УДК 004. 89 **Технические науки**

Мосин Евгений Дмитриевич, *студент магистр, Калужский филиал ФГБОУ ВО «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»*

Белов Юрий Сергеевич, *к.ф.-м.н., доцент, Калужский филиал ФГБОУ ВО «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»*

МОДЕЛЬ ВАРИАЦИОННОГО АВТОКОДИРОВЩИКА ДЛЯ ЖАНРОВОЙ ГЕНЕРАЦИИ МУЗЫКИ

Аннотация: Генерация музыки с использованием нейронных сетей является актуальной исследовательской темой в современной эпоху цифровой трансформации и развития искусственного интеллекта. Благодаря стремительному развитию компьютерных технологий и доступности больших объемов данных, нейронные сети предоставляют новые возможности для творческого процесса в музыкальной индустрии. С помощью моделей вариационных автокодировщиков можно добиться качественной генерации музыки в заданном стиле и жанре.

Ключевые слова: Жанровая генерация музыки, VAE, апостериорный коллапс, шум Гамбеля.

Abstract: Music generation using neural networks is a hot research topic in the modern era of digital transformation and the development of artificial intelligence. Thanks to the rapid development of computer technology and the availability of large amounts of data, neural networks provide new opportunities for the creative process in the music industry. With the help of variational autoencoder models, you can achieve high-quality generation of music in a given style and genre.

Keywords: Genre music generation, VAE, post-collapse, Gumbel noise.

Введение. Музыка, сгенерированная нейронными сетями, находит применение в различных областях и сферах деятельности. Она может быть использована в рекламных роликах, фоновой музыке для видеоигр, фильмов и телевизионных шоу. Ученые исследуют возможности моделей генерации музыки, исследуют новые алгоритмы и методы, а также разрабатывают инструменты для создания и редактирования музыки. Нейронные сети для генерации музыки могут быть интегрированы в синтезаторы или программы для создания музыки, чтобы предлагать пользователю новые идеи и варианты музыкальных фрагментов. С развитием технологий и исследований в этой области, появляются новые возможности для экспериментов с музыкальным искусством.

Архитектура предлагаемой модели

Упрощенная предлагаемая модель изображена на рис. 1, где пунктирная линия представляет собой обычный VAE. Чтобы изучить стиль музыки, в предлагаемой архитектуре скрытое пространство z разделено на две части, z_s и z_c. Идея, стоящая за этим, заключается в том, что кодировщик кодирует «содержание» входной музыки, такое как высота нот и длина нот, в z_s, в то время как z_c будет оптимизирован, чтобы содержать информацию о «стиле», которая направляет декодировщик для создания музыки в определенном стиле.

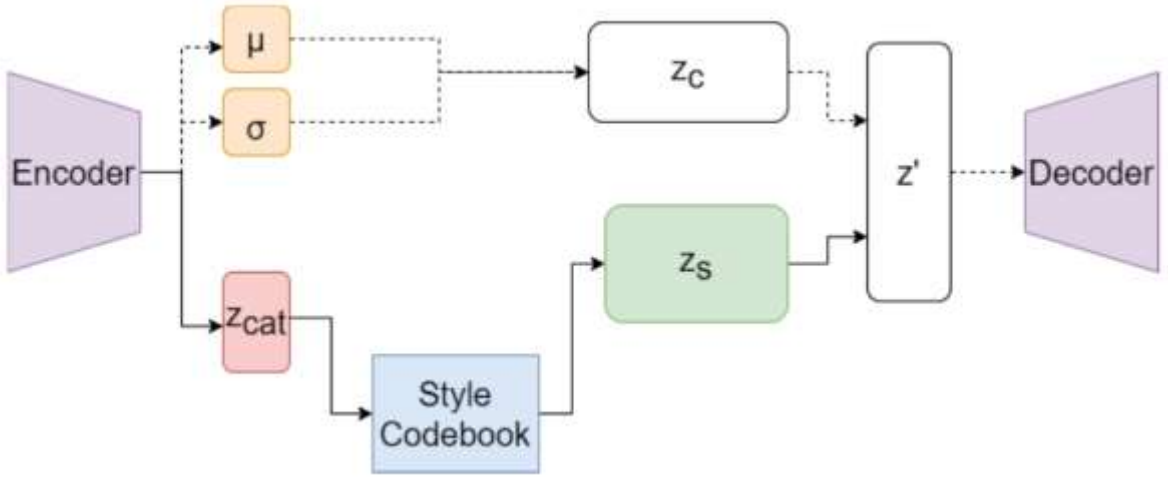

Рис. 1. Предлагаемая архитектура модели.

Чтобы получить z_s, кодировщик сначала генерирует однозначную категориальную переменную $z_{cat} \in \{0,1\}^s$, где s обозначает количество стилей в наборе данных. Из-за природы дискретных переменных градиенты z_{cat} трудно поддаются обработке. Поэтому для обработки используется прием распределения Гамбеля, сформулированный следующим образом:

$$G = -\log(-\log(Uinf[0,1]))$$

$$z_{cat} = softmax(\frac{\alpha + G}{\tau_{gumbel}})$$

где G — шум Гамбеля, α — логиты для z_{cat} от кодировщика, а τ_{gumbel} — температура. Температура является гиперпараметром, по мере приближения к нулю z_{cat} становится однократным вектором.

