Generowanie muzyki przy pomocy modelu transformer-GAN trenowanego metodą polityki gradientu

Music generation with transformer based generative adverserial network trained using policy gradient methods

Transformer

[Attention is all you need]

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

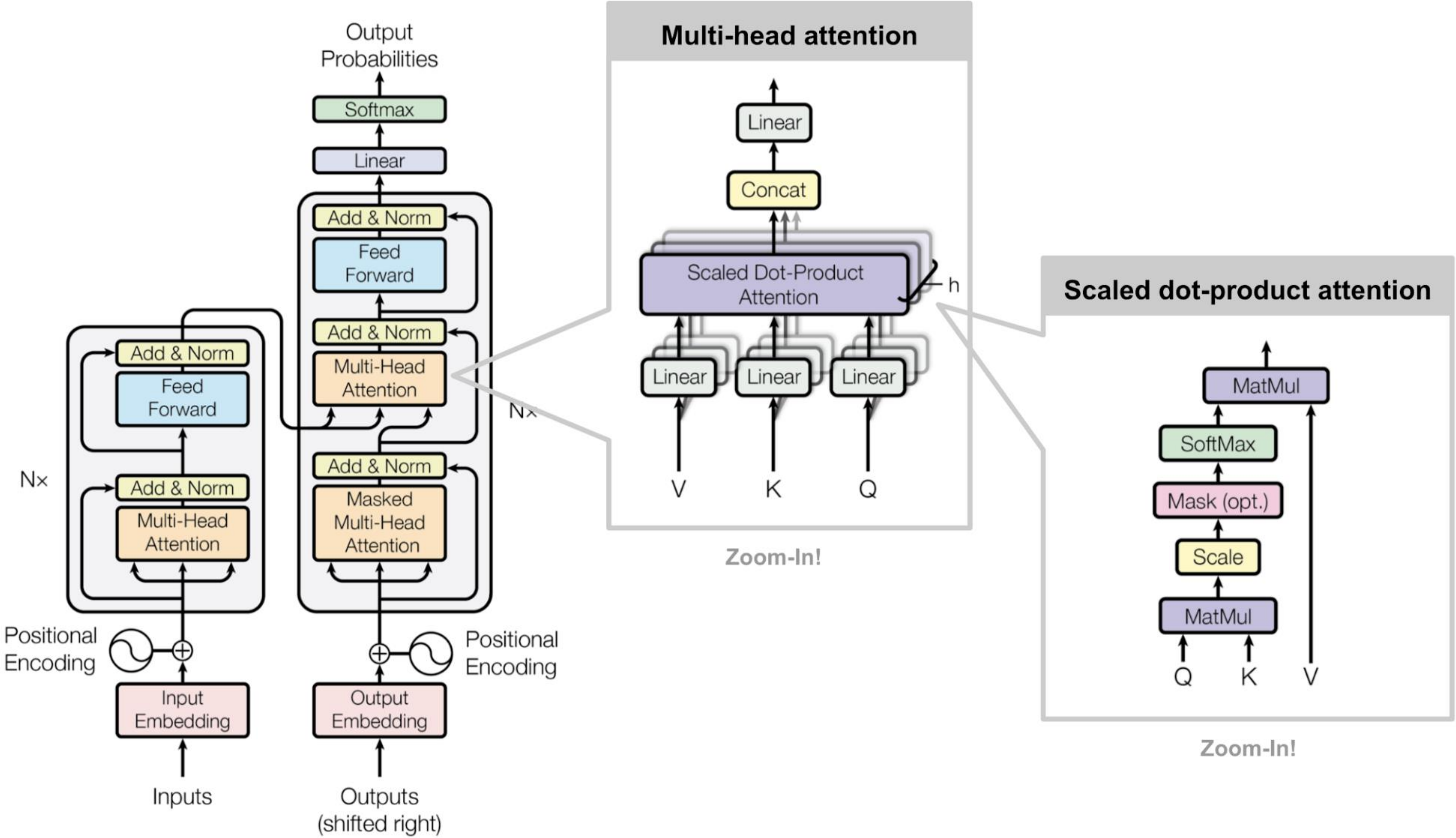
Mechanizm uwagi (attention)

W niebie spotkało się trzech fizyków: Newton, Pascal i Einstein i zaczynają bawić się w chowanego...

i.t.d

RNN, LSTM, GRU

W niebie spotkało się trzech fizyków: Newton, Pascal i Einstein i zaczynają bawić się w chowanego.


