Symbolic Music Generation with Transformer-GANs

Aashiq Muhamed[*], **Liang Li**[*], **Xingjian Shi, Rahul Suresh, Alexander J. Smola**
Amazon Web Services
muhaaash, mzliang, xjshi, surerahu@amazon.com, alex@smola.org

Abstract

Transformers have emerged as the dominant approach in music literature for generating minute-long compositions with compelling musical structure. These models are trained by minimizing the negative log-likelihood (NLL) of the observed sequence autoregressively. Unfortunately, the quality of samples from these models tends to degrade significantly for long sequences, a phenomenon attributed to exposure bias. Fortunately, we are able to detect these failures with classifiers trained to distinguish between real and sampled sequences. This motivates our Transformer-GAN framework that trains an additional discriminator to complement the NLL objective. We use a pre-trained SpanBERT model for the discriminator, which in our experiments helped with training stability. Using human evaluations and other objective metrics we demonstrate that music generated by our approach outperforms a baseline trained with likelihood maximization and the state-of-the-art Music Transformer.

1 Introduction

Recent advancements in Natural language Processing (NLP), especially the attention mechanism and the Transformer architecture (Vaswani et al., 2017), have helped push the state of the art in symbolic music generation (Huang et al., 2018; Payne, 2019; Donahue et al., 2019). These approaches represent a piece of music as a sequence of time-ordered events. A Transformer-based language model is then trained on top of the event sequences by maximizing likelihood. Music can be generated by sampling from this language model. Huang et al. (2018) used relative positional encoding (Shaw et al., 2018) within the original Transformer architecture to capture relative timing information. Payne (2019) used sparse kernels (Child et al., 2019) to remember the long term structure in the composition. More recent works in music generation (Donahue et al., 2019; Huang and Yang, 2020; Wu et al., 2020) adopt the Transformer-XL architecture (Dai et al., 2019) that uses recurrent memory to enable the model to attend beyond a fixed context.

Despite recent improvements, these approaches exhibit crucial failure modes which we argue arise from the training objective; Music Transformer (Huang et al., 2018) occasionally forgets to switch off notes and loses coherence beyond a few target lengths as stated by the authors. Sometimes it produces highly repetitive songs, sections that are almost empty, and discordant jumps between contrasting phrases and motifs. Consequently, music generated by such models can be distinguished from real music by a simple classifier. This suggests that a distribution distance, such as the adversarial objective of a GAN (Goodfellow et al., 2014) should improve the fidelity of the generative model.

Unfortunately, incorporating GAN losses for discrete sequences can be difficult as differentiating through the discrete sampling process is challenging. As such, many models (de Masson d'Autume et al., 2019; Nie et al., 2019) are limited to 20-40 token-length sentences, in contrast to the more than 1000 tokens required for minutes-long musical compositions. We leverage the Gumbel-Softmax (Kusner and Hernández-Lobato, 2016) trick to obtain a differentiable approximation of the sampling process. To address memory requirements, we use the Truncated Backpropagation Through Time (TBPTT) (Sutskever et al., 2014) for gradient propagation on long sequences.

Recent works on evaluation metrics in text generation (Salazar et al., 2019; Zhang et al., 2019; Montahaei et al., 2019) suggest that BERT-based

[*] Equal contribution

scores (Devlin et al., 2018) are well correlated with human rankings and jointly measure quality and diversity. As BERT is trained using a self-supervised loss on bidirectional contexts of all attention layers, it can be an effective way of extracting representations. Inspired by this idea, we propose using pretrained BERT as discriminator to send feedback to the generator.

Works that have explored GANs for sequence generation include Nie et al. (2019), where the authors use a relational memory based generator, the Gumbel-Softmax trick, and multiple embedded representations in the CNN discriminator— and Zhang (2020), where the author proposed an adversarial framework that used Transformers as both generator and discriminator, trained with a global and local loss. Our work differs in the following ways: (1) We use the Transformer-XL as our generator and pretrained BERT as the discriminator; (2) we pretrain the BERT discriminator in SpanBERT style (Joshi et al., 2020); (3) we design an algorithm using the Gumbel-Softmax trick and a variant of the TBPTT algorithm to train on long sequences.

