УДК 004.89

МОДЕЛИ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ ГЕНЕРАЦИИ МУЗЫКИ

Золотарев А.М., Белов Ю.С.

ФГБОУ ВО «Московский государственный технический университет имени Н.Э. Баумана», Калужский филиал, Калуга, e-mail: artemmaxzolotarev@yandex.ru

Аннотация. Одним из наиболее популярных направлений применения искусственного интеллекта на сегодняшний день является генерация музыки. Цель данной работы – рассмотреть существующие сегодня модели генерации аудиоданных, основанные на рекуррентных нейронных сетях, генеративных состязательных сетях, вариационных автоэнкодерах и трансформерах. В ходе исследования были проанализированы последние разработки в области генерации музыки. В статье были представлены наиболее современные архитектуры генерации аудиоданных. Модели, описанные в данной статье, используются для генерации мелодий, аранжировок, нот, сохраняя при этом стиль произведения и генерируя довольно продолжительные композиции. Представленные в тексте архитектуры, такие как MelodyRNN, C-RNN-GAN, MusicVAE и др., демонстрируют различные подходы и техники, используемые для создания музыкального контента на сегодняшний день. Каждая из этих моделей обладает как своими преимуществами, так и недостатками, а также зачастую используют комбинации различных подходов к построению алгоритмов. Данное исследование призвано показать многообразие подходов к генерации музыки на основе моделей глубокого обучения. Генерация музыки имеет большой потенциал применения, как помогая творческим людям, так и заменяя их.

Ключевые слова: глубокое обучение, рекуррентные нейронные сети, состязательные сети, вариационные автоэнкодеры, трансформеры

DEEP LEARNING MODELS FOR MUSIC GENERATION

Zolotarev A.M., Belov Yu.S.

Bauman Moscow State Technical University, Kaluga branch, Kaluga, e-mail: artemmaxzolotarev@yandex.ru

Annotation. One of the most popular applications of artificial intelligence today is music generation. The purpose of this work was to consider the currently existing models of audio data generation based on recurrent neural networks, generative adversarial networks, variational autoencoders and transformers. The research analyzed the latest developments in the field of music generation. The most modern architectures of audio data generation were presented in the article. The models described in this article are used to generate melodies, arrangements, and notes, while maintaining the style of the piece and generating fairly long compositions. The architectures presented in the text, such as MelodyRNN, C-RNN-GAN, MusicVAE and others, demonstrate the various approaches and techniques used to create musical content today. Each of these models has its own advantages and disadvantages, and often use combinations of different approaches to building algorithms. This study aims to show the variety of approaches to music generation based on deep learning models. Music generation has great potential for use both as an assistant to creative people and as a substitute for them.

Keywords: deep learning, recurrent neural networks, adversarial networks, variational auto-encoders, transformers

Все созданные на данный момент модели генерации музыки можно классифицировать в зависимости от вида использованного алгоритма на четыре вида: основанные на правилах, использующих марковскую модель, работающие на алгоритмах глубокого обучения и применяющие эволюционные вычисления. Наиболее распространенными из них являются модели, в основе которых лежат алгоритмы глубокого обучения.

Цель исследования – рассмотреть модели глубокого обучения для генерации музыки, такие как рекуррентные нейронные сети, генеративные состязательные сети, вариационнные автоэнкодеры и трансформеры.

Общая структура модели

Генерация контента – это расширенная область глубокого обучения. Благодаря достижениям Google black и CTRL в области создания музыки, глубокое обучение как метод создания музыки привлекает все большее внимание. В отличие от систем музыкальной генерации, которые оперируют грамматикой или правилами, системы, использующие глубокое обучение, способны анализировать распределение и актуальность сэмплов из различных музыкальных корпусов и создавать музыку в стиле исследуемого массива путем прогнозирования или классификации [1].

