Omnizart: A General Toolbox for Automatic Music Transcription

Yu-Te Wu[1], **Yin-Jyun Luo**[1], **Tsung-Ping Chen**[1], **I-Chieh Wei**[1], **Jui-Yang Hsu**[1], **Yi-Chin Chuang**[1], **and Li Su**[1]

1 Music and Culture Technology Lab, Institute of Information Science, Academia Sinica, Taipei, Taiwan

DOI: 10.21105/joss.03391

Software

- Review
- Repository
- Archive

Editor: Fabian-Robert Stöter

Reviewers:

- @hagenw
- @keunwoochoi

Submitted: 24 April 2021
Published: 21 June 2021

Summary

We present and release Omnizart, a new Python library that provides a streamlined solution to automatic music transcription (AMT). Omnizart encompasses modules that construct the life-cycle of deep learning-based AMT, and is designed for ease of use with a compact command-line interface. To the best of our knowledge, Omnizart is the first transcription toolkit which offers models covering a wide class of instruments ranging from solo, instrument ensembles, percussion instruments to vocal, as well as models for chord recognition and beat/downbeat tracking, two music information retrieval (MIR) tasks highly related to AMT. In summary, Omnizart incorporates:

- Pre-trained models for frame-level and note-level transcription of multiple pitched instruments, vocal melody, and drum events;
- Pre-trained models of chord recognition and beat/downbeat tracking;
- Main functionalities in the life-cycle of AMT research, covering from dataset downloading, feature pre-processing, model training, to sonification of the transcription result.

Omnizart is based on Tensorflow (Abadi et al., 2016). The complete code base, command-line interface, documentation, as well as demo examples can all be accessed from the project website.

Statement of need

AMT has been one of the core challenges in MIR because of the multifaceted nature of musical signals. Typically, streams of musical notes performed with various instruments overlap with each other and then create a hierarchy of abstraction. This complicates the task to identify the melodic, timbral, and rhythmic attributes of the music.

While the majority of the previous solution focuses on single-instrument transcription, Omnizart collects several state-of-tha-art (SoTA) models for transcribing multiple pitched and percussive instruments, as well as vocal out of the interference with rich music polyphony. Omnizart also finds it applicability for chord recognition and beat tracking. As such, the proposed library offers a unified solution to music transcription for multi-track and modalities.

In short conclusion, Omnizart represents an AMT tool which unifies multiple transcription utilities and enables further productivity. Omnizart can save one's time and labor in generating massive amount of multi-track MIDI files, which could have a great impact on music production, music generation, education, and musicology research.

Implementation Details

Piano solo transcription

The piano solo transcription model in Omnizart reproduces the implementation of (Wu et al., 2020). The model features a U-net which takes as inputs the audio spectrogram, generalized cepstrum (GC) (Li Su & Yang, 2015), and GC of spectrogram (GCoS) (Wu et al., 2018), and outputs a multi-channel time-pitch representation with time- and pitch-resolution of 20ms and 25 cents, respectively. For the U-net, implementation of the encoder and the decoder follows DeepLabV3+ (L.-C. Chen et al., 2018), and the bottleneck layer is adapted from the Image Transformer (Parmar et al., 2018).

The model is trained on the MAESTRO dataset (Hawthorne et al., 2019), an external dataset containing 1,184 real piano performance recordings with a total length of 172.3 hours. The model achieves 72.50% and 79.57% for frame- and note-level F1-scores, respectively, on the Configuration-II test set of the MAPS dataset (Kelz et al., 2016).

Multi-instrument polyphonic transcription

The multi-instrument transcription model extends the piano solo model to support 11 output classes, namely piano, violin, viola, cello, flute, horn, bassoon, clarinet, harpsichord, contrabass, and oboe, accessed from MusicNet (Thickstun et al., 2017). Notably, the model allows for *instrument-agnostic transcription* where the instruments to transcribe are unknown during inference (Wu et al., 2020). The evaluation on the test set from MusicNet (Thickstun et al., 2018) yields 66.59% for the note streaming task.

