MUSIC TRANSFORMER

Anonymous authors
Paper under double-blind review

ABSTRACT

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 is quadratic in the sequence length. We propose an algorithm that reduces the intermediate memory requirements to linear in the sequence length. This enables us to demonstrate that a Transformer with our modified relative attention mechanism can generate minute-long (thousands of steps) compositions with compelling structure, generate continuations that coherently elaborate on a given motif, and in a seq2seq setup generate accompaniments conditioned on melodies[1]. 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.

1 INTRODUCTION

A musical piece often consists of recurring elements at various levels, from motifs to phrases to sections such as verse-chorus. To generate a coherent piece, a model needs to reference elements that came before, sometimes in the distant past, repeating, varying, and further developing them to create contrast and surprise. Intuitively, self-attention (Parikh et al., 2016) appears to be a good match for this task. Self-attention over its own previous outputs allows an autoregressive model to access any part of the previously generated output at every step of generation. By contrast, recurrent neural networks have to learn to proactively store elements to be referenced in a fixed size state or memory, potentially making training much more difficult. We believe that repeating self-attention in multiple, successive layers of a Transformer decoder (Vaswani et al., 2017) helps capture the multiple levels at which self-referential phenomena exist in music.

In its original formulation, the Transformer relies on absolute position representations, using either positional sinusoids or learned position embeddings that are added to the per-position input representations. Recurrent and convolutional neural networks instead model position in relative terms: RNNs through their recurrence over the positions in their input, and CNNs by applying kernels that effectively choose which parameters to apply based on the relative position of the covered input representations.

Music has multiple dimensions along which relative differences arguably matter more than their absolute values; the two most prominent are timing and pitch. To capture such pairwise relations between representations, Shaw et al. (2018) introduce a relation-aware version of self-attention which they use successfully to modulate self-attention by the distance between two positions. We extend this approach to capture relative timing and optionally also pitch, which yields a dramatic improvement in sample quality and perplexity for JSB Chorales. As opposed to the original Transformer, samples

[1]Samples are available for listening at
https://storage.googleapis.com/music-transformer/index.html

from a Transformer with our relative attention mechanism maintain the regular timing grid present in this data set. The model furthermore captures timing on a more global level, giving rise to regular phrases.

The original formulation of relative attention (Shaw et al., 2018) requires $O(L^2D)$ memory where L is the sequence length and D is the dimension of the model's hidden state. This is prohibitive for long sequences such as those found in the Piano-e-Competition data set of human-performed virtuosic, classical piano music. In Section 3.4, we show how to reduce the memory requirements to $O(LD)$, making it practical to apply relative attention to long sequences.

The Piano-e-Competition dataset consists of MIDI recorded from performances of competition participants, bearing expressive dynamics and timing on the granularity of < 10 miliseconds. Discretizing time on a fixed grid will yield sequences that are too long. We adopt a sparse, MIDI-like, event-based representation from (Simon & Oore, 2017). As position in sequence no longer corresponds to time, a priori it is not obvious that relative attention should work as well with such a representation as with a dense representation with quantized time steps. However, we will show in Section 4.2 that it does improve perplexity and sample quality over strong baselines.

We speculate that idiomatic piano gestures such as scales, arpeggios and other motifs all exhibit a certain grammar and recur periodically, hence knowing their relative positional distances makes it easier to model this regularity. This inductive bias towards learning relational information, as opposed to patterns based on absolute position, suggests that the Transformers with relative attention could generalize beyond the lengths it was trained on, which our experiments in Section 4.2.1 confirm.

1.1 Contributions

We show the first successful use of Transformer in generating music that exhibits long-term structure. Our model achieves state-of-the-art perplexity on the Piano-e-Competition data set. The relative attention mechanism is essential to the model's quality. In listening tests (see Section 4.2.3) samples from models with relative self-attention were perceived as more coherent than the baseline Transformer model from Vaswani et al. (2017).

The space complexity of the relative self attention mechanism in its original formulation (Shaw et al., 2018) made it infeasible to train on sequences of sufficient length to capture long-range structure in longer musical compositions. Addressing this we present a crucial algorithmic improvement to the relative self attention mechanism, dramatically reducing its memory requirements from $O(L^2D)$ to $O(LD)$. For example, we reduce the memory consumption from 8.5 GB to 4.2 MB for a sequence of length $L = 2048$ and hidden-state size $D = 512$, allowing us to use GPUs to train the relative self-attention Transformer on long sequences.

