Optimization Techniques for t-SNE Visualizations

Liam Brinker

Supervisors: Prof. Ofra Amir and Dr. Nir Rosenfeld

Outline

1. Introduction

2. Attempts

3. Conclusion

Dimensionality Reduction

- Transformation of high dimensional data $X \in \mathbb{R}^{n \times p}$ into low dimensional data $Y \in \mathbb{R}^{n \times q}$, where $q < p$, while preserving much of the significant structure of X.
- Linear vs. Non-Linear transformations.
- Popular techniques: PCA, UMAP, and more.
- Our focus for today: t-SNE.

What Can We Expect For

- Question: can we always preserve distance in the low dimensional space?
- Answer: no.
- Any 4 equidistant points in 3-d space (tetrahedron) can't be transformed into 4 - equidistant points in 2-d space.

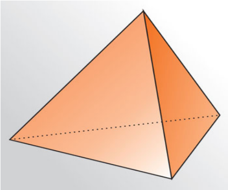


Figure: Tetrahedron

Stochastic Neighbor Embedding (SNE)

- Stochastic Neighbor Embedding (SNE) is a (non-linear) technique for dimensionality reduction.
- Aims to preserve local structure.
- Result is given by minimizing an objective function defined over distributions P, Q.

Mathematical Formulation of SNE

Step 1: Compute Pairwise Similarities

In the high-dimensional space, we define the conditional probability of two points $\mathbf{x}_i$ and $\mathbf{x}_j$ using a Gaussian distribution:

$$p_{j|i} = \frac{\exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_k\|^2}{2\sigma_i^2}\right)}$$

where σ_i is the variance of the Gaussian centered at $\mathbf{x}_i$.

Similarly, we define the joint distribution of two points $\mathbf{y}_i$ and $\mathbf{y}_j$ in the low-dimensional space:

$$q_{ij} = \frac{\exp\left(-\|\mathbf{y}_i - \mathbf{y}_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|\mathbf{y}_i - \mathbf{y}_k\|^2\right)}$$

Mathematical Formulation of SNE

Step 2: Minimize the Kullback-Leibler Divergence
The objective is to minimize the divergence between the high-dimensional probabilities p_{ij} and the low-dimensional probabilities q_{ij}:

$$\mathrm{KL}(P\|Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

Remarks:

- The original version uses conditional probabilities instead. This version is called symmetric SNE.
- Notice that σ_i are predetermined hyper-parameters.

From SNE to t-SNE

- Empirically proven to preserve global structure as well.
- The key idea is to redefine the low-dimensional distribution using a heavy-tailed Student t-distribution with one degree of freedom.
- Using this distribution, the joint probabilities q_{ij} are defined as:

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}}$$

- The gradient can be intuitively interpreted as a repulsive force between points y_i and y_j:

$$\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1}$$

where C is the cost function.

Improvements

Various methods have been empirically shown to improve the results:

- Introducing momentum term with adaptive learning rate scheme in the optimization of the cost function.
- Early Compression.
- Early Exaggeration.

The Problem

- Convergence to global minimum is not guaranteed.
- Results are sensitive to the initial points.

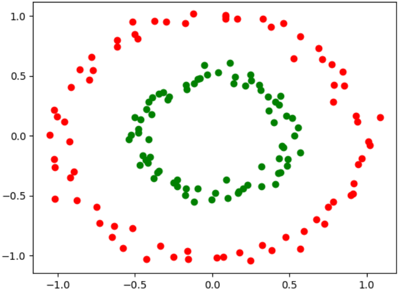


Figure: Original Dataset

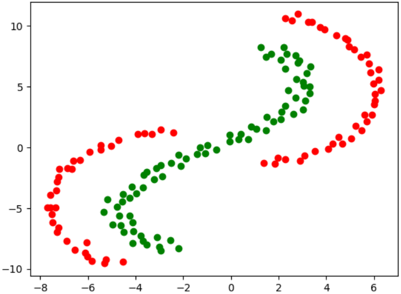


Figure: t-SNE Visualization for Bad Initialization

Outline

1. Introduction

2. Attempts

3. Conclusion

The Approach

- Train a model to learn the distribution of "good" points.
- Finding metrics to assess the quality and meaningfulness of t-SNE visualizations.
- Starting with three simple dataset classes:
    - Circles
    - Blobs
    - Clusters
- Defining good visualizations is easy.

