Modelling Multimodal Integration in Human Concept Processing with Vision-Language Models

Anna Bavaresco (a.bavaresco@uva.nl)

Marianne de Heer Kloots (m.l.s.deheerkloots@uva.nl)

Sandro Pezzelle (s.pezzelle@uva.nl)

Raquel Fernández (raquel.fernandez@uva.nl)

University of Amsterdam
Institute for Logic, Language and Computation
Science Park 107, 1098 XG Amsterdam, The Netherlands

Abstract

Text representations from language models have proven remarkably predictive of human neural activity involved in language processing. However, the word representations learnt by language-only models may be limited in that they lack sensory information from other modalities. Here, we leverage recent AI advancements in multimodal modelling to investigate whether current pre-trained vision-language models (VLMs) yield concept representations that are more aligned with human brain activity than those obtained by models trained with language-only input. Our results reveal that VLM representations correlate more strongly than those by language-only models with activations in brain areas functionally related to language processing. Altogether, our study indicates that vision-language integration better captures the nature of human concepts.

Keywords: Deep neural networks; fMRI; vision and language processing; representational similarity analysis

Introduction

How does the brain represent semantic knowledge? Whereas many computational models of language still build on the idea that meaning can be extracted from text corpora (Piantadosi & Hill, 2022), increasing evidence from cognitive science and neuroscience suggests that human semantic representations are in fact grounded in sensory experiences (Louwerse, 2011; Barsalou, 1999; Harnad, 1990; Bergen, 2012).

Here, we investigate the ability of vision-language models (VLMs) implemented as deep neural networks to capture multimodal aspects of human semantic processing. From the perspective of technical applications, there is no doubt that VLMs can perform tasks, such as visual question answering and image captioning, that are simply impossible for language-only models. This opens the intriguing question of whether VLMs also learn text representations that model human language processing more accurately. More concretely, we investigate the following key research question: *Are representations in current pretrained VLMs better models of brain activity involved in concept word processing than those in text-only language models?* Human concept representations seem to reflect knowledge from different modalities (Dirani & Pylkkänen, 2024). Since VLMs are trained to align input from visual and linguistic streams, we hypothesise they will exhibit an advantage over text-only language models in modelling brain activity during concept processing. In the following, we summarise our study and refer the reader to Bavaresco, de Heer Kloots, Pezzelle, and Fernández (2024) for more details.

Methods

Data To study the alignment between (V)LMs and brain responses, we focus on a publicly available fMRI dataset containing neural responses to concept words (Experiment 1 in Pereira et al., 2018). We consider two experimental conditions: (1) a language-only *sentence condition* where each word appears boldfaced in the context of a sentence that makes the relevant concept salient; participants see six sentences, one at a time; (2) a multimodal *picture condition* where each word is presented together with an image illustrating the relevant concept; again, for each concept word, participants are shown six different images, one at a time. 16 participants were scanned while viewing the words in these two conditions. Voxel activations for each participant are averaged across the six presentations of the same word per condition. We focus on two functionally localised brain networks: the *Language network* (Fedorenko, Behr, & Kanwisher, 2011), reporting results separately for the left hemisphere (LH) and the right hemisphere (RH), and the *Visual network* (Power et al., 2011; Buckner, Andrews-Hanna, & Schacter, 2008).

Models We employ two main types of deep neural network models: a set of VLMs trained on related visual and textual input and a set of language-only models trained exclusively on text. We focus on models widely applied for Natural Language Processing applications and use them off-the-shelf as pre-trained by their developers.

We test three families of VLMs corresponding to different vision-language integration strategies: (1) *Contrastive VLMs*: CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021); (2) *Vision-language encoders*: VisualBERT (Li, Yatskar, Yin, Hsieh, & Chang, 2019) and LXMERT (Tan & Bansal, 2019); (3) *Generative VLMs*: IDEFICS2 (Laurençon, Tronchon, Cord, & Sanh, 2024) and LLaVA NeXT (Liu et al., 2024).