2 Related work

Sequence models have been the canonical choice for modeling music, from Hidden Markov Models to RNNs and Long Short Term Memory networks (Eck & Schmidhuber, 2002; Liang, 2016; Simon & Oore, 2017), to bidirectional LSTMs (Hadjeres & Pachet, 2016). Successful application of sequential models to polyphonic music often requires serializing the musical score or performance into a single sequence, for example by interleaving different instruments or voices. Alternatively, a 2D pianoroll-like representation (see A.1 for more details) can be decomposed into a sequence of multi-hot pitch vectors, and their joint probability distributions can be captured using Restricted Boltzmann Machines (Smolensky, 1986; Hinton et al., 2006) or Neural Autoregressive Distribution Estimators. (NADE; Larochelle & Murray, 2011). Pianorolls are also image-like and can be modeled by CNNs trained either as generative adversarial networks (e.g., Dong et al., 2017) or as orderless NADEs (Uria et al., 2014; 2016; Huang et al., 2017).

Lattner et al. (2016) use self-similarity in style-transfer fashion, where the self-similarity structure of a piece serves as a template objective for gradient descent to impose similar repetition structure on an input score. Self-attention can be seen as a generalization of self-similarity; the former maps the input through different projections to queries and keys, and the latter uses the same projection for both.

Dot-product self-attention is the mechanism at the core of the Transformer, and several recent works have focused on applying and improving it for image generation, speech, and summarization (Parmar et al., 2018; Povey et al., 2018; Liu et al., 2018).

A key challenge encountered by each of these efforts is scaling attention computationally to long sequences. This is because the time and space complexity of self-attention grows quadratically in the sequence length. For relative self-attention (Shaw et al., 2018) this is particularly problematic as the space complexity also grows linearly in the dimension, or depth, of the per-position representations.

3 MODEL

3.1 DATA REPRESENTATION

We take a language-modeling approach to training generative models for symbolic music. Hence we represent music as a sequence of discrete tokens, with the vocabulary determined by the dataset. Datasets in different genres call for different ways of serializing polyphonic music into a single stream and also discretizing time.

The JSB Chorale dataset consists of four-part scored choral music, which can be represented as a matrix where rows correspond to voices and columns to time discretized to sixteenth notes. The matrix's entries are integers that denote which pitch is being played. This matrix can than be serialized in raster-scan fashion by first going down the rows and then moving right through the columns (see A.1 for more details).

Compared to JSB Chorale, the piano performance data in the Piano-e-Competition dataset includes expressive timing information at much finer granularity and more voices. For the Piano-e-Competition we therefore use the performance encoding proposed by Simon & Oore (2017) which consists of a vocabulary of 128 `NOTE_ON` events, 128 `NOTE_OFF`s, 100 `TIME_SHIFT`s allowing for expressive timing at 10ms and 32 `VELOCITY` bins for expressive dynamics (see A.2 for more details).

3.2 BACKGROUND: SELF-ATTENTION IN TRANSFORMER

The Transformer decoder is a autoregressive generative model that uses primarily self-attention mechanisms, and learned or sinusoidal position information. Each layer consists of a self-attention sub-layer followed by a feedforward sub-layer.

The attention layer first transforms a sequence of L D-dimensional vectors $X = (x_1, x_2, \ldots, x_L)$ into queries $Q = XW^Q$, keys $K = XW^K$, and values $V = XW^V$, where W^Q, W^K, and W^V are each $D \times D$ square matrices. Each $L \times D$ query, key, and value matrix is then split into H $L \times D_h$ parts or attention heads, indexed by h, and with dimension $D_h = \frac{D}{H}$, which allow the model to focus on different parts of the history.

The scaled dot-product attention computes a sequence of vector outputs for each head as

$$Z^h = \text{Attention}(Q^h, K^h, V^h) = \text{Softmax}\left(\frac{Q^h K^{h\top}}{\sqrt{D_h}}\right) V^h. \tag{1}$$

The attention outputs for each head are concatenated and linearly transformed to get Z, a L by D dimensional matrix. A upper triangular mask ensures that queries cannot attend to keys later in the sequence. For other details of the Transfomer model, such as residual connections and learning rates, the reader can refer Vaswani et al. (2017).

The feedforward (FF) sub-layer then takes the output Z from the previous attention sub-layer, and performs two layers of point-wise dense layers on the depth D dimension, as shown in Equation 2. W_1, W_2, b_1, b_2 are weights and biases of those two layers.

$$\text{FF}(Z) = \text{ReLU}(ZW_1 + b_1)W_2 + b_2 \tag{2}$$

3.3 RELATIVE POSITIONAL SELF-ATTENTION

As the Transformer model relies solely on positional sinusoids to represent timing information, Shaw et al. (2018) introduced relative position representations to allow attention to be informed by how far

two positions are apart in a sequence. This involves learning a separate relative position embedding E^r of shape (H, L, D_h), which has an embedding for each possible pairwise distance m-l between a query and key in position m and l respectively. The distances thus range from 0 to $-L+1$, and are learned separately for each head. In Shaw et al. (2018), the relative embeddings interact with queries and give rise to a S^{rel}, an $L \times L$ dimensional logits matrix which modulates the attention probabilities for each head as:

$$\text{RelativeAttention} = \text{Softmax}\left(\frac{QK^\top + S^{rel}}{\sqrt{D_h}}\right)V. \tag{3}$$

We dropped head indices for clarity. Our work uses the same approach to infuse relative distance information in the attention computation, while significantly improving upon the memory footprint for computing S^{rel}.

