

Transfer Learning for Underrepresented Music Generation

Anahita Doosti and Matthew Guzdial

Department of Computing Science, Alberta Machine Intelligence Institute (Amii)

University of Alberta

Edmonton, AB, Canada

{doostisa, guzdial}@ualberta.ca

Abstract

This paper investigates a combinational creativity approach to transfer learning to improve the performance of deep neural network-based models for music generation on out-of-distribution (OOD) genres. We identify Iranian folk music as an example of such an OOD genre for MusicVAE, a large generative music model. We find that a combinational creativity transfer learning approach can efficiently adapt MusicVAE to an Iranian folk music dataset, indicating potential for generating underrepresented music genres in the future.

Introduction

Automated music generation has a long history (Briot, Hadjeres, and Pachet 2017). In recent years, large-scale neural network models for music generation have arisen, trained on massive datasets and requiring significant computation (Civit et al. 2022). While these approaches have proven successful at replicating genres of music like those in their training sets, due to the nature of large-scale neural network models we expect this may not prove true for dissimilar genres. Specifically, we hypothesize that these large scale models will perform poorly for out-of-distribution (OOD) genres of music, those representing underrepresented or less globally popular types of music. We therefore conducted a study on one such large-scale neural network model to understand (1) how it performed on OOD music genres, and (2) how we might best adapt the model to an OOD music genre.

For this paper, we focus on Google Magenta’s MusicVAE model (Roberts et al. 2018). We lack the space for a full discussion of the model, but direct interested readers to the original paper. This hierarchical variational autoencoder is trained on an enormous dataset of roughly 1.5 million unique MIDI files collected from the web. While its exact dataset was not made public, online repositories of MIDI files are typically made up of fan-made annotations of popular songs. Thus, automatic and indiscriminate data collection would result in an unbalanced dataset in terms of genre diversity. This is due to the fact that popular chart-topping songs are much more likely to be annotated in the MIDI format. The training requirements for MusicVAE are a problem when it comes to generating underrepresented music, like experimental music or music from particular cultures with distinct musical traditions. These genres of mu-

sic are unlikely to have the massive datasets needed to train models like MusicVAE. Even if such datasets existed, new musical genres are constantly being invented, meaning we could never use this approach to generate all underrepresented genres of music.

If we want to be able to generate underrepresented music, one option outside of training MusicVAE from scratch is transfer learning (Tan et al. 2018). Transfer learning refers to the collection of approaches that can adapt knowledge from a model pre-trained on some source dataset (i.e., popular MIDI files) to a target domain with limited data (i.e., underrepresented music MIDI files). However, these approaches tend to require significant similarity between the source and target domains, which may not hold true for popular and underrepresented music genres (Marchetti et al. 2021). Combinational creativity, also sometimes combinatorial creativity, is a type of creative problem solving in which two conceptual spaces are combined to represent a third or new conceptual space (Boden 2009). While different musical genres may vary in terms of their local features (e.g., melodies), they are all still music. As such, we hypothesized that a combinational creativity-inspired transfer learning approach may be able to outperform traditional transfer learning approaches at the task of adapting MusicVAE to an underrepresented genre of music (Mahajan and Guzdial 2023).

In this paper, we explore the application of CE-MCTS (Mahajan and Guzdial 2023), a combinational creativity-based transfer learning approach to adapt MusicVAE to an underrepresented music genre. While there are many deep neural network (DNN) models like MusicVAE for music generation, applying transfer learning to a DNN model for music generation remains under-explored (Svegliato and Witty 2016; Marchetti et al. 2021). In addition, while combinational-creativity-based transfer learning approaches have been applied to many domains including image classification (Banerjee 2021) and financial health prediction (Mahajan and Guzdial 2023), they have never been applied to the music generation domain. In the remainder of this paper, we first demonstrate an experiment to identify an out-of-distribution (OOD) music genre for MusicVAE. We then introduce CE-MCTS and a number of more standard transfer learning baselines. Finally, we demonstrate their performance in terms of reconstruction accuracy for an OOD

music dataset and present a short discussion on their music generation performance.

