Assignment 4 – Task 2 Saleh Alkhalifa
Summarize a Relevant Paper 11/20/2024

Objective:
Provide a comprehensive summary of Visualizing Data using t-SNE" by Laurens van der Maaten and Geoffrey Hinton.

1. **Summary:**

- **Outline the main objectives and contributions of the paper.**
  The paper introduces a technique known as t-distributed Stochastic Neighbor Embedding, also known as tSNE, is designed to visualize the high dimensional data of a given dataset by mapping it to lower dimensional spaces such as 2D or 3D. The paper aims to demonstrate three main objectives relating to this technique. First, the authors present the model as a method that overcomes the current (at the time) limitations of visualization techniques by addressing what is known as the "crowding problem" which was inherent in methods like SNE, or Stochastic Neighbor Embedding. This is done through modifications of the SNE model to create a symmetric version with simpler gradients. This results in making the optimization process more efficient. Another modification is using a heavy-tailed Student t-distribution in the low-dimensional space to compute similarities. This helps reduce the crowding problem by allowing moderate distances in high-dimensional space to be represented by larger distances in the map. Next, they aim to demonstrate the model can effectively preserve both the local and global structures in data. This is accomplished by revealing clusters at multiple scales which is needed for data based on several different low-dimensional manifolds. Finally, they authors provide evidence supporting the previous claims in which they show that tSNE produces better visualizations compared to other non-parametric techniques. A few comparisons that can be drawn are Sammon mapping, Isomaps, and Local Linear Embeddings.

- **Explain the t-SNE algorithm and how it differs from other dimensionality reduction techniques.**
  The tSNE model is a nonlinear model for dimensionality reduction that is designed to visualize high-dimensional data by mapping it into a lower-dimensional space. The algorithm works by converting the similarities between data points in the high-dimensional space into what are known as joint probabilities. This is done using a Gaussian distribution to model these similarities. The model then places points in the low-dimensional space. This is done so that the pairwise similarities are represented by a Student t-distribution with one degree of freedom. This is generally a heavy-tailed distribution which allows t-SNE to allocate more space between dissimilar data points, effectively addressing what is known as the "crowding problem" where high-dimensional data cannot be appropriately represented in lower dimensions. One interesting aspect here, or at least I found to be interesting, is that the algorithm minimizes what is known as the Kullback-Leibler divergence between the high-dimensional and low-dimensional joint probability distributions using gradient descent. This results in a map that preserves local structures while revealing meaningful global patterns. This model differs from other dimensionality reduction techniques like Principal Component Analysis (PCA) and Isomap in several key ways. While PCA is a linear method that focuses on maximizing variance and may miss nonlinear relationships, tSNE model captures complex nonlinear structures by preserving the probability distributions of local neighborhoods. Unlike other models like the Isomap and Locally Linear Embedding, also known as LLE, which can struggle and face difficulty with data lying on multiple manifolds or suffer from "short-circuiting", the tSNE model uses t-distribution which allows it to balance the preservation of local and global structures quite effectively. This results in clearer separation of clusters and the ability to reveal patterns at multiple scales, making tSNE particularly powerful for visualizing datasets like handwritten digits, gene expression profiles, or word embeddings where the data's intrinsic geometry is complex.

2. **Critical Analysis:**

- **Discuss the advantages and limitations of t-SNE as presented in the paper.**
  The tSNE model is represented in this paper by notable advantages by producing high-quality visualizations that preserve local data structures and reveal meaningful global patterns, such as clusters, which are often hidden in high-dimensional data. We can clearly see this in the code in the other portion of this assignment where the data was nicely visualized using this model. By utilizing a Student t-distribution with one degree of freedom in the low-dimensional space, t-SNE effectively addresses the crowding problem found in methods like SNE. It does so by allowing for better separation of dissimilar data points and clearer cluster formation. Its symmetric formulation and simplified gradient computation enhance optimization efficiency and reduce susceptibility to poor local

minima. That said, the t-SNE model also has limitations, including a computational complexity of O(n^2) that challenges its scalability to very large datasets without modifications. I have seen this in past in industry when trying to apply this model larger datasets. The non-convex nature of its cost function can lead to convergence on different local minima based on initialization, resulting in variability across runs. Additionally, t-SNE's performance can be sensitive to hyperparameters like perplexity and learning rate. This requires careful tuning as I demonstrated in the code, and the global distances in the resulting map may not correspond accurately to those in the original high-dimensional space. This ultimately affects the interpretation of inter-cluster relationships.

