Music Modeling and Music Generation with Deep Learning.

Dr. Tristan Behrens, LinkedIn, April 30th 2023.

What is new in this version?

Papers:

- Added: Moûsai: Text-to-Music Generation with Long-Context Latent Diffusion
- Added: RAVE: A variational autoencoder for fast and high-quality neural audio synthesis
- Added: Musika! Fast Infinite Waveform Music Generation
- Added: WaveNet: A Generative Model for Raw Audio

Datasets: -

Music Modeling Timeline.

- 2023 February ERNIE-Music: Text-to-Waveform Music Generation with Diffusion Models
- 2023 January Moûsai: Text-to-Music Generation with Long-Context Latent Diffusion
- 2023 January MusicLM: Generating Music From Text
- 2022 September, AudioLM: a Language Modeling Approach to Audio Generation
- 20233 August Musika! Fast Infinite Waveform Music Generation
- 2022 May, Symphony Generation with Permutation Invariant Language Model.
- 2021 November, RAVE: A variational autoencoder for fast and high-quality neural audio synthesis
- 2021 November, Theme Transformer: Symbolic Music Generation with Theme-Conditioned Transformer.
- 2021 July, MidiBERT-Piano: Large-scale Pre-training for Symbolic Music Understanding.
- 2021 June, MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training.
- 2021 March Symbolic Music Generation with Diffusion Models.
- 2020 August, Learning Interpretable Representation for Controllable Polyphonic Music Generation.

- 2020 August, MMM : Exploring Conditional Multi-Track Music Generation with the Transformer.
- 2020 July, Transformer-XL Based Music Generation with Multiple Sequences of Time-valued Notes.
- 2020 May, Transformer VAE: A Hierarchical Model for Structure-Aware and Interpretable Music Representation Learning.
- 2020 April, Jukebox: A Generative Model for Music
- 2019 November, TonicNet, Improving Polyphonic Music Models with Feature-Rich Encoding. Omar Peracha.
- 2019 July, LakhNES: Improving multi-instrumental music generation with cross-domain pre-training. Chris Donahue et alia.
- 2019 July, R-Transformer: Recurrent Neural Network Enhanced Transformer.
- 2019 April, MuseNet. OpenAI.
- 2019 March, Counterpoint by Convolution
- 2018 September, Music Transformer.
- 2018 June, Learning a Latent Space of Multitrack Measures
- 2018 February, BachProp: Learning to Compose Music in Multiple Styles.
- 2017 October, Polyphonic Music Generation with Sequence Generative Adversarial Networks
- 2017 October, Automatic Stylistic Composition of Bach Chorales with Deep LSTM
- 2017 May, Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models
- 2017 March, MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation.
- 2016 December, DeepBach: a Steerable Model for Bach Chorales Generation.
- 2016 December, Generating Music by Fine-Tuning Recurrent Neural Networks with Reinforcement Learning
- 2016 November, C-RNN-GAN: Continuous recurrent neural networks with adversarial training. Olof Mogren.
- 2016 September, WaveNet: A Generative Model for Raw Audio
- 2016 September, SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient
- 2016 August, BachBot: Automatic composition in the style of Bach chorales. Feynman Liang.
- 2002, Finding Temporal Structure in Music: Blues Improvisation with LSTM Recurrent Networks Douglas Eck and Jürgen Schmidhuber.
- 1991, HARMONET: a neural net for harmonizing chorales in the style of J.S.Bach
- 1989 Winter, A Connectionist Approach To Algorithmic Composition.
- 1959, Experimental Music: Composition with an Electronic Computer. Lejaren Hiller.

Abstracts Collection.

ERNIE-Music: Text-to-Waveform Music Generation with Diffusion Models

> In recent years, there has been an increased popularity in image and speech generation using diffusion models. However, directly generating music waveforms from free-form text prompts is still under-explored. In this paper, we propose the first text-to-waveform music generation model that can receive arbitrary texts using diffusion models. We incorporate the free-form textual prompt as the condition to guide the waveform generation process of diffusion models. To solve the problem of lacking such text-music parallel data, we collect a dataset of text-music pairs from the Internet with weak supervision. Besides, we compare the effect of two prompt formats of conditioning texts (music tags and free-form texts) and prove the superior performance of our method in terms of text-music relevance. We further demonstrate that our generated music in the waveform domain outperforms previous works by a large margin in terms of diversity, quality, and text-music relevance.

Moûsai: Text-to-Music Generation with Long-Context Latent Diffusion

> The recent surge in popularity of diffusion models for image generation has brought new attention to the potential of these models in other areas of media synthesis. One area that has yet to be fully explored is the application of diffusion models to music generation. Music generation requires to handle multiple aspects, including the temporal dimension, long-term structure, multiple layers of overlapping sounds, and nuances that only trained listeners can detect. In our work, we investigate the potential of diffusion models for text-conditional music generation. We develop a cascading latent diffusion approach that can generate multiple minutes of high-quality stereo music at 48kHz from textual descriptions. For each model, we make an effort to maintain reasonable inference speed, targeting real-time on a single consumer GPU. In addition to trained models, we provide a collection of open-source libraries with the hope of facilitating future work in the field.

