Study doc: feature representation

how does the feature representation effect the embedding space

Fabian Gröger
fabian.groeger@stud.hslu.ch

Sunday 24th May, 2020

Abstract

The experiment aims to show how the embedding space will change when using different feature representations.

1 Introduction

The purpose of the feature representation is to represent the audio in a more compact form than the raw audio. The feature representation further determines the input size for the model. There are a lot of different ways to represent the an audio file in a more compact form. One of the most popular representations is the MFCCs, which is heavily used in the audio domain. However in the recent years, the trend leads more towards using the log Mel spectrogram, which is very similar to the MFCCs, but by omitting the last step of the calculation. This experiment aims to find the optimal feature representation for the thesis.

2 Hyperparameters

The hyperparameters used for this experiment are shown in table 1. The experiment will be conducted using a state of the art ResNet18 architecture on the DCASE dataset. The hyperparameters in section *Feature representation* as well as the sample rate are the default ones proposed by the organisers of the DCASE challenge within the baseline project. The optimal feature representation will be evaluated for `[LogMel, MFCCs]`.

Table 1: Hyperparameters used for the experiment

Hyperparameter	**value**
Dataset	DCASE
Model	ResNet18
Epochs	30-50
Batch size	128
Optimizer	Adam
Learning rate	1e-5
Margin	1.0
L2 regularisation factor	0.01
Embedding dimension	64
Prefetch batches	Autotune (-1)
Random selection buffer	64
Shuffle dataset	True
Random seed	1234
Multi threading	
Number of generators	16
Number of parallel calls	16
Audio sample	
Sample rate	16000
Sample size	10
Sample tile size	5
Sample tile range	5
Convert to mono	True
Feature representation	
Feature extractor	**[LogMel, MFCCs]**
Frame length	480
Frame step	160
FFT size	1024
Number of Mel bins	128
Number of MFCC bins	13

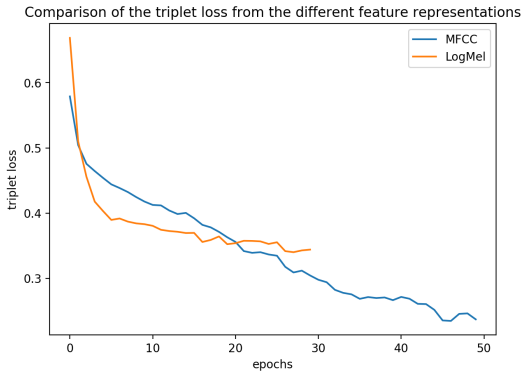

Figure 1: Plot of the triplet loss of the different feature representations

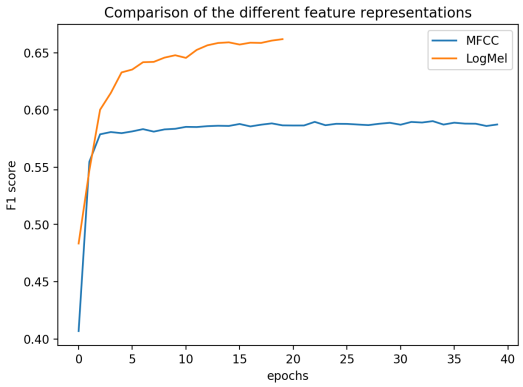

Figure 2: F1 score of the trained classifier using the trained embedding spaces from the different feature representations

3 Results

Comparing the effect of different regularisation factors on the embedding space is pretty hard, since comparing embedding spaces is not very straight forward and is a rather tricky task. This is mainly because of the fact, that to visualise the high dimensional embedding space in a way humans can perceive it, it has to be reduced to two or three dimensions. Therefore the original space can not be examined, and the visual representation is always an approximation of the space in a lower dimension. To still compare the feature representations, a simple logistic classifier is trained on top of the resulting embedding spaces, which aims to show how well a simple classifier works with the embedding space. This will show how good the resulting embedding space is. The classifier is trained for 40 epochs with the same parameters as the embedding model (table 1).

Figure 1 shows the difference between the triplet loss from the model using the different feature representations. It shows that the model trained using the MFCCs results in a significantly lower loss than the model using the log Mel spectrogram. The model using the log Mel spectrogram seem already converging at epoch 30.

Figure 2 shows the different F1 scores of the logistic classifier, which is trained using the embedding space as input. This provides an idea of how well the embedding space separates the classes and therefore gives a performance gain. The figure 2 shows clearly that the classifier trained on top of the log Mel spectrogram reached a higher F1 score. The metric further shows, that the classifier using the MFCCs embedding space fails to improve the F1 score over time, whereas the classifier for the log Mel spectrogram shows a definite increase over time.

4 Conclusion

From figure 1 it seems that the optimal feature representation is the MFCCs since it reached a significantly lower loss value. It further shows that the model would be able to benefit from further training since the model is still decreasing. However, when looking at the resulting F1 score of the classifiers (figure 2), the result shows, that even though the MFCCs representation reached a lower triplet loss, the classifier fails to separate the resulting clusters using a hyperplane.

This experiment shows that the optimal feature representation for the current thesis is using the log Mel spectrogram, since it resulted in a higher F1 score of the classifier, even though the triplet loss value is higher than the one using the MFCCs.

This is mainly due to the different nature of the feature representations. The MFCCs rep-

resentation input size (498, 13) is significantly lower than the log Mel spectrogram representation input size (498, 128). The MFCCs is a more compact representation, which seems to have a negative effect on the model's performance.

This experiment shows that the model benefits from using a representation which has more features and therefore, the log Mel spectrogram is used as the optimal representation for the thesis.

5 Next steps

For the next experiments, the feature representation log Mel spectrogram is chosen.