УДК 004.8

НЕЙРОННАЯ СЕТЬ ДЛЯ ГЕНЕРАЦИИ ИЗОБРАЖЕНИЙ НА ОСНОВЕ ТЕКСТА ПЕСЕН С ПРИМЕНЕНИЕМ МОДЕЛЕЙ OPENAI И CLIP

А. Р. Давлетгареева[1] [0009-0008-7258-470X], К. А. Едкова[2] [0009-0005-4706-2254]

[1,2] *Институт информационных технологий и интеллектуальных систем Казанского федерального университета, ул. Кремлевская, 35, г. Казань, 420008*

[1]alsudavletgareeva@gmail.com, [2]ksushka.e21@gmail.com

Аннотация

Исследована эффективность моделей ImageNet diffusion model и CLIP для генерации изображений по текстовому описанию. С использованием различных текстовых вводов на разных параметрах проведены два эксперимента для определения лучших параметров при генерации изображений на основе текстового описания. Результаты показали, что, хотя ImageNet хорошо справляется с созданием изображений, CLIP лучше обеспечивает соединение текстовых подсказок с релевантными изображениями. Полученные результаты характеризуют высокий потенциал объединения названных моделей для создания высококачественных и контекстно релевантных изображений на основе текстового описания.

Ключевые слова*: генерация изображений, ImageNet diffusion model, CLIP, глубокое обучение, нейронные сети, обработка естественного языка.*

ВВЕДЕНИЕ

Искусственный интеллект (ИИ) в области генерации изображений развивался с первых дней возникновения машинного обучения (МО). Первые попытки создания изображений на основе текстовых описаний были предприняты с использованием систем, основанных на правилах, которые были ограничены в своей способности улавливать нюансы и сложности человеческого языка [1]. Однако с появлением методов глубокого обучения, таких как генеративные состязательные сети (GAN) и вариационные автокодеры (VAEs) [2], в этой области наблюдается значительный прогресс.

Генерация изображений на основе текстового описания привлекла значительное внимание в последние годы благодаря своему потенциалу генерировать высококачественные изображения на основе текстовых описаний [2].

Одним из наиболее популярных подходов к созданию изображений из текста является использование сетей GAN [3], которые состоят из двух нейронных сетей: сети-генератора, которая генерирует изображения, и сети-дискриминатора, которая пытается отличить реальные изображения от сгенерированных. Объединяя эти сети в игровой среде, GAN способны создавать высококачественные изображения, которые трудно отличить от реальных.

Еще одним популярным направлением в этой области исследований стала разработка крупномасштабных языковых моделей, таких как GPT-3 [4]. Эти модели способны генерировать текст, который практически неотличим от текста, написанного человеком, и могут использоваться также для генерации текстовых описаний изображений. Это открыло новые возможности для исследований в области генерации изображений, поскольку позволяет генерировать изображения на основе описаний на естественном языке, а не полагаться на заранее определенные правила или шаблоны.

Отметим, что способность генерировать изображения на основе текстовых описаний значительно расширилась за последние годы благодаря достижениям в области методов глубокого обучения и крупномасштабных языковых моделей [2, 5]. Хотя предстоит преодолеть еще много проблем, таких, например, как создание изображений, точно отражающих нюансы человеческого языка, будущее этой области выглядит многообещающим, и можно ожидать много интересных разработок в ближайшие годы [6].

В настоящей статье мы предлагаем подход к созданию изображений на основе текстов песен и используем два алгоритма из репозитория OpenAI [7] для генерации изображений, которые строятся на основе текста. Чтобы достичь желаемого результата, мы представляем текст песни в виде последовательности слов, выделяя ключевые выражения, и кодируем их, используя предварительно обученную языковую модель.

АРХИТЕКТУРА НЕЙРОННОЙ СЕТИ

Для создания изображения на основе текста песни была предложена архитектура нейронной сети (Рисунок 1), построенная на основе генеративно-состязательной нейросети (GAN). GAN – это один из алгоритмов классического МО – обучения без учителя. Его суть заключается в комбинации двух нейросетей: генератора и дискриминатора [8]. Задача генератора (Рисунок 1а) заключается в генерации образов заданной категории, а задача дискриминатора – в распознавании созданных образов.