For each head, Shaw et al. (2018) instantiate an intermediate tensor R of shape (L, L, D_h), containing the embeddings that correspond to the relative distances between all keys and queries. Q is then reshaped to an $(L, 1, D_h)$ tensor, and $S^{rel} = QR^\top$ is computed with a batch matrix–vector product.[2] This incurs a total space complexity of $O(L^2D)$, restricting its application to long sequences.

3.4 Memory efficient implementation of relative position-based attention

We improve the implementation of relative attention by reducing its intermediate memory requirement from $O(L^2D)$ to $O(LD)$. We observe that all of the terms we need from $QR^\top$ are already available if we directly multiply Q with E^r, the relative embedding matrix. (Table 1). As Q is indexed by absolute query positions l and E^r by relative distances m-l, the result is a absolute-by-relative indexed tensor. However, the rows in $QK^\top$ are indexed by absolute query positions l and columns indexed by absolute key positions m, to compute S^{rel}, we need to "skew" $QE^{r\top}$ so that the relative distance terms are added to the query-by-key entries that bear that relative distance. In the next section, we describe our skewing procedure in detail.

Table 1: Our method reduces the dominating memory complexity while maintaining time complexity.

Implementation	Relative memory	Attention memory	Time Complexity
Shaw et al. (2018)	$O(L^2D + L^2)$	$O(L^2)$	$O(L^2D)$
Ours	$O(LD + L^2)$	$O(L^2)$	$O(L^2D)$

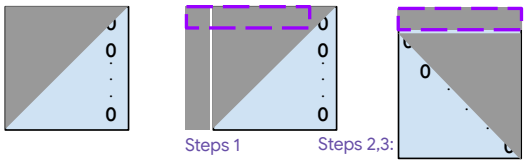


Figure 1: Steps for "skewing" an absolute-by-relative indexed matrix into absolute-by-absolute. Grey indicates self-attention masks or dummy entries introduced by the skewing procedure. Positions with relative distance zero are marked. Entries in the purple dotted rectangle will be removed.

3.4.1 The "skewing" procedure

Hence, we propose a "skewing" procedure to transform an absolute-by-relative indexed matrix to into an absolute-by-absolute indexed matrix. For simplicity the matrices illustrated in Figure 1 correspond to skewing the last two dimensions of QE, which is an absolute-by-relative indexed matrix is of shape (L, L), rows indexed by query positions l, and columns indexed by the pairwise distances between query and key positions $m - l$. Hence its leftmost column corresponds to $L + 1$ time steps back and the rightmost column corresponds to 0 time steps back. The goal is to skew this matrix

[2]We assume that the batch size is 1 here. With a batch size of B, Q would be reshaped to (L, B, D_h) and S^{rel} would be computed with a batch matrix–matrix product.

into one that is absolute-by-absolute with rows indexed by queries, columns by keys. The "skewing" procedure is as follows, illustrated in Figure 1:

1. Pad a dummy column vector of length L before the leftmost column.

2. Reshape the matrix to have shape $(L+1, L)$. (This step assumes NumPy-style row-major ordering.)

3. Slice that matrix to retain only the last l rows and all the columns, resulting in a (L, L) matrix again, but now absolute-by-absolute indexed, which is the S^{rel} that we need.

3.5 Relative local attention

For very long sequences, the quadratic memory requirement of even baseline Transformer is impractical. Local attention has been used for example in Wikipedia generation (Liu et al., 2018; Parmar et al., 2018) by chunking the input sequence into non-overlapping blocks. Each block then attends to itself and the one before, as shown by the smaller thumbnail on the top right corner of Figure 2. To extend relative attention to the local case, we first note that the right block has the same configuration as in the global case (see Figure 1) but much smaller: $(L/M)^2$ (where M is the number of blocks, and N be the resulting block length) as opposed to L^2. The left block is unmasked with relative indices running from -1 (top right) to -$2N + 1$ (bottom left). Hence, the learned E^r for the local case is of shape $(2N - 1, N)$. Similar to the global case, we first compute $QE^{r\top}$ and then use the following procedure to skew it to have the same indexing as $QK^\top$, as illustrated in Figure 2.

