Метрики качества генеративных моделей

К.И. Абросимов[1], Т.В. Львутина[1], А.С. Суркова[1]

[1] *Нижегородский государственный технический университет им. Р.Е. Алексеева, 24, ул.Минина, Нижний Новгород, 603095, Россия*

Аннотация

В рамках данной статьи рассматриваются современные метрики оценивания генеративных моделей. Особое внимание уделяется метрикам, которые применяются в области обработки естественного языка – BLUE (оценивает качество на основе сравнения полученного результата моделью и человека), NIST (основана на метрике BLUE), METEOR (основана на гармоническом среднем униграмм точности и полноты), ROUGE (. В статье представлена новая метрика, которая основана на субъективных оценках. Используемые в рассмотренной метрике субъективные оценки собираются с помощью попарного сравнения в виде шкал оценивания. Также в рамках статьи предложен алгоритм генерации музыки, построенный на основе автоматных моделей работы с ABC-нотацией, моделей дистрибутивной семантики и генеративных моделей глубоких нейронных сетей – Трансформеров. Новая метрика качества (SS-метрика), представленная в статье, применяется для оценки качества предложенного алгоритма генерации музыки в сравнении с решениями, которые предлагает человек и baseline-модели. Генерация музыки на основе baseline-модели строит продолжение музыкального фрагмента путем случайного выбора тактов из первой половины музыкального фрагмента. В ходе экспериментов удалось выяснить, что SS-метрика позволяет формализовать и обобщить субъективные оценки, это может быть использовано при оценке качества различных объектов.

Ключевые слова

Метрика, генеративные модели, анализ объектов сложной структуры, SS-метрика, генерация музыки, машинное обучение

Quality metrics of generative models

K.I. Abrosimov[1], T.V. Lvutina[1], A.S. Surkova[1]

[1] *Nizhny Novgorod State Technical University n.a. R.E. Alekseev (NNSTU), 24, Minin Street, Nizhny Novgorod, 603095, Russia*

Abstract

Within the framework of this article, modern metrics for evaluating generative models are considered. Particular attention is paid to metrics that are used in the field of natural language processing - BLUE (evaluates quality based on a comparison of the result obtained by a model and a person), NIST (based on the BLUE metric), METEOR (based on the harmonic mean of unigrams of accuracy and completeness), ROUGE (. The article presents a new metric, which is based on subjective assessments. The subjective estimates used in the considered metric are collected using pairwise comparison in the form of evaluation scales. The article also proposes an algorithm for generating music based on automatic models of working with ABC notation, models of distributive semantics and generative models of deep neural networks - Transformers. The new quality metric (SS-metric) presented in the article is used to assess the quality of the proposed algorithm for generating music in comparison with the solutions offered

ГрафиКон 2021: 31-я Международная конференция по компьютерной графике и машинному зрению, 27-30 сентября 2021 г., Нижний Новгород, Россия

EMAIL: abrosimov.kirill.1999@mail.ru (К.И. Абросимов); tat.lvutina@mail.ru (Т.В. Львутина); ansurkova@yandex.ru (А.С. Суркова)
ORCID: 0000-0001-9262-0474 (К.И. Абросимов); 0000-0002-2061-8858 (Т.В. Львутина); 0000-0003-0018-9053 (А.С. Суркова)

by humans and baseline models. Music generation based on the baseline model builds a continuation of a musical fragment by randomly selecting bars from the first half of the musical fragment. During the experiments, it was found out that the SS-metric allows you to formalize and generalize subjective assessments, this can be used to assess the quality of various objects.

Keywords
Metrics, generative models, analysis of objects of complex structure, SS-metric, music generation, machine learning

1. Введение

Задача оценивания картин, музыкальных произведений и текстов достаточно давно является предметом обсуждения научного сообщества. Первоначально подобными проблемами занимались специалисты соответствующих областях: искусствоведы, литературные и музыкальные критики и т.п. Однако с распространением компьютеров задачами получения численной оценки, отражающей объективные параметры объекта, стали заниматься математики-аналитики.