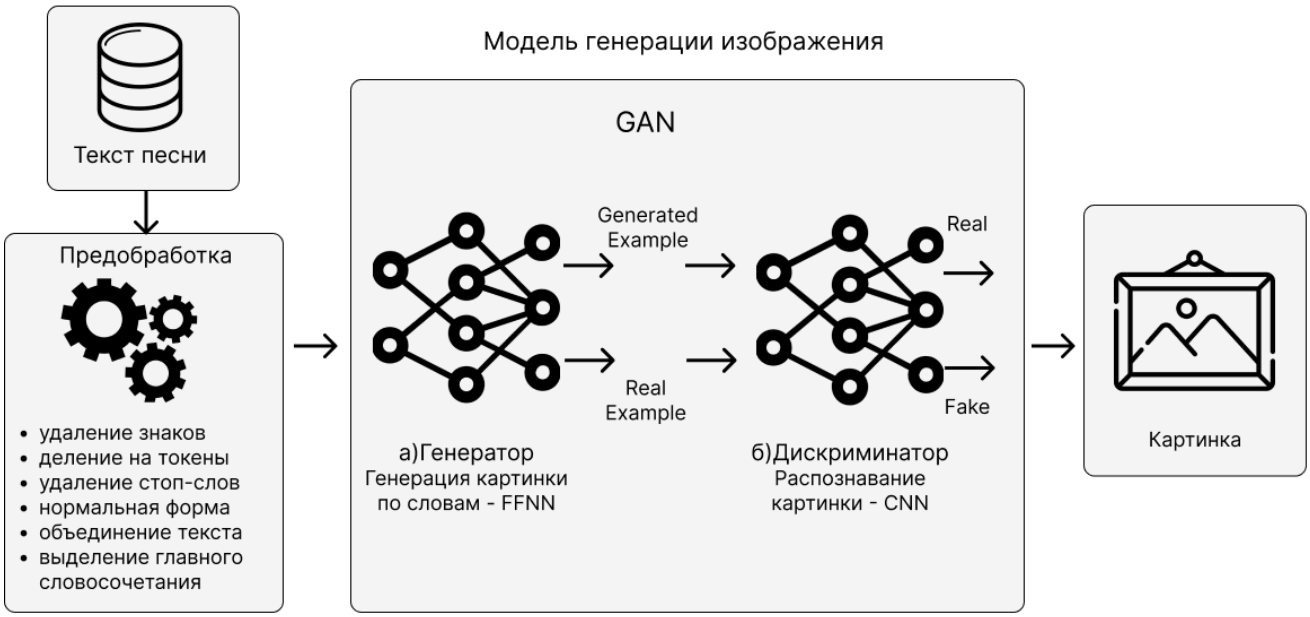

Рис. 1. Архитектура нейронной сети №1

Дискриминатор (Рисунок 1б) использует сверточные нейронные сети (CNN) [9] для распознавания образов на изображении. Чтобы обучить нейронную сеть распознавать образы, необходимо обработать большое количество изображений, на которых присутствуют искомые образы. Генератор начинает формирование изображений с создания произвольного шума, на котором постепенно появляются фрагменты искомого изображения. В качестве генерирующей нейронной сети могут использоваться сети FFNN – нейронные сети прямого распространения [10].

Шаги, которые проходит GAN: генератор получает случайное число и возвращает изображение, а затем передает это сгенерированное изображения дискриминатору вместе с потоком изображений, взятых из фактического набора данных. Дискриминатор принимает набор из реальных и сгенерированных изображений и возвращает вероятности, 0 или 1, где 1 – изображение подлинное, а 0 – фальшивое.

Из-за ограничений нашего вычислительного оборудования мы не смогли использовать предложенную ранее архитектуру нейронной сети для генерации изображений по тексту песни. В связи с этим мы обратили внимание на недавние достижения в области искусственного интеллекта, особенно на решения от компании OpenAI.