- **Relate the concepts from the paper to the t-SNE technique used in the practical component:**
  In the practical component, we applied t-SNE, PCA, and Autoencoders to the MNIST dataset to reduce its dimensionality to 2D and visualize the handwritten digits. The concepts from the paper were directly reflected in how t-SNE was employed and how it differed from the other techniques. Specifically, t-SNE's focus on preserving local neighborhoods by converting high-dimensional similarities into joint probabilities using a Gaussian kernel, and then mapping these to a low-dimensional space using a Student t-distribution, resulted in clear separation of digit clusters in the 2D visualization. This aligns with the paper's discussion on t-SNE's ability to handle the "crowding problem" and preserve both local and global data structures. In contrast, PCA, being a linear method, captured the largest variance but failed to separate the digits effectively due to its inability to model nonlinear relationships, leading to overlapping clusters. Autoencoders, while capable of learning nonlinear embeddings, did not specifically optimize for preserving the probability distributions of local neighborhoods, and thus their 2D representations were not as distinctly clustered as those produced by t-SNE. The practical use of t-SNE on the MNIST dataset showcased the algorithm's strengths highlighted in the paper, such as its simplified gradient optimization and effective handling of high-dimensional data manifolds, resulting in superior visualization of the handwritten digits compared to PCA and Autoencoders.

3. **Applications:**

- **Highlight potential applications of t-SNE beyond the scope of the assignment:**
  The tSNE model has extensive applications in biotechnology and pharmaceuticals, particularly for visualizing and analyzing high-dimensional biological data. Going beyond the scope of our application of this model with the MNIST data, it can be used in single-cell RNA sequencing to identify cell types, states, and developmental trajectories by providing insights into tissue heterogeneity and disease mechanisms. In the field of drug discovery tSNE facilitates chemical space mapping, helping researchers to identify novel compound scaffolds, predict molecular side effects, and analyze high-throughput screening data. Interestingly, personalized medicine benefits from t-SNE’s ability to stratify patients by clustering clinical and multi-omics data that can lead to tailored treatment strategies. Aside from pharmaceuticals, tSNE is essential in other industries: for example in natural language processing, it visualizes word embeddings to explore semantic relationships. In computer vision, it helps analyze learned features from neural networks.

- **Provide examples of how t-SNE has been utilized in other research or industries:**
  In pharma R&D, t-SNE has been applied in advancing cancer biology, where it has been used to analyze single-cell RNA sequencing data to identify immune cell populations within tumor microenvironments, helping the development of targeted immunotherapies. Pharmaceutical companies, especially here in the Boston area, use t-SNE to visualize chemical libraries by grouping compounds by structural or functional similarity to accelerate drug discovery efforts. In healthcare analytics, t-SNE has been used to cluster patient populations based on genetic and clinical data, helping uncover disease subtypes and inform precision medicine approaches. Outside life sciences, t-SNE has been applied in natural language processing to visualize word embeddings such as Word2Vec, uncovering relationships like synonyms or analogies. In cybersecurity it has been used to detect anomalies in network traffic, differentiating between normal and suspicious patterns, as showed by AWS in their yearly summits.

  - "Visualizing Data using t-SNE" by Laurens van der Maaten and Geoffrey Hinton (Journal of Machine Learning Research, 2008)
  - Curse of Dimensionality, t-SNE and Kullback-Leibler Divergence, Johannes Otterbach.