MusicLM: Generating Music From Text

> We introduce MusicLM, a model generating high-fidelity music from text descriptions such as "a calming violin melody backed by a distorted guitar riff". MusicLM casts the process of conditional music generation as a hierarchical sequence-to-sequence modeling task, and it generates music at 24 kHz that remains consistent over several minutes. Our experiments show that MusicLM outperforms previous systems both in audio quality and adherence to the text description. Moreover, we demonstrate that MusicLM can be conditioned on both text and a melody in that it can transform whistled

> and hummed melodies according to the style described in a text caption. To support future research, we publicly release MusicCaps, a dataset composed of 5.5k music-text pairs, with rich text descriptions provided by human experts.

AudioLM: a Language Modeling Approach to Audio Generation

> We introduce AudioLM, a framework for high-quality audio generation with long-term consistency. AudioLM maps the input audio to a sequence of discrete tokens and casts audio generation as a language modeling task in this representation space. We show how existing audio tokenizers provide different trade-offs between reconstruction quality and long-term structure, and we propose a hybrid tokenization scheme to achieve both objectives. Namely, we leverage the discretized activations of a masked language model pre-trained on audio to capture long-term structure and the discrete codes produced by a neural audio codec to achieve high-quality synthesis. By training on large corpora of raw audio waveforms, AudioLM learns to generate natural and coherent continuations given short prompts. When trained on speech, and without any transcript or annotation, AudioLM generates syntactically and semantically plausible speech continuations while also maintaining speaker identity and prosody for unseen speakers. Furthermore, we demonstrate how our approach extends beyond speech by generating coherent piano music continuations, despite being trained without any symbolic representation of music.

Musika! Fast Infinite Waveform Music Generation

> Fast and user-controllable music generation could enable novel ways of composing or performing music. However, state-of-the-art music generation systems require large amounts of data and computational resources for training, and are slow at inference. This makes them impractical for real-time interactive use. In this work, we introduce Musika, a music generation system that can be trained on hundreds of hours of music using a single consumer GPU, and that allows for much faster than real-time generation of music of arbitrary length on a consumer CPU. We achieve this by first learning a compact invertible representation of spectrogram magnitudes and phases with adversarial autoencoders, then training a Generative Adversarial Network (GAN) on this representation for a particular music domain. A latent coordinate system enables generating arbitrarily long sequences of excerpts in parallel, while a global context vector allows the music to remain stylistically coherent through time. We perform quantitative evaluations to assess the quality of the generated samples and showcase options for user control in piano and techno music generation. We release the source code and pretrained autoencoder weights at this http URL, such that a GAN can be trained on a new music domain with a single GPU in a matter of hours.

Symphony Generation with Permutation Invariant Language Model.

> In this work, we present a symbolic symphony music generation solution, SymphonyNet, based on a permutation invariant language model. To bridge the gap between text generation and symphony generation task, we propose a novel Multi-track Multi-instrument Repeatable (MMR) representation with particular 3-D positional embedding and a modified Byte Pair Encoding algorithm (Music BPE) for music tokens. A novel linear transformer decoder architecture is introduced as a backbone for modeling extra-long sequences of symphony tokens. Meanwhile, we train the decoder to learn automatic orchestration as a joint task by masking instrument information from the input. We also introduce a large-scale symbolic symphony dataset for the advance of symphony generation research. Our empirical results show that our proposed approach can generate coherent, novel, complex and harmonious symphony compared to human composition, which is the pioneer solution for multi-track multi-instrument symbolic music generation.

RAVE: A variational autoencoder for fast and high-quality neural audio synthesis

> Deep generative models applied to audio have improved by a large margin the state-of-the-art in many speech and music related tasks. However, as raw waveform modelling remains an inherently difficult task, audio generative models are either computationally intensive, rely on low sampling rates, are complicated to control or restrict the nature of possible signals. Among those models, Variational AutoEncoders (VAE) give control over the generation by exposing latent variables, although they usually suffer from low synthesis quality. In this paper, we introduce a Realtime Audio Variational autoEncoder (RAVE) allowing both fast and high-quality audio waveform synthesis. We introduce a novel two-stage training procedure, namely representation learning and adversarial fine-tuning. We show that using a post-training analysis of the latent space allows a direct control between the reconstruction fidelity and the representation compactness. By leveraging a multi-band decomposition of the raw waveform, we show that our model is the first able to generate 48kHz audio signals, while simultaneously running 20 times faster than real-time on a standard laptop CPU. We evaluate synthesis quality using both quantitative and qualitative subjective experiments and show the superiority of our approach compared to existing models. Finally, we present applications of our model for timbre transfer and signal compression. All of our source code and audio examples are publicly available.

Theme Transformer: Symbolic Music Generation with Theme-Conditioned Transformer.

