Speech processing in Deep Learning and Computational Neuroscience SE367A: Cognitive cience

Shefali Garg 11678 Dept. of CSE
Smith Gupta 11720 Dept. of EE
Indian Institute of Technology, Kanpur

Guide: Prof. Amitabha Muherjee
18 November, 2014

Abstract

Deep learning has being 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. In this project, Convolutional Deep Belief Network (CDBN) [1] has been applied on audio data for the task of digit classification. Audio features are extracted from the dataset which are later used for classification. We also present a comparison of our results with traditionally used approaches for audio classification i.e., MFCC and raw spectrogram methods.

1 Introduction

Processing of audio data in human brain has been of great interest since a long time. In the past there has been a lot of work done related to audio processing involving HMM and GMM for computation. In this work, we are applying deep learning methods to get more general model and to have an insight as to how the neurons in the brain processes speech input. It is possible to design filters and feature extraction methods which resembles processing done by human neurons. In sparse coding, the learnt representation of auditory data gave filters which are known to be similar to the neurons involved in audio processing in mammals' brains. For example, learnt representations of auditory data and cochlear filters in the auditory cortex were found strikingly similar [2].

2 Speech Processing in human brain

According to a research [3] in 2014, human brain breaks up a speech signal into phonemes and extract features corresponding to these phonemes. Different neurons

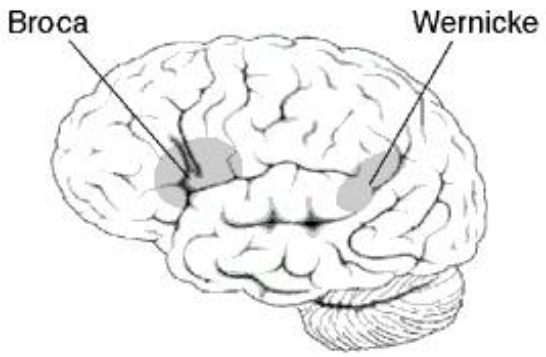

Figure 1: Human brain

in the brain are responsive to different type of sounds, like fricatives, plosives, vowels and nasals.

2.1 Wernicke's Area

Wernickes area is involved in the understanding of spoken and written language. It is located in posterior section of the superior temporal gyrus (STG) in the left cerebral hemisphere. Research using Transcranial magnetic stimulation suggests that the area corresponding to the Wernickes area in the non-dominant cerebral hemisphere has a role in processing and resolution of subordinate meanings of ambiguous words - such as ("river") when given the ambiguous word ("bank"). In contrast, dominant word meanings ("teller" given "bank") are processed by the Wernicke's area present in the dominant hemisphere. This area is also important in understanding jokes. [source : wikipedia]

3 Dataset

Audio dataset is used for digit classification, which is a version of TIDIGITS dataset. It contains speeches of 326 speakers (111 men, 114 women, 50 boys and 51 girls) each pronouncing 10 digit sequences.

4 Methodology

For the digit classification task, first the features are extracted from the audio data which are then used for classification using Support Vector Machines. Features are the components of an audio signal that are good for identifying its linguistic content and do not include irrelevant information and noise. They are obtained either using domain-expert knowledge (e.g., spectrogram, MFCC, etc.) or are generated automatically using deep

learning algorithms (DBN, CDBN, etc.) Following methods have been implemented to get good feature representation of the input data.

4.1 Spectrogram (RAW)

In this approach [4], the waveforms are converted into spectrograms by taking Fast Fourier Transformation (FFT). The spectrogram represents power of different frequency bands over time. For a sampling frequency of 8000, it gives 129 frequency bins (FFT coefficients) per frame. The log of magnitude of these complex coefficients represents power in each frequency bins. Half-overlapping 'Hann' window is used to get the frames.

4.2 Mel-frequency Cepstral Coefficients (MFCC)

MFCCs are widely used features in speech processing tasks and have been state-of-the-art. In this method [8], the input audio signal is first framed into short frames such that the signal is statistically stationary within a frame. The power spectrum of each frame is then calculated to identify the frequencies present in it. This is motivated by the human cochlea (an organ in the ear) which vibrates (and wobbles small hairs) at different spots depending on the frequency of incoming sounds. Different nerves fire depending on different locations of vibration which informs the brain what frequencies are present in the signal.

But, the cochlea cannot differentiate between two closely spaced frequencies. For this reason, the energy in each spot is calculated. The Mel filterbank is applied to the power spectra which gives the amount of energy present in each filter (frequency regions). The Mel scale tells about the width of filterbanks to be used and their spacing.

Then, the logarithm of filterbank energies is taken. This is also motivated by human hearing as in order to double the perceived volume of sound to our ear, eight times energy is required. Since the filterbanks are overlapping, to decorrelate the filterbank energies, Discrete Cosine Transform (DCT) of log energies is calculated. 12(2-13) of DCT coefficients are kept discarding the rest as the higher coefficients represent fast energy changes and degrade the performance. The delta and delta-delta features are then appended to keep some temporal information. Also, the frame energy is included. This gives the 42-dimensional feature vector for each input.

