MDX
Music Demixing Workshop

Source Separation: Explorations and Applications

Ethan Manilow[1]

1 Interactive Audio Lab, Northwestern University

Arxiv DOI: 10.21105/joss.01667

In partnership with

ISMIR2021

Abstract

This talk will consist of two parts, the first of which outlines exploratory work in using pretrained, general purpose music models for source separation. I will discuss an unsupervised method that uses large, pretrained music models for audio-to-audio tasks–like source separation and style transfer; all without any retraining. Inspired by the popular VQGAN+CLIP combination for making generative visual art, I accomplish audio tasks by pairing OpenAI's Jukebox with a pretrained music tagger in a system. called TagBox. I will showcase some fun and interesting results, contextualize this method within the rest of the literature, and discuss my excitement about the vast potential that lays relatively untapped in these large pretrained models.

The second part will focus on newly added deep learning capabilities in the Audacity audio editor, which simplifies the process of getting trained models into the hands of end-users. This work lets model creators a way to sidestep DAW-specific development work and enables them to upload pretrained PyTorch models to HuggingFace, where they will automatically be discoverable and runnable within Audacity's UI by end-users. In this talk, I will provide a high level overview of this software framework. We hope this work will reduce the gap between model builders and end-users.

TagBox

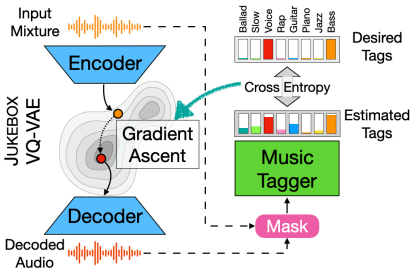

Figure 1: TagBox diagram.

The first part of this talk is about TagBox. TagBox is an method that repurposes deep models trained for music generation and music tagging for audio source separation, without any retraining. An audio generation model is conditioned on an input mixture, producing a latent encoding of the audio used to generate audio. This generated audio is fed to a pretrained music tagger that creates source labels. The cross-entropy loss between the tag distribution for the generated audio and a predefined distribution for an isolated source is used to guide gradient *ascent* in the (unchanging) latent space of the generative model. This system does *not* update the weights of the generative model *or* the tagger, and only relies on moving through the generative model's latent space to produce separated sources. We use OpenAI's Jukebox as the pretrained generative model, and we couple it with four kinds of

pretrained music taggers (two architectures and two tagging datasets). Experimental results on two source separation datasets, show this approach can produce separation estimates for a wider variety of sources than any tested supervised or unsupervised system. I will also show a fun example of how this setup can do unsupervised musical style transfer. This work points to the vast and heretofore untapped potential of large pretrained music models for audio-to-audio tasks like source separation.

Deep Learning Tools for Audacity

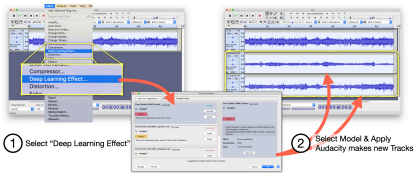

Figure 2: Audacity diagram.

In the second part of this talk, I will talk about our newly developed software framework that integrates neural networks into Audacity. Once a developer has a trained PyTorch model, they are able to compile it using torchscript and upload it to HuggingFace. Once on HuggingFace, the model will be available to users of Audacity, directly accessible through the GUI. This enables end-users to run deep learning models without learning how to code and enables model creators to put their work into the hands of end-users, without extra development. We hope this work will reduce the gap between model creators and artists.