t-SNE: A Guide

1 Overview

t-SNE (t-distributed Stochastic Neighbor Embedding) is a nonlinear dimensionality reduction technique for embedding high-dimensional data in a low-dimensional space (typically 2D or 3D) for visualization. It focuses on preserving local structure.

2 High-Dimensional Similarities

For each data point $\mathbf{x}_i$, define conditional probabilities:

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

These describe how likely point $\mathbf{x}_j$ is to be a neighbor of $\mathbf{x}_i$, based on a Gaussian centered at $\mathbf{x}_i$.

Perplexity

$$\text{Perp}(P_i) = 2^{H(P_i)}, \quad \text{where} \quad H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}$$

A binary search over σ_i ensures the perplexity matches a user-defined target (e.g., 30).
Note: $H(P_i)$ is the Shannon entropy.

Symmetrization

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

3 Low-Dimensional Similarities

Define joint probabilities q_{ij} in 2D/3D space:

$$q_{ij} = \frac{\left(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2\right)^{-1}}$$

This uses a Student-t distribution with 1 degree of freedom (Cauchy) to allow heavy tails.

4 Loss Function

KL divergence between high- and low-dimensional similarities:

$$\mathcal{L} = \sum_{i \neq j} p_{ij} \log \left(\frac{p_{ij}}{q_{ij}} \right)$$

This emphasizes preserving local structure. Discrepancies between distant pairs are penalized less.

5 Optimization

Initialize $\mathbf{y}_i \in \mathbb{R}^2$ randomly. Then perform gradient descent with momentum (which helps smooth convergence in a non-convex loss landscape):

$$\Delta \mathbf{y}_i^{(t)} = \eta \cdot \nabla_i \mathcal{L} + \alpha \cdot \Delta \mathbf{y}_i^{(t-1)}$$
$$\mathbf{y}_i^{(t+1)} = \mathbf{y}_i^{(t)} + \Delta \mathbf{y}_i^{(t)}$$

Gradient

$$\nabla_i \mathcal{L} = 4 \sum_{j \neq i} (p_{ij} - q_{ij}) \cdot \frac{(\mathbf{y}_i - \mathbf{y}_j)}{1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2}$$

Early exaggeration

Multiply p_{ij} by γ (e.g. 4 or 12) during early iterations (e.g., first 250) to strengthen attractive forces and help cluster formation.

6 Practical Notes

- Good for visualizing clusters and manifold unfolding
- Not suitable for preserving global distances
- Highly sensitive to perplexity and initialization

7 Common Applications

- Bioinformatics (e.g., single-cell RNA-seq)
- Visualizing CNN embeddings
- Exploring latent space in NLP
- ML model debugging and interpretability