ORES
Научно-практический электронный журнал
Оригинальные исследования (ОРИС), *ISSN 2222-0402*
www.ores.su
original-research.ru

УДК 004.89

ИСПОЛЬЗОВАНИЕ МОДЕЛИ ВАРИАЦИОННОГО АВТОКОДИРОВЩИКА ДЛЯ ГЕНЕРАЦИИ МУЗЫКИ

Мосин Евгений Дмитриевич
Калужский филиал государственного бюджетного образовательного учреждения высшего образования «Московский государственный технический университет имени Н. Э. Баумана (национальный исследовательский университет)»
Студент магистр
Lolko40rus@yandex.ru

Федоров Виктор Олегович
Калужский филиал государственного бюджетного образовательного учреждения высшего образования «Московский государственный технический университет имени Н. Э. Баумана (национальный исследовательский университет)»
Кандидат технических наук, доцент
Fedorov_vo@bmstu.ru

Аннотация

Музыкальная генерация нейронными сетями является интересной областью исследований. 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.

Ключевые слова: генерация музыки, сэмплирование, вариационный автокодировщик, MIDI формат

USING A VARIATIONAL AUTOENCODER MODEL TO GENERATE MUSIC

Eugeniy D. Mosin
Kaluga branch of the state budgetary educational institution of higher education "Bauman Moscow State Technical University (National Research University)"
Master's degree student
Lolko40rus@yandex.ru

Victor O. Fedorov
Kaluga branch of the state budgetary educational institution of higher education "Bauman Moscow State Technical University (National Research University)"
Candidate of Technical Sciences, Associate Professor
Fedorov_vo@bmstu.ru

ABSTRACT

Musical generation by neural networks is an interesting area of research. Scientists explore the capabilities of music generation models, investigate new algorithms and methods, and develop tools for creating and editing music. Neural networks for music generation can be integrated into synthesizers or music creation programs to offer users new ideas and variations of musical fragments.

Keywords: music generation, sampling, variational autoencoder, MIDI format

В эпоху больших данных спрос на короткие видеоролики и саундтреки к играм сильно вырос благодаря стремительному развитию стриминговых платформ. С одной стороны, пока она отвечает определенным стилистическим потребностям, людям не нужна музыка очень высокого художественного уровня; с другой стороны, спрос на этот тип музыки часто довольно высок, в то время как немногие музыканты захотят тратить свое время на написание большого количества малосодержательной музыки. Таким образом, автоматическая композиция с использованием технологий искусственного интеллекта и совместная композиция с музыкантами станут мейнстримом [1].

VAE (Variational Autoencoder / вариационный автокодировщик) — это модель глубокого обучения, которая может эффективно использоваться для жанровой генерации музыки.

VAE состоит из двух основных компонентов: кодировщика и декодировщика. Кодировщик преобразует входные данные, такие как аудиофрагменты или миди-сигналы, в скрытое представление низкой размерности, называемое латентным пространством [2]. Это латентное представление является статистическим распределением, определяемым параметрами среднего квадратичного отклонения и дисперсии.

Затем, используя методы сэмплирования, из латентного пространства выбирается случайный вектор, который передается в декодировщик. Декодировщик преобразует латентный вектор обратно в исходное пространство данных, производя генерацию новых музыкальных фрагментов [3].

Недавние достижения показали некоторый успех в музыкальных генерирующих моделях, которые позволяют пользователю больше контролировать генерацию музыкального контента. Например, большинство этих моделей интерактивны таким образом, что они позволяют учитывать такие атрибуты, как последовательность аккордов, плотность нот и ритмический стиль. Однако композиционный стиль сгенерированной музыки в основном определяется (ограничен) типом музыки в наборе обучающих данных, например, хоралы Баха, поп-музыка и джазовая музыка.

zc встраивает «контент» входной песни, а z_{cat} — это категориальная переменная, обозначающая стиль. Затем вложение для соответствующего стиля извлекается из кодовой книги стилей, что дает $z_{s.}$ Наконец, z_c и z_s объединяются вместе, чтобы сформировать z для декодировщика, чтобы синтезировать новую песню.