> Attention-based Transformer models have been increasingly employed for automatic music generation. To condition the generation process of such a model with a user-

> specified sequence, a popular approach is to take that conditioning sequence as a priming sequence and ask a Transformer decoder to generate a continuation. However, this prompt-based conditioning cannot guarantee that the conditioning sequence would develop or even simply repeat itself in the generated continuation. In this paper, we propose an alternative conditioning approach, called theme-based conditioning, that explicitly trains the Transformer to treat the conditioning sequence as a thematic material that has to manifest itself multiple times in its generation result. This is achieved with two main technical contributions. First, we propose a deep learning-based approach that uses contrastive representation learning and clustering to automatically retrieve thematic materials from music pieces in the training data. Second, we propose a novel gated parallel attention module to be used in a sequence-to-sequence (seq2seq) encoder/decoder architecture to more effectively account for a given conditioning thematic material in the generation process of the Transformer decoder. We report on objective and subjective evaluations of variants of the proposed Theme Transformer and the conventional prompt-based baseline, showing that our best model can generate, to some extent, polyphonic pop piano music with repetition and plausible variations of a given condition.

MidiBERT-Piano: Large-scale Pre-training for Symbolic Music Understanding.

> This paper presents an attempt to employ the mask language modeling approach of BERT to pre-train a 12-layer Transformer model over 4,166 pieces of polyphonic piano MIDI files for tackling a number of symbolic-domain discriminative music understanding tasks. These include two note-level classification tasks, i.e., melody extraction and velocity prediction, as well as two sequence-level classification tasks, i.e., composer classification and emotion classification. We find that, given a pre-trained Transformer, our models outperform recurrent neural network based baselines with less than 10 epochs of fine-tuning. Ablation studies show that the pre-training remains effective even if none of the MIDI data of the downstream tasks are seen at the pre-training stage, and that freezing the self-attention layers of the Transformer at the fine-tuning stage slightly degrades performance. All the five datasets employed in this work are publicly available, as well as checkpoints of our pre-trained and fine-tuned models. As such, our research can be taken as a benchmark for symbolic-domain music understanding.

MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training.

> Symbolic music understanding, which refers to the understanding of music from the symbolic data (e.g., MIDI format, but not audio), covers many music applications such

> as genre classification, emotion classification, and music pieces matching. While good music representations are beneficial for these applications, the lack of training data hinders representation learning. Inspired by the success of pre-training models in natural language processing, in this paper, we develop MusicBERT, a large-scale pre-trained model for music understanding. To this end, we construct a large-scale symbolic music corpus that contains more than 1 million music songs. Since symbolic music contains more structural (e.g., bar, position) and diverse information (e.g., tempo, instrument, and pitch), simply adopting the pre-training techniques from NLP to symbolic music only brings marginal gains. Therefore, we design several mechanisms, including OctupleMIDI encoding and bar-level masking strategy, to enhance pre-training with symbolic music data. Experiments demonstrate the advantages of MusicBERT on four music understanding tasks, including melody completion, accompaniment suggestion, genre classification, and style classification. Ablation studies also verify the effectiveness of our designs of OctupleMIDI encoding and bar-level masking strategy in MusicBERT.

Symbolic Music Generation with Diffusion Models.

> Score-based generative models and diffusion probabilistic models have been successful at generating high-quality samples in continuous domains such as images and audio. However, due to their Langevin-inspired sampling mechanisms, their application to discrete and sequential data has been limited. In this work, we present a technique for training diffusion models on sequential data by parameterizing the discrete domain in the continuous latent space of a pre-trained variational autoencoder. Our method is non-autoregressive and learns to generate sequences of latent embeddings through the reverse process and offers parallel generation with a constant number of iterative refinement steps. We apply this technique to modeling symbolic music and show strong unconditional generation and post-hoc conditional infilling results compared to autoregressive language models operating over the same continuous embeddings.

Learning Interpretable Representation for Controllable Polyphonic Music Generation

> While deep generative models have become the leading methods for algorithmic composition, it remains a challenging problem to control the generation process because the latent variables of most deep-learning models lack good interpretability. Inspired by the content-style disentanglement idea, we design a novel architecture, under the VAE framework, that effectively learns two interpretable latent factors of polyphonic music: chord and texture. The current model focuses on learning 8-beat long piano composition segments. We show that such chord-texture disentanglement provides a controllable generation pathway leading to a wide spectrum of applications,

> including compositional style transfer, texture variation, and accompaniment arrangement. Both objective and subjective evaluations show that our method achieves a successful disentanglement and high quality controlled music generation.

MMM : Exploring Conditional Multi-Track Music Generation with the Transformer.

> We propose the Multi-Track Music Machine (MMM), a generative system based on the Transformer architecture that is capable of generating multi-track music. In contrast to previous work, which represents musical material as a single time-ordered sequence, where the musical events corresponding to different tracks are interleaved, we create a time-ordered sequence of musical events for each track and concatenate several tracks into a single sequence. This takes advantage of the Transformer's attention-mechanism, which can adeptly handle long-term dependencies. We explore how various representations can offer the user a high degree of control at generation time, providing an interactive demo that accommodates track-level and bar-level inpainting, and offers control over track instrumentation and note density.

Transformer-XL Based Music Generation with Multiple Sequences of Time-valued Notes.

