DL Midsem Evaluation Winter 2023
Topic: Music Beat Position Estimation

Anonymous ACL submission

Abstract

This document is the mid-semester project evaluation report of group number 4, whose members are namely Aadit Kant Jha (2020001), Niranjan Sundararajan (2020090) and Sahas Marwah (2020237) of Indraprastha Institute of Information Technology, Delhi (IIIT-D). The task is to predict the beat and downbeat positions given a music audio sample.

1 Problem Statement

The problem statement is to accurately and efficiently detect beats and downbeats in audio recordings, a critical task for various applications such as music production, performance, and analysis.

Traditional methods based on handcrafted features and statistical models like HMMs (Hidden Markov Models) or DBNs (Dynamic Bayesian Networks) have limitations in dealing with the complexity and variability of musical rhythm for genres with intricate and irregular patterns. (Böck et al., 2019) Hence, the motivation of Deep Learning (DL) methods for beat estimation is an important improvement in the field.

The papers (Böck et al., 2020) propose novel solutions using DL techniques, including Temporal Convolutional Networks (TCN) and Multi-task Learning (MTL) Frameworks, to capture the temporal dynamics and interdependencies among different rhythm-related features in time and frequency domains like the non-linear change of beats with respect to the melody, harmony, or lyrics.

Other approaches (Heydari et al., 2021) adopt a Convolutional Recurrent Neural Network (CRNN) to establish causality in the time domain using recurrence and employ frequency-based learning methods using convolutions.

2 Related Work and Existing Baselines

2.1 Baseline 1

2.1.1 Existing Baselines

For the first baseline, we have followed a series of three papers:

- Böck and Davies (2019)- TCN for Musical Audio Beat Tracking
- Böck et al. (2019) - Multi-task Learning of Tempo and Beat
- Böck and Davies (2020) - Deconstruct, Analyze, Reconstruct

Böck and Davies (2019) propose a novel approach based on TCNs for beat tracking, which outperforms previous baselines based on handcrafted features and statistical models.

Böck et al. (2019) propose an MTL approach that learns beat and tempo estimation together and shows that learning one task can improve the accuracy of the other.

Böck and Davies (2020) propose a modular framework that decomposes the problem of tempo, beat, and downbeat estimation into three sub-tasks and combines the results in a consistent and efficient manner.

If seen together in a flow of work that progresses towards creating SOTA models one after another, the baseline focuses on DL-based approaches using CNNs and TCNs for music analysis. They discuss various architectures and techniques that have been proposed for beat and tempo tracking, including Recurrent Neural

Networks (RNNs) and graph-based models for detecting the beats from activations emitted by the network.

2.1.2 Related Work

For Böck and Davies (2019):

- Böck and Schedl (2017) proposed a beat tracking algorithm using a CNN-based approach that relied on hand-crafted features.
- McFee et al. (2018) proposed a beat-tracking algorithm using a hybrid approach that combined deep learning and traditional machine learning methods.
- Daudet et al. (2019) proposed a beat-tracking algorithm using an RNN-based approach incorporating temporal dynamics.

For Böck et al. (2019):

- Böck et al. (2019) proposed a DNN-based MTL approach accompanied by separate networks for each task but also used hand-crafted features.
- Sun et al. (2020) proposed an RNN-based MTL approach that jointly learned tempo and beat from audio signals without relying on hand-crafted features.

For Böck and Davies (2020):

- Krebs and Böck (2013) proposed a beat-tracking algorithm incorporating additional information such as meter and tempo changes.
- Bozkurt et al. (2018) proposed a downbeat estimation algorithm using a CNN-based approach that relied on hand-crafted features.

2.2 Baseline 2 - BeatNet

2.2.1 Existing Baselines

(Heydari et al., 2021) present an online (causal) system for beat, downbeat, and meter tracking that uses causal convolutional and recurrent layers. The system performs better than other online beat/downbeat tracking systems and performs comparably to other offline methods.

BeatNet comprises two stages: One NN stage that produces activations and one particle filtering stage for inference (we haven't used this in our implementation of baseline as our task is only for offline inference). The NN includes convolutional and recurrent fully connected layers and utilizes the CRNN architecture to process input features and get beat/downbeat activation functions.

