Correlating Speech Processing in Deep Learning and Computational Neuroscience
Shefali Garg, Department of Computer Science and Engineering
Smith Gupta, Department of Electrical Engineering

Abstract

Deep learning has been extensively used in various fields like computer vision, natural language processing, etc. but it has not been explored potentially in the field of auditory data classification. We are building up on the work of Lee[1] and presenting here an application of Convolutional Deep Belief Network (CDBN) on audio data for the task of digit classification.

Features are extracted from an unlabeled audio dataset, which is a version of TIDIGITS which contains speeches of 326 speakers (111 men, 114 women, 50 boys and 51 girls) each pronouncing 10 digit sequences.

We also present a comparison of our results with traditionally used approaches for audio classification i.e., MFCC and raw spectrogram methods.

Methodology

Convolutional Deep Belief Network (CDBN)

- PCA whitening (with 80 components)
- 300 first-layer bases (maps), Filter size 6, Max-pooling ratio of 3
- Local receptive field, weight sharing, pooling
- Invariance to small frequency shifts

Audio Data

Unsupervised Feature Learning

Classification

• SVM is used as a classifier

Speech Processing in human brain

Wernicke’s Area

This is located in posterior section of the superior temporal gyrus (STG) in the left cerebral hemisphere and is involved in the understanding of written and spoken language.

Feature Extraction:

According to a research[3] in 2014, human brain breaks up a speech signal into phonemes and extracts features corresponding to these phonemes. Different neurons in the brain are responsive to different types of sounds, like plosives, fricatives, vowels and nasals.

So far, the task of digit classification is accomplished through use of baseline features (MFCC and spectrogram) and use of Deep Belief Network[4]. Here we apply CDBN on an unlabeled audio dataset and use the learned features on the task of digit classification. We compare our results with the traditional approaches.

Results

CDBN is giving lower accuracy than MFCC (need more experiments and training)
MFCC features are specifically designed for audio data and particular task
CDBN is general model, can be used for other speech processing tasks. Maybe, because CDBN model is a bit closer to the human brain model.

In future, improve accuracy of CDBN by increasing number of training epochs and modifying parameters. Use GPU to improve speed.
Combine other deep learning methods like Self Taught Learning, etc.
Try the model with other datasets and other speech classification tasks like speaker identification, phone classification etc.
Relate the activation of neurons in the Wernicke’s area to the hidden layer activations in deep belief network.

References

4. Audio Feature Extraction with Deep Belief Networks, API 2011