Рекуррентные нейронные сети

Рекуррентные нейронные сети (RNN) представляют собой класс нейронных сетей, предназначенных для анализа временных рядов, что делает их удобными для работы с музыкальными данными. Од-

нако проблема длительных временных зависимостей остается актуальной для RNN из-за проблемы градиентного исчезновения или взрыва, которая возникает при обучении на длинных временных последовательностях. LSTM, как вариант RNN, эффективно решает проблему длительной временной зависимости RNN, вводя состояния ячеек и используя три типа элементов управления, а именно входные элементы, элементы забывания и выходные элементы, предназначенные для хранения информации и управления ею [2].

Google Brain разработала MelodyRNN (рис. 1), где использовались lookback RNN и attention RNN для повышения способности RNN к запоминанию структур в длинных последовательностях. Кирти и его коллеги внедрили механизм внимания и использовали метод отсева для снижения переобучения при создании джазовой музыки [3]. Rl Tuner использует сеть RNN для предоставления частичных значений вознаграждения для модели обучения с подкреплением. Нейронная сеть Anticipation-RNN для генерации мелодий интерактивных припевов в стиле Баха. Макрис и соавт. разработали индивидуальные сети LSTM для различных видов барабанов, при этом сеть прямой связи (FF) играла роль условного уровня [4]. StructureNet создает базовую монофоническую сопровождающую музыку, используя сети LSTM.

Что касается генерации аранжировок, Folk-RNN впервые применила LSTM для создания музыкальных последовательностей, представленных в формате ABC для создания народной музыки. DeepBach использует Bi-LSTM для создания хоровой музыки в стиле Баха, учитывая двунаправленный поток времени: одно направление учитывает прошлое, а другое – будущее. XiaoIce Band предлагает сквозную китайскую многодорожечную платформу для генерации поп-музыки, которая использует сеть GRU для обработки низкоразмерных аккордов и получение скрытых состояний с помощью кодеров и декодеров. Amadeus применяет явный подход к кодированию длительности, представляя несколько аудиопотоков в виде длительности нот и используя механизм вознаграждения RL для улучшения структуры создаваемой музыки. JamBot использует LSTM сеть аккордов для предсказания последовательностей аккордов и полифоническую LSTM сеть для генерации полифонической музыки на основе этих предсказанных последовательностей аккордов [5].

PerformanceRNN преобразует MIDI-файл живой фортепианной пьесы в музыкальное представление нескольких one-hot векторов с 413 измерениями и определяет временной «шаг» фиксированного размера (10 мс) вместо значения времени ноты. PerformanceNet – это первая попытка преобразования партитуры в аудио с использованием полностью сверточной нейронной сети с символическим представлением музыки в качестве входных данных и звуковым представлением в качестве выходных данных [6].

