EVALUATING A MULTI-TASK AND A SINGLE-TASK CRNN MODEL FOR AUTOMATIC DRUM TRANSCRIPTION

A PREPRINT

Teresa Pelinski
Master in Sound and Music Computing
Music Technology Group
Universitat Pompeu Fabra, Barcelona
teresa.pelinski01@estudiant.upf.edu

March 29, 2021

ABSTRACT

Multi-task learning can improve the generalization and robustness of data-driven models, assuming a statistical relationships between the tasks. Recent approaches to automatic drum transcription (ADT) have exploited the locally periodic nature of Western drum performances by training ADT systems jointly with the task of beat detection. While most ADT vocabularies contain only the three most common drum instruments (kick-drum, snare and hi-hat), training such models for larger vocabulary transcriptions encounters the difficulty that the available datasets do not offer a significant frequency of appearance of less usual instruments. A recent study trains a CRNN model with a synthetic dataset along with real datasets in order to balance the instrument occurrences. In this project, I compare the performance of the multi-task (MT) version of this CRNN model and its single-task (ST) version on the Groove MIDI Dataset. The results show no significant improvement in performance.

1 Introduction

Automatic Drum Transcription (ADT) is a subtask of Automatic Music Transcription (AMT). While many authors have proposed different models for the transcription of melodic instruments performances, there is not such abundant literature on the topic of ADT [1]. In general, drum instrument sounds are unpitched, percussive and transient, which makes algorithms tailored for melodic instrument transcription (e.g., based on fundamental frequency detection), unsuitable for the task of ADT. In this report, I will refer to transcription as the task of the onset detection and classification of drum sound events on the context of Western music.

The detection and classification of drum sound events can be of high complexity due to the strong overlap of different drum sound events in both time and frequency domain. This overlap is caused by the usual simultaneous playing of drum instruments as well as their broadband and noise-like spectra. For this reason, some approaches in ADT, such as the one that has been implemented in this project, do not only rely on spectral information but also on the locally periodic characteristics of drum performances, using methods similar to the implemented in speech and language processing problems [1].

Multi-task models jointly learn to perform several tasks. Assuming that there is an statistical relationship between the tasks (i.e., there is knowledge to be shared across them), such models often yield to a greater statistical strength and generalization [2]. In this project, a multi-task and a single-task implementation of the same model are compared for the task of ADT.

2 The CRNN architecture

The ADT system implemented in this project uses a CRNN architecture, that combines both a convolutional neural network (CNN) and a recurrent neural network (RNN). The CNN block focuses on the acoustic modeling of the events, while the RNN learns the temporal structures of the features. The CRNN architecture was proposed by Vogl et al. in 2017 [3] and jointly learns to the task of 3-voice transcription as well beat and down-beat detection.

Drum-transcription data-driven methods usually have a 3-voice vocabulary: kick-drums, snare-drums and hi-hats. These are the most frequent drum in performances and recordings, and despite some datasets have larger vocabularies, instruments different these three appear scarcely, which complicates the training of data-driven ADT systems. In order to address the data paucity for these less frequent instruments, Cartwright and Bello [4] create a synthetic dataset with a balanced vocabulary of 14 voices. In their work, a CRNN model is trained in both the synthetic and other available, and a single-task as well as a multi-task versions of the model are implemented[1]. The multi-task approach jointly learns three tasks: 14-voice transcription, 3-voice transcription, and beat-downbeat detection. The task of 3-voice transcription is included for the purpose of benefitting from 3-voice vocabulary datasets. The single-task implementation addresses the 14-voice transcription task.

