Chapter 1

Introduction

Neural networks have always been an attractive area of research since 1960. The attempt at simulating the human brain has always been intriguing. Since the first perceptron model, Neural Networks have evolved in many ways, in which layers of perceptrons grew wider and deeper. However, until recently, it was only possible to train shallow networks, because of the vanishing gradient problem. The vanishing gradient is a phenomena where the error information starts to decay when propagated through many layers, and therefore the learning process is no longer doable. A remedy was made by Hinton[3], in which the network is trained a layer at a time, instead of trying to train all layers at once.

Deep Learning is the new trend in Machine Learning field. Recently, there has been many applications that uses Deep Learning. Training these deep networks is very expensive computationally, they require heavy computations on the GPU, and so far, several frameworks that facilitate training Deep Neural Networks (DNN) and Convolutional Neural Networks (CNN) —a deep network that uses Convolutional filters— have been developed by many research labs around the world. CNNs are usually very popular with the computer vision applications.

Recently, DNNs and CNNs have been applied to the field of speech recognition with very promising results. CNNs have the ability to reduce spectral variations and model spectral correlations which exist in signals, therefore CNNs are a more effective model for speech compared to DNNs [6]. In this work, we experiment with CNNs and apply them to a small scope of acoustic modeling which is phoneme classification, using a part of the TIMIT dataset due to the expensive nature of training CNNs. We use the CNN training library, Caffe to train our network. Since Caffe is intended for training images, we will train on spectrograms i.e. images of FFT of the phonemes. We will base our work on the architecture described in [6].

Chapter 2

Related Work

There has been many approaches for speech recognition with Neural Networks. The classical approach was always to combine Hidden Markov Models (HMMs) with NNs, such as [5]. This approach has been very successful and popular. Recently, there has been attempts to remove the need for HMMs. [2] used Recurrent Neural Networks (RNNs) with good results for speech recognition and has yielded promising result. [1] have done similar work with RNNs.

[6] uses HMMs in their model, however, we only build a CNN based on their architecture, and since we do not do speech recognition, there is no need for an HMM.

There has been attempts to use both DNNs and CNNs in speech recognition, however, DNNs have difficulty modeling transitional variance within speech signals, which exists due to difference in speaking styles [4]. Various speaker adaptation techniques are required to reduce this variation. Therefore, we have preferred to use CNNs for this task, since TIMIT consists of a wide range of speakers.

Chapter 3

Methodology

Chapter 4

Experimentation and Results

Bibliography

[1] Graves, A., and Jaitly, N. Towards end-to-end speech recognition with recurrent neural networks. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)* (2014), pp. 1764–1772.

[2] Graves, A., Mohamed, A.-R., and Hinton, G. Speech recognition with deep recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (2013), IEEE, pp. 6645–6649.

[3] Hinton, G. E., and Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science 313*, 5786 (2006), 504–507.

[4] LeCun, Y., and Bengio, Y. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks 3361* (1995), 310.

[5] Robinson, A. J. An application of recurrent nets to phone probability estimation. *Neural Networks, IEEE Transactions on 5*, 2 (1994), 298–305.

[6] Sainath, T. N., Mohamed, A.-r., Kingsbury, B., and Ramabhadran, B. Deep convolutional neural networks for lvcsr. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (2013), IEEE, pp. 8614–8618.