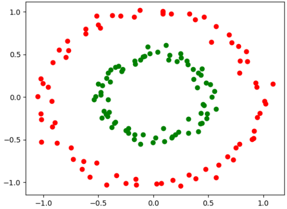


Figure: Circles

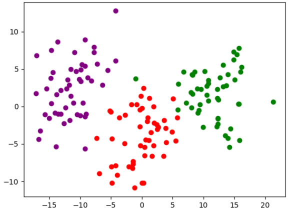


Figure: Blobs

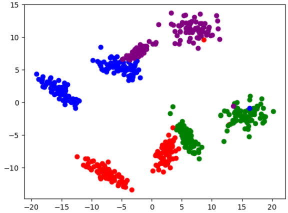


Figure: Clusters

Loss Function Landscape

- On an intuitive level, we want to "learn" the loss landscape.

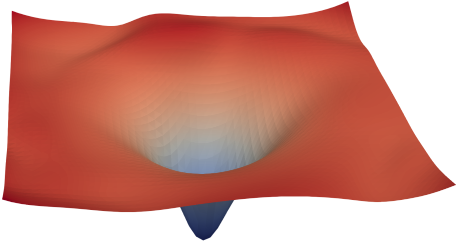


Figure: Easy to Learn Landscape

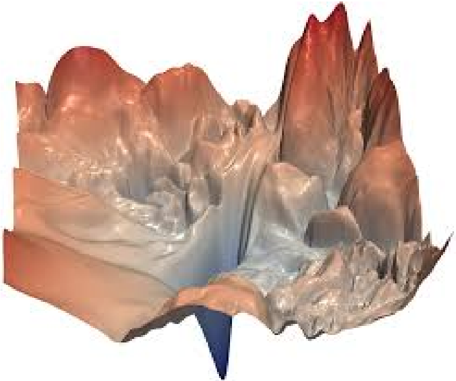


Figure: Complicated Landscape

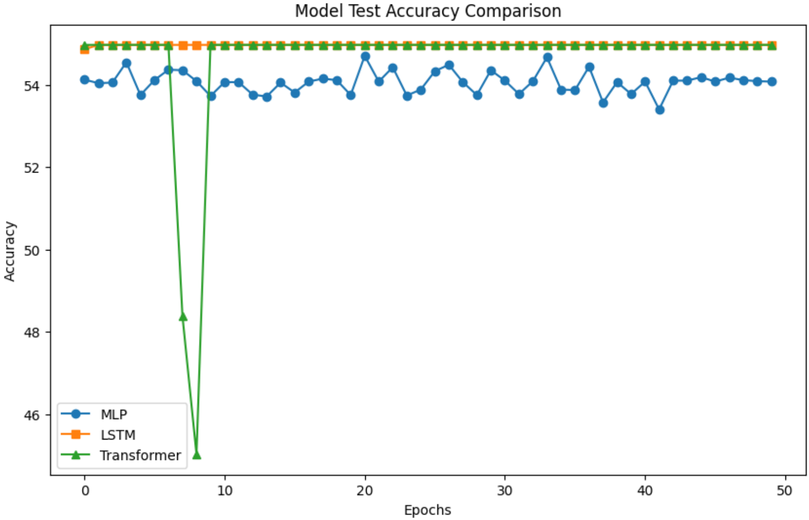
First Attempt - Setting

- For now, focus on dimensionality reduction for the same dimension.
- Good results should replicate original data.
- Easy to check.

First Attempt - Setting

- Working with 2-d circles dataset instance.
- Dataset of 40,000 examples of initial points.
- Points sampled from isotropic Gaussian.
- Deterministic and exact gradient decent for t-SNE objective function.
- Automated labeling done via db-scan and KL-divergence values.
- No class imbalance.

First Attempt - Results



Further Work in This Area

- Trying different architectures, hyper-parameters and optimization schemes.
- Increasing the dataset size to 200,000 examples.
- Generating new dataset of points sampled from a distribution with lower variance.
- Included knowledge of high dimensional points and their probabilities caused a minor improvement.

Reformulating The Objective

- t-SNE is a specific instance of the family of dimensionality reduction methods known as Neighbors Embedding (NE).
- Methods differ in the choice of similarities and objective functions.
- Under a change of the high-dim kernel (similarity), the objective function of t-SNE can be written as an instance of Maximum Likelihood Estimation (MLE).

Notations for MLE Formulation

- $X \in \mathbb{R}^{n \times 2}$: Parameters in low-dimensional space (embedding coordinates)
- P: High-dimensional distribution defined over **pairs of points** (i, j)
- Q: Low-dimensional distribution
- $S(x_i, x_j)$: Similarity measure in low-dimensional space
- $Z(X)$: Partition function (normalizing factor for the low-dimensional distribution).

