Neural Symbolic Music Genre Transfer Insights

Gino Brunner(✉), Mazda Moayeri, Oliver Richter, Roger Wattenhofer, and Chi Zhang

ETH Zurich, Zurich, Switzerland
{brunnegi,moayerim,richtero,wattenhofer,czhang}@ethz.ch

Abstract. Transferring a song from one genre to another is most difficult if no instrumentation information is provided and genre is only defined by the timing and pitch of the played notes. Inspired by the CycleGAN music genre transfer presented in [2] we investigate whether recent additions to GAN training like spectral normalization and self-attention can improve transfer. Our preliminary results show that spectral normalization improves audible quality, while self-attention hurts content retention due to its non-locality. We further provide insights into genre attribution, showing that often only few notes are genre-decisive.

1 Introduction

What if you could listen to your favourite Beethoven symphony as a Jazz interpretation at the press of a button? Humans are capable of performing such transcription tasks, but it requires considerable skill, effort and creativity. The goal of music genre transfer is to automate this task by training deep neural networks on large amounts of music data. Unsupervised methods excel at this task by allowing us to find structure in complex data in the absence of explicit ground truth labels. Deep generative models have been particularly successful, exemplified by methods such as Variational Autoencoders [5] and Generative Adversarial Networks [3]. One natural application of deep generative models is domain transfer, in which we learn a mapping function between two domains and thus implicitly parts of the underlying data generating distributions. This has led to many impressive applications such as rendering photographs in the style of different painters [13]. However, most applications have focused on images and only recently approaches for other types of data such as music have been proposed. In this work we focus on the task of transferring pieces of music in the MIDI format between different genres, e.g., from classic to jazz. For that purpose we extend the architecture from [2] with recent advances in GANs, in particular spectral normalization [8] and self-attention layers [12], and present respective transfer performance as measured by an automatic classifier-based metric, as well as inherent problems of using self-attention in domain transfer. With this,

P. Cellier and K. Driessens (Eds.): ECML PKDD 2019 Workshops, CCIS 1168, pp. 437–445, 2020.
https://doi.org/10.1007/978-3-030-43887-6_36

we introduce a simple content change metric to quantify content retention in transferred pieces. We further give insights in the decisions made by a neural network based genre classifier using a gradient-based attribution method [10] to better understand genre differences.

2 Related Work

Most neural network based domain transfer approaches are build on either VAEs [5] or GANs [3], or a combination of the two. Liu et al. [7] use a pair of GANs to learn the joint distribution of observations. Their method cannot directly perform domain transfer, but it can generate multiple versions of the same image in different domains (e.g., the same face with different hair color). Liu et al. [6] use a VAE architecture with a shared latent space to perform unsupervised image-to-image translation. Zhu et al. [13] introduce an architecture called CycleGAN which consists of a pair of GANs and is trained to perform domain transfer using a cycle consistency loss. While aforementioned methods are generally applicable, they focus their empirical evaluation on images, where best practices are well established.

In contrast, we focus on domain transfer in music. Mor et al. [9] use an autoencoder based architecture with a shared domain-invariant latent space to transfer input sounds to different instruments. While instruments can be indicative of genre, we focus on the task of genre transfer in absence of any instrumentation information. Brunner et al. [1] force one dimension of the latent space of a VAE to encode the genre by attaching a style classifier. Genre transfer can then be achieved by manipulating this latent genre label. They also propose a classifier-based metric to automatically evaluate the genre transfer. In a follow up work, Brunner et al. [2] adapt the original CycleGAN architecture to perform music genre transfer and achieve good results as measured by a slightly improved classifier-based metric. However, GANs are known to be difficult to train and there are many common failure modes, such as mode collapse or the discriminator overpowering the generator. We therefore investigate the effect of two recent advances in GANs that have been shown to improve GAN performance. In particular, we apply spectral normalization [8] to both the generator and discriminator. We further incorporate self-attention, a recent advance in neural network architectures that has been applied successfully for language modeling [11] and music generation [4]. Self-attention has been incorporated into GANs and together with spectral normalization was shown to improve training stability and overall performance [12]. We investigate both self-attention and spectral normalization in our setup and compare with the genre transfer performance of [2] as measured by a classifier-based metric. We further evaluate their individual impact and show that self attention hinders content retention.