В последнее время широкое распространение получили такие генеративные модели, как генеративные состязательные сети (GAN), глубокие нейронные сети Трансформеры и другие техники, которые могут быть использованы для создания новых объектов, таких как изображения, текст или музыкальные произведения. В связи с этим одной из важных задач стала задача оценки качества генеративных моделей и в целом задача формализации и объективизации оценки генерируемых объектов.

2. Метрики качества генеративных моделей

Как уже было сказано, остро встает вопрос оценки качества генеративных моделей. Для классических задач существует исчерпывающее количество метрик, например: правильность, точность, полнота, и т. д., но вот с генеративными моделями все достаточно тяжело, ведь мы не просто хотим свериться с правильным ответом, но и понять, а не смогла ли она даже в чем-то превзойти человека.

Метрики можно классифицировать на две категории – субъективные метрики и объективные метрики. Субъективные метрики основаны на оценке восприятия человеком-наблюдателем, тогда как объективные метрики основаны на вычислительных моделях, которые пытаются сопоставить результат с некоторым эталоном. Субъективные метрики часто являются более «точными для восприятия», однако большая часть этих метрик неудобна, трудоемка или дорога для вычисления. Другая проблема заключается в том, что эти две категории метрик могут не соответствовать друг другу. Следовательно, исследователи часто анализируют результаты, используя метрики из обеих категорий.

2.1. Метрики качества NLP

В рамках задачи оценки музыкальных фрагментов, в том числе сгенерированных на основе ABC-нотации, можно пользоваться объективными метриками, которые используются в области обработки естественного языка (Natural Language Processing, NLP).

1. Метрика BLUE - оценивает качество на основе сравнения полученного результата моделью и человека, изначально применялась именно в оценке качества машинного перевода, центральной идей метрики являлась: «Чем ближе машинный перевод к профессиональному человеческому переводу, тем он лучше». Во-первых, вычисляется среднее геометрическое модифицированных n-граммовых точностей p_n, используя n-граммы до длины N и положительно-определенные веса, сумма которых ровняется единице. Во-вторых, вычисляется штраф за краткость BP, на основе длины результата модели с и эффективной длиной, на основе

корпуса r. В результате произведения штрафа за краткость и экспоненты сум, получаем значение метрики [1], вычисляемой по формулам

$$p_n = \frac{\sum_{C\epsilon\{Candidates\}} \sum_{n\_gram\epsilon C} Count_{clip}(n\_gram)}{\sum_{C`\epsilon\{Candidates\}} \sum_{n\_gram`\epsilon C`} Count_{clip}(n\_gram`)} \quad (1)$$

$$BP = \begin{cases} 1, & if\ c > r \\ e^{(1-\frac{r}{c})}, & if\ c \leq r \end{cases} \quad (2)$$

$$BLUE = BP * \exp(\sum_{i=1}^{n} w_n log p_n) \quad (3)$$

2. Метрика NIST - основанная на метрике BLUE, с разницей лишь в том, где BLUE при вычислении точности n-грамм использует не одинаковый вес w_n, а вычисляет его, насколько информативный конкретный n-грамм. Например, если биграмма «но и» найдена, то она получит меньший вес, чем биграмма «вычислительная лингвистика», так как у первой биграммы больше априорная вероятность.

