Visual Interpretation of DNN-based Acoustic Models using Deep Autoencoders

Tamás Grósz, Mikko Kurimo

**Department of Signal Processing and Acoustics,
Aalto University, Finland**

Aalto University
School of Electrical
Engineering

Introduction

DNNs have become the state-of-the-art in many fields

They are viewed as black boxes

It is important to know how and why they work

- So that users can trust the system
- Also, to avoid learning issues

Visual interpretation means that we show the hidden representations of the data

- We need to transform high dimensional data into 2D

T-SNE

Well-known and widely used tool

It has two steps:

1. Calculation of distances between points (probabilistic approach)
2. Optimization of the layout by minimizing the KL-divergence

One major issue: the optimization is performed on the data that we want to visualize

- Changing some of the data modifies the entire layout

UMAP

An alternative to t-SNE

A bit more complicated than t-SNE

The steps of the algorithm:

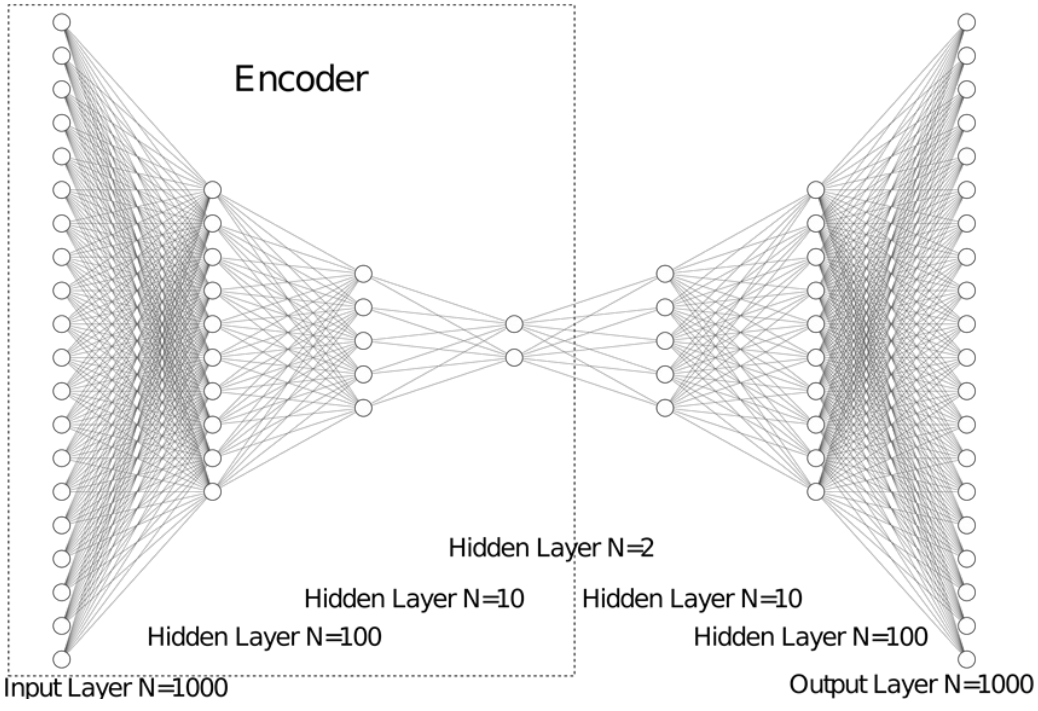
1. Finding a manifold on which the data-points are uniformly distributed (Riemann metric is needed)
2. Fuzzy topological representation of the original and embedded data is calculated
3. Optimization, where cross-entropy between the two topological representations is minimized

It also optimizes using the visualization data

Deep Autoencoder

A general tool for dimension reduction

After training, only the encoder part is used (no optimization)



Evaluation metrics

An ideal criteria to compare the methods doesn't exist

Procrustes Distance

- how well can we reconstruct the high-dimensional data from the embeddings?