3 Methodology

3.1 Dataset

Our dataset is based on polyphonic multi-instrument MIDI (Musical Instrument Digital Interface) files from three genres: jazz, classic and pop. We use the same dataset and preprocessing steps as [2]. That is, we remove the drum track and merge the remaining instrument tracks into a single piano track, resulting in a two dimensional matrix usually referred to as a *piano roll*, where the one dimension represents time steps and the other represents pitches. Each matrix entry indicates whether that note is played at the corresponding time. To acquire a homogeneous dataset, we omit songs whose time signature is not consistently $\frac{4}{4}$. We choose a sampling rate of 16 time steps per bar and combine 4 consecutive bars into one training example. This means that the shortest possible note we consider is the 16th note. While music in MIDI files can have pitch values between 0 and 127, i.e., note pitches ranging from C_{-1} to C_9, a standard piano can only play pitches between 21 to 108, i.e., notes ranging from A_0 to C_8. Since we merge all tracks into a single-instrument piano track we discard pitches beyond that range. Therefore, each input piano roll matrix has dimensions 64×84, corresponding to $16 * 4$ timesteps and 84 possible pitches respectively.

3.2 Architecture

Our neural network architecture is based on Generative Adversarial Networks (GANs [3]), where a generator and a discriminator are optimized by playing a minimax game. Since we want to perform style transfer in two directions, i.e., from domain A to domain B and vice versa, two GANs are arranged in a CycleGAN architecture [13]. In particular we use as baselines the "full" models from [2] with the additional discriminators.[1] We add two self-attention layers each to the discriminator and generator. For the generator, we add them after the second to last and the last residual blocks. For the discriminator, we add the attention layers after both hidden layers. Spectral normalization is applied to all convolution layers of each discriminator and generator. We use a batch size of 16 and the Adam optimizer with a learning rate of 0.0002. The generators and discriminators are both updated at each step. While training the discriminator, we add Gaussian noise with mean 0 and standard deviation σ_D as this was found to improve genre transfer performance in [2].

3.3 Metrics

As discussed in [1,2], human genre transfer evaluation is time consuming and cannot be applied continuously during development. Thus, a classifier based metric was introduced in [1] and slightly adapted in [2]. The classifier is a 5-layer

[1] See https://github.com/sumuzhao/CycleGAN-Music-Style-Transfer for more details on the baseline architecture.

CNN that performs binary classification between two genres. To evaluate genre transfer, the classifier is applied before and after transfer. For example, when performing transfer from A to B, the original piece should be classified as A, the transferred piece as B, and the transferred-back piece again as A. We report the transfer strength S^D_{tot}, a measure of average difference in correctly classified samples [2]. Specifically, let $P_A(x)$ be the empirical probability of classifying x as genre A. We then calculate the $A \rightarrow B$ transfer strength as

$$S^D_{A \rightarrow B} = \frac{1}{2}(P_A(x_A) + P_A(\tilde{x}_A) - 2 \cdot P_A(\hat{x}_B))$$

where x_A is a sample from domain A, $\hat{x}_B$ is the same sample transferred to domain B and $\tilde{x}_A$ is the sample transferred back to domain A. S^D_{tot} is then calculated as

$$S^D_{tot} = \frac{1}{2}(S^D_{B \rightarrow A} + S^D_{A \rightarrow B})$$

where $S^D_{B \rightarrow A}$ is defined symmetrically to $S^D_{A \rightarrow B}$. For the sake of brevity we refer to [2] for more details.