3. Метрика METEOR - основанная на гармоническом среднем униграмм точности и полноты. Она также имеет несколько особенностей, которые не встречаются в других метриках, таких как совпадение стемминга и синонимии, а также стандартное точное совпадение слов. Появился новый штраф: более длинные n-граммовые совпадения используются для вычисления штрафа p за выравнивание. Чем больше сопоставлений, которые не являются смежными в ссылке и предложении-кандидате, тем выше будет штраф, что позволяет учитывать конгруэнтность не только по отношению к отдельным словам, но и по отношению к более крупным сегментам [2]. Метрика METEOR вычисляется по формулам

$$P = \frac{\text{количество n} - \text{грамм в сгенерированном объекте, которые также были найдены в эталонном объекте}}{\text{количество n} - \text{грамм в сгенерированном объекте}} \quad (4)$$

$$R = \frac{\text{количество n} - \text{грамм в сгенерированном объекте, которые также были найдены в эталонном объекте}}{\text{количество n} - \text{грамм в эталонном объекте}} \quad (5)$$

$$F = \frac{10PR}{R + 9P} \quad (6)$$

$$p = \left(\frac{\text{число групп n} - \text{грамм}}{\text{количество n} - \text{грамм, которые объединили в группы(словосочетания)}}\right)^3 \quad (7)$$

$$METEOR = F(1 - p) \quad (8)$$

4. Семейство метрик ROUGE. Рассмотрим две основные метрики из данного семейства: ROUGE-N и ROUGE-SKIP2. Формально, ROUGE-N-это n-граммовая полнота между кандидатом и реальным продолжением, которая вычисляется как отношение полученных n-грамм кандидата к максимальному количеству n-граммов, которые есть и у кандидата, и у эталона. ROUGE-SKIP2. Скип – биграмма - это любая пара слов в их порядке предложения, допускающая произвольные пробелы. Статистика совпадений скип-биграмм измеряет перекрытие скип-биграмм между кандидатом и эталоном [3].

2.2. Метрики качества музыкальных произведений

Однако, все эти метрики имеют главный минус – все они сравниваются с некоторым эталоном, а в музыке достаточно тяжело найти такой «эталон», ведь даже тому, что написал

человек нельзя поставить такую метку, так как завтра найдется другой человек, который, возможно, сделает лучше. Поэтому необходимо определить иную метрику.

Одной из первых работ по формализации оценок качества музыкальных произведений стала статья «Оценка сложности песен» [4], в которой рассмотрена эволюция популярных песен с точки зрения теории вычислительной сложности.

2.3. Scale Score метрика (SS-метрика)

Однако, все эти метрики имеют главный минус – все они сравниваются с некоторым эталоном, а в музыке достаточно тяжело найти такой «эталон», ведь даже тому, что написал человек нельзя поставить такую метку, так как завтра найдется другой человек, который, возможно, сделает лучше. Поэтому необходимо определить иную метрику.

Одной из первых работ по формализации оценок качества музыкальных произведений стала статья «Оценка сложности песен» [4], в которой рассмотрена эволюция популярных песен с точки зрения теории вычислительной сложности.

У каждого человека есть свой собственный музыкальный вкус, который зависит от огромного количества факторов, но особенно от характера человека [5,6,7]. Поэтому давать оценку конкретному произведению разным людям нецелесообразно, однако давать на оценку пару произведений, покажет, что именно больше нравится человеку в сравнении.

Поэтому была предложена метрика попарной оценки объектов, названная Scale Score метрика (SS-метрика). SS-метрика рассчитывается по формуле

$$\mathrm{SS}_{ke}{}^{scale} = \frac{\sum_{i=1}^{n} score_{ke_i}}{n * (\min(scale) + \frac{\max(scale) - \min(scale)}{2})} \quad (9)$$

где n - количество оценок, k - предлагаемое решение, относительно решения e; scale - выбранная шкала с неотрицательными оценками.

Данная метрика возвращает значение от 0 до 2.

0 – решение k абсолютно хуже решения e.

1 – решение k сопоставим решению e.

2 – решение k абсолютно лучше решения e.

