Deep learning for Speech Commands

Juan Carlos Morales[1], Guillem París[2], Somayeh Jafari Dinani[3]
[1] (Affiliation): dept. Signal Theory and Communications, UPC, Barcelona, Spain,
[2] (Affiliation): dept. Signal Theory and Communications, UPC, Barcelona, Spain,
[3] (Affiliation): dept. Signal Theory and Communications, UPC, Barcelona, Spain, somayeh.jafaridinani@alu-upc.edu

***Abstract*—This paper indicates one of the speech recognition problem which is recognizing speech commands and gives several solutions for it. It is shown that the best solution is changing the kernel dimension with accuracy of 95.9 % and the loss of 0.3423 %. Although we believe that the fourth method which is the concatenation of the LSTM method and VGG method may leads to better results.**

***Index Terms*—LSTM, Mixup, Loss, VGG**

I. INTRODUCTION

According to [1], the problem of automatically recognizing speech with the help of a computer is a difficult problem and the reason for this is the complexity the human language. Speech recognition, or more commonly known as *automatic speech recognition* (ASR), is the process of interpreting human speech in a computer. We will consider *speaker independent* ASR, i.e. systems that have not been adapted to a single speaker, but in some sense all speakers of a particular language. In this way, another problem will arise which is "Speaker Variability" meaning that all speaker have their special voices due to their unique physical body and personality. The main goal of speech recognition is to get efficient ways for humans to communicate with computers. One of the application of the ASR is speech commands recognition which can be used for environmental control like turning on the light, controlling the TV and etc.

Deep Learning is a new area of Machine Learning research, which has been introduced with the objective of moving Machine Learning closer to one of its original goals means Artificial Intelligence.

In this article, we use Deep Learning method to alleviate the speech commands recognition problem. In the section II, the related speech recognition problem has been mentioned. In the section III, several solutions have been introduced for the problem. In the following section, section IV, the results of the experiments which have been done based on the aforementioned solutions depicted. At the end in the final section, there is a conclusion for the whole experiment.

II. DEFINING THE PROBLEM AND ANALYZING

Here [2], we want to use deep learning for one of the speech recognition problems which is detecting speech commands. The used dataset includes 65,000 one-second long utterances of 30 words recorded by thousands of different people, and have created by the TensorFlow and AIY teams. It is asked to determine the related word for each recording file based on its spectrogram. As the outputs are 30 words and the inputs are different spectrograms, the problem is a supervised deep learning type and also classification one. The dataset has been divided into three parts: one for training, one for testing and another one for validation, according to the standard method for supervised deep learning. About 80% of the dataset allocated to the training, 10% to the testing and 10% to the validation.

III. TOWARDS THE SOLUTION

To solve the classification task we have used the spectrogram of the audio signal as input in several deep learning models. We have tried four different models, being the first one a VGG baseline. The VGG used for the baseline (shown in fig. 1) is a simple VGG16 with no further modifications.

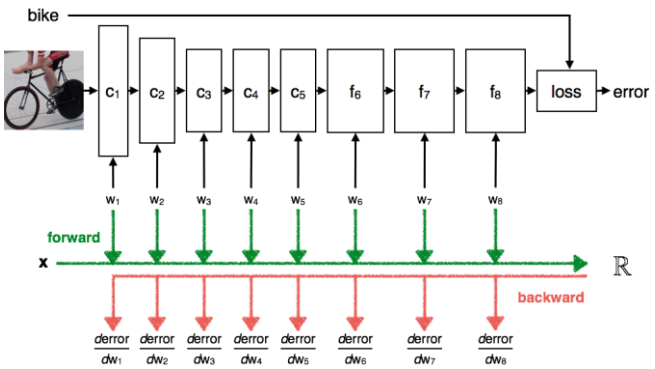

Fig. 1. VGG16 model

A final fully connected layer and a softmax was used for classification over the 31 classes.

The second and the third model use two different ways of mixing over the input data. The second model stacks two input samples to create a new input, doubling the dimensionality of the input data. The new label is computed as the mean of the one-hot vectors. The third model sums two input samples weighted using a beta distribution. The labels are also computed as the weighted sum of the one-hot labels of the original samples. In this case the dimensionality of the input is maintained.

For the fourth and last model the baseline was used in the same way than in the first model, but a LSTM model running in parallel to the baseline was added, as shown in next figure. The LSTM is run over the spectrogram taking at each step the frequency vector at time 't'. The output of the VGG is passed through another convolutional layer to reduce the dimensionality and make it similar to the one of the LSTM. These two outputs are concatenated in the same vector and a final fully connected layer and a softmax are used for the classification in the 31 classes.

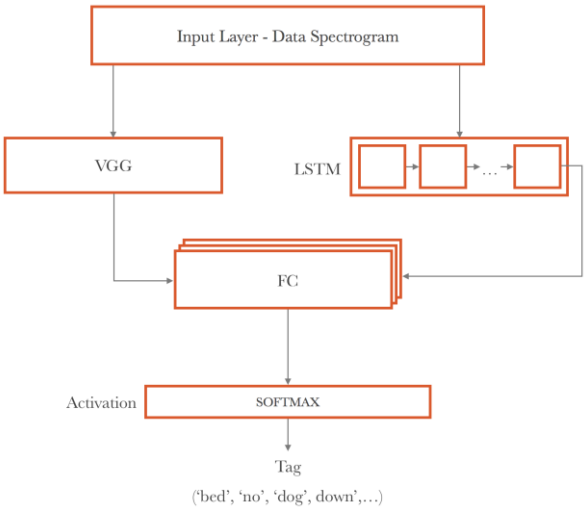

Fig. 2. **. Parallel structure joining both VGG and LSTM.**

IV. RESULTS OF THE EXPERIMENTS

Three main changes have been carried out on the given model: (1) A change on the convolutional kernel's shape, (2) the dataset augmentation using the Mixup [4] procedure and (3) the addition of a LSTM structure parallel to the already programmed VGG.