Прием распределения Гамбеля используется для генерации случайных категориальных переменных z_{cat}, которые затем используются для генерации музыкальных событий (например, нот и темпа).

Кодировщик в данном случае генерирует логиты α для каждой категории z_{cat}, где каждая категория соответствует определенному стилю музыки. Однако градиенты дискретных переменных z_{cat} трудно обрабатывать, поэтому используется прием распределения Гамбеля, который позволяет генерировать плавные категориальные переменные.

Uinf[0,1] - равномерно распределенная случайная величина на интервале [0, 1]. Затем полученный шум Гамбеля G используется для преобразования логитов α для каждой категории zcat с помощью формулы softmax((α+G)/τ_gumbel).

Таким образом, прием распределения Гамбеля позволяет получить плавную версию дискретной категориальной переменной z_{cat}, что упрощает обработку градиентов в контексте генерации музыки. Предполагаемая модель показана на рис. 2.

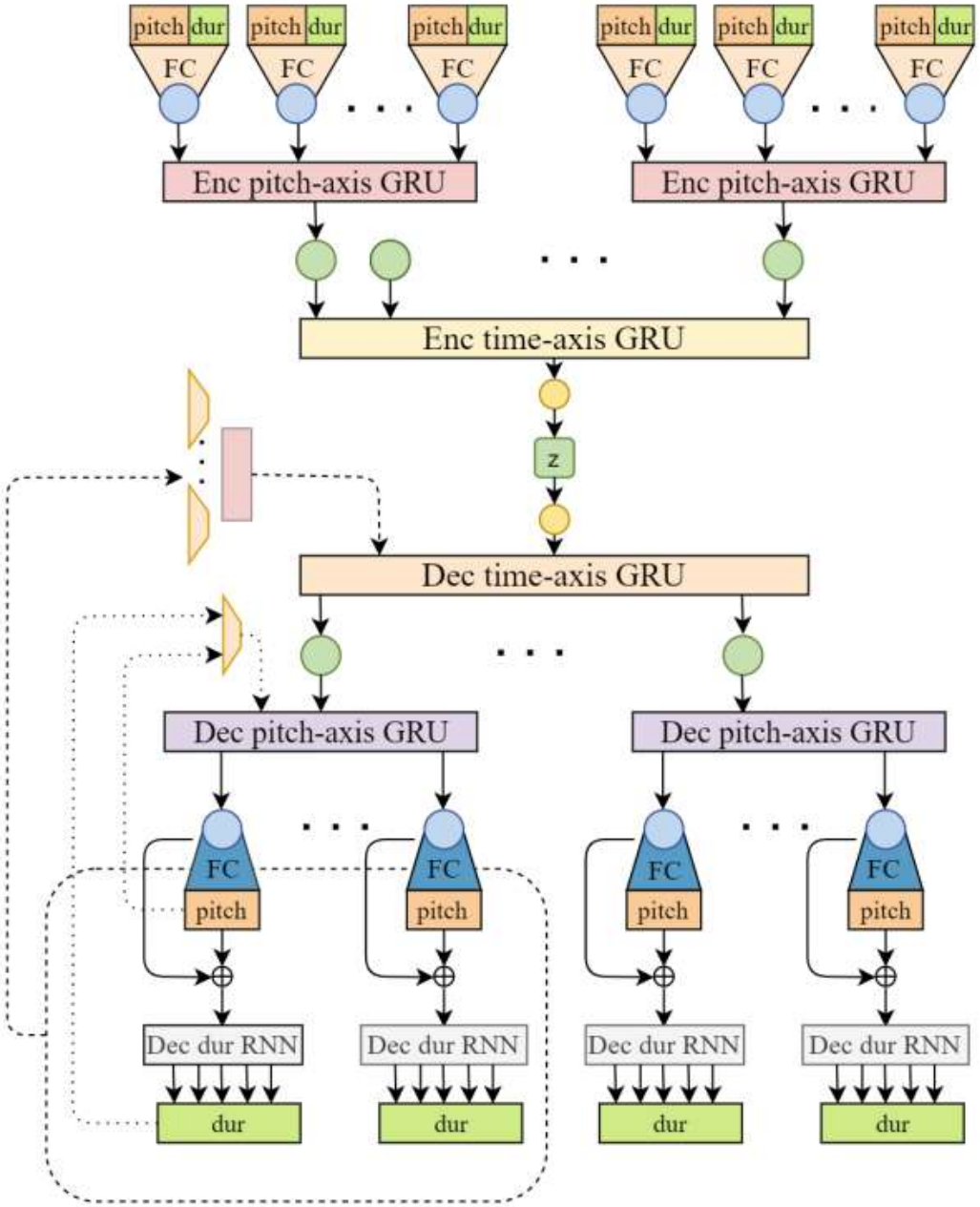

Рис. 2. Предполагаемая модель.