Одним из таких решений является модель CLIP (Contrastive Language-Image Pre-Training) [11], которая была разработана для связи естественного языка и изображений. Эта модель использует большой корпус данных текстов и изображений для обучения своей архитектуры, что позволяет ей понимать связь между текстом и изображением. Благодаря этому CLIP может использоваться для различных задач, в том числе для генерации изображений по тексту.

Вторым решением в репозитории компании OpenAI является модель ImageNet diffusion [12], которая позволяет генерировать высококачественные изображения с помощью диффузии. Она обучается на большом наборе изображений (1,3 миллиардах изображений из набора данных ImageNet), что позволяет ей генерировать разнообразные изображения с высокой детализацией и качеством: пиковое отношение сигнал/шум (PSNR) – диффузионная модель ImageNet генерирует изображения с разрешением до 40 дБ, Индекс структурного сходства (SAM) – генерирует изображения с суммой до 0,95, средняя оценка мнений (MOS) – до 5.

Диффузионные модели – это тип генерирующих моделей, которые обучают генерировать изображения путем постепенного добавления шума к изображению, а затем итеративного удаления шума до тех пор, пока изображение не станет неотличимым от реального изображения. Поэтому, в качестве альтернативы модели GAN, мы решили применить интеграцию решений от OpenAI, такие как

модели CLIP (Рисунок 2а) и ImageNet (Рисунок 2б) diffusion для генерации изображений по тексту песни (Рисунок 2). Эти модели показали высокую эффективность в задачах генерации изображений и позволяют создавать качественные визуальные интерпретации текстовых данных [12].

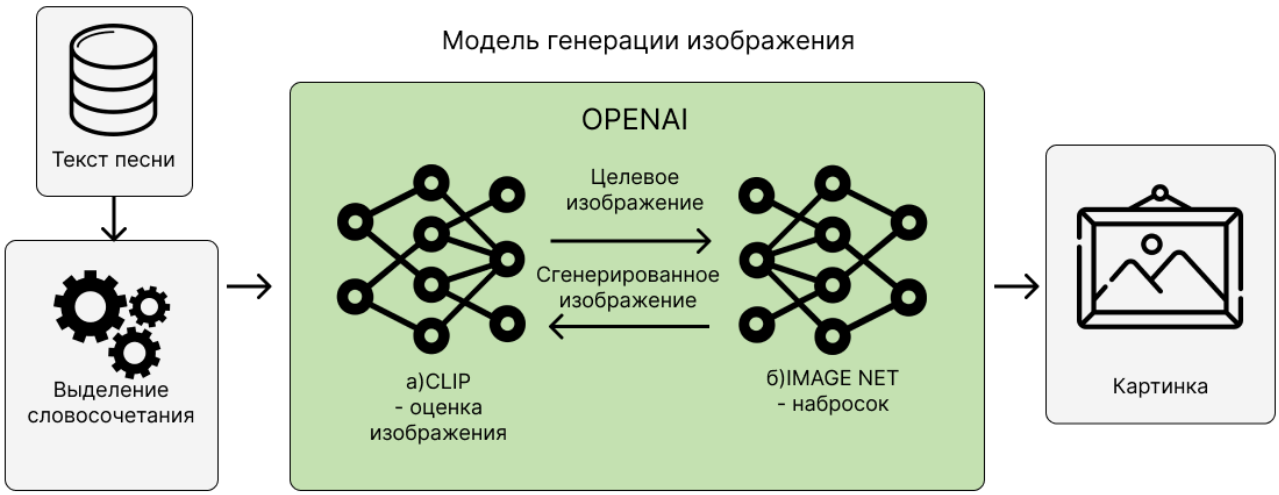

Рис. 2. Архитектура нейронной сети №2

ImageNet – это нейронная сеть, предназначенная для генерации изображений с высоким разрешением. Для генерации изображения использовалась обученная диффузионная модель ImageNet 256 x 256, которая превосходит [12] генеративную модель GAN при генерации изображений.

CLIP [11] – это нейросеть обученная на парах текст-изображение. Таким образом CLIP выстраивает ассоциации на основе текстового описания с соответствующими изображениями.

