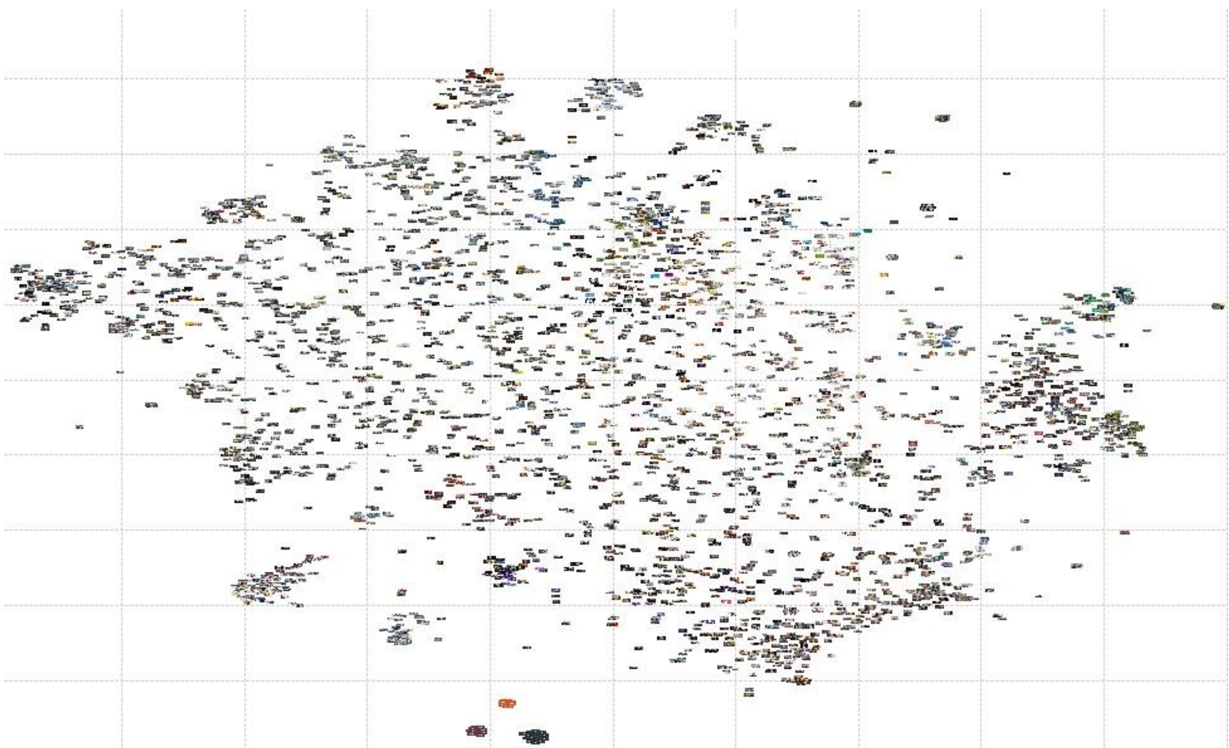
T-SNE

Radek Bartyzal
Let's talk ML in Prague
8. 12. 2016
(rbartyzal1@gmail.com)

What is it?

- t-Distributed Stochastic Neighbor Embedding
- dimensionality reduction technique



Stochastic Neighbor Embedding

1. convert the high-dimensional Euclidean distances between datapoints into conditional probabilities

= similarity of x_i to x_j = cond. prob. that x_j would be picked from all other points picked in proportion to their probability density under a Gaussian at x_i

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$

Stochastic Neighbor Embedding

2. calculate the same cond. prob. from distances of points in low-dimensional space

$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i - y_k\|^2\right)}$$

Stochastic Neighbor Embedding

3. if simialrity of y_i and y_j corresponds to similarity of x_i and x_j then their cond. prob. would be also similar

= we are trying to minimize the mismatch between them = minimize Kullback-Leibler divergences over all datapoints using a gradient descent method

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

t-SNE

- calculate the cond. prob. as a Student t-distribution with one degree of freedom

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

Why Student t-distribution?

- large clusters of points that are far apart interact in the same way as individual points
- much faster than Gaussian (no exp)
- t-SNE gradient more strongly repels dissimilar datapoints that are modeled by a small pairwise distance in the low-dimensional representation

Gradient descent

- equivalent to springs between y_i and y_j whose force is proportional to its length $(y_i - y_j)$, and also its stiffness = mismatch =

  $(p_{j|i} – q_{j|i} + p_{i|j} – q_{i|j})$

Optimization tricks

- early compression

= force the map points to stay close together at the start of the optimization

=> it is easy for clusters to move through one another so it is much easier to explore the space of possible global organizations of the data

Optimization tricks

- early exaggeration

= multiply all of the p_{ij}'s

=> q_{ij}'s are relatively smaller

=> large p_{ij}'s modeled by large q_{ij}'s

=> natural clusters in the data tend to form tight, widely separated clusters in the map

=> a lot of empty space in the map, which makes it much easier for the clusters to move around

Sources

- http://jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf
- https://lvdmaaten.github.io/tsne/