Speech Recognition/Classification using Deep Neural Networks

Etai Wanger & Eyal Ben David

Supervised by Guy Adam

Abstract

While todays interaction between man and machine is done many through screens a much more natural way to do so is via speech. Many industry leaders are already developing and improving algorithms that are meant to mimic and understand human language. Achieving human like performance will open a completely new way to interact with devices and offer new applications we haven't thought of before. In this Project we test state of the art methods in machine learning using deep neural networks for "understanding" words and develop new architectures that try to capture all the features of human language.

Keywords: Deep Neural Networks, Speech Recognition, Speech Classification

Model

We aim to build a Deep neural network that will be able to distinguish between 12 classes: 'Yes', 'No', 'Up', 'Down', 'Left', 'Right', 'On', 'Off', 'Stop', 'Go', Silence and Unknown. The input data to the net are 1 second recordings of different people saying a word. Below are schematics of the Networks tested.

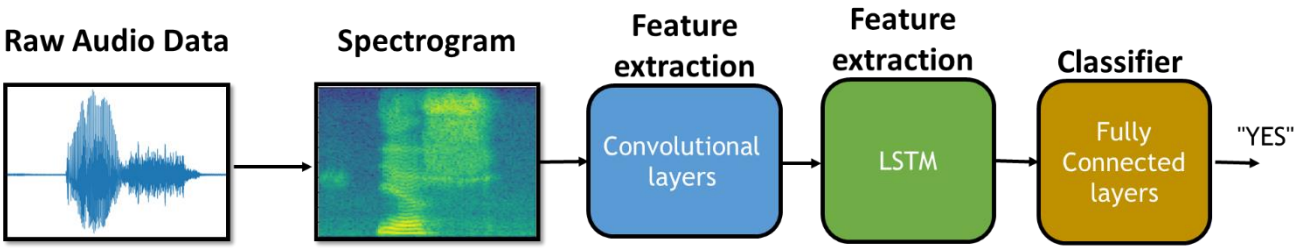


Figure 1 Topology of the Basic network tested – raw data is transformed into spectrogram and then inserted into 8 convolutional layers each channel is then inserted into an LSTM and afterwards into a fully connected classifier

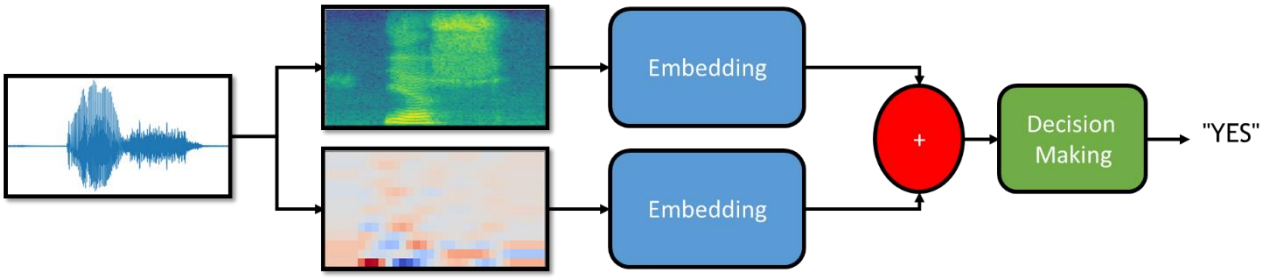


Figure 2 Two input convolutional network followed by an LSTM layer per channel

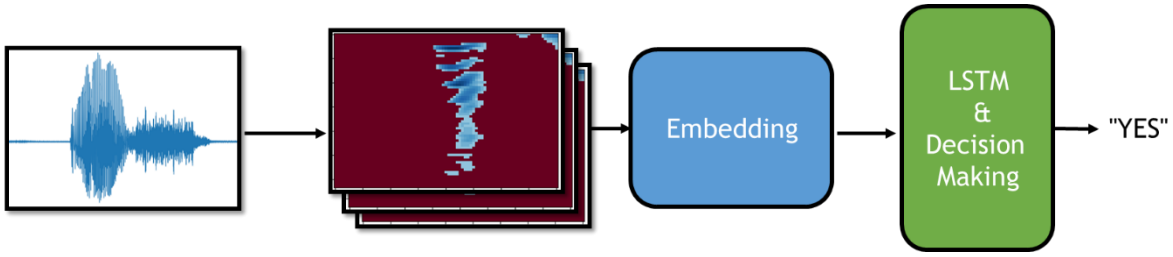


Figure 3 the input is a 3 channel RGB like picture where each channel is created with different window sizes for the STFT

Results

The Networks where tested on the Kaggle TensorFlow Speech Recognition Challenge – data set

