Martin Skrodzki

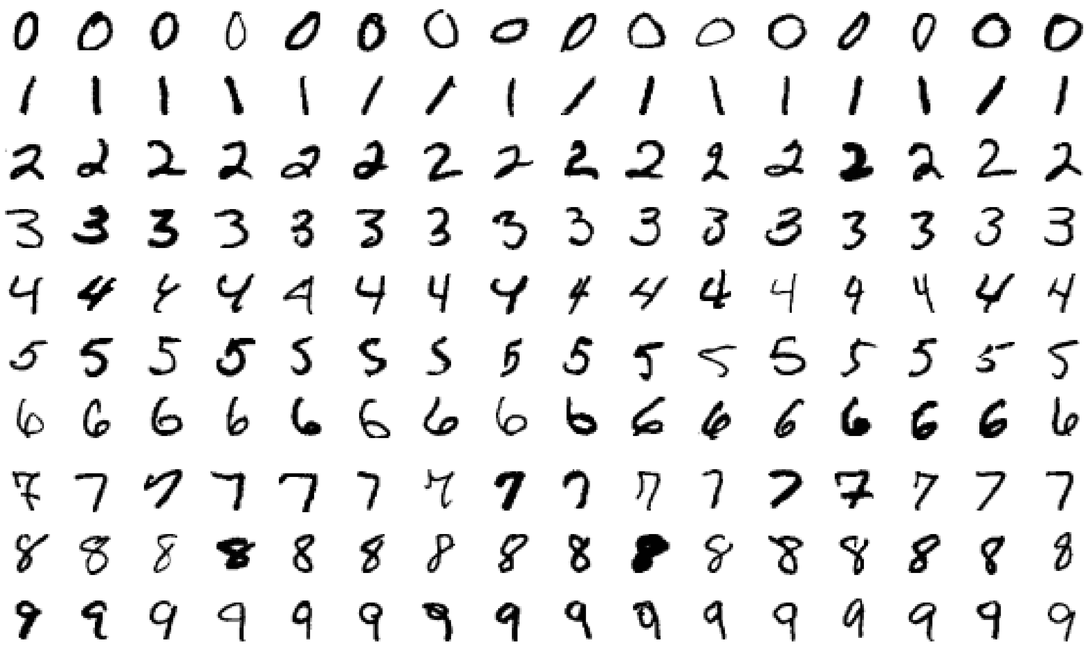
Computer Graphics and Visualization, TU Delft

Contact: m.skrodzki@tudelft.nl

Explorative Analysis via t-SNE Visualization of High-Dimensional Data

Data Example: MNIST

The MNIST data consists of 70,000 images with gray scale, anti-aliased 28×28 pixel images, i.e., elements in $\mathbb{R}^{784}$, image by [1].



Goal

Provide a two-dimensional visualization of the data in a way such that:

1. Points that are close in the high-dimensional space are close in the two-dimensional embedding.
2. Points in the two-dimensional embedding are sufficiently far away from each other (no "crowding").

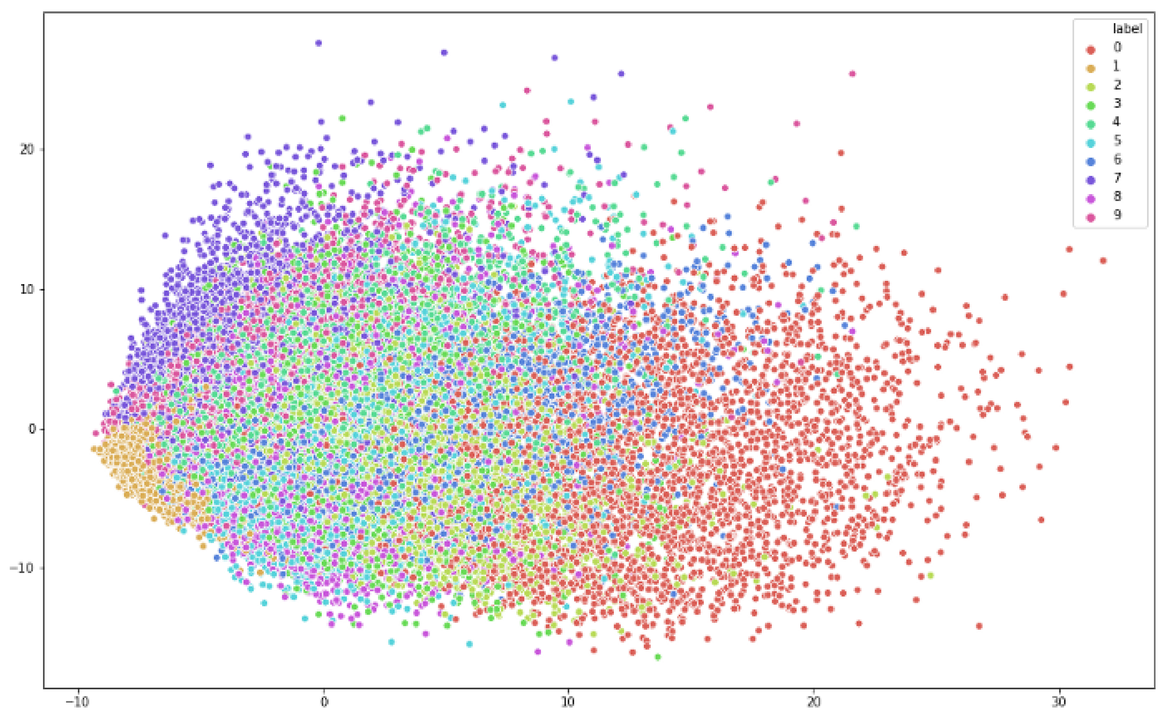
Method

Take the similarities $\boldsymbol{P}$ of the high-dimensional points and find similarities $\boldsymbol{Q}$ of the two-dimensional points such that

