Research Project Report

Speech Recognition Study: Apply Deep Learning to Speech Data with TensorFlow

Ke Wang (wang.ke2@husky.neu.edu)

Keywords

Speech recognition, convolutional neural networks, connectionist temporal classification lost function

Abstract

Among all the new models unveiled this decade, Convolutional Neural Networks (CNNs) are effective for reducing spectral variations and modeling spectral correlations in acoustic features [1]. In CNNs, the nodes of each layer are clustered, the clusters overlap, and each cluster feeds data to multiple nodes (orange and green) of the next layer [2]. CNNs are widely used in signal or image processing, computer vision, etc.

Background

While artificial intelligence is developing, machine learning and deep learning come to people's lives. We are using them every single day without even noticing them sometimes. Recent years, recognition systems that dealing biometric modalities, like fingerprint, iris, voice, are blossoming. Among them, speech recognition is introduced to us by cell phone manufacturers and becoming more and more popular. As machine learning thrives in not only science field, but also our daily lives, neural networks have been applied in various industries and approved to be sometimes efficient to solve some tricky problems.

This paper will apply existing algorithms to datasets to automatically translate speech into text, then use loss function to evaluate this model. There are many algorithms and methodologies to analyze and recognize speech, to better understand neural networks and algorithms, mainly CNNs and are used in this system.

Introduction

1. Speech Recognition

Speech, in other words, sound wave, is a bunch of continues signals. To analyze that, we have to convert it into digits with same space for computers to understand. This process is called sampling, a part of signal processing. It is the reduction of a continuous-time signal to a discrete-time signal [3].

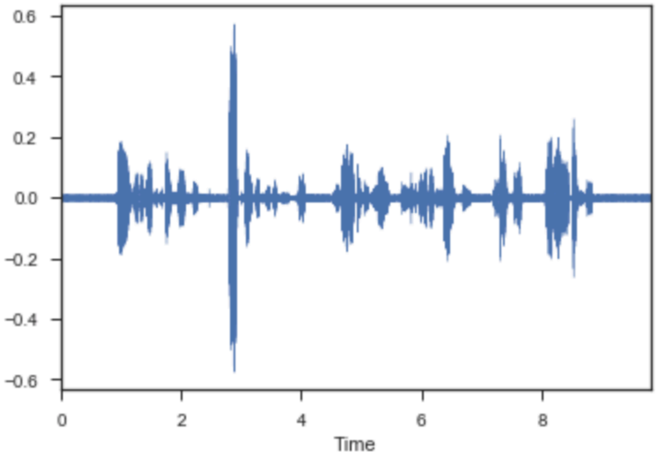


Fig. 1 Wave plot of a speech

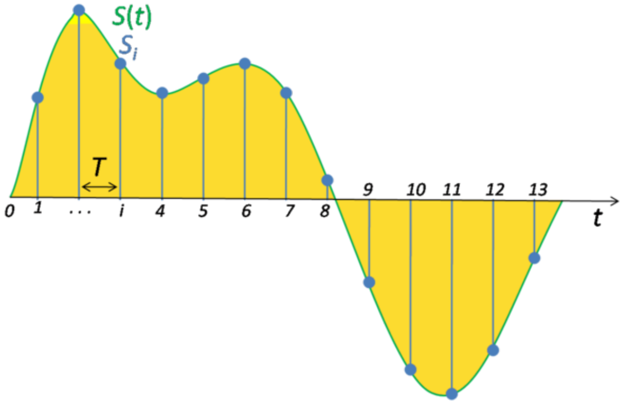


Fig. 2 Sampled Signal [3]

Then based on a spectrogram, we can visually see the frequency at each time, and then can be transformed into digits for computer to process.