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)
Multi-head attention
Linear
Concat
Scaled Dot-Product Attention
h
Linear
Linear
Linear
V
K
Q
Zoom-In!
Scaled dot-product attention
MatMul
SoftMax
Mask (opt.)
Scale
MatMul
Q
K
V
Zoom-In!

MUSIC TRANSFORMER: GENERATING MUSIC WITH LONG-TERM STRUCTURE

Cheng-Zhi Anna Huang* **Ashish Vaswani** **Jakob Uszkoreit** **Noam Shazeer**
Ian Simon **Curtis Hawthorne** **Andrew M. Dai** **Matthew D. Hoffman**
Monica Dinculescu **Douglas Eck**
Google Brain

[Music Transformer]

Poza użyciem architektury transformera w celach generacji muzyki, artykuł używa metody uwagi nazywanej „Relative Attention” czyli wzbogacony tradycyjny algorytm. Polega on na dodaniu możliwości patrzenia na relatywną pozycje tokenu w sekwencji.

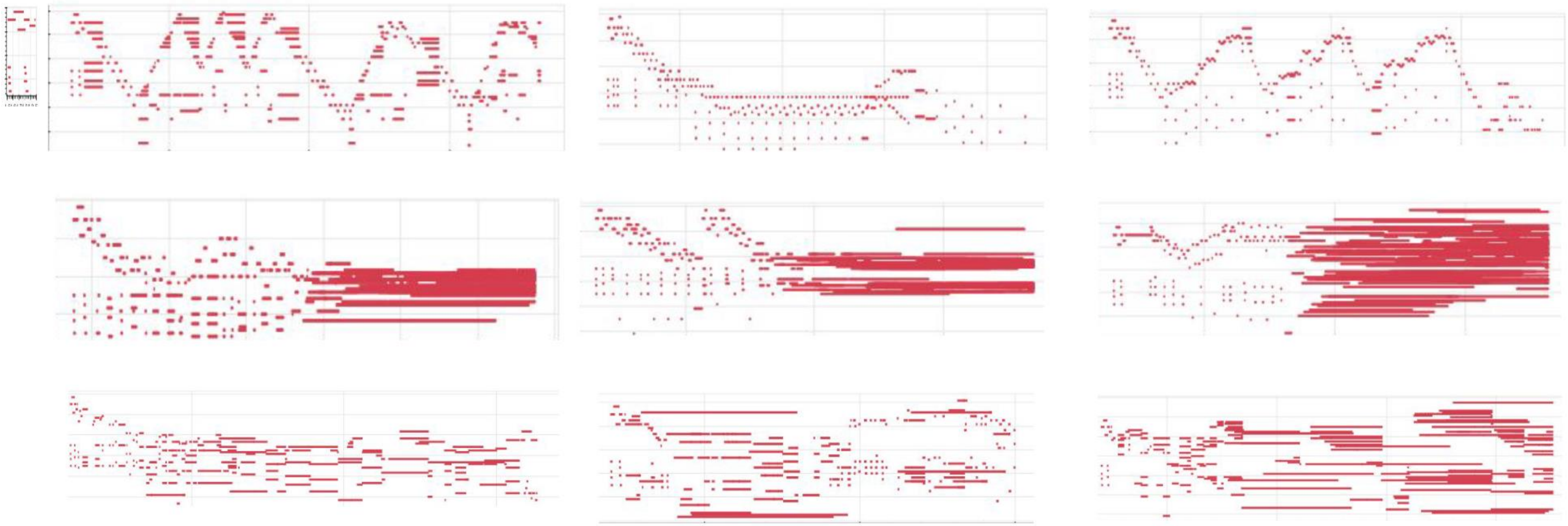
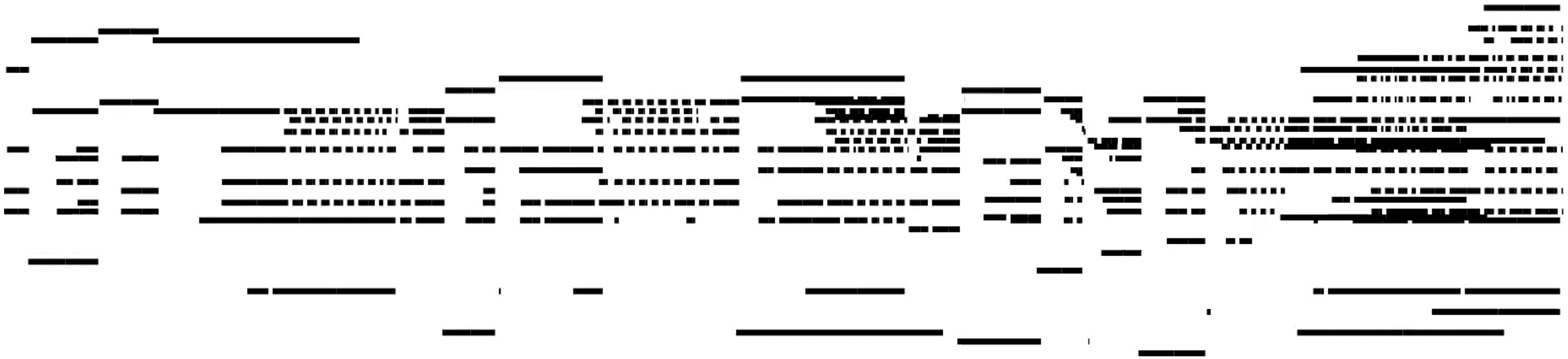


