IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 6, March 2024

An In-depth Review on Music Source Separation

Prof. K. G. Jagtap[1], Vivek Deshmukh[2], Maviya Mahagami[3], Himanshu Lohokane[4], Rashi Kacchwah[5]

Professor, Department of AI & ML[1]

Students, Department of AI & ML[2,3,4,5]

AISSM Polytechnic, Pune, India

Abstract*: The study in Music Source Separation (MSS) raises a fundamental question: Is there any benefit in considering broader contextual information, or are local acoustic features adequate? In various domains, attention-based Transformers [1] have demonstrated their capacity to assimilate information across extensive sequences. In our research, we introduce Hybrid Transformer Demucs (HT Demucs), a hybrid temporal/spectral bi-U-Net based on Hybrid Demucs [2]. Here, the innermost layers are substituted with a cross-domain Transformer Encoder, utilizing self-attention within one domain and cross-attention across domains. Although its performance is lacking when exclusively trained on MUSDB [3], we illustrate that it surpasses Hybrid Demucs (trained on the same data) by 0.45 dB of Signal-to-Distortion Ratio (SDR) when provided with an additional 800 training songs. By employing sparse attention kernels to broaden its receptive field and undertaking per-source fine-tuning, we attain state-of-the-art results on MUSDB with extra training data, achieving a remarkable 9.20 dB of SDR*.

Keywords: Music Source Separation, Transformers.

REFERENCES

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). "Attention is all you need." CoRR, vol. abs/1706.03762.

[2] De´fossez, A. (2021). "Hybrid spectrogram and waveform source separation." In Proceedings of the ISMIR 2021 Workshop on Music Source Separation.

[3] Rafii, Z., Liutkus, A., Sto¨ter, F. R., Mimilakis, S. I., & Bittner, R. (2017). "The musdb18 corpus for music separation."

[4] Ono, N., Rafii, Z., Kitamura, D., Ito, N., & Liutkus, A. (2015). "The 2015 Signal Separation Evaluation Campaign." In International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA).

[5] Zafar Rafii, Antoine Liutkus, Fabian-Robert Sto¨ter, Stylianos Ioannis Mimilakis, and Rachel Bittner, "Musdb18-hq - an uncompressed version of musdb18," Aug. 2019.

[6] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Herve´ Je´gou, "Going deeper with image transformers," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjo¨rn Ommer, "High-resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022, pp. 10684–10695.

[8] Tom B. Brown et al., "Language models are few-shot learners," 2020.

[9] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019.

[10] Alexandre De´fossez, Nicolas Usunier, Le´on Bottou, and Francis Bach, "Music source separation in the waveform domain," 2019.

[11] F.-R. Sto¨ter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-unmix - a reference implementation for music source separation," Journal of Open Source Software, 2019.

[12] Takahashi, N., & Mitsufuji, Y. (2020). "D3net: Densely connected multidilated densenet for music source separation."

DOI: 10.48175/IJARSCT-16841

[13] Choi, W., Kim, M., Chung, J., & Jung, S. (2021). "Lasaft: Latent source attentive frequency transformation for conditioned source separation." In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[14] Luo, Y., & Yu, J. (2022). "Music source separation with band-split RNN."

[15] Yi Luo, Zhuo Chen, and Takuya Yoshioka, “Dualpath rnn: efficient long sequence modeling for timedomain single-channel speech separation,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 46–50.

[16] Daniel Stoller, Sebastian Ewert, and Simon Dixon, “Wave-u-net: A multi-scale neural network for end- toend audio source separation,” arXiv preprint arXiv:1806.03185, 2018.

[17] Minseok Kim, Woosung Choi, Jaehwa Chung, Daewon Lee, and Soonyoung Jung, “Kuielab-mdx-net: A twostream neural network for music demixing,” 2021.

[18] Yuki Mitsufuji, Giorgio Fabbro, Stefan Uhlich, FabianRobert Sto¨ter, Alexandre De´fossez, Minseok Kim, Woosung Choi, Chin-Yun Yu, and Kin-Wai Cheuk, “Music demixing challenge 2021,” Frontiers in Signal Processing, vol. 1, jan 2022.

[19] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam, “Spleeter: a fast and efficient music source separation tool with pre-trained models,” Journal of Open Source Software, 2020.

[20] Zelun Wang and Jyh-Charn Liu, “Translating math formula images to latex sequences using deep neural networks with sequence-level training,” 2019.

[21] Diederik P. Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” 2014.

[22] Fabian-Robert Sto¨ter, Antoine Liutkus, and Nobutaka Ito, “The 2018 signal separation evaluation campaign,” 2018

DOI: 10.48175/IJARSCT-16841