Random partition t-SNE

Szu-Han Lin

Academia Sinica

Abstract

Data visualization has been recognized as an important tool in exploring the heterogeneity of high dimensional data, for which many statistical methods have been used including Principal Component Analysis (PCA), Multidimensional Scaling (MDS) and Laplacian Eigenmaps (2003, Belkin and Niyogi). Along the development of data visualization methods, t-Distributed Stochastic Neighbor Embedding (tSNE) proposed in 2008 by Laurens van der Maaten and Geoffrey Hinton has been the first to successfully separate the 10 digit groups of MNIST data set into 10 clusters, yielding a total of 5997 citations to date. There are two key features in tSNE for its success: 1) it transforms the similarity matrix into a distribution (e.g. Gaussian distribution and t-distribution) for both the input data of high dimension and the visualization in two dimensions; 2) it minimizes the KL divergence between these two distributions by the gradient descent algorithm. However, as the data volume becomes huge, the computation for similarity matrix becomes a burden and the application faces an overwhelming barrier. In this talk, we propose a random partition algorithm for t-SNE, which we name RP-tSNE, to accommodate large volume data sets for data visualization. In addition to providing a proof for the consistency of RP-tSNE, we will demonstrate the usage of RP-tSNE on a cryo-electron microscopy image data set with 103,363 images.

This is a joint work with Ting-Li Chen and I-Ping Tu.