3. Расчеты Scale Score метрики при оценивании музыкальных фрагментов

Были проведены эксперименты для оценки качества музыкальных фрагментов. Генерировались фрагменты алгоритмом на основе ABC-нотации и дистрибутивной семантики [8]. Алгоритм состоит из нескольких моделей:

1. С помощью автоматной модели происходит разделение ABC-нотации на мелодические и ритмические конструкции.
2. В мелодических конструкциях, с помощью поиска коллокаций, находятся часто встречаемые аккорды, которые объединяются в пару.
3. С помощью моделей дистрибутивной семантики (Word2Vec (CBOW) и FastText) производится векторизация мелодических и ритмических конструкций.
4. Полученные сжатые векторные представления подаются генеративным моделям - Трансформерам.
5. Полученные продолжения с помощью еще одной автоматной модели сцепляются в ABC-нотацию, в результате получаем продолжение музыкального фрагмента.

Датасет представлен компанией YandexCloud, в котором представлены 182 000 музыкальных произведений различных жанров, тональностей, длительностей в формате ABC-нотаций [9]. Для

тестирования и оценки модели были выбраны 100 ABC-нотации из указанного датасета. На Рисунке 1, представлены круговые диаграммы тестовой репрезентативной выборки.

Рисунок 1: Характеристики тестовой выборки

Для оценки качества полученного алгоритма использовалась предложенная SS-метрика: людям предлагали попарно сравнить три фрагмента, у которых первая половина была одинаковой. Вторая половина отличалась: первый фрагмент - это продолжение, которое придумал человек, второй фрагмент - это то, что сгенерировал предложенный алгоритм, описанный выше, третий фрагмент - это продолжение, сгенерированное baseline-моделью, которая в случайном порядке с повторениями выбирала такты из первой половины произведения.

Для удобства сбора субъективных оценок был разработан web-сервис, который автоматизирует и упрощает пользователю взаимодействовать с музыкальными фрагментами, а также равномерно распределять их между самими пользователями.

Была выбрана шкала от 0 до 4, то есть SS-метрика для сравнения объектов будет рассчитывается по формуле

$$\mathrm{SS}_{model-human}{}^{[0..4]} = \frac{\sum_{i=1}^{n} score_{model-human_i}}{n * 2} \tag{10}$$

Интересное свойство данной метрики: сумма SS-метрика для решения A, относительно решения Б и SS-метрика для решения Б, относительно решения А равна 2, или $\mathrm{SS}_{model-human}{}^{[0..4]} = 2 - \mathrm{SS}_{human-model}{}^{[0..4]}$.

Для каждого значения шкалы предусмотрен следующий смысл:

- **score = 0** - предлагаемое решение k явно хуже решения e;
- **score = 1** - предлагаемое решение k скорее хуже решения e;
- **score = 2** - предлагаемое решение k также хорошо/плохо как решение e;
- **score = 3** - предлагаемое решение k скорее лучше решения e;
- **score = 4** - предлагаемое решение k явно лучше решения e.

На Рисунках 2–4 представлены посчитанные SS-метрики для каждого музыкального фрагмента, на которых считались результирующие метрики, которые рассмотрены выше.

