SCHMUBERT: A SYMBOLIC CREATIVE HARMONIC MUSIC UNMASKING BIDIRECTIONAL ENCODER REPRESENTATION TRANSFORMER

Matthias Plasser JKU Linz Computational Perception **Silvan Peter**

ABSTRACT

Denoising Diffusion Probabilistic Models (DDPMs) have shown great success generating high quality samples in both discrete and continuous domains [1–3]. However, Discrete Denoising Diffusion Probabilistic Models (D3PMs) have not yet been shown to be directly applicable to the domain of Symbolic Music. In this work we present the direct generation of Polyphonic Symbolic Music using D3PMs. Our model does not only exhibit state of the art sample quality, but also allows for various conditioning methods at sample time without extra training. As the model is trained to reconstruct randomly masked out tokens, conditioning on an existing piece of symbolic music is possible. Such conditioning scenarios include, but are not limited to, accompaniment (one track is provided, accompaniment tracks are masked out) and infilling/completion (one or multiple tracks with temporal gaps are provided). We provide our implementation, trained model weights and some selected samples at `https://github.com/plassma/symbolic-music-discrete-diffusion`.

1. INTRODUCTION

DDPMs [2,4] are a relatively new class of generative models, which outperform previous state of the art generative models in various generation tasks since they were proposed in 2015 [1–3]. DDPMs are inspired by Langevin-dynamics; they are only naturally defined on continuous domains [5]. Austin et al. [5] extended DDPMs to discrete domains and Taylor et al. [2] demonstrated the capabilites of D3PMs on discrete sequences. Mittal et al. [3] successfully applied continuous DDPMs to the domain of symbolic music, albeit indirectly by encoding the sequence of discrete tokens into a continuous latent space before applying a continuous DDPM.

In this work we propose a model that is applied directly to the domain of Symbolic Music.

2. MODEL AND REPRESENTATION

We use an Absorbing State Discrete Denoising Diffusion Probabilistic Model [5] similar to the model Taylor et al. use in [2]. DDPMs learn to create data by training a denoising function (typically a neural network) to reverse a step-wise diffusion process. The diffusion process corrupts tokens of a fixed length sequence, by replacing them by the Absorbing State token. This Absorbing State [MASK] is an artificial token that does not occur in the domain, but indicates corruption of data. In the forward diffusion process, each token either stays in its state, or transitions to [MASK] with a fixed probability. In a training step, data is partially masked, and the denoising function is optimized to reconstruct the original, unmasked sample.

The sampling process on the other hand starts with a sample consisting of only [MASK] tokens. Given a fully or partially masked sample as input, the model predicts a full musical piece. The fewer [MASK] tokens there are in the input sample, the more reliable the full predicted musical piece becomes. Thus sampling is performed in S steps: Instead of using the full predicted piece, only a small fraction of the masked piece is unmasked in each step. The partially unmasked sample is then used as input to condition the next unmasking step.

Tracks of musical pieces are represented as time-quantized series of discrete tokens, each timestep of fixed duration represents either pitch, pause or note-off in the corresponding interval. In the model, for each track, pitch indices are vector embedded, before a convolutional layer summarizes sets of 4 adjacent embeddings into one vector, effectively reducing the sequence length to a quarter of its original length. The summarized embeddings are then passed through a stack of transformer blocks, before being decompressed again using a transpose-convolution. For each track, a head then predicts the logits of all tokens in the unmasked sample.

3. EXPERIMENTS

3.1 Dataset and Preprocessing

Like Mittal et al. [3], we use the Lakh MIDI Dataset (LMD) [6] for all our experiments. We extract $\frac{4}{4}$ monophonic melodies and trios with lengths between 16 and 64 bars using Magenta's MusicVAE [7] pipelines. The trios consist of a monophonic melody, a monophonic bassline

Late-Breaking / Demo Session Extended Abstract, ISMIR 2022 Conference

and a polyphonic drum track. The drum track itself includes 9 canonical drums, resulting in 512 (2^9) different events. In all sequences, all notes' onsets and durations were quantized to 16^{th} notes.

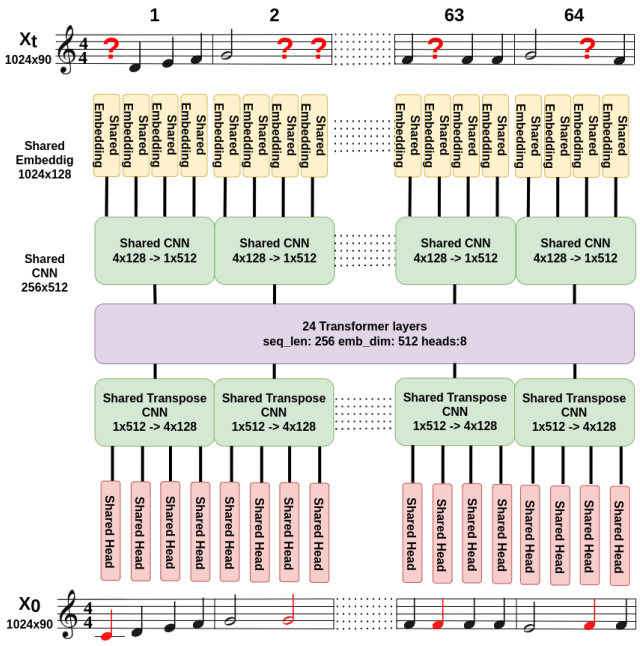

Figure 1. SCHmUBERT Architecture
The staves in the upper half represent the model's corrupted input, question marks symbolize [MASK] tokens. Actually all depicted notes are formed by a concatenation of 16^{th} notes. The denoising function reconstructs all notes, but in the sampling process, only some of the uncovered notes are transferred to X_{t-1}

3.2 Training

This section briefly outlines the training hyperparameters for the configuration of the 64 bar melody model which is sketched in Figure 1.