Related Work

There have been many recent applications of deep neural networks (DNN) to music generation (Civit et al. 2022). For instance, sequence-based approaches are popular in this field due to their ability to learn long-term dependencies in musical pieces. Multiple studies have combined sequence-based models such as Long Short-term Memory (LSTM) Recurrent Neural Networks (RNN) with autoencoders and achieved good results (Oore et al. 2017). Alternatively, Generative Adversarial Networks (GAN) have been employed to generate novel music (Yang, Chou, and Yang 2017). These models typically are also trained from scratch. However, given the data imbalance across different genres, approaches like transfer learning that can adapt knowledge from one domain to another might be useful. One such example attempts to test a pre-trained Generative Adversarial Networks (BinaryMuseGAN) with traditional Scottish music and improves its performance using finetuning (Marchetti et al. 2021). To the best of our knowledge, finetuning is the only transfer learning approach that has been applied to DNN music generation (Svegliato and Witty 2016). We use it as a baseline.

In regards to Iranian (Persian) traditional or folk music, the prior work focuses on generation of music via traditional non-machine learning approaches and/or training from scratch. In (Arshi 2018), the author uses a combination of evolutionary algorithms, Boltzmann machine models and cellular automata to generate music. They evaluate their work by the use of surveys targeted to both general and professional audiences. Alternatively, researchers have employed RNNs trained on a dataset of traditional Iranian music to generate music (Ebrahimi, Majidi, and Eshghi 2019). We note our goal is not to generate Iranian music specifically, but to explore the best ways to adapt large DNN music generation models to OOD genres like Iranian music.

Genre Analysis

MusicVAE as a music generation model boasts a very impressive performance. Specifically, it achieves 95.1% over its test dataset. However, our hypothesis was that MusicVAE would do poorly for OOD music.

To examine this question, we collected four experimental datasets of 10 songs each. These were small in size as we only required a general approximation of whether the genre was out-of-distribution for MusicVAE. We selected these songs according to two criteria: (1) if they were published after MusicVAE or were otherwise unlikely to be included in the original dataset (Roberts et al. 2018) and (2) if their genres were distinct from a melodic standpoint. Melodies can differ in many ways such as contour, range and scale and these characteristics are different across different genres (DeLone and Wittlich 1975).

Our four datasets are as follows:

- **Synth pop**, songs from a 2021 Netflix special, Inside by Bo Burnham, which musically fall into the synth pop category. This dataset serves as a comparison point, since we expected MusicVAE to perform well on this genre.

Dataset	Accuracy(%)
Synth pop	95.83
Iranian folk	43.75
Video game	84.38
Horror score	87.92

Table 1: MusicVAE accuracy on 4 datasets of different genres

- **Iranian folk**, arose from a region with a long-standing history of composing music with independent roots from western music. As such, we anticipated this would be the most challenging for MusicVAE.
- **Video game**, consists of Nintendo Entertainment System video game music. These songs have limited polyphony as only 3 notes can be played on the NES at once.
- **Horror scores**, were designed to build suspense and create a sense of foreboding. Musically this genre frequently uses dissonant notes or chords, atonality (not having a clear scale), sudden changes of tempo, and other effects to induce a sense of eeriness and dread.

In these experiments, we fed melody sequences extracted from the songs into the pre-trained MusicVAE and report the reconstruction accuracy. We include the results of our analysis in Table 1. As we expected, the model performed best on the first dataset. The 95.83% accuracy is in line with what was reported for the test accuracy on the original MusicVAE dataset. Predictably, the accuracy is noticeably lower for the other three datasets, with the Iranian dataset standing at a mere 43%, a major drop in performance compared to the rest. As such, we focused on Iranian folk music for the remainder of our study.

Iranian Folk Music Dataset

We gathered a new dataset of Iranian folk MIDI files in order to evaluate the possibility of adapting MusicVAE to this out-of-distribution (OOD) genre. This dataset consists of 100 MIDI files from both Farsi-speaking websites and from musescore.com which is a free sheet music sharing website. These files contain different instruments and varying levels of polyphony. We collected 100 MIDI files as this is in line with the target genre dataset size for prior finetuning-based transfer learning approaches with music generation models (Svegliato and Witty 2016; Marchetti et al. 2021). However, we anticipate that CE-MCTS could perform well with fewer samples (Mahajan and Guzdial 2023). During our experiments we used a five-fold cross validation, meaning we split the data into five train-test splits (80 songs for training, 20 for testing), which helped ensure that we did not just get a “lucky” train-test split.

Conceptual Expansion Monte Carlo Tree Search (CE-MCTS)

We hypothesized that a combinational creativity-based transfer learning approach could most effectively adapt MusicVAE to an OOD genre. While there have been sev-

eral prior examples of combinational creativity-based transfer learning approaches (Banerjee 2021; Singamsetti, Mahajan, and Guzdial 2021), we chose Conceptual Expansion Monte Carlo Tree Search (CE-MCTS) (Mahajan and Guzdial 2023). None of these combinational creativity-based transfer learning approaches have been applied to music data, but CE-MCTS demonstrated the ability to adapt to the behaviours of distinct groups of humans across problem domains. As such, we anticipated it would be the best for adapting to other types of human expression, like music.