> Current state-of-the-art AI based classical music creation algorithms such as Music Transformer are trained by employing single sequence of notes with time-shifts. The major drawback of absolute time interval expression is the difficulty of similarity computing of notes that share the same note value yet different tempos, in one or among MIDI files. In addition, the usage of single sequence restricts the model to separately and effectively learn music information such as harmony and rhythm. In this paper, we propose a framework with two novel methods to respectively track these two shortages, one is the construction of time-valued note sequences that liberate note values from tempos and the other is the separated usage of four sequences, namely, former note on to current note on, note on to note off, pitch, and velocity, for jointly training of four Transformer-XL networks. Through training on a 23-hour piano MIDI dataset, our framework generates significantly better and hour-level longer music than three state-of-the-art baselines, namely Music Transformer, DeepJ, and single sequence-based Transformer-XL, evaluated automatically and manually.

Transformer VAE: A Hierarchical Model for Structure-Aware and Interpretable Music Representation Learning.

> Structure awareness and interpretability are two of the most desired properties of music generation algorithms. Structure-aware models generate more natural and coherent

> music with long-term dependencies, while interpretable models are more friendly for human-computer interaction and co-creation. To achieve these two goals simultaneously, we designed the Transformer Variational AutoEncoder, a hierarchical model that unifies the efforts of two recent breakthroughs in deep music generation: 1) the Music Transformer and 2) Deep Music Analogy. The former learns long-term dependencies using attention mechanism, and the latter learns interpretable latent representations using a disentangled conditional-VAE. We showed that Transformer VAE is essentially capable of learning a context-sensitive hierarchical representation, regarding local representations as the context and the dependencies among the local representations as the global structure. By interacting with the model, we can achieve context transfer, realizing the imaginary situation of "what if" a piece is developed following the music flow of another piece.

Jukebox: A Generative Model for Music.

> We introduce Jukebox, a model that generates music with singing in the raw audio domain. We tackle the long context of raw audio using a multi-scale VQ-VAE to compress it to discrete codes, and modeling those using autoregressive Transformers. We show that the combined model at scale can generate high-fidelity and diverse songs with coherence up to multiple minutes. We can condition on artist and genre to steer the musical and vocal style, and on unaligned lyrics to make the singing more controllable. We are releasing thousands of non cherry-picked samples, along with model weights and code.

Improving Polyphonic Music Models with Feature-Rich Encoding.

> This paper explores sequential modelling of polyphonic music with deep neural networks. While recent breakthroughs have focussed on network architecture, we demonstrate that the representation of the sequence can make an equally significant contribution to the performance of the model as measured by validation set loss. By extracting salient features inherent to the training dataset, the model can either be conditioned on these features or trained to predict said features as extra components of the sequences being modelled. We show that training a neural network to predict a seemingly more complex sequence, with extra features included in the series being modelled, can improve overall model performance significantly. We first introduce TonicNet, a GRU-based model trained to initially predict the chord at a given time-step before then predicting the notes of each voice at that time-step, in contrast with the typical approach of predicting only the notes. We then evaluate TonicNet on the canonical JSB Chorales dataset and obtain state-of-the-art results.

LakhNES: Improving multi-instrumental music generation with

cross-domain pre-training.

> We are interested in the task of generating multi-instrumental music scores. The Transformer architecture has recently shown great promise for the task of piano score generation; here we adapt it to the multi-instrumental setting. Transformers are complex, high-dimensional language models which are capable of capturing long-term structure in sequence data, but require large amounts of data to fit. Their success on piano score generation is partially explained by the large volumes of symbolic data readily available for that domain. We leverage the recently-introduced NES-MDB dataset of four-instrument scores from an early video game sound synthesis chip (the NES), which we find to be well-suited to training with the Transformer architecture. To further improve the performance of our model, we propose a pre-training technique to leverage the information in a large collection of heterogeneous music, namely the Lakh MIDI dataset. Despite differences between the two corpora, we find that this transfer learning procedure improves both quantitative and qualitative performance for our primary task.

R-Transformer: Recurrent Neural Network Enhanced Transformer.

> Recurrent Neural Networks have long been the dominating choice for sequence modeling. However, it severely suffers from two issues: impotent in capturing very long-term dependencies and unable to parallelize the sequential computation procedure. Therefore, many non-recurrent sequence models that are built on convolution and attention operations have been proposed recently. Notably, models with multi-head attention such as Transformer have demonstrated extreme effectiveness in capturing long-term dependencies in a variety of sequence modeling tasks. Despite their success, however, these models lack necessary components to model local structures in sequences and heavily rely on position embeddings that have limited effects and require a considerable amount of design efforts. In this paper, we propose the R-Transformer which enjoys the advantages of both RNNs and the multi-head attention mechanism while avoids their respective drawbacks. The proposed model can effectively capture both local structures and global long-term dependencies in sequences without any use of position embeddings. We evaluate R-Transformer through extensive experiments with data from a wide range of domains and the empirical results show that R-Transformer outperforms the state-of-the-art methods by a large margin in most of the tasks. We have made the code publicly available at \url{this https URL}.

Counterpoint by Convolution.