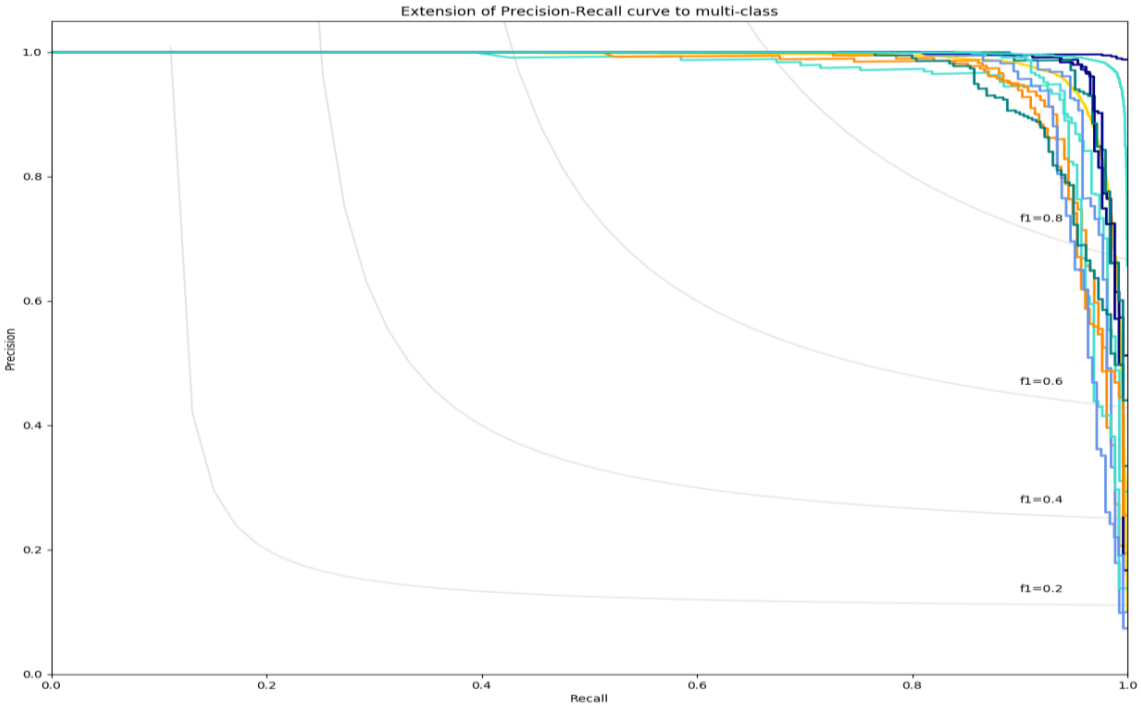
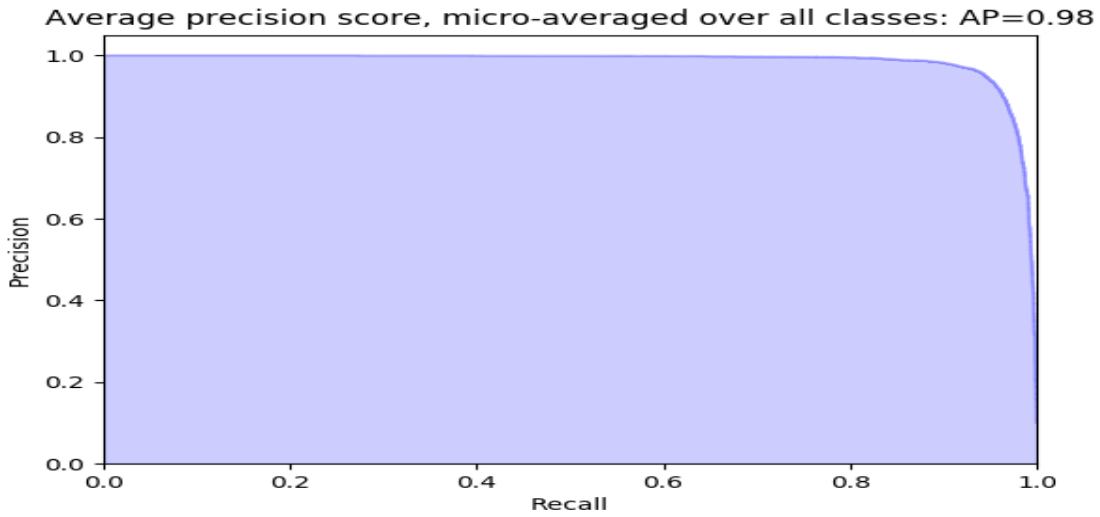


Figure 4 precision Vs recall per class on the validation set for our best model



Results on the Speech recognition data set Test set

Architecture	Data	Private Score	Public Score
CNN-NN	Basic	0.75930	0.75575
CNN-LSTM-NN*	Aug, 'Silence'	0.89157	0.88185
2 Inputs *	Aug, 'Silence'	0.88582	0.88130
3D input*	Aug, 'Silence'	0.84564	0.83771
Best Score*	CNN	0.91060	0.90296

GitHub:

https://github.com/eyalbd2/Kaglle-Tensorflow-Speech-Recognition