ЭКСПЕРИМЕНТЫ И ИХ РЕЗУЛЬТАТЫ

Были проведены два эксперимента на основе различных параметров, для каждого эксперимента использовались 10 текстовых выборок. Модель ImageNet diffusion была обучена на большом наборе данных изображений для генерации качественных изображений из текстовых подсказок. Нейросеть CLIP, с другой стороны, была обучена понимать взаимосвязь между текстом и изображениями, что позволило ей найти наиболее релевантное изображение для данного текстового

описания. Сгенерированные изображения оценивались на основе их качества и соответствия текстовому описанию**.**

Мы подобрали песни из различных музыкальных жанров на основе личных предпочтений, а при выборе словосочетаний руководствовались их смысловой составляющей, предпочитая использовать конкретные выражения, которые могут вызывать у большинства людей представление ясной и определенной картинки, а не абстрактные конструкции (см. Таблицу 1).

Таблица 1. Песни, выбранные для проведения экспериментов.

№	Исполнитель	Название песни	Выбранный отрывок
1	Smash Mouth	All Star	“shooting stars”
2	Rufus Wainwright	Hallelujah	“hallelujah”
3	The neighborhood	Sweater Weather	“sweater weather”
4	Sia	Snowman	“cry snowman”
5	Sia	Courage To Change	“news on TV”
6	Lady Gaga	Bloody Mary	“bloody mary”
7	Ariana Grande	7 rings	“bottles of bubbles”
8	Artem Kolpakov	The Blue Tractor	“blue tractor on big wheels”
9	Ryan Gosling	City Of Stars	“crowded restaurant”
10	Mark Philippe	Dancer in the Dark	“Dancer in the dark”

Для каждого отрывка было проведено по 2 эксперимента. Было замечено, что с определенными параметрами генерируемые изображения абсолютно не соответствуют предполагаемому результату. В экспериментах рассматривались 2 параметра: clip_guidance_scale и range_scale.

Параметр clip_guidance_scale отвечает за масштабирование влияния функции потерь, связанной с моделью CLIP. Модель CLIP используется для оценки сходства между сгенерированными изображениями и их текстовыми описаниями. Чем выше значение clip_guidance_scale, тем больший учет будет приниматься вклад функции потерь от модели CLIP при оптимизации генерации изображений. Более высокое значение этого параметра означает, что сгенерированные изображения будут более точно соответствовать заданным текстовым описаниям.

Параметр range_scale отвечает за масштабирование влияния функции потерь, связанной с диапазоном значений пикселей в сгенерированных изображениях. Функция потерь, связанная с диапазоном, гарантирует, что значения пикселей в сгенерированных изображениях остаются в допустимом диапазоне, например, от 0 до 1. Чем выше значение range_scale, тем больше будет учитываться влияние функции потерь, связанной с диапазоном, при оптимизации генерации изображений, и сгенерированные изображения будут более ограничены в диапазоне значений пикселей.

Для экспериментов значения параметров были выбраны случайным образом, а именно: для первого эксперимента clip_guidance_scale = 1000, range_scale = 100; для второго эксперимента clip_guidance_scale = 1500, range_scale = 150.

Рассмотрим пример с генерацией изображения по тексту “shooting stars”. На шагах 4–6 (Рисунок 3) можно было предположить, что алгоритм пытается нарисовать морскую звезду, т. е. он выделяет из словосочетания слово “stars”. Далее с каждым шагом изображение всё меньше похоже на звезду, и результат (Рисунок 4) уже мало похож на морскую звезду, не говоря уже о падающих (“shooting”) звездах.

Рис. 3. Результаты генерации изображения для “shooting stars” на шагах 4–6

Рис. 4. Результат эксперимента “shooting stars”

Далее на Рисунке 5 приведены изображения, сгенерированные в результате двух экспериментов для каждой песни. Номер на изображении расшифровывается следующим образом:

<номер песни из таблицы 1>.<номер эксперимента>.

