Compositional Meaning in Vision-Language Models and the Brain

Maithe van Noort (maithe.van.noort@student.uva.nl)
Amsterdam Brain and Cognition, University of Amsterdam
Nieuwe Achtergracht 129-B, 1018 WT Amsterdam, The Netherlands

Luke Korthals (l.korthals@uva.nl)
Amsterdam Brain and Cognition, University of Amsterdam
Nieuwe Achtergracht 129-B, 1018 WT Amsterdam, The Netherlands

Giacomo Aldegheri (giacomo.aldegheri@gmail.com)
Amsterdam Brain and Cognition, University of Amsterdam
Nieuwe Achtergracht 129-B, 1018 WT Amsterdam, The Netherlands;
Experimental Psychology, Justus Liebig University, Otto-Behagel-Str. 10
Giessen, Hessen 35394, Germany;
Center for Mind, Brain and Behavior (CMBB), Hans-Meerwein-Str. 6
Marburg, Hessen 35032, Germany

Marianne de Heer Kloots (m.l.s.deheerkloots@uva.nl)
Institute for Logic, Language, & Computation, University of Amsterdam
Science Park 107, 1098 XG Amsterdam, The Netherlands

Micha Heilbron (m.heilbron@uva.nl)
Amsterdam Brain and Cognition, University of Amsterdam
Nieuwe Achtergracht 129-B, 1018 WT Amsterdam, The Netherlands

Abstract

What is the role of compositional structure in the alignment of visual and linguistic brain areas to computational semantic embeddings? Vision-language models (VLMs) have shown meaningful alignment to the brain in their representations of semantic structure, for both images and text. However, the extent to which these representations capture *compositional structure* – i.e. changes in meaning based on changes to the combinatorial structure of parts – remains uncertain. Here we leverage Winoground, a dataset designed to test compositionality in multimodal representations, to compare the compositional structure captured by different model embeddings, as well as fMRI responses collected as part of a larger study on multi-modal meaning (with 2760 image and 2760 semantically equivalent language trials). In contrast to VLM embeddings, neural representations in the brain show a striking absence of compositional processing (chance level performance) when evaluated on the Winoground benchmark – despite robust semantic encoding of individual concepts as measured by voxel activity predictions. This is intriguing as distinctions between stimuli in Winoground are trivial to any English-speaking human, highlighting the challenge of identifying the substrates of compositional processing in the brain. Our targeted dataset and evaluation pipeline lay the foundation for systematic, cross-modal evaluations of compositionality in both artificial and biological neural representations.

Keywords: vision-language models; compositionality; fMRI

Introduction

Multimodal semantic embeddings have shown strong performance at vision-language tasks, including image-caption matching, zero-shot classification, and visual question answering (Radford et al., 2021; Li, Li, Xiong, & Hoi, 2022; Singh et al., 2022). Moreover, semantic embeddings can also predict neural responses in high-level visual and language areas of the brain with unprecedented accuracy (Wang, Kay, Naselaris, Tarr, & Wehbe, 2023; Doerig et al., 2022). This has led some to suggest a meaningful degree of alignment between the semantic structure in these vectors and semantic representations in the brain, in both visual and language areas.

However, besides similarity structure, linguistic meaning heavily depends on compositionality – the derivation of meaning as a function of individual elements and the way they are combined (Partee, 2004; Fodor & Lepore, 1992). In natural language processing, multiple benchmarks have been created to specifically evaluate the compositionality of multi-modal embeddings and vision-language models (VLMs) (Thrush et al., 2022; Yuksekgonul, Bianchi, Kalluri, Jurafsky, & Zou, 2022). These typically comprise image-caption pairs that consist of identical words yet differ in composition (e.g., 'a cup in some grass' versus 'some grass in a cup'; Fig. 1A). Generally, these benchmarks of compositional meaning have found stark performance gaps, where models that perform well on common semantic benchmark tasks (such as visual question answering or image-caption matching) often showed a strikingly low degree of compositionality (Parcalabescu et al., 2021; Yuksekgonul et al., 2022).