Further, as genre classification does not capture content retention, we introduce a new *content change metric*. We quantify content change by counting the number of added/removed notes in the piano roll, divided by the number of non-zero entries in the source piano roll. Specifically, for input sample $x \in \{0, 1\}^{64 \times 84}$ we calculate the content change $c(x)$ as

$$c(x) = \frac{\sum_{t,p} |x_{t,p} - \hat{x}_{t,p}|}{\sum_{t,p} x_{t,p}}$$

where $\hat{x}$ is the transferred sample and t and p are the indices into the time and pitch dimension. For a more fine grained analysis we can additionally look at added/removed notes individually:

$$c_{added}(x) = \frac{\sum_{t,p} \max(\hat{x}_{t,p} - x_{t,p}, 0)}{\sum_{t,p} x_{t,p}} \qquad c_{removed}(x) = \frac{\sum_{t,p} \max(x_{t,p} - \hat{x}_{t,p}, 0)}{\sum_{t,p} x_{t,p}}$$

Note that instead of looking at all notes one could also apply a heuristic for melody extraction, e.g., taking the skyline notes, to quantify melody change (as opposed to overall content change). However, we show that the simple metric based on all nodes already correlates well with human ranking of content retention.

3.4 Genre Attribution

We note that genre is ill defined, but the decisions of deep neural networks could provide insights into its nature. We therefore apply a gradient based input attribution method to the trained genre classifier in order to highlight notes that are most important in deciding genre. For instance, a 1-entry in the piano roll matrix corresponds to a played note, and if the back propagated class activation

gradient is high for that note, then removing it would decrease the confidence in the corresponding genre classification, indicating that the presence of the note was significant in determining its genre. We use the saliency map attribution method [10], which multiplies all positive gradients with the original sample element-wise.

4 Experiments and Results

4.1 Genre Transfer

We fix $\sigma_D = 1$ as this worked best in [2] and compare our new models with the corresponding re-trained *full* model from [2], here referred to as *Baseline*. The genre transfer results in Fig. 1 show that self-attention (SA) and spectral normalization (SN) – individually and combined – improve the transfer in two out of the three genre pairings. Moreover, we see that transfer strength mainly depends on the genre pair investigated, as the boundary between some genres is ill defined. Further, the classifier metric does not measure content retention and audible quality, two aspects we are also interested in when performing genre transfer. To preliminarily investigate these aspects we took the classic vs. pop models and asked 12 people of our lab to rank the anonymized and randomly ordered transfers of the 4 models (Baseline, SN, SA, SN + SA) on 8 song snippets (4 transferred from classic to pop and 4 from pop to classic). Each participant thereby ordered for each song the transfers according to (a) content retention and (b) audible quality with respect to the target domain. We aggregated the rankings linearly into a normalized *human ranking score* s_{hr}^M by scoring each model M according to

$$s_{hr}^M = \frac{1}{N} \sum_{r=1}^{K} \#\{\text{rank of } M = \text{ r}\} \frac{K-r}{K-1}$$

where K is the number of models compared (4 in our case) and N is the number of participants. Note that rank one corresponds to the best and rank K to the worst transfer. Figure 1 (right) shows that our content change metric introduced above correlates negatively (Pearson correlation -0.805) with the human content retention ranking, indicating that this is a good heuristic to quantify content retention. Also visible in the figure is that models with self-attention score worse on content retention. This is also reflected in the content change metric over all test samples reported in Table 1, which shows an average content change of 0.92 for the Baseline model, 0.52 for SN, 2.11 for SA and 2.19 for SN + SA.

We therefore suspect that the use of self-attention can actually be harmful, as the generators can encode information in a global manner, as every time step and every pitch level attends to all other time-pitch cells, and hence the generators can alter the content of the source piece more strongly while still being able to achieve cycle-consistency. Explicit regularization techniques to retain parts of the content, e.g., the melody, could be developed in future work. As for audible

Table 1. Results of the content change metric for the different models. Shown is the mean and standard deviation over the test set samples.

