Hybrid Transformers for Music Source Separation

Rashi Kacchwah[1] Himanshu Lohokane[2] Maviya Mahagami[3] Vivek Deshmukh[4]
[1,2,3,4]Student
[1,2,3,4]Department of Artificial Intelligence and Machine Learning
[1,2,3,4]All India Shri Shivaji Memorial Society Polytechnic Pune, Maharashtra, 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

I. Introduction

(SiSEC) in 2015 [4], the Music Source Separation (MSS) community has primarily focused on training supervised models for the task of separating songs into four stems: drums, bass, vocals, and other (representing all other instruments). The benchmark dataset used in MSS is MUSDB18 [3, 5], consisting of 150 songs in two versions (HQ and non-HQ). The training set comprises 87 songs, a relatively small corpus compared to other deep learning tasks like vision [6, 7] or natural language processing [8], where Transformer [1]-based architectures have found success. In source separation, both short and long context inputs are relevant. Conv-Tasnet [9] utilizes about one second of context, relying solely on local acoustic features for separation. On the other hand, Demucs [10] can use up to 10 seconds of context to address input ambiguities. In this study, our goal is to explore how Transformer architectures can effectively utilize this context and determine the amount of data needed to train them. In Section 3, we introduce a novel architecture called Hybrid Transformer Demucs (HT Demucs), which replaces the innermost layers of the original Hybrid Demucs architecture [2] with Transformer layers. These layers are applied in both the time and spectral representation, using self-attention within one domain and cross-attention across domains. Transformers typically require substantial data, so we augment the MUSDB dataset with an internal dataset of 800 songs, detailed in Section 4. Our second contribution, presented in Section 5, involves an extensive evaluation of this new architecture under various settings (depth, number of channels, context length, augmentations, etc.). We demonstrate a notable improvement over the baseline Hybrid Demucs architecture (retrained on the same data) by 0.35 dB. Finally, we experiment with increasing the context duration using sparse kernels based on Locally Sensitive Hashing to overcome memory issues during training and fine-tuning procedures. This results in a final Signal-toDistortion Ratio (SDR) of 9.20 dB on the MUSDB test set

II. Literature survey

Demucs, a deep extractor for music sources, is a prominent framework in audio signal processing that has been used to isolate individual sound sources from complex mixtures. However, as the field advances and challenges persist, researchers are increasingly exploring hybrid approaches, integrating Demucs with complementary methodologies to unlock new potentials and address existing limitations. This literature survey explores the evolution, principles, and implications of hybrid Demucs architectures for audio source separation, examining their combinations with various deep learning models, signal processing techniques, and data-driven methodologies. The survey also delves into the intricate interplay between hybrid architectures, evaluation methodologies, and benchmark datasets, providing insight into the complex landscape of performance assessment in audio source separation. By synthesizing insights from various scholarly contributions, the survey illuminates the state-of-the-art in hybrid Demucs, identifies emerging trends, unresolved challenges, and future research directions, serving as a comprehensive resource for researchers, practitioners, and enthusiasts in the ever-evolving field of audio source separation.

III. Proposed system

In the realm of Music Source Separation (MSS) methods, there's a traditional categorization between spectrogram-based and waveform-based models. Spectrogram-based models include approaches like Open-Unmix [11], employing a biLSTM with fully connected components to predict a mask on the input spectrogram. Another example is D3Net [12], which utilizes dilated convolutional blocks with dense connections. More recently, there has been a preference for using complex spectrogram as both input and output [13], offering a more comprehensive representation and eliminating the top-line constraint imposed by the Ideal-Ratio-Mask. The Band-Split RNN [14], the latest spectrogram model, incorporates this concept along with multiple dual-path RNNs [15], each operating in carefully designed frequency bands. Currently, it holds the state-of-the-art performance on MUSDB with 8.9 dB. On the other hand, waveform-based models originated with Wave-U-Net [16], forming the foundation for Demucs [10], a time-domain U-Net with a bi-LSTM positioned between the encoder and decoder. Around the same period, Conv-TasNet demonstrated competitive results [9, 10] by using residual

dilated convolution blocks to predict a mask over a learned representation. A recent trend involves combining both temporal and spectral domains, either through model blending, exemplified by KUIELAB-MDX-Net [17], or by adopting a biU-Net structure with a shared backbone, as seen in Hybrid Demucs [2]. Despite Hybrid Demucs being the top-ranked architecture in the latest MDX MSS Competition [18], it has now been surpassed by Band-Split RNN.