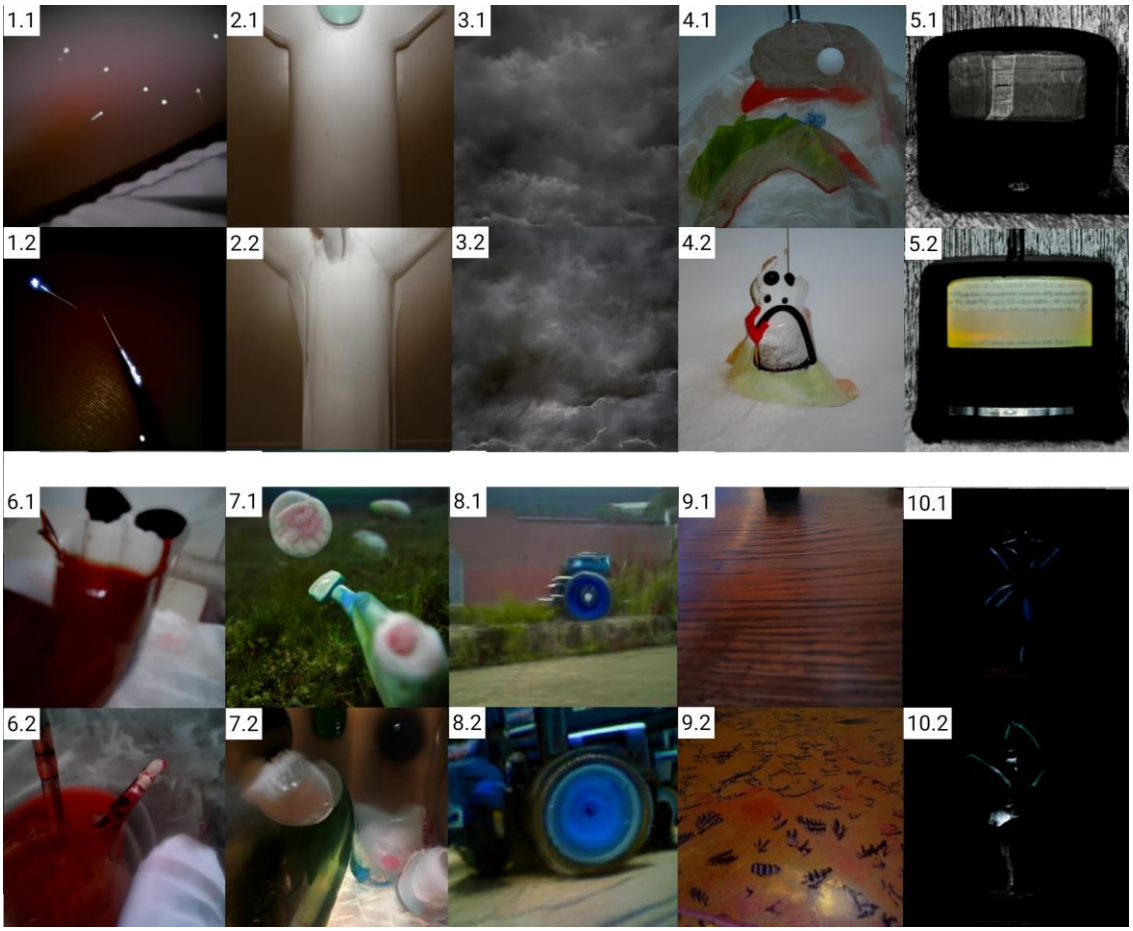

Рис. 5. Результаты работы алгоритма.

Рассмотрим результаты генерации изображений (Рисунок 5):

- 1.1 и 1.2: можно заметить падающие звезды, при этом картинка 1.1 как будто больше похожа на “shooting stars”;
- 2.1 и 2.2: учитывая, что выражение “hallelujah” очень абстрактно, не совсем понятно, что именно пыталась изобразить нейросеть;
- 3.1 и 3.2: “sweater weather” можно перевести как прохладная погода (погода для надевания свитера), алгоритм сгенерировал мрачную погоду, тучи, молнии;
- 4.1 и 4.2: не похожи на снеговика, на первой картинке – пингвин, на втором – что-то отдаленно похожее на снеговика;
- 5.1 и 5.2: можно увидеть телевизор, на втором изображении более четкий текст;
- 6.1 и 6.2: просматривается коктейль «Кровавая Мэри»;