	Baseline		SN	
	$C \rightarrow P$	$P \rightarrow C$	$C \rightarrow P$	$P \rightarrow C$
Added	0.82 ± 0.45	0.28 ± 0.17	0.57 ± 0.38	0.08 ± 0.09
Removed	0.27 ± 0.17	0.46 ± 0.12	0.91 ± 0.07	0.3 ± 0.11
Total	1.10 ± 0.5	0.75 ± 0.25	0.66 ± 0.38	0.38 ± 0.16
	SA		SN + SA	
Added	1.47 ± 0.41	0.85 ± 0.79	1.78 ± 1.33	0.72 ± 0.66
Removed	0.95 ± 0.04	0.95 ± 0.04	0.95 ± 0.05	0.93 ± 0.05
Total	2.42 ± 0.42	1.79 ± 0.78	2.73 ± 1.33	1.65 ± 0.66

	J vs. P	C vs. P	J vs. C
Baseline	28.49%	**64.62%**	57.64%
SN	32.16%	61.88%	63.98%
SA	**44.85%**	59.35%	63.56%
SN+SA	33.23%	53.07%	**66.76%**

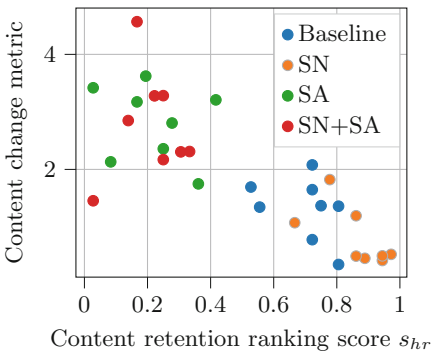

Fig. 1. Top: Genre transfer performance S^D_{tot}. J: Jazz, C: Classic, P: Pop, SN: With spectral normalization, SA: With self-attention. **Right:** Content change metric to human evaluation correlation.

quality, the results of our small user study were less homogeneous. On average, models with spectral normalization where slightly preferred over the others: the Baseline scored 0.47, SN 0.60, SA 0.42 and SN + SA 0.51, where scores are between 1 (always ranked best) and 0 (always ranked worst). Note that the user study only reflects relative audio quality among the studied models, and that there is room for improvement in terms of absolute fidelity. In particular, the genre transfer seems to introduce quite many dissonant notes. However, note that audible quality is already an issue with the original pre-processed pieces, because we reduce music pieces to single-instrument tracks, remove the drums and get rid of some of the dynamics (ignoring velocity, constant tempo). Using a richer input representation as, e.g., done in [1], would already result in more pleasing audio.[2]

[2] Additional results and audio samples can be found here: http://bit.ly/31VnTxS.

4.2 Attribution

Figure 2 depicts source piano rolls from the jazz and classical genres, along with the corresponding attributed piano roll. Intensity-thresholded instance normalized saliency maps [10] are presented. Pixels with intensity less than one-fourth of the maximum were removed in order to reduce clutter around more significantly attributed notes. Attribution was conducted on correctly classified samples with high gradient magnitudes to show interesting examples.

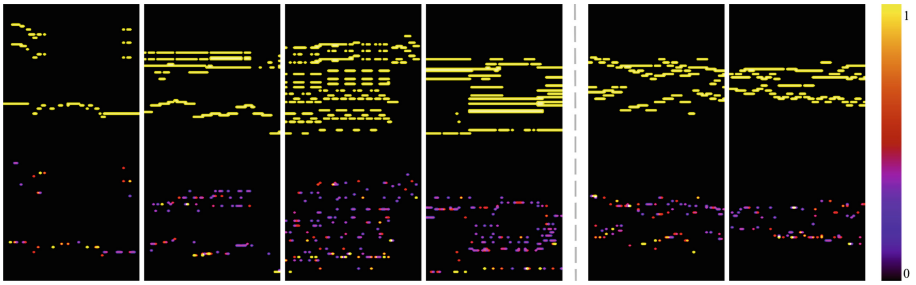

Fig. 2. Piano rolls (top) with corresponding saliency maps (bottom). Jazz samples are to the left of the delimiter, classical to the right.

