Deep Neural Networks for Acoustic Modeling

November 2015

1 Project Proposal

- **Topic of group:** Deep Neural Networks for Acoustic Modeling
- **Group members:** Bajibabu Bollepalli, Hieu Nguyen, Rakshith Shetty, and Pieter Smit (Mentor).
- **Work plan:**
  1. Frame based phoneme recognition using simple multilayer perceptrons. Quantify performance against no of layers, type of non-linearity used (For eg. sigmoid neurons, Rectified linear units).
  2. Experiments with various input features i.e augmenting two or three neighbour frames of of MFCCs as single frame, and spectrogram features as input to the DNNs.
  3. Compare the results with Gaussian Mixture Model method.
  4. Weight vectors initialization: randomly assigned vs pretraining or other methods (if time permits)
  5. Experiments using different neural network architecutres i.e simple recurrent neural networks (RNNs) and long term short term memory networks (if time permits).
- **Implementation language:** Python with Theano library
- **Possible papers** (**???**)

2 Introduction

The goal of automatic speech recognition (ASR) is to convert a speech signal into corresponding text form. Since five decades the progress of ASR has been improved from recognition of isolated digits to telephone-conversational speech. However, the performance of ASR is still beyond humans level. Humans are good at recognizing the speech in different accents, dialects, or pronunciations, and speech in different styles, at different rates, and in different environmental

conditions. The variability in speech signals pose a big challenge to the conventional ASR system. The system simply fails if you provide new speech signal, which is coming from different souce than in training data.

Traditional ASR system employs the hidden Markov models (HMMs) to model the speech signal where each acoustic state is modeled by a Gaussian mixture model (GMM).

2.1 Recurrent Network based Speech recognition

Speech signal is inherently sequential and it would be beneficial to use models which can handle features with larger time contexts. Recurrent Neural networks(RNN) are a such a class of models as they have infinite depth in time, allowing an RNN based ASR model to utilize preceding context when making predictions. This is because the current state of an RNN, h_t depends on its previous state, h_{t-1}, and thereby on all its previous inputs, x_t. This is described in equations **??** and **??**. Here N is the hidden layer non-linearity and y_t is the network output at time t.

$$h_t = N(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \quad (1)$$

$$y_t = W_{hy}h_t + b_y \quad (2)$$

A more powerful class of recurrent networks called Long Short-Term Memory (LSTM) models **?** are widely used today. Each LSTM unit has a memory cell whose update is controlled by a set of non-linear gates as shown in figure **??**. This allows the network to store important information for a long duration in time, giving them ability to better handle longer sequences. This model also addresses the problem of vanishing gradients observed when training vanilla RNN models **??**.

Idea to use RNNs in speech recognition is not very new and was already explored in 1994 in **?** for phoneme recognition. With renewed interest in neural networks in general, RNN based speech recognition systems have also seen an upturn recently. In **?**, a deep (multi-layered) recurrent neural network is used to de-noise the speech features before using them in traditional speech recognizers. The model is trained to predict clean MFCC features given an input of MFCC features corrupted with noise. Such tandem model, where RNN is used to encode/ the features and its output is used as input to speech recognizer models like HMM is also used in **?**. Alternatively in **?**, LSTM based model is used in **?** end-to-end speech recognition, with model taking acoustic MFCC features as input and predicting phoneme classes as output. LSTM based models get computationally expensive with larger vocabularies. A efficient way to utilize model parameters to facilitate using LSTM models on datasets with large vocabulary is presented in **?**. RNN based models are also used other areas of speech recognition like voice activity detection **?**.

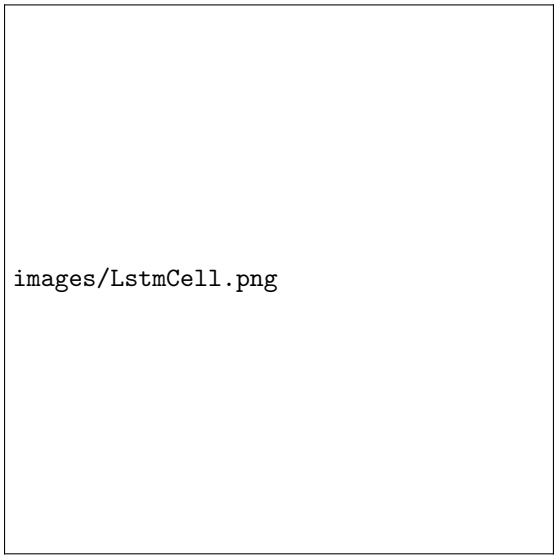

Figure 1: Block diagram of a single LSTM cell. dotted lines indicate gate controls and full lines are data flow. Triangle indicates sigmoid non linearity.