Mutual Information

- measures mutual dependence between two sets
- Kraskov-Stögbauer-Grassberger entropy estimation

Distance Correlation

- correlation between paired vectors of arbitrary dimension

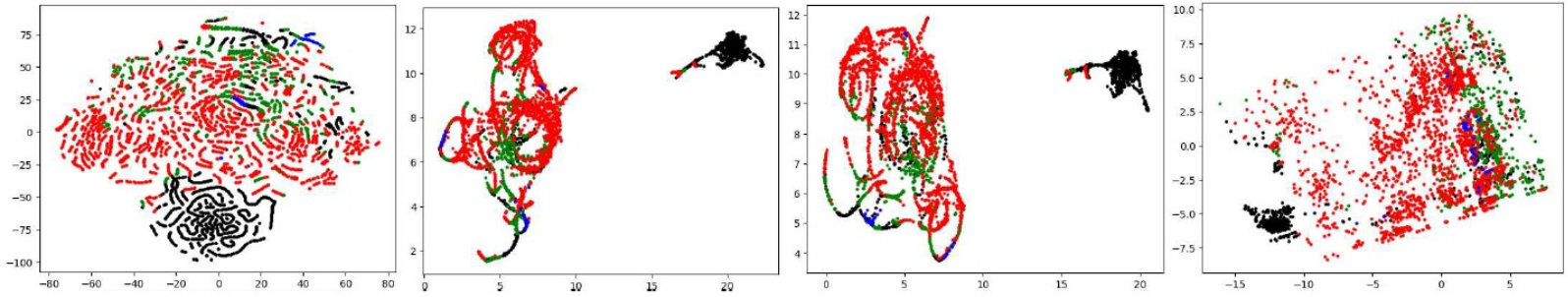
Experimental setup

- An acoustic DNN was inspected (5 layers, 1000 relu)
- Trained on WSJ English corpus
- 3 test files randomly chosen for visualization
- T-SNE, UMAP parameters tuned using the first layer
- AE trained using the training data
- UMAP-train: 1st step performed with extra data

Results

layer	metric	t-SNE	UMAP	UMAP-train	AE
1	PD	0.909	0.919	0.923	**0.896**
	MI	0.341	**0.460**	0.429	0.392
	DC	0.755	0.823	**0.828**	0.793
2	PD	0.936	0.948	0.939	**0.928**
	MI	0.402	0.470	**0.559**	0.544
	DC	0.640	0.813	0.838	**0.871**
3	PD	0.970	0.960	0.955	**0.948**
	MI	0.407	0.485	0.453	**0.503**
	DC	0.497	0.804	0.807	**0.882**
4	PD	0.976	0.968	0.965	**0.952**
	MI	0.499	**0.586**	0.491	0.495
	DC	0.431	0.797	0.804	**0.871**
5	PD	0.977	0.968	0.966	**0.938**
	MI	0.433	0.497	0.502	**0.518**
	DC	0.461	**0.805**	0.796	0.791

Visualizaton of the first layer

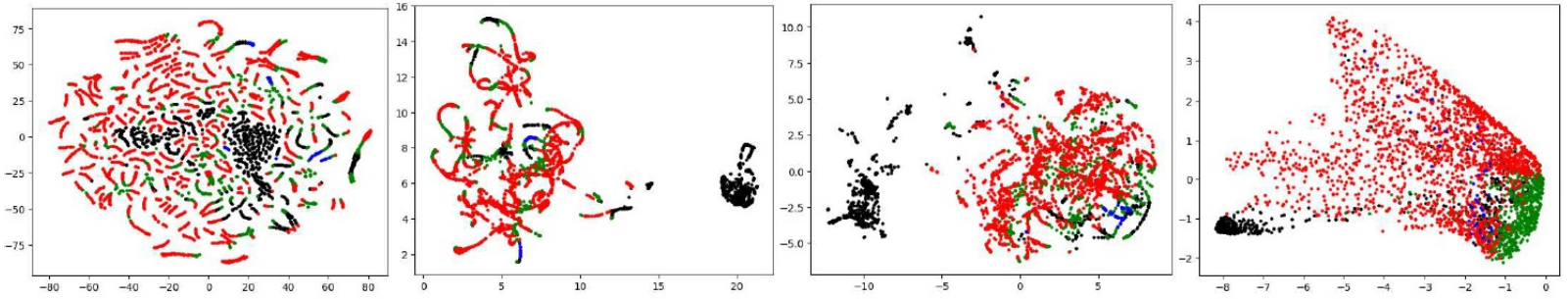


Silence is well separated from the rest

Consonants and vowels clusters start to form

UMAP-train: the extra data impacts the layout

Visualization of the last layer



AE shows that 3 clusters are formed

Inspecting the consonant group of AE, two additional sub-clusters can be found (nasal and fricative)

Conclusions

- AEs could be used to visually interpret DNNs.
- We can train AEs with a large amount of data, and after training, a fixed transformation is used.
- There is a great need for a good evaluation metric.

Thank you!

aalto.fi

Aalto University
School of Electrical
Engineering