Multimodal Human Perception of Object Dimensions: Evidence from Deep Neural Networks And Large Language Models

Florian Burger (F.Burger@westernsydney.edu.au)
The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Sydney, Australia

Manuel Varlet (M.Varlet@westernsydney.edu.au)
The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Sydney, Australia
School of Psychology, Western Sydney University, Sydney, Australia

Genevieve Quek (G.Quek@westernsydney.edu.au)
The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Sydney, Australia

Tijl Grootswagers (T.Grootswagers@westernsydney.edu.au)
The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Sydney, Australia
School of Computer, Data and Mathematical Sciences, Western Sydney University, Sydney, Australia

Abstract

Human object recognition relies on both perceptual and semantic dimensions. Here, we examined how deep neural networks (DNNs) and large language models (LLMs) capture and integrate human-derived dimensions of object similarity. We extracted layer activations from CORnet-S and obtained BERT embeddings for 1853 images from the THINGS dataset. We used support vector regression (SVR) to quantify explained variance in human-derived dimensions. Results showed that multimodal integration improved predictions in early visual processing but offers limited additional benefits at later stages, suggesting that deep perceptual processing already encodes meaningful object representations.

Keywords: Object Recognition, Multimodal Integration, Deep Neural Networks, Large Language Models, Human-Derived Dimensions

Introduction

Human object recognition depends on both perceptual dimensions, such as color or shape, and semantic dimensions, such as category or conceptual relationships. To investigate the relevance of perceptual dimensions, previous research has often used deep neural networks (DNNs) to identify layers that correspond to human object recognition (Cichy et al., 2016; Kriegeskorte, 2015) and found that early layers in DNNs tend to capture on more simple, perceptual features, while higher layers align more closely with complex, semantic features (Guclu & Van Gerven, 2015). Other studies have shown that large language models (LLMs) capture semantic aspects of dimensions underlying object recognition, with high similarity to human judgments (Grand et al., 2022).

Combining LLMs with DNNs has been shown to outperform each modality individually (Marjieh et al., 2023), reinforcing the importance of multimodal integration in object recognition (Martin, 2016). To understand how perception and meaning interact in object recognition, it is essential to determine at which processing stages semantic information contributes.

In this study, we integrated LLM embeddings with individual DNN layers and identified how perceptual and conceptual representations align along the visual hierarchy. This allowed us to quantify where semantic knowledge enhances predictions of human-derived object dimensions, revealing the dynamics of multimodal integration in object recognition.

Methods

We used a publicly available dataset of 49 human-derived dimensions underlying human object recognition (Hebart et al., 2020) for 1853 images from the THINGS database (Hebart et al., 2019). Individual layer activations from CORnet-S (Kubilius et al., 2019) were extracted using a forward-pass per image, separately for the V1, V2, V4, and IT layers. To reduce dimensionality and equate feature space size between the DNN and LLM, probabilistic PCA (Halko et al., 2010) was applied separately to each DNN layer retaining the first 200 components. Next, LLM embeddings were derived from the BERT model (Devlin et al., 2019) by averaging hidden states for prompts based on concept names from the THINGS dataset (e.g., "fish"), followed by probabilistic PCA retaining the first 200 components. Finally, a combined predictor set was created by concatenating the top 100 PCA components from the DNN with the top 100 PCA components from the LLM embeddings for each image. This resulted in three distinct predictor sets, each containing 200 features per image: (1) ***DNN predictors*** (first 200 PCA components from DNN activations), (2) ***LLM predictors*** (first 200 PCA components from BERT embeddings), and (3) ***Combined predictors*** (concatenated top 100 PCA components from both the DNN and LLM).

Using these three different predictor sets, a 10-fold cross-validated support vector regression (SVR) with a radial-basis function kernel was used to predict the loading for each image on 49 dimensions (Hebart et al., 2020). To evaluate model performance and determine whether predictions exceeded chance-level accuracy, we calculated the explained variance (R^2) for each individual dimension separately. Statistical significance was assessed using 1000 permutations to generate a null distribution of R^2 values. The same procedure was applied to other DNNs (AlexNet, VGG) and LLMs (RoBerta, BERT) to assess the generalizability of the results across different architectures.

