Music Inter-class Disentanglement

Matan Halfon

Abstract

Research In style transfer and domain translation has demonstrated the ability of deep learning algorithms to manipulate images using a robust learnt representation of the domain using disentanglement methods, Several works have tried to extend such approaches to speech and music(Mor et al. [2018]). Translating Music across musical instruments Is an inherently hard task, as human perception is sensitive to both global structure and fine scale waveform coherence and to preform a conversing translation both small scale and big scale have to match. In this work we Adjust the disentanglement method LORD(Gabbay and Hoshen [2020]) to music to create a new disentangled representation of instrument and content.

1 Introduction

The ability to learn and replicate music using a different instrument is as old as music itself. This skill is not unique only to humans as there are various species of animals that can mimic sounds upon hearing them. nevertheless the ability to separate a body of music into separate instruments and the general content of the music track (tempo, harmonics, pitch, Scale, etc..) in order to later translate the music content to a new instrument, and creating a novel track, is a challenging open problem in machine learning. this is due to the high temporal resolution of the data and the presence of a global structure over different timescales.Due to these challenging issues instead of modelling a raw temporal audio directly it is common to simplify the problem to a lower resolution representation that can be efficiently computed from the raw data and inverse back to raw audio with a small loss of information.

Nevertheless deep learning models have shown great success with working on the direct waveform such as WaveNet (van den Oord et al. [2016]) to syntheses new audio tracks (speech/music) which solve the scale problem by focusing on the finest scale possible (a single audio sample) and rely upon external conditioning signals for global structure. This comes at the cost of slow sampling speed, since they rely on inefficient ancestral sampling to generate waveform one audio sample at a time. Due to their high quality, a lot of research has gone into speeding up generation, but the methods introduce significant overhead such as training a secondary student network or writing highly customized low-level kernels (van den Oord et al. [2018]). Furthermore, since these large models operate at a fine timescale, their autoencoder variants are restricted to only modeling local latent structure due to memory constraints (Engel et al. [2017]).

As for audio synthesis also in music translation there are great works that employ waveform models on this task(Mor et al. [2018]), this is also demand an heavy computational cost. Others have tried to work on a symbolic with representation (Cífka et al. [2019]) or frequency domain (Pasini [2019]) to preform a convincing translation, some more or less successfully than others.

The music translation problem is essentially a disentanglement problem s.t the labeled attribute is the instrument identity and the unlabeled attributes are the track content that we assume that is uncorrelated with the instrument,

Disentanglement in computer vision has greatly benefited image translation in which the goal is to translate a given input image from a source domain to an analogous image in a target domain.we employed this methods to music translation.

Preprint. Under review.

