Deep learning architectures in speech processing

January 28, 2017

1 Introduction

Deep learning as a broad framework of methods has taken the speech and natural language processing system building by a storm. The deluge of work that has happened in the last 25 years or so has shown remarkable promise and also underscores the need to present here the progress that has been made since the late 80's. One of the primary goals of our paper, therefore, is to highlight some of the most significant approaches and the findings therein. While our approach mostly will be chronological, we will also focus on how deep neural networks (DNN) have fundamentally revolutionized our approach to both Automatic Speech Recognition (ASR) and Text-to-Speech Synthesis (TTS). In addition, we want to focus on a very specific aspect within the use of DNNs in speech processing, namely the integration of linguistic knowledge in achieving some of the remarkable successes in the core tasks of speech processing.

In a series of seminal papers, **??** outline the use of multilayered neural networks in ASR and speaker recognition, respectively.

2 The architecture of Deep Neural Netowrks

Typically, DNNs refer to feedforward multi-layered artificial neural networks (ANN) with more than one layer of hidden units with a logistic function to traverse between the hidden layers and the output. Here we rely on **?** to outline the general architecture of DNNs. We will illustrate the functioning of the algorithms and the processes with an acoustic modeling task as discussed in **?**.

$$y_j = logistic(x_j) = \frac{1}{1 + e^{-x_j}}, x_j = b_j + \sum_i y_i w_{ij},$$

(1)

$$p_j = \frac{exp(x_j)}{\sum_{k exp(x_k)}} (2)$$

$$C = -\sum_{j} d_j log p_j (3)$$

$$\Delta w_{ij}(t) = \alpha \Delta w_{ij}(t-1) - \epsilon \frac{\delta C}{\delta w_{ij}(t)} \tag{4}$$