In summary, our main contributions include: (a) A novel Transformer-GAN approach for generating long music sequences of over 1000 tokens, using a pretrained SpanBERT as the discriminator; (b) the investigation of the influence of pretraining, loss functions, regularization, and number of frozen layers in the discriminator on music quality; (c) A number of critical tricks for adversarial training.

2 Transformer-GAN

2.1 Modeling choice

Given a music sequence $\mathbf{x} = [x_1, x_2, ..., x_n]$, we choose to model the unknown data distribution $p_{data}(\mathbf{x})$ autoregressively, i.e., $p_\theta(\mathbf{x}) = \prod_{t=1}^{n} p_\theta(x_t|x_1, ..., x_{t-1})$. The most common approach to train such a model is Maximum Likelihood that finds a θ to minimize,

$$L_{\text{mle}} = \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})} - \log(p_\theta(\mathbf{x})). \quad (1)$$

Despite its attractive *theoretical* properties, Maximum Likelihood training suffers from many limitations, e.g. whenever the model is misspecified. This issue is illustrated by (Isola et al., 2017) in image-to-image translation, where no explicit loss function is available. Furthermore, teacher forcing introduces exposure bias (Bengio et al., 2015; Holtzman et al., 2019)—a distributional shift between training sequences used for learning and model data required for generation. This amplifies any errors in the estimate, sometimes creating degenerate, repetitive outputs. We address this problem by incorporating an adversarial loss into our objective. Denoting our model as *generator* p_θ, we ensure that the sequences obtained from it match those on the training set as assessed by an appropriate *discriminator* D_ϕ. We also regularize D_ϕ to prevent overfitting.

$$L_{\text{G}} = L_{\text{mle}}[p_\theta] + \lambda L_{\text{gen}}[p_\theta] \quad (2)$$

$$L_{\text{D}} = L_{\text{disc}}[D_\phi] + \gamma L_{\text{reg}}[D_\phi] \quad (3)$$

Here $\lambda, \gamma > 0$ are hyperparameters. The loss functions $L_{\text{gen}}[.]$, $L_{\text{disc}}[.]$ and $L_{\text{reg}}[.]$ are determined by the specific GAN loss, such as the Relativistic GAN (RSGAN) (Jolicoeur-Martineau, 2018) or Wasserstein GAN with gradient penalty (WGAN-GPen) (Gulrajani et al., 2017).

Learning $p_\theta(\mathbf{x})$ using (2) and (3) requires backpropagating through a sampling operation that is non differentiable. We choose the low variance, biased gradient estimator using the Gumbel-Softmax trick. We use the exponential inverse-temperature decay policy proposed in (Nie et al., 2019) for balancing exploration and exploitation during generator sampling.

2.2 Selecting the discriminator

Training GANs using the Transformer as generator (Chen et al., 2020; Zhang, 2020) is a difficult problem since training dynamics, memory overhead and generator and discriminator losses need to be carefully balanced. In prior work, CNNs (Kim, 2014) have proven to be useful discriminators for text generation (Nie et al., 2019). In this work, we propose to use BERT as discriminator to extract sequence embeddings followed by a pooling and linear layer. The bidirectional transformer is comparable in capacity to the transformer-based generator and uses the self-attention mechanism that captures meaningful aspects of the input music sequence. We speculate that this would help the discriminator provide informative gradients to the generator and stabilize the training process. Our findings on the importance of discriminator regularization align with prior work (Gulrajani et al., 2017) on image GANs. We found that discriminator regularization in the form of layer normalization, dropout and L2 weight decay provided a less significant performance boost than using gradient penalty.

Like Mo et al. (2020) observed for images, we conjecture that freezing earlier layers (closer to generator) of the pretrained discriminator can be viewed as transfer learning, where we transfer music representations useful for generation from a different dataset. Unlike Mo et al. (2020), where the discriminator is transferred between trained GANs on different datasets, we pretrain our discriminator on the same dataset using a self-supervised loss.

