Museformer: Transformer with Fine- and Coarse-Grained Attention for Music Generation

Botao Yu[†], Peiling Lu[‡], Rui Wang[‡],
Wei Hu[†*], Xu Tan[‡*], Wei Ye[§],
Shikun Zhang[§], Tao Qin[‡], Tie-Yan Liu[‡]

[†] Nanjing University [‡] Microsoft Research Asia [§] Peking University [*] Corresponding authors

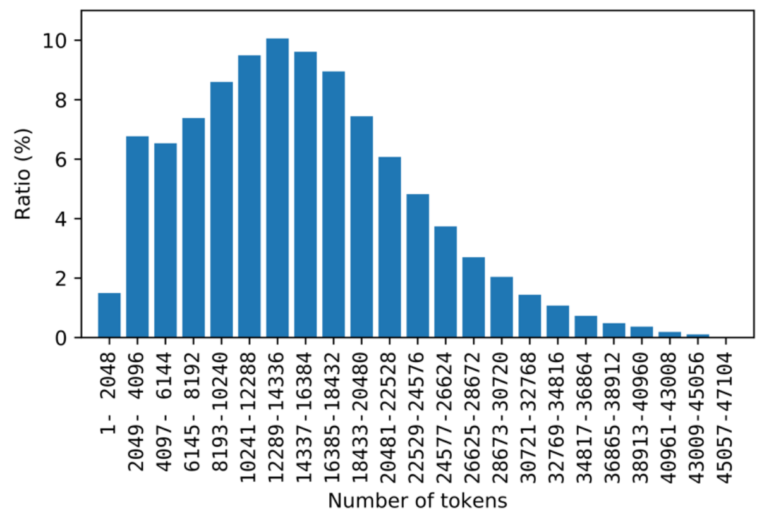
btyu@foxmail.com, {peil,ruiwa,xuta,taoqin,tyliu}@microsoft.com, whu@nju.edu.cn, {wye,zhangsk}@pku.edu.cn

NEURAL INFORMATION PROCESSING SYSTEMS

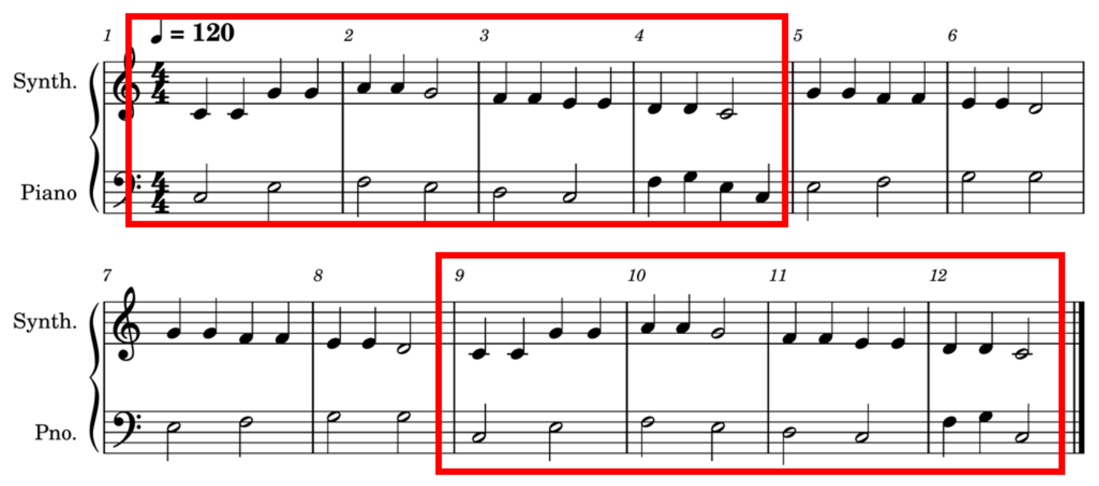
Background

Symbolic music generation aims to compose music scores automatically. Since music is similar to text, Transformer-based models, which have been demonstrated to work well on text generation, are increasingly applied in music generation. However, There are two ubiquitous challenges:

- *Long sequence modeling*: how to model typically long music sequences (over 10,000 tokens) efficiently.
- *Music structure modeling*: how to generate music with realistic music repetition structures. In human-made music, music structures embodied as repetitions and variations are common in music.



Statistics of sequence length.



A music score as an example to show repetition structures.