Drum transcription

The model for drum transcription is a re-implementation of (Wei et al., 2021) which predicts percussive events from a given input audio. Building blocks of the network include convolutional layers and the attention mechanism.

The model is trained on a dataset with 1,454 audio clips of polyphonic music with synchronized drum events (Wei et al., 2021). The model demonstrates SoTA performance on two commonly used benchmark datasets, i.e., 74% for ENST (Gillet & Richard, 2006) and 71% for MDB-Drums (Southall et al., 2017) in terms of the note-level F1-score.

Vocal transcription in polyphonic music

The system for vocal transcription features a pitch extractor and a module for note segmentation. The inputs to the model are composed of spectrogram, GS, and GCoS derived from polyphonic music recordings (Wu et al., 2018).

A pre-trained Patch-CNN (L. Su, 2018) is leveraged as the pitch extractor. The module for note segmentation is implemented with PyramidNet-110 and ShakeDrop regularization (Yamada et al., 2019), which is trained using Virtual Adversarial Training (Miyato et al., 2018) enabling semi-supervised learning.

The training includes labeled data from TONAS (Mora et al., 2010) and unlabeled ones from MIR-1K (Hsu & Jang, 2009). The model yields the SoTA F1-score of 68.4% evaluated with the ISMIR2014 dataset (Molina et al., 2014).

Chord recognition

The harmony recognition function of Omnizart is implemented using the Harmony Transformer (HT) (T.-P. Chen & Su, 2019). The HT model is based on an encoder-decoder architecture, where the encoder performs chord segmentation on the input, and the decoder recognizes the chord progression based on the segmentation result.

The original HT supports both audio and symbolic inputs. Currently, Omnizart supports only audio inputs. A given audio input is pre-processed using Chordino VAMP plugin (Mauch & Dixon, 2010) as the non-negative-least-squares chromagram. The outputs of the model include 25 chord types, covering 12 major and minor chords together with a class referred to the absence of chord, with a time resolution of 230ms.

In an experiment with evaluations on the McGill Billboard dataset (Burgoyne et al., 2011), the HT outperforms the previous SoTAs (T.-P. Chen & Su, 2019).

Beat/downbeat tracking

The model for beat and downbeat tracking provided in Omnizart is a reproduction of (Chuang & Su, 2020). Unlike most of the available open-source projects such as `madmom` (Böck et al., 2016) and `librosa` (McFee et al., 2015) which focus on audio, the provided model targets symbolic data.

The input and output of the model are respectively MIDI and beat/downbeat positions with the time resolution of 10ms. The input representation combines piano-roll, spectral flux, and inter-onset interval extracted from MIDI. The model composes a two-layer BLSTM network with the attention mechanism, and predict probabilities of the presence of beat and downbeat per time step.

Experiments on the MusicNet dataset (Thickstun et al., 2018) with the synchronized beat annotation show that the proposed model outperforms the SoTA beat trackers which operate on synthesized audio (Chuang & Su, 2020).

Conclusion

Omnizart represents the first systematic solution for the polyphonic AMT of general music contents ranging from pitched instruments, percussion instrument, to voice. In addition to note transcription, Omnizart also includes high-level MIR tasks such as chord recognition and beat/downbeat tracking. As an ongoing project, the research group will keep refining the package and also extending the scope of transcription in the future.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., & others. (2016). Tensorflow: A system for large-scale machine learning. *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 265–283.

Böck, S., Korzeniowski, F., Schlüter, J., Krebs, F., & Widmer, G. (2016). Madmom: A new python audio and music signal processing library. *Proceedings of the 24th ACM International Conference on Multimedia*, 1174–1178.

The Journal of Open Source Software

Burgoyne, J. A., Wild, J., & Fujinaga, I. (2011). An expert ground truth set for audio chord recognition and music analysis. *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, 633–638.

