Deep learning architectures in speech processing: Modeling dynamic speech gestures

March 4, 2017

1 Introduction

Deep learning as a broad framework of methods has taken speech and natural language processing system-building by a storm. The deluge of work that has happened in the last 25 or so years has shown remarkable promise, and growth. Therefore we don't need to underscore 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 speech processing systems in general, and Automatic Speech Recognition (ASR) and Text-to-Speech Synthesis (TTS) in particular. More particularly, we will focus on a specific sub-task within broad speech processing and that is the articulatory-acoustic domain mapping. In that sense, the objective of this paper is to cover the most significant advances made in speech processing though deep learning models and architectures, but also focus on how deep learning architectures have helped unearth crucial generalizations between speech articulatory gestures and their acoustic manifestations. 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. At the outset, we would like to outline that the goals of this paper are not to introduce the concepts of machine learning, but to specifically treat a class of learning algorithms that variously appear in the literature under the cover term deep learning. Essentially, all deep learning systems and architectures are a specific form of artificial neural networks which have been in existence for a significant amount of time, but regained currency in the mid-80s with publications emerging from the parallel distributed processing group at San Diego. The paper is organized as follows: In the following section, we motivate the need to look closely into the various deep learning initiatives and outline the importance of the major impacts of these learning mechanisms. In Section 3, we briefly, and very generally, discuss the various architectures that help build deep learning

models. In section 4, we concentrate on a specific use-case and outline the successes in this use-case, namely, the articulatory-acoustic mapping problem. In section 5, we offer some perspective and perhaps, hazard some speculation as to the future of these learning models.

2 Why Deep Learning?

We will begin this section with Manning [2015] who in his paper points to a few crucial changes that have come as a consequence of the expansion of deep learning systems in natural language processing. One of his primary concerns and engagement in this paper has been to point out how natural language processing takes center stage as far as being one of the most important challenges for machine learning scientists and deep learning enthusiasts, alike. Manning's 2015 entreatment is a clarion call for linguists, NLP engineers and data scientists to shift focus away from beating benchmark dataset tasks and challenges, and to concentrate on "problems, approaches, and architectures". The import of Manning's 2015 appeal to re-engage with the cognitive and design goals of the study of human languages has been felt and responded to by work that has sought to understand deep learning architectures in the context of language cognition, and essentially has managed to re-center focus on NLP tasks as not just an engineering challenge, but also to re-imagine the goals of NLP research broadly within the cognitive and language sciences. In this paper, and especially **??**, we discuss in detail how these goals have been achieved, and in which direction NLP research is moving armed with deep learning tools. We will treat the use of deep learning methods on solving acoustic variation problems that come about of an essentially non-linear problem that of articulatory-acoustic mapping and articulatory-acoustic inversion. The non-linearities that arise in the articulatory-acoustic mapping are a direct consequence of the way speech articulators overlap with each other in a non-discrete fashion to produce contrastive sequences of segments.

3 The architecture of Deep Neural Netowrks

While the most basic functions of the artificial neural network or perceptron remain the same and lot of advancement has been made in the way the basic ingredient, in this case, the perceptron has been used to create architectures that are remarkable improvements over the initial attempts to use these machines for both classification and regression tasks.

In a series of seminal papers, Bengio et al. [1989a,b] outline the use of multilayered neural networks in ASR and speaker recognition, respectively.

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 Hinton et al. [2012] 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 Hinton et al. [2012]. Information from each hidden unit, j, is used along with a logistic function in order to map the total input from the previous layer, x_j, to a scalar state, y_j which is then sent to the following layer.

Here, as in 1 below, b_j refers to the bias associated with the unit j

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

(2)

(3)

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

3.1 Restricted Boltzmann Machines

3.2 Recurrent Neural Networks

3.3 Sequence learning: Special case of LSTMs

4 Modelling dynamic processes in speech processing

In this section we will cover a special use case of applying deep learning methods to an essentially non-linear set of problems, namely, modeling the dynamic nature of the articulatory-acoustic mapping (AAM). Since Stevens [1968], it has been known that the relationship between the articulatory and acoustic parameters is non-linear and

References

Christopher D. Manning. Computational linguistics and deep learning. *Comput. Linguist.*, 41(4):701–707, December 2015. ISSN 0891-2017.

Y. Bengio, R. Cardin, R. De Mori, and E. Merlo. Programmable execution of multi-layered networks for automatic speech recognition. *Commun. ACM*, 32 (2):195–199, February 1989a. ISSN 0001-0782.

Yoshua Bengio, Renato De Mori, and Regis Cardin. Speaker independent speech recognition with neural networks and speech knowledge. In *Proceedings of the 2nd International Conference on Neural Information Processing Systems*, NIPS'89, pages 218–225, Cambridge, MA, USA, 1989b. MIT Press.

Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition. *Signal Processing Magazine*, 2012.

K.N. Stevens. *The Quantal Nature of Speech: Evidence from Articulatory-acoustic Data*. 1968.