ECS7013P - Deep Learning for Audio and Music Assignment Report: Chord + Melody Estimation

Joe Zacaroli

April 2020

1 Introduction

A chord estimation model has been created, trained and combined with a predominant melody extraction algorithm to create a chord+melody representation of the opening passage of a Chopin Ballade.

2 Data

The training data consisted of 890 chroma files extracted from songs in the US charts in the 1960s-1990s, with corresponding labelled chord files. The training data was taken from the McGill Billboard Project website[1] [1] and is one of the two datasets used in the MIREX Audio Chord Estimation task[2]. There are differing complexities of chord vocabulary labelling available for download. This work looked at three of the simpler sets of vocabularies which were 1) Major and Minor {N, maj, min}, 2) Seventh chords {N, maj, min, maj7, min7, 7}, 3) Major and Minor with inversions {N, maj, min, maj/3, min/b3, maj/5, min/5}. Each vocabulary includes the 'N' chord which represents no chord being played. There is a final category which wasn't analysed, consisting of Major and Minor seventh chords with inversions. With more time, this is an area that would be interesting to explore.

The Chopin Ballade that was analysed was downloaded from the OrangeFreeSounds website[3]. To extract chroma values in the same format as the training data, the Chordino VAMP plugin was used [2].

3 Machine Learning Methods

Two architectures were considered to perform the chord estimation. The first is a **SISO** (Single-Input Single-Output) model that takes a vector of 24 chroma values (12 bass register, 12 higher register) for a single time step and outputs a single chord label. The second architecture is a **MIMO** (Multiple-Input Multiple-Output) model that takes a time sequence of input chroma vectors and outputs a time sequence of corresponding chords. As the output data is categorical, the Pytorch Cross Entropy Loss function was used for both models when computing the loss. The total accuracy (correct chord predictions / total number of predictions) was also measured for training and test sets and when these started to diverge significantly, training was manually stopped.

[1] https://ddmal.music.mcgill.ca/research/The_McGill_Billboard_Project_(Chord_Analysis_Dataset)/
[2] https://www.music-ir.org/mirex/wiki/2019:Audio_Chord_Estimation
[3] http://www.orangefreesounds.com/chopin-ballade-4/

3.1 SISO Model

Training the SISO model was performed using a set of 50,000 training data points which were randomly sampled from the input data, which comprised of an estimated 3,000,000 points. A test set of 10,000 points was also randomly sampled from the same selection, therefore it is possible that some duplicates occurred. However, the size of the set that was being sampled meant that this was likely not much of an issue. A not insignificant amount of chords fell into the category of 'X', which meant that the current system of chord vocab was not complex enough to represent that chord. These were excluded from training and testing the SISO model.

Using one or two dense layers was investigated for the SISO model architecture. The reasoning behind this is that all the SISO model is doing is effectively template matching, therefore adding a more complex network is likely going to be detrimental to the performance. The SISO model managed to achieve an impressive test accuracy of up to 92% on the MajMinInv chord vocabulary. by inspecting some of the predictions, it was possible to see that most of the mistakes fell into one of a few fairly predictable categories:

- Same key center chord mistakes (F:maj vs C:maj)
- Major/minor substitutions (F:maj vs F:min)
- Relative major/minor substitutions (F:maj vs D:min)
- Semitone errors (F:maj vs Gb:maj)
- Inversion errors (F:maj vs F:maj/3)

Apart from the semitone error, all of these errors have a commonality which is that the predicted chord often shares a good proportion of notes with the ground truth, suggesting that the network is learning a reasonable mapping of chroma vectors to chord labels. Interestingly, much poorer results were obtained from the MajMin7 chord vocabulary set. This was probably due to the relative scarcity with which 7th notes are played in popular music, even when the chords are described as 7th chords. This observation lends itself well to the idea that some sort of temporal smoothing is required to extract chord information over longer periods of time.

Number Of Dense Layers	Test Accuracy
MajMin Vocab	
1	84%
2	72%
MajMin7 Vocab	
1	11%
2	3%
MajMinInv Vocab	
1	92%
2	87%

Table 1: SISO model test results for the different chord vocabularies.

3.2 MIMO Model

A natural way by which we can include temporal information in a network is through the use of recurrent layers. The main architecture changes I experimented with were whether or not to include a time-distributed (TD) layer before the recurrent layers, whether or not to include a fully connected (FC) layer after the recurrent layers, and how many recurrent layers to use, as well as the bidirectionality of the layers. Also

Input Size	Pre TD layer	LSTM Layers	LSTM Directions	Post FC layer	Test Accuracy
50	1	0	0	1	38%
50	0	1	1	0	55%
50	0	2	1	0	57%
50	0	1	2	1	53%
50	0	2	2	1	55%
50	0	3	1	1	56%
50	0	3	1	0	60%
50	0	4	1	0	61%
50	0	5	1	0	60%
50	1	1	1	0	56%
100	1	0	0	1	31%
100	0	1	1	0	58%
100	0	2	1	0	57%
100	0	1	2	1	51%
100	0	2	2	1	54%
100	0	3	1	1	53%
100	0	3	1	0	61%
100	1	3	1	0	57%

Table 2: MIMO Hyperparameter search results on the simple Major/Minor chord vocabulary.

investigated is the length of input data used. I experimented with input sizes of 50, 100 and 200, which correspond to about 2s, 4s and 9s of input audio. Input sequences were again randomly sampled from the middle of songs but this time no duplicates occurred in training and test.

Due to the complexity of the task, a larger empirical evaluation was only conducted for the MajMin chord vocabulary and sequence sizes of 50 and 100. Results are shown in Table 2. A smaller evaluation was conducted for the MajMin7 and MajMinInv chord vocabularies. These are shown in Table 3.

