Audio Segmentation with U-Net architecture: Ablation study of Depth Variations in 1-D and 2-D Audio Representations

Andrew Romero (aromero6@stanford.edu)
Stanford University: EE367

Stanford | ENGINEERING
Electrical Engineering

Abstract

In the study of Music Information Retrieval (MIR), audio source separation is an important subject that's been propelled by the invention of the U-Net CNN architecture. Source separation is used for removing noise from sound, enhancing the quality of vocal recordings, and augmenting text-audio pair datasets for controlled audio generation tasks. In this project, I explore how spectrogram images compare to raw audio when used for the task of audio segmentation. These two formats come from the same dataset, but use two different U-Net architecture types, shown on the right. I show progress towards implementing, training, testing, and comparing the two model types to select an optimal model.

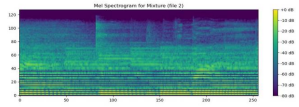
Data and Pre-Processing

Dataset Source

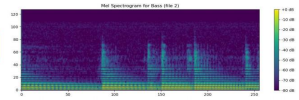
- Musdb18 –150 full length songs, 10 hours of audio
  - Each song is divided into labeled stems:
  - Vocals, Bass, Drums, Accompaniment, and Other
  - 44.1 KHz Sample Rate

Trimming and STFT / Mel Spectrogram

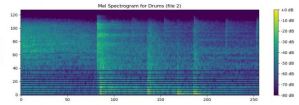
Mix (Input):



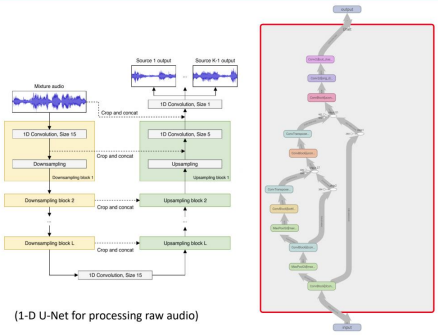
Bass (Output):



Drums (Output):



U-Net Model Architectures



(1-D U-Net for processing raw audio)

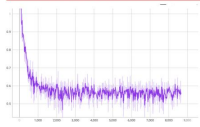
(2-D Unit for processing audio spectrograms)

Training Model

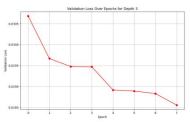
$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

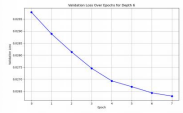
Note: Prediction type is regressive – I attempt to generate a spectrogram for each instrument class in the mix.

Loss Curve for 2D U-Net

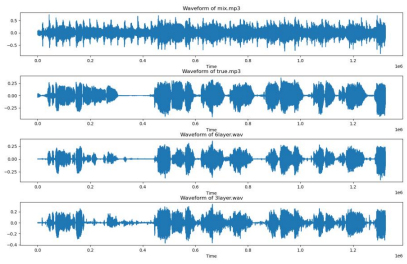


Loss Curve for 1D U-Net





Generation Results (1D)




```
MSE between True and 3-Layer: 0.0016592073952779174
MSE between True and 6-Layer: 0.0014940850025939941
```

Conclusions and Future Work

- The 1D U-Net model with 6 layers, where each layer is a downsample and upsample step, performs surprisingly well. Best performance is heard with drums.
- I notice a significant decrease in audio segmentation quality as the model layer count decreases to 3 from 6. This is consistent with my expectations going into this project.
- For the 2D U-Net model, I observe successful training based on the MSE loss function. The next step is to decode the outputs back into an audio format. The final and stretch goal is to test the U-Net with various layers.
- The goal is to balance computational energy and model accuracy to select a near optimal U-Net type (1D vs 2D, layer depth, etc.) that generative audio engineers can use for dataset augmentation.
- With google Colab, and a V100 GPU, training the 1D model was not possible with a layer depth of 7 or more. I use this to define the upper limit of the layer depth, and I expect this to be lower for the 2D U-Net.

References and Acknowledgements

- Z. Rafii, A. Liutkus, F.-R. Stoter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," Dec. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1117372
- J. Oh, D. Kim, and S.-Y. Yun, "Spectrogram-channels u-net: a source separation model viewing each channel as the spectrogram of each source," 2018
- 1D U-Net implementation from https://github.com/f90/Wave-U-Net
- Special thanks to Ashwin Alinkil for guidance on 2-D U-Net and compute resources.