Generative Neural Networks for the Sciences

WS24/25

Project Proposal

Cristi Andrei Prioteasa
(Uni-ID: qf323, Matrikel-Nr.: 4740844)
Theo Stempel-Hauburger
(Uni-ID: fk323, Matrikel-Nr.: 4740729)

A GNN WS24/25 Homework Assignment

January 21, 2025

Project Proposal

We propose the following two project topics for the final project.

Our group prefers proposal 1. However, we are open to suggestions and feedback before we start working on it.

We have participated in the course in the WS23/24 and did not pass the final project but had the points from the exercises already from that semester.

Proposal 1: Music Style Transfer With Diffusion Model

The approach proposed in this paper is to use Latent Diffusion Models conditioned on instrument to do style transfer (instrument transfer) to short pieces of instrument only samples of music. In order to apply a Latent Diffusion Model for this task, the audio samples are encoded into spectrograms, which offer a spatial representation that can be harnessed by the underlying architecture of the LDMs.

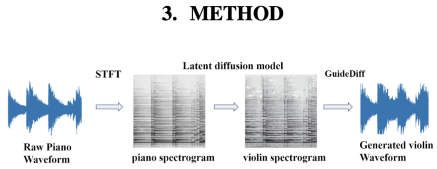

Figure 1. Piano to violin style transfer.

(a) Style transfer for audio samples.

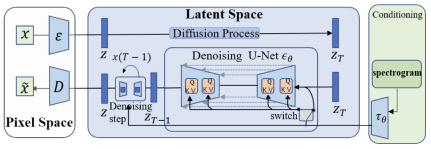

Figure 3. Models of transfer.

(b) Architecture.

Our project will consist of:

1) **Data Collection**: we plan to collect around 30 000 WAV samples, lasting 5 seconds each, coming from different individual instruments such as piano, guitar, harp, trumpet etc. This can be easily acquired from publicly available sources and since we will be mainly dealing with single instrument pieces like piano or guitar we can generate many samples from a single piece, even if that compromises our diversity. The genre of the music being played is not relevant for our task, but the samples should contain only individual instruments. The dataset will be then preprocessed to spectograms which will be used as the input to our model.

2) **Architecture**: The main architecture will be an LDM operating on the spectograms and conditioned on the spectograms of the other instrument for style transfer (instrument transfer). The original paper also employs a quality enchancing network which interpolates out of place pixels in the generated spectogram before converting it back to audio. This quality enchancement network will not be the main part of our project, and we will try to implement it if we have the time.

3) **Training**: In terms of training we will stick close to the details proposed in the paper, but downscale so it can fit on our training devices. We want to experiment with the generated sample length but keep it between 2 and 5 seconds. The paper uses 3

NVIDIA RTX3090Ti, 500k training steps and a batch size of 100. We have access to a RTX3060Ti and our personal laptops. We hope than downscaling the generation time and training steps will fit into our GPU size.

4) **Experiments and Training**: we want to assess the quality of the generated samples in terms of standard metrics such as the Fréchet inception distance, but also look into what metrics would be more suitable considered that we deal with audio signal instead images and music style transfer. We will try to stay close to the experiments and metrics used in the proposed paper, but also challenge them and see if we can come out with something else.

We find this paper interesting because it uses LDMs traditionally designed to generate images and applies it to audio data. Possible motivation for this could be ringtone generation, instrument transfer and missing section interpolation / enchancement (for example when a part of the audio of a song is affected by noise). We expect to encounter some problems regarding the training capabilities, but LDMs with DDIM should be efficient enough to work on our hardware for small length samples. We are looking forward to feedback on this and we would appreciate if you took a short look on the mentioned paper, since a lot of the actual implementation details were left out from this report but are present in the paper.

Proposal 2: Neural Discrete Representation Learning - Aaron van den Oord et al. (2017)

The paper introduces the innovative VQ-VAE framework with compelling results in generative tasks. Demonstrates clear advantages in interpretability and scalability. There is also a newer paper, "Generating Diverse High-Fidelity Images with VQ-VAE-2" (Razavi et al., 2019) which extends VQ-VAE to hierarchical architectures, achieving state-of-the-art image synthesis, but gives it more computational costs which is not feasable for our resources. We want to build upon the original VQ-VAE paper by replicating its results and extending its application to a new dataset (something you may perhaps propose as feedack).

Our project will consist of:

1) **Method**: Implement the VQ-VAE framework as described in the original paper.

2) **Dataset**: Cifar-10, CelebA or other data you may propose as feedback. In terms of quantity, we are thinking of 60 000 - 100 000 images of maximum 32x32 resolution.

3) **Experiments**: Experiment with different codebook sizes to understand how varying the size of the codebook affects the quality of the learned representations, model performance, and reconstruction accuracy.

4) **Interpretability**: specific codes may correspond to distinct visual features like color, shape, or texture, which we want to explore in terms of interpretability.

5) **Hardware**: same as in the last proposal.

We are open to more suggestion for this proposal and others things we could do, but as stated previously we are more excited about the first proposal.