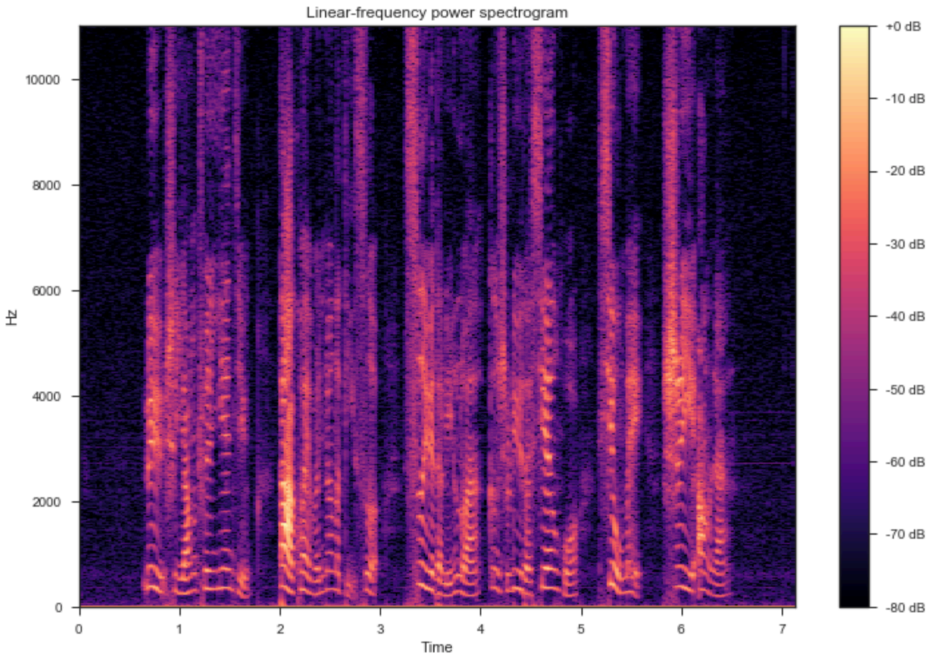
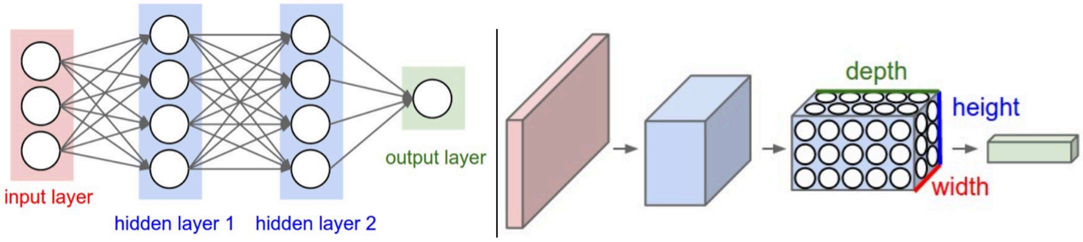


Fig. 3 Frequency – time spectrogram

Each language has a dictionary to convert sound digits to words. For example

2. Convolutional Neural Networks

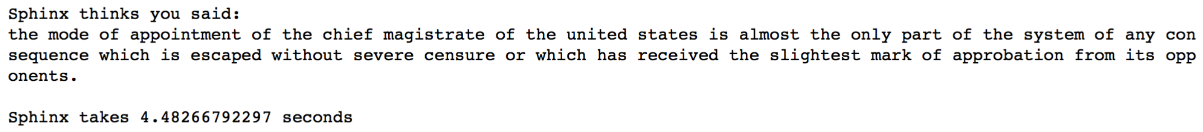
Convolutional Neural Networks are very similar to ordinary Neural Networks from the previous chapter: they are made up of neurons that have learnable weights and biases. Each neuron receives some inputs, performs a dot product and optionally follows it with a non-linearity. The whole network still expresses a single differentiable score function: from the raw image pixels on one end to class scores at the other [6].



