Speech Commands

Fabio Zamboni

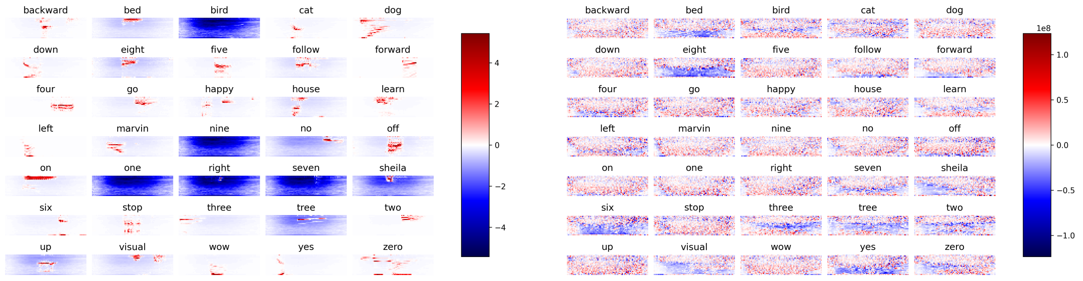
17-05-2023

1 Introduction

In this paper we are going to look at a speech recognition case concerning the identification of 35 words. Child's play, right? More like baby's play! I'm not kidding at age one, children recognize about 50 words; by age three, they recognize about 1,000 words; and by age five, they recognize at least 10,000 words[1]. The data set includes 105 829 recordings from different speakers. It has been divided into a 84 291 audio clips training set, a 12 162 validation set and a 9376 clips test set. Feature extraction has already been performed and these are the wave spectrogram of the audio clips, in the form of 20 frequencies times 80 time periods array. To achieve our goal we are going to use a MLP (MultiLayer Perceptron).

2 Results

A good idea is first of all to look at spectrograms Fig.(1a), we can easily see that words with a similar sound have similar spectrograms, we expect some mistakes among them. We start training a MLP with 1600 inputs, 35 outputs and without hidden layers, in Fig.(1b) we can see it's weights, deep red parts indicate areas in favor of that particular class, while dark blue parts indicate areas against that class. To give the same range of value to all the features we can normalize the data, there are various ways to normalize our data, the graph in Fig.(2a) shows us how different types of normalization affect the training accuracy. Chosen the *meanvar* as normalization we start training our MLP, we choose stochastic gradient descend with batches of 40 given its good trade-off between loss and training time Fig(2b). Fixed the batch size now

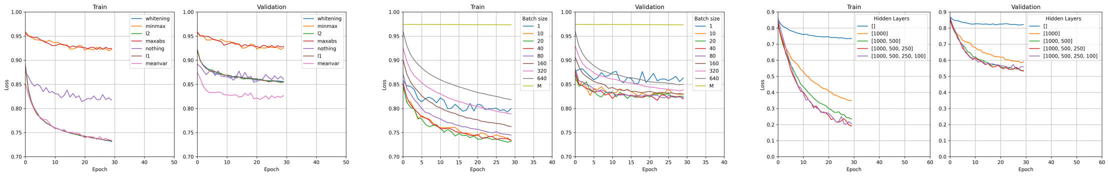


(a) Randomly taken spectrogram for each class

(b) MLP weights without hidden layers

Figure 1

[1] Shipley, K. G., & McAfee, J. G. (2019). Assessment in speech-language pathology: A resource manual. (Plural Publishing)



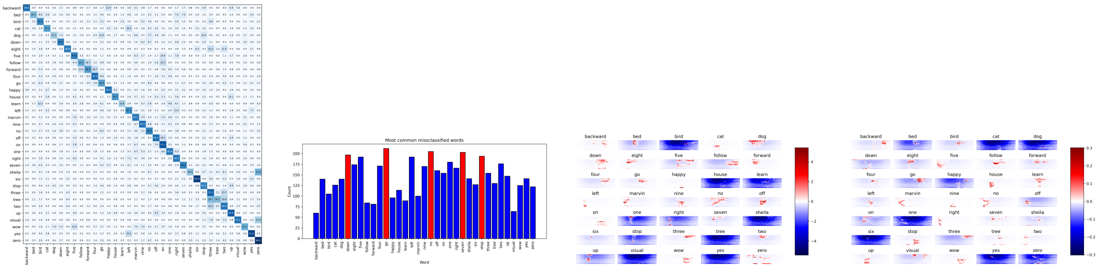
(a) Normalization comparison

(b) Batch size comparison

(c) Layer comparison

Figure 2

we can try to modify the structure of the MLP adding hidden layers Fig.(2c), we see that the 3 hidden layers configuration $[1000, 500, 250]$ works well. Chosen the model now we can perform some analysis, first of all let's look at the accuracy, evaluated on the test set, of our MLP: 46.33%, after that we can plot the confusion matrix Fig.(3a) in order to see the distribution of errors among the classes, here we can see that some of those are confused, in particular there are 2 groups that used to be confused, *three*, *tree*, *two* and $follow$, $forward$. Instead *down*, *go*, *no*, *seven* and *stop* are the top 5 miss-classified words Fig.(3b). Looking at the spectrograms of the miss-classified words Fig.(3c), we toke one for each classes, we see some strange behavior with respect to the well-classified spectrograms Fig.(3d). Listening to the audios of some miss-classified samples, one can see that they are either pronounced with particular accents or recorded in a way that cuts off a small part of the spoken word. We can infer that these are some of the reasons why MLP fails to classify them.



(a) Confusion Matrix

(b) Top5 miss-classified

(c) Miss-classified

(d) Well-classified

Figure 3

I affirm that this report is the result of my own work and that I did not share any part of it with anyone else except the teacher.