Figure 4: Comparing how models continue a prime (top left). Repeated motives and structure are seen in samples from Transformer with relative attention (top row), but less so from baseline Transformer (middle row) and PerformanceRNN (LSTM) (bottom row).



TTS-GAN: A Transformer-based Time-Series Generative Adversarial Network

Xiaomin Li, Vangelis Metsis, Huangyingrui Wang, Anne Hee Hiong Ngu
x_l30, vmetsis, h_w91, angu @txstate.edu

Texas State University, San Marcos TX 78666, USA

TTS-GAN

Architektura i trening sieci

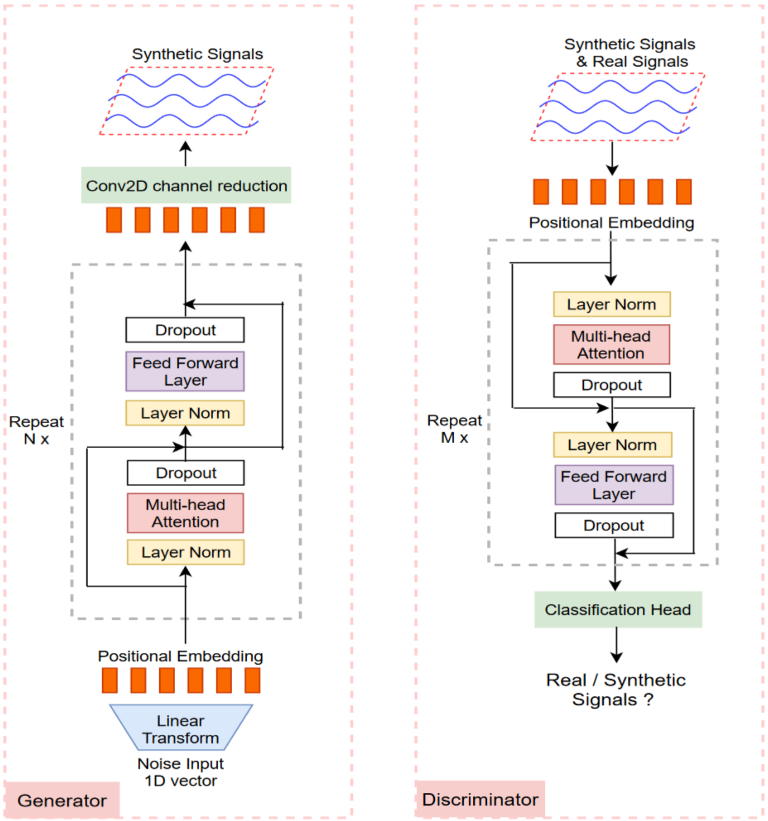
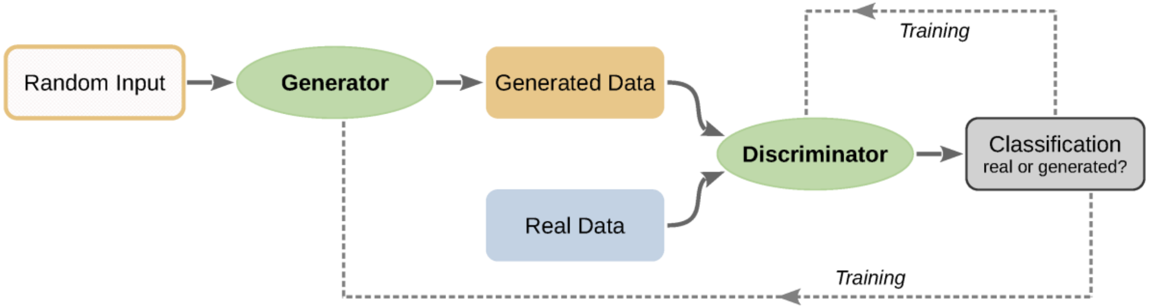