Music Source Separation (MSS) benefits greatly from large datasets, as demonstrated by Spleeter [19], a U-Net architecture trained on 25,000 30-second song extracts. Spleeter set a benchmark upon its release and showcased the advantages of leveraging extensive data. Further improvements were achieved with models like D3Net and Demucs, which showed enhanced performance on datasets like MUSDB when provided with additional training data. Band-Split RNN introduced an innovative unsupervised augmentation technique, boosting performance by 0.7 dB of Signal-to-Distortion Ratio (SDR) using only mixes.

Transformers, known for their effectiveness in various tasks, have also made strides in speech source separation with models like SepFormer [20]. However, SepFormer's demanding memory requirements render it unsuitable for longer inputs at higher frequencies. In response to this challenge, the Hybrid Transformer Demucs model was introduced, building upon the architecture of Hybrid Demucs [2]. The new model preserves key aspects of Hybrid Demucs while incorporating cross-domain Transformer Encoder layers, enhancing flexibility and adaptability.

Hybrid Transformer Demucs maintains the basic structure of Hybrid Demucs, featuring two U-Nets operating in the time and spectrogram domains, respectively. However, it replaces the innermost layers in both encoder and decoder with cross-domain Transformer Encoder layers. These layers enable concurrent processing of 2D spectral data and 1D waveform data, improving the model's overall performance.

The architecture of Hybrid Transformer Demucs includes a depiction of a single self-attention Encoder layer of the Transformer, featuring normalizations before the Self-Attention and FeedForward operations. The cross-attention Encoder layer follows a similar structure but involves cross-attention with the representation from the other domain. The entire architecture incorporates a double U-Net encoder/decoder structure, maintaining the essence of the original Hybrid Demucs while incorporating Transformer elements.

To address memory consumption and attention speed issues as sequence length increases, the model employs sparse attention kernels introduced in the xformer package. Additionally, a Locally Sensitive Hashing (LSH) scheme dynamically determines the sparsity pattern. By conducting multiple rounds of LSH and selecting elements that match across these rounds, the model achieves a desired sparsity level, enhancing efficiency without sacrificing performance. This modified version is referred to as Sparse HT Demucs.

In summary, the Hybrid Transformer Demucs model represents a significant advancement in Music Source Separation, combining the strengths of Hybrid Demucs with Transformer architecture. By introducing cross-domain Transformer Encoder layers and implementing sparsity techniques, the model achieves enhanced flexibility, adaptability, and efficiency, addressing key challenges in MSS tasks.