SpanBERT style self-supervised pretraining, where we predict spans of masked tokens, enables the model to learn span representations. We hypothesize that span representations are better inductive biases for modeling coherence in music, as music is composed in spans of notes or chords. Given a masked span $x_{\text{mask}} := (x_s, ..., x_e) \in \mathbf{x}$, the SpanBERT Masked Language Model (MLM) objective for each token $x_i \in x_{\text{mask}}$ is

$$L_{MLM}(x_i) = -\log(P(x_i|x_{\backslash\text{mask}})), \quad (4)$$

where $x_{\backslash\text{mask}} = \{y \mid y \in \mathbf{x} \text{ and } y \notin x_{\text{mask}}\}$. Freezing the discriminator also reduces the number of trainable parameters and training memory requirements.

2.3 Adversarial training

We train our Transformer-GAN by alternating updates between the generator (2) and discriminator (3) objectives. Like training other language GANs (Lu et al., 2018), we found that pretraining our generator with Maximum Likelihood before beginning adversarial training stabilized training dynamics and reduced the variance of gradient estimates. Since training involves a sequential sampling step, this can quickly become a compute and memory bottleneck on long sequences. We address the memory bottleneck by feeding our generated sequence into SpanBERT in segments and calling backward on each segment (TBPTT). The truncation improves memory efficiency as it avoids holding all forward computation graphs during sampling. The length of the subsequence is also well suited to BERT since it is trained to accept a smaller fixed length sequence. Additionally, we observed that generating samples conditioned on the same context as the corresponding real samples helped reduce variance in gradient estimation.

3 Experiments and Results

We benchmarked our models on the MAESTRO MIDI v1 dataset (Hawthorne et al., 2019). We used the same data augmentation as Music Transformer, where we augmented the data with uniform pitch transposition from {-3, -2, ..., 2, 3} and time stretch from {0.95, 0.975, 1.0, 1.025, 1.05}. For the Music Transformer baseline, we used their tensor2tensor implementation (Vaswani et al., 2018).

Hyperparameters	Music Transformer	Transformer-XL
Layers	8	6
Dropout	0.2	0.1
Hidden size	384	500
Target length	2048	128
Train Memory length	-	1024
Eval Memory length	-	2048
Number of heads	8	10
Number of parameters	16253189	13677310
Learning rate	0.2	0.004
Scheduler	inv-sqrt	inv-sqrt
Optimizer	Adam	Adam

Table 1: Hyperparameters

3.1 Implementation Details

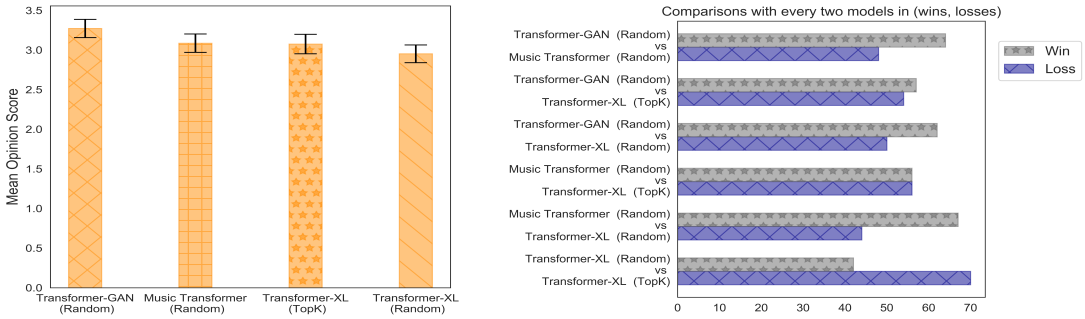
We serialized each MIDI into a sequence of tokens using the event-based representation in Oore et al. (2018). Table 1 lists a few critical hyperparameters for the baseline Transformer-XL. When training the GAN, we cycled between training the generator five times with the NLL and one time with the GAN loss. We updated the generator using a smaller learning rate (0.002) and set the sequence length of generated samples to 128 (equal to target length in Table 1). We used a maximal inverse temperature of 100. Additionally, we cleared the Transformer memory at the start of a MIDI. To generate musical samples, we set memory length equal to the number of tokens to generate, and we implemented Random sampling and TopK sampling (Holtzman et al., 2019).

