IMPACT OF DATA AUGMENTATION ON CLASSIFICATION OF SPOKEN DIGITS

DT2112 SPEECH TECHNOLOGY FINAL BID

Student 1 Robert Rey,
robrey@kth.se

Student 2 Yassine Elkheir,
elkheir@kth.se

Student 3 Simone Porcu,
porcu@kth.se

Student 4 Alex Norlin,
alexnor@kth.se

Student 5 Giovanni Castellano,
giocas@kth.se

Date October 3, 2022

Abstract

The goal of this project is to train a convolutional neural network to perform audio digits recognition. In particular, we are interested in comparing the performances of the model after enriching the training set using different data augmentation techniques. The initial training set is created by ourselves. The test set instead, contains samples from another dataset available online. We observed that while individual data augmentation methods improves our performance, the biggest performance increase was observed when combining them. This brought the test accuracy from 50.3% to 73.1%. There was also no big difference between the accuracy of the test set with only male speakers and the one with female and male speakers.

Introduction

Today, numbers are everywhere, they are the most relevant form of identification, such as personal identity numbers, social security codes, bank accounts, etc. Many ways exist to communicate numbers to a computer, for instance, it could be done using a keyboard or a mouse; with the remarkable improvement in audio recognition, we propose an automatic recognition method for spoken digits.

The recognition of digits from audio files is of relevant interest since it allows to have a smoother interaction between the machine and the user. As part of more complicated tasks, digits recognition could be the key factor for speeding up processes in several scenarios. Some examples where this tool could be beneficial are:

- getting access through security codes;
- submitting digits through a call by means of the voice instead of the keypad
- read loud digits rather than keeping them in mind to transcribe them somewhere

The aforementioned examples are just a few cases of the situations where this project can enhance human computer interaction. Sometimes a constraint could be that no other interactions are possible with the machine: the absence of a keyboard or a mouse on a computer, the impossibility of a person to use similar tools, etc.

In this project, we use a convolutional neural network to output the number (from 0 to 9) that was spoken from an audio input, we use three different data augmentation techniques (**noise injection**, **changing pitch**, **changing speed**) and study how they affect the training process, our deep learning model and the overall recognition process. The model is trained many times on different datasets created adding different combinations of augmentation techniques on a dataset created by ourselves, and it is tested on another dataset available online. [1].

Similar projects have already been done in the past using several different techniques, which again validates the importance of having a tool like this in place. Our interest in the project is due to the appeal that digit recognition gained over the years, along with our willing to apply deep learning models to a task like this one. We are thrilled to play around machine learning techniques for a series of reasons: we are Machine Learning Master students with a shared interest in deepening our knowledge around Machine Learning methodologies in Speech recognition; we would like to compare our results with the ones from trained models already in place, analysing the differences and see how far we can get using just our dataset with some augmentation. Finally, we do think that spoken digits recognition are an important problem to be solved, hence here is why we put our efforts to better understand this task.

Background and related work

There has been a lot of work done regarding both audio classification with neural networks and audio data augmentation. There exists a handful of popular methods for data augmentation. How one goes about augmenting audio data can vary a great deal depending on if one wants to directly modify the spectrogram, waveform and on how to feed the data to the model. Things that could be applied to most audio data include adding noise, speeding it up, adding a low pass filter, simulating echo or reverberation, increasing the volume, increasing the pitch and so on. Feeding the model with images of spectograms we could also use any other technique for data augmentation that one could use for normal images. There is already a lot of previous work done in this field. Classifying audio samples dates back quite some time but a relatively recent paper where deep learning has been used to classify audio samples has been published by P. Suppakitjanusant et al. [2]. A more relevant paper to ours is one by L. Nanni et al. where these methods of data augmentation has been tested on cat and bird sounds [3].

Method

In this project the convolutional neural network AlexNet was used. To train the model a "base dataset" containing Wave audio files was created with the help of the software Audacity. For the base dataset 20 samples per digit, per person were recorded. This means that the total number of samples in the base set is 1000 samples. In particular, each recording is about one second and was sampled using a sampling frequency of 8KHz with a monochannel 16-bits encoding. The base set will also be referred to as "the original data" from here on out.

Samples from the base data set were then used to augment the data.
The following techniques were used to augment the data:

-Changing pitch
-Changing speed
-Noise injection

In particular, the following datasets were created:

1) only original data (size: 1000 samples)

2) original data + white noise (size: 1000 + 1000 = 2000 samples)

3) original data + pink noise (size: 1000 + 1000 = 2000 samples)

4) original data + random noise (size: 1000 + 1000 = 2000 samples)

5) original data + changed speed 1 (size: 1000 + 1000 = 2000 samples)

6) original data + changed speed 2 (size: 1000 + 1000 = 2000 samples)

7) original data + changed pitch 1 (size: 1000 + 1000 = 2000 samples)

8) original data + changed pitch 2 (size: 1000 + 1000 = 2000 samples)

9) original data + random noise + changed speed + changed pitch (size: 1000 + 1000 + 2000 + 2000 = 6000 samples)
Dataset number 4 contains the original samples plus the same samples with the addition of one of these noises: white, pink, violet, blue, Brownian. Moreover, the intensity of the noises is drawn randomly from a uniform distribution between 0.2 and 0.8. Speed changing was performed with two different speeds, 0.75 and 1.5 . Meaning samples were slowed down to 75% of their original speed or sped up to 150% of their

original speed. Two different types of pitch shifts were also used, shifting the pitch up by a half third tone and shifting the pitch up by a whole third tone.

