Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering
InES Institute of Embedded Systems

# GAN-based Audio Style Transfer

Diplomierende
Gabriel Koch
Raphael Mailänder
Michael Schaufelberger

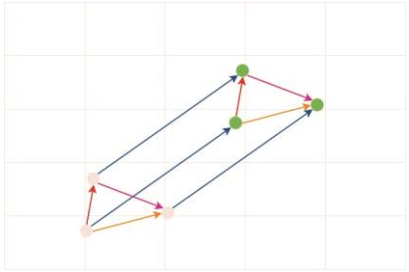
Dozierende
Martin Loeser
Matthias Rosenthal

Jüngst wurden Generative Adversarial Networks (GAN) erfolgreich eingesetzt, um Style-Transfer auf Bildern umzusetzen. In zwei Teilen zeigt diese Arbeit, wie GANs auch eingesetzt werden können, um Musik in verschiedene Kanäle wie Gesang und Instrumente aufgeteilt werden können (Audio Source Separation), und wie Style-Transfer auf Musik (Audio Style Transfer) möglich ist.

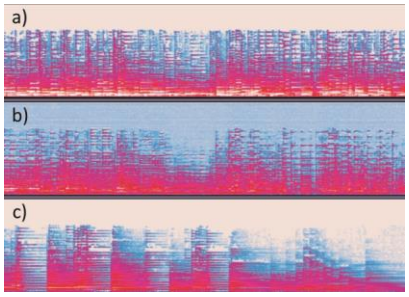
Im ersten Teil wird anhand des Audio Source Separation Problems untersucht, ob sich das Umwandeln der Audio-Dateien in Spektrogramme lohnt oder ob die Arbeit mit den ursprünglichen Audio-Dateien ebenfalls möglich ist. Die Resultate zeigen, dass die Nutzung von Spektrogrammen in der Anwendung praktikabler ist, aber dass Audio-Dateien auch direkt mit neuronalen Netzwerken erfolgreich verarbeitet werden können. Die Implementation eines GANs auf Basis der Conditional GAN (cGAN) Architektur für Audio Source Separation zeigt, dass GANs ähnliche Resultate liefern wie klassische, neuronale Netzwerke und für die Manipulation von Audio-Daten genutzt werden können.

Im zweiten Teil wird der aktuelle Stand der Technik von GANs für Audio Style Transfer untersucht. Die TravelGAN Architektur nutzt ein siamesisches Netzwerk, um semantische Information der Musik in einem Vektorraum abzubilden und darauf Transformationsvektoren zu ermitteln, wo andere Ansätze wie das CycleGAN oft auf Cycle-Consistency setzen. Das TravelGAN wurde schon erfolgreiche eingesetzt, um Style-Transfer zwischen sehr unterschiedlichen Domänen auf Bildern zu implementieren. In dieser Arbeit wird ein MelGAN, eine Anwendung der TravelGAN Architektur für Audio, implementiert. Ein MelGAN wurde schon erfolgreich für Style-Transfer eingesetzt und wird in dieser Arbeit mit einer Implementation eines CycleGANs verglichen. Die Resultate zeigen, dass ein MelGAN nach kurzer Zeit gute Ergebnisse liefert, während ein CycleGAN kein Style-Transfer lernt.

Experimente, um die Grenzen der MelGAN Architektur auszuloten, zeigen, dass die Auswahl der Trainingsdaten essenziell für die Performance des Netzwerks ist und insbesondere, dass eine kleinere Varianz innerhalb der Quell- und Zieldomänen wie etwa verschiedene Klavierstücke statt das ganze Genre der klassischen Musik bessere Resultate liefert. Das Trainieren eines GANs auf einer kleineren Domäne dauert weniger lang und die Ergebnisse klingen besser. Unsere MelGAN-Implementation wurde auf GitHub veröffentlicht.



Das siamesische Netzwerk weist einem Song einen Punkt in einem Vektorraum zu. Der Transformationsvektor (blau) sollte für jedes Paar von originalem Song (grau) und transformiertem Song (grün) gleich oder ähnlich sein.




a) zeigt ein Gitarren-Song, b) zeigt den selben Song nach der Transformation durch unser Netzwerk zu einem Klavierstück und c) zeigt ein zufälliges Beispiel aus der Zieldomäne von Klaviermusik.