Кодовая книга стиля

Затем вводится кодовая книга обучаемого стиля, которая случайным образом инициализируется в начале обучения. Кодовая книга используется для извлечения вложений стилей на основе z_{cat}. Кодовая книга состоит из s вложений, каждое из которых имеет размерность dim_s, что дает ему размер $s \times dim_s$. Чтобы получить z_s, выполняется матричное умножение между z_{cat} и кодовой книгой стилей:

$$z_s = z_{cat} \times style\_codebook$$

z_s, будет иметь форму dim_s. Таким образом, можно гарантировать, что разные песни одного стиля будут иметь одинаковые вложения стиля z_s. Кроме того, чтобы модель правильно запоминала стиль, z_{cat} оптимизирован с помощью меток стиля y с использованием кросс-энтропии:

$$L_s = -\sum_{s=1}^{S} y log z_{cat}$$

где s — количество стилей в наборе данных.

z_c аналогично исходному скрытому представлению $z = \mu + e\sigma$. Затем z_s и z_c объединяются, чтобы сформировать z'. Наконец, z' проходит через линейный слой для формирования начального состояния декодировщика RNN.

Проблема апостериорного коллапса

Апостериорный коллапс — это проблема, когда декодировщик игнорирует скрытые векторы, которыми в данном случае является z_c, в результате чего член KL падает до нуля и делает z_c бесполезным. Эта проблема часто встречается в настройках seq2seq VAE из-за особенностей авторегрессионного декодера. Для решения этой проблемы используется μ-форсинг. В контексте VAE, μ-форсинг используется для сохранения информации о скрытых переменных zc в процессе генерации x из z. В частности, μ-форсинг добавляет регуляризирующий член Lμ в формулировку VAE, который заставляет выборочную дисперсию μ контролироваться на уровне β_μ. Здесь β_μ — это гиперпараметр, который определяет силу регуляризации. Член Lμ вынуждает выборочную дисперсию μ контролироваться на уровне β_μ, что поддерживает взаимную информацию между входом x и скрытыми переменными z_c, необходимыми для восстановления ввода. В результате форсинга модель VAE становится более способной к моделированию зависимостей между входом x и скрытыми переменными z_c, что позволяет ей эффективно решать задачу восстановления ввода x. Итоговая формулировка Lμ выглядит так:

$$L_\mu = \max(0, \beta_\mu - \frac{1}{2N}\sum_{n=1}^{N}(\mu^n - \bar{\mu})^T(\mu^n - \bar{\mu}))$$

где β_μ — сдвиг, а N — размер партии. μ — это средний вектор, смоделированный кодировщиком для параметризации распределения z_c. Термин L_μ заставляет выборочную дисперсию μ контролироваться на уровне β_μ, который поддерживает взаимную информацию X и z'.

В целом окончательная формулировка предложенной модели выглядит следующим образом:

$$L'_V = L_r - \beta D_{KL}(q(z_c|x)\big|\big|p(z_c)\big) + L_s + L_\mu$$

Во время генерации можно указать z_{cat} в соответствии с желаемым стилем, взять образцы z_c из распределения Гаусса и передать их в декодировщик для создания новой песни.

Заключение. Модификация вариационного автоэнкодера между последовательностями, которая позволяет пользователю определять стиль создаваемой выходной музыки. Модель включает непрерывное внедрение стилей для каждого стиля в наборе данных. С помощью экспериментов показано, что предлагаемый метод превосходит базовый уровень, который напрямую передает метки дискретного стиля в модель, аналогичную другим работам в литературе.

Библиографический список:

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

2. П. Е. Белоножко. Модификации архитектуры WaveNet для реализации вокодера в генеративной модели преобразования текста в речь // Научное обозрение. Технические науки. 2022. № 6. с. 37-42.

3. K. Zhao, S. Li, J. Cai, H. Wang, J. Wang «An Emotional Symbolic Music Generation System based on LSTM Networks», 2019 IEEE 3rd Information

Technology, Networking, Electronic and Automation Control Conference (ITNEC).

4. K. Chen, W. Zhang, S. Dubnov, G. Xia, W. Li «The Effect of Explicit Structure Encoding of Deep Neural Networks for Symbolic Music Generation», 2019 International Workshop on Multiplayer Music Representation and Processing (MMRP).

5. A. Emin, «Piano Music Generation with a Text Based Musical Note Representation using LSTM Models», 2011 29th Signal Processing and Communications Applications Conference.

6. H. H. Mao, T. Shin, G. W. Cottrell «DeepJ: Style-Specific Music Generation», 2018 12th IEEE International Conference on Semantic Computing.

7. D. Liu, X. Yang, F. He, Y. Chen, J. Lv, «mu-forcing: Training variational recurrent autoencoders for text generation», 2019 arXiv preprint arXiv:1905.10072.

8. A. Roberts, J. H. Engel, C. Raffel, C. Hawthorne, D. Eck, «A hierarchical latent vector model for learning long-term structure in music», 2018 in ICML.