Regarding the language-only models, we include architectures that either provide informative baselines or useful comparisons with specific VLMs. Concretely, we experiment with one encoder-only language model —BERT (Devlin, Chang, Lee, & Toutanova, 2019), which underpins the language components of ALIGN, VisualBERT, and LXMERT— and two generative large language models — Mistral (Jiang et al., 2023), which is the language model used in IDEFICS2, and Llama3 (Meta, 2024), used in LLaVA NeXT. In addition, we include a simpler distributional semantic model —GloVe (Pennington, Socher, & Manning, 2014)— for reference.

Procedure & Evaluation To extract model representations that we can compare to the fMRI responses, we feed the models with the same stimuli presented to the participants. In the sentence condition, we input the sentences read by participants. In the picture condition, both the target word and each image accompanying it are fed to the VLMs, and only the target word is fed to the language-only models. In both conditions, we average the representations extracted from the 6 sentences or the 6 images per word, repeating this procedure for each model layer whenever possible.

We quantify model alignment to neural activity using Representational Similarity Analysis (RSA; Kriegeskorte, Mur, and Bandettini, 2008), based on cosine-distance Representational Dissimilarity Matrices (RDMs) for both the model and brain representations. We average across participants on the fMRI

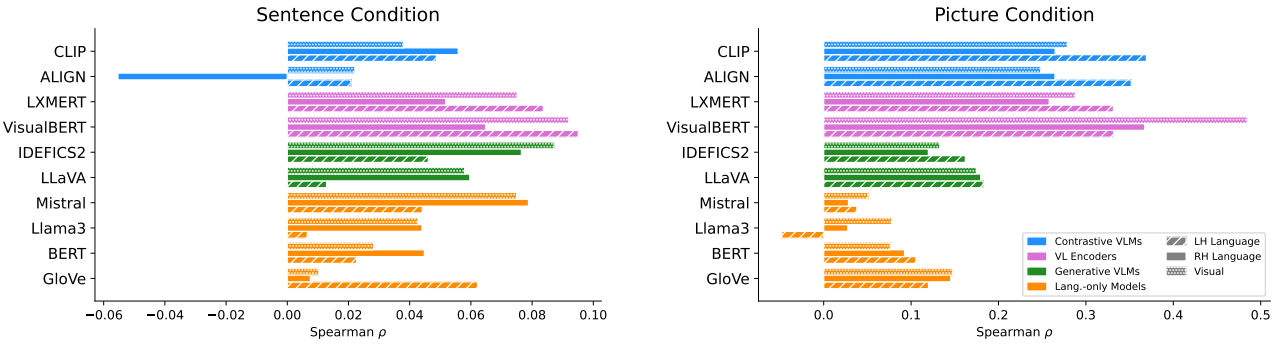

Figure 1: Representational Similarity Analysis results for the sentence condition (left) and the picture condition (right). Correlations are reported for the best layers, which differ across brain networks.

side and compute RDMs separately for each layer on the model side. Here, we only report results from the best layer. Alignment is measured by the Spearman correlation between RDMs. In both the main experiment and ablation studies, we conduct appropriate statistical tests to verify that correlation differences between models are significant.

Main Results

Overall, our results in the sentence condition (Figure 1, left) indicate that the VLMs IDEFICS2 and VisualBERT tend to exhibit the strongest brain alignment across all brain areas. Considering VLM families, VL encoders are significantly superior to contrastive VLMs in the LH —but not the RH— language network and in the visual one, and to generative VLMs in the LH language network. Lastly, VL encoders consistently outperform their language-only counterpart (BERT).

In the picture condition (Figure 1, right), we observe higher correlations than in the sentence condition, which can be attributed to higher signal-to-noise ratio (i.e., higher inter-participant similarities) in the fMRI responses. VLMs are significantly more brain-aligned than their unimodal counterparts across all brain networks in this condition.

Ablation Studies

To complement the findings provided by RSA, we conduct two ablation analyses aimed at answering the following questions.

1) How much of the VLMs' brain alignment in the sentence condition can be attributed to semantic information already present in their language encoder *prior to* any multimodal training? To investigate this question, we conduct a partial correlation analysis aimed at removing from VLMs' representational spaces the information shared with LLMs' representational spaces. We highlight two findings: First, in the LH language network, all differences between partial and initial correlations, except for IDEFICS2, are *not* statistically significant, suggesting that the brain alignment achieved by these models is mainly attributable to semantic information acquired during multimodal pretraining. Second, results from the visual network reveal that, for generative VLMs, the differences between initial and partial correlations are statistically significant, while, for VL encoders, they are not. This suggests that part of the information relevant for alignment with visual brain responses was already present in Mistral and Llama3 before any vision-language training.