Fig. 1: TTS-GAN model architecture



Trening odbywa się w następujący sposób: z losowego szumu generator nadbudowuje syntetyczne dane, które następnie są przekazywane do dyskryminatora razem z prawdziwymi danymi z odpowiednimi oznaczeniami. Oba modele poprawiają sobie wagi. Trening odbywa się do momentu, w którym dyskryminator nie jest w stanie odróżnić prawdziwych danych od sztucznie wygenerowanych.

Trenowanie przy pomocy gradient polityki

[SeqGAN]

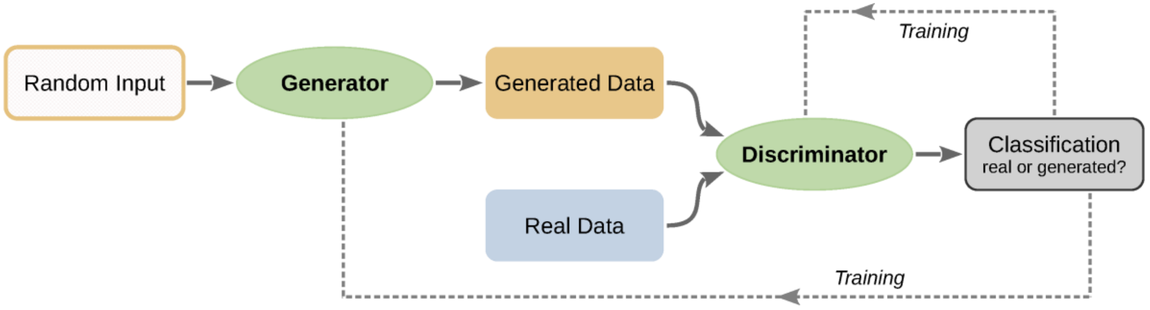
SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient

Lantao Yu[†], Weinan Zhang[†*], Jun Wang[‡], Yong Yu[†]
[†]Shanghai Jiao Tong University, [‡]University College London
{yulantao,wnzhang,yyu}@apex.sjtu.edu.cn, j.wang@cs.ucl.ac.uk

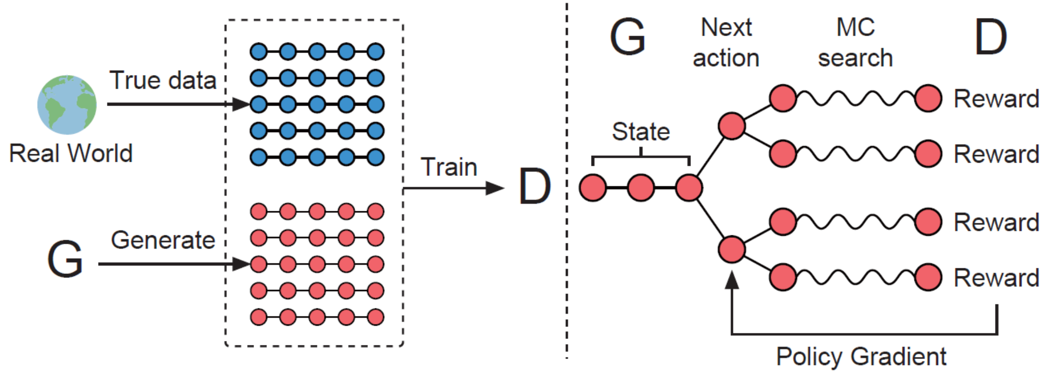
Problem:

Typowy GAN jest trenowany według schematu przedstawionego na rysunku. Niestety w naszym przypadku bloczek *Random Input*, jest sekwencyjne losowanym wektorem na podstawie poprzednich tokenów, co uniemożliwia przeprowadzenie gradientu.