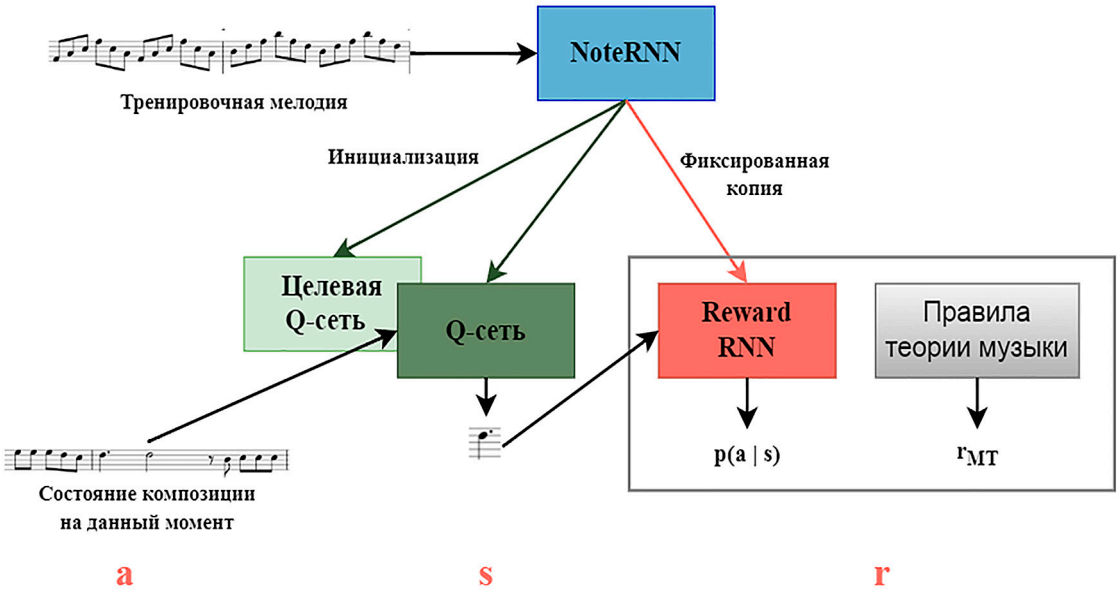

Рис. 1. Архитектура модели MelodyRNN

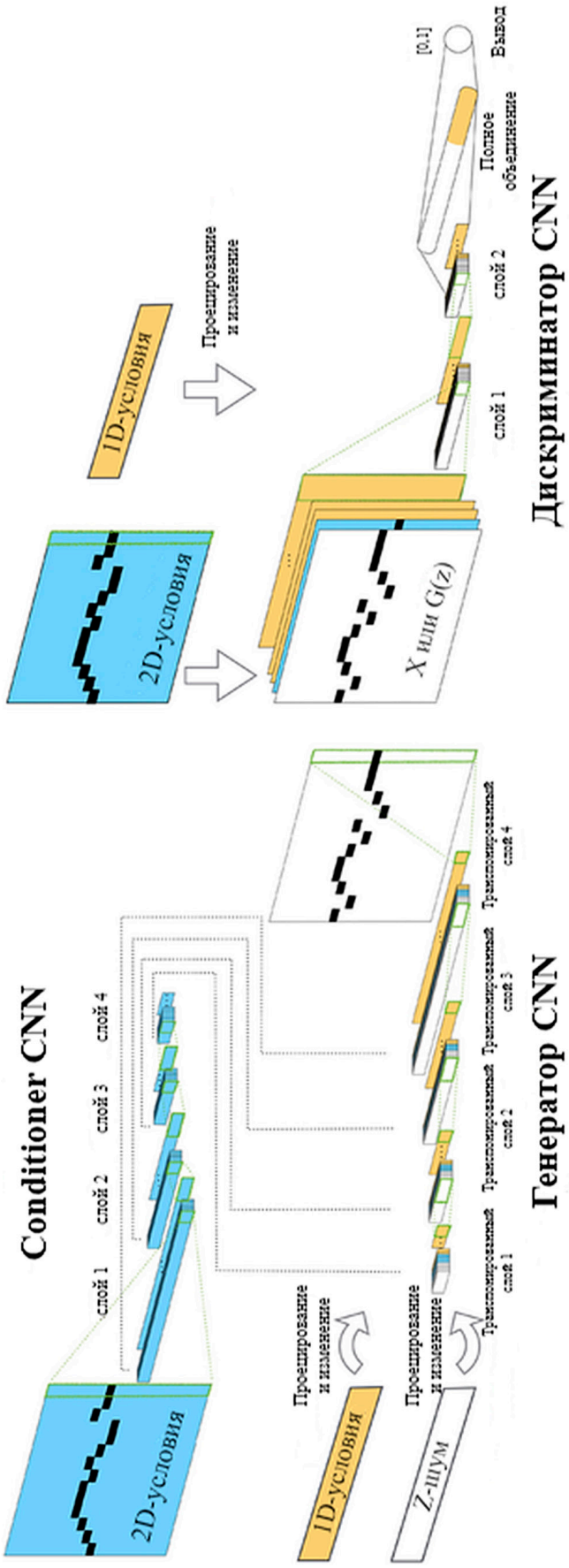

Рис. 2. Системная схема модели MidiNet для генерации музыки в символьном представлении

Генеративные состязательные сети

Генеративные состязательные сети (GAN) работают по принципу соревнования между двумя нейронными сетями: генератором и дискриминатором. Генератор создает новые данные из случайного шума или других данных. Он стремится создать данные, которые похожи на обучающие. Дискриминатор принимает как настоящие обучающие данные, так и данные, созданные генератором, и пытается отличить их друг от друга.