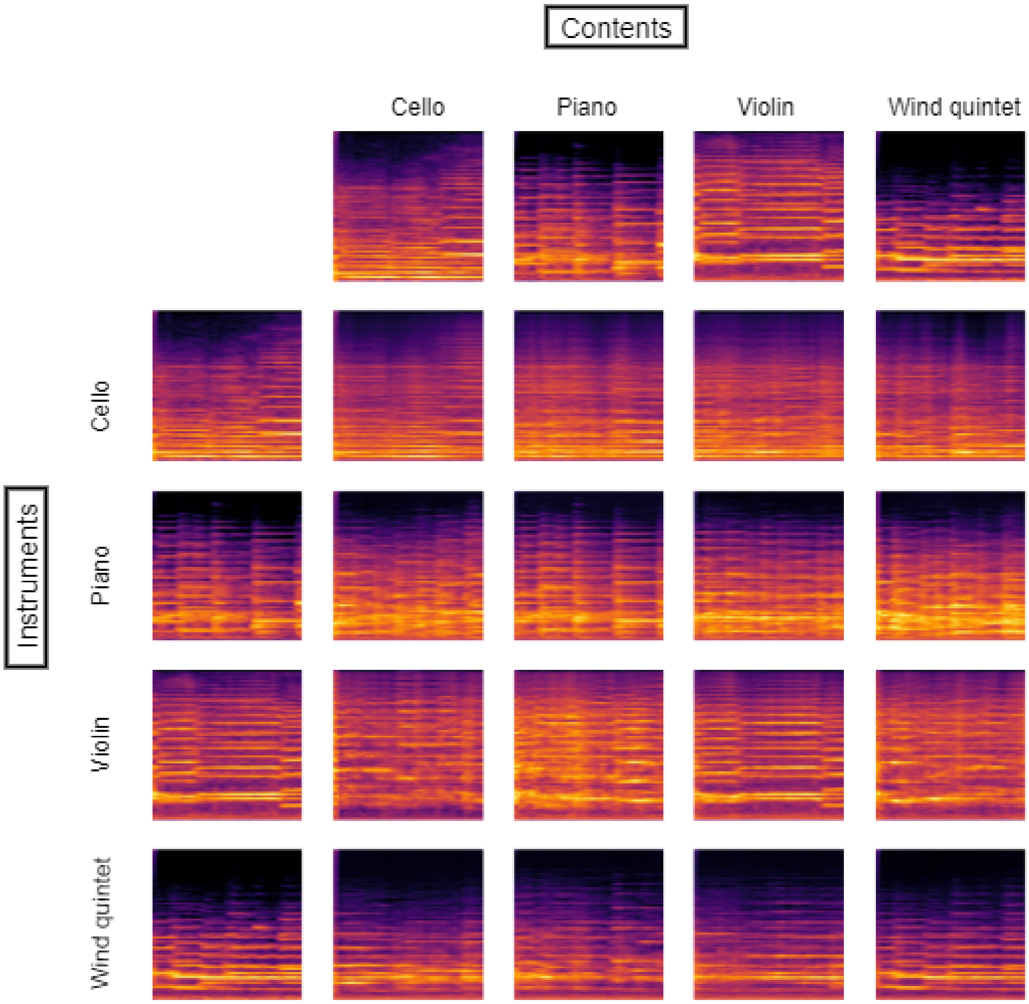

Figure 1: music translation Lord model, the original spectugrams appeared in the the first row and column,The inner parts of the figure are the spectugram created by the model s.t the class code is derived from the spectogram appeared in it row and the content row is from the column spectugram.

In this work we apply a latent optimization framework for disentanglement LORD(Gabbay and Hoshen [2020]), the main challenges we met were the the lack of good perceptual loss to measure the successes of the reconstruction from the regularized embedding,more over a "fuzzy" image is a more damaging to our result then for image translation that keep the overall structure even when the result are more blurry, therefor we tested several variations for the latent decoder to generate a more accurate results the work can be found in the the following git repository [1].

2 Previous Work

domain translation and and disentanglement have gained great achievements in the field of computer vision in resent years with the use of of adversarial learning Choi et al. [2018],Denton and Birodkar [2017]) and non-adversarial constraints such as cycle (Jha et al. [2018]) or variational group codes (Bouchacourt et al. [2018]). many works try to learn disentangled representation for style and content Abdal et al. [2019]. Domain translation is not restricted to a single pair of domains and can map between multiple domains(Choi et al. [2018],Denton and Birodkar [2017]), and even unseen domains(Gabbay and Hoshen [2020],Mor et al. [2018]) .

[1] https://github.cs.huji.ac.il/matanhalfon/Lord-music

Audio synthesis has fallowed the rapid advances in computer vision and there have been great works in this field using auto-regressive generation (van den Oord et al. [2016],Dhariwal et al. [2020]) and unconditional generation (Engel et al. [2019]) of new music in the several representation of audio. works in the audio-translation have followed this path and translate voices between people and melodies across genres and instruments (Mor et al. [2018]) inspired by image translation methods and potentially enabling the creation of instrumental music by untrained humans.

3 Methods

3.1 Representation and Data

We are using a log Mel frequency representation "images" as it is a representation that can easily inversion back to audio and better capture the nature of different scale structure of music, the Mel-frequency representation captures the global and local variation appeared in music nevertheless this representation has some drawbacks, the fact that the phase is lost makes the recovery of the Mel-spectogram far from perfect this recovery is even more noticeable when working with music in comparison to speech due to the frequency rich nature of music, We dealt with this issue by using a smaller width of windows that make the loss of phase less severe, and also work with a relatively high number of Mel bands to deal with the variations of high frequency which that can be found in music. To recover the spectogram back into the wav form we used Griffen-lim(Griffin and Lim [1984]) recovery method with 64 iterations of the librosa packege (McFee et al. [2015]).

