Metric-Learning Encoding Models Identify Processing Profiles of Linguistic Features in BERT's Representations

Louis Jalouzot
PSL University, EHESS, LSCP, CNRS; ENS de Lyon
jalouzot.louis@gmail.com

Robin Sobczyk
PSL University, EHESS, LSCP, CNRS; ENS Paris-Saclay

Bastien Lhopitallier
PSL University, EHESS, LSCP, CNRS; ENS Paris-Saclay

Jeanne Salle
PSL University, EHESS, LSCP, CNRS; Télécom Paris

Nur Lan
Ecole Normale Supérieure

Emmanuel Chemla[1]
PSL University, EHESS, LSCP, CNRS

Yair Lakretz[1]
PSL University, EHESS, LSCP, CNRS
yair.lakretz@gmail.com

[1] Equal contribution

Abstract

We introduce Metric-Learning Encoding Models (MLEMs), a new framework to learn a feature-based metric explaining the geometry of neural representations[2]. Applying MLEMs to BERT, we track various linguistic features (e.g., tense, subject number) and find distinct importance profiles across layers. For a given layer, feature importance ranking corresponds to a hierarchical geometry of representations. A univariate variant of our model reveals remarkable spontaneous disentanglement: in all layers, distinct neuron groups specialize in encoding single, specific linguistic features. MLEMs are more robust than popular decoding methods, offering a powerful tool for analyzing representations in artificial and biological neural systems.

Keywords: Feature Attribution; Sentence Embedding; Multivariate Encoding; BERT; Metric-Learning

Introduction

A central question in neuroscience and AI is how neural networks encode and process language. While modern language models offer unprecedented access to neural activity, their internal representational principles remain largely unknown. A key step is understanding where and how fundamental *linguistic features*—like grammatical number or syntactic structure—are encoded.

Two main approaches exist: *decoding* and *encoding* (King et al., 2020). Decoding (or 'diagnostic probes') predicts features from neural activity (Hupkes & Zuidema, 2017; Tenney et al., 2019). However, high decodability doesn't imply causality; a feature might be decodable due to correlation with a truly encoded one. Encoding models reverse this, predicting activity from features. This allows controlling confounds, but traditional encoding models are often univariate, predicting single unit activity (e.g., one electrode/voxel), thus failing to capture distributed representations.

We introduce Metric-Learning Encoding Models (MLEMs), a multivariate encoding approach preserving benefits of both methods. MLEMs model *distances* between neural representations of stimuli (e.g., sentences) as a weighted function of their linguistic feature differences. This quantifies each feature's contribution to representational geometry while accounting for distributed encoding. We apply MLEMs to BERT (Devlin et al., 2019) to trace sentence processing with complex relative clauses, revealing how syntactic information is encoded across layers.

Methods

Datasets

To study complex syntactic structure encoding, we created the **Relative-Clause Dataset** (7,680 sentences) probing syntactic ambiguity and clause embedding. They are varied and annotated for 14 linguistic features, including clause attachment site (center vs. peripheral), relative clause type (subject vs. object), and noun grammatical number. For the univariate analysis, we used a separate **Short-Sentence Dataset** contrasting basic features like tense, question vs. declarative.

[2]A preprint of this work is available (c.f. Jalouzot et al. (2024)).

