Deep Learning for Automatic Downbeat Tracking

Dale Luginbuhl
Department of Computer Science
University of Cincinnati
Cincinnati, USA
luginbdr@mail.uc.edu

Atharva Pingale
Department of Computer Science
University of Cincinnati
Cincinnati, USA
pingalac@mail.uc.edu

Rithvik Reddy Sama
Department of Computer Science
University of Cincinnati
Cincinnati, USA
samary@mail.uc.edu

Lewis Thelen
Department of Computer Science
University of Cincinnati
Cincinnati, USA
thelenlr@mail.uc.edu

I. Introduction

By endowing computers with the ability to listen to music we unlock a number of applications for computer automation in music that were previously unavailable. This field of study is referred to as Music Information Retrieval (MIR). MIR is broken down into a number of overlapping problems as described in [1]. Examples include:

- Beat and downbeat tracking
- Tempo, time signature, and key estimation
- Melody tracking
- Chord detection
- Music structure analysis
- Musical onset/offset detection
- Mood or emotion recognition

We explore one of these problems, downbeat tracking, in more detail. Music is typically organized temporally around a rhythmic pulse. Each occurrence of the pulse is referred to as a *beat*. The goal of beat tracking is to annotate the locations of these beats in an excerpt of music. Beats can further be organized into *measures*, where each measure contains the same fixed number of beats. The first beat of each bar is typically emphasized, and this beat is known as the *downbeat*. The number of beats per measure, and the emphasis on the downbeat give a piece of music a certain feel, and are important qualities that shape how we interpret the music. The goal of downbeat tracking is to annotate the location of downbeats in an excerpt of music.

Successfully annotating the location of downbeats is important for two reasons. One, it can be used directly in applications. For example it could be used to analyze the performance of a practicing musician and provide feedback on how well they were able to maintain a consistent rhythmic pulse. Or it could be used to create sync points to allow an editor to easily synchronize a video or other multimedia application to the music. The second reason that downbeat tracking is important, is that it can be used as a pre-processing step in solving higher-level MIR tasks. If we have annotations of downbeat locations in an excerpt of music, these locations provide us with a temporal framework of the excerpt. This can simplify tasks like music structure analysis, melody tracking, or musical onset/offset detection, because these events typically occur on or around these beat and downbeat locations.

II. Related Work

A. Datasets

Various kinds of dataset are available from the various sites. All the datasets extracted from the various sites are in Waveform Audio File Format (.wav format). Ballroom with a duration of 05 hours 57 minutes and Harmonix dataset with a duration of 56 hours, SMC dataset has 02 hours 25 minutes, RWC-POP dataset has 06 hours 47 minutes, Beatles dataset has 08 hours 09 minutes, Hainsworth dataset has 03 hours 19 minutes, Simac dataset has 03 hoours 18 minutes, HJDB dataset has 03 hours 19 minutes and GTZAN dataset has 08 hours 20 minutes . Each dataset contains the same feature/label configuration. They consist of .wav audio files, each with a corresponding label file that lists the location of each beat in seconds. An example audio file from the Ballroom dataset with only the downbeat labels shown is given below. Each dataset will be divided into 60%- 20%-20%, where 60% is used for the training and the 20% is used for testing and rest 20% is used for validation purpose.

B. Models

Previous models have employed a wide variety of strategies in both beat and downbeat tracking (we include mention of beat tracking only because beat tracking techniques often generalize to the special case of downbeat tracking or inadvertantly occur as an intermediary step in the process of downbeat tracking). Earlier models [2], [3], [4] relied heavily on feature engineering to extract relevant events from the music (ie chord changes, accents etc.) in order to identify beat patterns within the audio file. More recent models have taken advantage of ”data-driven” approaches, made possible by the rise of deep learning techniques, which allow for the model itself to learn and then extract relevant features from the raw audio file.

Unsurprisingly, of the deep learning approaches used in downbeat tracking, sequential models have found the best results. Earlier sequential models have used more vanilla sequential techniques such as Long Short Term Memory (LSTM) applied directly to the raw audio data as in the case of [5] others have found success with more complicated approaches such as Convolutional Recurrent Neural Networks (C-RNNs) applied to spectrogram representations of the audio data as in the case of [6]. However, such models suffer from issues with vanishing gradient when faced with longer music segments as is common with recurrent models operating on time series data. More contemporary models have tried to overcome these issues by employing the then novel Temporal Convolutional Network (TCN) architecture most notably in the case of [7]. The model employed convolutional filters across the temporal dimension of the audio file and achieved what was until recently state of the art performance.

Current state of the art models have taken advantage of recent advances with transformer based architectures to achieve even higher performance at downbeat tracking as well as more general MIR tasks such as chord change detection and music tagging than previous RNN and TCN models. Within transformer based downbeat tracking two divergent strategies appear to have emerged. The first engages in more explicit feature extraction as in the case of [8] where the data from individual isntruments within the audio track are extracted and then fed together into a muli-headed attentiont transformer. The seconds pushes further with the "data-driven" approach by relying on the model architecture to extract relevant features from the raw audio as in the case of [9] where novel Transformer in Transformer (TNT) blocks were utilized to mutually inform both spectral level and temporal level analysis of a raw audio file. Our current study aims to push the performance of this so called SpecTNT by re-incorporating some degree of data engineering.

