BMAT Assignment

28/04/2021

Nicolás Schmidt

Part 1 justification

The assignment was pretty straightforward and it makes clear for the applicants what skills are you looking for:

- database familiarity
- audio processing
- dataset creation
- deep learning using keras

My main concern in this assignment was the keras schema given, which involves 2D convolutions and MaxPooling operations. Given this information I understood that you wanted me to implement a convolutional neural network. However, I was confused about the 2D operations. Since most of the assignment is based on the construction of a featured based dataset, I understood that the audio file itself should be left out of the training process and just use the dataset created with the selected features computed by essentia to train and evaluate the model. I could be wrong on my assumption, and I would love to have some feedback regarding this point.

S1

Regarding the S1 part of the assignment, my main assumptions and the reasons that I took them were:

- I used torchaudio instead of librosa in order to make the script able to process .mp3 in a more flexible way. This is useful in case the dataset included some of this files, and since the assignment is against the clock, I prefered to use this powerful tool to avoid potential problems.
- I added tqdm library to have better feedback about the processing of the files
- I added all the arguments as ArgumentParse, to make the system more flexible
- I added an optional format parameter, in case that the script would be used to write files in another format.
- I added the database query to make more flexible the query over the SQLite database

S2

Regarding S2, I used Essentia to analyze the frames of each one of the audio files resampled in S1. For this I used a part of a code written by Frederic Font to split the audio files in equal length segments and process them individually to create a dataframe of the features for each frame. My final work considers 5 different features: Loudness, MFCC, MelBands, OnsetDetection and HPCP.

The most important feature for this particular task is the Loudness since the sum of speech and music will probably have a higher loudness for the audible class then the inaudible class. On the other hand, the MFCC features are known to be a very good timbre descriptor. Since the loudsense task will include music, the timbre characteristic and the diversity of the dataset could be added to the model by the use of these filters. Later I decided to add a MelBand analysis of the frames. This one was used to have a complement to the MFCC timbre features, but in terms of the energy on the Mel frequency bands. The onset estimator was chosen in order to allow a better frame selection according to the global and local beat. This makes sense specially for the audible class category. Finally, I incorporated a last timbre estimation based on the Harmonic Pitch Class Profile computed over the pitch peaks in the spectrogram.

With these features, which in total makes almost 80 features for each frame, I jumped into the model construction.

S3

The model is a basic fully connected neural network. I think this is my weak point: since we are working with numerical features for each frame, it did not make sense for me to make 2D convolutions. The main reason for this is that each datapoint is an independent audio frame with the respective features (1D). Maybe I missed something, but I would love to have some feedback about how 2D convolutions + max pooling could be used for this task.

The model is a 2 hidden layer fully connected neural network with batch normalization and dropout of 0.3. The activation function used was the ReLU for the hidden layers and I used a softmax activation function for the output layer. The first hidden layer has 120 neurons and the second 60 neurons. The configuration was chosen based on some experimentation.

The model was compiled using adam optimizer and a categorical cross entropy loss function, since this is a classification task.

S4

Part 4 was pretty straightforward, since the annotations and the previously generated featured based dataset generated in S2 were clear. The main responsibility of this script is to create the final dataset including the target class in the annotations. For this was really important to provide the model with the sample_rate used in S1 to resample the audios, in order to be able to map the start and end frames in the annotation csv to frames. With this information the only remaining work to do was to filter the frames and the respective features

by file and start/end frames to create the definitely dataset. The target classes were saved using the following dictionary:

```
class_dict = {
   "audible": 2,
   "barely": 1,
   "inaudible": 0,
}
```

S5

The train script was responsible for loading the dataset, compiling the model, train and save the trained model. Here, the script receives the directory of the final dataset compiled in S4, the directory to save the trained model, the number of epochs, the validation proportion and the batch size for training.

The method to load the data separated the dataset in the features and the target, makes the dummies for the target and drops the unused columns. Then the model is called and the training process begins.

Part 2: Questions:

Q1. What do you think are the problems that one should overcome when trying to manually annotate a dataset for the Audibility task?

A: The main problems are that this is a very exhausting task and is very difficult for the subjects to keep consistent annotations throughout the whole process. Also is very time consuming and expensive, since a lot of subjects are needed. I think that also the different loudness levels of the foreground/background sounds between two consecutive audios could affect the perception of the subject to be objective.

Q2. What solutions can you think of to the problems of questions Q1?

A: A source separation system between background and foreground audios could be used to measure the loudness of both signals. Using this information the dataset could be normalized to an equally tempered loudness between all the dataset. With this, the subject will be less affected by the different loudness across the dataset.

Q3. Synthetic audio can be generated by mixing two audio files where one contains music and the other contains non-music sounds such as speech, sound effects, environment and audience sounds, etc. What do you think are the advantages and problems of using a dataset where the music and non-music parts are synthetically mixed together?

A: This approach created dry signals that do not consider the acoustic factors of the environment where the two audio sources were recorded. In an extrem example one could

end up mixing a speech sound with a lot of room reverb with a very dry sound of birds. This kind of generated mix won't help the model to learn real-world sounds with its respective acoustic shared features between foreground and background sounds.

Q4. What solutions can you think of to the problems of Q3?

A: The source audios should be de-reverberated and then reverberated using the same impulse response of a real space. This way it would be like if both sounds were recorded in the same space.

Q5. If we could not solve the problems of Q1, how would they impact the training of an algorithm for the Audibility task with a manually annotated dataset?

A: Probably the accuracy of the system will be very bad and will not be able to correctly classify the audios. The model will not be able to correctly discriminate between audios with different loudness levels, and the differences between the loudness of the background and the foreground audios will not be learned by the model.

Q6. If we could not solve the problems of Q3, how would they impact the training of an algorithm for the Audibility task with a synthetically generated dataset?

A: If the test set used to evaluate the model is also created using a synthetic way, probably the evaluation will be very good. However the model will fail in real-life applications, where both the foreground and background audios are recorded using the same device and in the same acoustic environment.