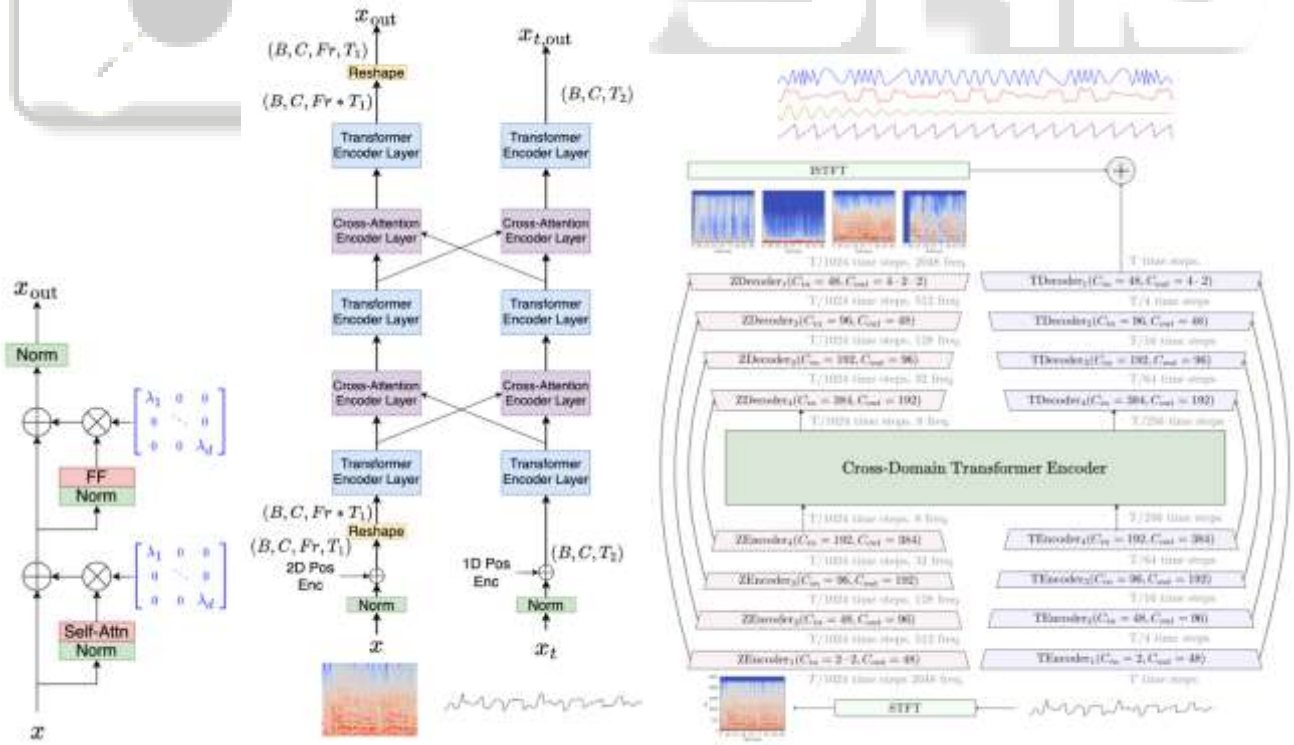

Fig. 1: Details of the Hybrid Transformer Demucs architecture. (a): the Transformer Encoder layer with self-attention and Layer Scale [6]. (b): The Cross-domain Transformer Encoder treats spectral and temporal signals with interleaved Transformer Encoder layers and cross-attention Encoder layers. (c): Hybrid Transformer Demucs keeps the outermost 4 encoder and decoder layers of Hybrid Demucs with the addition of a cross-domain Transformer Encoder between them.

A. Dataset:

The dataset comprises 3,500 songs with stems from 200 artists spanning various genres, each stem labeled with one of four sources, though subjectivity and naming ambiguity may introduce noise. Manual validation of 150 tracks was conducted. An initial Hybrid Demucs model was trained on a subset of 150 tracks from the dataset and MUSDB. Preprocessing retained stems with all four sources active for at least 30% of the time and identified silent segments. For each song, denoted 'xi' for drums, bass, other, and vocals, and model 'f,' 'yi,j = f(xi)j,' where 'j' indicates the output when separating stem 'i.' In an ideal scenario, 'yi,j' equals 'xiδi,j.' Volume in dB over 1-second segments is defined as V(z) = 10.(AveragePool(, 1sec)). For each source pair, 'Pi,j' measures the proportion of segments where (V(yi,j) - V(xi) > -10 dB), yielding a square matrix 'P' in [0, 1] with dimensions 4×4. Retention criteria include 'Pi,i > 70%' for all sources and 'Pi,j < 30%' for pairs. This process selected 800 songs. The approach integrates diverse datasets and a structured split to improve the model's ability to distinguish between authentic and manipulated content across sources.[1]

B. Preprocessing:

The dataset preprocessing stage involves several key steps to enhance the efficiency of subsequent model training. Initially, each video is split into frames, followed by face detection to identify and isolate facial regions within each frame. Subsequently, to maintain consistency in the number of frames across the dataset, the mean frame count is computed. A new processed dataset is then generated, containing frames equal to this computed mean. During this process, frames lacking detected faces are excluded to ensure the dataset's focus on facial features.

Recognizing the computational demands associated with processing an entire 10-second video, especially at a frame rate of 30 frames per second (resulting in a total of 300 frames), we propose a pragmatic approach for experimental purposes. Specifically, for training the model, we suggest utilizing only the initial 100 frames from each video. This reduction in frame count aims to alleviate computational requirements during the experimental phase while still providing sufficient data for training and validating the deep fake detection model effectively. This approach allows for a more manageable and resource-efficient exploration of the model's performance within the experimental constraints.