Here, we extend this logic to the domain of fMRI, evaluating the compositionality of multi-voxel semantic representations and semantic encoding models, in the context of a large-scale study of visual-linguistic meaning. In this study, participants are shown a large number of diverse images and, on separate trials, their corresponding captions (for a total of 2760 image and 2760 sentence trials). The majority of these trials were taken from Visual Genome Dataset (the images that are also in COCO (Krishna et al., 2017; Lin et al., 2014)) and are intended to serve as a diverse sampling of the semantic space. A subset of the trials were extracted from one particular benchmark of compositionality, Winoground (Thrush et al., 2022); these pairs contain the same lexical items but create different compositional meanings (Fig. 1A). This allows us to ask 4 questions: 1) How compositional are VLM embeddings, commonly used to model meaning in the brain? 2) How compositional are multi-voxel semantic representations? 3) Is compositional meaning, to the extent present in original embeddings, picked up by semantic encoding models, or do these models only rely on bag-of-words meaning? 4) Does the degree of semantic compositionality differ between semantic representations in different brain areas, e.g., visual versus language areas?

Methods

fMRI data This is part of a large-scale (low-N) fMRI study on visual and linguistic meaning. Over the course of at least 13 sessions, participants (2 participants fully collected) are presented at least 2760 image and 2760 sentence trials (4s each); for each image there is a corresponding 8-word semantically equivalent sentence. Data was collected using a 3T (Philips Achieva DS) fMRI scanner; preprocessing was performed using fMRIprep (Esteban et al., 2019) and single-trial fMRI responses were estimated using GLMSingle (Prince et al., 2022).

Models We compare 4 VLMs: CLIP, a Transformer-based contrastive model trained to map images and corresponding captions to a joint space (Radford et al., 2021); NegCLIP, an extension of CLIP fine-tuned with hard negatives to improve its compositional capacity (Yuksekgonul et al., 2022); SigLIP an optimised version of CLIP using a sigmoid loss (Zhai, Mustafa, Kolesnikov, & Beyer, 2023); and FLAVA (Singh et al., 2022) a VLM trained on various multi- and unimodal tasks, including masked language modelling.

Embedding extraction & evaluation To extract **model representations** for each image and caption, we take each model's final layer embedding. By design, these embeddings

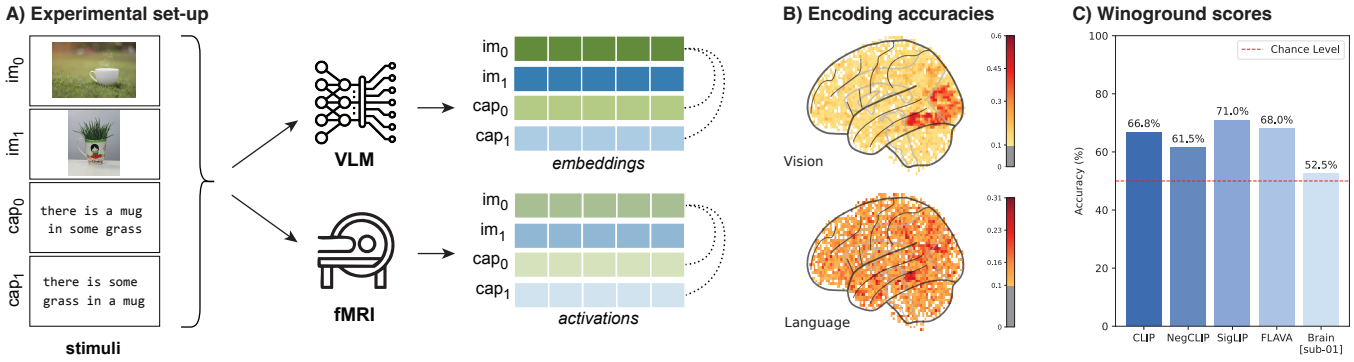

Figure 1: **A)** Two image-caption pairs, close in form yet different in meaning, get turned into vector representations through both a VLM (embeddings) and fMRI (brain activations). Matching and mismatching image-caption pairs are then compared in similarity (e.g., im0-cap0 vs. im1-cap1). **B)** Encoding model performance (cross-validated correlation coefficient) for vision and language trials of subject-01. **C)** Bar chart of Winoground accuracy scores across various models and the brain.

exist in a joint image-text space. Therefore, we can evaluate their compositionality by testing whether cosine similarities between matching image-text pairs are greater than between their non-matching counterparts, for each set of visio-linguistic minimal pairs in the Winoground benchmark (Fig. 1A). To apply the same evaluation pipeline to **brain representations**, we first learn a linear transformation to map fMRI activity recorded from visual and language areas to a joint space. In this joint space, we can apply an identical procedure to evaluate the compositionality of activation vectors.