> Machine learning models of music typically break up the task of composition into a chronological process, composing a piece of music in a single pass from beginning to

end. On the contrary, human composers write music in a nonlinear fashion, scribbling motifs here and there, often revisiting choices previously made. In order to better approximate this process, we train a convolutional neural network to complete partial musical scores, and explore the use of blocked Gibbs sampling as an analogue to rewriting. Neither the model nor the generative procedure are tied to a particular causal direction of composition. Our model is an instance of orderless NADE (Uria et al., 2014), which allows more direct ancestral sampling. However, we find that Gibbs sampling greatly improves sample quality, which we demonstrate to be due to some conditional distributions being poorly modeled. Moreover, we show that even the cheap approximate blocked Gibbs procedure from Yao et al. (2014) yields better samples than ancestral sampling, based on both log-likelihood and human evaluation.

Music Transformer.

Music relies heavily on repetition to build structure and meaning. Self-reference occurs on multiple timescales, from motifs to phrases to reusing of entire sections of music, such as in pieces with ABA structure. The Transformer (Vaswani et al., 2017), a sequence model based on self-attention, has achieved compelling results in many generation tasks that require maintaining long-range coherence. This suggests that self-attention might also be well-suited to modeling music. In musical composition and performance, however, relative timing is critically important. Existing approaches for representing relative positional information in the Transformer modulate attention based on pairwise distance (Shaw et al., 2018). This is impractical for long sequences such as musical compositions since their memory complexity for intermediate relative information is quadratic in the sequence length. We propose an algorithm that reduces their intermediate memory requirement to linear in the sequence length. This enables us to demonstrate that a Transformer with our modified relative attention mechanism can generate minute-long compositions (thousands of steps, four times the length modeled in Oore et al., 2018) with compelling structure, generate continuations that coherently elaborate on a given motif, and in a seq2seq setup generate accompaniments conditioned on melodies. We evaluate the Transformer with our relative attention mechanism on two datasets, JSB Chorales and Piano-e-Competition, and obtain state-of-the-art results on the latter.

Learning a Latent Space of Multitrack Measures.

Discovering and exploring the underlying structure of multi-instrumental music using learning-based approaches remains an open problem. We extend the recent MusicVAE model to represent multitrack polyphonic measures as vectors in a latent space. Our approach enables several useful operations such as generating plausible measures from scratch, interpolating between measures in a musically meaningful way, and manipulating specific musical attributes. We also introduce chord conditioning, which

> allows all of these operations to be performed while keeping harmony fixed, and allows chords to be changed while maintaining musical "style". By generating a sequence of measures over a predefined chord progression, our model can produce music with convincing long-term structure. We demonstrate that our latent space model makes it possible to intuitively control and generate musical sequences with rich instrumentation (see this https URL for generated audio).

BachProp: Learning to Compose Music in Multiple Styles.

> Hand in hand with deep learning advancements, algorithms of music composition increase in performance. However, most of the successful models are designed for specific musical structures. Here, we present BachProp, an algorithmic composer that can generate music scores in any style given sufficient training data. To adapt BachProp to a broad range of musical styles, we propose a novel normalized representation of music and train a deep network to predict the note transition probabilities of a given music corpus. In this paper, new music scores sampled by BachProp are compared with the original corpora via crowdsourcing. This evaluation indicates that the music scores generated by BachProp are not less preferred than the original music corpus the algorithm was provided with.

Polyphonic Music Generation with Sequence Generative Adversarial Networks.

> We propose an application of sequence generative adversarial networks (SeqGAN), which are generative adversarial networks for discrete sequence generation, for creating polyphonic musical sequences. Instead of a monophonic melody generation suggested in the original work, we present an efficient representation of a polyphony MIDI file that simultaneously captures chords and melodies with dynamic timings. The proposed method condenses duration, octaves, and keys of both melodies and chords into a single word vector representation, and recurrent neural networks learn to predict distributions of sequences from the embedded musical word space. We experiment with the original method and the least squares method to the discriminator, which is known to stabilize the training of GANs. The network can create sequences that are musically coherent and shows an improved quantitative and qualitative measures. We also report that careful optimization of reinforcement learning signals of the model is crucial for general application of the model.

Automatic Stylistic Composition of Bach Chorales with Deep LSTM.

> This paper presents “BachBot”: an end-to-end automatic composition system for

> composing and completing music in the style of Bach's chorales using a deep long short-term memory (LSTM) generative model. We propose a new sequential encoding scheme for polyphonic music and a model for both composition and harmonization which can be efficiently sampled without expensive Markov Chain Monte Carlo (MCMC). Analysis of the trained model provides evidence of neurons specializing without prior knowledge or explicit supervision to detect common music-theoretic concepts such as tonics, chords, and cadences. To assess BachBot's success, we conducted one of the largest musical discrimination tests on 2336 participants. Among the results, the proportion of responses correctly differentiating BachBot from Bach was only 1% better than random guessing.

Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models.