Кодировщик принимает входные данные x и преобразует их в скрытый вектор z с помощью двух RNN слоев. Затем с помощью параметров среднего μ и стандартного отклонения σ скрытого распределения, кодировщик вычисляет аппроксимированное апостериорное распределение $q(z|x)$ на скрытый вектор z.

Вместо использования градиентов для обратного распространения ошибки, в модели VAE используется прием репараметризации, который позволяет сделать модель дифференцируемой и получить градиенты. Этот прием заключается в генерации шума e из нормального распределения, и затем получении скрытого вектора z путем преобразования шума и параметров μ и σ, т.е. z = μ + eσ.

Декодировщик, также использующий два RNN слоя, получает скрытый вектор z и восстанавливает исходные данные x с помощью распределения p(x|z).

В модели VAE применяется априорное распределение p(z), которое придает скрытым переменным z определенный вид. В частности, это распределение определяет границы для значений z, что может способствовать лучшей обучаемости модели. Для минимизации ошибки модели используется функция потерь LV, которая состоит из двух компонент: первая компонента Lr отвечает за восстановление исходных данных, а вторая компонента βDkl отвечает за разницу между аппроксимированным апостериорным распределением q(z|x) и априорным распределением p(z). Значение β задает вес второй компоненты в функции потерь [4].

Таким образом, VAE позволяет получить непрерывное представление данных, сохраняя при этом их семантическую информацию. Эта модель может использоваться для задач генерации и реконструкции данных, а также для изучения скрытых признаков данных. Общая формулировка VAE выглядит следующим образом:

$$L_V = L_r - \beta D_{kl}(q(z|x)||p(z))$$

где Lr — потери при реконструкции, которые представляют собой перекрестную энтропию между реконструированными выходными данными $\bar{X}$ и X. Второй член — это расхождение Кульбака-Лейблера (KL) между апостериорным и априорным распределением. Оптимизация члена KL заставит скрытое распределение быть близким к распределению Гаусса. Термин β представляет собой вес, чтобы сбалансировать компромиссы между реконструкцией и термином KL.

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 [5].

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 zcat 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 z_{cat} 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 [6].

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 dims, что дает ему размер s × dims. 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 dims. 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 = μ + eσ. Затем z_s и z_c объединяются, чтобы сформировать z'. Наконец, z' проходит через линейный слой для формирования начального состояния декодировщика RNN.

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 z_c 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 βμ. Здесь βμ — это гиперпараметр, который определяет силу регуляризации. Член Lμ вынуждает выборочную дисперсию μ контролироваться на уровне βμ, 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 βμ — сдвиг, а N — размер партии. μ — это средний вектор, смоделированный кодировщиком для параметризации распределения z_c. Термин Lμ заставляет выборочную дисперсию μ контролироваться на уровне βμ, который поддерживает взаимную информацию X и z'.

В целом окончательная формулировка предложенной модели выглядит следующим образом:

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

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.

На рис. 1. показаны тенденции потерь от соперничества L(θdis) и L(θenc | θdis). На ранней стадии декодировщик учится восстанавливать мелодии на основе z_c, поэтому зеленая кривая уменьшается. Однако по мере продолжения рекламной процедуры z_c постепенно освобождается от сигналов, связанных с мелодией. Следовательно, декодировщик требуется все меньше и меньше релевантной информации для восстановления мелодии, и, таким образом, зеленая кривая увеличивается. Красная кривая демонстрирует обратный тренд. Когда каждая кривая потерь сходится, это интерпретируется как равновесие.

