Revisiting High Dimension Data Visualization Measures

Alexander Kiefer & Md. Khaledur Rahman

Indiana University Bloomington

Introduction

Dimensionality reduction is an important approach to extract meaningful information from high-dimensional data and represent it in a low dimensional space for visualization. There are different methods in the literature to perform this task; however, the quality of the visualization may not be the same. In our past experiment, we focused upon comparing the effectiveness of several dimensionality reduction techniques through their runtime, memory usage, and qualitative observations. Our new results, this year, hope to add more clarity to the ones already discussed and provide a clearer picture of the strengths and weaknesses of the various algorithms.

Objectives

For this experiment, we will be combining our previous findings with new measures to better evaluate the use cases of four dimensionality reduction algorithms, TSNE, LargeVis, UMAP, and Trimap. In addition to memory consumption and runtime, we will be adding two quality metrics, those being the David Bouldin score and the Silhouette score.

The David-Bouldin score measures the quality of the clustering through the average ratio of within-cluster distances to between-cluster distances, with the best score being 0. With $\boldsymbol{s_i}$ being the average distance between each point of cluster $\boldsymbol{i}$ and the centroid of that cluster and $\boldsymbol{d_{ij}}$ the distances between the centroids of clusters $\boldsymbol{i}$ and $\boldsymbol{j}$, we choose a similarity measure

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

Using this similarity measure, we can then calculate the David-Bouldin score as

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j} R_{ij}$$

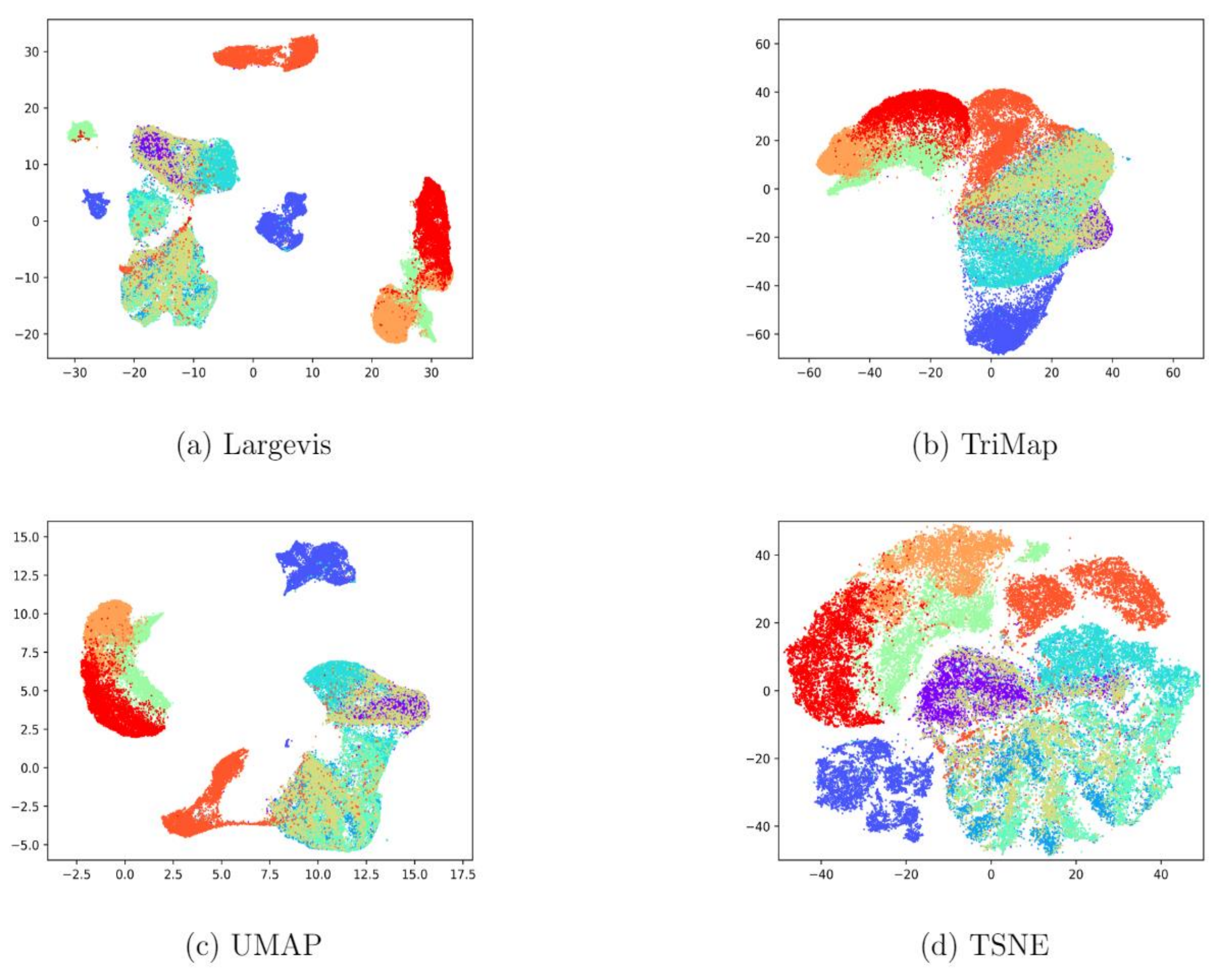
Where $\boldsymbol{k}$ is the total number of clusters in the embedded dataset.

