Extended Abstract Track

Towards identifiable Latent Structures in Language Embeddings

Editors: List of editors' names

Abstract

This work aims to identify interpretable low-dimensional structure inherent in high-dimensional language embeddings. Prior studies have shown that *linear ICA* can uniquely transform embeddings into spaces with *semantically meaningful axes*, with evidence pointing toward extensions beyond language (including vision) (Yamagiwa et al., 2023; Li et al., 2024). As a natural extension, we consider *nonlinear ICA* to capture the latent structure of nonlinear internal representations; however, generic nonlinear ICA suffers from the long-standing *identifiability* problem (Hyvärinen et al., 2019). To address this, we adopt *CEBRA*, a contrastive-learning framework that achieves theoretical *identifiability up to linear transformations* by leveraging *auxiliary variables* (Schneider et al., 2023). In preliminary experiments using *emotion labels* as auxiliary variables, CEBRA maps sentence embeddings into a *low-dimensional, linearly separable* space, consistent with the view that its InfoNCE loss behaves as a *multiclass discriminative objective* under discrete labels. Moreover, across random initializations, the learned embeddings exhibit *high alignment up to linear transforms*, empirically supporting identifiability in practice. We discuss open questions regarding the *choice of auxiliary variables*, the *interpretation* of linearly equivalent solutions, and the *topology* of the learned low-dimensional manifolds. As a longer-term goal, we plan *brain-encoding* studies (fMRI) to test whether the discovered structures correspond to neural representations involved in affective language processing.

Keywords: Contrastive learning, CEBRA, Identifiability, Representation learning

1. Introduction

A central challenge in representation learning is the *interpretability* of high-dimensional vector spaces. For language embeddings, *linear ICA* has emerged as a powerful tool: prior work indicates that one can construct a *unique* linear transformation yielding *independent, interpretable semantic axes*, with signs of cross-modal applicability (Yamagiwa et al., 2023; Li et al., 2024). Yet Language Embedding model implement intrinsically *nonlinear* mappings, suggesting that their internal representations may organize meaning along nonlinear factors that linear methods only partially capture.

A natural next step is *nonlinear ICA*, aimed at recovering latent structure after nonlinear mixing. However, *generic nonlinear ICA lacks identifiability*: without additional assumptions or supervisory signals, true latent factors cannot be uniquely recovered (Hyvärinen et al., 2019). We therefore turn to *CEBRA* (Schneider et al., 2023), which circumvents this issue by *contrastive learning with auxiliary variables*, providing *identifiability up to linear transformations* under realistic conditions. This makes CEBRA an attractive, practically deployable alternative to unconstrained nonlinear ICA for embedding analysis.

2. Material and Method

2.1. Objective: Manifold Discovery

Our objective is to recover, within the high-dimensional observation space (LLM embeddings), a *low-dimensional manifold* that is *diffeomorphic* to the latent variable space responsible for observed variation. We experimentally apply *CEBRA* to this end.

2.2. Datasets

We evaluated our approach on two publicly available emotion-labeled text datasets. dair-ai/emotion corpusSaravia et al. (2018), which consists of 20000 English Twitter messages categorized into six basic emotions: anger, fear, joy, love, sadness, and surprise.

2.3. Embeddings

we obtained a fixed-dimensional vector representation using pre-trained transformer language models from Hugging Face. In particular, we extracted text embeddings from three models: BERT (Bidirectional Encoder Representations from Transformers)Devlin et al. (2019), RoBERTa (a robustly optimized BERT variant)Liu et al. (2019), and all-MiniLM-L6-v2**?**. BERT and RoBERTa were used in their base uncased versions, each producing a 768-dimensional embedding for an input sentence (corresponding to the model's hidden size). The all-MiniLM-L6-v2 model (a distilled sentence transformer) yields a more compact 384-dimensional sentence embedding.

2.4. CEBRA with Discrete Auxiliary Variables

CEBRA optimizes an *InfoNCE*-based objective. When the *auxiliary variable is a discrete label* (e.g., emotion), the loss *acts as a multiclass classifier*: it *attracts* same-label pairs and *repels* different-label pairs in the learned space. Consequently, we expect a *low-dimensional representation* in which classes are *linearly separable*.

2.5. Preliminary Findings

Our preliminary results demonstrate that CEBRA can successfully overcome the identifiability problem of nonlinear ICA Using emotion-labeled text, we extract sentence embeddings and train CEBRA with labels as the auxiliary variable. The resulting latent space shows *clearly separable clusters* aligned with labels. This is shown in Figure 1. Crucially, training from different random seeds yields embeddings that are *highly congruent up to linear transforms* (rotations/scalings), supporting *stable, identifiable* latent structure in practice.

