Towards the Identification of Latent Structures in Language Embeddings

Ryunosuke Abe ABE.RYUNOSUKE.AT2@NAIST.AC.JP
Takatomi Kubo TAKATOMI-K@IS.NAIST.JP
Kazushi Ikeda I.KAZUSHI@NAIST.AC.JP
Nara Institute of Science and Technology, Ikoma, Japan

Editors: Editor's name

Abstract

This work aims to identify *the latent structure* of high-dimensional language embeddings by applying a theoretically grounded framework for *identifiable* representation learning. Prior studies have shown that linear ICA can transform embeddings into spaces with semantically meaningful axes. As a natural extension, nonlinear ICA has been proposed to recover the latent structure generated through nonlinear mixing in the data-generating process. We adopt *CEBRA*, a contrastive learning framework grounded in nonlinear ICA theory, which ensures *identifiability* of the latent structure up to linear transformations by leveraging auxiliary variables. In our preliminary experiments on an emotion-labeled text dataset, where we use emotion labels as auxiliary variables, the resulting CEBRA embeddings form a low-dimensional space that exhibits linear separability. Moreover, across random initializations, the learned embeddings exhibit consistency up to linear transformations, empirically supporting practical identifiability of the learned representation. We discuss open questions regarding the interpretation of the label-related latent representations and future directions, including their potential alignment with human neural processing.

Keywords: Identifiability, Representation learning, Language embedding, CEBRA, Contrastive 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 the ability to construct a space with independent, interpretable semantic axes (Yamagiwa et al., 2023; Li et al., 2024). However, language embedding models intrinsically implement nonlinear mappings, suggesting that linear methods may have limitations in capturing their internal representations. A natural next step is *nonlinear ICA*, aimed at recovering latent sources after nonlinear mixing. By contrast, generic nonlinear ICA lacks identifiability, meaning that the true latent factors cannot be uniquely recovered. Conditioning on auxiliary variables provides a solution by leveraging the conditional independence of the latent sources. This approach enables identifiability (Hyvärinen et al., 2019). We turn to *CEBRA* (Schneider et al., 2023), which builds directly on this theoretical foundation and provides *identifiability up to linear transformations* under several conditions. In this preliminary work, we use emotion labels as auxiliary variables in CEBRA, to form informative positive pairs from samples sharing the same label, thereby providing the conditional structure needed for identifiability.

2. Related works: CEBRA

CEBRA, originally developed for neural data, is a library for representation learning with rich configurability. Assuming that the data distribution is sufficiently diverse, the learned embeddings from independent runs, denoted as $f(\boldsymbol{x})$ and $\tilde{f}(\boldsymbol{x})$, are consistent up to an invertible linear transformation: $f(\boldsymbol{x}) = \boldsymbol{L}\tilde{f}(\boldsymbol{x})$, where $\boldsymbol{L}$ is an invertible linear matrix. Crucially, under appropriate assumptions about the generative process, CEBRA may recover representations that reflect the ground-truth latent structure up to a linear transformation. This identifiability is based *"on the linear identifiability of learned representations"* Roeder et al. (2020) for a broad model family with a canonical discriminative form. This perspective is also compatible with empirical findings that simple linear probes often perform surprisingly well in large language models (e.g., BERT), in part because aspects of their latent structure are linearly recoverable. This suggests that a framework based on linear identifiability may offer a useful lens for studying such embeddings.

3. Method

3.1. Experiment overview

We transform the dataset below using multiple language embedding models to obtain the embeddings. We construct the sentence embeddings with an 80/20 train–validation split. Then, we apply CEBRA to learn latent representations. We train a CEBRA model on the training subset five times with different random seeds for each language embedding model to assess the consistency of the learned representations. We use the discrete emotion labels for training the CEBRA models with either a cosine-based or a Euclidean-based contrastive objective.

Additionally, we visualize the latent spaces using PCA—applied when the CEBRA output dimensionality exceeds three—to qualitatively inspect cluster structure and to compare these projections with those of the original language embeddings. To quantify consistency across runs, we compute pairwise linear regressions between all five embeddings obtained from different seeds and report the average coefficient of determination R^2.

3.2. Datasets

We evaluated our approach on an emotion-labeled short-text corpus: **dair-ai/emotion** Saravia et al. (2018) Twitter corpus with six basic emotions (anger, fear, joy, love, sadness, and surprise).

3.3. Embeddings

We obtain fixed-dimensional vector representations using pretrained Transformer based language embedding models from Hugging Face. Specifically, we use **BERT-base-uncased** Devlin et al. (2019) and **RoBERTa-base** Liu et al. (2019), both with 768-dim embeddings, as well as the Sentence-Transformer model **all-MiniLM-L6-v2** Wang et al. (2020) with 384-dim embeddings. For all models, we compute sentence embeddings by mean-pooling the final-layer token representations over non-padding tokens using the attention mask, avoiding any special pooling heads or the [CLS] token for comparison.