The silhouette score measures the quality of the delineation between each point in each cluster on average, where positive values indicate superior delineation of values between clusters, values near 0 indicate overlapping clusters, and more negative values indicate points mapped to incorrect clusters. With $\boldsymbol{a}$ being the average distance between a sample and all other points in the same cluster and $\boldsymbol{b}$ being the average distance between a sample and all other points in the next nearest cluster, we find the mathematical formulation of the score to be

$$s = \frac{b-a}{max(a,b)}$$

Results

Fashion MNIST 2D Embeddings



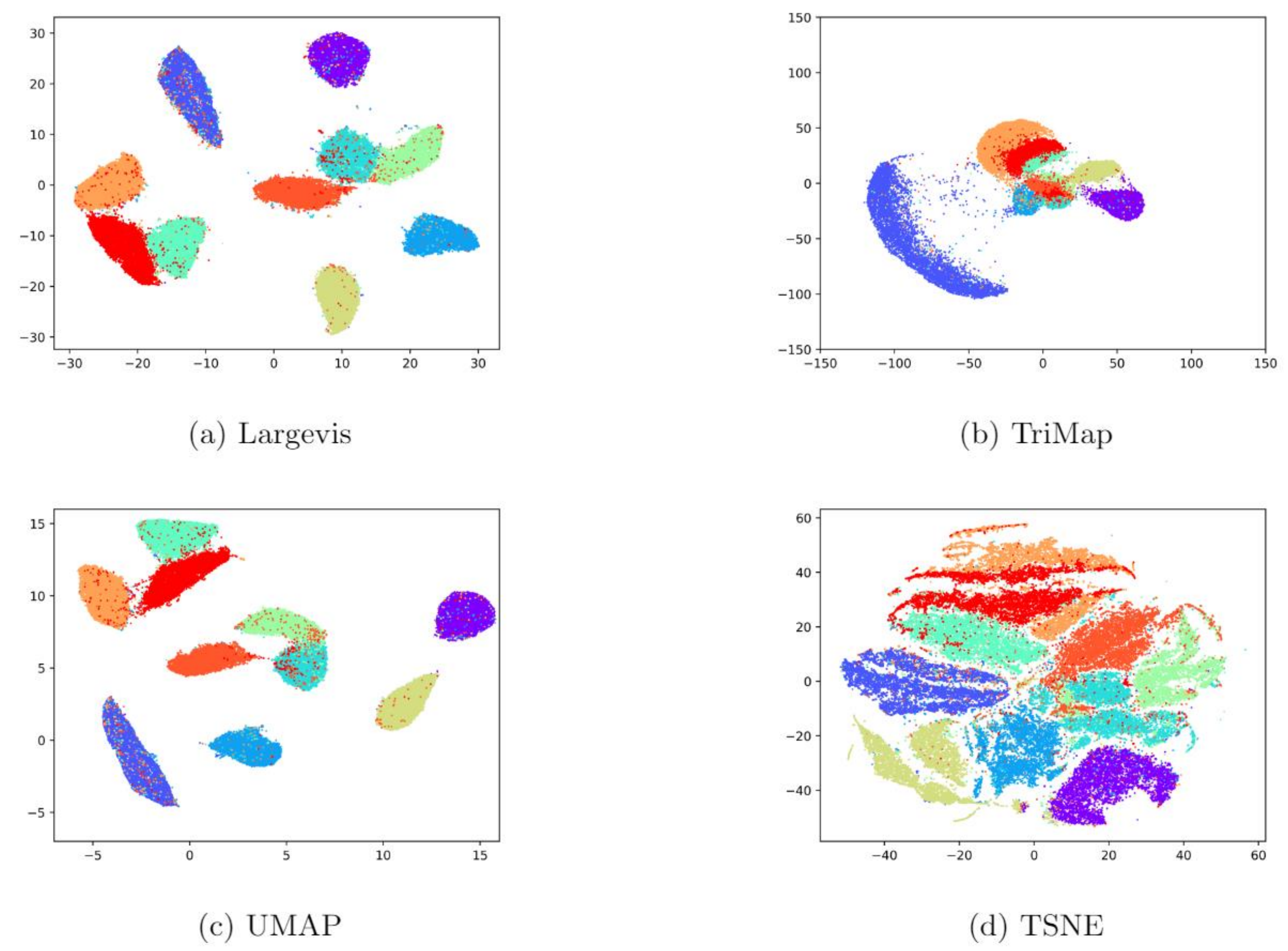
(a) Largevis

(b) TriMap

(c) UMAP

(d) TSNE

MNIST Digits 2D Embeddings



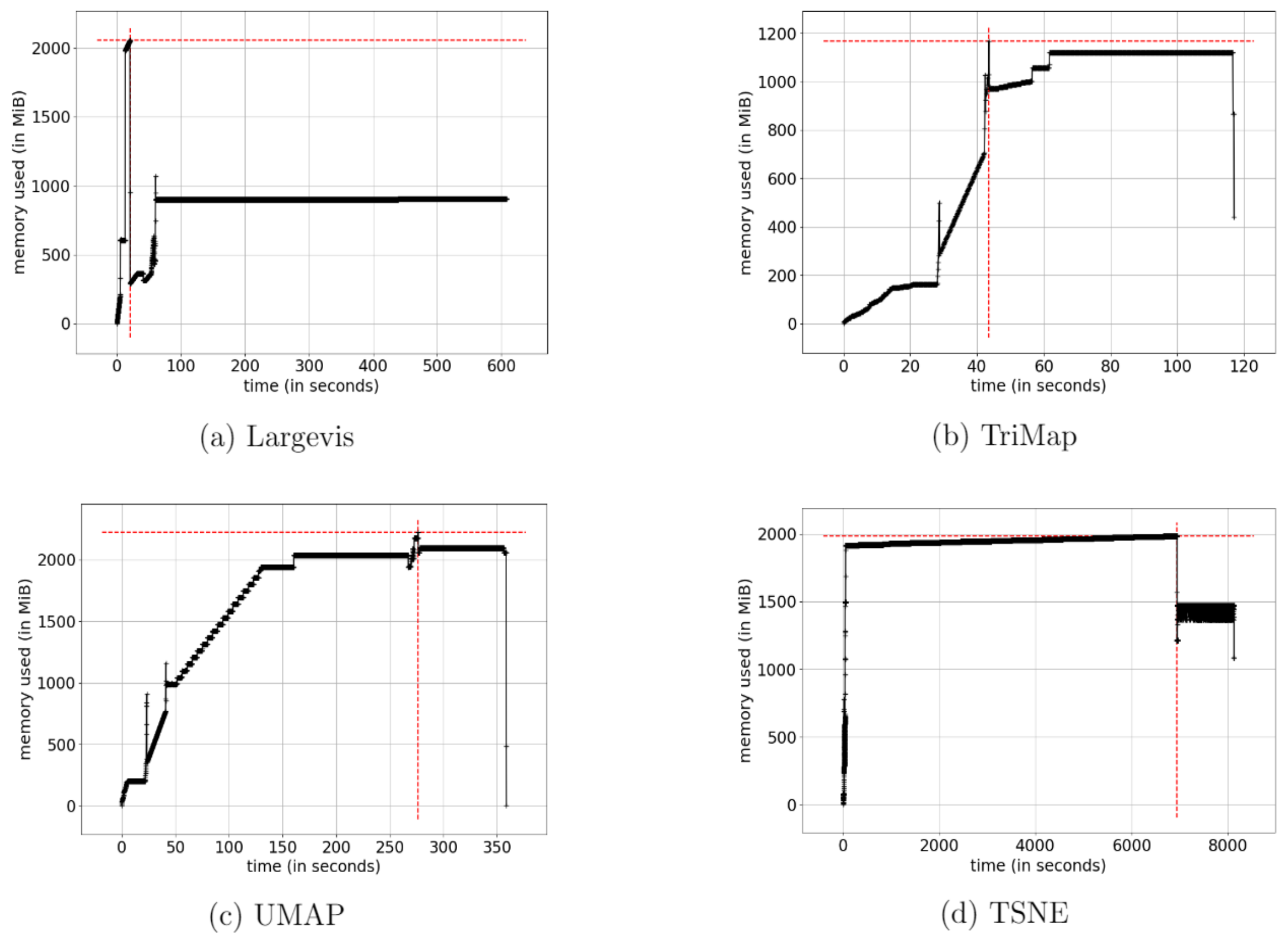
(a) Largevis

(b) TriMap

(c) UMAP

(d) TSNE

Fashion MNIST Memory Usage



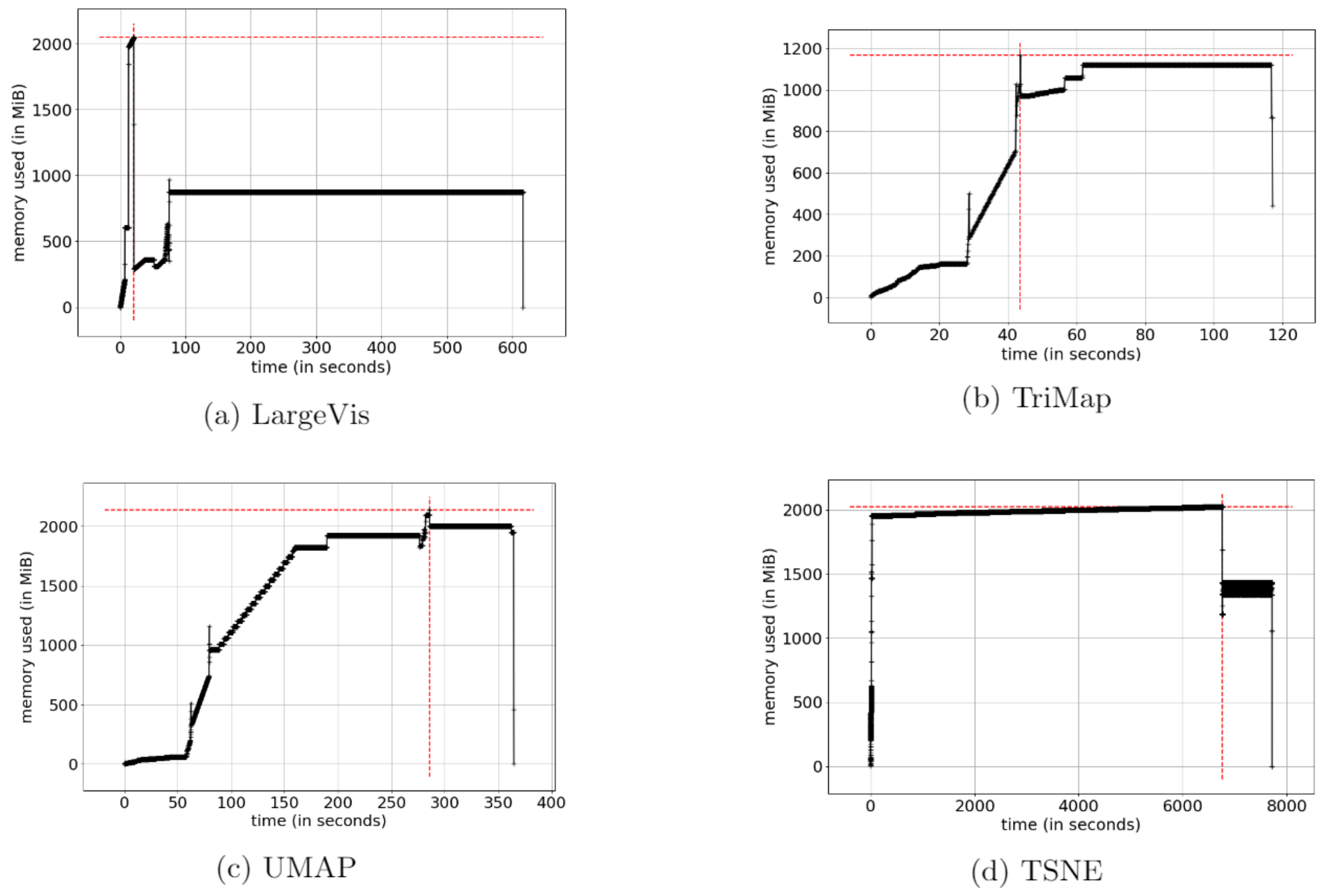
(a) Largevis

(b) TriMap

(c) UMAP

(d) TSNE