Results

We first assessed how perceptual features (DNN layers), semantic features (LLM embeddings), and their combination (DNN+LLM) predict human-derived dimensions**.** Explained variance (R^2) increased progressively along the DNN hierarchy from lower-level layers (V1, V2) to higher-level layers (V4, IT), reflecting a clear hierarchical structure of visual feature complexity (Fig. 1A). Semantic features from LLM embeddings alone also explained substantial variance (Fig. 1A).

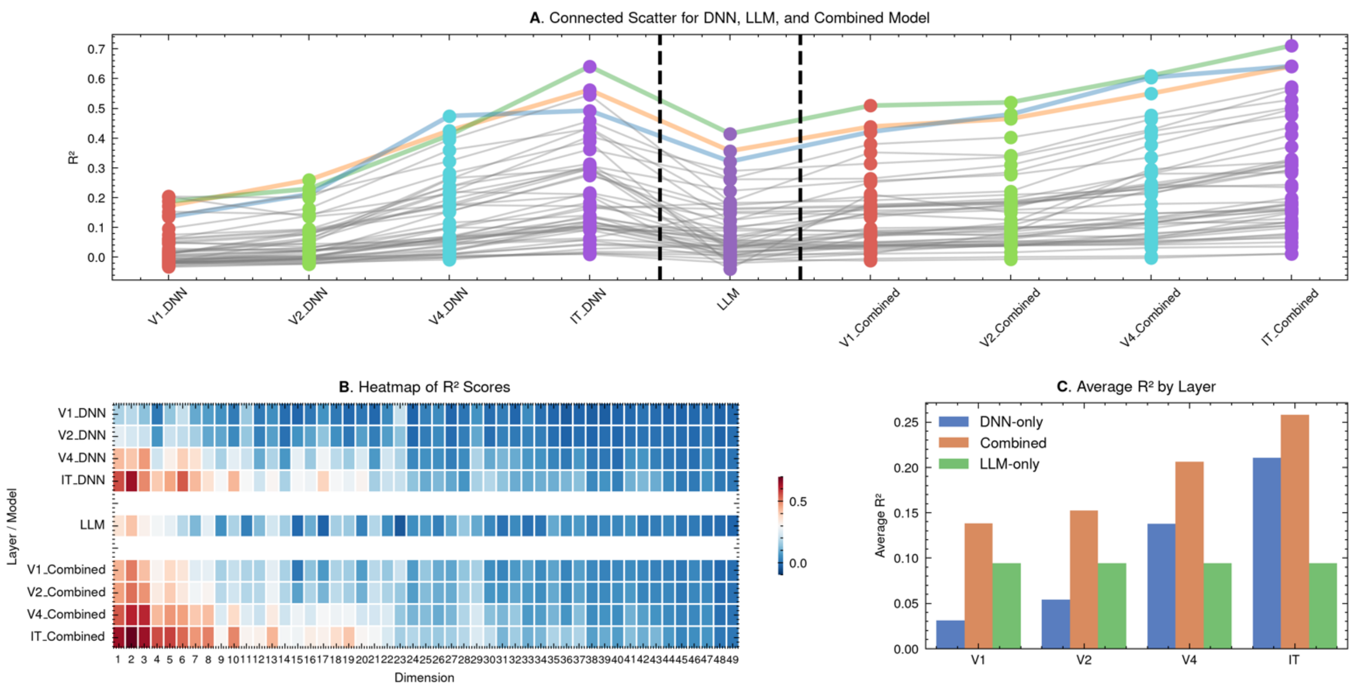

Figure 1. Multimodal prediction of human object dimensions. **A:** R^2 scores for each dimension (dots) across DNN layers, LLM, and combined models; gray lines connect the same dimension across models. Colored lines show trajectory of top 3 dimensions averaged across all methods. **B:** Heatmap of R^2 scores per model/layer (rows) and dimension (columns); each square represents one prediction. **C:** Average R^2 by layer, showing improved performance for combined models. Shown trend was observed across a variety of methodological choices.

