Lucerne University of
Applied Sciences and Arts

HOCHSCHULE
LUZERN

Informatik

Deep Embedded Music

Topic: Creation of an embedding space for noise detection and music
Student: Fabian Gröger
Advisor: Daniel Pfäffli
Expert: Dr. Jeremy Callner
Client: Hochschule Luzern - Informatik
Keywords: Audio signal, neural networks, machine learning, embedding, noise detection

1 Problem definition

For humans to distinguish between music genres is from easy to difficult, depending on how similar the given genres are. It is rather simple to distinguish music songs, for example, of classical music and techno or country and hip-hop rap. However, the difficulty increases if songs within the same genre are compared. When it comes to finding the similarity between songs or to sort a list of songs by their similarity, it is hard, even for humans.

In this thesis, an alternative learning strategy is developed that exploits basic semantic properties of sound that are not grounded by an explicit label. This approach will then be applied to a noise detection and a music dataset to evaluate its applicability and performance. Further, it attempts to show that the representation learned, provides a similarity space for both datasets, which can then be used to compare specific audios, sort them by similarity or find similarities between categories.

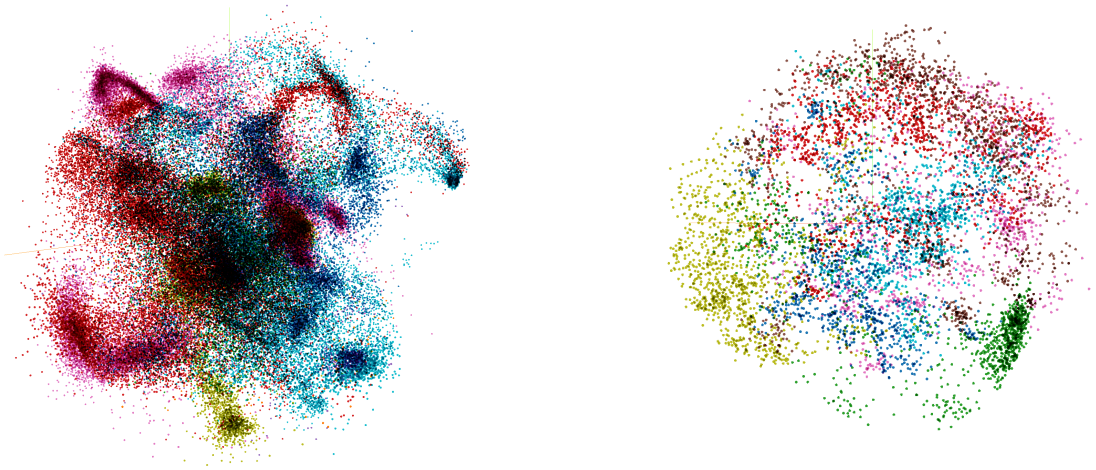
This thesis intends to adapt Tile2Vec, an image embedding method, to audio streams and evaluate its performance on the «SINS, DCASE 2018: task 5» noise detection dataset as well as on a music dataset, consisting of songs of seven sub-genres. Tile2Vec is an adaption of triplet loss to an unsupervised setting using the distributional hypothesis from natural language, which states that words appearing in similar contexts tend to have similar meanings. The idea of triplet loss is to create so-called triplets, which consist of an anchor, neighbour and opposite sample, and to learn a representation, where the distance from the anchor to the neighbour is smaller than the distance from the anchor to the opposite, which aims to represent similarity and dissimilarity. Therefore, triplet loss tries to minimise the distances between the anchor and the neighbour while maximising the distance between the anchor and the opposite.

2 Results

A simple linear logistic classifier is trained on the resulting embeddings, to compare the results to the ones from the DCASE challenge. The best classifier reached an F1 score of 62.19%, while most of the models submitted to the challenge accomplished results 80% and higher. However, the comparison to the challenge was just one part of the evaluation. The other part is the manual examination of the embedding space (1a), which aims to check if it has successfully learned a similarity space. The conducted error analysis revealed misclassified samples and microphones malfunctions in the dataset. Moreover, the embedding space revealed clusters of similar-sounding segments, regardless of its label. Figure 1a shows the resulting embedding space, where the colours represent the labels. For example, pink represents the label *watching_tv*.

For the music dataset, the same architecture was used to train an embedding space, which showed similar

results as for the noise detection space. It revealed clear clusters when applying k-Means to it and further indicated that the optimal amount of clusters to use for k-Means was seven, which is the number of classes in the dataset. Moreover, each cluster showed a significantly higher amount of segments of a specific class, which demonstrates that the model succeeded in building clusters for each genre. Further, it was observed that the neighbours of each segment, even if from a different category, build a neighbourhood, which has been confirmed in the thesis by a DJ. When training the simple classifier from above on the music embedding space, it achieved an astonishing F1 score of approximately 84% on unseen test data. The DJ, who performed the qualitative analysis, was thrilled and impressed with the results. Figure 1b shows the embedding space for music, where for example yellow represents the sub-genre *Techno_PeakTime_Driving_Hard*.