Here we briefly describe CE-MCTS and how we adapted it in this work. However, for a full description of the approach we direct interested readers to (Mahajan and Guzdial 2023). The “Conceptual Expansion” (CE) in the name refers to the fact that we are searching over combinations (Banerjee 2021). In this case, different combinations of the learned features from the original MusicVAE model. While it may seem unintuitive to combine features from the same model, this can allow us to approximate unseen features. This is equivalent to combining different musical patterns in the source domain (i.e., popular MIDI music) to approximate patterns in the target domain (i.e., Iranian folk music). We then employ Monte Carlo Tree Search (MCTS) to search over the space of these possible combinations. As is typical with MCTS, we build up a tree to search through this space. The root node represents the original trained MusicVAE model (no combinations) and every subsequent node represents a different set of feature combinations, with closer nodes representing more similar combinations.

For our implementation of CE-MCTS we largely employed the same setup from the original paper (Mahajan and Guzdial 2023). However, we made a number of changes for this domain. For the fitness, we used the reconstruction accuracy of the training split (80 songs from Iranian folk music dataset). We ran 10 iterations, each with 10 rollouts ($L = 5$) and $\epsilon = 0.5$. We did this to bias the search towards exploration near the original MusicVAE model as we did not want to risk catastrophic forgetting, in which a model loses useful features. Ultimately, we output the three best performing models according to the fitness and report their average performance over the test split (20 songs). The strategy for the final model selection varies based by domain.

Transfer Learning Baselines

We hypothesized that CE-MCTS would outperform standard transfer learning approaches at adapting MusicVAE to our Iranian folk music dataset. Here we introduce the transfer learning baselines we used for comparison purposes. Other transfer learning approaches were not appropriate as we lacked access to MusicVAE’s original training dataset.

- **Non-transfer**, where we train a randomly initialized MusicVAE on the Iranian music dataset alone. This represents the standard approach to this problem without transfer learning. We did not expect this to work given the limited amount of training data.
- **Zero-shot**, which uses the pre-trained weights of MusicVAE with no additional training on the Iranian music

dataset. We know MusicVAE does poorly when reconstructing 10 Iranian folk songs, but this won’t necessarily hold for our larger 100 song dataset.

- **Finetuning (all)**, in which we use finetuning, a traditional network-based transfer learning approach that has been applied in prior music generation DNN transfer learning work (Tan et al. 2018; Svegliato and Witty 2016; Marchetti et al. 2021). In this baseline, we applied finetuning by continuing to train MusicVAE on a train split of our dataset until convergence. The (all) indicates retraining all of MusicVAE’s layers. This is unusual as it can lead to catastrophic forgetting, where useful features are lost in the adaptation process.
- **Finetuning (last)**, which is the same as Finetuning (all) except that we freeze the weights of all but the last layer. This is the more typical approach when applying finetuning as it assumes that the earlier layers contain useful features (e.g., musical patterns) and we can just adapt the last layer to apply these patterns more appropriately for our target dataset.

We also developed a knowledge distillation approach called student-teacher learning (Romero et al. 2015). In this approach we trained a MusicVAE (student network) on Iranian music, through a combination of its loss and the loss of another MusicVAE with pre-trained weights (teacher network). This model proved worse than all other baselines, thus we do not include its results.

Evaluation

For our evaluation, we compared the reconstruction accuracy of our approach and baselines. Each model was fed our test set and we measured the percentage of correctly reconstructed notes. Clearly, this method of evaluation does not assess the ability of the model to actually generate new musical sequences, which is the main objective of a music generation model. However, this approach allows us to run an initial quantitative evaluation. While there are objective metrics employed by other researchers, they are not consistently defined, making it difficult to compare outputs across different generation systems. Furthermore, there is no correlation between qualitative and quantitative metrics of evaluation, making it difficult to draw implications. We expect this to be even more difficult for a genre like Iranian music or other regional genres. Therefore, correctly evaluating the quality of generation would require a human subject study with experts in the target genre (i.e. Iranian folk music). (Ji, Luo, and Yang 2020) We leave this to future work, but include a small case study below.