C-RNN-GAN использует LSTM сети для создания музыкальных мелодий, однако он не имеет механизма для создания музыки с учетом определенного начального условия. MidiNet (рис. 2) улучшает метод, вставляя информацию о мелодии и аккорде, сгенерированных на предыдущем этапе, в качестве условного механизма на среднем уровне свертки генератора для ограничения генерации типов нот. JazzGAN разработал модель GAN для монофонической джазовой музыки, используя LSTM для музыкальной импровизации на основе последовательности аккордов. SSMGAN применил модель GAN для создания самоподобной матрицы (SSM) для отражения музыкальных самоповторов, которая затем была подана на сеть LSTM для генерации мелодий. Для генерации мелодии из текстов песен был разработан условный LSTM-GAN. Он содержит генератор LSTM и дискриминатор LSTM, оба с текстами песен в качестве условных входных данных.

Для генерации аранжировки MusGAN предложил модель GAN с тремя типами генераторов для построения корреляций между несколькими дорожками. BinaryMuseGAN усовершенствовал вышеупомянутый метод, введя бинарные нейроны (BN) в качестве входных данных для генератора.

Лян и соавт. провели исследования, в котором попытались воссоздать сианьскую барабанную музыку, обучив генеративную сеть на правилах теории музыки и характеристиках китайской народной музыки [7].

Когда речь идет о передаче стиля и создании звука, CycleGAN применяет функцию потери стиля для изменений и функцию потери контента для сохранения согласованности контента [8]. CycleBEGAN использует сеть BEGAN для стабилизации процесса обучения, а также вводит переходные соединения для улучшения четкости мелодии и текстов песен, а также рекурсивные слои для повышения точности высоты тона для достижения преобразования мужского и женского голоса. Кроме того, Джин и др. использовали сеть LSTM в качестве генератора, добавив музыкальные правила в качестве функции вознаграждения при обучении с подкреплением в управляющую сеть [9]. В области генерации звука WaveGAN был первым, кто попытался использовать GAN для создания аудиосигналов в их необработанном виде, а GANSynth усовершенствовал этот подход, генерируя всю последовательность параллельно.

Вариационный автоэнкодер

Вариационный автоэнкодер (VAE) – это алгоритм сжатия для кодеров и декодеров, который способен анализировать и генерировать такую информацию, как динамику высоты тона и инструментовку в полифонической музыке.

Примерами таких моделей являются MIDI-VAE и MusicVAE (рис. 3). MIDI-VAE использует три пары кодеров/декодеров, которые вместе используют латентное пространство для автоматического восстановления высоты тона, интенсивности и инструментовки музыкальной композиции, с целью изменения музыкального стиля. MusicVAE использует иерархический декодер для улучшения моделирования последовательностей с долговременной структурой, используя двунаправленный RNN в качестве кодера. Вэй и соавт. не применяли сквозной подход к изучению иерархических представлений, а вместо этого представили новую модель, основанную на EC2-VAE [10]. Дубнов и соавт. разработали vanilla polyphonic VAE, используя только линейные слои, чтобы изучить скрытое представление музыкальной поверхности [11].

MG-VAE – первое исследование, где были использованы глубокие генеративные модели и методы состязательного обучения для создания восточной популярной и фолк-музыки [12].

MahlerNet построил условный VAE для моделирования распределения латентных состояний. Две двунаправленные сети RNN образуют кодер, а декодер выводит длительность, высоту звука и инструмент [13]. MIDI-Sandwich2 представляет иерархическую мультимодальную сеть, порожденную слиянием VAE (MFG-VAE), основанную на RNN. MusAE впервые применил состязательные автоэнкодеры для генерации музыки.

Трансформер

Основная идея этой архитектуры заключается в использовании механизма внимания для указания корреляций между входными данными.

Музыкальный трансформер существенно снижает пространственную сложность промежуточных векторов, представляющих относительное положение в порядке очередности длины, что делает его применимым для фортепианных музыкальных композиций.