C. Model:

The Hybrid Transformer Demucs model builds upon the architecture of the original Hybrid Demucs model by incorporating cross-domain Transformer Encoders, enhancing its flexibility and performance. The original Hybrid Demucs consists of two U-Nets, one operating in the time domain and the other in the spectrogram domain. Each U-Net comprises 5 encoder layers and 5 decoder layers, with their outputs combined before reaching a shared layer. In contrast, Hybrid Transformer Demucs retains the outermost 4 layers of this architecture but replaces the innermost 2 layers in both the encoder and decoder with a cross-domain Transformer Encoder, which includes local attention and bi-LSTM.

This adaptation enables the model to concurrently process 2D signals from the spectral branch and 1D signals from the waveform branch. Unlike the original model, which required meticulous parameter tuning to align time and spectral representations, the cross-domain Transformer Encoder adapts to heterogeneous data shapes, enhancing flexibility. The architecture consists of self-attention Encoder layers with normalizations before self-attention and feedforward operations, combined with Layer Scale initialization for stabilization.

The input/output dimension of the Transformer is set to 384, with linear layers used for dimension conversion when necessary. The attention mechanism incorporates 8 heads, and the feedforward network's hidden state size is four times the Transformer's dimension. The cross-attention Encoder layer follows a similar structure but involves cross-attention with representations from the other domain. The cross-domain Transformer Encoder interleaves self-attention and cross-attention Encoder layers in both spectral and waveform domains, with 1D and 2D sinusoidal encodings added to scaled inputs and spectral representation reshaped to treat it as a sequence.

To address memory consumption and attention speed decline as sequence length increases, the model implements sparse attention kernels introduced in the xformer package. Locally Sensitive Hashing (LSH) is employed to dynamically determine the sparsity pattern. A sparsity level of 90% is defined as the proportion of elements removed in the softmax operation. This pattern is established through 32 rounds of LSH, each with 4 buckets. Elements matching at least k times across all rounds of LSH are selected, with k chosen to achieve the desired sparsity level. This modified version is referred to as Sparse HT.

D. Predict:

The sparse kernels, as described in Section 3, are tested to increase the depth to 7 and the training segment duration to 12.2 seconds, with a dimension of 512. This simple adjustment results in an additional 0.14 dB of Signal-to-Distortion-Ratio (SDR), reaching 8.94 dB. The fine-tuning per source further improves the SDR by 0.25 dB, achieving a total of 9.20 dB, and notably requiring only 50 epochs to train. An attempt was made to extend the receptive field of the Transformer Encoder to 15 seconds during the fine-tuning stage by reducing the batch size. However, this did not lead to an improvement in SDR, maintaining the same value of 9.20 dB. It's worth noting that training from scratch with such a context might yield different results

IV. RESULT

The introduction of Hybrid Transformer Demucs marks a significant advancement in audio source separation techniques, extending the Hybrid Demucs architecture with Transformers at its core. This variant replaces inner convolutional layers with a Cross-domain Transformer Encoder, integrating self-attention and cross-attention mechanisms for capturing complex dependencies in audio signals. By combining the strengths of convolutional and transformer architectures, the model achieves superior performance over the baseline Hybrid Demucs, surpassing it by 0.45 dB. Sparse attention techniques enable efficient

scaling for processing longer input lengths during training, reaching up to 12.2 seconds. This scalability not only enhances the model's capacity for longer audio sequences but also improves performance by an additional 0.4 dB, making it applicable to a broader range of real-world scenarios.

Looking ahead, our exploration into splitting the spectrogram into subbands, as proposed in [14], presents an exciting avenue for further enhancement. By processing different frequency subbands separately, we aim to tailor the model's processing to better suit the characteristics of each frequency range. This approach has the potential to further boost separation performance and enhance the model's adaptability to diverse audio sources and environments.

Hybrid Transformer Demucs represents a significant step forward in audio source separation, leveraging the synergy between convolutional and transformer architectures to achieve state-of-the-art performance. With ongoing research and development efforts, we are committed to advancing the boundaries of audio processing and empowering applications across domains such as music production, speech enhancement, and more.

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."
[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