C. Challenges

Beat and downbeat tracking in music analysis pose several significant challenges. Early heuristic-based methods faced limitations in accurately detecting beats, relying on manually encoded rules that may not generalize well across diverse musical styles. Data-driven approaches, while effective, require substantial annotated data for training, which can be both costly and limited. Downbeat tracking is a challenging task because it often relies on other sub tasks such as beat tracking, tempo and time signature estimation and also because of the difficulty to state an unambiguous ground truth [8] [18]. Additionally, incorporating local spectral information and facilitating the exchange of critical local details across disparate temporal positions are formidable tasks. Both beat and downbeat tracking also grapple with the scarcity of annotated data, a prevalent issue in training data-intensive models like Transformers [8] [14]. The generalization problem of deep learning models, which may fail to adapt to unseen or novel music pieces with different time signatures, tempo ranges, or musical characteristics than the training data [8].

D. Future Work

In the context of enhancing the performance of SOTA model for downbeat, several crucial strategies have been employed. Firstly, a pivotal step involves demixing audio sources to effectively isolate distinct instruments prior to feeding the data into the network [8]. Moreover, we seek to enhance beat tracking by introducing instrumental attention among drum, piano, bass, vocal, and other demixed sources. This preprocessing step proves instrumental in disentangling overlapping sounds, providing the network with clearer and more discernible input. Additionally, dilated self attention models have been claimed to demonstrate powerful sequential modelling with linear complexity, potentially adaptable to more general MIR tasks [8]. Due to the limited datasets with beat and downbeat annotations, data augmentation might be required in curbing overfitting. These augmentations serve to diversify the training data, thereby enhancing the model's adaptability to a broader range of musical contexts. Some of fairly common data augmentations, each of which has an associated probability p of being applied to each training example during a training epoch. This includes the applications of highpass and lowpass filters with random cutoff frequencies (p = 0.25), random pitch shifting between -8 and 8 semitones (p = 0.5),additive white noise (p = 0.05), applying a tanh nonlinearity (p = 0.2), shifting the beat locations forward or back by a random amount between ± 70 ms (p = 0.3), dropping a contiguous block of audio frames and beats of no more than 10% of the input (p = 0.05), as well as a random phase inversion (p = 0.5) [13].

References

[1] Choi, K., Fazekas, G., Cho, K., and Sandler, M., "A tutorial on deep learning for music information retrieval." arXiv preprint arXiv:1709.04396, 2017.

[2] Dixon, Simon (2001) Automatic Extraction of Tempo and Beat From Expressive Performances, Journal of New Music Research, 30:1, 39-58

[3] A. P. Klapuri, A. J. Eronen and J. T. Astola, "Analysis of the meter of acoustic musical signals," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 1, pp. 342-355, Jan. 2006

[4] Goto, Masataka. (2002). An Audio-based Real-time Beat Tracking System for Music With or Without Drum-sounds. Journal of New Music Research

[5] S. Böck, F. Krebs and G. Widmer, "Joint beat and downbeat tracking with recurrent neural networks", Proc. Int. Soc. Music Inf. Retrieval Conf., pp. 255-261, 2016

[6] Tian Cheng, Satoru Fukayama, Masataka Goto, "Joint Beat and Downbeat Tracking Based on CRNN Models and a Comparison of Using Different Context Ranges in Convolutional Layers", July 2021

[7] MatthewDavies, E. P., and Böck, S., "Temporal convolutional networks for musical audio beat tracking." In 2019 27th European Signal Processing Conference (EUSIPCO) (pp. 1-5). IEEE, September 2019.

[8] Zhao, J., Xia, G., and Wang, Y., "Beat Transformer: Demixed beat and downbeat tracking with dilated self-attention." arXiv preprint arXiv:2209.07140, 2022.

[9] Hung, Y. N., Wang, J. C., Song, X., Lu, W. T., and Won, M., "Modeling beats and downbeats with a time-frequency transformer." In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 401-405). IEEE, May 2022.

[10] Jia, B., Lv, J., and Liu, D., "Deep learning-based automatic downbeat tracking: a brief review." Multimedia Systems, 25, 617-638, 2019.

[11] Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., and Wang, Y., "Transformer in transformer." Advances in Neural Information Processing Systems, 34, 15908-15919, 2021.

[12] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K., "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805, 2018.

[13] Steinmetz, C. J., and Reiss, J. D., "WaveBeat: End-to-end beat and downbeat tracking in the time domain." arXiv preprint arXiv:2110.01436, 2021.

[14] Chiu, C. Y., Ching, J., Hsiao, W. Y., Chen, Y. H., Su, A. W. Y., and Yang, Y. H., "Source separation-based data augmentation for improved joint beat and downbeat tracking." In 2021 29th European Signal Processing Conference (EUSIPCO) (pp. 391-395). IEEE, August 2021.

[15] Cheng, T., and Goto, M., "U-Beat: A Multi-Scale Beat Tracking Model Based on Wave-U-Net." In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE, June 2023.

[16] Di Giorgi, B., Mauch, M., and Levy, M., "Downbeat tracking with tempo-invariant convolutional neural networks." arXiv preprint arXiv:2102.02282, 2021.

[17] Hung, Y. N., Wang, J. C., Won, M., and Le, D., "Scaling Up Music Information Retrieval Training with Semi-Supervised Learning." arXiv preprint arXiv:2310.01353, 2023.

[18] S. Durand, J. P. Bello, B. David, and G. Richard, "Robust downbeat tracking using an ensemble of convolutional networks," IEEE/ACM Trans. Audio, Speech, and Language Process., vol. 25, no. 1, 2016.