Inspecting individual dimension predictions revealed notable variation, with some dimensions benefiting more from perceptual than semantic features (Fig. 1B).

However, across dimensions and layers, the combination of DNN and LLM features had some added benefit. For early DNN layers, the combination with LLM added to the performance while the benefit decreased as we moved up the hierarchy (Fig. 1C).

Conclusion

We investigated how individual DNN layers and LLM embeddings correspond to human-derived object dimensions, and whether combining both modalities enhances this correspondence. While multimodal integration benefits early visual processing, the strongest predictions occurred at later stages, where DNNs already capture high-level object representations effectively. Our results suggest that linguistic knowledge does not consistently enhance DNN-based representations in IT. This highlights that deep perceptual processing in DNNs already incorporates meaningful structure at higher levels of the visual hierarchy.

Acknowledgments

This work was supported by the Australian Research Council (DE230100380).

References

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1), 27755. https://doi.org/10.1038/srep27755

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, 4171–4186.

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour*, *6*(7), 975–987. https://doi.org/10.1038/s41562-022-01316-8

Guclu, U., & Van Gerven, M. A. J. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, *35*(27), 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Halko, N., Martinsson, P.-G., & Tropp, J. A. (2010). *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions* (No. arXiv:0909.4061). arXiv. https://doi.org/10.48550/arXiv.0909.4061

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, *14*(10), e0223792. https://doi.org/10.1371/journal.pone.0223792

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*(11), 1173–1185. https://doi.org/10.1038/s41562-020-00951-3

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*(1), 417–446. https://doi.org/10.1146/annurev-vision-082114-035447

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., Issa, E., Bashivan, P., Prescott-Roy, J., Schmidt, K., Nayebi, A., Bear, D., Yamins, D. L., & DiCarlo, J. J. (2019). *Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs*.

Marjieh, R., Rijn, P. van, Sucholutsky, I., Sumers, T. R., Lee, H., Griffiths, T. L., & Jacoby, N. (2023). *Words are all you need? Language as an approximation for human similarity judgments* (No. arXiv:2206.04105). arXiv. https://doi.org/10.48550/arXiv.2206.04105

Martin, A. (2016). GRAPES—Grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain. *Psychonomic Bulletin & Review*, *23*(4), 979–990. https://doi.org/10.3758/s13423-015-0842-3

Supplementary Material

A: Description of Human-Derived Dimensions (Hebart et al., 2020)

1	'made of metal / artificial / hard'
2	'food-related / eating-related / kitchen-related'
3	'animal-related / organic'
4	'clothing-related / fabric / covering'
5	'furniture-related / household-related / artifact'
6	'plant-related / green'
7	'outdoors-related'
8	'transportation / motorized / dynamic'
9	'wood-related / brownish'
10	'body part-related'
11	'colorful'
12	'valuable / special occasion-related'
13	'electronic / technology'
14	'sport-related / recreational activity-related'
15	'disc-shaped / round'
16	'tool-related'
17	'many small things / course pattern'
18	'paper-related / thin / flat / text-related'
19	'fluid-related / drink-related'
20	'long / thin'
21	'water-related / blue'
22	'powdery / fine-scale pattern'
23	'red'
24	'feminine (stereotypically) / decorative'
25	'bathroom-related / sanitary'
26	'black / noble'
27	'weapon / danger-related / violence'
28	'musical instrument-related / noise-related'
29	'sky-related / flying-related / floating-related'
30	'spherical / ellipsoid / rounded / voluminous'
31	'repetitive'
32	'flat / patterned'
33	'white'
34	'thin / flat'
35	'disgusting / bugs'
36	'string-related'
37	'arms/legs/skin-related'
38	'shiny / transparent'
39	'construction-related / physical work-related'
40	'fire-related / heat-related'
41	'head-related / face-related'
42	'beams-related'
43	'seating-related / put things on top'
44	'container-related / hollow'
45	'child-related / toy-related'
46	'medicine-related'
47	'has grating'
48	'handicraft-related'
49	'cylindrical / conical'