Features were then extracted from the samples in order to feed them to our models as input. The spectrogram were chosen as the feature to represent each sample. Samples that were shorter than 1 second were extended by adding silence on both the end and beginning of the sample until it was long enough. Samples that were longer were trimmed so that they were 1 second.

Finally, the training of the network was performed, we have used Adam optimizer to perform task training and parameters update, we have selected the best optimizer with the optimal learning rate, and decay rate after testing many optimizers on a standard dataset. During the training phase, we have used a tracking technique of accuracy and loss values, and this tracking technique updated parameters only if the validation loss decreases and the validation accuracy increases, this method helps us to avoid overfitting problem as our model weights generalizes on unseen data during training (Validation data 20% randomly selected samples from current dataset, and the model get trained on the remaining 80%).

In order to test the performance of our model, another dataset from a research paper relating to the classification of Audio Signals with the help of Deep Neural Networks [1] was used. This data set contains approximately 30,000 audio recordings of spoken digits (0-9), such that each of them is repeated 50 times by 60 different speakers. The speakers consists of 12 females and 48 males, ranging from the age of 22 to 61. For our testing purposes two test sets were created. One which only contained male speakers, sampled from all speakers totalling 6000 samples and another which contained a mix of both male and female speakers, also totalling 6000 samples.

Results and Discussion

Thanks to data augmentation, we have been able to improve the generalisation capabilities of the AlexNet model. In terms of evaluating our results, we only made use of the accuracy metric, since we believed that including the other metrics was not relevant in this experiment as we were focused on comparing the performance of the model on different data sets. To obtain the results, we trained each model with respective data set and tested them on another data set available online, in order to make it a fair comparison. The results are presented in Table 1. Our interpretation of the results will be given in the next paragraph.

Before the experiment, we expected the performance to be the worst on the Original data set, and so we were surprised to see that we actually obtained a worst performance on the data set which contained the Original Data + Pitch Shift 4, however the initial idea still holds for all the other data sets. We can deduce that having a Pitch Shift of 4 deteriorates the generalisation capabilities of the model, by contaminating the training samples. As far as which type of data augmentation gave the best results, we can see that the results with Pink Noise and Random Noise were the best methods of data augmentation.

Lastly, as we expected, we obtained the best results with the data set which contained all the types of data augmentation, with a final accuracy of 0.7312. This can be explained by the fact that this data set contains a larger training sample size as well as containing a more diverse data set, both contributing to the better generalisation performance of the model.

As a final note, we also performed testing on a data set which contained a mix of male and female speakers, with training being done on the data set which contained all the different types of data augmentation. The results were very similar to the test data of only male speakers, hence we can say that gender doesn't

have a big impact on performance in this case.

Dataset	Accuracy Score
Original Data	0.503
Original Data +White Noise	0.531
Original Data + Pink Noise	0.604
Original Data + Random Noise	0.602
Original Data + 0.7 Speed	0.563
Original Data + 1.5 Speed	0.557
Original Data + Pitch Shift 2	0.541
Original Data + Pitch Shift 4	0.470
Original data + random noise + changed speed + pitch shift	0.7312

Table 1: Accuracy results of all the different datasets

Conclusion

We can conclude that it is possible to get very good results just starting from a small data set with only 1000 samples from 5 different speakers. It is clear from our results that data augmentation has a big impact on the performances. The original data set is not enough to get good results, but adding some data augmentation we can get a decent model which can guess the spoken digits most of the times. In general, augmenting the data we get better results, however, it is interesting to notice that sometimes the performances could also decrease. In our project we observed this behaviour augmenting with pitch shift 4 (although the differences between the accuracies is not significant). Lastly, we want to mention that using a much better training set without any augmentation it is possible to get much better results on the same test set (95.82%) [1]. This means that although augmenting the data helps to improve the model, we cannot get as good results as starting from a good training set.

References

[1] S. Becker, M. Ackermann, S. Lapuschkin, K.-R. Müller, and W. Samek, "Interpreting and explaining deep neural networks for classification of audio signals," *CoRR*, vol. abs/1807.03418, 2018.

[2] P. Suppakitjanusant, S. Sungkanuparph, T. Wongsinin, S. Virapongsiri, N. Kasemkosin, L. Chailurkit, and B. Ongphiphadhanakul, "Identifying individuals with recent covid-19 through voice classification using deep learning," *Scientific Reports*, vol. 11, pp. 2045–2322, 2021.

[3] L. Nanni, G. Maguolo, and M. Paci, "Data augmentation approaches for improving animal audio classification," 2020.