Each sample is a 2 second interval sampled in a 44kHz that is resembled to 16kHz and then transformed to a spectogram representation of 2048 fft with a window size of 10 ms using an Hann windowing(*)for overlapping, and used the Mel bank filter to quantize the frequency's to 128 Mel levels s.t every sample is an image of [128,128,1].

The data from the MusicNet(Thickstun et al. [2016]) data set an only used the 4 different solo instruments: Cello,Piano,Violin and Wind Quintet to avoid to avoid Wave interference with in ensembles with overlapping domains. We have tried to disentangle 4200 samples.

4 disentanglement model

4.1 Class and Content disentanglement

First we define the section disentanglement problem we are dealing with. Assume that we are given a collection of n tracks $x_1, x_2, ..., x_n \in X$. For each track xi , we are given a class label $y_i \in [k]$. We assume that every image belongs to a single class. Note that many images may share the same class label. We denote the embedding of a given class y as e_y. We assume that the tracks can be disentangled into representations in two latent spaces $\mathcal{Y}$ and $\mathcal{C}$. Therefore our objective it to find a class representation $e_{y_i} \in Y$ and a content representations $c_i \in C$ for each 2 image x_i . Let us define the information that we wish each representation to contain. The track class representation e_{y_i}, needs to include all information that is shared by all images sharing the same class e.g. The content representation c_i includes all the information that is unchanged if the image is transferred between classes. This information must be independent of the class-information. Besides the class and content representations. We denote the style representation as $s_i \in S$. We define a generator G, a neural network parameterized by θ, which transforms the disentangled representations into an track. Given our definitions above each image can be modeled by:

$$x_i = G_\theta(e_{y_i}, s_i, c_i) \quad s.t \quad x_i \in X \;\; e_{y_i} \in Y \;\; s_i \in S \;\; c_i \in C \tag{1}$$

The content must be independent of the class and style, In many cases, it can be assumed that inter-class variation is significantly larger that intra-class variation. Many approaches were devised to learn disentangled representations for this scenario, in which s_i contains both class and content information of an track x_i.

4.2 LORD

We learn the content representation by optimizing over per-sample content embedding directly using latent optimization and not in an amortized fashion using an image to content encoder. As we show