1. Pad a dummy column vector of length N after the rightmost column.

2. Flatten the matrix and then pad with a dummy row of length $N - 1$.

3. Reshape the matrix to have shape $(N + 1, 2N - 1)$.

4. Slice that matrix to retain only the first N rows and last N columns, resulting in a (N, N) matrix.

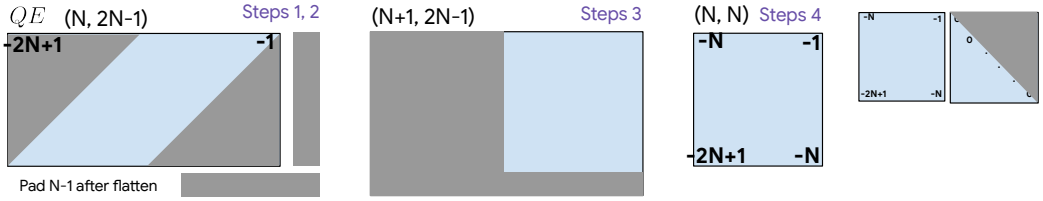


Figure 2: Steps for "skewing" an $l \times (m - l)$ matrix with $2N - 1$ ranged relative indices into $l \times m$.

4 Experiments

4.1 J.S. Bach Chorales

J.S. Bach chorales is a canonical dataset used for evaluating generative models for music [3] (e.g., Allan & Williams, 2005; Boulanger-Lewandowski et al., 2012; Liang, 2016; Hadjeres et al., 2016; Huang et al., 2017). It consists of score-based four-part chorales. We first discretize the scores onto a 16th-note grid, and then serialize it by iterating through all the voices within a time step and then advancing time (see A.1 for more details). As there is a direct correspondence between position in sequence and position on the timing/instrument grid in a piece, adding relative position representations could make it easier to learn this grammar. We indeed see relative attention drastically improve negative log-likelihood (NLL) over baseline Transformer (Table 2). This improvement is also reflected in sample quality. The samples now maintain the necessary timing/instrument grid, always advancing four steps before advancing in time. As local timing is maintained, the model is able to capture timing on a more global level, giving rise to regular phrasing, as shown in Figure 3.

[3] J.S. Bach chorales dataset: `https://github.com/czhuang/JSB-Chorales-dataset`

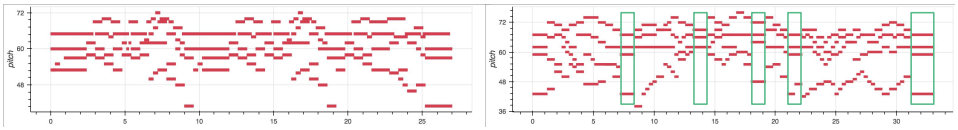


Figure 3: Unconditioned samples from Transformer without (left) and with (right) relative self-attention. Green vertical boxes indicate the endings of (sub)phrases where cadences are held.

In addition to relative attention, we also explored enhancing absolute timing through concatenating instead of adding the sinusoids to the input embeddings. This allows the model to more directly learn its absolute positional mapping. This further improves performance for both the baseline and relative transformer (Table 2). We compare against COCONET as it is one of the best-performing models that has also been evaluated on the 16-note grid using the canonical dataset split. To directly compare, we re-evaluated COCONET to obtain note-wise losses on the validation set [4]. For the Transformer models (abbreviated as TF), we implemented our attention mechanisms in the Tensor2Tensor framework (Vaswani et al., 2018). We use 8 heads, and keep the query, key (att) and value hidden size (hs) fixed within a config. We tuned number of layers (L in {4,5,6}), attention hidden size (att in {256, 512}) and pointwise feedforward hidden size (ff in {512, 1024}).

4.1.1 GENERALIZING RELATIVE ATTENTION TO CAPTURE RELATIONAL INFORMATION

A musical event bears multiple attributes, such as timing, pitch, instrument etc. To capture more relational information, we extend relative attention to capture pairwise distances on additional attributes. We learn separate relative embeddings can be learned for timing E_t and also pitch E^p. E^t has entries corresponding to how many sixteenth notes apart are two positions, while E^p embeds the pairwise pitch interval. However this approach is not directly scalable beyond J.S. Bach Chorales because it involves explicitly gathering relative embeddings for R^t and R^p, resulting in a memory complexity of $O(L^2D)$ as in Shaw et al. (2018). This is due to relative information being computed based on content as opposed to content-invariant information such as position in sequence. It was sufficient to add the extra timing signals to the first layer, perhaps because is closest to the raw input content. Here, the relative logits are computed from three terms, $S^{rel} = \text{Skew}(QE^r) + Q(R^t + R^p)$ in contrast with other layers that only have one term, $\text{Skew}(QE^r)$.

4.2 PIANO-E-COMPETITION

We use the first 6 years of of Piano-e-Competition because these years have corresponding MIDI data released [5], resulting in about 1100 pieces, split 80/10/10. Each piece is MIDI data capturing a classical piano performance with expressive dynamics and timing, encoded with the MIDI-like representation described in Section A.2. We trained on random crops of 2000-token sequences and employed two kinds of data augmentation: pitch transpositions uniformly sampled from $\{-3, -2, \ldots, 2, 3\}$ half-steps, and time stretches uniformly sampled from the set $\{0.95, 0.975, 1.0, 1.025, 1.05\}$.

