Interpreting Feature Importance in Autoencoders for Dimensionality Reduction via MEDAL

Name: Vicky Jung (Mentors: Claire He, Irene Chang, Dr. Genevera Allen, 2026 IICD SRP)

The recent proliferation of big data technology development underpinned by modern computational infrastructure have expanded the capabilities and performance of machine learning and deep learning models fit in scientific disciplines. Consequently, much data in recent years are high-dimensional with numerous features, calling forth the need for dimension reduction techniques. Widely used dimension reduction methods include linear methods such as Principal Component Analysis (PCA) and non-linear methods such as t-SNE, UMAP, and PHATE. These techniques aim to preserve the intrinsic structure of high-dimensional data in a manifold embedding, enabling downstream analysis and interpretation for single-cell genomics. However, dimension reduction methods are oftentimes validated simply through visual appeal alone or by relying on default parameters, which may not guarantee the most faithful geometric representation of the initial data. Furthermore, assessing the validity of nonlinear dimensional reduction techniques is a nontrivial task as they do not provide an invertible map. Thus, MEDAL (Manifold EMbedding Distillation via Autoencoder Learning) was formulated. MEDAL is a framework that trains the encoder such that the latent representation matches the teacher embedding (i.e. fitted manifold embedding) through knowledge distillation while the decoder learns how to reconstruct the original given data. As a result, instead of a static manifold embedding, MEDAL provides an explicit map for unseen data samples and quantifies point-wise reconstruction error evaluating the distortion level of the manifold space. This ultimately enables fair and rigorous comparison among different dimensionality reduction techniques and supports informed hyperparameter optimization.

However, it is equally important to understand which features of the data most strongly contribute to the choice of the low–dimensional embedding. As with most deep learning models, autoencoders are typically perceived as black-box models that lack interpretability of the decision-making process. Thus, practicing feature importance, also known as model interpretability, can unveil latent relationships underlying the data that may be necessary for further analysis. We aim to discover which features are significant for preserving global and local data structures, compare which features are preserved between different dimension reduction methods, and develop a methodology to interpret the manifold embedding via the uncovered features, integrated with the workflow of MEDAL. To do so, we refer to explainable artificial intelligence algorithms such as DeepLIFT and DeepSHAP for interpretation. Deciphering important features will further empower understanding of the cluster structures and separations as well as downstream conclusions.

References

Chang, I.; Zikry, T.; Allen, G. MEDAL: Manifold Embedding Distillation via Autoencoder Learning. *arXiv*, 10.48550/arXiv.2605.24244, (2026).