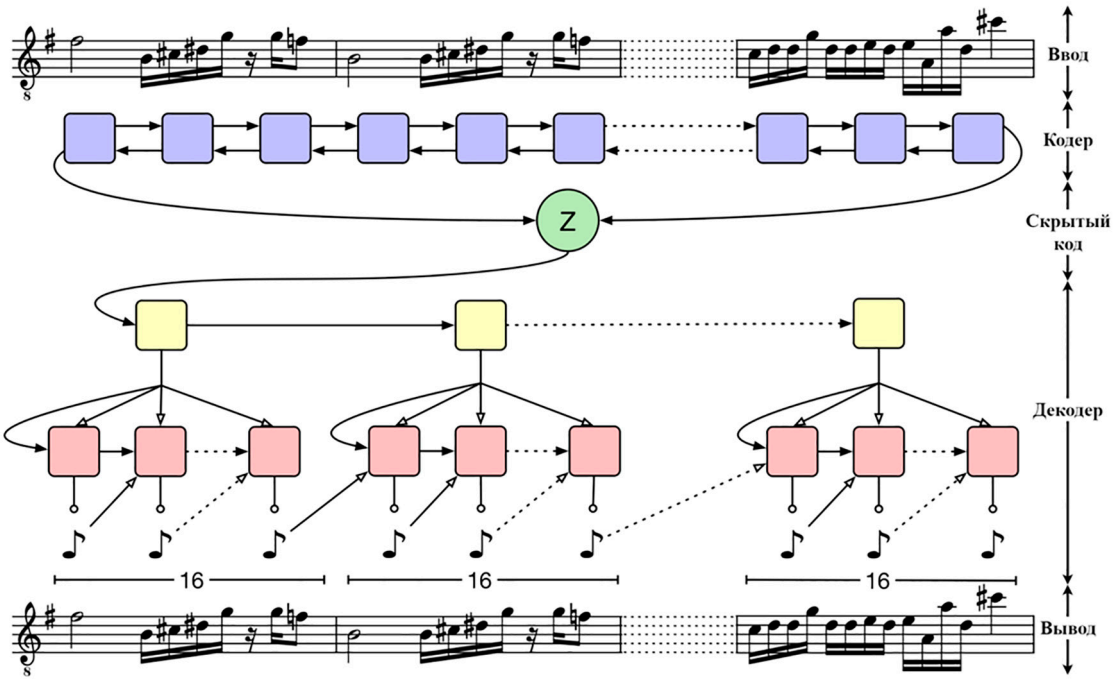

Рис. 3. Схема иерархической рекуррентной вариационной модели автоэнкодера MusicVAE

MuseNet использовала ту же сеть, что и GPT-2, которая обучается с использованием пересчитанного и оптимизированного ядра трансформера для генерации четырехминутной музыкальной композиции, состоящей из десяти различных инструментов. На рис. 4 показана блок-схема этого трансформаторного самокодера [5].

Цао с соавт. предложили метод, основанный на трансформере, для генерации высококачественной мультиинструментальной музыки из входных данных аккордовой последовательности [14]. Muse Morphose предлагает соединить воедино трансформер и вариационный автоэнкодер для того, чтобы добиться передачи стиля длинных пьес для фортепиано поп-музыки, где могут быть указаны различные атрибуты композиции. Transformer VAE объединяет трансформер с VAE, эффективно устраняя ограничения VAE в обработке структур временных рядов и невыясненной природы скрытых состояний трансформера. Чой и соавт. аналогичным образом использовали трансформер (декодер) для получения глобального представления музыки [15].