Left: A regular 3-layer Neural Network. Right: A ConvNet arranges its neurons in three dimensions (width, height, depth), as visualized in one of the layers. Every layer of a ConvNet transforms the 3D input volume to a 3D output volume of neuron activations. In this example, the red input layer holds the image, so its width and height would be the dimensions of the image, and the depth would be 3 (Red, Green, Blue channels).

3. Existing Speech Recognition APIs

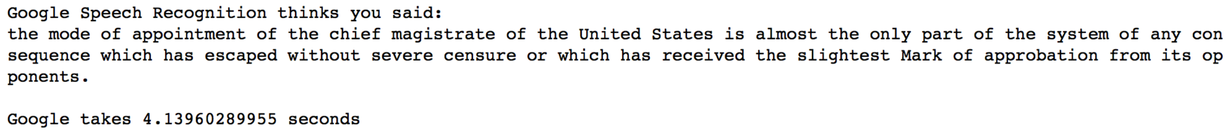
CMU Sphix



```
Sphinx thinks you said:
the mode of appointment of the chief magistrate of the united states is almost the only part of the system of any con
sequence which is escaped without severe censure or which has received the slightest mark of approbation from its opp
onents.

Sphinx takes 4.48266792297 seconds
```

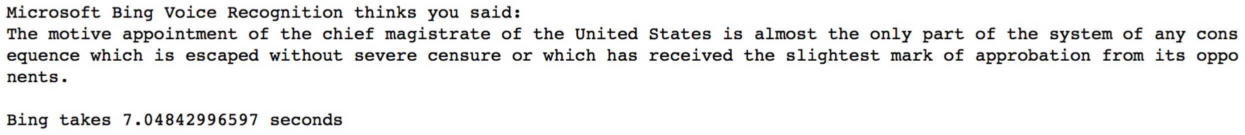
Google



```
Google Speech Recognition thinks you said:
the mode of appointment of the chief magistrate of the United States is almost the only part of the system of any con
sequence which has escaped without severe censure or which has received the slightest Mark of approbation from its op
ponents.

Google takes 4.13960289955 seconds
```

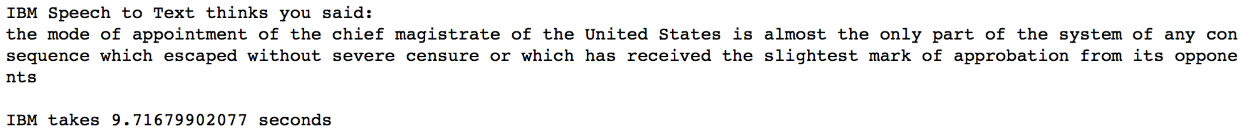
Microsoft Bing



```
Microsoft Bing Voice Recognition thinks you said:
The motive appointment of the chief magistrate of the United States is almost the only part of the system of any cons
equence which is escaped without severe censure or which has received the slightest mark of approbation from its oppo
nents.

Bing takes 7.04842996597 seconds
```

IBM



```
IBM Speech to Text thinks you said:
the mode of appointment of the chief magistrate of the United States is almost the only part of the system of any con
sequence which escaped without severe censure or which has received the slightest mark of approbation from its oppone
nts

IBM takes 9.71679902077 seconds
```

Improvements

1. Although loss decreases as training times increase, it still takes a very long time to train the system, and the error rate is not low.
2. Can apply Recurrent Neural Networks (RNNs) to speech data
3. Can apply sentiment analysis on the text

Reference:

[1] https://arxiv.org/pdf/1701.02720.pdf
[2] http://news.mit.edu/2017/explained-neural-networks-deep-learning-0414
[3] https://en.wikipedia.org/wiki/Sampling_(signal_processing)
[4] https://gab41.lab41.org/speech-recognition-you-down-with-ctc-8d3b558943f0
[5] https://medium.com/@ageitgey/machine-learning-is-fun-part-6-how-to-do-speech-recognition-with-deep-learning-28293c162f7a
[6] http://cs231n.github.io/convolutional-networks/