Although many Transformer variants have been proposed to handle long sequences (the first challenge), they cannot well model the music structures (the second challenge).

Museformer

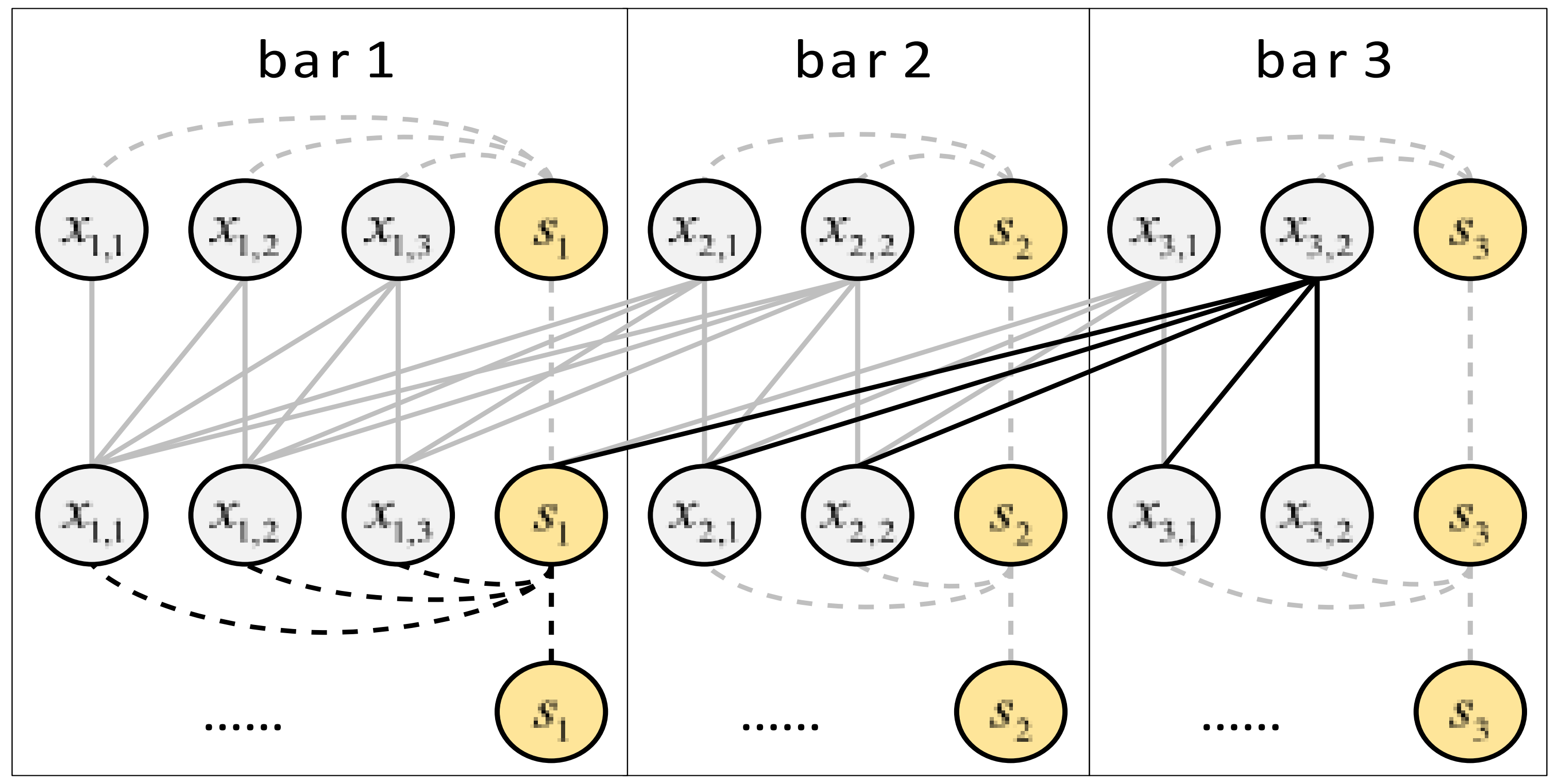
Basically, it follows the architecture of Transformer, and the full attention is replaced by our fine- and coarse-grained attention (FC-Attention).

Fine- and Coarse-Grained Attention (FC-Attention)

The general idea of FC-Attention is straightforward: focus more on the music structure-related content and less on the others. To this end, we first append a summary tokens at the end of each music bar, which is depicted in yellow in the following figure.