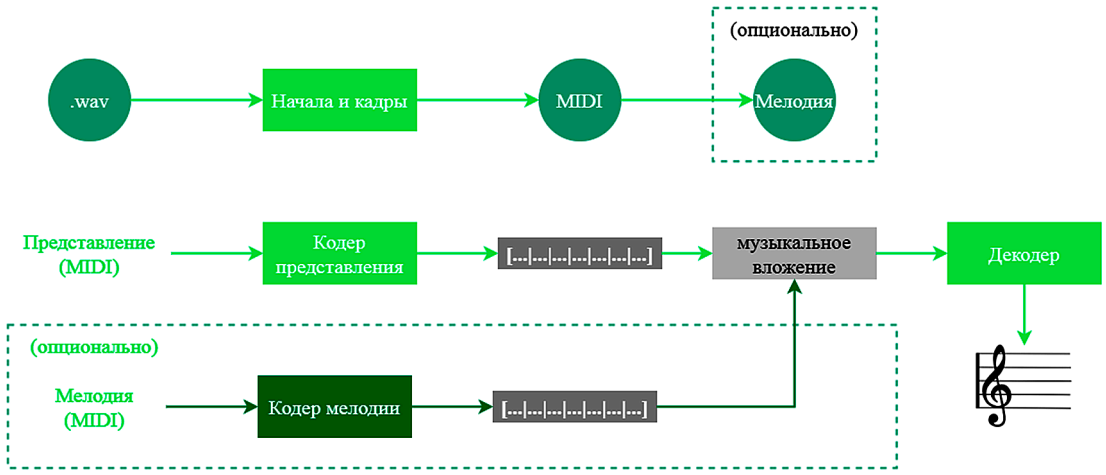

Рис. 4. Технологическая схема самокодера Transformer MuseNet

Заключение

В данной статье были рассмотрены существующие на сегодняшний день модели для генерации музыки, основанные на рекуррентных нейронных сетях генеративных состязательных сетях, вариационных автоэнкодерах и трансформерах.

Список литературы

1. Мосин Е.Д., Белов Ю.С. Основы генерации музыки в волновом и нотном форматах // Наукоемкие технологии в приборо- и машиностроении и развитие инновационной деятельности в вузе: материалы Всероссийской научно-технической конференции. М., 2022. С. 71–75.

2. Мосин Е.Д., Белов Ю.С. Генерация музыки с использованием двунаправленной рекуррентной нейронной сети // Научное обозрение. Технические науки. 2023. № 1. С. 10–14.

3. Keerti G., Vaishnavi V.A., Mukherjee P. Attentional networks for music generation // Multimedia tools and applications. 2020. P. 1–5.

4. Makris D., Kaliakatsos-Papakostas M., Karydis I. Conditional neural sequence learners for generating drums' rhythms // Neural Computing and Applications. 2019. № 31 (6). P. 1793–1804.

5. Natsiou A., O'Leary S. Audio representations for deep learning in sound synthesis: A review // IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA). 2021. P. 1–8.

6. Oore S., Simon I., Dieleman S. This time with feeling: learning expressive musical performance // Neural Computing and Applications. 2020. № 32 (4). P. 955–967.

7. Liang T., Li P., Cao Y. Research on Generating Xi'an Drum Music Based on Generative Adversarial Network // 2023 IEEE International Conference on Sensors, Electronics and Computer Engineering (ICSECE). 2023. P. 1398–1402.

8. Brunner G., Wang Y., Wattenhofer R. Symbolic Music Genre Transfer with CycleGAN // IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI). 2018. P. 786–793.

9. Jin C., Tie Y., Bai Y. A Style-Specific Music Composition Neural Network // Neural Processing Letters. 2020. № 52 (3). P. 1893–1912.

10. Wei S., Xia G. Learning long term music representations via hierarchical contextual constraints // 22nd International Society for Music Information Retrieval Conference. 2021. P. 738–745.

11. Wang T., Liu J., Jin C. An intelligent music generation based on Variational Autoencoder // International Conference on Culture-oriented Science & Technology (ICCST). 2020. P. 394–398.

12. Luo J., Yang X., Ji S. MG-VAE: Deep Chinese Folk Songs Generation with Specific Regional Styles // Proceedings of the 7th Conference on Sound and Music Technology (CSMT), 2020. P. 93–106.

13. Lousseief E., Sturm B. MahlerNet: Unbounded Orchestral Music with Neural Networks // Nordic Sound and Music Computing Conference. 2019. P. 57–63.

14. Cao B., Fukumori T., Yamashita Y. Multi-Instruments Music Generation Based on Chord Input // 2023 IEEE 12th Global Conference on Consumer Electronics (GCCE). 2023. P. 1082–1083.

15. Choi K., Hawthorne C., Simon I. Encoding musical style with transformer autoencoders // International Conference on Machine Learning. 2020. P. 1899–1908.