3 Groove MIDI Dataset (GMD)

The multi-task and the single-task versions of the CRNN model trained on the synthetic and real data [4] will be evaluated on the Groove MIDI Dataset[2] [5]. This dataset contains 13.6 hours of aligned MIDI and synthesized drum performances, with more of the 80% of the duration coming from hired professionals covering a wide variety of styles. The performances are separated in two classes: 'beats' and 'fills'. While the 'beats' durations are of the order of minutes, the 'fills' correspond to shorter improvisations of just a couple of bars. Since these do not show a clear pattern or rhythmic structure, the model will be evaluated only on the 'beat' files: 444 performances of 10.3 hours in total. It should be noted that the CRNN model has not been trained in this dataset, so both the train, test and validation dataset split can be used for the evaluation.

Voice reduction Since the CRNN 14-voice model instrument vocabulary does not directly match with the 9 voices present in the GMD dataset, the CRNN 14 voices have been down-mapped to 8 voices. One of the voices present in the GMD dataset was discarded since it does not have a direct correspondence to the voices present in the CRNN model.

4 Evaluation metrics

In order to evaluate the performance of each variation of the model, the onsets in the automatic transcription will be compared against the ground truth (the MIDI transcriptions). This comparison is done for each instrument, comparing the obtained array of onsets for the instrument against the corresponding ground truth. Finally, the overall metrics of each model are computed by averaging the metrics across the instruments. The average is weighted by the occurrence of each instrument (i.e., how much onsets of each instrument are present in the ground truth).

The metrics used for the evaluation are F-measure (F), Precision (P) and Recall (R), defined by Eq. 1. A detected onset is counted as True Positive (TP) if its deviation from the corresponding ground truth onset is less than a pre-determined tolerance window. If a detected onset does not correspond to any annotated onset for the instrument, it is counted as False Positive (FP). Finally, if a ground truth onset does not correspond to any transcribed onset, it is counted as false negative (FN) [1].

$$F = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \qquad P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \tag{1}$$

These quantities have been computed using the onset evaluation function in the `mir_eval`[3] library [6], using a tolerance window of 50ms, which is a common value in the literature [1, 4] as well as the default value in the `mir_eval` onset evaluation function.

Threshold optimization For each timeframe and instrument, the CRNN model outputs an activation value between 0 and 1. In order to determine the threshold above which it can be considered that a drum sound event has been detected, an threshold optimization for each model and instrument has been conducted. Following an informal search for reasonable values, 29 threshold candidates are proposed, and for each model, 20 dataset performances are randomly selected. For each instrument, the threshold giving the best F-measure is averaged. Doing so, an optimal threshold for each model and instrument is obtained.

[1] https://github.com/mcartwright/dafx2018_adt

[2] https://magenta.tensorflow.org/datasets/groove

[3] https://craffel.github.io/mir_eval/#module-mir_eval.onset

5 Results and discussion

As it has been already mentioned, the dataset used for the evaluation contains 444 audio files and a total of 10.2 hours of expressive drumming. Table 1 presents the results of the evaluation by instrument and the average scores for both the multi-task (MT) and the single-task (ST) models.

Table 1: Evaluation scores for both Multi-task (MT) and Single-task (ST) CRNN models. W-Average refers to the average across instruments weighted by the occurrence of each instrument.

Drum Instrument	F-Measure		Precision		Recall		Occurrence
	MT	ST	MT	ST	MT	ST	
Kick-Drum	0.22	0.23	0.18	0.21	0.30	0.28	67335
Snare-Drum	0.36	0.35	0.43	0.45	0.33	0.29	102642
Crash	0.02	0.02	0.08	0.12	0.02	0.01	3742
Ride + Bell	0.16	0.18	0.40	0.40	0.11	0.12	40068
Open Hi-Hat	0.06	0.02	0.21	0.19	0.04	0.01	9538
Closed Hi-Hat	0.37	0.32	0.35	0.38	0.39	0.29	88101
Low/Mid-Tom	0.06	0.01	0.12	0.07	0.06	0.01	3976
High-Tom	0.03	0.04	0.12	0.11	0.02	0.03	9574
W-Average	0.28	0.27	0.33	0.35	0.29	0.25	