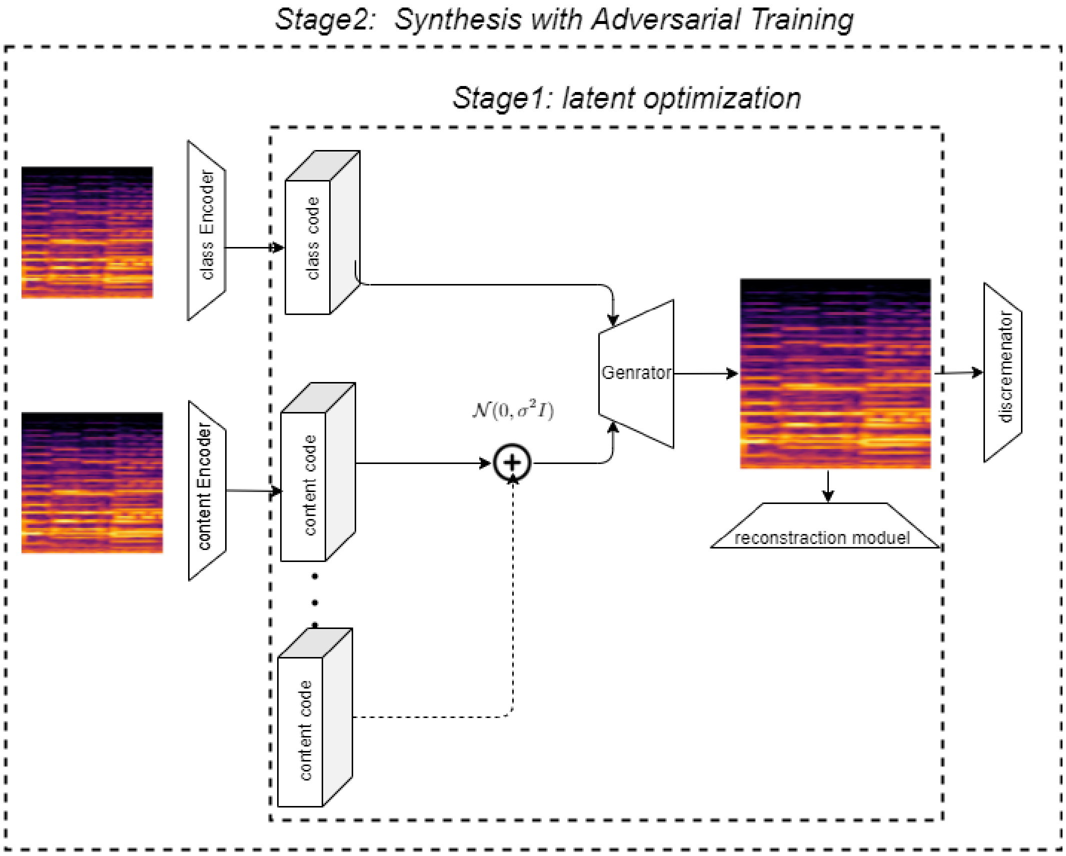

Figure 2: For disentanglement class and content embedding and the generator are jointly optimized. All tracks of the same instrument share a single class embedding. The content embeddings are regularized by a Gaussian noise and a dimension bottleneck . By the end of this stage, the latent space of the training set is disentangled . For synthesis, we tune all modules in an amortized fashion using the learned embeddings as targets of the new encoders. During this stage, an adversarial discriminator is trained to increase the spectogram quietly.

in the experimental section, a model trained with latent optimization preserves a very high degree of disentanglement along the training and is less sensitive to hyperparameter choices. Asymmetric Noise Regularization: Latent optimization over the class embedding ensures that no content information is present in the class representation. To ensure that class information does not leak into the content representation, we regularize the content code to enforce minimality of information. Previous approaches attempted to minimize content information by setting a bottleneck of a small content code or by matching the content distribution to a prior normal distribution using *KL-divergence*. Using a small noiseless bottleneck, does not however reduce information significantly. In our approach, we regularize the content code with an additive Gaussian noise of a fixed variance,and an activation decay penalty. In contrast to a variational auto-encoder, we do not learn the variance, but rather keep it fixed. This prevents the possibility of the variance decreasing to a small value, ensuring that noise is applied equally on all components. Our objective function becomes:

$$\mathcal{L} = \sum_{n=1}^{n} ||G_\theta((e_{y_i}), 0, C_i + Z_i) - x_i|| + \lambda||c_i|| \quad \mathcal{N}(0, \sigma^2 I) \tag{2}$$

The first loss terms is a perceptual loss we found that the best results we got were when using a combined loss that will be discussed in the audio loss compression part . Unless stated otherwise, we optimize over class and content codes (e_{y_i} and c_i) directly using latent optimization.

5 Audio loss compression

In the original Lord paper the perceptual loss a VGG perceptual loss as implemented by (Hoshen et al. [2019]) which is very robust perceptual loss that is known to work for images from distinct domains that does not appear in the original data set, one of main issues that had to be dealt was the lack of quality perceptual loss for music and for audio in general and for music in particular. the need for perceptual loss is critical for the success of the model therefor we tested several options.

5.0.1 Vector norm loss

the use of general matrix norm methods such as $\mathcal{L}_1$ or $\mathcal{L}_2$ tend to "average" the outputted spectugram and create a "fuzzy" spectogram with a lot of additive noise in all frequncey channels equally, moreover this kind of loss will not take care of any global changes as only work "pixel-wise" on the spectogram and loss the global structure of the melody and there for will not enforce a good reconstruction-loss alone.

5.0.2 Perceptual vision loss

In the original implementation of Lord the reconstruction loss was the Vgg perceptual loss as described in(), this loss was preforming well on the spectogram data but had few drawbacks, first the compatibility with the domain (RGB-images to Mel-spectograms) that is not strongly connected and there for there is no real cause the features extracted from this network that will be very meaningful for a spectugram. furthermore the lower frequencies (e.g the lower rows of the spectugram image) have a bigger impact on the recovered sound then for the higher frequencies(e.g the higher rows of the spectugram) this impotent bias is not expressed at all in the Vgg loss.