Wykorzystanie polityki gradientowej zaciągniętej z działu RL, pozwala na policzenie gradientu, traktując ciąg dyskretnych tokenów, jako ciąg przyszłych akcji dla danego stanu, maksymalizowania szansa na „oszukanie” modelu dyskryminatora.



Trening GAN-a



Trening SeqGAN-a

Music Generation

For music composition, we use Nottingham[6] dataset as our training data, which is a collection of 695 music of folk tunes in midi file format. We study the solo track of each music. In our work, we use 88 numbers to represent 88 pitches, which correspond to the 88 keys on the piano. With the pitch sampling for every 0.4s[7], we transform the midi files into sequences of numbers from 1 to 88 with the length 32.

To model the fitness of the discrete piano key patterns, BLEU is used as the evaluation metric. To model the fitness of the continuous pitch data patterns, the mean squared error (MSE) (Manaris et al. 2007) is used for evaluation.

From Table 4, we see that SeqGAN outperforms the MLE significantly in both metrics in the music generation task.

[5] https://github.com/samim23/obama-rnn

[6] http://www.iro.umontreal.ca/~lisa/deep/data

Algorithm 1 Sequence Generative Adversarial Nets

Require: generator policy G_θ; roll-out policy G_β; discriminator D_ϕ; a sequence dataset $\mathcal{S} = \{X_{1:T}\}$

Initialize G_θ, D_ϕ with random weights θ, ϕ.
Pre-train G_θ using MLE on $\mathcal{S}$
$\beta \leftarrow \theta$
Generate negative samples using G_θ for training D_ϕ
Pre-train D_ϕ via minimizing the cross entropy
repeat
    for g-steps **do**
        Generate a sequence $Y_{1:T} = (y_1, \ldots, y_T) \sim G_\theta$
        for t in $1 : T$ **do**
            Compute $Q(a = y_t; s = Y_{1:t-1})$ by Eq. (4)
        end for
        Update generator parameters via policy gradient Eq. (8)
    end for
    for d-steps **do**
        Use current G_θ to generate negative examples and combine with given positive examples $\mathcal{S}$
        Train discriminator D_ϕ for k epochs by Eq. (5)
    end for
    $\beta \leftarrow \theta$
until SeqGAN converges

GETMusic: Generating Music Tracks with a Unified Representation and Diffusion Framework

Ang Lv[‡†], Xu Tan[†*], Peiling Lu[†], Wei Ye[§], Shikun Zhang[§], Jiang Bian[†], Rui Yan[‡*]
[†]Microsoft Research Asia
[‡]Gaoling School of Artifical Intelligence, Renmin University of China
[§]National Engineering Research Center for Software Engineering, Peking University
{anglv, ruiyan}@ruc.edu.cn, {xuta, peil, jiabia}@microsoft.com,
{wye,zhangsk}@pku.edu.cn

https://github.com/microsoft/muzic

GETMusic

Artykuł ten pokazuje innowacyjne podejście do generacji muzyki, ponieważ zamiast tradycyjnych sekwencyjnych modeli używa modeli dyfuzyjnych. Pliki muzyczne są reprezentowane jako struktura 2D, gdzie osią Y jest ilość ścieżek dźwiękowych, a osią X kolejne timestamy.