The saliency maps are dominated by a few hyperintense pixels. Therefore, what distinguishes a sample's genre from the perspective of a deep classifier is truly subtle. We find that in jazz samples, often a sequence of notes in the lower pitch ranges are highlighted. This is somewhat similar to how humans recognize jazz, where genre becomes clear upon hearing a bass play a simple, rhythm-keeping line, over which different melodies are played.

One limitation of gradient-based attribution is that it is only a first order approximation and it is unable to capture complex dependencies across notes. Furthermore, patterns are not always obvious or provable. Nonetheless, the attribution provides a qualitative insight into the decisions of the deep classifier, highlighting certain musical motifs and revealing the nuance of musical genre in its ability to be determined mostly by a small number of notes. Identifying and isolating such motifs would make for fascinating future work in better defining genre and extracting genre specific features.

5 Conclusion

We presented preliminary qualitative insights on automated music genre transfer using MIDI files. We start from the CycleGAN model presented in [2] and show the effect of adding spectral normalization and self-attention on transfer as measured by a classifier-based metric. Further, we find on subsequent inspection that self-attention often makes the transferred songs less recognizable from

a human viewpoint, which is emphasized by our simple content change metric which seems to correlate well with human perception. To the best of our understanding this is due to the global attention mechanism scrambling the pitch/time locality of notes. We further show that genre is often a matter of changing a few notes by looking at the attribution of our genre classifier. Our work offers many directions for follow up work, including the development of a better metrics for genre transfer as well as a quantitative analysis of motifs that make up a genre using attribution on classifiers. To stimulate further research in this direction make our code publicly available.[3]

References

1. Brunner, G., Konrad, A., Wang, Y., Wattenhofer, R.: MIDI-VAE: modeling dynamics and instrumentation of music with applications to style transfer. In: Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, 23–27 September 2018, pp. 747–754 (2018)
2. Brunner, G., Wang, Y., Wattenhofer, R., Zhao, S.: Symbolic music genre transfer with cyclegan. In: IEEE 30th International Conference on Tools with Artificial Intelligence, ICTAI 2018, 5–7 November 2018, Volos, Greece, pp. 786–793 (2018)
3. Goodfellow, I., et al.: Generative adversarial nets. Adv. Neural Inf. Process. Syst. **27**, 2672–2680 (2014)
4. Huang, C.A., et al.: An improved relative self-attention mechanism for transformer with application to music generation. CoRR abs/1809.04281 (2018)
5. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, 14–16 April 2014, Conference Track Proceedings (2014)
6. Liu, M., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4–9 December 2017, Long Beach, CA, USA, pp. 700–708 (2017)
7. Liu, M., Tuzel, O.: Coupled generative adversarial networks. In: Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, 5–10 December 2016, Barcelona, Spain, pp. 469–477 (2016)
8. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April – 3 May 2018, Conference Track Proceedings (2018)
9. Noam Mor, Lior Wold, A.P., Taigman, Y.: A universal music translation network. In: International Conference on Learning Representations (ICLR) (2019)
10. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. In: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, 14–16 April 2014, Workshop Track Proceedings (2014)
11. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4–9 December 2017, Long Beach, CA, USA, pp. 6000–6010 (2017)

[3] https://github.com/czhang0808/Music-Genre-Transfer-with-Deep-Learning.

12. Zhang, H., Goodfellow, I.J., Metaxas, D.N., Odena, A.: Self-attention generative adversarial networks. In: Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA, pp. 7354–7363 (2019)
13. Zhu, J., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017, pp. 2242–2251 (2017)