Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/113/2024.18302

Transformer-based Note level Automatic Drum-Set Transcription

Shijie Hao
Basis Bilingual School Shenzhen

matthew.hao2010@hotmail.com

Abstract. Automatic Drum-set Transcription (ADT) aims to convert drum performance audio into corresponding musical notes. Unlike ordinary instruments, drum performances are characterized by higher discreteness, faster tempos, and shorter note durations. To address these challenges, we propose a novel method for achieving precise drum-set music transcription. Our approach employs a Transformer model as the feature extractor and applies the SemiCRF loss function to guide the prediction probabilities of all potential notes. Given the scarcity of drum-set transcription datasets within the community, we have collected and curated a high-quality, detailed-labeled dataset of drum performances spanning various styles and rhythms, totaling over 1000 minutes. Comparative experimental results demonstrate the efficacy of our proposed method.

Keywords: Automatic Drum-set Transcription, Transformer Encoder, Semi-CRF

1. Introduction

Automatic Music Transcription (AMT) is the process of converting sound signals captured from musical performances into corresponding musical notations [1,2]. With the advancements in deep learning [3–5], significant progress has been made in the field of AMT. Most research efforts have concentrated on pitch-based instruments, such as the piano, violin, and guitar [6, 7]. In this paper, we focus on the task of Automatic Drum-set Transcription (ADT).

Compared to transcription of ordinary musical instruments, the transcription system of drum instruments faces the following difficulties. Firstly, discreteness. There is only a difference in timbre between different drums, without continuity in frequency. Secondly, density. In a short period of time, drum performers often exhibit higher frequency playing behaviors. This means that the algorithm needs to predict more events at the same time. Last but not least, Short duration. Compared to other instruments, each note of the drum has a shorter duration. This requires our algorithm to process at higher resolutions.

Existing music transcription systems can be categorized into two types: frame-based [8–10] and note-based [6, 7, 10, 11]. Frame-based methods initially divide the audio sequence into multiple time frames, predict the notes contained within each frame, and subsequently apply post-processing to obtain a complete representation of each note based on these predictions. In contrast, note-based approaches directly predict the pitch and duration of notes within an audio segment.

Due to the presence of numerous short-duration drum sounds in drum kits and the simultaneous activation of multiple drums, frame-based prediction is more challenging for drums compared to other

DOI: 10.54254/2755-2721/113/2024.18302

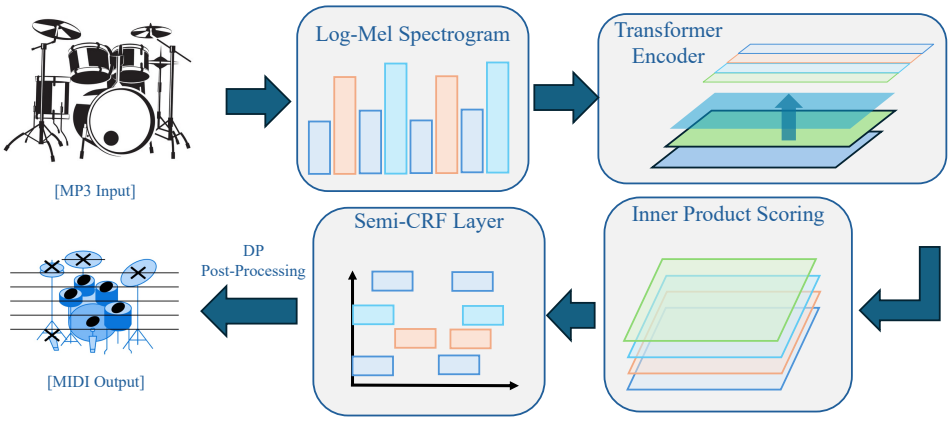

Figure 1. The architecture of our model.

instruments. Therefore, in this paper, we opt for a note-based prediction approach. Specifically, to better capture temporal relationships, we employ an encoder-only Transformer [12] model as the feature extractor, which processes each segmented audio clip to extract features. Subsequently, these extracted features are fed into a SemiCRF [13] layer to score every possible combination of audio events within that time segment. Finally, we apply the Viterbi algorithm to post-process the scores across different segments, thereby obtaining the complete annotation of the musical piece.

