Automatic Music Transcription

Дмитрий Протасов

Научный руководитель: Иван Матвеев
МФТИ

16 декабря 2023

Постановка задачи

Проблема

Генеративные музыкальные модели довольно удобно строить в пространстве MIDI-файлов. Проблема – нет большого количества таких MIDI-датасетов, для большинства песен в интернете есть только их аудиоформат. Эту проблему предлагается решать алгоритмом преобразования аудио-представления песен в её MIDI-представление.

Цель

Исследование и улучшение существующих алгоритмов извлечения MIDI из песен

Задачи работы

- Собрать базу данных песен и их MIDI-представлений, сгенерировать свои датасеты
- Изучить и протестировать существующие модели, понять их главные недостатки
- Реализовать свои методы извлечения MIDI из аудио

Постановка задачи

Сама задача распознавания нот делится на три этапа

- Разделение на отдельные инструментальные дорожки (Music-Source-Separation)
- Распознавание инструмента (Instrument-Recognition)
- Транскрибация в ноты (Note-Transcription)

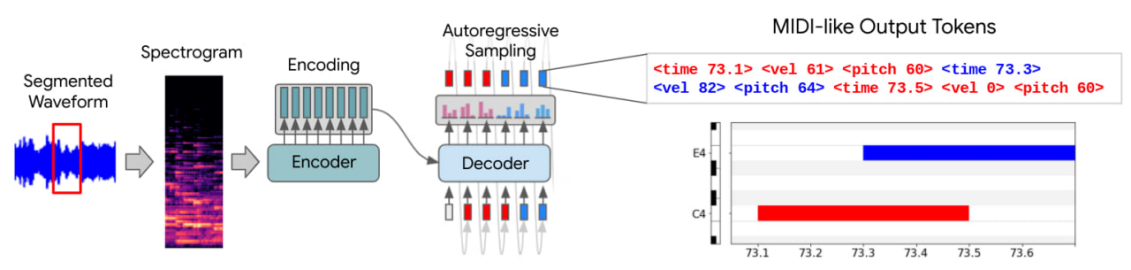
Рассмотрим основные работы, посвященные одному или нескольким из этих этапов

Figure 1. The proposed Jointist framework. Our actual framework can transcribe/separate up to 39 different instruments as defined in Table 7 of Appendix. B: batch size, L: audio length, C: instrument classes, T: number of time steps, K: number of predicted instruments. Dotted lines represent iterative operations for K times. Best viewed in color.

Обзор литературы: Transcription

MT3 (2022) [link]

SOTA в Multi-instrument, основана на модели T5



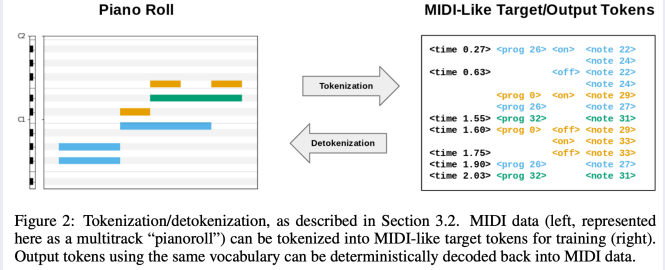


Figure 2: Tokenization/detokenization, as described in Section 3.2. MIDI data (left, represented here as a multitrack "pianoroll") can be tokenized into MIDI-like target tokens for training (right). Output tokens using the same vocabulary can be deterministically decoded back into MIDI data.

Jointist (2022) [link]

Instrument Recognition: CNN + Transformer
Note Transcription: опирается на onsets-and-frames

Crepe (2018) [link]

Находит фундаментальную частоту по аудио. Может быть полезно для извлечения нот из вокала

Обзор литературы: Music Source Separation

Benchmarks and leaderboards for sound demixing tasks [link]

Demucs [github]

Based on a U-Net convolutional architecture

MDX-Net [github]

Two-stream neural network for music demixing
MDX-Net consists of six networks, all trained separately

Band-split RNN [github]

В основе две RNN-ки по оси частот и по оси времени

MUSDB18 Dataset [link]

150 music tracks (10h duration) with isolated drums, bass, vocals, others

Обзор литературы: другие работы

Matrosov (2015) [link]

Генеративная модель на MIDI в пространстве 4095 аккордов

Eronen (2001) [link]

Рассмотрено много методов различных методов без инструментов глубокого обучения, в основном методе используются cepstral coefficients

Encodec (2022) [link]

В основе лежит VQ-VAE – интересно понять есть ли связь MIDI-пространства с латентным пространством, выуичвыаемым в этой модели

Проведенные на данный момент эксперименты

Ссылка на github с черновыми скриптами и запусками готовых моделей

- генерация синтетического датасета
- audio to specetrogram
- separate with demucs

TODO: Вставить картинки результатов из demucs Вставить картинки которые вычленяли ноты с помощью CREPE, и других алгоритмов (моих, самописных)

Literature

1. **MT3** MT3: MULTI-TASK MULTITRACK MUSIC TRANSCRIPTION
2. **Jointist** JOINTIST: JOINT LEARNING FOR MULTI-INSTRUMENT TRANSCRIPTION AND ITS APPLICATIONS
3. **Demucs** HYBRID TRANSFORMERS FOR MUSIC SOURCE SEPARATION
4. **CREPE** A CONVOLUTIONAL REPRESENTATION FOR PITCH ESTIMATION
5.

Future Work

1. Сделать эксперименты по замеру качества существующих методов
2. Собрать базу данных песен и их MIDI-представлений, сгенерировать свои датасеты
3. Реализовать и протестировать несколько своих методов извлечения MIDI из аудио