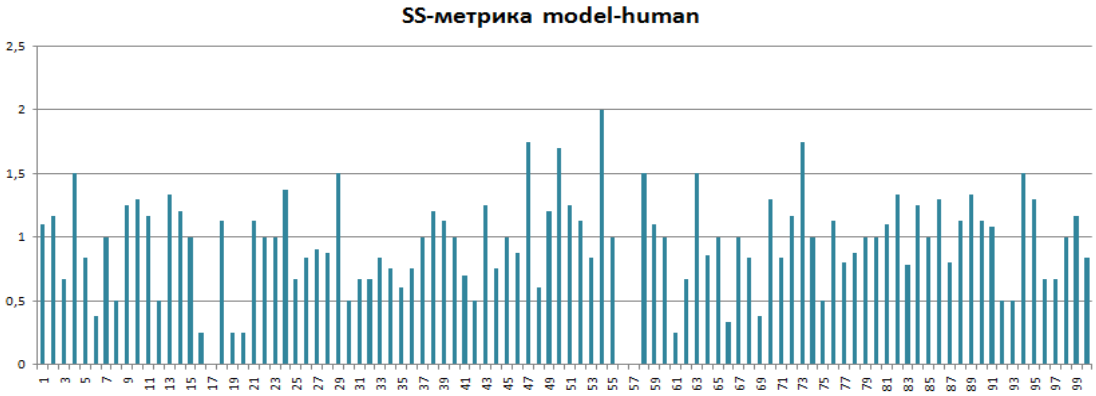

Рисунок 2: SS-метрика model-human по каждому фрагменту

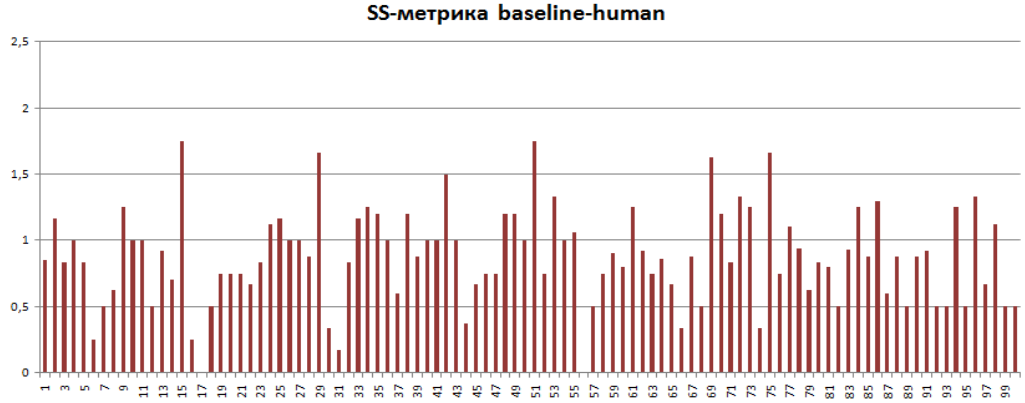

Рисунок 3: SS-метрика baseline-human по каждому фрагменту

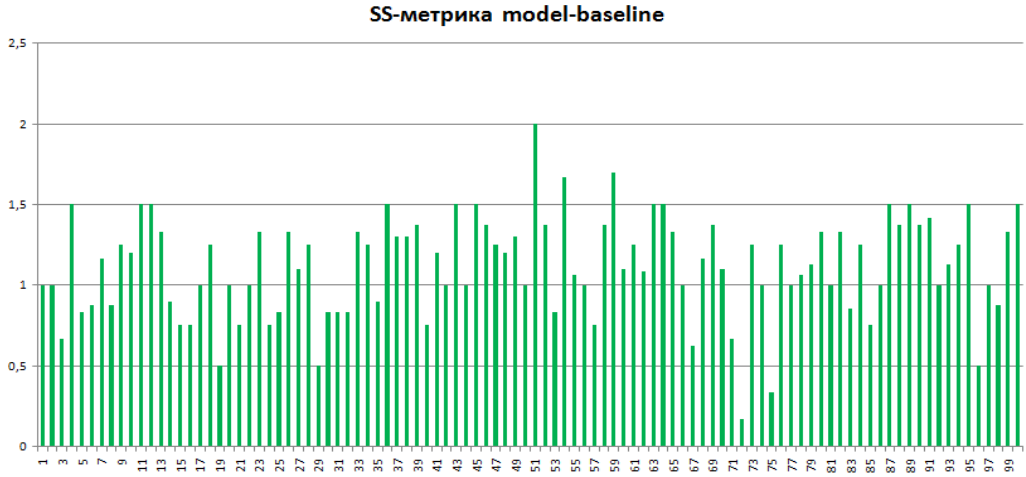

Рисунок 4: SS-метрика model-baseline по каждому фрагменту

Предложенный алгоритм показал себя хорошо, что видно по Рисунку 5, результирующая метрика показала, что предложенный алгоритм сгенерировал лучше музыкальные фрагменты, чем baseline-модель, основанная на случайном выборе тактов из подаваемого начала. Однако, алгоритм все-таки хуже предлагал продолжения, чем написанные фрагменты человеком, но тем не менее метрика очень близка к 1, это означает, что модель генерирует чуть хуже продолжения, чем человек, но все равно довольно достойно.

