Rey-Jouanchicot Jordan

Report :

Advanced Machine Learning : Urban Sound 8K

Rapport d'écoute et conduite d'entretien

Contents

1 Introduction and problem description

This a report regarding a project for the Advanced Machine Learning class in Algebra University. For this project I decided to work on a sound dataset, which contains 8732 sounds of less (but mostly around) 4 seconds. They is one CSV file giving info about files where to find them in the dataset,their original file, their location in the original files, and of course their classes in ID but also with the name associated to it, also if the sound to detect is in foreground or background.

There are ten classes of sound, which are : airconditioner, carhorn, childrenplaying, dogbark, drilling, engineidling, gunshot, jackhammer, siren, street music.

2 My approach to solve this problem

My aim is to use a feature like spectrogram which generates a 2D representation of sound and to use this representation with a Convolutionnal neural network to learn from it and be able to classify the sounds.

The sounds are split in 10 folders, it is recommended from the different sources related to this dataset to use cross validation and training through the 10 folders.
I followed recommendation and also split the dataset in three dataset :

- Training : 0.6
- Validation : 0.2
- Testing : 0.2

3 Experiments and evaluation

3.1 Fast representation

The first was to decide when to generates the images from the sound at the beginning I did it inside the dataloader, it was a bit too slow and was really slowing down training. I could have move this to training and validation loops and I would have been able to even use GPU for processing but I decided that I wanted to move this to preprocessing and decided to store the list of images in memory and give it to the Pytorch Datasets objects. I created my own Custom Pytorch Dataset class to make things clean.

3.2 Representation choice

I studied multiple representation from images, including MFCC, Spectrogram, and "compressed Spectrogram". According to multiple reports MFCC is a bit decorrelated with sound classification and the results I got during my first test seems to validate it on our dataset.

I made some test on a compressed feature that I will call compressed spectrogram the idea is to do a spectrogram and for each row (each row representing a frquency range and all spectogram having the same number of row) calculating the mean value (so mean of frequency) finally I get a 1D representation with this of the number of row (in my case 400) and I can reshape it to a 2D image, matrix (in my case square image of 20 by 20 pixels). This representation is really slow and generate with small convolutional networks with only four "classics" convolutionnal and pooling layers and one fully connected layer allowed me to get around 90% of accuraccy on validation set and test set.

I made a version taking spectrogram but as the spectogram size changes a lot depending of sounds I decided to just resize it to an image of 100 by 100 whatever the input size. Then I used a pretrained model from Torchvision, I tested mainly on wide_resnet_50_2, I replaced the last layer and did not freeze training of others layers because it is pretrained on Imagenet which purpose is really far from our case, so with freezing unfortunately the model was not able to have so good results but using pretrained weight allow us to have better inititialization than random one, so allow faster convergence of the weights.

3.3 Some optimisations

I choose my optimizer to try to optimize results of the network. I decided to use AdamW which is a modificated version of Adam which should be able to generalize better tham Adam but also include weight decay and converges faster than optimizer like Stochastic Gradient Descent.

I also tested Amsgrad variant of AdamW, it seems to improve the loss), so I decided to keep it used for the final version.

3.4 Results

During this project I had the opportunity to compare multiple representation of sounds, I was really surprised to see how good was able to be the small convolutionnal neural network on the "compressed spectrogram". Also, the small number of epochs required to get fine-tuned wide_resnet_50_2, is also really impressive and I got a validation loss of less than 0.240, showing a really good confidence and good prediction of the model and a validation accurracy of around 0.93. We can see with confusion matrix that results seems pretty balanced, the class with the "lowest" good prediction rate seems to be car horn with biggest confusion with street music.

Following, two confusion matrix, first one on wide_resnet_50_2 trained, second one on smaller neural network on "compressed visualization".

Figure 1: wide_resnet_50_2 trained confusion matrix

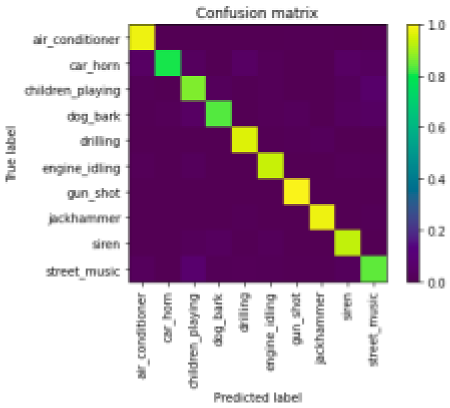
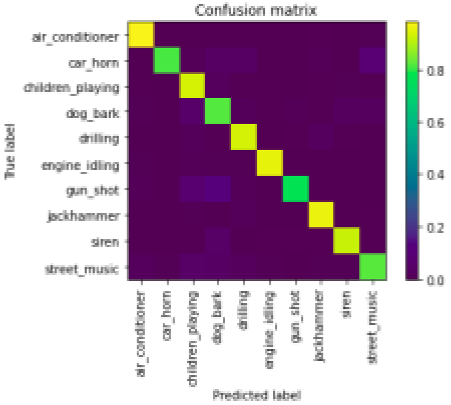


Figure 2: Smaller neural network on compressed representation confusion matrix



4 Conclusion

During this project I learned a lot about sound processing and gained experience with pytorch, it was really interesting. I got really good results. I also liked that we could apply something like this to detect some kinds of sounds like gunshot to be able to detect them automatically, and for example inform the police, using microphones from cities.

5 Sources

For this project, I used to compare my results but also to get some ideas of processing this ressources :
https://arxiv.org/pdf/2007.11154.pdf
https://www.kaggle.com/prabhavsingh/urbansound8k-classification/notebook
https://pytorch.org/
https://www.fast.ai/2018/07/02/adam-weight-decay/