Optimizing Space Complexity of EmbedOR Embedding Creation via Subsampling and Alignment

Derrick Ma and Samyah Ahmed (Mentor: Dr. Andrew Blumberg, 2026 IICD SRP)

EmbedOR is a Stochastic Neighbor Embedding (SNE) algorithm that utilizes elements of the previous tSNE and UMAP algorithms via Olliver-Ricci Curvature (ORC) based metric learning [1]. It is designed to highlight cluster structure more effectively than the aforementioned traditional SNE algorithms and is not only resistant against fragmentation of continuous, high density regions of data, but also effective at providing deeper geometric info on existing datasets. However, EmbedOR has a space complexity bottleneck of $O(N^2)$ as it is necessary to store the relationship from one point to every other point, making it more difficult to use on larger, more complex datasets. Previously, researchers have attempted to mitigate this– in different contexts– by subsampling larger sets of data via a uniform probability distribution, but found that subsampling this way degraded the quality of the embeddings produced. To attempt to address these limitations, we will explore techniques that improve the space complexity via subsampling whilst still maintaining the quality of the embeddings. One technique we will use is experimenting with different probabilistic distributions when subsampling, aside from simply sampling using a uniform distribution. Additionally, we introduce Procrustes distance, which is "a matching distance for (partially defined) vector-valued functions on a given finite set" [2]. These two ideas combined, we will analyze the relationship between EmbedOR performance when subsampling using various probability distributions and using an alignment algorithm related to Procrustes distance to make the subsamples more robust.

References

[1] Blumberg, A.J. *et al*. EmbedOR: Provable Cluster-Preserving Visualizations with Curvature-Based Stochastic Neighbor Embeddings. *arXiv* 2509.03703 (2025).
[2] Blumberg, A.J. *et al*. Resampling and Averaging Coordinates on Data. *arXiv* 2408.01379 (2024).