3.2 Human Evaluations

The participants were presented with samples from Transformer-GAN (Random), Transformer-XL (Random), Music Transformer (Random) and Transformer-XL (TopK). Each sample is a 30 second extension of a 10 second prime. A survey comprises 7 questions, with each question comparing the 4 models on the same priming melody. We use 6 sets of surveys evenly distributed among people. The participants were then asked in each question (a) to rate the sample from 0 (low quality) to 5 (high quality) (b) to rank the samples based on their coherence and consistency. In total, we received 448 ratings for (a) and 672 pairwise comparisons for (b). The results are shown in Fig. 1.

	NLL ↓	PCU∼	NPU∼	EBR∼	ISR∼	PRS∼	TUP∼	PR∼	APS∼	IOI∼	CA↓	PLL ∼
Training Set	-	7.810	65.848	0.985	0.586	0.399	65.28	67.34	11.531	0.1334	-	2.0203
Music Transformer	1.79	7.230	**54.950**	0.999	0.589	0.570	55.23	61.74	11.340	0.1156	0.8443	2.5666
Transformer-XL	**1.74**	7.020	52.190	0.983	0.567	0.265	52.95	60.39	11.117	0.1072	0.8377	2.1531
GAN (RSGAN)	1.75	7.200	53.000	0.984	0.592	0.330	53.68	62.55	11.037	0.1179	0.8615	2.1084
GAN (RSGAN-GPen)	**1.74**	7.300	53.970	0.975	**0.585**	0.334	54.78	63.17	**11.647**	**0.1387**	0.8307	2.2766
GAN (WGAN-GPen)	1.75	**7.330**	54.830	**0.987**	0.611	**0.405**	**55.56**	**63.23**	11.935	0.145	**0.8179**	**2.1020**

Table 2: Quantitative music metrics: NLL (Negative likelihood); PCU (No. of unique pitch classes in a piano roll); NPU (No. of unique pitches); EBR (Ratio of empty beats to the total no. of beats); ISR (Ratio of the no. of nonzero entries that lie in a specific scale to the total no. of nonzero entries in a piano roll); PRS (Ratio of the no. of time steps where the no. of pitches being played is larger than 3 to the total no. of time steps); TUP (No. of different pitches within a sample); PR (Avg. difference of the highest and lowest pitch in semitones); APS (Avg. semitone interval between two consecutive pitches); IOI (Inter-onset-interval or the time between two consecutive notes); CA (SpanBERT classifier accuracy distinguishing real and generated data); PLL (Pseudo-log-likelihood score). Bolded values are better. Metrics marked with ∼ are better when closer to the dataset.



(a) (b)

Figure 1: Human Evaluation: (a) the average rating for each model. (b) the pairwise comparisons. For 'A vs B', 'Win' stands for the counts of A beating B and 'Loss' stands for the counts of B beating A. (Best viewed in color).

3.3 Results

We run our Transformer-GAN with three GAN losses, i.e. RSGAN, WGAN-GPen, and RSGAN-GPen, and we compare them with the Transformer-XL baseline and Music Transformer using human evaluation and several quantitative metrics. The results in Table 2 led to the following findings. (i) We achieve comparable results overall between Music Transformer (Random) and our Transformer-XL (TopK). (ii) Transformer-GAN with WGAN-GPen performs especially better at polyphonic generation and is seen to be closest to the training dataset on multiple quantitative metrics (Dong et al., 2018; Yang and Lerch, 2020). (iii) A Kruskal Wallis test on the ratings yields $\mathcal{X}^2(2) = 3.272, p = 0.031$, and Transformer-GAN (Random) outperforms the Transformer-XL (Random) with 62 wins and 50 losses. (iv) Transformer-GAN (Random) outperforms all other models sampled with either Random or TopK sampling.