2.2.2 Related Work

- Durand et al. (2016) obtained downbeats through features and CNN structures.
- Giorgi et al. (2020) did downbeat tracking by tempo-invariant convolutional filters.
- Böck and Krebs (2016) employed an RNN structure to do beat and downbeat tracking jointly.
- Fuentes et al. (2018) when taking input as a tatum matrix, CRNNs perform better than RNNs in downbeat tracking.
- Böck and Davies (2019) enhanced the performance of offline beat and downbeat tracking using CNNs and TCN.

3 Dataset Details

The GTZAN dataset is a collection of labeled audio files that can be used for classifying music genres. It has 1000 audio clips of 30 seconds each, spread into 10 different genres. These audio clips were sourced from commercial sources and were processed to have a bit depth of 16 bits and a sampling rate of 22050 Hz.

4 Experiment Setup and Results

The reproducibility of our models is contingent on the training which was done for the original model. Since the original models were trained over different datasets, exact reproducibility was not achieved. However, all the results are well within the ballpark measures.

4.1 Baseline 1

4.1.1 Experiment Setup

For baseline-1, we followed the ISMIR 2021 Tutorial for Beat, Downbeat tracking that utilizes the modified architecture proposed in Böck et al., (2019).
The entire experimentation was done on Google

Colab implemented using TensorFlow. The network was trained for 100 epochs with a patience of 20. The train-val-test split of 70-15-15. RAdam optimizer was used with a learning rate of 0.05, with a reduction applied in case of plateauing. The objective function used was a modified masked Binary Cross Entropy Loss over all 3 channels of beat, downbeat, and tempo.

The original paper reported an F-measure of 0.883, CMLt of 0.808, and an AMLt of 0.930 for beat tracking. For downbeat tracking, the reported values are 0.654, 0.619 and 0.817, respectively, for the GTZAN dataset.

4.1.2 Results

The model outputs activations for beats, downbeats, and tempos. These activations are passed through a DBN Beat and Downbeat Tracking Processor proposed in Böck et al., (2016). We then obtain the timestamp and position of beats and downbeats. On evaluation, the system roughly matches the baseline scores reported in the results of the paper Böck et al., (2020). The F-measure for beat is 0.882, and that for downbeat is 0.579.

4.2 Baseline 2

4.2.1 Experiment Setup

The entire experimentation was done on Kaggle implemented using PyTorch. The network was trained for 100 epochs with a patience of 20. The train-val-test split of 70-15-15. Adam optimizer was used with a learning rate of 5 x 10^{-4}. The objective function used was a weighted Cross Entropy Loss over all 3 channels of beat, downbeat, and nonbeat.

The original paper reported an F-measure of 0.8064 for beats and 0.5407 for downbeats on offline inference over GTZAN Dataset.

4.2.2 Results

The model outputs activations for beats, downbeats, and nonbeats. These activations are passed through a DBN Beat and Downbeat Tracking Processor proposed in Böck et al., (2016). We then obtain the timestamp and position of beats and downbeats. On evaluation, the system roughly matches the baseline scores reported in the results of the paper (Heydari et al., 2021). The F-measure for beat is 0.818, and that for downbeat is 0.571.

	Baseline-1		Baseline-2	
	Beats	Downbeats	Beats	Downbeats
F-measure	0.882	0.579	0.818	0.571
Cemgil	0.829	0.548	0.758	0.532
CMLc	0.785	0.552	0.668	0.562
CMLt	0.809	0.552	0.681	0.562
AMLc	0.888	0.817	0.855	0.815
AMLt	0.919	0.819	0.869	0.818

Table 1: Results of our implementation of the baseline.

5 Observations and Future Work

5.1 Baseline 1

As stated previously, the first baseline follows a series of three papers. The future work of the first paper is the approach/improvement of the second paper, and so on.

