MIREX 2018 DRUM TRANSCRIPTION CHALLENGE: SOFT-SUPERVISED LEARNING

Julien Schroeter
Cardiff University
schroeterj1@cardiff.ac.uk

Kirill Sidorov
Cardiff University

David Marshall
Cardiff University

ABSTRACT

This extended abstract summarizes the key contributions and results of our MIREX 2018 drum transcription submission. The main novelty resides in the softly-supervised nature of the implemented training regime and architecture.

1. TASK

The MIREX 2018 drum transcription challenge contains two tasks: the classical 3-class (bass-drum, snare-drum, hit-hat) introduced in 2005 and a new extended 8-class challenge. Methods are assessed on their ability to predict as accurately as possible – within a 30ms margin – the occurrence of hits for each of the drum types.

Dataset The dataset is comprised of a rich collection of smaller datasets. The detailed list of datasets provided for the challenge can be found on the official MIREX website [1] .

It seems relevant to point out that the new 8-class dataset is quite imbalanced. Indeed, the 5 additional drum classes only account for approximately 15% of the overall number of hits; in contrast to the 50% occurrence rate of the hi-hat class. Moreover, tom-toms, cowbells and slaves are nearly absent from the MEDLEY subdataset with only 30, 16 and 0 occurrences respectively. A more diverse and balance dataset seems key to make this new task as relevant as the classical 3-class challenge.

Benchmark Benchmarks are only available for the 3-class task, as the 8-class challenge has been newly introduced this year. The following two benchmarks will be considered:

- *Vogl'17*: The current best submission [2].
- *MIR'05*: The previous task winning results from MIREX 2005 [2] .

	M'05	V'17	TOP'18	JS2
IDMT	**75**	66	66	59
KT	61	**65**	**65**	63
RBMA	-	**72**	**72**	57
MDB	-	**73**	68	59
GEN	-	78	**81**	73
Overall	67	**71**	69	62

Table 1. F-measure results for the 3-class task [%]. Only the performances on the MIREX 2018 evaluation dataset are reported.

2. SUBMISSION

As the soft-learning approach implemented for this challenge has not been published yet, this section will only outline the main concepts behind the method. Of course, more detailed explanations will appear in a future works.

Main Idea The proposed model is characterized by two main components: a soft prior learning unit and a mass-shrinking unit. More precisely, the former roughly learns the localization of notes with a certain level of tolerance, while the latter compresses the mass toward single points to obtain precise temporal localizations. In the context of this challenge, this soft approach seems more natural than forcing the model to learn the exact time of each occurrence. Indeed, as the biggest part of the dataset has been hand-labeled, expecting a millisecond precision seems not only optimistic, but can certainly affect the training negatively.

Data augmentation Data augmentation is a key component when training robust deep learning pipelines; models for audio-based tasks are no different [1]. Consequently, our submission includes pitch shifting and time stretching transformations. In order to obtain more robust results, individual predictions made over a wide range of transformation are aggregated to produce the final output.

Model ensembling As an additional effort to reduce the variance of predictions, an ensemble of models is used for inference. However, as our data augmentation is computationally intensive, eventually only two models have been selected for each of the two tasks.

[1] http://www.music-ir.org/mirex/wiki/2018:Drum_Transcription

[2] http://www.music-ir.org/mirex/wiki2005:Audio_Drum_Detection_Results

3. RESULTS

The results are summarized in Table 1 and Table 2. A complete review of this year's results with additional extended abstracts can be found on the MIREX website [3]. Overall, the results are promising except for the two most challenging classes RBMA and MDB. The fragility of the proposed model under rich and loud background noise likely indicates a lack of generalization of the representation learning part of the network. It seems indeed that the initial choice of selecting a minimal number of filters to avoid overfitting has restricted the model too much in this case. Hence, future submission will attempt to remedy this issue.

	M'05	V'17	TOP'18	JS2
RBMA	-	-	**55**	50
MDB	-	-	**65**	61
MIDI	-	-	**66**	61
Overall	-	-	**62**	58

Table 2. F-measure results for the 8-class task [%]. Only the performances on the MIREX 2018 evaluation dataset are reported.

4. DISCUSSION

The overall results are fair for a first participation. Some potential improvements include a better architecture for the representation learning part of the network, a more efficient implementation to allow more extensive model ensembling as well as an improved initial feature preprocessing pipeline.

5. REFERENCES

[1] Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283, 2017.

[2] Richard Vogl, Matthias Dorfer, Gerhard Widmer, and Peter Knees. Drum transcription via joint beat and drum modeling using convolutional recurrent neural networks. *In Proc. Int. Soc. Music Inf. Retrieval Conf.*, pages 150–157, 2017.

[3] http://www.music-ir.org/mirex/wiki/2018:Drum_Transcription_Results