- 7.1 и 7.2: если постараться, то можно разглядеть пузырьки и бутылку, но в целом получилось не очень понятное изображение;
- 8.1 и 8.2: на первой картинке нейронная есть сгенерировала что-то похожее на трактор, на второй можно увидеть только часть картинки;
- 9.1 и 9.2: на обоих изображениях можно разглядеть какой-то узор, что совершенно не подходит под описание “crowded restaurant”, что в переводе означает «переполненный ресторан»;
- 10.1 и 10.2: на первой картинке можно увидеть контур балерины-танцовщицы в темноте, на второй же более четкий силуэт;

После завершения экспериментов мы приняли решение провести опрос, чтобы оценить соответствие визуального восприятия людей относительно выражений, выделенных в тексте, с изображениями, созданными нейронной сетью. Для оценки сгенерированных изображений использовалась шкала от 0 до 10, где 0 – нет схожести, а 10 – точно совпадает. Опрашиваемая группа: 18 человек в возрасте от 18 до 67. Среди них 2 мужчин и 16 женщин. Результаты опроса представлены в Таблице 2.

Таблица 2. Анализ оценки результатов экспериментов по опросу

	среднее значение	максимум	минимум	среднее отклонение	дисперсия	медиана
1.1 НАЗВАНИЕ - All Star из мультфильма «Шрэк» Smash Mouth ОТРЫВОК - shooting stars Параметры: 1000	7,27	10,00	3,00	1,88	6,02	7,14
1.2 НАЗВАНИЕ - All Star из мультфильма «Шрэк» Smash Mouth ОТРЫВОК - shooting stars Параметры: 1500	6,82	10,00	2,00	2,05	6,56	6,91
1.3 НАЗВАНИЕ - All Star из мультфильма «Шрэк» Smash Mouth ОТРЫВОК - shooting stars Параметры: 500	1,82	8,00	**0,00**	1,80	6,16	1,81
2.1 НАЗВАНИЕ - Hallelujah из мультфильма «Шрэк» Rufus Wainwright ОТРЫВОК - Hallelujah Параметры: 1000	4,18	9,00	**0,00**	3,26	12,76	2,63

2.2 НАЗВАНИЕ - Hallelujah из мультфильма «Шрэк» Rufus Wainwright ОТРЫВОК - Hallelujah Параметры: 1500	4,36	9,00	**0,00**	2,69	9,25	2,35
3.1 НАЗВАНИЕ - The neighbourhood Sweather Weather ОТРЫВОК - Sweather weather Параметры:1000	7,36	10,00	4,00	2,33	7,05	4,50
3.2 НАЗВАНИЕ - The neighbourhood Sweather Weather ОТРЫВОК - Sweather weather Параметры:1500	**8,00**	10,00	4,00	1,82	4,60	7,00
4.1 НАЗВАНИЕ - Snowman Sia ОТРЫВОК - cry snowman Параметры: 1000	4,45	10,00	1,00	2,58	9,87	3,29
4.2 НАЗВАНИЕ - Snowman Sia ОТРЫВОК - cry snowman Параметры: 1500	6,45	10,00	2,00	2,23	6,87	5,23
5.1 НАЗВАНИЕ - Courage to Change Sia ОТРЫВОК - news on TV Параметры: 1000	7,18	10,00	2,00	2,50	8,36	6,09
5.2 НАЗВАНИЕ - Courage to Change Sia ОТРЫВОК - news on TV Параметры: 1500	**8,27**	10,00	3,00	1,34	4,02	8,00
6.1 НАЗВАНИЕ - Bloody Mary Lady Gaga ОТРЫВОК - Bloody Mary Параметры: 1000	5,64	10,00	**0,00**	2,58	9,85	3,50
6.2 НАЗВАНИЕ - Bloody Mary Lady Gaga ОТРЫВОК - Bloody Mary Параметры: 1500	6,91	10,00	2,00	2,48	8,69	7,45
7.1 НАЗВАНИЕ - 7 rings Ariana Grande ОТРЫВОК - bottles of bubbles Параметры: 1000	5,00	8,00	3,00	1,45	3,20	4,50
7.2 НАЗВАНИЕ - 7 rings Ariana Grande ОТРЫВОК - bottles of bubbles Параметры: 1500	5,00	9,00	1,00	2,18	7,00	5,00
8.1 НАЗВАНИЕ - Blue tractor ОТРЫВОК - blue tractor on big wheels Параметры: 1000	6,82	10,00	3,00	1,87	5,16	6,91