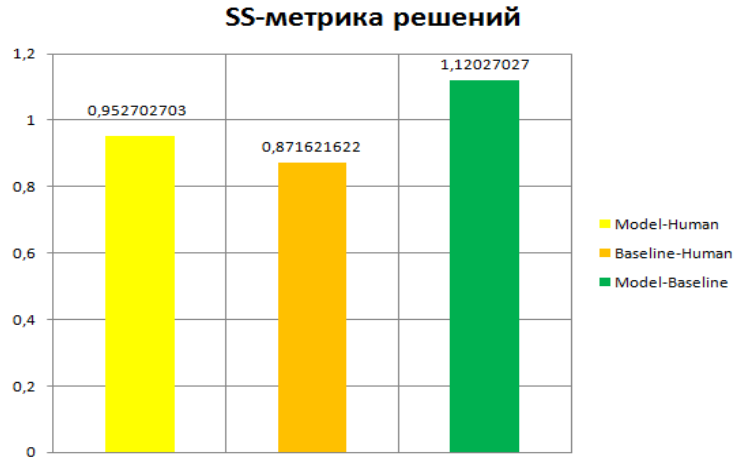

Рисунок 5: Значение SS-метрики для разных алгоритмов

4. Заключение

В настоящее время все быстрее развиваются технологии генерации контента, которые крайне тяжело оценить. В рамках этой статьи были рассмотрены основные метрики оценок моделей NLP, описана предложенная SS-метрика, а также применена для оценивания предложенного алгоритма генерации музыки.

Предложенная метрика позволяет формализовать и обобщить субъективные оценки, что может быть использовано при оценивании качества различных объектов, таких как музыкальных произведений, изображений, текстов, в том числе полученных генеративными методами.

5. References

[1] K. Papineni, S. Roukos, T. Ward, W. J. Zhu, BLEU: a method for automatic evaluation of machine translation. ACL-2002: 40th Annual meeting of the Association for Computational Linguistics, 2002, p. 311–318.

[2] S. Banerjee, A. Lavie, «METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments» in Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005), 2005.

[3] Lin, Chin-Yew. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, 2004.

[4] D. Knuth, The Complexity of Songs, SIGACT News, 1977, p.17–24.

[5] R.A. Brown, Music preferences and personality among Japanese university students // International Journal of Psychology: journal, volume 47, 2012, p. 259–268.

[6] Langmeyer, Alexandra; Guglhör-Rudan, Angelika & Tarnai, Christian. What do music preferences reveal about personality: a cross-cultural replication using self-ratings and ratings of music samples // Journal of Individual Differences: journal, volume 33, 2012, p. 119–130.

[7] Chamorro-Premuzic, Tomas; Gomà-i-Freixanet, Montserrat, Furnham, Adrian & Muro, Anna. Personality, self-estimated intelligence, and uses of music: A Spanish replication and extension using structural equation modeling // Psychology of Aesthetics, Creativity, and the Arts: journal, volume 3, 2009, p. 149–155.

[8] К. И. Абросимов, А.С. Суркова, Алгоритм генерации музыки на основе ABC-нотации и дистрибутивной семантики // Информационные системы и технологии ИСТ-2021. Сборник материалов XXVII Международной научно-технической конференции. Нижегородский государственный технический университет им. Р.Е. Алексеева. 2021. С. 906–912.

[9] Хакатон-соревнование от Yandex.Cloud. URL: https://ds2020.ai-community.com/.