There are 2 steps in FC-Attention, namely *summarization* and *aggregation*.

In the summarization step, the local information of each bar summarized onto the corresponding summary token. This is achieved by a standard attention, where the query is each summary token, and the key and the value are the local music tokens within the bar as well as the summary token itself. This process is illustrated with the dashed lines in the figure.


bar 1
bar 2
bar 3
$x_{1,1}$ $x_{1,2}$ $x_{1,3}$ s_1
$x_{2,1}$ $x_{2,2}$ s_2
$x_{3,1}$ $x_{3,2}$ s_3
......

The information flow of FC-Attention. It shows a toy example with 3 music bars, and for each bar, only the previous 1st bar is regarded as the structure-related bar.

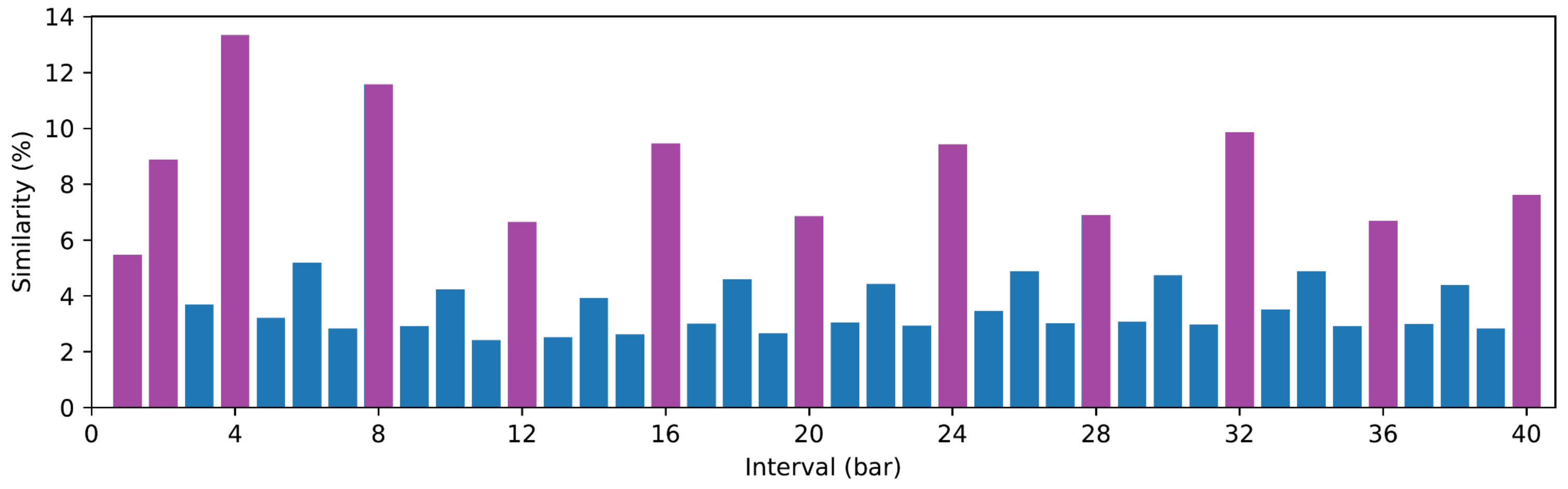
In the aggregation step, there is another scaled dot-product attention, where each music token attends to previous tokens selectively, which shows the idea of fine- and coarse-grained attention.

- *Fine-grained attention*: Each music token only directly attends to those music tokens in the structure-related bars, which are the bars that tend to be repeated by the current bar being generated.
- *Coarse-grained attention*: Each music token only attends to their summary tokens to get a sketch of them.

This process is illustrated with the solid lines in the figure.

Structure-related Bar Selection

The structure-related bars are determined by similarity statistics over human-made music. Specifically, we calculate the average similarities of bar pairs w.r.t. their intervals. The following figure visualizes the distribution.



The similarity distribution of the melody track of the training data.

The similarity statistics on various datasets show that music structure generally has a periodical pattern: a music bar tends to be more similar to its previous 2 bars, and also to the previous 4-th bar or its multipliers in most cases. Based on this pattern, we select 8 bars as the default structure-related bars: the previous 1st, 2nd, 4th, 8th, 12th, 16th, 24th, 32nd bars.