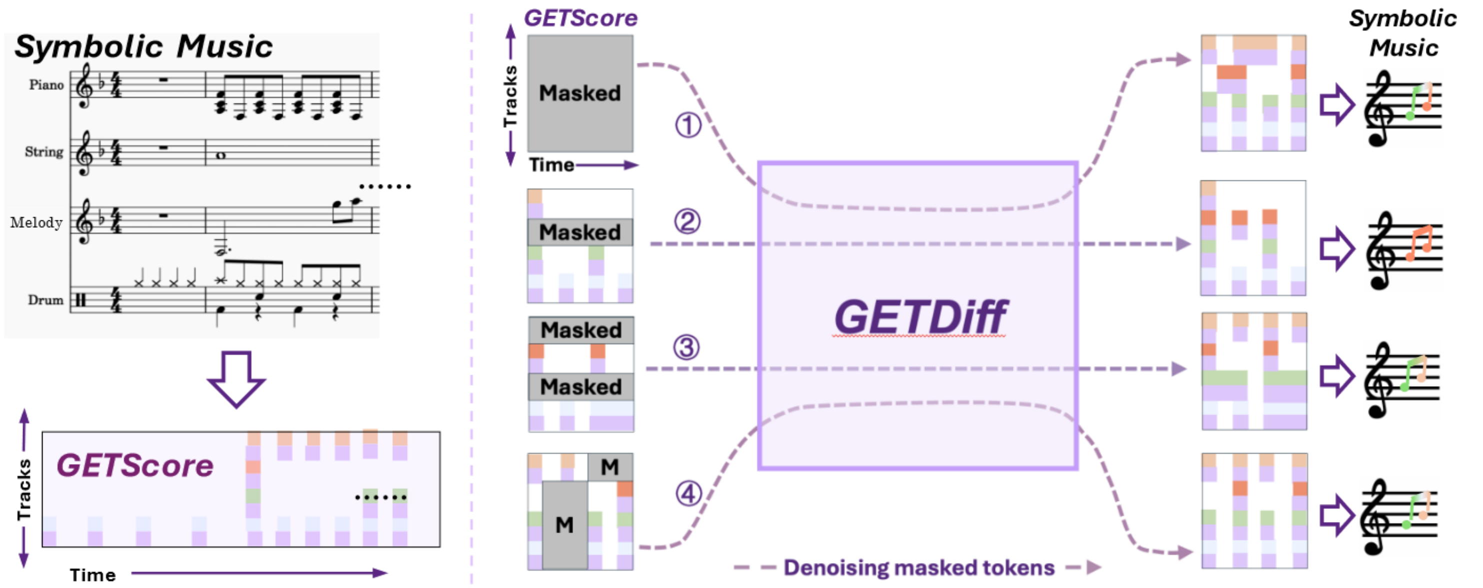
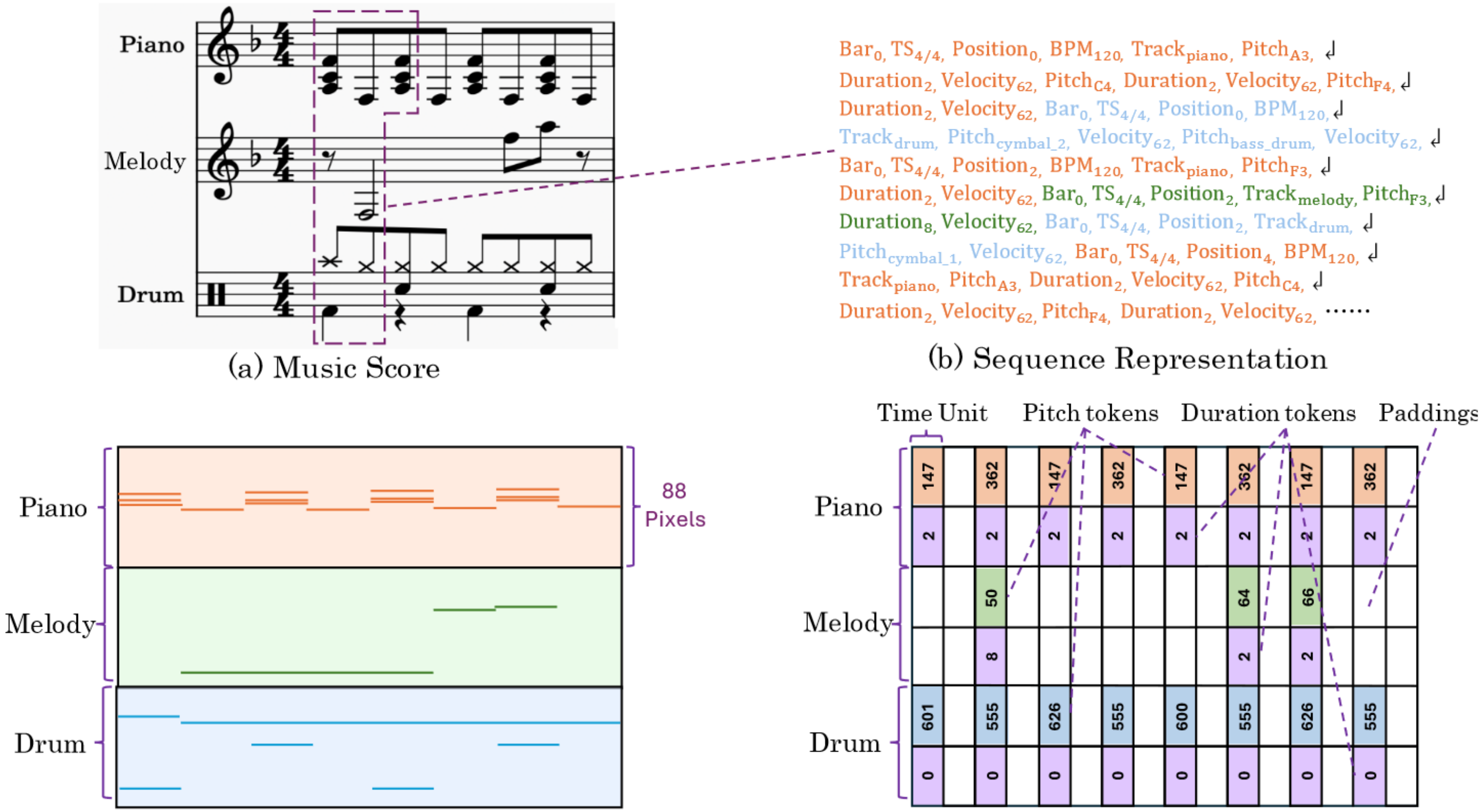