5.1 Perceptual sound loss

As for the drawbacks of the Vgg-loss we had tried to find a good feature extractor network that use the special biases of the of the frequency domain, I have tried two feature extractor network that are trust worthy pre-trained audio classifier in the frequency domain : 1.Vggish network(Gemmeke et al. [2017]),2.Rho(Koutini et al. [2020]). We have found that those classifiers that worked well for audio classification do not tend to capture the entire structure of the spectugram but the only the general timbre of the sound and there for enforce only lower or higher frequency but not accurately the similarity of two spectugrams. i can only guess that the task of audio classification does not require to capture the entire spectogram structure and the classification can be based on more global features (e.g the power of the lower frequencies, the ratio between them,ect...), never the less the Rho network had given better results for classification tasks and also gave better results when used as a perceptual loss.

5.2 Loss combination

After numerous compassion's of suggested losses metrics we found the best result when applying a combined loss metric of the Vgg-loss, Rho-loss and the spectral norm ($L_{2,2}$) as following:

$$\mathcal{L}_{rec} = w_{vis} \cdot \mathcal{L}_{Vgg} + w_{sound} \cdot \mathcal{L}_{Rho} + w_f \cdot \mathcal{L}_{2,2} \tag{3}$$

5.3 Losses function tried

We have tried several other perceptual loss that did not preform well anther the model and optimization implemented,

self-supervised loss- we tried to harness the self supervised embedding that was presented cola(Saeed et al. [2021]) and was trained on gtzan(Sturm [2013]) data-set data-set, the embedding that resulted from the training got $\approx 95\%$ accuracy for a linear classifier but had not preformed better then the supervised methods (e.g Rho and Vggish) as a sound perceptual loss, we assume that a self-supervised embedding trained on a very big data-set will maybe preform better due to lack of time and resources we choose not pursue this way. We also tried several sound losses function suggested in the package aurora-loss (**) but they didn't add a additional value to the resulted sound.

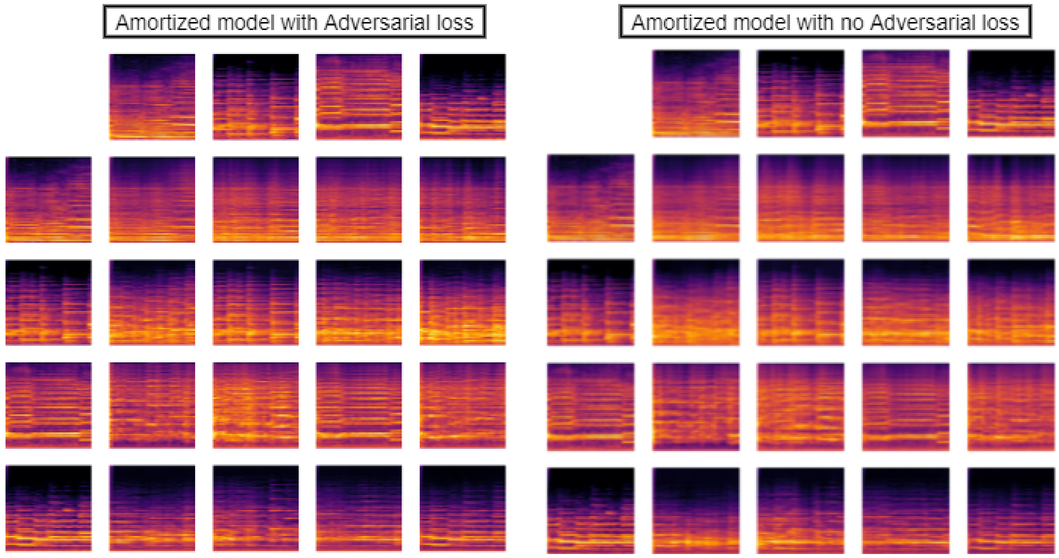

Figure 3: compering the results for the amortized model with and without using adversarial loss, the defenses are not very clear at first sight, but there is an noticeable deference in the high frequencies due to the discriminator being able to detect abnormal volume of high frequencies as not realistic.

6 Generating spectograms

One of The main objective in the music translation task is to use a spectugram generator that will employ as much inductive bias to the Lord framework as passable nevertheless producing and "sharp" and spectogram to avoid ambient-noise that will be added as a result of a blurred spectugram.

The original Lord generator did not preform well due to the fact in lacks inductive bias and also quite a big generator that enforce harder optimization and require more data, while the spectograms are 1 channel sparse images and do not require such a expressive generator.

We Choose to use the genrator of the Adain-Vc(Chou et al. [2019]) paper which include a more adapted generator for the task due to higher inductive bias and less parameters and enables the use of smaller data and faster optimization.