Input Size	Pre TD layer	LSTM Layers	LSTM Directions	Post FC layer	Test Accuracy
MajMin7 Vocab					
50	0	1	1	0	25%
50	1	1	1	0	10%
50	0	1	1	1	60%
MajMinInv Vocab					
50	1	1	1	1	50%
50	0	1	1	0	33%
50	0	1	1	1	60%

Table 3: MIMO Hyperparameter search on the Major/Minor/7th and Major/Minor/Inversion chord vocabularies.

By inspecting some of the output chord sequences for the testing set, it was possible to see a lot of the same sorts of errors cropping up as for the SISO model. However, more un-explainable wrong errors also seemed to be appearing. It's worth noting that whereas the SISO model had some un-guessable chord labels ('X') removed, this processing was not performed for the MIMO model. This could partially explain the worse performance of the MIMO model. However, on the whole the MIMO model did seem to predict a lot of chord sequences correctly, achieving a test accuracy of 60% across all chord vocabularies and input sizes. Note that it also managed to significantly outperform the SISO model on the MajMin7 vocabulary. The additions of the time distributed and fully connected layers did not seem to have a significant effect on the performance, and often made the over-fitting problem much worse than helping with the test accuracy.

4 Chopin Evaluation

For tractability, evaluation on Chopin has been reserved to just the first 70 seconds of the piece. SISO and MIMO models trained on two of the chord vocabularies were fed the Chopin file's chroma representation as their input and the outputs were recorded. The chord outputs were then smoothed using a median filter (kernel size = 9) to remove some of the spurious fluctuations. The overall amount of time that the correct chord was selected was then calculated (using the PerformanceTesting spreadsheet included in the submission). Results are shown in Table 4.

Architecture	Chord Vocabulary	Accuracy
SISO	Major/Minor	77%
SISO	Major/Minor/Inversions	78%
MIMO	Major/Minor	25%
MIMO	Major/Minor/Inversions	30%

Table 4: Chord Estimation Accuracy of the two architectures evaluated on the first 70s of the Chopin piece.

The Chopin piece contained quite a few diminished chords in the first 70 seconds, which were not included in the vocabulary. To account for this, if a chord was predicted that shared at least 2 notes with the diminished chord, it was counted as a match.

A reason why perhaps the MIMO method did not perform as well is that the size of the network is much larger, and a training set of comparable size to the SISO model was used in training. The MIMO method also will have learned temporal dependencies of chord sequences that are typical of 20th century pop music, which are undeniably much different to that of 19th Century Romantic piano works. It is interesting that temporal smoothing using a median filter was still able to improve the output.

A short 10s section in the middle of the piece with very slow and simple inverted chords was also briefly analysed to see if any chord inversions could be detected. One of the chord inversions was correctly detected, but most weren't, suggesting this is an important area that the model could improve upon.

Another area to note is that the prediction of onsets was not significantly harmed by the median filtering, and that both methods, when they guessed the right chord (which was not always), were very successful at determining the right chord onset times.

5 Chord + Melody Estimation

A robust melody extraction technique [3] was taken and the first 70s of the Chopin piece was processed with it. This method takes Time-Frequency representations of data and extracts the predominant melody. The output was put side by side with the highest performing chord outputs from the SISO model and placed in a spreadsheet in a 'scroll-down play-along' style (see the Chopin_TimesChordsAndMelody_Evaluation spreadsheet included in the submission). I played through and noted down which parts represented the piece well and which ones didn't. Overall I found that 65% of the opening 70 seconds of Chopin was well represented by my representation, meaning the top melody line and underlying chord progression sounded like that of the Chopin piece. The melody estimation algorithm did struggle with much of the piece, however, as there are often many different interleaved 'voices' playing at once. Unfortunately due to time constraints I was not able to perform more analysis of the two techniques together.

6 Discussion

There are a few notable areas that could be improved upon. Firstly, the Chopin chord and melody estimation evaluation in Sections 4 and 5 was done manually, which meant I could not do much of it. This really should have been automated, and would allow for a much more rigorous empirical evaluation of the deep learning methods created, and the combination of the two.

A larger chord vocabulary, especially one containing diminished chords, needs to be used in order to be effective at representing complex classical pieces. Additionally, the training set used would need to have a much more classically-representative selection of chords in order to be worth using.

The inability of the algorithm to detect inversions could potentially be fixed by using more and better training data, however it would be interesting to see if the network could be split into two parts, one deciding the bass note, and another deciding the rest of the chord name. This would have the added benefit of more closely matching the input chroma representation.

Finally, the output of the two algorithms could be presented in a much nicer way, such as in a scrolling format, so that you could watch it and play along with it yourself.

7 Conclusion

A chord estimation model has been trained using a set of pop music from the 1960s-1990s. It has then been combined with a predominant melody estimation algorithm and a chord+melody representation of the opening section of a Chopin Ballade has been created.

References

[1] John Ashley Burgoyne, Jonathan Wild, and Ichiro Fujinaga, 'An Expert Ground Truth Set for Audio Chord Recognition and Music Analysis', in *Proceedings of the 12th International Society for Music Information Retrieval Conference*, ed. Anssi Klapuri and Colby Leider (Miami, FL, 2011), pp. 633–38

[2] Matthias Mauch and Simon Dixon, 'Approximate Note Transcription for the Improved Identification of Difficult Chords', in *Proceedings of the 11th International Society for Music Information Retrieval Conference*, ed. J. Stephen Downie and Remco C. Veltkamp (Utrecht, the Netherlands, 2010), pp. 135–40

[3] D. Basaran, S. Essid, and G. Peeters, "Main melody extraction with source-filter nmf and crnn," in Proc. ISMIR, 2018.