4.3 Restricted Boltzmann Machine (RBM)

RBM [4] is a stochastic neural network and a complete bipartite undirected probabilistic graphical model. It contains hidden and visible units whose joint configuration has an energy given by:

$$\text{E(v,h)} = \sum_{i \epsilon visible} \frac{{v_i - a_i}^2}{2\sigma_i^2} - \sum_{j \epsilon hidden} b_i h_j - \sum_{i,j} \frac{v_i}{\sigma_i} h_j w_{ij}$$

where h_j, v_i are the states of hidden unit j and visible unit i, a_i, b_j are their biases, w_{ij} is the weight between them and σ_i is the standard deviation of the Gaussian noise for visible unit i. The network assigns a probability to every pair of a visible and a

hidden vector via this energy function:

$$p(v,h) = \frac{1}{Z} e^{-E(v,h)}$$

The RBM is trained by Contrastive Divergence in which the weights are updated such that $p(v)$ is high. In this task, sliding window approach is used so that each context window is the training input for the RBM. The RBM used in our experiment has 6 x 129 visible units which train 300 hidden units (bases). The activation of hidden units is used as a feature vector for the context window.

4.4 Convolutional Deep Belief Network (CDBN)

CDBN [1] is composed of many layers of Convolutional RBMs (CRBM), where CRBM is an extension of RBM to a convolutional setting. In the convolutional network [5], all the neurons of same feature map share the same weights but receive different inputs shifted in frequency. Each neuron is connected to small number of neurons in the previous layer representing features of limited frequency range. This is inspired from Hubel and Weisels experiment on cats that the cells of a visual cortex in cats respond to only local receptive field.

For input layer consisting of nv dimensional array of binary units and K n_w dimensional filter weights W_k (bases), the hidden layer consists of n_h-dimensional arrays ($n_h = n_v - n_w + 1$) with units in group k sharing the weights W_k and bias b_k and c. Their energy function is defined as:

$$\text{E(v,h)} = -\sum_{k=1}^{K}\sum_{j=1}^{n_H}\sum_{r=1}^{n_w} h_j^k W_r^k v_{j+r-1} - \sum_{k=1}^{K} b_k \sum_{j=1}^{n_H} j_j^k - c\sum_{i=1}^{n_v} v_i$$

After the convolutional layer, a max-pooling layer is added which computes the maxima over small neighborhoods of hidden units probabilistically. The weight sharing and pooling gives invariance to small frequency shifts and reduces overfitting.

For training CDBN, contrastive divergence is used and sparsity penalty term is added to prevent overfitting. In this task, firstly PCA whitening is done to reduce dimensionality to 80 components. The kernel size (filter length) of 6 and max-pooling ratio of 3 is used to train 300 first layer bases. [Figure 2] These bases (activations) serves as the feature representation of input audio.

5 Results

The results are shown in the Table 1. MFCC features have given the best accuracy among all other feature representations outperforming all the methods.

6 Conclusions and Future Work

According to our experiments, MFCC outperforms all other methods and gave good representation of audio data. But, these hand-tuned features are time-consuming, re-

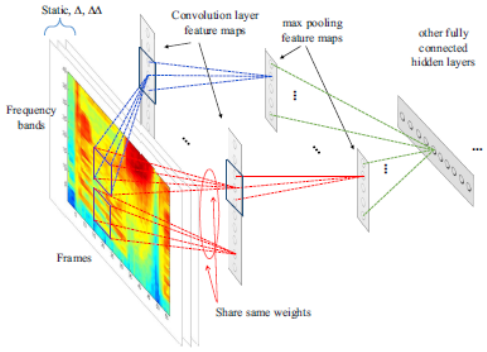

Figure 2: CDBN architecture

Method	Accuracy(%)
RAW	86.68%
MFCC	**92.76%**
RBM	36.99%
CDBN	53.42%

Table 1: `Results`

quire domain-expert knowledge and do not generalize well to other domains and tasks. On the other hand, deep learning methods do automatic feature extraction and gave more general features. They have been shown to perform quite well in the past.

Thus, it is not straightforward to obtain good feature representation from deep learning methods. More experiments are needed to tune parameters, number of layers in the network and number of training epochs. Since the training time taken by deep networks is enormous, GPUs will be used to improve speed. We will also explore large datasets and more challenging tasks. Other deep learning methods like Self Taught Learning, etc. will also be explored. Also we would like to relate the activation of neurons in the Wernicke's area to the hidden layer activations in deep belief network.

7 Acknowledgement

We would like to acknowledge the efforts of Prof. Dr. Amitabha Mukherjee for his support and guidance in this project.

References

[1] H. Lee, P. Pham, Y. Largman, and A. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in Advances in Neural

Information Processing Systems 2, Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, Eds. Cambridge, MA: MIT Press, 2009, pp. 10961104.

[2] E. C. Smith and M. S. Lewicki. Efficient auditory coding. Nature, 439:978982, 2006.

[3] Nima Mesgarani, Connie Cheung, Keith Johnson, Edward F. Chang, Phonetic Feature Encoding in Human Superior Temporal Gyrus. (2014)

[4] Audio Feature Extraction with Deep Belief Networks, API 2011

[5] Abdel-Hamid, Ossama, Li Deng, and Dong Yu. "Exploring convolutional neural network structures and optimization techniques for speech recognition."INTERSPEECH. 2013.

[6] Abdel-Hamid, Ossama, et al. "Convolutional Neural Networks for Speech Recognition." (2014).

[7] G. Hinton, L. Deng, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath and B. Kingsbury, & ldquo, Deep Neural Networks for Acoustic Modeling in Speech Recognition,& rdquo, IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82-97, Nov. 2012

[8] http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/