In addition, due to the relatively limited research on automatic recognition of drums, it is difficult to find a suitable dataset for drum-set transcription. For this purpose, we collected and curated over 1000 minutes of high-quality drum performance data for the study of this task.

The main contribution of this paper can be summarized in the following three aspects.

- We have developed a note-level drum-set transcription algorithm based on transformer context encoder and supervised by SemiCRF loss, which can efficiently transcribe notes from drum music records.
- We collected and curated a new dataset for automatic drum-set transcription, which contains over 1000 minutes high-quality pure drum-set music records and corresponding note annotation.
- We conduct detailed experiments to validate the effectiveness and superiority of our method. The ablative experiments also demonstrated the role of different modules.

2. Methodology

2.1. Problem Definition

Assuming the sampling rate of the audio is F, then T seconds of audio can be represented as a one-dimensional array of length TF. A complete drum performance audio typically ranges from several minutes to over ten minutes, and we usually divide it into multiple segments, each a few seconds long. For each segment $X \in R^n$, our task is to predict the sequence of start and end times $Y \in \{(s_i, e_i)|i = 0, 1, 2, ...\}$ for each event of every possible event type within that segment.

2.2. Data Preprocessing

In order to process the audio data effectively, we apply log-mel spectrogram preprocessing. The log-mel spectrogram is a widely used technique in audio signal processing that converts the raw audio waveform into a time-frequency representation. This transformation involves first computing the short-time Fourier transform (STFT) of the audio signal to obtain its frequency spectrum over short intervals. The resulting frequency bins are then mapped onto the mel scale, which is designed to mimic the human auditory

DOI: 10.54254/2755-2721/113/2024.18302

system's perception of sound. After mapping, the logarithm of the power values is taken to compress the dynamic range of the spectrogram, making it more suitable for machine learning models. This preprocessing step not only enhances the features relevant to human perception but also reduces the computational complexity of subsequent processing tasks.

2.3. Transformer Feature Encoder

To extract meaningful features from audio data, we employ the Transformer [12] architecture, a powerful deep learning model originally proposed for natural language processing tasks. The Transformer model is based on the self-attention mechanism, which allows it to capture long-range dependencies and context within sequences efficiently. Unlike traditional recurrent neural networks (RNNs) that process data sequentially, Transformers parallelize the computation across all elements in the input sequence, significantly improving training speed and scalability. The self-attention mechanism allows the model to weigh the importance of different parts of the input sequence, enabling it to focus on relevant features and ignore noise. The feed-forward sub-layer applies a fully connected neural network to each position independently, further refining the extracted features.

2.4. Semi-CRF Layer

To predict events for each type of drum in a drum kit performance, we employ SemiCRF [13] (Conditional Random Fields with Segmental Structure), an advanced sequence modeling technique. SemiCRF is particularly well-suited for tasks where events or segments have internal dependencies and complex structures, making it an ideal choice for drum event detection.

SemiCRF is a probabilistic graphical model that defines a conditional probability distribution $P(\mathbf{Y} \mid \mathbf{X})$, where $\mathbf{X} = \{x_1, x_2, \ldots, x_T\}$ represents the input sequence (here we use the features extracted by the Transformer encoder) and $\mathbf{Y} = \{y_1, y_2, \ldots, y_S\}$ represents the segment labels. Each segment y_i corresponds to a contiguous subsequence of input frames that share a common label, such as a specific drum hit. A segment s_i is defined as a contiguous subsequence of input frames $\{x_j, x_{j+1}, \ldots, x_{j+k}\}$ that share a common label y_i. For drum event detection, a segment could represent a single drum hit or a combination of hits.

The conditional probability of the segment labels given the input sequence is defined as:

$$P(\mathbf{Y} \mid \mathbf{X}) = \frac{1}{Z(\mathbf{X})} \exp\left(\sum_{i=1}^{S}\sum_{k=1}^{K} \lambda_k f_k(\mathbf{X}, \mathbf{Y}, i)\right) \tag{1}$$

where $Z(\mathbf{X})$ is the normalization factor, ensuring that the probabilities sum to one; λ_k are the weights associated with each feature function f_k.

