Comparing neural networks for word recognition

Michał Kotarba

Abstract

In this article 4 different neural networks are used to recognize and classify words. Three of the four networks are taken from Tensorflows tutorial to simple audio recognition[1] and one is proposed by the author of this article. Both training and validation datasets are rather small, so even with dropout as high as 50% overfitting was practically impossible to bypass. Because of that even though training accuracy is between 64 – 99% validation accuracy is between 50-60%.

Introduction

Word and speech recognition became a priority in many companies such as Google, Microsoft, Amazon, Apple and many others. Recently shown Duplex[2] by Google is a system that can communicate with the world just as humans do. People watching presentation were both thrilled and terrified of possibilities. Therefore it is a good idea to try building a simple system, that recognizes and classifies words spoken to it as a part of the project.

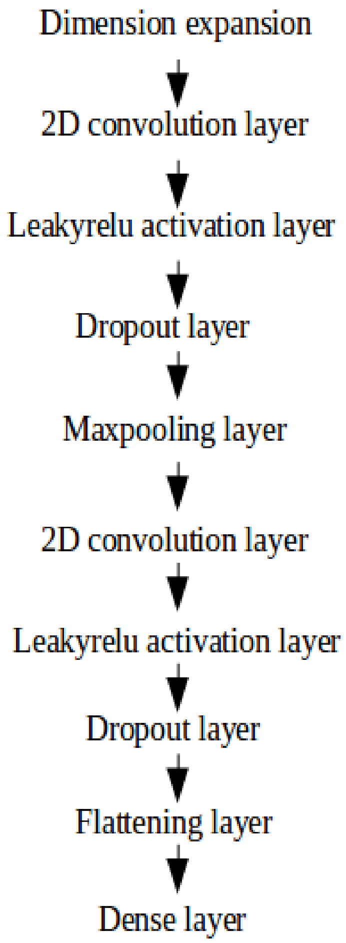
In speech recognition domain many different solutions are tried with varying accuracy obtained by them. Some of them are: hidden Markov models, methods based on dynamic time warping, convolutional and recurrent neural networks[3].

Preprocessing data

The data collected by people in my group is in uncut .wav files with many words in one. Files with labels are provided with them. The files are cut into single words based on label files. After that these files are randomly split into train and validation sets with ratio 4:1. Both training and validation sets are read in main file, they are extended to maximum length and MFCC representation is obtained. These MFCC representations are put into neural network.

Authors neural network

Authors neural network consists of 10 layers as shown below.



Drawing 1: Layers of authors network

Tuning the network

When selecting best parameters for this network 3 values were the subject of change: first convolution filter number, second convolution filter number and dropout value. Train accuracy and validation accuracy in table 1. are the biggest values during training for particular parameters. It’s important to note very small training and validation datasets, therefore results indicate overfitting.

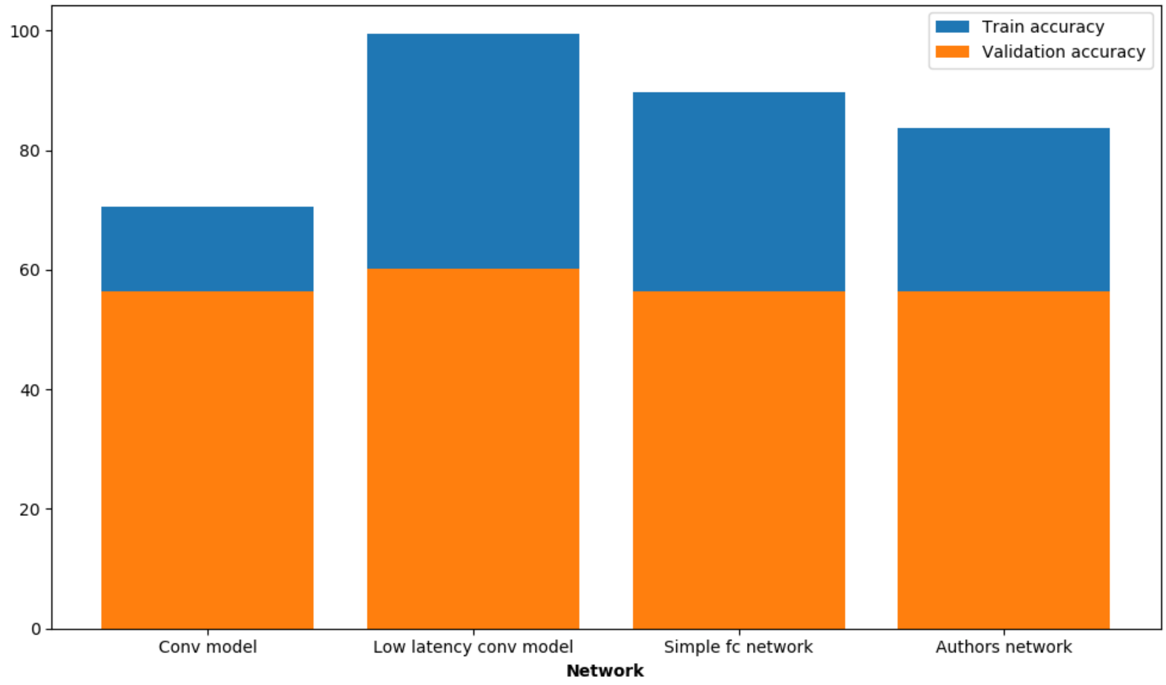
Table 1: Train and validation accuracies for particular parameters

First convolution filters	**Second convolution filters**	**Dropout**	**Train accuracy**	**Validation accuracy**
64	64	0.1	83,01	50,00
64	64	0.2	83,65	56,41
64	64	0.3	72,11	52,56
64	64	0.5	76,28	55,13
80	80	0.35	69,87	50,00
80	80	0.4	68,26	52,56
100	100	0.1	64,74	50,00
100	100	0.3	69,87	47,43
100	100	0.4	73,08	48,72

The best validation accuracy was acquired for 64 first and second convolution filters and 0.2 dropout.

Networks comparison

As can be seen at drawing 2 authors network is competitive in validation accuracy with all 3 others networks. Biggest validation values were taken from training. All the networks were tested using the same dataset both for training and validation. It’s also important to note that all 4 networks got similar results ranging 55-60 validation accuracy.



Drawing 2: Network comparison

Conclusion

Authors network proposed in this article came out as a success. It's competitive with other networks and trains similarly fast.

Bibliography:

1. https://www.tensorflow.org/versions/master/tutorials/audio_recognition
2. https://www.youtube.com/watch?v=bd1mEm2Fy08
3. https://en.wikipedia.org/wiki/Speech_recognition