Hence, for each vector representation and each set of Winoground pairs (e.g. the two images and captions in Fig. 1A), we obtain 4 binary accuracy values: accuracy at matching the correct caption for each image, and the correct image for each caption.

Results & Discussion

We first assessed whether we could replicate the sensitivity to higher-level (semantic) features. To this end, we fit voxel-wise encoding models, predicting fMRI activations based on final-layer CLIP embeddings. This revealed strong encoding performance, in both image and language trials (Fig. 1B).

Having established data quality and replicated the sensitivity to semantic features in high-level visual and language areas, we then evaluated the compositionality of these embeddings and neural activation patterns. For the VLM embeddings, we observe above-chance accuracies for all models, but a few differences are worth highlighting. First, despite its fine-tuning tailored to improving compositional representations, NegCLIP does not outperform CLIP; instead, compositionality is improved by more general optimizations of the CLIP pre-training procedure (SigLIP), and somewhat by extending pre-training to multiple tasks (FLAVA).

We then applied the same test to fMRI activation patterns; deriving the joint-space fMRI activations from a linear mapping between IT and the language network. Strikingly, despite the apparently high sensitivity to semantics as measured from voxel-wise encoding models, we do not observe any evidence for compositional meaning in the joint fMRI embeddings, with chance-level performance on the Winoground test.

Since the Winoground evaluation is cognitively trivial for any English-speaking human (e.g. distinguishing 'some grass in a cup' from 'a cup in some grass'), this discrepancy provokes some interesting possibilities. First, it could be that the fMRI data itself does not adequately capture compositional structure and instead primarily reflects lexical-semantic content (Kauf, Tuckute, Levy, Andreas, & Fedorenko, 2024). Secondly, it could be that compositional structure is lost during the linear mapping we use to derive joint-space fMRI activations. We are currently implementing alternative representational transformations and control analyses to evaluate this possibility. Third, it could be that the exact ROIs we selected are primarily sensitive to the lexical semantics but not the compositional meaning; we will address this with a searchlight equivalent of the analysis to evaluate the degree of compositionality across the brain.

Together, we present a targeted dataset and evaluation for probing compositional structure in both artificial and neural representations of meaning. This resource, combining carefully selected minimal pairs with large-scale whole-brain fMRI recordings to semantically equivalent visual and linguistic stimuli, enables systematic evaluation of how different representational spaces capture the building blocks of meaning.

References

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2022). Visual representations in the human brain are aligned with large language models. *arXiv preprint arXiv:2209.11737*.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., ... others (2019). fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, *16*(1), 111–116.

Fodor, J. A., & Lepore, E. (1992). Holism: A shopper's guide. *Philosophical Books*, *33*(2), 65–68.

Kauf, C., Tuckute, G., Levy, R., Andreas, J., & Fedorenko, E. (2024). Lexical-semantic content, not syntactic structure, is the main contributor to ann-brain similarity of fmri responses in the language network. *Neurobiology of Language*, *5*(1), 7–42.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., ... others (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, *123*, 32–73.

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *Proceedings of the 39th International Conference on Machine Learning*, *162*, 12888–12900.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014* (pp. 740–755). Cham: Springer International Publishing. doi: 10.1007/978-3-319-10602-1$_48$

Parcalabescu, L., Cafagna, M., Muradjan, L., Frank, A., Calixto, I., & Gatt, A. (2021). Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*.

Partee, B. H. (2004). *Compositionality in formal semantics: Selected papers by Barbara Partee*. Oxford: Blackwell Publishing.

Prince, J. S., Charest, I., Kurzawski, J. W., Pyles, J. A., Tarr, M. J., & Kay, K. N. (2022). Improving the accuracy of single-trial fmri response estimates using glmsingle. *Elife*, *11*, e77599.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, *139*, 8748–8763.

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., & Kiela, D. (2022). Flava: A foundational language and vision alignment model. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 9635–9644).

Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., & Ross, C. (2022). Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 5238–5248).

Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J., & Wehbe, L. (2023). Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, *5*(12), 1415–1426.

Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., & Zou, J. (2022). When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*.

Zhai, X., Mustafa, B., Kolesnikov, A., & Beyer, L. (2023, October). Sigmoid Loss for Language Image Pre-Training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 11941–11952). Paris, France: IEEE. doi: 10.1109/ICCV51070.2023.01100