Implementation of t-SNE and its comparison with other Dimensionality Reduction Models

Team Kaaju
Aaradhya Gupta (2019114010)
Akhilesh Aravapalli (2019114016)
Chayan Kochar (2019114008)
Jayant Panwar (2019114013)

Github Link: https://github.com/ChayanK2000/SMAI_Project

Problem Statement

Dimensionality reduction is a very important aspect of visualization of data and intends to transform data from a higher dimensional space to a lower dimensional space while retaining meaningful properties and features of the original data. Reducing the dimensions of data is very important as analyzing high dimensional data is restricted to availability of computational resources. Moreover, dimension reductionality finds its use in many areas today like signal processing, speech recognition, bioinformatics, etc.

There are various dimensionality reduction models that are able to reduce the dimensions of the given data efficiently and some of them include: Sammon mapping, Isomap, Locally Linear Embedding, and t-SNE. Our problem statement will be to implement each of these four dimensionality reduction models and assess each of them against the latest of the batch, i.e., t-SNE model. t-SNE model will be implemented from scratch whereas library implementations for other models will be used for comparative study.

Goals and Approach

Visualization of high-dimensional data is an important problem in many different domains, and deals with data of widely varying dimensionality. Important techniques of such visualization include iconographic displays such as Chernoff faces (Chernoff, 1973), pixel-based techniques (Keim, 2000), and techniques that represent the dimensions in the data as vertices in a graph (Battista et al., 1994). Most of these techniques simply provide tools to display more than two data dimensions, and leave the interpretation of the data to the human observer. This severely limits the applicability of these techniques to real-world data sets that contain thousands of high-dimensional data points.

In contrast to the visualization techniques, dimensionality reduction methods convert the high-dimensional data set X = {x 1 , x 2 , ..., x n } into two or three-dimensional data Y = {y 1 , y 2 , ..., y n } that can be displayed in a scatterplot.
The aim of dimensionality reduction is to preserve as much of the significant structure of the high-dimensional data as possible in the low-dimensional map.

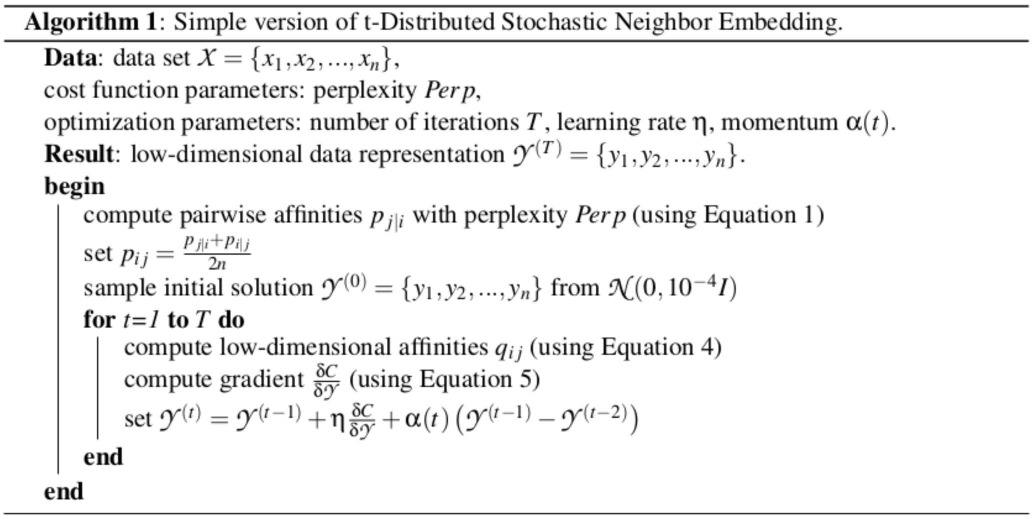
So in this project, we work on the way of converting a high-dimensional data set into a matrix of pairwise similarities as described in the paper introducing a technique, called "t-SNE", for visualizing the resulting similarity data. t-SNE is capable of capturing much of the local structure of the high-dimensional data very well, while also revealing global structure such as the presence of clusters at several scales.