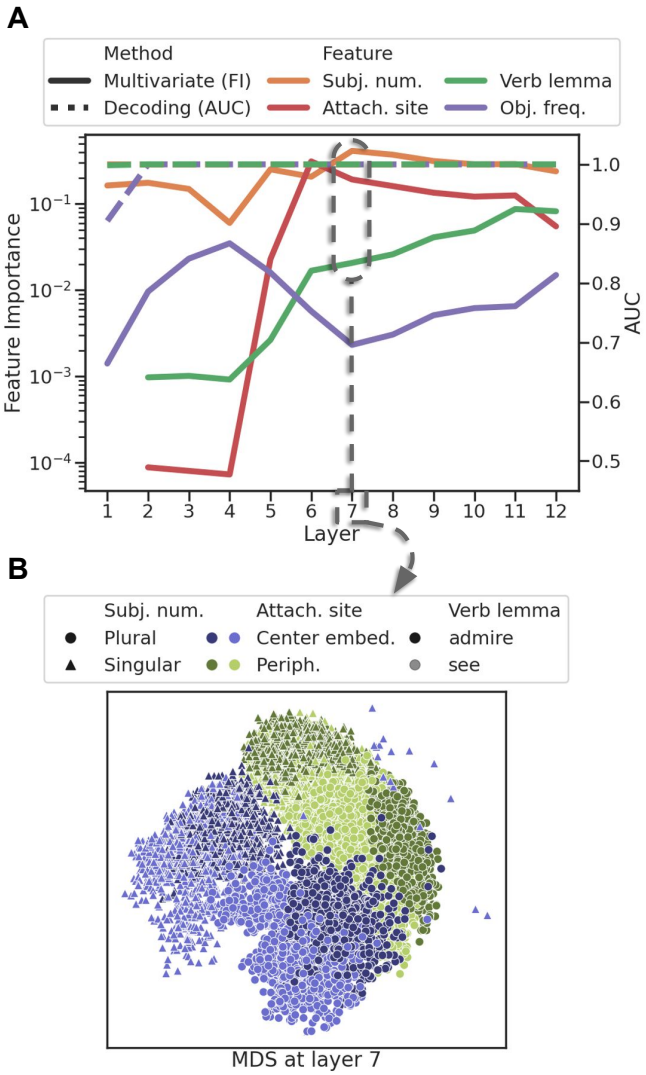

Figure 1: **Processing Profile and Hierarchical Geometry for Relative Clauses. A:** Feature Importances (FIs) for top features in the Relative-Clause dataset. 'Attachment site' FI (red) increases dramatically in middle layers. Decoding accuracy (AUC, dashed lines) is at ceiling. **B:** MDS plot of layer 7 sentence representations. Representations form nested clusters following FI order: *Subject number* (circles/triangles) $\subset$ *Attachment site* (colors) $>$ *Verb lemma* (shades).

Language Model and Representations

We used `bert-base-uncased`. For each sentence, we extracted hidden states from the `[CLS]` token of each of the 12 layers, yielding a 768-dimensional vector per layer as an aggregated, sentence-level representation.

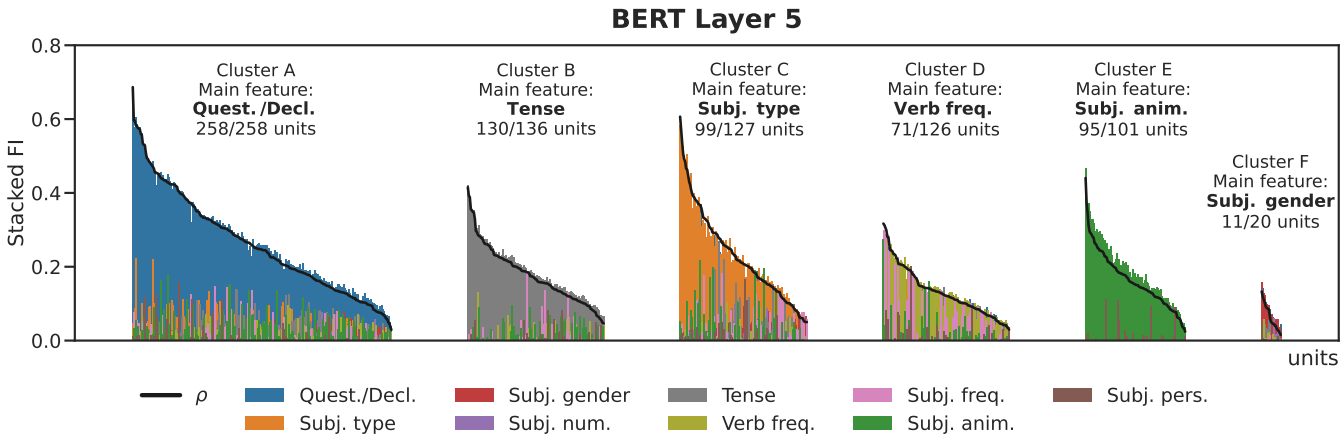

Figure 2: **Disentanglement of Linguistic Features in BERT Layer 5.** Units in layer 5 were clustered by univariate FI profiles. Each stacked bar is a unit's Feature Importance profile (colors are features). Units per cluster are sorted by univariate model performance (black line). Clusters show strong specialization; most units in a cluster are selective for one dominant feature.

Metric-Learning Encoding Models (MLEMs)

MLEMs assume neural representation geometry is informative. We explain pairwise distances between neural sentence representations using their linguistic features.