3.4 Ablation Studies

We study the effect of freezing layers in our pretrained discriminator. In Table 3, the first row corresponds to the randomly initialized BERT without any pretraining. We observe that (i) The randomly initialized BERT discriminator does not perform as well as the discriminator with 5 frozen layers (ii) Freezing more layers in the pretrained discriminator tends to improve CA.

Model	Frozen layers(BERT)	NLL ↓	CA ↓
GAN (WGAN-Gpen)	Random-init	1.74	0.8359
GAN (WGAN-Gpen)	['emb', '0']	1.74	0.8852
GAN (WGAN-Gpen)	['emb', '0', '1', '2']	1.74	0.8586
GAN (WGAN-Gpen)	['emb', '0', '1', '2','3']	1.75	0.8394
GAN (WGAN-Gpen)	['emb', '0', '1', '2','3','4']	1.75	**0.8179**

Table 3: Ablation Studies: BERT discriminator has 6 layers denoted ['emb','0','1','2','3','4']

4 Conclusion

The results obtained from various experiments demonstrate that our Transformer-GAN achieves better performance compared to other transformers trained by maximizing likelihood alone. By sampling during training, the adversarial loss helps bridge the discrepancy between the training objective and generation. In future work, we plan to extend our work by pretraining on larger datasets where our idea can be beneficial.

References

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179.

Xingyuan Chen, Ping Cai, Peng Jin, Hongjun Wang, Xinyu Dai, and Jiajun Chen. 2020. A discriminator improves unconditional text generation without updating the generator. *arXiv preprint arXiv:2004.02135*.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Chris Donahue, Huanru Henry Mao, Yiting Ethan Li, Garrison W Cottrell, and Julian McAuley. 2019. Lakhnes: Improving multi-instrumental music generation with cross-domain pre-training. *arXiv preprint arXiv:1907.04868*.

Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. 2018. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777.

Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. 2019. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *International Conference on Learning Representations*.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. 2018. Music transformer. *arXiv preprint arXiv:1809.04281*.

Yu-Siang Huang and Yi-Hsuan Yang. 2020. Pop music transformer: Generating music with rhythm and harmony. *arXiv preprint arXiv:2002.00212*.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Alexia Jolicoeur-Martineau. 2018. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Matt J Kusner and José Miguel Hernández-Lobato. 2016. GANs for sequences of discrete elements with the gumbel-softmax distribution. *arXiv preprint arXiv:1611.04051*.

Sidi Lu, Yaoming Zhu, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Neural text generation: Past, present and beyond. *arXiv preprint arXiv:1803.07133*.

Cyprien de Masson d'Autume, Shakir Mohamed, Mihaela Rosca, and Jack Rae. 2019. Training language GANs from scratch. In *Advances in Neural Information Processing Systems*, pages 4300–4311.

Sangwoo Mo, Minsu Cho, and Jinwoo Shin. 2020. Freeze discriminator: A simple baseline for fine-tuning gans. *arXiv preprint arXiv:2002.10964*.

Ehsan Montahaei, Danial Alihosseini, and Mahdieh Soleymani Baghshah. 2019. Jointly measuring diversity and quality in text generation models. *arXiv preprint arXiv:1904.03971*.

Weili Nie, Nina Narodytska, and Ankit Patel. 2019. RelGAN: Relational generative adversarial networks for text generation. In *ICLR*.

Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck, and Karen Simonyan. 2018. This time with feeling: Learning expressive musical performance. *Neural Computing and Applications*, pages 1–13.

Christine Payne. 2019. MuseNet.

Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2019. Masked language model scoring. *arXiv preprint arXiv:1910.14659*.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems (NeurIPS)*.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for neural machine translation. *CoRR*, abs/1803.07416.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Xianchao Wu, Chengyuan Wang, and Qinying Lei. 2020. Transformer-XL based music generation with multiple sequences of time-valued notes. *arXiv preprint arXiv:2007.07244*.

Li-Chia Yang and Alexander Lerch. 2020. On the evaluation of generative models in music. *Neural Computing and Applications*, 32(9):4773–4784.

Ning Zhang. 2020. Learning adversarial transformer for symbolic music generation. *IEEE Transactions on Neural Networks and Learning Systems*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.