GAN-based Audio Style Transfer

In the past few years, GANs have shown impressive performance on various image-related transfer tasks. In two separate parts, the thesis applies GAN concepts to the tasks of blind audio source separation and audio style transfer.

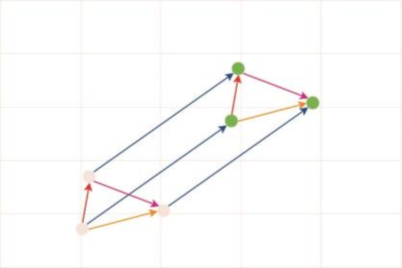
In the first part, we work on audio source separation to establish whether the usage of time-frequency spectrograms or the usage of raw audio waveform data is more expedient to process audio data using CNNs. We determine that it is currently more practical to work on spectrograms, as many practices common in image transfer tasks can be applied on spectrograms as well. However, we note that working with raw audio waveform data is possible and may be more effective with further advances in the field. We then apply GANs on audio source separation tasks and establish that cGANs can be used for this task but note that our implementation does not perform better than classical CNNs.

In the second part, we apply GANs to transfer music styles to establish the state of the art. We investigate the TravelGAN, which uses transformation vector learning using a siamese network to preserve high-level semantic information instead of cycle consistency. TravelGAN has been shown to perform well in translation tasks for image domains with only very high-level semantics. We implement a MelGAN, an application of a TravelGAN, which has shown to work on audio style transfer and compare it with a CycleGAN, which is used for various image style transfer tasks. We determine that a MelGAN is the most promising network architecture for transferring audio style.

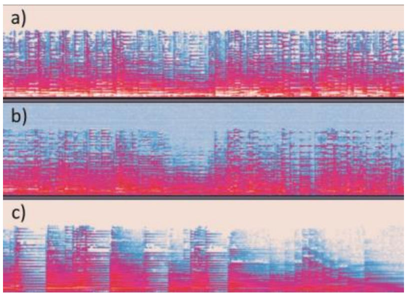
We experiment with different domain sizes to establish the boundaries of MelGAN capabilities and conclude that it works better when both domains are somewhat narrowly defined, e.g. Piano and Guitar music as opposed to Jazz and Classical music. Training time was shorter and perceived music quality was significantly better with reduced domain sizes. We publish our best implementation of a style transfer GAN on GitHub.

Diplomierende
Gabriel Koch
Raphael Mailänder
Michael Schaufelberger

Dozierende
Martin Loeser
Matthias Rosenthal



The siamese network learns to assign each song a point in a vector space. Gray dots represent the original song and the green dots the transformed song. The transformation vector (blue) should be the same for all songs.



a) shows a guitar song given as input for the network, b) represents the song transformed to piano music by our network, and c) shows a random sample from the piano music domain.