The accuracy of the given VGG model was already formidable: 95.9%, with an average loss of 0'2482.

The VGG used in the given model operates with a squared kernel, of dimensions 3x3. Therefore, the first proposed improvement has been to change this to a more adequate shape, in which the frequential information of the signal is better gathered. Taking into account the data's spectrograms (shown in figure 3), a rectangular shape of 5x3 has been chosen. For this, a total accuracy of 95.9% has been achieved, with no difference to the original one whatsoever, and an average loss of 0'3423. Figure 4 shows the loss curve comparison between the original model and the 5x3-kernel model.

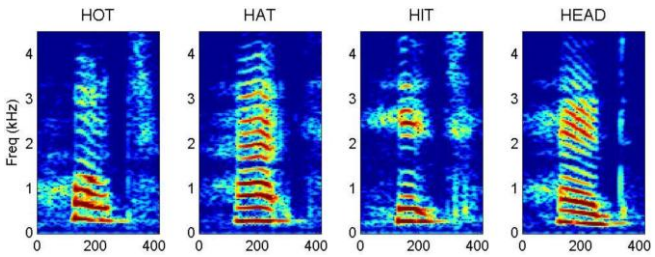

Fig. 3. From "Deep Learning With Keras". J.A. Rodriguez Fonollosa

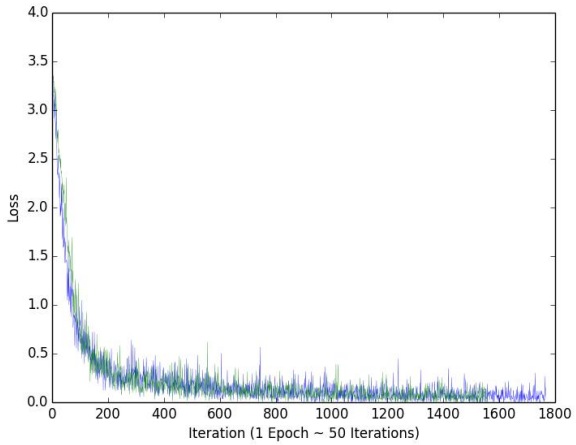

Fig. 4. Loss curve comparison. In blue the original one, in green the 5x3-kernel's one.

The data augmentation technique -following the Mixup procedure- has been performed in two ways: (1) stacking both source matrices and (2) averaging over them.

For the already explained stacking procedure, the change resulted in a worsened accuracy of 94.7% and an average loss of 0'538. The pure-mixup procedure, averaging -with adequate weights- both inputs and both targets over their one-hot representation gave the same results we already obtained in the original model (95.9% and 0'3423 average loss). Figure 5 shows the loss curve comparison between the original model, the mixing-stack model and the mixing-sum model.

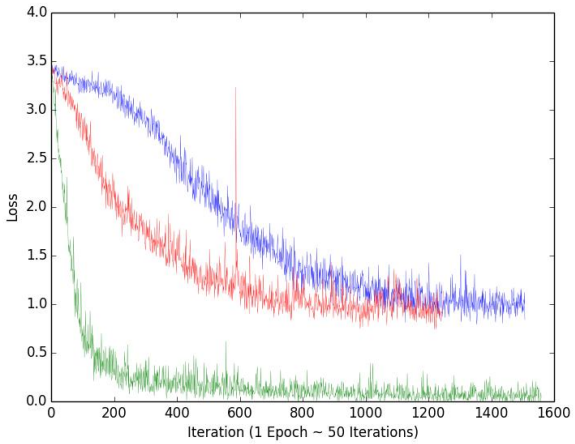

Fig. 5. Loss curve comparison between the original (green) model and the mixing (stack - blue, sum - red) models.

Notice that in the previous figures, the loss curves may have different lengths. That is to avoid overfitting.

Finally, a LSTM was added to the architecture, so memory features of the speech signal could be exploited. However, it couldn't be added after or before the VGG since either one of them would lose its original sense. Therefore, a parallel structure (already shown in fig. 2) has been designed. For this model, there has been no results since there was no time to end its integration on the given baseline.

V. CONCLUSION

The proposed improvements have been quite unsuccessful. Part of it because the given model had already a formidable behavior.

As a possible future improvements, the kernel's shape augmentation could be the path to follow, since its motivation to capture a better frequential resolution of the spectrogram gave good results -at least, it did not worsened the ones we already had. Furthermore, parallel combination of systems, such as the fourth model (VGG parallel to LSTM), usually produce improvement on the results, at the expense of computational complexity.

REFERENCES

[1] Markus Forsberg," Why is Speech Recognition Difficult? ",February 24, 2003,Technical University of Chalmers

[2] https://github.com/jarfo/gcommands

[3] https://research.googleblog.com/2017/08/launching-speech-commands-dataset.html

[4] H. Zhang et al. "Mixup. Beyond Empirical Risk Minimization". https://arxiv.org/abs/1710.09412