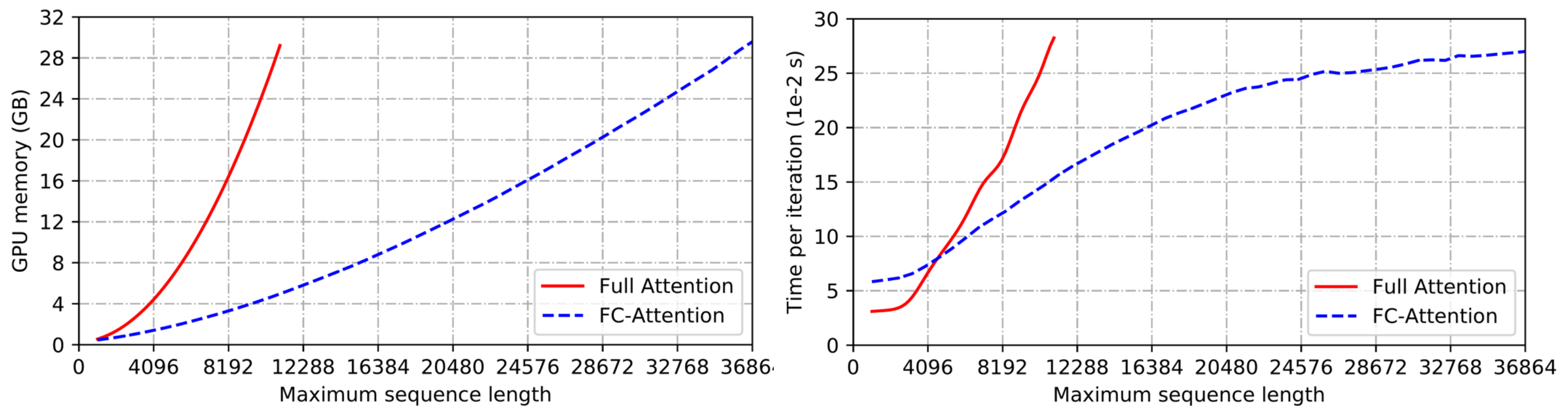
Experiments

	PPL (1,024)	PPL (5,120)	PPL (10,240)	SE (%)
Music Transformer	1.66	1.77	2.55	2.49
Transformer-XL	**1.64**	1.45	1.43	15.66
Longformer	1.65	1.46	1.45	5.25
Linear Transformer	1.86	1.67	1.64	1.97
Museformer (ours)	**1.64**	**1.41**	**1.35**	**0.95**
w/o coarse-grained	1.65	1.42	1.38	1.08
w/o bar selection	1.65	1.43	1.39	6.39

The results of objective evaluation and ablation study.

	Musicality	ST structure	LT structure	Overall	Pref
Music Transformer	6.00 ± 2.21	6.90 ± 1.76	5.30 ± 2.58	5.90 ± 1.90	0.20
Transformer-XL	6.10 ± 2.19	7.40 ± 1.81	6.26 ± 2.78	6.44 ± 2.01	0.34
Longformer	6.46 ± 1.81	7.60 ± 1.47	6.18 ± 2.54	6.44 ± 1.72	0.24
Linear Transformer	6.06 ± 1.99	6.92 ± 2.03	5.78 ± 2.64	6.30 ± 1.84	0.24
Museformer (ours)	**6.88 ± 1.95**	**7.86 ± 1.51**	**6.72 ± 2.74**	**7.12 ± 1.81**	**0.46**

The results of subjective evaluation.



(a) GPU memory consumption.

(b) Running time per iteration.

Efficiency test of Museformer with FC-Attention compared to its full attention counterpart.

Conclusion

Museformer combines a fine-grained attention for learning the structure-related correlations, and a coarse-grained attention for preserving other necessary information. Advantages:

- The combination of the two schemes lets Museformer focus on the important and preserve other information at a low cost.
- The structure-related bar selection explicitly ensure that the repetition structure-related content can be directly attended, which promotes the generation of the music structures.

References

Huang et al. (2018). “Music transformer: Generating music with long-term structure.” In: ICLR.

Dai et al. (2019). “Transformer-XL: Attentive language models beyond a fixed-length context.” In: ACL.

Beltagy et al. (2020). “Longformer: The long-document transformer.” In: arXiv.

Katharopoulos et al. (2020). “Transformers are RNNs: Fast autoregressive transformers with linear attention.” In: ICML.