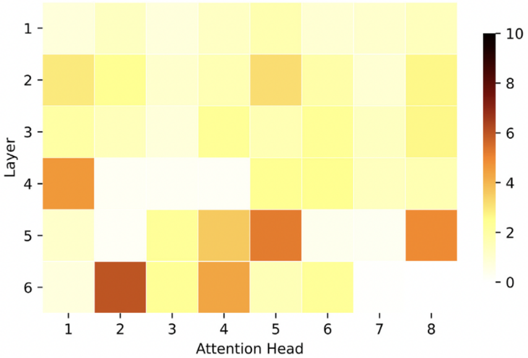
GÖTEBORGS UNIVERSITET




1. attention converges to fewer objects in deeper layers
2. disperse and dissimilar in earlier layers vs. focused and concentrated in deeper layers

**Q**: What are the reasons for lower attention entropy in deeper layers?

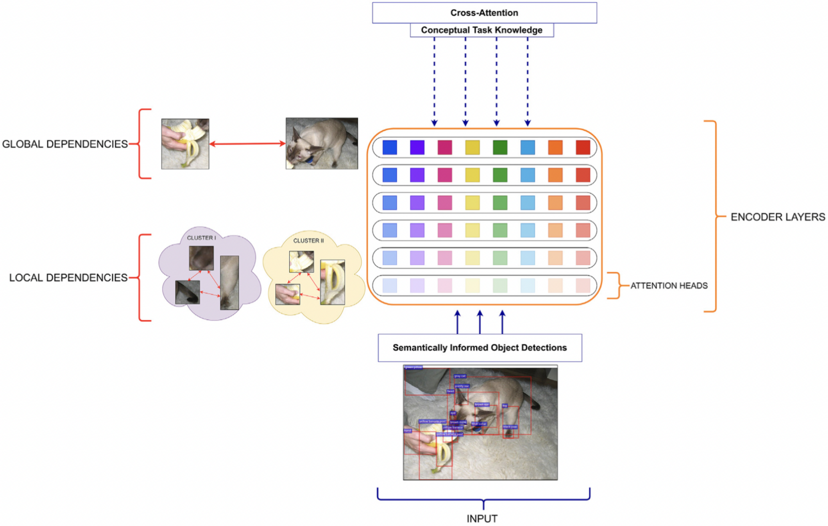
# Analysis IV: Visual Grounding




**Result**: The pragmatic nature of the task (knowledge from language) affects learning in deeper layers since they often learn to mirror pairings of objects and noun phrases that correspond to them.
**Follow-Up Experiment**: Looking for mappings between nouns and objects that would correspond to a different, hypothetical description.

# Results of this Study





# Resembling Mammal Cognition

Hierarchical processing of visual information has been observed in mammals:

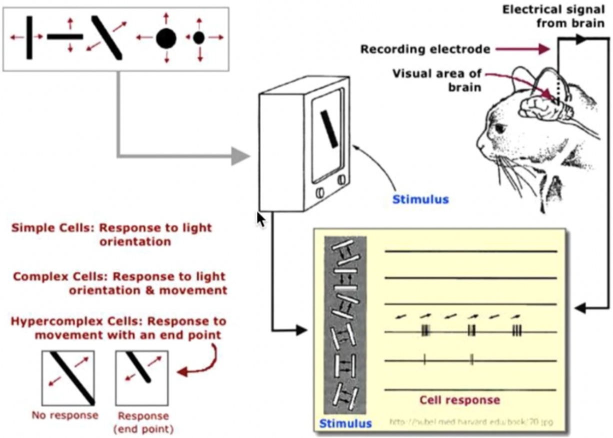



Figure: (Hubel and Wiesel, 1959): deeper cells build on simpler cells to process complex patterns

# Resembling Human Cognition I

Ullman (1984): humans process visual information in sequential order, starting from simpler relresentations and applying task-dependent rules at a later point.

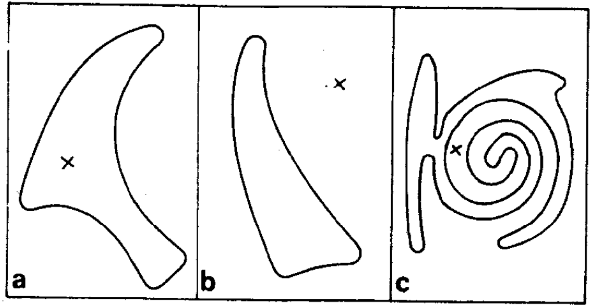
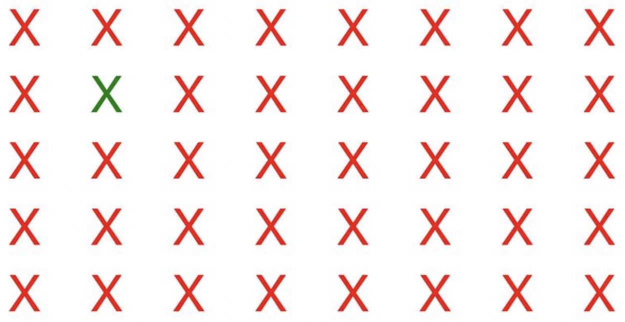