5.1.1 Observations

- The use of TCNs allows the model to capture temporal dependencies so it is able to accurately track beats of many musical genres, indicating its robustness (Böck et al., 2019).
- The proposed MTL (Böck et al.,2019)approach allows the model to leverage the correlation between tempo and beat information, improving accuracy and robustness in both tasks.
- The proposed approach (Böck et al., 2020) decomposes the problem into three sub-tasks, allowing the model to focus on each sub-task separately and combine the results efficiently, outperforming previous SOTA method's accuracy and robustness.

5.1.2 Future Work

- Further research could investigate using variants of TCNs, and improve the model's robustness to tempo and rhythm variations.
- Techniques such as data augmentation or transfer learning could be used to improve the generalizability of the model to different musical genres and cultures.
- The proposed framework could be evaluated and compared with other SOTA approaches in affiliated domains, such as speech or natural language processing, to identify similarities and differences in the underlying principles and techniques.

5.2 Baseline 2

5.2.1 Observations

- The proposed BeatNet model achieves SOTA performance on the task of joint beat, downbeat, and meter tracking and achieves superior performance for both online.

- The CRNN architecture results in a more robust model than traditional methods.

5.2.2 Future Work

- The current implementation of BeatNet requires a large amount of computational resources. Future work could focus on optimizing the model for faster and more efficient processing.

- The proposed method is evaluated on a single dataset, and it remains to be seen how well it generalizes to other musical genres and styles.

6 References

1. Davies, E. P.; Bock, S. Temporal Convolutional Networks for Musical Audio Beat Tracking. *2019 27th European Signal Processing Conference (EUSIPCO) 2019.*

2. Böck, S.; Davies, M.; Knees, P. MULTI-TASK LEARNING of TEMPO and BEAT: LEARNING ONE to IMPROVE the OTHER.

3. Böck, S.; Davies, M. DECONSTRUCT, ANALYSE, RECONSTRUCT: HOW to IMPROVE TEMPO, BEAT, and DOWNBEAT ESTIMATION.

4. Heydari, M.; Cwitkowitz, F.; Duan, Z. BEAT-NET: CRNN and PARTICLE FILTERING for ONLINE JOINT BEAT DOWNBEAT and METER TRACKING.

5. M. Fuentes, B. McFee, H. C. Crayencour, S. Essid, and J. P. Bello, "Analysis of common design choices in deep learning systems for downbeat tracking," in *Proc. of the 19th Intl. Society for Music Information Retrieval Conf., 2018, pp. 106–112.*

6. S. Durand, J. P. Bello, B. David, and G. Richard, "Feature Adapted Convolutional Neural Networks for Downbeat Tracking," in *Proc. of the 41st IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, 2016, pp. 296–300.*

7. S. Böck, F. Krebs, and G. Widmer. Joint beat and downbeat tracking with recurrent neural networks. *In Proc. of the 17th Intl. Society for Music Information Retrieval Conf. (ISMIR), pages 255–261, 2016.*

8. J. Hockman, M. E. P. Davies, and I. Fujinaga, "One in the Jungle: Downbeat detection in Hardcore, Jungle, and Drum and Bass," in *Proc. of the 13th Intl. Society for Music Information Retrieval Conf., 2012, pp. 169– 174.*

9. D. Ellis, "Beat tracking with dynamic programming," *Journal of New Music Research, vol. 36, no. 1, pp. 51– 60, 2007.*

10. S.Böck and S. Schedl, "Enhanced beat tracking with context-aware neural networks," in *In Proc. of the 14th International Conference on Digital Audio Effects (DAFx-11), 2011, pp. 135–140.*

11. B. D. Giorgi, M. Mauch, and M. Levy, "Downbeat tracking with tempo-invariant convolutional neural networks," in *In Proc. of the 17th Intl. Conf. on Music Information Retrieval (ISMIR), 2020, pp. 216–222.*

12. F. Krebs, S. Böck, and G. Widmer., "Rhythmic pattern modeling for beat and downbeat tracking in musical audio," in *Proc. of the 14th Int. Society for Music Information Retrieval Conf., 2013.*

13. F. Krebs, S. Böck, M. Dorfer, and G. Widmer, "Downbeat tracking using beat-synchronous features and recurrent neural networks," in *In Proc. of the 17th Intl. Conf. on Music Information Retrieval (ISMIR), 2016.*