> In unsupervised data generation tasks, besides the generation of a sample based on previous observations, one would often like to give hints to the model in order to bias the generation towards desirable metrics. We propose a method that combines Generative Adversarial Networks (GANs) and reinforcement learning (RL) in order to accomplish exactly that. While RL biases the data generation process towards arbitrary metrics, the GAN component of the reward function ensures that the model still remembers information learned from data. We build upon previous results that incorporated GANs and RL in order to generate sequence data and test this model in several settings for the generation of molecules encoded as text sequences (SMILES) and in the context of music generation, showing for each case that we can effectively bias the generation process towards desired metrics.

MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation.

> Most existing neural network models for music generation use recurrent neural networks. However, the recent WaveNet model proposed by DeepMind shows that convolutional neural networks (CNNs) can also generate realistic musical waveforms in the audio domain. Following this light, we investigate using CNNs for generating melody (a series of MIDI notes) one bar after another in the symbolic domain. In addition to the generator, we use a discriminator to learn the distributions of melodies, making it a generative adversarial network (GAN). Moreover, we propose a novel conditional mechanism to exploit available prior knowledge, so that the model can generate melodies either from scratch, by following a chord sequence, or by conditioning on the melody of previous bars (e.g. a priming melody), among other possibilities. The resulting model, named MidiNet, can be expanded to generate music with multiple MIDI channels (i.e. tracks). We conduct a user study to compare the

> melody of eight-bar long generated by MidiNet and by Google's MelodyRNN models, each time using the same priming melody. Result shows that MidiNet performs comparably with MelodyRNN models in being realistic and pleasant to listen to, yet MidiNet's melodies are reported to be much more interesting.

DeepBach: a Steerable Model for Bach Chorales Generation.

> This paper introduces DeepBach, a graphical model aimed at modeling polyphonic music and specifically hymn-like pieces. We claim that, after being trained on the chorale harmonizations by Johann Sebastian Bach, our model is capable of generating highly convincing chorales in the style of Bach. DeepBach's strength comes from the use of pseudo-Gibbs sampling coupled with an adapted representation of musical data. This is in contrast with many automatic music composition approaches which tend to compose music sequentially. Our model is also steerable in the sense that a user can constrain the generation by imposing positional constraints such as notes, rhythms or cadences in the generated score. We also provide a plugin on top of the MuseScore music editor making the interaction with DeepBach easy to use.

Generating Music by Fine-Tuning Recurrent Neural Networks with Reinforcement Learning.

> Supervised learning with next-step prediction is a common way to train a sequence prediction model; however, it suffers from known failure modes and is notoriously difficult to train models to learn certain properties, such as having a coherent global structure. Reinforcement learning can be used to impose arbitrary properties on generated data by choosing appropriate reward functions. In this paper we propose a novel approach for sequence training, where we refine a sequence predictor by optimizing for some imposed reward functions, while maintaining good predictive properties learned from data. We propose efficient ways to solve this by augmenting deep Q-learning with a cross-entropy reward and deriving novel off-policy methods for RNNs from stochastic optimal control (SOC). We explore the usefulness of our approach in the context of music gener- ation. An LSTM is trained on a large corpus of songs to predict the next note in a musical sequence. This Note-RNN is then refined using RL, where the reward function is a combination of rewards based on rules of music theory, as well as the output of another trained Note-RNN. We show that this combination of ML and RL can not only produce more pleasing melodies, but that it can significantly reduce unwanted behaviors and failure modes of the RNN.

C-RNN-GAN: Continuous recurrent neural networks with adversarial training.

> Generative adversarial networks have been proposed as a way of efficiently training deep generative neural networks. We propose a generative adversarial model that works on continuous sequential data, and apply it by training it on a collection of classical music. We conclude that it generates music that sounds better and better as the model is trained, report statistics on generated music, and let the reader judge the quality by downloading the generated songs.

WaveNet: A Generative Model for Raw Audio

> This paper introduces WaveNet, a deep neural network for generating raw audio waveforms. The model is fully probabilistic and autoregressive, with the predictive distribution for each audio sample conditioned on all previous ones; nonetheless we show that it can be efficiently trained on data with tens of thousands of samples per second of audio. When applied to text-to-speech, it yields state-of-the-art performance, with human listeners rating it as significantly more natural sounding than the best parametric and concatenative systems for both English and Mandarin. A single WaveNet can capture the characteristics of many different speakers with equal fidelity, and can switch between them by conditioning on the speaker identity. When trained to model music, we find that it generates novel and often highly realistic musical fragments. We also show that it can be employed as a discriminative model, returning promising results for phoneme recognition.

SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient.

> As a new way of training generative models, Generative Adversarial Nets (GAN) that uses a discriminative model to guide the training of the generative model has enjoyed considerable success in generating real-valued data. However, it has limitations when the goal is for generating sequences of discrete tokens. A major reason lies in that the discrete outputs from the generative model make it difficult to pass the gradient update from the discriminative model to the generative model. Also, the discriminative model can only assess a complete sequence, while for a partially generated sequence, it is non-trivial to balance its current score and the future one once the entire sequence has been generated. In this paper, we propose a sequence generation framework, called SeqGAN, to solve the problems. Modeling the data generator as a stochastic policy in reinforcement learning (RL), SeqGAN bypasses the generator differentiation problem by directly performing gradient policy update. The RL reward signal comes from the GAN discriminator judged on a complete sequence, and is passed back to the intermediate state-action steps using Monte Carlo search. Extensive experiments on synthetic data and real-world tasks demonstrate significant improvements over strong baselines.