Results

Tables 2 and 3 contain our training and test reconstruction accuracy results, respectively. Overall we can observe that CE-MCTS consistently outperforms other methods in both training and test accuracy. Although Finetuning (last) performs similarly on average during training, CE-MCTS is better at reconstructing the test data.

Approach	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Non-transfer	68.75	68.75	68.75	68.75	68.75	68.75
Zero-shot	93.75	90.62	84.37	87.50	87.50	88.75
Finetune (all)	87.50	84.37	78.12	78.12	75.00	80.62
Finetune (last)	96.87	90.62	93.75	100	100	96.25
CE-MCTS	98.96	94.80	98.97	94.84	100	97.52

Table 2: Training reconstruction accuracy of each approach on each fold

Approach	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Non-transfer	37.50	37.00	37.50	6.25	53.12	34.27
Zero-shot	90.62	87.50	75.00	37.50	90.62	76.24
Finetune (all)	81.25	68.75	25.00	34.37	65.62	55.00
Finetune (last)	96.87	96.87	65.62	40.62	93.75	78.75
CE-MCTS	93.75	97.9	83.34	51.07	93.77	83.97

Table 3: Test reconstruction accuracy of each approach on each fold

As we expected, Finetuning (all) produces inferior results to Finetuning (last). In fact, it seems that the original pre-trained MusicVAE (Zero-shot) outperforms this method. This is likely due to catastrophic forgetting. As for the non-transfer method, it is not surprising that the small quantity of data available is unable to effectively train the network. The initial dataset size used to train MusicVAE is roughly 15,000 times larger than our dataset.

CE-MCTS outperforms both the pre-trained MusicVAE and last layer finetuning on test accuracy. Therefore we can deduce that by recombining the features in the earlier layers, CE-MCTS is able to create better features for the target dataset. Based on the performance of the Zero-shot and Finetuning (last) baselines, we can infer that the features present in original model are not sufficient to represent Iranian folk music. This suggests we can usefully approximate Iranian folk music features via a combination of features from popular western music.

Case Study

In this section, we provide a brief qualitative analysis of some of the models in terms of music generation instead of reconstruction. We include figures with representative outputs from the three best performing approaches. These examples were chosen by the first author, who has expertise in Iranian music, to be generally representative of the characteristics of the outputs from these approaches. This is obviously highly subjective and susceptible to confirmation bias. A study with expert participants is needed to make reliable assertions about the quality of generation, but we leave this to future work.

In each corresponding figure, the x-axis represents time in seconds, limited to 4 seconds which is the length of all 2-bar outputs by MusicVAE. The y-axis represent the pitch for the notes in the MIDI format which ranges between 0 and 127. Each red rectangle in the figure represents a continuous note.

Figure 1 represents a typical melody generated by the Zero-shot pre-trained MusicVAE. The notes in this melody sound harmonious and follow a somewhat cohesive progression. They also gradually move from a higher pitch to a lower one, spanning somewhat evenly across the melodic

range (distance between the highest and the lowest pitch). Figures 2 and 3, were generated using the Finetuning (last) and CE-MCTS approaches, respectively. In the first author’s subjective opinion, samples generated by these two models sound more similar to, and evocative of, the type of melodies present in Iranian folk music. This is hard to qualify but here we point out a number of characteristics commonly seen in traditional and folk Iranian (Persian) music according to (Farhat 2004) and (Iran Chamber Society 2001).

- Melodies have a narrow register (pitch range).
- Melodic movement is often achieved with conjunct steps.
- There is an emphasis on cadence, symmetry, and repetition of musical motifs at varying pitches.
- Rhythmic patterns are generally kept uncomplicated and rhythmic changes are infrequent.
- The tempo is often fast, with dense ornamentation. Similar to this, it is common to see repetitive and rapid use of the same note/pitch.

As shown in both figures, the register is more limited locally and patterns that repeat the same note appear.

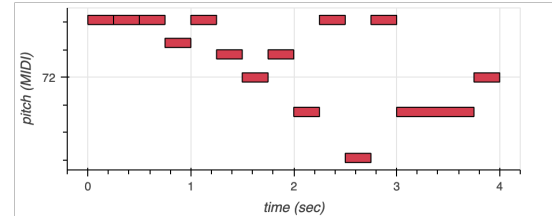


Figure 1: Visualization of a melody generated by the pre-trained MusicVAE model

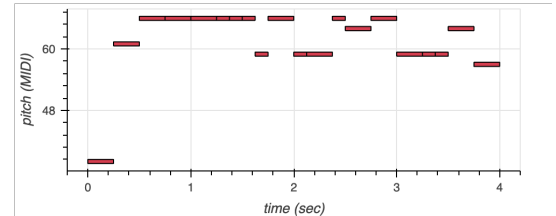


Figure 2: Visualization of a melody generated by the model finetuning the last layer

