Study doc: tile size experiment

how does the size of the audio segments affect the embedding space

Fabian Gröger
fabian.groeger@stud.hslu.ch

Tuesday 5th May, 2020

Abstract

The experiment aims to show the effect of the size of the audio segments which will be used as the size of the triplets.

1 Introduction

The size of each triplet, which will be fed to the network, is an essential hyperparameter which needs to be carefully chosen. Because it specifies how much information each segment contains and is therefore fed to the network. If the segments are chosen too small, it does not contain enough information to distinguish between categories. If the size is too big, the segment contains too much information, and therefore, the model needs to work with a lot more data and gets a lot heavier. This experiment is conducted to search an optimal segment size for the triplets.

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 sample tile size will be evaluated for three different values `[1, 2, 4]` in seconds.

Table 1: Hyperparameters used for the experiment

Hyperparameter	**value**
Dataset	DCASE
Model	ResNet18
Epochs	20
Batch size	[32, 256]
Optimizer	Adam
Learning rate	1e-4
Margin	1.0
L2 regularisation amount	0.0
Embedding dimension	512
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	[1, 2, 5]
Sample tile range	[4, 5]
Convert to mono	True
Feature representation	
Feature extractor	LogMelExtractor
Frame length	480
Frame step	160
FFT size	1024
Number of Mel bins	128
Number of MFCC bins	13

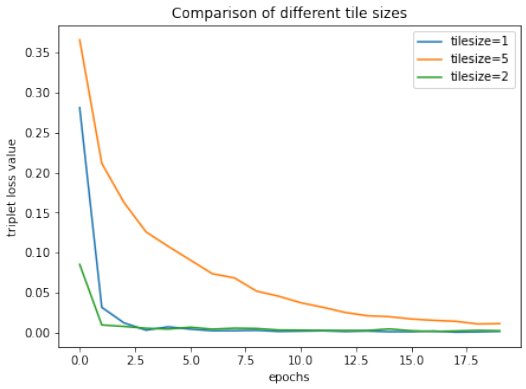

Figure 1: Plot of the triplet loss of the different tile sizes

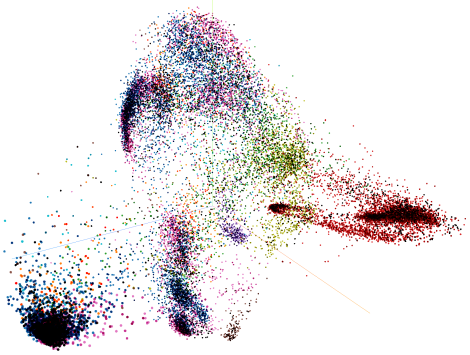

Figure 2: Visualisation of the embedding space from the tile size 5s

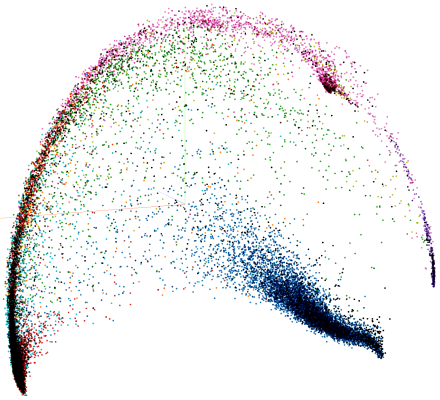

Figure 3: Visualisation of the embedding space from the tile size 1s

3 Results

Three models with the same hyperparameters, shown in table 1, were trained for 20 epochs. The value of the triplet loss is the primary evaluation criteria which will be used to compare the different triplet sizes since it has the most effect on this value because it should show how much of an audio sample is needed to distinguish between different classes. Figure 1 shows all the trained models in a single plot to visualise the impact of changing the sample size. Nevertheless, it is quite hard to interpret the different graphs since all of them are near zero.

The figure 1 shows that the tile size 5s has the highest loss value while the tile sizes 1s and 2s have relatively similar values. However, this does not mean that the tile size of 5s is the worst out of the three, it instead means that the model finds it harder to distinguish audio files when a larger sample is available, which is pretty evident since longer samples also contain more information.

The figure 1 also shows that the tile sizes of 1s and 2s have a very steep graph at the beginning and then hardly change their value. This indicates that it is relatively simple to achieve a good loss value with small audio samples, which means that the model can easily distinguish between small samples.

Both of these interpretations of the plot is pretty straight forward, but when the resulting embedding space is further examined using the Embedding Projector from the tensorboard, it can be seen that the tile size of 5s (2) results in much clearer clusters than the tile size of 1s (3).

4 Conclusion

If the optimal parameter for the sample tile is only chosen from the loss value and therefore, from the plot 1, it would be quite hard. However, since the visualisation of the embedding space shows a clear benefit in using a larger sample, the **sample tile size of 5s** is chosen to be the optimal one for the DCASE

dataset. This can be explained because smaller samples much often contain sounds which do not indicate a specific sound in that class, say for example there is a two-second silence in a sound file of the class eating, it would be projected in the nearby region of silence in a sound of a different class, which is useful for other applications but since the goal is to achieve a best possible embedding space, this is not a satisfying result. Therefore, the larger sound segments are more robust to such problems, since they hold much more information about the resulting class.

If the thesis focused on supervised triplet loss, it would make sense to cut the audio files in much smaller segments than in the unsupervised setting, since in supervised learning the triplet selection makes sure that the clustering focuses on the classes and not some other arbitrary criteria, which happens in unsupervised learning. In the unsupervised triplet loss, it is challenging to examine what exactly is being clustered, because there can be an underlying structure which can not be seen for us humans.

5 Next steps

The experiment has found an optimal parameter for the dataset and will be used from now on for the next experiments.