BachBot: Automatic composition in the style of Bach chorales.

> This thesis investigates Bach's composition style using deep sequence learning. We develop BachBot: an automatic stylistic composition system for composing polyphonic music in the style of Bach's chorales. Our approach encodes music scores into a sequential format, reducing the task to one of se- quence modeling. Traditional N - gram language models are found to be insufficient, prompting the use of RNN sequence models. We find a 3-layer stacked LSTM performs best and conduct analyses and evaluations to understand its success and failure modes. Unlike many previous works, we avoid allowing prior assumptions about music impact model design, opting instead to build systems that learn rather than ones which encode prior hypotheses. While this is not the first application of deep LSTM to Bach chorales, our work consists of the following novel contributions. First, we devise a sequential encoding for polyphonic music which resolves issues noted by prior work, including: the ability to determine when notes end and a time-resolution exceeding all prior work by at least 2×. Second, we identify neurons which, without any prior knowledge or supervision, have learned to specifically detect musically meaningful concepts such as chords and cadences. To our knowledge, this is the first reported reesult demonstrating LSTM is capable of learning high-level musically-meaningful concepts automatically from data. Finally, we build a web-based musical Turing test (www.bachbot.com) and evaluate on a participant pool more than 3× larger than the next-closest comparable study [91]. We find that a human evaluation study promoted over social media can yield responses from a signif- icant number (165 at time of writing) of domain experts. After evaluating BachBot on 721 participants, we found that participants could only differentiate BachBot's generated chorales from Bach's originals works only 9% better than random guessing. In other words, generat- ing stylistically successful Bach chorales is more closed (as a result of BachBot) than open a problem.

Finding Temporal Structure in Music: Blues Improvisation with LSTM Recurrent Networks.

> We consider the problem of extracting essential ingredients of music signals, such as a well-defined global temporal structure in the form of nested periodicities (or meter). We investigate whether we can construct an adaptive signal processing device that learns by example how to generate new instances of a given musical style. Because recurrent neural networks (RNNs) can, in principle, learn the temporal structure of a signal, they are good candidates for such a task. Unfortunately, music composed by standard RNNs often lacks global coherence. The reason for this failure seems to be that RNNs cannot keep track of temporally distant events that indicate global music structure. Long short-term memory (LSTM) has succeeded in similar domains where other RNNs have failed, such as timing and counting and the learning of context sensitive languages. We show that LSTM is also a good mechanism for learning to compose

> music. We present experimental results showing that LSTM successfully learns a form of blues music and is able to compose novel (and we believe pleasing) melodies in that style. Remarkably, once the network has found the relevant structure, it does not drift from it: LSTM is able to play the blues with good timing and proper structure as long as one is willing to listen.

HARMONET: a neural net for harmonizing chorales in the style of J.S.Bach.

> HARMONET, a system employing connectionist networks for music pro- cessing, is presented. After being trained on some dozen Bach chorales using error backpropagation, the system is capable of producing four-part chorales in the style of J .s.Bach, given a one-part melody. Our system solves a musical real-world problem on a performance level appropriate for musical practice. HARMONET's power is based on (a) a new coding scheme capturing musically relevant information and (b) the integration of backpropagation and symbolic algorithms in a hierarchical system, com- bining the advantages of both.

A Connectionist Approach To Algorithmic Composition.

> With the advent of von Neumann-style computers, widespread exploration of new methods of music composition became possible. For the first time, complex sequences of carefully specified symbolic operations could be performed in a rapid fashion. Composers could develop algorithms embodying the compositional rules they were interested in and then use a computer to carry out these algorithms. In this way, composers could soon tell whether the results of their rules held artistic merit. This approach to algorithmic composition, based on the wedding between von Neumann computing machinery and rule-based software systems, has been prevalent for the past thirty years. The arrival of a new paradigm for computing has made a different approach to algorithmic composition possible. This new computing paradigm is called parallel distributed processing (PDP), also known as connectionism. Computation is performed by a collection of several simple processing units connected in a network and acting in cooperation (Rumelhart and McClelland 1986). This is in stark contrast to the single powerful central processor used in the von Neumann architecture. One of the major features of the PDP approach is that it replaces strict rule-following behavior with regularity-learning and generalization (Dolson 1989). This fundamental shift allows the development of new algorithmic composition methods that rely on learning the structure of existing musical examples and generalizing from these learned structures to compose new pieces. These methods contrast greatly with the majority of older schemes that simply follow a previously assembled set of compositional rules, resulting in brittle systems typically unable to appropriately handle unexpected musical situations. To be sure, other algorithmic composition methods in the past have been

> based on abstracting certain features from musical examples and using these to create new compositions. Techniques such as Markov modeling with transition probability analysis (Jones 1981), Mathews' melody interpolation method (Mathews and Rosler 1968), and Cope's EMI system (Cope 1987) can all be placed in this category. However, the PDP computational paradigm provides a single powerful unifying approach within which to formulate a variety of algorithmic composition methods of this type. These new learning methods combine many of the features of the techniques listed above and add a variety of new capabilities. Perhaps most importantly, though, they yield different and interesting musical results. This paper presents a particular type of PDP network for music composition applications. Various issues are discussed in designing the network, choosing the music representation used, training the network, and using it for composition. Comparisons are made to previous methods of algorithmic composition, and examples of the network's output are presented. This paper is intended to provide an indication of the power and range of PDP methods for algorithmic composition and to encourage others to begin exploring this new approach. Hence, rather than merely presenting a reduced compositional technique, alternative approaches and tangential ideas are included throughout as points of departure for further efforts.