We compare to Magenta's PerformanceRNN (LSTM, which first used this dataset) (Simon & Oore, 2017) and LookBack RNN (LSTM with attention) (Waite, 2016). LookBack RNN uses an input representation that requires monophonic music with barlines which is information that is not present in performed polyphonic music data, hence we simply adopt their architecture. Table 3 shows that Transformer-based architectures fits this dataset better than LSTM-based models.

We implemented our attention mechanisms in the Tensor2Tensor framework (Vaswani et al., 2018), and use the default hyperparameters for training, with 0.1 learning rate, 0.1 dropout, and early stopping. We compare four architectures, varying on two axes: global versus local, and regular versus relative attention. We found that reducing the query and key hidden size (att) to half the hidden size

[4] Some earlier papers report frame-wise losses to compare to models such as RNN-RBM which model "chords". Coconet can be evaluated under note-wise or frame-wise losses.

[5] Piano-e-Competition dataset (competition history): `http://www.piano-e-competition.com/`

[6] COCONET is an instance of OrderlessNADE, an ensemble over orderings. The chronological loss evaluates the model as autoregressive, from left to right. We can also evaluate the model as a mixture, by averaging its losses over multiple random orderings. This is a lower bound on log-likelihood. It is intractable to sample from exactly but can be approximated through Gibbs sampling.

Table 2: Note-wise validation NLL: J.S.Bach Chorales at 16th notes.

Model variation	Validation NLL
COCONET (CNN, chronological, 64L, 128 3x3f)	0.436
COCONET (CNN, orderless, 64L, 128 3x3f)	≤ 0.238 [6]
Transformer (TF) baseline (Vaswani et al., 2017) (5L, 256hs, 256att, 1024ff, 8h)	0.417
TF baseline + concat positional sinusoids (cps)	0.398
TF baseline + concat positional sinusoids, instrument labels (cpsi)	0.370
Relative Transformer (Shaw et al., 2018) (5L, 512hs, 512att, 512ff, 256r, 8h)	0.357
Relative Transformer + cpsi	0.347
Relative Music Transformer + cpsi	0.335

Table 3: Validation NLL for Piano-e-Competition dataset, with event-based representation

Model variation	Validation NLL
PERFORMANCE RNN (LSTM) (3L, 1024hs)	1.969
LSTM with attention (3L, 1024hs, 1024att)	1.959
Transformer (TF) baseline (6L, 256hs, 512att, 2048fs, 1024r, 8h)	1.861
TF with local attention (Liu et al., 2018) (8L, 1024fs, 512bs)	1.863
TF with relative global attention (our efficient formulation) (6L, 2048fs, 1024r)	**1.835**
TF with relative local attention (ours) (6L, 1024fs, 2048r, 512bs)	**1.840**

(hs) works well and use this relationship for all of the models, while tuning on number of layers (L) and filter size (fs). We use block size (bs) 512 for local attention. We set the maximum relative distance to consider to half the training sequence length for relative global attention, and to the full memory length (which is two blocks) for relative local attention. Table 3 show that relative attention (global or local) outperforms regular self-attention (global or local). All else being equal, local and global attention perform similarly. Each though local attention does not see all the history at once, it can build up a larger receptive field across layers. This can be an advantage in the future for training on much longer sequences, as local attention requires much less memory.

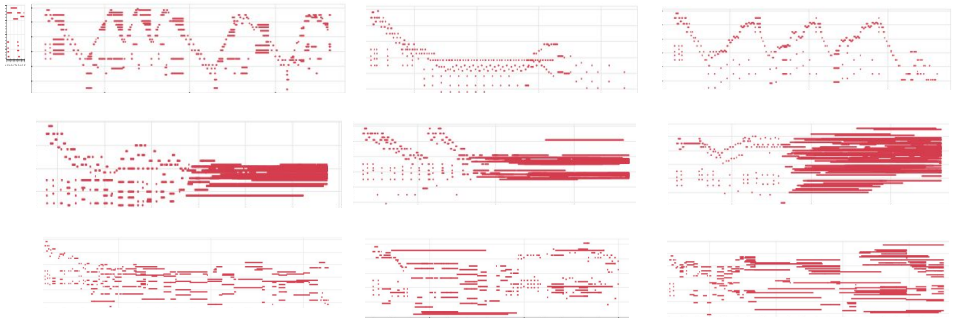


Figure 4: Comparing how models continue a prime (top left). Repeated motives and structure are seen in samples from Transformer with relative attention (top), but less so for baseline Transformer (middle) and PerformanceRNN (LSTM).