To improve the resolution of the spectogram we added an adversarial loss to the amortized stage of, as a discriminator we used the model suggested in the GaNsynth(Engel et al. [2019]) paper and decreased its size, and a used Wassersatin loss (Gulrajani et al. [2017]) with a gradient penalty,The use of adversarial loss helped the generator to avoid abnormal high frequencies in the generated spectugrams (see figure 3) .

7 disentangle criteria

To assess the disentanglement of our learned representations, we tried to cluster the content vectors using t-SNE clustering method(Van der Maaten and Hinton [2008]) we hope that if the representations are not disentangled the t-SNE algorithm could preform any clustering on the content vectors as seen in some works(Husnain et al. [2019],Dupont et al. [2013],Benjamin and Altosaar [2015]) that could cluster spectugrams according to music style and instruments using t-sne, our learned representation could not be clustered by t-SNE algorithm which employ that the representation are somehow disentangled the results can be seen in figure 2, at first we wanted to asses the disentanglement learned representations using the protocol in (Jha et al. [2018]), but the lack of any compression in this domain would make it redundant, with no base line classifier and comparable results combined with the small data-set given the accuracy will be a factor of the size of the used classifier.

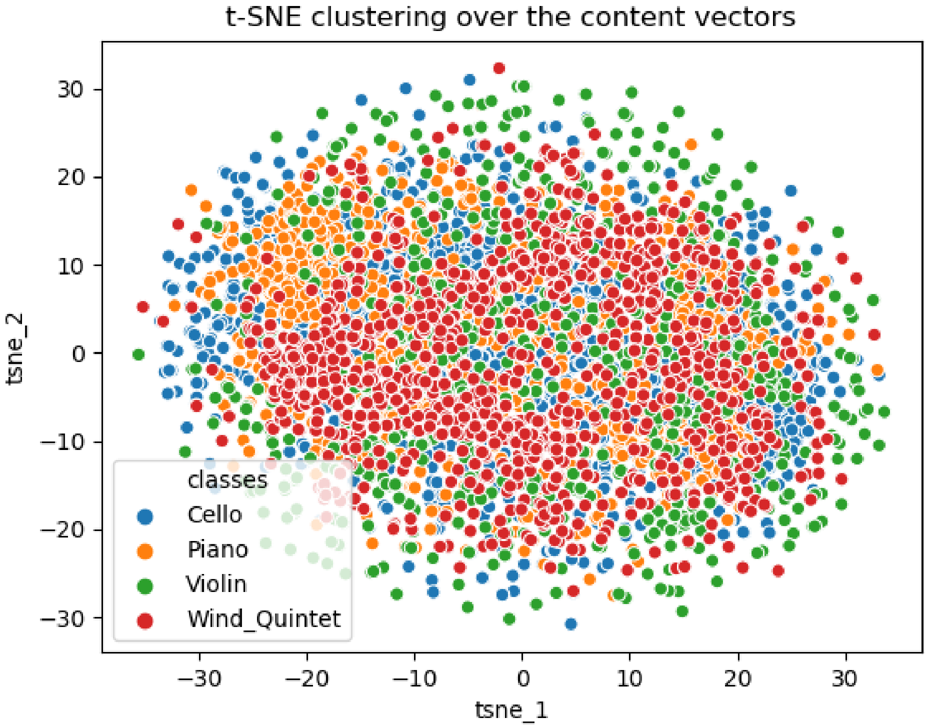

Figure 4: clustering of the content vectors that is colored by instrument as can been seen t-sne could not cluster any good results .

8 Discussion and future work

music-representation There are numerous ways to represent audio some are more adapted to music and some are less. The Mel-spectogram is a very known method to visualize and analyse audio tracks and that is why i choose to use this representation. Never the less this representation is more suited to speech tasks. There are less common representation that could be more suited for music oriented tasks such as CQT (Brown and Puckette [1992]) or Rainbowgrams (Engel et al. [2017])variations that weighted frequencies by octaves and better capture the audio and especially music (Fuentes et al. [2012],Huang et al. [2018]) that we as human beings can notice,and could inject even more inductive bias to the model.

Perceptual loss Another drawback that is even more noticeable in the representation mentioned above is the lack of strong similarity metric and specially perceptual similarity metrics. A better feature extractor methods for the music domain could dramatically improve the results.This could potentially be derived via self-supervised methods such as the one tried with no success in this work (see section *5.3*). using this methods on very big data could create the much needed quality embedding for audio and music in particular.