Other inspiring works

Timeline.

- 2017 August, Learning Musical relations using Gated Autoencoders. Stefan Lattner, Maarten Grachten, Gerhard Widmer.

Abstracts Collection.

Learning Musical relations using Gated Autoencoders.

> Music is usually highly structured and it is still an open question how to design models which can successfully learn to recognize and represent musical structure. A fundamental problem is that structurally related patterns can have very distinct appearances, because the structural relationships are often based on transformations of musical material, like chromatic or diatonic transposition, inversion, retrograde, or rhythm change. In this preliminary work, we study the potential of two unsupervised learning techniques - Restricted Boltzmann Machines (RBMs) and Gated Autoencoders (GAEs) - to capture pre-defined transformations from constructed data pairs. We evaluate the models by using the learned representations as inputs in a discriminative task where for a given type of transformation (e.g. diatonic transposition), the specific relation between two musical patterns must be recognized (e.g. an upward

> transposition of diatonic steps). Furthermore, we measure the reconstruction error of models when reconstructing musical transformed patterns. Lastly, we test the models in an analogy-making task. We find that it is difficult to learn musical transformations with the RBM and that the GAE is much more adequate for this task, since it is able to learn representations of specific transformations that are largely content-invariant. We believe these results show that models such as GAEs may provide the basis for more encompassing music analysis systems, by endowing them with a better understanding of the structures underlying music.

Datasets.

The Lakh MIDI Dataset v0.1.

- Paper
- Download

> The Lakh MIDI dataset is a collection of 176,581 unique MIDI files, 45,129 of which have been matched and aligned to entries in the Million Song Dataset. Its goal is to facilitate large-scale music information retrieval, both symbolic (using the MIDI files alone) and audio content-based (using information extracted from the MIDI files as annotations for the matched audio files).

JSB Chorales.

- Paper
- Download
- Download

JS Fake Chorales.

- Download

> A MIDI dataset of 500 4-part chorales generated by the KS_Chorus algorithm, annotated with results from hundreds of listening test participants.

The MAESTRO Dataset.

- Paper
- Download

> MAESTRO (MIDI and Audio Edited for Synchronous TRacks and Organization) is a dataset composed of about 200 hours of virtuosic piano performances captured with

> fine alignment (~3 ms) between note labels and audio waveforms.

CrestMuse.

- Paper
- Download

RWC Music Database.

- Paper
- Download

Nottingham Database.

- Download
- Download
- Download

> This is a collection of 1200 British and American folk tunes, (hornpipe, jigs, and etc.) that was created by Eric Foxley and posted on Eric Foxley's Music Database.

POP909.

- Paper
- Download

> Music arrangement generation is a subtask of automatic music generation, which involves reconstructing and re-conceptualizing a piece with new compositional techniques. Such a generation process inevitably requires reference from the original melody, chord progression, or other structural information. Despite some promising models for arrangement, they lack more refined data to achieve better evaluations and more practical results. In this paper, we propose POP909, a dataset which contains multiple versions of the piano arrangements of 909 popular songs created by professional musicians. The main body of the dataset contains the vocal melody, the lead instrument melody, and the piano accompaniment for each song in MIDI format, which are aligned to the original audio files. Furthermore, we provide the annotations of tempo, beat, key, and chords, where the tempo curves are hand-labeled and others are done by MIR algorithms. Finally, we conduct several baseline experiments with this dataset using standard deep music generation algorithms.

DadaGP.

- Paper

- Download

> DadaGP is a dataset of 26,181 GuitarPro songs in 739 genres, converted to a token sequence format suitable for generative language models like GPT2, TransformerXL, etc.

js-fakes-4bars.

- Download

> This is a tokenized version of the JS-Fakes dataset by Omar Peracha. The representation is four tracks with four bars per track.

MetaMIDI Dataset.

- Download
- GitHub

> We introduce the MetaMIDI Dataset (MMD), a large scale collection of 436,631 MIDI files and metadata. In addition to the MIDI files, we provide artist, title and genre metadata that was collected during the scraping process when available.

GiantMIDI-Piano.

- GitHub

> GiantMIDI-Piano is a classical piano MIDI dataset contains 10,855 MIDI files of 2,786 composers. The curated subset by constraining composer surnames contains 7,236 MIDI files of 1,787 composers. GiantMIDI-Piano are transcribed from live recordings with a high-resolution piano transcription system.