Figure 1: The overview of GETMusic, involving a novel music representation "GETScore" and a discrete diffusion model "GETDiff." Given a predefined ensemble of instrument tracks, GETDiff takes GETScores as inputs and can generate any desired target tracks conditioning on any source tracks (①, ②, and ③). This flexibility extends beyond track-wise generation, as it can perform zero-shot generation for any masked parts (④).

Podejście jest o tyle innowacyjne, że muzyka jest reprezentowana jako obrazek, a nie sekwencja dyskretnych tokenów.



(a) Music Score (b) Sequence Representation

(c) Pianoroll (d) GETScore

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[‡]
[†]State Key Laboratory for Novel Software Technology, Nanjing University, China
[‡]Microsoft Research Asia
[§]National Engineering Research Center for Software Engineering, Peking University, China
btyu@foxmail.com, {peil,ruiwa,xuta,taoqin,tyliu}@microsoft.com,
whu@nju.edu.cn, {wye,zhangsk}@pku.edu.cn

Museformer

Podobnie jak Music Transformer, ten model próbuje rozwiązać problem z ilością tokenów, które w algorytmie attention, każdy inny token ma brać pod uwagę. Ten artykuł wprowadza nowy rodzaj algorytmu nazywany *„fine- and coarse-grained attention (FC-Attention)”*. Podstawową ideą FC-Attention jest to, że zamiast bezpośrednio zwracać uwagę na wszystkie tokeny, co powoduje złożoność kwadratową, token określonego taktu zwraca uwagę tylko na takty związane ze strukturą, które są niezbędne do generowania muzyki strukturalnej (uwaga drobnoziarnista), a w przypadku innych taktów token zwraca uwagę tylko na ich tokeny podsumowujące, aby uzyskać skoncentrowane informacje (uwaga gruboziarnista). Aby to osiągnąć, najpierw podsumowujemy lokalne informacje każdego taktu w kroku podsumowania, a następnie agregujemy drobnoziarniste i gruboziarniste informacje w kroku agregacji.