4.2.1 QUALITATIVE PRIMING EXPERIMENTS

When primed with an initial motif (Chopin's Étude Op. 10, No. 5, known as Black Key Black Étude) shown in the top left corner of Figure 4, we see the models perform qualitatively differently. Transformer with relative attention elaborates the motif and creates phrases with clear contour which are repeated and varied. Baseline Transformer uses the motif in a more uniform fashion, while LSTM uses the motif initially but soon drifts off to other material. Note that the generated samples are twice as long as the training sequences. Relative attention was able to generalize to lengths longer

than trained but baseline Transformer deteriorates beyond its training length. See Appendix C for visualizations of how the Relative Transformer attends to past motifs.

4.2.2 CONDITIONING ON MELODY

To explore the sequence-to-sequence setup of Transformers, we experimented with a conditioned generation task where the encoder takes in a given melody and the decoder has to realize the entire performance, i.e. melody plus accompaniment. The melody is encoded as a sequence of tokens as in Waite (2016), quantized to a 100ms grid, while the decoder performance uses the encoding described in Section 3.1 (and further illustrated in A.2). We use relative attention on the decoder side and demonstrate that it also improves performance, as shown in Table 4.

Table 4: Validation conditional NLL given ground-truth melody for Piano-e-Competition dataset.

Model variation	Validation NLL [7]
Baseline Transformer	2.066
Relative Transformer with our efficient formulation	1.786

4.2.3 HUMAN EVALUATIONS

To compare the perceived sample quality of models trained on the Piano-e-Competition dataset, and their ability to generate a continuation for a priming sequence, we carried out a listening test study comparing the baseline/vanilla Transformer, Transformer with relative-attention, PerformanceRNN, and validation set Piano-e-Competition. Participants were presented with two musical excerpts (from two different models that were given the same priming sequence) and asked to rate which one is more musical on a Likert scale. 180 such ratings were collected, with each source involved in 30 pairwise comparisons and each comparison completed by 3 different participants.

Figure 5 shows the number of comparisons in which an excerpt from each model was selected as more musical. The improvement in sample quality from using relative attention over the baseline Transformer model was statistically significant (See Appendix B for the analysis). LSTMs performed better in the study than the Transformer, despite having higher perplexity. This may be due to samples being short (10s to 20s), weakening the comparison on long-term structure.

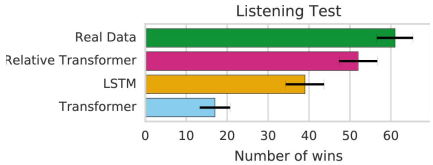


Figure 5: Results of our listening tests, showing the number of times each model/real data won in a pairwise comparison. Black error bars indicate estimated standard deviation of means.

5 CONCLUSION

In this work we demonstrated that the Transformer equipped with relative attention is very well-suited for generative modeling of symbolic music. The quality of samples from our model leaves us enthusiastic about this direction of research. Moreover, the ability to expand upon a primer, in particular, suggests potential applications as creative tool.

The significant improvement from relative attention highlights a shortcoming of the original Transformer that might also limit its performance in other domains. Improving the Transformer's ability to capture periodicity at various time scales, for instance, or relations between scalar features akin to pitch could improve time-series models. Our memory-efficient implementation enables the application of relative attention to much longer sequences such as long texts or even audio waveforms, which significantly broadens the range of problems to which it could be applied.

REFERENCES

Moray Allan and Christopher KI Williams. Harmonising chorales by probabilistic inference. *Advances in neural information processing systems*, 17:25–32, 2005.

Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. *International Conference on Machine Learning*, 2012.

Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. Musegan: Symbolic-domain music generation and accompaniment with multi-track sequential generative adversarial networks. *arXiv preprint arXiv:1709.06298*, 2017.

Douglas Eck and Juergen Schmidhuber. Finding temporal structure in music: Blues improvisation with lstm recurrent networks. In *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*, pp. 747–756. IEEE, 2002.

Gaëtan Hadjeres and François Pachet. Deepbach: a steerable model for bach chorales generation. *arXiv preprint arXiv:1612.01010*, 2016.

Gaëtan Hadjeres, Jason Sakellariou, and François Pachet. Style imitation and chord invention in polyphonic music with exponential families. *arXiv preprint arXiv:1609.05152*, 2016.

Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

Cheng-Zhi Anna Huang, Tim Cooijmans, Adam Roberts, Aaron Courville, and Doug Eck. Counterpoint by convolution. In *International Conference on Music Information Retrieval*, 2017.

Hugo Larochelle and Iain Murray. The neural autoregressive distribution estimator. In *AISTATS*, volume 1, pp. 2, 2011.

Stefan Lattner, Maarten Grachten, and Gerhard Widmer. Imposing higher-level structure in polyphonic music generation using convolutional restricted boltzmann machines and constraints. *arXiv preprint arXiv:1612.04742*, 2016.

Feynman Liang. Bachbot: Automatic composition in the style of bach chorales. *Masters thesis, University of Cambridge*, 2016.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*, 2018.

Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*, 2016.