(a) SINS, DCASE dataset, where the colours represent the labels (e.g. pink is *watching_tv*)

(b) Music dataset, where the colors represent the genres (e.g. yellow is *Techno_PeakTime_Driving_Hard*)

Figure 1: Visualisations of the embedding spaces

3 Solution concept

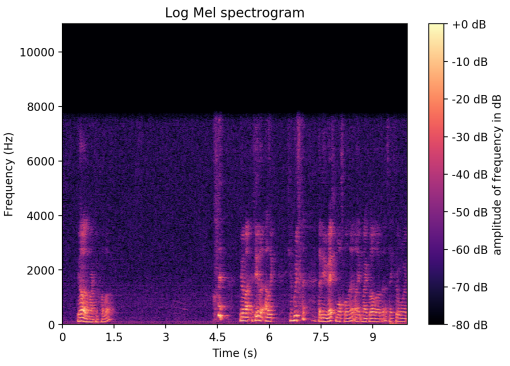


Figure 2: Log Mel spectrogram of a 10s audio from the label *social_activity*

The log Mel spectrogram is used, to represent each audio segment in a more compact form. It aims to illustrate the audio in the same form how humans perceive it. Figure 2 shows a log Mel spectrogram of an audio file from the DCASE dataset of the label *social_activity*, where one person is on the phone with somebody else. The x-axis of the figure shows the time, the y-axis the frequency and the colour represents the amplitude of each frequency at a given time. This representation was chosen because, when comparing different features, the use of the log Mel spectrogram showed a significant performance gain. Since the feature is a two-dimensional array, convolution neural networks can be used to create an even lower-dimensional representation. This lower-dimensional space is called the embedding space. This process is adapted from image processing and applied to audio streams.

Triplet loss is used to train the model to represent a meaningful embedding space. However, when using the general triplet loss function and calculating the loss of a batch during training, the resulting loss can swiftly converge to zero after a few training steps. Therefore, the model fails to learn the underlying structure of the data. To mitigate that problem, a triplet loss function with zero-filtering is used to train the model, which only calculates the mean of the triplet losses in each batch with non-zero entries. This results in a much higher and more consistent loss function, which can be used to train the model even for a very long period.

4 Special challenges

The comparison of different embedding architectures was challenging since there was no single score which could be compared. Therefore, the models had to be evaluated using a combination of metrics. Hence, the triplet loss value, the classifier F1 score and a manual examination of the embedding space were used for comparison.

One of the biggest challenges when training a triplet loss architecture is how these triplets are sampled. The triplet selection process has to be adapted to unsupervised learning since no labels are present. The adapted selection splits the audio samples into smaller segments and samples the anchor and the neighbour to belong to the same sample, whereas the opposite segment is chosen from a different random audio sample. Through the adaption, the algorithm tries to find an underlying structure to represent a similarity space. Therefore, it is inevitable that the model finds structures which are not represented by the label. For example, in individual experiments, it was observed that the model clusters «silence» from different labels in the nearby region of the embedding space. This led to a performance decrease when training the classifier. However, this behaviour is not wrong, since all of the segments indeed represent silence even if they correspond to different labels.

5 Outlook

The focus of the thesis was to provide a good baseline architecture for creating an embedding space for noise detection and music. Therefore, extensive hyperparameter tuning was not possible due to the lack of time towards the end of the thesis. Further research has to show how extensive hyperparameter tuning affects the embedding space. Furthermore, more extensive convolutional neural networks or recurrent convolutional neural networks should be used to train the embedding space to see if there is a performance gain in using them.

When comparing the achieved results from the classifier to the ones from the DCASE challenge 2018 task 5, they were near the results from the model of the last place. Which is a noticeable result, since the models' submitted to the challenge, use much larger classifier architectures than the ones used in this thesis. The results of the embedding space could be improved when using a more extensive classifier rather than a one layered softmax model. However, when looking at the result differently, they are quite impressive because the embedding space was able to find misclassified segments and even microphone malfunctions in approximately 200 hours of data.

The results of the music embedding space are astonishing, mainly because the model succeeds in finding the underlying structures in the songs, even in such entangled sub-genres from the music dataset. The results when classifying the embedded samples using a classifier show how well the clusters are separable, which further demonstrates the advantage of using the embedding space.