Given sentences, we compute two distance matrices. First, the *neural distance matrix* $D^{\mathcal{N}}$, where $D^{\mathcal{N}}_{ij}$ is the Euclidean distance between neural representations of sentences s_i and s_j. Second, the *feature distance matrix* $D^{\mathcal{F},W}$, based on feature difference vectors $\Delta(s_i, s_j) = (\mathbb{1}_{f(s_i)\neq f(s_j)})_{f\in\mathcal{F}}$, indicating differing features between s_i and s_j from a predefined set of features $\mathcal{F}$. Feature distances are the weighted norm of these difference vectors: $(D^{\mathcal{F},W}_{ij})^2 = ||\Delta(s_i, s_j)||^2_W = \Delta(s_i, s_j)^T W \Delta(s_i, s_j)$, where W is a learned symmetric positive definite matrix. An MLEM optimizes W to best align feature and neural distances. For simplicity, we assume W is diagonal. We learn these weights via non-negative least-squares optimization.

To assess each feature's contribution, we compute its **Feature Importance (FI)** via a permutation test: the drop in model performance (Spearman's ρ) on a held-out set when that feature's values are randomly shuffled.

Results

Processing Profiles of Linguistic Features. Figure 1A shows FI profiles for the most important linguistic features across BERT's 12 layers. While subject number is important throughout, attachment site (distinguishing center-embedding from right-branching) shows a striking pattern: negligible importance in early layers, increasing over three orders of magnitude at layer 5, peaking in middle layers. This sharp rise suggests complex computations for this structural ambiguity dominate at this stage. In contrast, standard decoding (linear classifier) shows near-perfect accuracy for all features across all layers (dashed lines), failing to reveal this layer-specific specialization and highlighting decoding's false positive risk.

Hierarchical Organization of Neural Representations. The MLEM-identified feature importance order directly reflects neural representation geometry. Figure 1B shows a 2D MDS projection of layer 7 sentence representations, revealing a clear hierarchical structure. Representations first separate by the most important feature, subject number (circles vs. triangles). Within these, sub-clusters form by the second-most important, attachment site (blue vs. green). These sub-clusters are further organized by verb lemma (light vs. dark). This nested structure would be hard to uncover without MLEM's feature ranking.

Disentanglement of Features into Specialized Units. We used a univariate variant of MLEM, which considers independently each of the 768 units per layer, then clustered units by univariate FI profiles. Figure 2 shows the results for BERT's layer 5 on the short sentence dataset. Units segregate into distinct clusters, each specialized for a single linguistic feature. This reveals remarkable linguistic information disentanglement: BERT dedicates separate neuron groups to different features without explicit supervision during training.

Discussion

Our findings show that MLEMs are a powerful tool for interpreting neural network representations. By modeling representational geometry, they offer more nuanced and robust analysis than standard decoding, which can misinterpret correlations. As multivariate models, they capture distributed patterns missed by univariate approaches, while their comparison reveals single-unit feature specialization. The MLEM framework is general, applicable to other domains (e.g., vision), and neural systems (e.g., human brain), offering a promising path to understanding neural computation principles.

Acknowledgements

This project was provided with computer and storage resources by GENCI at IDRIS thanks to the grant 2023-AD011013783R1 on the supercomputer Jean Zay's V100 partition.

References

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *North american chapter of the association for computational linguistics.* Retrieved from https://api.semanticscholar.org/CorpusID:52967399

Hupkes, D., & Zuidema, W. H. (2017). Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *J. Artif. Intell. Res.*, *61*, 907-926. Retrieved from https://api.semanticscholar.org/CorpusID:5013113

Jalouzot, L., Sobczyk, R., Lhopitallier, B., Salle, J., Lan, N., Chemla, E., & Lakretz, Y. (2024). *Metric-Learning Encoding Models Identify Processing Profiles of Linguistic Features in BERT's Representations.* Retrieved 2024-03-28, from http://arxiv.org/abs/2402.11608 doi: 10.48550/arXiv.2402.11608

King, J.-R., Gwilliams, L., Holdgraf, C., Sassenhagen, J., Barachant, A., Engemann, D., ... Gramfort, A. (2020). Encoding and decoding framework to uncover the algorithms of cognition. *The Cognitive Neurosciences (Sixth Edition)*, *58*. Retrieved from https://direct.mit.edu/books/edited-volume/5456/chapter-abstract/3967019/Encoding-and-Decoding-Framework-to-Uncover-the?redirectedFrom=fulltext

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., ... Pavlick, E. (2019). What do you learn from context? probing for sentence structure in contextualized word representations. *ArXiv*, *abs/1905.06316*. Retrieved from https://api.semanticscholar.org/CorpusID:108300988