2) To what extent is the VLMs' advantage over their language-only counterparts in the picture condition driven by the input images at inference time? We address this issue with an ablation study where we pass the same input (concepts without pictures) to both VLMs and language-only models. We find that, despite changes in model rankings, the most brain-aligned architectures in all brain networks remain multimodal. More specifically, LXMERT is statistically significantly more brain-aligned than other models across all three networks, and VisualBERT statistically significantly outperforms all language-only models in the LH language network and in the visual one. In summary, while some architectures rely heavily on the input images, others yield strong brain correlations even without meaningful visual input.

Conclusion

Our study advances our understanding of the brain-relevant conceptual information learnt by multimodal and unimodal models and makes the following contributions: 1) It provides a broad investigation of the brain alignment achieved by multiple recent pretrained vision-language models from different model families; 2) It shows evidence that the highest brain alignment is consistently achieved by one of the VLMs (and not a language-only model), although not the same architecture across all conditions and brain networks; 3) It reveals that vision-language encoders tend to exhibit higher brain alignment than the more recent generative VLMs; 4) It demonstrates that the superior brain alignment achieved by vision-language encoders stems from learning novel multimodal semantic information.

Acknowledgments

Anna Bavaresco and Raquel Fernández are funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455). Marianne de Heer Kloots is funded by the Netherlands Organization for Scientific Research (NWO), through a Gravitation Grant 024.001.006 to the Language in Interaction Consortium.

References

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and brain sciences*, *22*(4), 577–660.

Bavaresco, A., de Heer Kloots, M., Pezzelle, S., & Fernández, R. (2024). Modelling multimodal integration in human concept processing with vision-and-language models. *arXiv preprint arXiv:2407.17914*.

Bergen, B. K. (2012). *Louder than words: The new science of how the mind makes meaning*. Basic Books.

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: anatomy, function, and relevance to disease. *Annals of the new York Academy of Sciences*, *1124*(1), 1–38.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers).* Minneapolis, Minnesota: Association for Computational Linguistics.

Dirani, J., & Pylkkänen, L. (2024, May). MEG evidence that modality-independent conceptual representations contain semantic and visual features. *Journal of Neuroscience*. doi: 10.1523/JNEUROSCI.0326-24.2024

Fedorenko, E., Behr, M. K., & Kanwisher, N. (2011). Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, *108*(39), 16428–16433.

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

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., ... Duerig, T. (2021). Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *International conference on machine learning* (pp. 4904–4916).

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., ... others (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 249.

Laurençon, H., Tronchon, L., Cord, M., & Sanh, V. (2024). What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., & Chang, K.-W. (2019). VisualBERT: A simple and Performant Baseline For Vision and Language. *arXiv preprint arXiv:1908.03557*.

Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., & Lee, Y. J. (2024, January). *LLaVA-NeXT: improved reasoning, OCR, and world knowledge.* Retrieved from https://llava-vl.github.io/blog/2024-01-30-llava-next/

Louwerse, M. M. (2011). Symbol Interdependency in Symbolic and Embodied Cognition. *Topics in Cognitive Science*, *3*(2), 273–302. doi: 10.1111/j.1756-8765.2010.01106.x

Meta. (2024). *Llama3 model card.* Retrieved from https://ai.meta.com/blog/meta-llama-3/

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., ... Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, *9*(1), 963.

Piantadosi, S., & Hill, F. (2022, October). Meaning Without Reference in Large Language Models.. Retrieved 2024-01-01, from https://openreview.net/forum?id=nRkJEwmZnM

Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., ... others (2011). Functional network organization of the human brain. *Neuron*, *72*(4), 665–678.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning Transferable Visual Models From Natural Language Supervision. In *International conference on machine learning* (pp. 8748–8763).

Tan, H., & Bansal, M. (2019). LXMERT: Learning Cross-Modality Encoder Representations from Transformers. 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).*