Pozalabs Assignment

The assignment was to design a generative model that takes inputs of violin and flute voices and generates a new voice with the timbre of a combination of violin and flute, with the restriction that the training data, taken from MusicNet [1], consists of separate solo violin and flute recordings, each of which were from different pieces. The inference task was to convert a given flute/violin piece into the new voice. Due to the lack of time and resources, the proper training and inference steps were not completed, but the theoretical background and base implementation are provided.

Many approaches were considered for this task - most notably Magenta's NSynth [2]. This model uses a WaveNet-style autoencoder [3], which allows for end-to-end audio generation, and its application in timbre interpolation has been successfully showcased. Initial attempts involved both recreating the model architecture or re-training the base model, but the constraints on time and GPU resources of the assignment and lack of suitable data made it so that these options were not viable. This would, however, be the ideal and state-of-the-art approach.

The approach taken instead was to implement custom generative models that complied with the given restrictions. Examples of such models include Variational Autoencoders (VAE) and Generative Adversarial Networks (GAN), both of which have been applied to audio generation tasks in previous works [4, 5]. In both approaches, however, there is a limitation regarding their ability to support end-to-end generation, due to performance issues when feeding and generating raw audio formats. The alternative is to use spectrograms, obtained by applying a short-time fourier transform, either as images or raw arrays. The theory and implementation details of both approaches are detailed below.

Variational Autoencoder (VAE)

A VAE is a generative model that learns to encode and decode data while capturing a probabilistic distribution of latent representations. They are a potential candidate for the task of combining timbres because they can learn meaningful and continuous latent representations of audio data from both the violin and flute recordings, making it possible to interpolate between different timbres and create smooth transitions. In theory, a VAE should be able to compress the high-dimensional audio data, including timbre, into a low-dimensional latent space, and use this latent space to interpolate and generate new sounds. A similar idea has been applied for music synthesis and timbre transfer [4], but timbre combination is a rather unexplored topic.

In terms of the implementation, the audio data firstly needs to be preprocessed so that it can be used for training. This was done using the Librosa library, which converted the raw audio into spectrograms of two formats - numpy arrays and images. This was done to support training of both VAE and CycleGAN (explained below). Parameters such as hop size and n_fft were determined based on values that would most optimise the training process, not considering the loss in quality, as the training takes place with limited GPU resources. The audio samples were segmented into 5-second snippets, which were each converted into the respective formats. The VAE implemented was a simple architecture with input size (1024, 128, 1), 5 convolutional layers, and 128 latent space dimensions. These parameters will likely require tweaking as the model gets trained and tested.

CycleGAN

Another suitable approach that has been explored is the CycleGAN [6]. CycleGAN is a type of generative adversarial network that employs the idea of cycle consistency. This is the concept of optimising two generators F and G that map between two domains A and B (in this case flute and violin timbres), such that $x \approx F(G(x))$ and $y \approx G(F(x))$. In other words, if a violin track is converted to flute via F, G should map the converted track back to violin. Since the CycleGAN architecture was designed for unpaired image-to-image translation tasks, it suits the design of this assignment well, as the provided violin and flute soundtracks are unpaired. This architecture has been used for audio generation tasks such as genre transfer successfully [5]. Ultimately, the generators will learn the mappings between either instrument, which makes the task of timbre combination more feasible. The task can then take a variety of approaches. One way is to use the now paired violin and flute data to mix the sounds using audio blending techniques, which can then be used as ground truth labels to condition the training process of another model (either a VAE or vanilla GAN) for timbre transfer. This process hasn't been fully implemented in the code, but is a very promising avenue to explore and is theoretically feasible.

Limitations & Future Improvements

There were various limitations to this project, both inherently and logistically. Overall, the lack of time and resources had an impact on the choice of model and how robustly they were implemented. Only the VAE was fully implemented (but not trained and optimised), while the CycleGAN approach needs further implementation. The models themselves may also have limitations in the context of this project. For the VAE, the fact that the training data is not uniform can lead to unstable training and thus unpredictable results. One way to account for this is to employ a multi-encoder VAE (ME-VAE) architecture [7]. This would allow for separate encoders to be used for each instrument, which could stabilise the training process. The main challenge of the task was the unpaired, raw dataset, which many models struggle to handle. Although the CycleGAN accounts for that, the model needs to be trained and thoroughly tested to evaluate its viability for this task.

References

1. Thickstun J, Harchaoui Z, Kakade S. Learning Features of Music from Scratch. arXiv; 2017. Available from: http://arxiv.org/abs/1611.09827
2. Engel J, Resnick C, Roberts A, Dieleman S, Eck D, Simonyan K, et al. Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. arXiv; 2017. Available from: http://arxiv.org/abs/1704.01279
3. Oord A van den, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, et al. WaveNet: A Generative Model for Raw Audio. arXiv; 2016. Available from: http://arxiv.org/abs/1609.03499
4. Caillon A, Esling P. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. arXiv; 2021. Available from: http://arxiv.org/abs/2111.05011
5. Brunner G, Wang Y, Wattenhofer R, Zhao S. Symbolic Music Genre Transfer with CycleGAN. arXiv; 2018. Available from: http://arxiv.org/abs/1809.07575
6. Zhu JY, Park T, Isola P, Efros AA. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. arXiv; 2020. Available from: http://arxiv.org/abs/1703.10593
7. Ternes L, Dane M, Gross S, Labrie M, Mills G, Gray J, et al. A multi-encoder variational autoencoder controls multiple transformational features in single-cell image analysis. Commun Biol. 2022 Mar 23;5(1):1–10.