MNIST Digits Memory Usage



(a) LargeVis

(b) TriMap

(c) UMAP

(d) TSNE

Visualization Program Quality Measures

		TSNE	UMAP	LargeVis	Trimap
MNIST Digits	Silhouette	-0.04781	0.11363	0.05754	-0.01781
	Davies-Bouldin	2.92073	2.14496	78.92549	2.34602
Fashion MNIST	Silhouette	-0.11374	-0.04145	-0.11132	-0.00672
	Davies-Bouldin	6.38772	6.22085	11.15644	2.97314

Visualization Program Runtimes

	TSNE	UMAP	LargeVis	Trimap
MNIST Digits Actual	129:40.967	5:06.650	9:35.643	1:39.526
Fashion MNIST Actual	132:09.443	5:42.680	9:57.080	2:08.569
Theoretical	$O(N^2)$	$O(N^{1.14})$	$O(smN)$	$O(N)$

Conclusions

For TSNE in the Digits dataset, the scores, as compared to the others, are within a close margin of the others and reflect the fact that, on average, there exists some amount overlap of among the clusters, while still maintaining a very clear delineation between clusters. In the Fashion MNIST embeddings, we see similar qualities, with slightly less delineation among clusters than in the Digits data-sets. With this being said, the goal of mitigating the crowding-out problem of dimensional reduction was successful for t-SNE.

For LargeVis, we can see that in the Digits dataset, it provides very clear clustering among similar data points, however, fails to correctly delineate certain clusters clearly. With many clusters grouping together in the Digits dataset as well as the Fashion dataset, LargeVis has low performance in both when considering this metric.

For UMAP, we see that on the Digits dataset, it provides a very high-quality embedding, having the top Silhouette and David-Bouldin scores in this dataset. With a positive Silhouette score, this means that, on average, the points in the dataset are clustered correctly. Pairing this with the low David-Bouldin score, we have clearly delineated and accurate clustering. Moving to the Fashion MNIST dataset, we see that the quality of the embedding has decreased, with more overlap and less delineation between. Placing it at second best in this category, this more modern dataset proves to be more difficult to embed for UMAP.

For Trimap, we can see that it produces embeddings in both the MNIST Digits and Fashion MNIST data-sets that are quite different from all the others. In the Digits dataset, Trimap performs third best in Silhouette and second best in David-Bouldin scores. This puts it somewhere in the middle of the road in term of quality for this dataset, with accurate clustering and lower delineation. However, on the newer Fashion MNIST dataset, Trimap outperforms all other algorithms. With accurate clustering and quality delineation, this shows great promise for the use of Trimap overall.

References

Project GitHub: https://github.com/alexk101/dimensionality_reduction_measures

TSNE: https://lvdmaaten.github.io/tsne/

LargeVis: https://github.com/lferry007/LargeVis

UMAP: https://github.com/lmcinnes/umap

Trimap: https://github.com/eamid/trimap