Diffusion Model timesteps	1024
Optimizer	Adam @ lr= $5 * 10^{-5}$
Batch size	64
Transformer layers	24
Transformer embedding size	512
Transformer attention heads	8
Total parameters	77M
Train steps	100,000
GPU	4x NVIDIA 2080 Ti
Duration	10h

3.3 Sampling and Evaluation

We trained our model for melodies and trios, for sequence lengths of 16 and 64 bars. For all configurations, the models were able to generate realistic, appealing samples with excellent long-time coherence. Once trained, the model can be used for unconditional generation, but also allows for conditional infilling without extra training. Any token in a musical piece can be replaced by [MASK], allowing for example the interpolation between two pieces. Another use case is accompaniment generation: given a melody track and two tracks consisting only of [MASK] for drum and bass, the trio model can fill the masked tracks. Although this might be a less common use case, any arbitrary combination of the previous infilling tasks can be performed. To evaluate long-time coherence in the sampled musical pieces, we use the framewise self-similarity metrics Mittal et al. describe in [3]:

In a frame of 4 bars (64 tokens), mean and variance of pitch and duration are calculated, defining two Gaussian probability density functions (PDFs) for each frame. Using a hop size of 2 bars, the overlap area OA_i between Gaussian PDFs of adjacent frames is calculated. For each piece, μ_{OA} and σ^2_{OA} are then calculated for pitch and duration. Calculating the same statistics for a ground truth (training data) enables calculating *Consistency* and *Variance* metrics for our sampled pieces:

$$\text{Consistency} = \max(0, 1 - \frac{\mid \mu_{OA} - \mu_{GT} \mid}{\mu_{GT}}) \quad (1)$$

$$\text{Variance} = \max(0, 1 - \frac{\mid \sigma^2_{OA} - \sigma^2_{GT} \mid}{\sigma^2_{GT}}) \quad (2)$$

For comparability with Magenta's Diffusion on MusicVAE (MDMVAE) latents [3], we used only a subset of 1,000,000 64 bar melodies for training and evaluation. Like Mittal et al. [3], we masked out the central half of all tokens in each piece, and evaluated our metrics on batches of 1,000 pieces.

Setting	Unconditional				Infilling			
Quantity	Pitch		Duration		Pitch		Duration	
Metric	C	Var	C	Var	C	Var	C	Var
Train Data	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Test Data	1.00	0.96	1.00	0.96	1.00	0.96	1.00	0.91
MDMVAE 64 bar	0.99	0.90	0.96	0.92	0.97	0.87	0.97	0.80
Melody 64 bar (ours)	**0.99**	**0.90**	**0.99**	**0.94**	**0.99**	**0.98**	**0.99**	**0.96**

Our implementation, trained model weights and some selected samples are available at `https://github.com/plassma/symbolic-music-discrete-diffusion`.

4. CONCLUSION AND FUTURE WORK

To our knowledge, SCHmUBERT is the first model that directly and successfully applies D3PMs to symbolic music. The proposed model is able to generate appealing, diverse samples, that match the state-of-the-art quality in the framewise self-similarity metrics [3]. In infilling, SCHmUBERT clearly outperforms the combination of MusicVAE and DDPM, which is plausible given that the DDPM can only use a fraction of the latents of MusicVAE in Mittal et al. [3]. Additionally, our model can be used for infilling or accompaniment generation without extra training. Future research includes the exploration of the possibility for guided diffusion using (adversarially robust) classifiers [1]. To finally assess the generation quality of the model, more hyperparameter tuning is necessary.

5. REFERENCES

[1] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 8780–8794. [Online]. Available: https://proceedings.neurips.cc/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf

[2] S. Bond-Taylor, P. Hessey, H. Sasaki, T. P. Breckon, and C. G. Willcocks, "Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 170–188.

[3] G. Mittal, J. H. Engel, C. Hawthorne, and I. Simon, "Symbolic music generation with diffusion models," in *ISMIR 2021*, 2021. [Online]. Available: https://archives.ismir.net/ismir2021/paper/000058.pdf

[4] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 2256–2265. [Online]. Available: https://proceedings.mlr.press/v37/sohl-dickstein15.html

[5] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. van den Berg, "Structured denoising diffusion models in discrete state-spaces," *CoRR*, vol. abs/2107.03006, 2021. [Online]. Available: https://arxiv.org/abs/2107.03006

[6] C. Raffel, "Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching," Ph.D. dissertation, COLUMBIA UNIVERSITY, 2016. [Online]. Available: https://colinraffel.com/projects/lmd/

[7] A. Roberts, J. H. Engel, C. Raffel, C. Hawthorne, and D. Eck, "A hierarchical latent vector model for learning long-term structure in music," *CoRR*, vol. abs/1803.05428, 2018. [Online]. Available: http://arxiv.org/abs/1803.05428