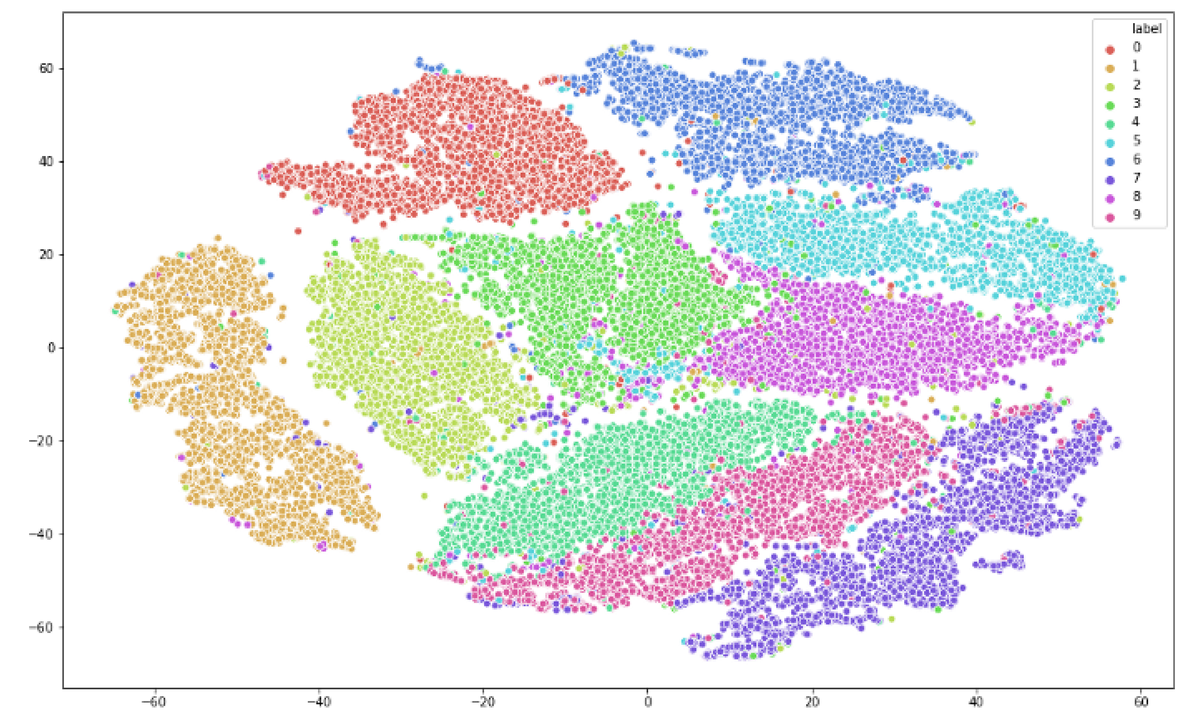
$$\boldsymbol{KL}(\boldsymbol{P}||\boldsymbol{Q}) = \sum_i \sum_{j \neq i} \boldsymbol{p}_{i|j} \log\left(\frac{\boldsymbol{p}_{i|j}}{\boldsymbol{q}_{ij}}\right)$$

is minimized [2].

Embedding the MNIST data via PCA [3]



Embedding the MNIST data via t-SNE [3]



Definition of the Similarities

If we choose

$$\boldsymbol{p}_{j|i} = \frac{\exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 / 2\boldsymbol{\sigma}_i^2)}{\sum_{j' \neq j} \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_{j'}\|^2 / 2\boldsymbol{\sigma}_i^2)},$$

$$\boldsymbol{q}_{ij} = \frac{(1 + \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2)^{-1}}{\sum_k \sum_{\ell \neq k} (1 + \|\boldsymbol{y}_k - \boldsymbol{y}_\ell\|^2)^{-1}}$$

for some specific bandwidth $\boldsymbol{\sigma}_i$, then:

- large $\boldsymbol{p}_{i|j}$ and small $\boldsymbol{q}_{ij}$ give big penalty,
- small $\boldsymbol{p}_{i|j}$ and large $\boldsymbol{q}_{ij}$ give small penalty.

Minimization

Perform gradient descent on

$$\frac{\boldsymbol{KL}(\boldsymbol{P}||\boldsymbol{Q})}{\boldsymbol{\partial y}_i} = 4 \sum_{j \neq i} \boldsymbol{p}_{i|j} \boldsymbol{q}_{ij} \boldsymbol{Z}(\boldsymbol{y}_i - \boldsymbol{y}_j) - 4 \sum_{j \neq i} \boldsymbol{q}_{ij}^2 \boldsymbol{Z}(\boldsymbol{y}_i - \boldsymbol{y}_j),$$

where $\boldsymbol{Z} = \sum_{k \neq \ell} (1 + \|\boldsymbol{y}_k - \boldsymbol{y}_\ell\|^2)^{-1}$ normalizes and

- the first term represents attractive forces, while
- the second term represents repulsive forces.

One exact evaluation of the gradient takes $\mathcal{O}(n^2)$ steps for n points. Thus, utilize acceleration data structures.

Project Questions:

- Development of multi-grid approaches to accelerate the optimization.
- Investigation of alternative embeddings spaces, e.g., hyperbolic space.
- Derivation of corresponding acceleration structures.
- Optimized user-interaction with the embeddings.

References

[1] Steppan, L. (2017). A few samples from the MNIST test dataset. Wikimedia commons.

[2] Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(11).

[3] Zhang, D. (2020). Dimensionality Reduction using t-Distributed Stochastic Neighbor Embedding (t-SNE) on the MNIST Dataset. Towards Data Science.