Our Approach for implementing the t-SNE can broken roughly as:

1. Implementation of SNE
    - Involves getting a conditional probability such that one data point x_i picks x_j as its neighbour denoted by $p_{j|i}$
    - For the low-dimensional counterparts y_i and y_j of the high-dimensional data points x_i and x_j , it is possible to compute a similar conditional probability, which we denote by $q_{j|i}$
    - Getting appropriate Cost function and Perplexity and minimizing Cost by gradient descent.
2. Problems and disadvantages with it.
3. Overcoming them using the modified version of SNE ,i.e., t-SNE

The cost function used by t-SNE differs from the one used by SNE in two ways: (1) it uses a symmetrized version of the SNE cost function with simpler gradients that was briefly introduced by Cook et al. (2007) and (2) it uses a Student-t distribution rather than a Gaussian to compute the similarity between two points in the low-dimensional space.
t-SNE employs a heavy-tailed distribution in the low-dimensional space to alleviate both the crowding problem and the optimization problems of SNE.

The basic algorithm for t-SNE can be noted as:



```
Algorithm 1: Simple version of t-Distributed Stochastic Neighbor Embedding.
  Data: data set $\mathcal{X} = \{x_1, x_2, ..., x_n\}$,
  cost function parameters: perplexity $Perp$,
  optimization parameters: number of iterations $T$, learning rate $\eta$, momentum $\alpha(t)$.
  Result: low-dimensional data representation $\mathcal{Y}^{(T)} = \{y_1, y_2, ..., y_n\}$.
  begin
      compute pairwise affinities $p_{j|i}$ with perplexity $Perp$ (using Equation 1)
      set $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$
      sample initial solution $\mathcal{Y}^{(0)} = \{y_1, y_2, ..., y_n\}$ from $\mathcal{N}(0, 10^{-4}I)$
      for $t=1$ to $T$ do
          compute low-dimensional affinities $q_{ij}$ (using Equation 4)
          compute gradient $\frac{\delta C}{\delta \mathcal{Y}}$ (using Equation 5)
          set $\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\delta C}{\delta \mathcal{Y}} + \alpha(t) \left( \mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)} \right)$
      end
  end
```

Dataset

The dataset which we plan to use is the MNIST dataset which contains the grayscale images of handwritten images. In case, we face any problem with the dataset, we will opt for one of the following datasets:

1. Olivetti data set,
2. COIL-20 data set,
3. Word-features data set,
4. Netflix data set

Expected Deliverables

We hope to submit the following deliverables with regard to the project:

1. A working model of t-SNE implemented from scratch.
2. Working models of Sammon mapping, Isomap, and Locally Linear Embedding.
3. A comparative study of t-SNE against: Sammon mapping, Isomap, and Locally Linear Embedding in the form of a report.
4. Presentation of the summary of implementation and comparative study

Work Distribution

- Akhilesh and Jayant : Working on t-SNE
- Aaradhya and Chayan: Sammon mapping, Isomap, and Locally Linear Embedding
- Together: Comparative Study and final presentation.

Timeline

We hope to complete the project in 2 major work runs:

1. Mid-evaluation: Aim to complete the basic implementation of t-SNE model
2. Final-evaluation: Make any necessary changes required for correct implementation of t-SNE and to complete the comparative study of t-SNE vs other models

References

- L. van der Maaten, and G. Hinton: Visualizing Data using t-SNE. *Journal of Machine Learning Research* (*2008*)
- J.A. Lee and M. Verleysen: Nonlinear dimensionality reduction. *Springer, New York, NY, USA, 2007*
- S.T. Roweis and L.K. Saul: Nonlinear dimensionality reduction by Locally Linear Embedding. *Science, 290(5500):2323–2326, 2000.*
- L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik: Dimensionality reduction: A comparative review. *Online Preprint, 2008.*