Chen, L.-C., Zhu, Y., George, P., Florian, S., & Hartwig, A. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv Preprint arXiv:1802.02611*. https://doi.org/10.1007/978-3-030-01234-2_49

Chen, T.-P., & Su, L. (2019). Harmony transformer: Incorporating chord segmentation into harmony recognition. *Proc. ISMIR*.

Chuang, Y.-C., & Su, L. (2020). Beat and downbeat tracking of symbolic music data using deep recurrent neural networks. *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 346–352.

Gillet, O., & Richard, G. (2006). ENST-drums: An extensive audio-visual database for drum signals processing. *ISMIR*, 156–159.

Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C.-Z. A., Dieleman, S., Elsen, E., Engel, J. H., & Eck, D. (2019). Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset. *7th International Conference on Learning Representations (ICLR)*.

Hsu, C.-L., & Jang, J.-S. R. (2009). On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(2), 310–319. https://doi.org/10.1109/tasl.2009.2026503

Kelz, R., Dorfer, M., Korzeniowski, F., Böck, S., Arzt, A., & Widmer, G. (2016). On the Potential of Simple Framewise Approaches to Piano Transcription. *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, 475–481.

Mauch, M., & Dixon, S. (2010). Approximate note transcription for the improved identification of difficult chords. *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, 135–140.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). Librosa: Audio and music signal analysis in python. *Proceedings of the 14th Python in Science Conference*, *8*, 18–25. https://doi.org/10.25080/majora-7b98e3ed-003

Miyato, T., Maeda, S., Koyama, M., & Ishii, S. (2018). Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE TPAMI*, *41*(8), 1979–1993. https://doi.org/10.1109/tpami.2018.2858821

Molina, E., Barbancho-Perez, A. M., Tardón, L. J., Barbancho-Perez, I., & others. (2014). Evaluation framework for automatic singing transcription. *Proc. ISMIR*.

Mora, J., Gómez, F., Gómez, E., Escobar-Borrego, F., & Díaz-Báñez, J. M. (2010). Characterization and melodic similarity of a cappella flamenco cantes. *Proceedings of ISMIR*, 9–13.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., & Tran, D. (2018). Image Transformer. *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 4052–4061.

Southall, C., Wu, C.-W., Lerch, A., & Hockman, J. (2017). *MDB drums: An annotated subset of MedleyDB for automatic drum transcription*.

Su, L. (2018). Vocal melody extraction using patch-based CNN. *Proc. ICASSP*, 371–375. https://doi.org/10.1109/icassp.2018.8462420

Su, Li, & Yang, Y.-H. (2015). Combining Spectral and Temporal Representations for Multipitch Estimation of Polyphonic Music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(10), 1600–1612. https://doi.org/10.1109/taslp.2015.2442411

Thickstun, J., Harchaoui, Z., Foster, D. P., & Kakade, S. M. (2018). Invariances and Data Augmentation for Supervised Music Transcription. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2241–2245. https://doi.org/10.1109/icassp.2018.8461686

Thickstun, J., Harchaoui, Z., & Kakade, S. M. (2017). Learning features of music from scratch. *International Conference on Learning Representations (ICLR)*.

Wei, I.-C., Wu, C.-W., & Su, L. (2021). Improving automatic drum transcription using large-scale audio-to-midi aligned data. *Proc. ICASSP*. https://doi.org/10.1109/icassp39728.2021.9414409

Wu, Y.-T., Chen, B., & Su, L. (2018). Automatic Music Transcription Leveraging Generalized Cepstral Features and Deep Learning. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 401–405. https://doi.org/10.1109/icassp.2018.8462079

Wu, Y.-T., Chen, B., & Su, L. (2020). Multi-instrument automatic music transcription with self-attention-based instance segmentation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 2796–2809. https://doi.org/10.1109/taslp.2020.3030482

Yamada, Y., Iwamura, M., Akiba, T., & Kise, K. (2019). Shakedrop regularization for deep residual learning. *IEEE Access*, *7*, 186126–186136. https://doi.org/10.1109/access.2019.2960566