Bibliography

Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019.

Ethan Benjamin and Jaan Altosaar. Musicmapper: interactive 2d representations of music samples for in-browser remixing and exploration. In *NIME*, pages 325–326, 2015.

Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Judith C Brown and Miller S Puckette. An efficient algorithm for the calculation of a constant q transform. *The Journal of the Acoustical Society of America*, 92(5):2698–2701, 1992.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

Ju-chieh Chou, Cheng-chieh Yeh, and Hung-yi Lee. One-shot voice conversion by separating speaker and content representations with instance normalization. *arXiv preprint arXiv:1904.05742*, 2019.

Ondřej Cífka, Umut Şimşekli, and Gaël Richard. Supervised symbolic music style translation using synthetic data. *arXiv preprint arXiv:1907.02265*, 2019.

Emily Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. *arXiv preprint arXiv:1705.10915*, 2017.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.

Stéphane Dupont, Thierry Ravet, Cécile Picard-Limpens, and Christian Frisson. Nonlinear dimensionality reduction approaches applied to music and textural sounds. In *2013 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2013.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In *International Conference on Machine Learning*, pages 1068–1077. PMLR, 2017.

Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. Gansynth: Adversarial neural audio synthesis. *arXiv preprint arXiv:1902.08710*, 2019.

Benoit Fuentes, Antoine Liutkus, Roland Badeau, and Gaël Richard. Probabilistic model for main melody extraction using constant-q transform. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5357–5360. IEEE, 2012.

Aviv Gabbay and Yedid Hoshen. Demystifying inter-class disentanglement. In *International Conference on Learning Representations (ICLR)*, 2020.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017.

Yedid Hoshen, Ke Li, and Jitendra Malik. Non-adversarial image synthesis with generative latent nearest neighbors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5811–5819, 2019.

Sicong Huang, Qiyang Li, Cem Anil, Xuchan Bao, Sageev Oore, and Roger B Grosse. Timbretron: A wavenet (cyclegan (cqt (audio))) pipeline for musical timbre transfer. *arXiv preprint arXiv:1811.09620*, 2018.

Mujtaba Husnain, Malik Muhammad Saad Missen, Shahzad Mumtaz, Muhammad Muzzamil Luqman, Mickaël Coustaty, and Jean-Marc Ogier. Visualization of high-dimensional data by pairwise fusion matrices using t-sne. *Symmetry*, 11(1), 2019. ISSN 2073-8994. doi: 10.3390/sym11010107. URL `https://www.mdpi.com/2073-8994/11/1/107`.

Ananya Harsh Jha, Saket Anand, Maneesh Singh, and VSR Veeravasarapu. Disentangling factors of variation with cycle-consistent variational auto-encoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 805–820, 2018.

Khaled Koutini, Hamid Eghbal-Zadeh, Verena Haunschmid, Paul Primus, Shreyan Chowdhury, and Gerhard Widmer. Receptive-field regularized cnns for music classification and tagging. *arXiv preprint arXiv:2007.13503*, 2020.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25. Citeseer, 2015.

Noam Mor, Lior Wolf, Adam Polyak, and Yaniv Taigman. A universal music translation network, 2018.

Marco Pasini. Melgan-vc: Voice conversion and audio style transfer on arbitrarily long samples using spectrograms. *arXiv preprint arXiv:1910.03713*, 2019.

Aaqib Saeed, David Grangier, and Neil Zeghidour. Contrastive learning of general-purpose audio representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3875–3879. IEEE, 2021.

Bob L Sturm. The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use. *arXiv preprint arXiv:1306.1461*, 2013.

John Thickstun, Zaid Harchaoui, and Sham Kakade. Learning features of music from scratch. *arXiv preprint arXiv:1611.09827*, 2016.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *CoRR*, abs/1609.03499, 2016. URL `http://arxiv.org/abs/1609.03499`.

Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, Norman Casagrande, Dominik Grewe, Seb Noury, Sander Dieleman, Erich Elsen, Nal Kalchbrenner, Heiga Zen, Alex Graves, Helen King, Tom Walters, Dan Belov, and Demis Hassabis. Parallel WaveNet: Fast high-fidelity speech synthesis. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3918–3926. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/oord18a.html`.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.