Overall, the results are remarkably low. In the evaluation conducted by Cartwright and Bello [4], the scores lie between 0.67 and 0.70. These do not reflect an excellent performance, but are still considerably better than the results presented here. The closed hi-hat (cHH), snare-drum (SD) and kick-drum (KD) show the highest F-measures for both the MT and the ST models, which also occurs in [4]. This can be explained by the abundance of these three instrument in the real (non synthetic) datasets used in the training. For the rest of the instruments, the F-measures decrease below 0.10, with the exception of the Ride+Bell instrument, which has a greater frequency of occurrence. The low overall scores for the drum instruments different than the KD, SD and cHH also agree with the results presented in [4], where it is argued that in order to exploit the synthetic data it needs to be used jointly with real data. The differences in performance between the MT and the ST model are not large: the MT model shows only an improvement of 0.01 in the F-measure and 0.04 in the Recall, while it is surpassed by 0.02 in the Precision score. In [4], it is stated that the MT appears to hinder performance of the 14-voice ADT task. Here, a clear decrease in the performance is unclear, since the F-score is slightly better for the MT model, but has a lower precision than its counterpart.

The considerably low evaluation scores suggest that there might be errors in the processing of the data or the evaluation procedure. This has been carefully revised and the process can be reproduced in the jupyter notebook[4] accompanying this report. The time window used for the evaluation was 50ms, a reasonable value according to the literature and generous to the observed misalignment between the MIDI transcription and the model's transcription (10ms). The threshold has also been optimized for each model and instrument. For instruments not appearing in the data subset used for the optimization, the lowest threshold among the other instruments was chosen. In order to improve the optimization for such instruments, a better approach could be to look for a significant number of performances where those instruments appear and perform the threshold optimization among them. A bad down-mapping could also justify the low evaluation scores in some instruments, but these factors do not explain the bad performance in the most usual instruments (KD, SD, cHH), that were present in the threshold optimization subset and not down-mapped.

Regarding the evaluation metrics, it should be noted that this evaluation does not address the misclassification of an instrument. Since the model does not first detect the onset and then classify the instrument, but jointly detects and classifies the drum sound event, an instrument misclassification can not be directly evaluated. Moreover, the literature of ADT transcription systems does not usually report a misclassification evaluation [1, 3, 4]. A possibility for future work could be to evaluate the models' performances with recordings with only one instrument and check if the transcription includes other instruments.

In conclusion, the CRNN model does not provide acceptable results for the task of ADT. Moreover, the MT approach does not seem to show a significant improvement on the ADT task. The results presented here are significantly lower than the original paper [4], and the aforementioned difficulties in ADT regarding limited instrument vocabularies persist. Some factors that could have induced to errors in the evaluation have been identified, and a procedure for evaluating the misclassification of instruments has been proposed.

[4] https://github.com/pelinski/crnn_multitask_adt/blob/main/automatic_drum_transcription.ipynb

References

[1] Chih Wei Wu, Christian Dittmar, Carl Southall, Richard Vogl, Gerhard Widmer, Jason Hockman, Meinard Muller, and Alexander Lerch. A Review of Automatic Drum Transcription. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 26(9):1457–1483, 2018.

[2] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[3] Richard Vogl, Matthias Dorfer, Gerhard Widmer, and Peter Knees. Drum transcription via joint beat and drum modeling using convolutional recurrent neural networks. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017*, pages 150–157, 2017.

[4] Mark Cartwright and Juan Pablo Bello. Increasing drum transcription vocabulary using data synthesis. In *DAFx 2018 - Proceedings: 21st International Conference on Digital Audio Effects*, pages 72–79, 2018.

[5] Jon Gillick, Adam Roberts, Jesse Engel, Douglas Eck, and David Bamman. Learning to groove with inverse sequence transformations. *arXiv*, 2019.

[6] Colin Raffel, Brian McFee, Eric J. Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel P.W. Ellis. mir_eval: A transparent implementation of common MIR metrics. In *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014*, pages 367–372, 2014.