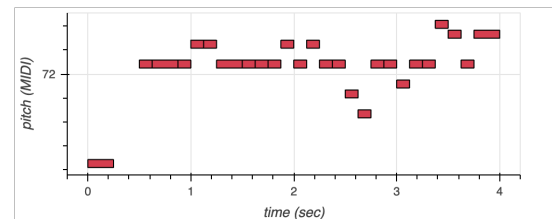


Figure 3: Visualization of a melody generated by the CE-MCTS model

Conclusions

In this paper, we investigated how MusicVAE, a music generation model, can be adapted to OOD music. We identified that MusicVAE in particular struggles with Iranian folk music. We then explored different transfer learning methods in order to improve MusicVAE’s performance on a newly collected Iranian folk music dataset. Based on our results, we observed that CE-MCTS, a combinational creativity-based transfer learning approach, is better able to produce reconstructions of this genre of music. This suggests that we can successfully adapt these large music generation models for underrepresented genres of music, and that combinational creativity can be an especially helpful tool in this task.

References

- Arshi, S. 2018. *Creating Persian-like music using computational intelligence*. Ph.D. Dissertation. HydraID: hull:17739 Hydra Discover Access Group: public ETD Collection: ETDComputerScience.
- Banerjee, A. 2021. Combinets v2: improving conceptual expansion using sgd. In *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*, 413–413.
- Boden, M. A. 2009. Computer models of creativity. *AI Magazine* 30(3):23–23.
- Briot, J.-P.; Hadjeres, G.; and Pachet, F.-D. 2017. Deep learning techniques for music generation—a survey. *arXiv preprint arXiv:1709.01620*.
- Civit, M.; Civit-Masot, J.; Cuadrado, F.; and Escalona, M. J. 2022. A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends. *Expert Systems with Applications* 209:118190.
- DeLone, R., and Wittlich, G. E. 1975. *Melody: Linear Aspects of Twentieth-Century Music*. Prentice-Hall. 270–301.
- Ebrahimi, M.; Majidi, B.; and Eshghi, M. 2019. Procedural composition of traditional persian music using deep neural networks. In *2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI)*, 521–525.
- Farhat, H. 2004. *The dastgah concept in Persian music*. Cambridge University Press.
- Iran Chamber Society. 2001. Iranian classical music. www.iranchamber.com/music/articles/iranian_classical_music.php. Accessed: April 27, 2023.
- Ji, S.; Luo, J.; and Yang, X. 2020. A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions. *arXiv preprint arXiv:2011.06801*.
- Mahajan, A., and Guzdial, M. 2023. *Modeling Individual Humans via a Secondary Task Transfer Learning Method*. Cham: Springer International Publishing. 259–281.
- Marchetti, F.; Wilson, C.; Powell, C.; Minisci, E.; and Riccardi, A. 2021. Convolutional generative adversarial network, via transfer learning, for traditional scottish music generation. In *Artificial Intelligence in Music, Sound, Art and Design: 10th International Conference, EvoMUSART 2021, Held as Part of EvoStar 2021, Virtual Event, April 7–9, 2021, Proceedings 10*, 187–202. Springer.
- Oore, S.; Simon, I.; Dieleman, S.; and Eck, D. 2017. Learning to create piano performances. In *NIPS 2017 Workshop on Machine Learning for Creativity and Design*.
- Roberts, A.; Engel, J.; Raffel, C.; Hawthorne, C.; and Eck, D. 2018. A hierarchical latent vector model for learning long-term structure in music. In *International conference on machine learning*, 4364–4373. PMLR.
- Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2015. Fitnets: Hints for thin deep nets.
- Singamsetti, M.; Mahajan, A.; and Guzdial, M. 2021. Conceptual expansion neural architecture search (cenas). *International Conference on Computational Creativity (ICCC)*.
- Svegliato, J., and Witty, S. 2016. Deep jammer: a music generation model. *Small* 6:67.
- Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; and Liu, C. 2018. A survey on deep transfer learning. In *International conference on artificial neural networks*, 270–279. Springer.
- Yang, L.-C.; Chou, S.-Y.; and Yang, Y.-H. 2017. Midinet: A convolutional generative adversarial network for symbolic-domain music generation. *arXiv preprint arXiv:1703.10847*.