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, and Alexander Ku. Image transformer. *arXiv preprint arXiv:1802.05751*, 2018.

Daniel Povey, Hossein Hadian, Pegah Ghahremani, Ke Li, and Sanjeev Khudanpur. A time-restricted self-attention layer for asr. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5874–5878. IEEE, 2018.

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

Ian Simon and Sageev Oore. Performance rnn: Generating music with expressive timing and dynamics. `https://magenta.tensorflow.org/performance-rnn`, 2017.

Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. Technical report, DTIC Document, 1986.

Benigno Uria, Iain Murray, and Hugo Larochelle. A deep and tractable density estimator. In *International Conference on Machine Learning*, pp. 467–475, 2014.

Benigno Uria, Marc-Alexandre Côté, Karol Gregor, Iain Murray, and Hugo Larochelle. Neural autoregressive distribution estimation. *arXiv preprint arXiv:1605.02226*, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 6000–6010, 2017.

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. Tensor2tensor for neural machine translation. *CoRR*, abs/1803.07416, 2018. URL `http://arxiv.org/abs/1803.07416`.

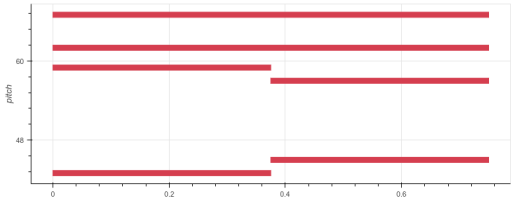
Elliot Waite. Generating long-term structure in songs and stories. `https://magenta.tensorflow.org/2016/07/15/lookback-rnn-attention-rnn`, 2016.

A DOMAIN-SPECIFIC REPRESENTATIONS

Adapting sequence models for music requires making decisions on how to serialize a polyphonic texture. The data type, whether score or performance, makes certain representations more natural for encoding all the information needed while still resulting in reasonable sequence lengths.

A.1 SERIALIZED INSTRUMENT/TIME GRID (J.S.BACH CHORALES)

The first dataset, J.S. Bach Chorales, consists of four-part score-based choral music. The time resolution is sixteenth notes, making it possible to use a serialized grid-like representation. Figure 6 shows how a pianoroll (left) can be represented as a grid (right), following (Huang et al., 2017). The rows show the MIDI pitch number of each of the four voices, from top to bottom being soprano (S), alto (A), tenor (T) and bass (B), while the columns is discretized time, advancing in sixteenth notes. Here longer notes such as quarter notes are broken down into multiple repetitions. To serialize the grid into a sequence, we interleave the parts by first iterating through all the voices at time step 1, and then move to the next column, and then iterate again from top to bottom, and so on. The resulting sequence is $S_1A_1T_1B_1S_2A_2T_2B_2$..., where the subscript gives the time step. After serialization, the most common sequence length is 1024. Each token is represented as onehot in pitch.




```
S: 67, 67, 67, 67
A: 62, 62, 62, 62
T: 59, 59, 57, 57
B: 43, 43, 45, 45
```

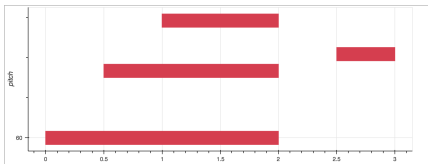
Figure 6: The opening measure of BWV 428 is visualized as a pianoroll (left, where the x-axis is discretized time and y-axis is MIDI pitch number), and encoded in grid representation with sixteenth note resolution (right). The soprano and alto voices have quarter notes at pitches G4 (67) and D4 (62), the tenor has eighth notes at pitches B3 (59) and A3 (57), and the bass has eighth notes at pitches A2 (45) and G2 (43).

A.2 MIDI-LIKE EVENT-BASED (PIANO-E-COMPETITION)

The second dataset, Piano-e-Competition, consists of polyphonic piano performances with expressive timing and dynamics. The time resolution here is on the millisecond level, so a grid representation would result in sequences that are too long. Instead, the polyphonic performance is serialized into a sequence of one hot encoded events as proposed in (Simon & Oore, 2017).

First, the input MIDI files are preprocessed to extend note durations based on sustain pedal control events. The sustain pedal is considered to be down whenever a sustain control change is encountered with a value $>= 64$; the sustain pedal is then considered up after a control change with a value < 64. Within a period where the sustain pedal is down, the duration of each note is extended to either the beginning of the next note of the same pitch or the end of the sustain period, whichever happens first. If the original duration extends beyond the time when the sustain pedal is down, that original duration is used.

