Visualization Tool to Understand the Learning of a Deep Network

J. Sun, I. Kastanis

Visualizing the learning process of a deep network is of great value to interpret this novel technique in industrial applications. The t-Stochastic Neighbor Embedding (t-SNE) technique is utilized to display graphically the learning data as well as the activity of the hidden layers of a trained deep network. The visualization shows how the deep network learns the natural cluster characteristics of the training data.

Deep learning has achieved success in various application domains, such as image classification and speech recognition. As a pioneer in applying deep learning algorithms in industrial automation, CSEM developed Tileye for image recognition and surface inspection tasks as well as Tilear for precision-machinery quality inspection. While these systems performed outstandingly, it remained challenging to explain their working principles concisely to new customers. A deep network assembles diverse non-linear functions with millions of parameters, which are difficult to interpret, remaining a "black box" for the customer. Visualization of the learning process and learned features are thus necessary to assist the interpretation of this new technique.

Tilear is a precision-machinery quality-inspection software system based on a deep auto-encoder. The acquired signals from product samples are always high-dimensional. t-Stochastic Neighbor Embedding (t-SNE) [1] was selected to create 2D visualization maps due to its ability to embed high-dimensional data into a low-dimensional space while preserving local relations between data points (i.e. data points close to each other in the original high-dimensional space remain close to each other in the embedded low-dimensional space). The algorithm can be summarized in two steps (refer to [1] for a thorough mathematical description):

- Model the neighborhood relations between data points by using joint probability distributions over pairs of data in both the original high-dimensional and the mapped low-dimensional space. Data points close to each other in the space have thus a high probability of being picked up as neighbors, whilst data points far away from each other have a low probability of being categorized as neighbors.
- Minimize the Kullback-Leibler divergence between the probabilities in high-dimensional and low-dimensional spaces with respect to the locations of the points.

In order to visualize the learning process of a well-trained deep auto-encoder, the following steps were performed:

- Retrieve the well-trained deep auto-encoder.
- Propagate samples in the test dataset through the well-trained auto-encoder, obtaining the hidden layer activity vectors of each sample.
- Apply t-SNE to these hidden layer activity vectors and the raw samples.
- Display the t-SNE mapped low-dimensional data in 2D.

Figure 1 shows an example visualization of the raw data and hidden activity vectors of different hidden layers of a well-trained Tilear network based on t-SNE. The blue circles represent good samples while the red crosses represent defective samples. Visualizations of raw data and hidden layer activities of a well-trained network are ordered from *a* to *e*.

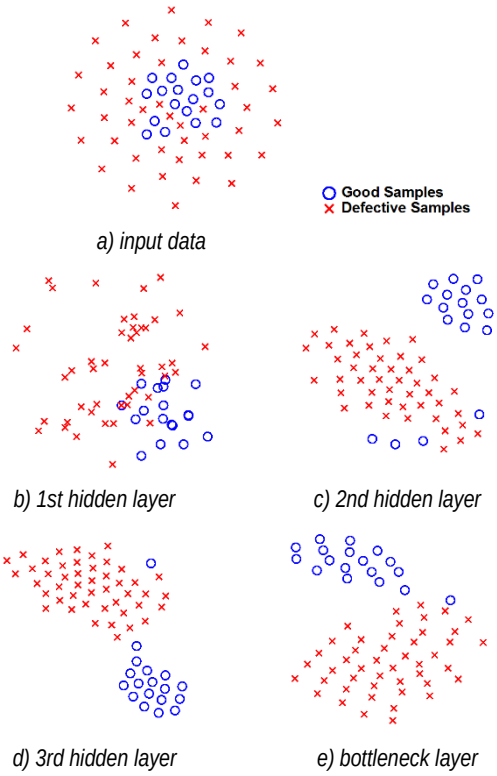

Figure 1: t-SNE 2D maps of layer activity vectors of a well-trained deep auto-encoder network based on raw audio signals.

Figure 1 shows how raw sample data in the dataset are mixed up and not separable. However, clusters of good and defective samples emerge as the learning progresses from low-level to high-level hidden layers (*b* to *e*). The learned clusters in the bottleneck layer form the basis to perform fault detection.

The visualization of the learning process clearly shows how the natural clusters buried in the raw high-dimensional space are learned by the network. This type of visualization is an invaluable tool to demonstrate the working principles of Tileye and Tilear to potential customers, facilitating the industrialization of the technique.

We thank MCCS, the Cantons of Central Switzerland and the Swiss Confederation for supporting this work.

[1] M.Laurens van der, G.Hinton, "Visualizing data using t-SNE", Journal of Machine Learning Research (2008) 2579.