Executive summary

Objective of the study

The purpose of the report is to investigate the behavior of t-Distributed Stochastic Neighbor Embedding (t-SNE) projections on a couple of image datsets viewed through the encodings perspective of some popular CNN architectures. We are aiming to highlight the role of this dimensionality reduction technique in data analysis and computer vision, and reveal potential uses, benefits and interpretations of results provided by such a method.

Proposed approach

Three datasets were selected as a study support:

- Fashion MNIST, due to its size and curated, balanced structure
- Muffin vs Chihuahua, a dataset of raw scrapped images from the internet, chosen for the diversity of visual contexts
- A Large Scale Fish Dataset, which contains preaugmented images taken in a controlled environment, featuring increased visual similarity

The experimental setup contains of an autoencoder trained on the identity function, a pretrained VGG-16 and a simple CNN softmax classifier network. After optimizing the trainable models, t-SNE plots are created from the latent space of the autoencoder, respectively the last feautre maps activations from the CNN models. In the case of the classfier, separate projections are created from train and validation sets. Each t-SNE projection is numerically evaluated with stress metrics and the Shepard goodness score along with the Shepard diagram.

Conclusion & futher directions

Analysis of the results concluded that t-SNE performs great at estimating the complexity of a dataset and finding common structural occurences in images, while it struggles to properly understand the presence of classes in contextual variety, leading to cluster fragmentation and amalgamation of clusters of different provenience. It can also pin-point potential defects in a dataset, such as duplicate samples (visible in the Shepard diagram) or the reason for overfitting, like a different distribution appearing in the validation set which the model is not able to generalize.

Improving the current study is possible by involving more datasets in the analysis, emplying local and per-cluster metrics as well as querying other projection algorithms (MDS, PCA, UMAP) looking for the same observations or others that currently escaped our vigilance. A drawback of the current method is the high computational requirements scaled to the size of the dataset. Furtherly, improving the actual model architectures and embeddings would also provide an idea of how a very good, easily separable embeddings projection would look like.