The model parameters λ_k are learned by maximizing the log-likelihood of the training data:

$$\mathcal{L}(\lambda) = \sum_{n=1}^{N} \log P(\mathbf{Y}^{(n)} \mid \mathbf{X}^{(n)}) \tag{2}$$

where N is the number of training examples.

During inference, the most likely sequence of segment labels $\mathbf{Y}^*$ is found using the Viterbi algorithm:

$$\mathbf{Y}^* = \arg\max_{\mathbf{Y}} P(\mathbf{Y} \mid \mathbf{X}) \tag{3}$$

3. Experiments

3.1. Dataset

Due to the limited research on drum transcription tasks, relevant datasets are scarce and difficult to obtain. To address this issue, we have collected and curated a new dataset specifically for drum transcription. During the collection process, we excluded data containing other instruments, retaining only pure drum

DOI: 10.54254/2755-2721/113/2024.18302

performances. The dataset comprises 106 drum performance tracks along with their corresponding MIDI annotations, totaling over 1000 minutes of audio. We randomly selected 90 tracks for the training set and reserved the remaining 16 tracks for the validation set.

3.2. Comparasion with other methods

Table 1. Table 1: Comparison with other methods

Method	**mAP (%)**
Frame-based method [8]	48.5
CNN-based method [6]	67.2
Ours	76.3

Table 2. Table 2: Accuracy of Different Drum Kit Components

Component	**mAP (%)**
Bass Drum	95.3
Snare Drum	93.5
Hi-Hat Cymbals	92.1
Ride Cymbal	90.7
Tom Toms	88.5
Floor Tom	86.3
Crash Cymbals	85.4
Cymbal Stands	80.0
Bass Drum Pedal	65.9
Hi-Hat Stand	60.3
Drum Key	55.5

We evaluated various methods on the aforementioned dataset, and the results are presented in Table 1. It is evident that our method, which leverages the temporal modeling capabilities of the Transformer and the event understanding capabilities of SemiCRF, significantly outperforms methods that use CNN for feature extraction and frame-based prediction methods.

3.3. Different Instruments Analysis

According to the contents of Table 2, it is evident that our method performs better on commonly used drums with longer durations, while it exhibits poorer performance on less frequently used drums or those with very high strike frequencies. This can primarily be attributed to two reasons. First, the limited size of our dataset results in some less common drums being underrepresented and thus not adequately recognized. Second, drums with very high strike frequencies are inherently more challenging to predict.

4. Conclusion

In this paper, we propose a drum transcription model based on Transformer and SemiCRF. The model first extracts audio features using a Transformer and then performs note-level predictions directly using SemiCRF. Experimental results demonstrate that our method achieves promising performance in drum recognition tasks. Additionally, we introduce a new drum transcription dataset to facilitate future research in this area.

DOI: 10.54254/2755-2721/113/2024.18302

References

[1] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: An overview," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, 2018.

[2] B. Bhattarai and J. Lee, "A comprehensive review on music transcription," *Applied Sciences*, vol. 13, no. 21, p. 11882, 2023.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[6] Y. Yan, F. Cwitkowitz, and Z. Duan, "Skipping the frame-level: Event-based piano transcription with neural semi-crfs," *Advances in Neural Information Processing Systems*, vol. 34, pp. 20 583–20 595, 2021.

[7] Y. Yan and Z. Duan, "Scoring intervals using non-hierarchical transformer for automatic piano transcription," *arXiv preprint arXiv:2404.09466*, 2024.

[8] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. Engel, S. Oore, and D. Eck, "Onsets and frames: Dual-objective piano transcription," *arXiv preprint arXiv:1710.11153*, 2017.

[9] Q. Kong, B. Li, X. Song, Y. Wan, and Y. Wang, "High-resolution piano transcription with pedals by regressing onset and offset times," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3707–3717, 2021.

[10] T. Kwon, D. Jeong, and J. Nam, "Polyphonic piano transcription using autoregressive multi-state note model," *arXiv preprint arXiv:2010.01104*, 2020.

[11] R. Kelz, S. Böck, and G. Widmer, "Deep polyphonic adsr piano note transcription," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 246–250.

[12] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[13] S. Sarawagi and W. W. Cohen, "Semi-markov conditional random fields for information extraction," *Advances in neural information processing systems*, vol. 17, 2004.