Madison Pahl

Augusto Partida

SE 480

Jan 5, 2018

Alex Acero: Deep Learning in Speech Recognition

<https://youtu.be/RBgfLvAOrss> &  [http://web.stanford.edu/class/ee380/Abstracts/171129.htmll](http://web.stanford.edu/class/ee380/Abstracts/161012.html)

This seminar discusses the history and progression of machine learning, and the lead up to neural networks dealing with speech recognition. Alex Acero discusses the lead up to neural networks through the beginning of artificial intelligence, to the recent breakthroughs with neural networks and cognitive computing. Specifically, Acero discusses the algorithms behind these algorithms, and shows how speech recognition works from a mechanical and algorithmic standpoint.

At the beginning of the discussion, Acero goes over a brief history of artificial intelligence and machine learning, in specific the birth of artificial intelligence in 1950 with the “Turing Test”. The very first breakthrough of artificial intelligence, however, was from Arthur Samuel in 1956 with his checkers machine. This understanding of new computing possibility soon lead to a wide range of programs such as ELIZA in 1966, and movies considering new possibilities on the front of computer. Before any other breakthroughs happened on the front there was an artificial intelligence winter 1970s until 1997, where people did not know much about neural networks or had a lack of interest in its development. This began to chance once Deep Blue, the machine playing chess was created by IBM, and later hosted challenges for autonomy in vehicles by Stanford University. This then comes to what we now know to be prominent fields of artificial intelligence including products like Siri and Alexa on the speech recognition front.

Acero then proceeds to discuss the importance of deep learning, and mathematically how it works and why. Some of the first commercial uses for these algorithmic processes were to categorize emails as spam. The task was to “improve on task T with respect to performance metric P based on experience. In short, taking an email with its content including labels, content, address, and correctly identifying if it is spam or legitimate. There were then algorithms for binary classifications, such as identifying a breast cancer tumor either benign or malignant, the first variable for this type of program may be tumor size, so from input data, there is a decision line that is put in the middle of the benign and malignant data, in order to identify one over the other. As more variables are added to create accuracy, so are mathematical dimensions. This means if we add age into identifying a tumor, then we must take into consideration how we are to place the decision line. This is where perceptron learning comes in, which in sort is a process of algorithms to find a line that accurately fits the plotted data by rotating its axis. From this first type of algorithmic, machine learning approach came more detailed approaches to training systems such as probabilistic frameworks, stochastic gradient decent, multilayer perceptron, to what we now know as neural networks.

When using neural networks and deep learning to deal with speech recognition, it is based on many levels of abstraction. The first is sound frequency, so much like visual sound-bytes, they are broken down and analyzed in small components. They are then combine into large sets of frequencies which identify words, then are able to be analyzed for actual meaning. This process is more complex but similar to the processes that are behind photo analysis and recognition, which includes breaking down a single image into smallest pieces called pixels, then identifying the image from there using lines, which create features within the picture, then it is able to use prior learned information from that and other pictures to identify an object from an image.

At the end of the seminar speech, Acero went over the framework used for Siri, Apple’s voice recognition software, which I found very interesting, since a phone can hold such an expansive neural network, but still be able to maintain stable battery life. This is done by running on two layers of framework called two-pass detection. In the first level, which is low compute and is always on consists of the MFCC computation, frame buffer, small DNN and HMM scorer, which allows for activation of the larger layer if relevant acoustic information is identified. This second layer is more accurate, and is the main processor, which hosts the large DNN and HMM scorer, which takes care of the main tasks of acoustic and speech recognition. This input of speech is then sent and TTS is activated for response. This process takes in text which is the response to the Siri activation, analyzes it, generates prosody, and selects groupings of sounds based on the database information, concatenates the information, thus leading to Siri’s voice responding to requests. Overall, the learning and understanding of the algorithmic history behind the development of today’s speech recognition software is interesting and informative to learning how to push speech technology further.