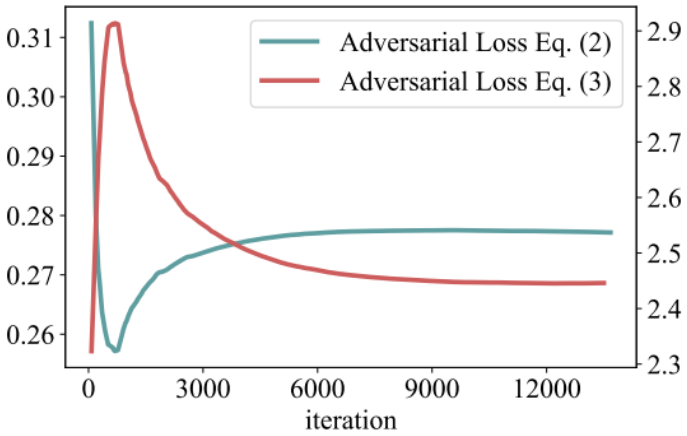

Рис. 1. График функций потерь при обучении

Для VAE ошибка восстановления является частью целевой функции, которая указывает, насколько хорошо модель декодирует свои скрытые коды в исходное пространство. Однако полное достижение поставленной цели не гарантирует получение высококачественных образцов [7]. VAE выдает много зашумленных выборок даже после достижения минимальной ошибки восстановления при $\beta \approx 0$. Это указывает на то, что апостериорное значение не соответствует гауссову. Проведя тщательный поиск, обнаружилось, что наилучшее условие с точки зрения музыкальных показателей $\beta = 1$.

Заключение

Модель вариационного автоэнкодера между последовательностями, которая позволяет пользователю определять стиль создаваемой выходной музыки включает непрерывное внедрение стилей для каждого стиля в наборе данных. Предлагаемый метод превосходит базовый уровень, который напрямую передает метки дискретного стиля в модель.

Список литературы:

1. K. Zhao, S. Li, J. Cai, H. Wang and J. Wang, An Emotional Symbolic Music Generation System based on LSTM Networks. – 2019. // IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chengdu, China, pp. 2039-2043.
2. K. Chen, W. Zhang, S. Dubnov, G. Xia and W. Li, The Effect of Explicit Structure Encoding of Deep Neural Networks for Symbolic Music Generation. – 2019. // International Workshop on Multilayer Music Representation and Processing (MMRP), Milan, Italy, pp. 77-84.

3. A. E. Memiş and v. H. Yalim Keles, Piano Music Generation with a Text Based Musical Note Representation using LSTM Models. - 2021. // 29th Signal Processing and Communications Applications Conference (SIU), Istanbul, Turkey, pp. 1-4.
4. H. H. Mao, T. Shin and G. Cottrell, DeepJ: Style-Specific Music Generation. – 2018. // IEEE 12th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, pp. 377-382.
5. Панина Е.А., Белов Ю.С. Анализ алгоритмов нахождения характерных точек изображений// Всероссийская научно-техническая конференция. – 2022. – Т.1. – С.55-57.
6. Мосин Е.Д., Белов Ю.С. Генерация музыки с использованием двунаправленной рекуррентной нейронной сети // Научное обозрение. Технические науки. 2023. № 1. с. 10-14.
7. Белоножко П.Е., Белов Ю.С. Модификации архитектуры WaveNet для реализации вокодера в генеративной модели преобразования текста в речь // Научное обозрение. Технические науки. 2022. № 6. с. 37-42.

References:

1. K. Zhao, S. Li, J. Cai, H. Wang and J. Wang, An Emotional Symbolic Music Generation System based on LSTM Networks. – 2019. // IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chengdu, China, pp. 2039-2043.
2. K. Chen, W. Zhang, S. Dubnov, G. Xia and W. Li, The Effect of Explicit Structure Encoding of Deep Neural Networks for Symbolic Music Generation. – 2019. // International Workshop on Multilayer Music Representation and Processing (MMRP), Milan, Italy, pp. 77-84.
3. A. E. Memiş and v. H. Yalim Keles, Piano Music Generation with a Text Based Musical Note Representation using LSTM Models. - 2021. // 29th Signal Processing and Communications Applications Conference (SIU), Istanbul, Turkey, pp. 1-4.
4. H. H. Mao, T. Shin and G. Cottrell, DeepJ: Style-Specific Music Generation. – 2018. // IEEE 12th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, pp. 377-382.
5. Panina E.A., Belov Yu.S. Analysis of algorithms for finding characteristic points of images // All-Russian Scientific and Technical Conference. – 2022. – T.1. – P.55-57.
6. Mosin E.D., Belov Yu.S. Music generation using a bidirectional recurrent neural network // Scientific Review. Technical science. 2023. No. 1. p. 10-14.
7. Belonozhko P.E., Belov Yu.S. Modifications of the WaveNet architecture to implement a vocoder in a generative model of text-to-speech conversion // Scientific review. Technical science. 2022. No. 6. p. 37-42.