Microsoft: muzic

- [1] ***MusicBERT***: *Symbolic Music Understanding with Large-Scale Pre-Training*, Mingliang Zeng, Xu Tan, Rui Wang, Zeqian Ju, Tao Qin, Tie-Yan Liu, **ACL 2021**.
- [2] ***PDAugment***: *Data Augmentation by Pitch and Duration Adjustments for Automatic Lyrics Transcription*, Chen Zhang, Jiaxing Yu, Luchin Chang, Xu Tan, Jiawei Chen, Tao Qin, Kejun Zhang, **ISMIR 2022**.
- [3] ***DeepRapper***: *Neural Rap Generation with Rhyme and Rhythm Modeling*, Lanqing Xue, Kaitao Song, Duocai Wu, Xu Tan, Nevin L. Zhang, Tao Qin, Wei-Qiang Zhang, Tie-Yan Liu, **ACL 2021**.
- [4] ***SongMASS***: *Automatic Song Writing with Pre-training and Alignment Constraint*, Zhonghao Sheng, Kaitao Song, Xu Tan, Yi Ren, Wei Ye, Shikun Zhang, Tao Qin, **AAAI 2021**.
- [5] ***TeleMelody***: *Lyric-to-Melody Generation with a Template-Based Two-Stage Method*, Zeqian Ju, Peiling Lu, Xu Tan, Rui Wang, Chen Zhang, Songruoyao Wu, Kejun Zhang, Xiangyang Li, Tao Qin, Tie-Yan Liu, **EMNLP 2022**.
- [6] ***ReLyMe***: *Improving Lyric-to-Melody Generation by Incorporating Lyric-Melody Relationships*, Chen Zhang, LuChin Chang, Songruoyao Wu, Xu Tan, Tao Qin, Tie-Yan Liu, Kejun Zhang, **ACM Multimedia 2022**.
- [7] ***Re-creation of Creations***: *A New Paradigm for Lyric-to-Melody Generation*, Ang Lv, Xu Tan, Tao Qin, Tie-Yan Liu, Rui Yan, arXiv 2022.
- [8] ***MeloForm***: *Generating Melody with Musical Form based on Expert Systems and Neural Networks*, Peiling Lu, Xu Tan, Botao Yu, Tao Qin, Sheng Zhao, Tie-Yan Liu, **ISMIR 2022**.
- [9] ***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, **NeurIPS 2022**.
- [10] ***PopMAG***: *Pop Music Accompaniment Generation*, Yi Ren, Jinzheng He, Xu Tan, Tao Qin, Zhou Zhao, Tie-Yan Liu, **ACM Multimedia 2020**.
- [11] ***HiFiSinger***: *Towards High-Fidelity Neural Singing Voice Synthesis*, Jiawei Chen, Xu Tan, Jian Luan, Tao Qin, Tie-Yan Liu, arXiv 2020.
- [12] ***CLaMP***: *Contrastive Language-Music Pre-training for Cross-Modal Symbolic Music Information Retrieval*, Shangda Wu, Dingyao Yu, Xu Tan, Maosong Sun, **ISMIR 2023**, ***Best Student Paper Award***.
- [13] ***GETMusic***: *Generating Any Music Tracks with a Unified Representation and Diffusion Framework*, Ang Lv, Xu Tan, Peiling Lu, Wei Ye, Shikun Zhang, Jiang Bian, Rui Yan, arXiv 2023.
- [14] ***MuseCoco***: *Generating Symbolic Music from Text*, Peiling Lu, Xin Xu, Chenfei Kang, Botao Yu, Chengyi Xing, Xu Tan, Jiang Bian, arXiv 2023.
- [15] ***MusicAgent***: *An AI Agent for Music Understanding and Generation with Large Language Models*, Dingyao Yu, Kaitao Song, Peiling Lu, Tianyu He, Xu Tan, Wei Ye, Shikun Zhang, Jiang Bian, **EMNLP 2023 Demo**.

Google: Magenta

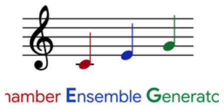
Magenta Studio 2.0 — AUGUST 24, 2023

Magenta Studio has been upgraded to more seamlessly integrate with Ableton Live. It is a collection of music creativity tools built on Magenta's open source models, using cutting-edge machine learning techniques for music generation.

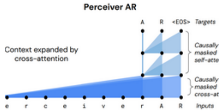
The 2023 I/O Preshow – Composed by Dan Deacon (with some help from MusicLM) — JUNE 21, 2023

A look into Dan Deacon's creative process for the 2023 Google I/O preshow.

The Wordcraft Writers Workshop: Creative Co-Writing with AI — DECEMBER 1, 2022

We invited 13 professional writers to explore the limits of co-writing with LaMDA and foster an honest and earnest conversation about the rapidly changing relationship between technology and creativity.

The Chamber Ensemble Generator and CocoChorales Dataset — SEPTEMBER 30, 2022

We combine Coconet and MIDI-DDSP into a system called the Chamber Ensemble Generator, which we use to make a giant dataset of four-part Bach chorales called CocoChorales.



Autoregressive long-context music generation with Perceiver AR — JUNE 16, 2022

We generate music with clear long-term coherence and structure in both symbolic and audio domains, by attending to inputs spanning up to several minutes.

DDSP-VST: Neural Audio Synthesis for All — JUNE 8, 2022

DDSP-VST is a neural audio synthesizer for your digital audio workstation, powered by DDSP.