Audio Segmentation with Modified U-Net Architecture

Andrew Romero

1 Motivation

For this project, I want to explore how the U-Net model is used in audio applications. Modern music generation technologies are based on either diffusion or transformers (token regression). Diffusion has shown to have the best performance in audio quality so far, and is based on the U-Net architecture. In the realm of Music Information Retrieval (MIR), the task of separating sources from a song is foundational yet challenging due to the complexity of sound. [1]

Source separation, as a form of ”feature extraction,” allows for numerous applications, including automatic lyric transcription, singer identification, and music genre classification. The advent of deep learning, particularly Convolutional Neural Networks (CNNs) and U-Net architectures, has shown promising results in various fields, including image segmentation and, more recently, sound domain applications. The ability of these models to extract and interpret complex features from data makes them great for advancing music source separation techniques.

2 Related Work

Jaehoon et. al. explains how they use a modified U-Net model for performing audio segmentation. They also shift towards using other convolutional structures, including Generative Adversarial Networks (GANs) for tasks like voice enhancement and singing voice separation, which underscores the potential of deep learning to revolutionize source separation within MIR. In this project, I want to focus on the UNet architecture to see how changing the architecture changes the quality of the segmentation.

3 Project Overview

I will implement a total of three different UNet architectures and compare their performance using the MUSDB18. [2]. This is a dataset of songs and their stems, the different instruments separated into different audio files. I want the UNet models to determine which values in the spectrogram of the full song belong to which instrument. I understand this research has been done before, but I would like to see how ”shallow” the UNet model can be while still maintaining an accurate classification and segementation of instruments in a full audio file.

Because this project feels vague, I will continue to do research to narrow in on the tasks.

The following tasks for this project are the following:

- Project Proposal 2/23
- Implement various UNet Architectures 3/1
- Finish Dataset - spectrogram images of dataset 3/3
- Train models with the dataset and compare the results 3/7
- Prepare results for presentation 3/11

References

[1] 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.

[2] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, “The MUSDB18 corpus for music separation,” Dec. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1117372