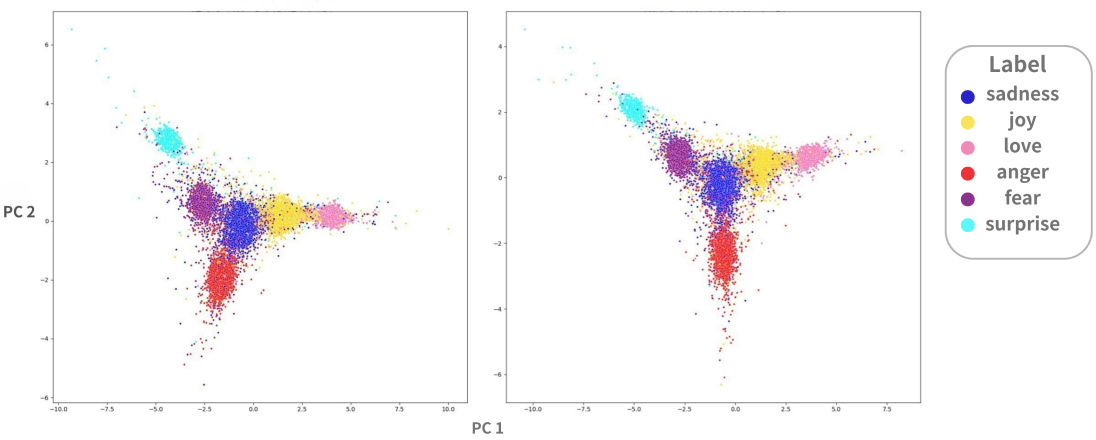

Figure 1: CEBRA Embedding visualization with PCA (Language Embedding: all-MiniLM-L6-v2). (Left) CEBRA output: 4d. (Right) CEBRA output: 5d.

4. Preliminary results

Our results demonstrate that CEBRA can find latent structure consistent across runs. When we align the learned embeddings by CEBRA from different runs using an optimal linear transformation, we obtain high coefficients of determination R^2, indicating that the learned latent structure is stable up to linear transformations. Through experiments, the Euclidean distance exhibits higher consistency than cosine similarity. Additionally, the choice of language embedding models substantially influences the resulting latent structure; sentence transformer models show high consistency across runs. Qualitatively, 2D visualizations of the CEBRA embeddings reveal clearer cluster structures that correspond to the emotion labels with high separability (Figure 1), whereas the PCA projections of the original language embeddings remain largely entangled. In addition, we observe that the overall cluster structure is largely preserved when varying the output dimensionality of CEBRA.

5. Discussion

Our preliminary results reveal that CEBRA produces latent embeddings with clear and well-separated cluster structures aligned with the emotion labels. While the model leverages these labels during training, the resulting organization is not merely a product of label-based discrimination. Instead, CEBRA appears to learn a latent representation that preserves the relational configuration among the labels, rather than collapsing the space into clusters optimized solely for classification.

This indicates that CEBRA uses the emotion labels to carve out a latent structure organized along label-relevant factors, refining—rather than merely reproducing—the geometry present in the original language embeddings. Furthermore, the arrangement of clusters remains highly consistent across language models, suggesting that the label-related organization may reflect a stable latent structure associated with the data-generating process.

In this work, we focus on encoder-style language models, assuming their representations to be more interpretable for this analysis. Future work will extend the investigation to decoder-only transformer models and examine layer-wise representations.

6. Open questions and future directions

We highlight questions that serve as the basis for our ongoing investigation and discussion.

1. **Choice of auxiliary variables with minimal inductive bias**
   Although this work uses discrete emotion labels to satisfy the conditions required for identifiability, an important open question is how to select auxiliary variables that introduce only minimal inductive bias. Identifying variables that genuinely capture the underlying structures while introducing only minimal inductive bias remains highly challenging.

2. **Interpretation of representations under linear indeterminacy**
   A critical issue for interpretation is that CEBRA representations are linearly indeterminate, meaning they are identifiable only up to an invertible linear transformation. This raises a fundamental question: how can we define robust and meaningful metrics that depend solely on the relative geometric structure — even when the latent space may form diverse and complex nonlinear manifolds — rather than absolute coordinate values?

3. **Geometry and topology of high dimensional latent structure**
   To enable both the interpretation of individual embedding spaces and the comparison of representations across models, datasets, and even languages, we aim to characterize latent structure from two complementary perspectives:
   Geometric structure: Analyzing distance/similarity patterns and relative geometric arrangements to characterize the local and global organization of the latent space.
   Topological structure: Examining cluster relationships, connectivity patterns, and manifold-level organization to capture higher-level structural regularities.

4. **Extensions to multimodal and neural data** Finally, we plan to extend this framework to multimodal settings, examining whether similar structural properties emerge when comparing text representations with those derived from image or audio data. Furthermore, we aim to connect this line of work back to its neuroscience origins by applying the analysis to neural recordings, exploring whether analogous structures appear in representations derived from neural activity.

Acknowledgments

This work was supported by JST, CREST Grant Number JPMJCR22P2, Japan.

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.

Geoffrey Roeder, Luke Metz, and Diederik P. Kingma. On linear identifiability of learned representations, 2020. URL https://arxiv.org/abs/2007.00810.

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.

W. Wang, N. Reimers, and I. Gurevych. all-minilm-l6-v2 (sentencetransformers, microsoft). https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2, 2020.

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. Consistency score across output dimensions

To examine how the CEBRA output dimensionality affects consistency, we evaluated multiple latent dimensions of CEBRA's embeddings. Using BERT embeddings with the Euclidean objective, we found that consistency remains high even at low dimensionalities and becomes stable by dimension 3 or 4.

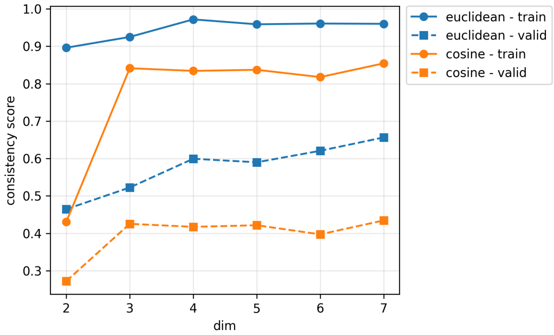

Figure 2: Consistency score of bert-base-uncased