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[1], 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. 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 [2].

Chapter 2

Related Work

Chapter 3

Methodology

Chapter 4

Experimentation and Results

Bibliography

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

[2] 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.