Next, the MIDI note events are converted into a sequence from the following set of vocabulary: 128 `NOTE_ON` events for starting a note of with one of the 128 MIDI pitches, 128 `NOTE_OFF` events for ending a note with one of the 128 MIDI pitches, 100 `TIME_SHIFT` events representing forward time shifts in 10ms increments from 10ms to 1s, and 32 `SET_VELOCITY` events representing the velocity for future `NOTE_ON` events in the form of the 128 possible MIDI velocities quantized into 32 bins. An example performance encoding is illustrated in Figure 7.




```
SET_VELOCITY<80>, NOTE_ON<60>
TIME_SHIFT<500>, NOTE_ON<64>
TIME_SHIFT<500>, NOTE_ON<67>
TIME_SHIFT<1000>, NOTE_OFF<60>, NOTE_OFF<64>,
NOTE_OFF<67>
TIME_SHIFT<500>, SET_VELOCITY<100>, NOTE_ON<65>
TIME_SHIFT<500>, NOTE_OFF<65>
```

Figure 7: A snippet of a piano performance visualized as a pianoroll (left) and encoded as performance events (right, serialized from left to right and then down the rows). A C Major chord is arpeggiated with the sustain pedal active. At the 2-second mark, the pedal is released, ending all of the notes. At the 3-second mark, an F is played for .5 seconds. The C chord is played at velocity 80 and the F is played at velocity 100.

B SUPPLEMENT OF LISTENING TEST

B.0.1 STUDY PROCEDURE

Participants were presented with two musical excerpts that shared a common priming sequence. For each excerpt, the priming sequence was played, followed by 2.5 seconds of silence, followed by the priming sequence again and a continuation of that sequence. The continuations were either sampled from one of the models or extracted from our validation set. We evaluated all possible pairs in the space of data and model samples, except from the same model. Each continuation had a length of 512 events using the encoding described in Section A.2. This corresponds to the length the models were trained on to remove the deteriorating effect that happens with baseline Transformer when asked to generate beyond the length it was trained on. Participants were asked which excerpt they thought was more musical on a Likert scale. 180 such ratings were collected, with each source involved in 30 pair-wise comparisons and each comparison completed by 3 different participants.

B.0.2 ANALYSIS

Further, a Kruskal-Wallis H test of the ratings showed that there was a statistically significant difference between the models: $\chi^2(2) = 63.84, p < 0.001$. A post-hoc analysis using the Wilcoxon signed-rank test with Bonferroni correction showed that participants rated samples from the relative Transformer as more musical than samples from the baseline Transformer with $p < 0.01/6$. We did not observe a statistically significant difference between samples from our validation set and the relative Transformer model or between samples from the relative Transformer and the LSTM. The effects are probably less pronounced between relative Transformer and LSTM because we used samples half the length of those shown in Figure 4 to prevent the baseline Transformer from deteriorating. This weakens the comparison on long-term structure.

C VISUALIZING SOFTMAX ATTENTION

One advantage of attention-based models is that we can visualize its attention distribution 3. This gives us a glimpse of how the model might be building up recurring structures and how far it is attending back. The pianorolls in the visualizations below is a sample generated from Transformer with relative attention. Each figure shows a query (the source of all the attention lines) and previous memories being attended to (the notes that are receiving more softmax probabiliy is highlighted in). The coloring of the attention lines correspond to different heads and the width to the weight of the softmax probability.

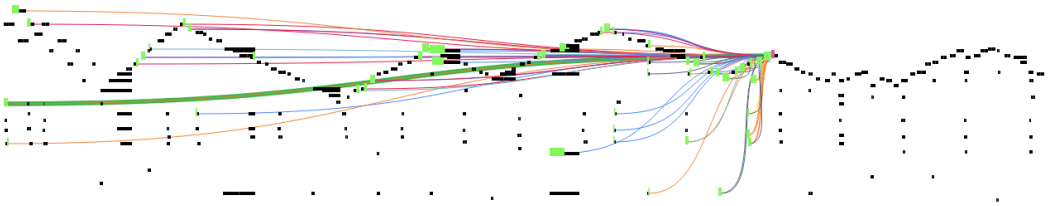


Figure 8: This piece has a recurring triangular contour. The query is at one of the latter peaks and it attends to all of the previous high notes on the peak, all the way to beginning of the piece.

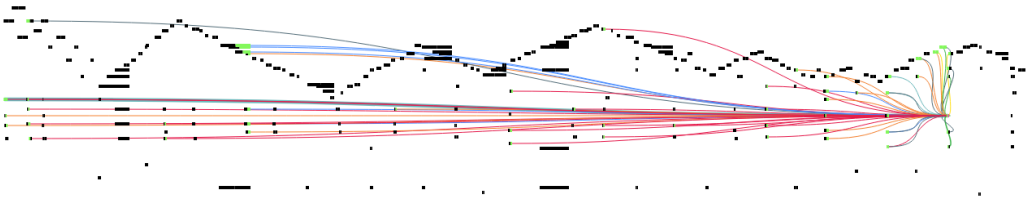


Figure 9: The query a note in the left-hand, and it attends to its immediate past neighbors and mostly to the earlier left hand chords, with most attention lines distributed in the lower half of the pianoroll.