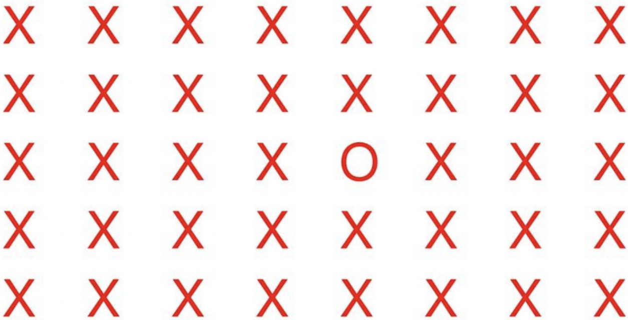

Figure: For the first two images, we apply *base routines*, while for the third image we use *visual routines* due to the complexity of the task.

# Find the green letter:

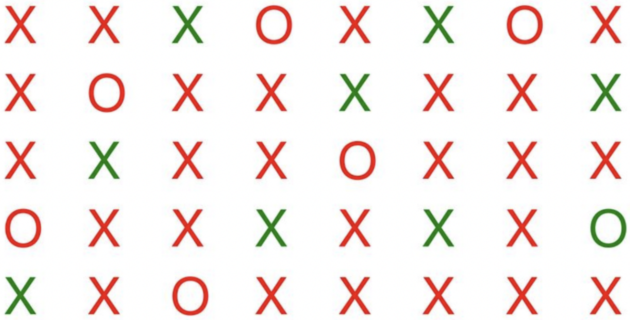


# Find the O:



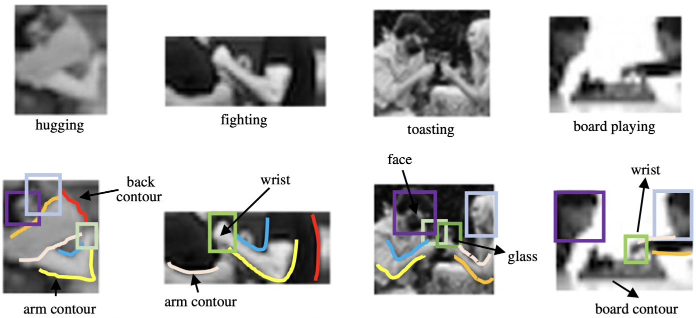


# Find the green O:



# Resembling Human Cognition II

Ben-Yosef and Ullman (2018): humans differentiate between elements 'below' objects and 'after'. This ability allows us to frequently bring complex meaning to parts of the objects.




Note: Context principle by Frege ('it is enough if the sentence as whole has meaning; thereby also its parts obtain their meanings') can be also observed in human vision and computational vision (Geman et al., 2002).

# What did we learn?



- visual representations in image captioning transformer are structured hierarchically
- the observed hierarchy and structure can be linked with human way of processing visual information: build more complex representations 'beyond' objects based on the knowledge 'before' objects
- important factors that shape multi-modal learning: task pragmatics and model's architecture

# References I


Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Guy Ben-Yosef and Shimon Ullman. 2018. Image interpretation above and below the object level. *Interface Focus*, 8.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *Computer Vision – ECCV 2020*, pages 565–580, Cham. Springer International Publishing.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660.

Thomas Deselaers and Vittorio Ferrari. 2011. Visual and semantic similarity in imagenet. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 1777–1784. IEEE Computer Society.

# References II


Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Stuart Geman, Daniel Potter, and Zhiyi Chi. 2002. Composition systems. *Quarterly of Applied Mathematics*, 60.

David H. Hubel and Torsten N. Wiesel. 1959. Receptive fields of single neurons in the cat's striate cortex. *Journal of Physiology*, 148:574–591.

Alessandro Raganato and Jörg Tiedemann. 2018. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Brussels, Belgium. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Eleanor Rosch. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104:192–233.

# References III




Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285.

Shimon Ullman. 1984. Visual routines. *Cognition*, 18(1-3):97–159.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.