3. Discussion and Open Questions

CEBRA enables us to sidestep the identifiability bottleneck of generic nonlinear ICA, yielding *low-dimensional, linearly separable* structure from language embeddings while remaining *unique up to linear transformations*. Several issues merit further study:

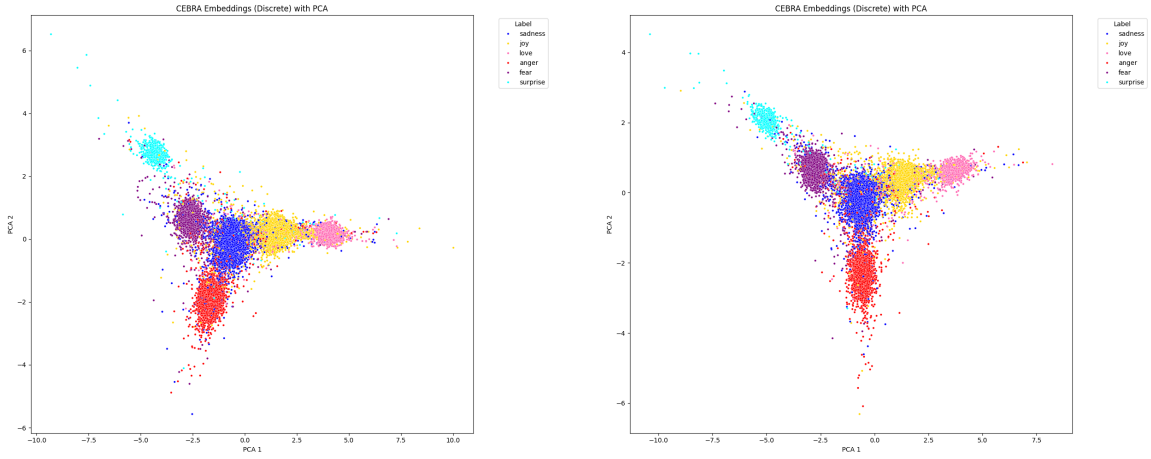


Figure 1: CEBRA Embedding visualization Left:cebra outputdimension 4dim with PCA Right:cebra outputdimension 5dim with PCA

1. **Choice of auxiliary variables.** Beyond emotion labels, which supervisory signals (topic, style, syntax, speaker/context) best expose semantically meaningful latent geometry?

2. **Interpreting linearly equivalent embeddings.** Because solutions are defined up to linear transforms, *axes need not carry fixed semantics* individually; interpretation should leverage *global geometry* and *pairwise relations* among clusters/trajectories.

3. **Manifold topology.** CEBRA extracts *nonlinear low-dimensional manifolds* embedded in the observation space. Characterizing their *topological properties* (e.g., connectivity, holes, loops) is key to understanding how LLMs "fold" meaning.

4. Future Work: Neuroscientific Validation

The ultimate goal of this research is to establish that the computationally identified low-dimensional structures correspond to the mechanisms of language processing in the human brain. To this end, we plan to validate our findings using a brain encoding mode framework. The experimental paradigm will involve two main stages. First, we will use CEBRA to derive low-dimensional coordinates for a set of linguistic stimuli (e.g., sentences). Concurrently, we will record brain activity (via fMRI) from human subjects as they listen to or read the same set of stimuli. Subsequently, we will train a regression model to predict the recorded brain activity in each voxel using the low-dimensional coordinates from CEBRA as input features. This interdisciplinary validation experiment would bridge the gap between representation learning in artificial intelligence and cognitive neuroscience, paving the way for a deeper understanding of both artificial and natural language processing.

5. Citations and Bibliography

The bibliography is displayed using `\bibliography`.

Acknowledgments

Acknowledgements go here.

References

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 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 (NAACL-HLT)*, pages 4171–4186, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

A. Hyvärinen, H. Sasaki, and R. E. Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pages 859–868. PMLR, April 2019. doi: 10.48550/arXiv.1805.08651. URL https://proceedings.mlr.press/v89/hyvarinen19a.html.

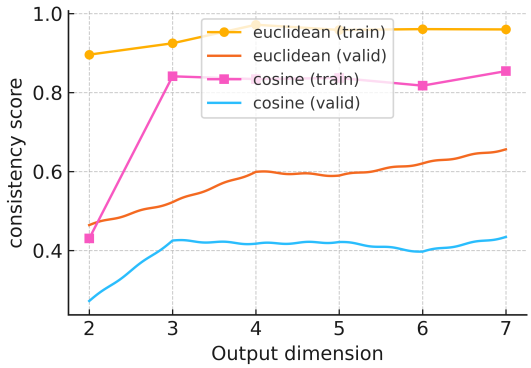
R. Li, T. Matsuda, and H. Yanaka. Exploring intra and inter-language consistency in embeddings with ICA. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 19104–19111, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1065. URL https://aclanthology.org/2024.emnlp-main.1065/.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, July 2019. URL https://arxiv.org/abs/1907.11692.

E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3687–3697, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1404. URL https://aclanthology.org/D18-1404/.

S. Schneider, J. H. Lee, and M. W. Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617:360–368, May 2023. doi: 10.1038/s41586-023-06031-6. URL https://www.nature.com/articles/s41586-023-06031-6.

H. Yamagiwa, M. Oyama, and H. Shimodaira. Discovering universal geometry in embeddings with ICA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4647–4675, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.283. URL https://aclanthology.org/2023.emnlp-main.283/.



Appendix A. First Appendix

Results with Metrics