Objective Function as MLE

Using the notations we can write the loss function:

$$L(X) = -\mathbb{E}_{(i,j)\sim P}(log(Q_X(i,j))) = -\sum_{i\neq j\in[n]} P_{i,j} \cdot log(Q_X(i,j))$$

$$= -\sum_{i\neq j\in[n]} P_{i,j} \cdot log\left(\frac{S(x_i,x_j)}{\sum_{k\neq l\in[n]} S(x_k,x_l)}\right)$$

$$= -\sum_{i\neq j\in[n]} \Big(P_{i,j} \cdot log(S(x_i,x_j))\Big) + log(Z(X))$$

Insights

This formulation allows us to implement computationally efficient tricks and derive new methods:

- Variants of MLE such as: regularized MLE, weighted MLE.
- Approximation methods for dealing with the partition function.
- Reduction to Noise Contrastive Estimation (NCE) which is a supervised learning task.

NC-t-SNE Objective Function

In NC-t-SNE, we aim to distinguish data samples from noise using the following objective function:

$$L = -\sum_{i} \sum_{j \neq i} [P(i,j) \log Pr(D = 1 \mid i,j) + mP_{noise}(i,j) \log Pr(D = 0 \mid i,j)]$$

Where:

$$Pr(D = 1 \mid i,j) = \frac{\tilde{q}_{\mathbf{x}}(i,j)}{\tilde{q}_{\mathbf{x}}(i,j) + mP_{noise}(i,j)}$$

$$Pr(D = 0 \mid i,j) = \frac{mP_{noise}(i,j)}{\tilde{q}_{\mathbf{x}}(i,j) + mP_{noise}(i,j)}$$

Theorem

Theorem (Gutmann & Hyvarinen, 2010; 2012)

Let ξ have full support and suppose there exists some θ^* such that $q_{\theta^*} = p$. Then θ^* is a minimum of

$$L_{NCE}(\theta) = -\mathbb{E}_{s\sim p} \log\left(\frac{q_\theta(s)}{q_\theta(s) + m\xi(s)}\right) - m\mathbb{E}_{t\sim\xi} \log\left(1 - \frac{q_\theta(t)}{q_\theta(t) + m\xi(t)}\right)$$

and the only other extrema of L_{NCE} are minima $\tilde{\theta}$ which also satisfy $q_{\tilde{\theta}} = p$.

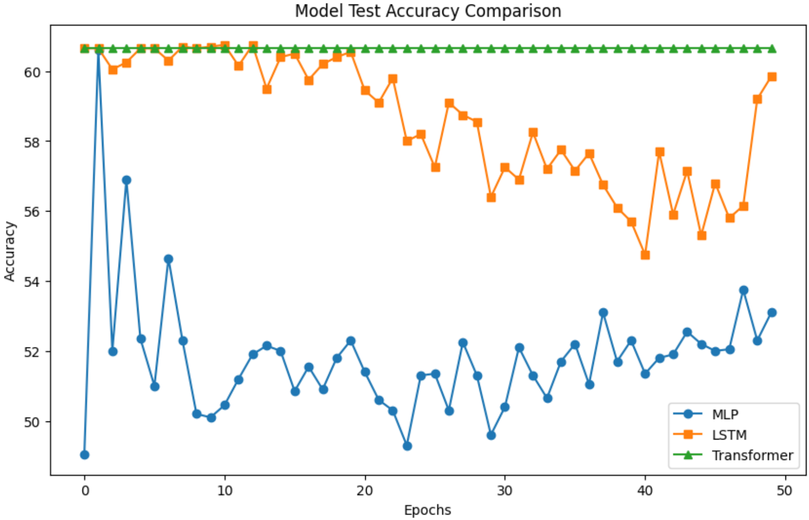
Second Attempt

- Theorem ensures convergence of NC-t-SNE to global minimum on instances where t-SNE failed.
- What can we say on instances where theorem doesn't apply?
- Conduct a similar analysis for NC-t-SNE.

Second Attempt - Setting

- Working with 3-d blobs dataset instance.
- Automated labeling determined by silhouette coefficient values.
- The rest of the setting stays the same...

Second Attempt - Results



Outline

1. Introduction

2. Attempts

3. Conclusion

Discussion Points

- Original problem remains unsolved.
- Objective function landscape is too hard to learn even in well structured instances.
- Method lacks explainability.
- Addressing the problem from different perspective:
  - Finding a more suitable optimization scheme than gradient descent for this specific class of objective functions.

Questions?

Thank you for your attention!