8.2 НАЗВАНИЕ - Blue tractor ОТРЫВОК - blue tractor on big wheels Параметры: 1500	**8,09**	10,00	5,00	1,04	2,49	8,00
9.1 НАЗВАНИЕ - City Of Stars из фильма «Ла-Ла Ленд» Ryan Gosling, Emma Stone ОТРЫВОК - crowded restaurants Параметры: 1000	3,18	8,00	**0,00**	2,08	7,76	3,00
9.2 НАЗВАНИЕ - City Of Stars из фильма «Ла-Ла Ленд» Ryan Gosling, Emma Stone ОТРЫВОК - crowded restaurants Параметры: 1500	3,82	8,00	**0,00**	2,38	8,36	4,91
10.1 НАЗВАНИЕ - Dancer in the Dark Marc philippe ОТРЫВОК - Dancer in the dark Параметры: 1000	7,64	9,00	4,00	1,32	3,45	7,82
10.2 НАЗВАНИЕ - Dancer in the Dark Marc philippe ОТРЫВОК - Dancer in the dark Параметры: 1500	***8,45***	10,00	2,00	1,52	5,47	8,23

В проведенном опросе по оценке качества генерируемых изображений получены следующие результаты:

- Максимальная оценка изображения составила 10, а минимальная – 0.
- Среднее отклонение для всех песен составило примерно 2–3, а медиана – от 4 до 7.
- Дисперсия для всех песен была выше 4.
- Самое высокое среднее значение оценки (8,27) было достигнуто при параметрах clip_guidance_scale = 1500 для песни "Courage to Change" исполнителя Sia, текст “news on TV”.
- Самое низкое среднее значение оценки (1,81) было получено при параметрах clip_guidance_scale = 500 для песни "All Star" исполнителя Smash Mouth текст “shooting stars”.
- Качество генерируемых изображений зависит от параметров генерации. Например, для большинства песен при значении clip_guidance_scale 1500 средняя оценка изображения выше, чем при значении 500. Однако есть исключения, например, для песни "All Star" с отрывком "shooting stars" при параметрах 500 средняя оценка значительно ниже, чем при других параметрах.

- Некоторые изображения получили в целом высокие оценки (например, по тексту для "Sweather Weather" и "Courage to Change"), тогда как другие изображения респонденты оценили гораздо ниже (например, "All Star" от Smash Mouth при параметрах 500).
- Среднее отклонение и дисперсия могут указывать на то, что в некоторых случаях мнения о качестве генерируемых изображений сильно расходятся. Например, для песни "Hallelujah" от Rufus Wainwright при параметрах 1000 оценки достаточно разнообразны, среднее отклонение составляет 3,26, а дисперсия – 12,76.
- Медиана может быть более надежным показателем центральной тенденции, особенно когда данные распределены неравномерно. Например, для песни "All Star" от Smash Mouth при параметрах 1000 медиана равна 7,14, что указывает на то, что большинство оценок находится в интервале от 7 до 10.

Общая тенденция по результатам опроса показывает, что качество генерируемых изображений не было однородным и имело достаточно большой разброс. Несмотря на разброс, были получены высокие оценки для некоторых изображений, что может говорить об эффективности генеративных моделей при использовании определенных параметров (clip_guidance_scale). В целом результаты опроса могут быть использованы для дальнейшего улучшения качества генерируемых изображений. Общим выводом можно считать, что качество генерируемых изображений может сильно варьироваться в зависимости от параметров генерации и контекста. В дальнейшем может быть полезным провести более широкий и разнообразный опрос, чтобы получить более надежные результаты.

Для наглядности приведем график по средним значениям оценок опрошенных по сгенерированным изображениям (Рисунок 6) и рейтинг изображений по среднему значению оценки (Рисунок 7). Синим цветом обозначены результаты опроса для изображений из первого эксперимента, красным – второго. “Параметр 1000” – это 1-й эксперимент, “параметр 1500” – 2-й эксперимент.

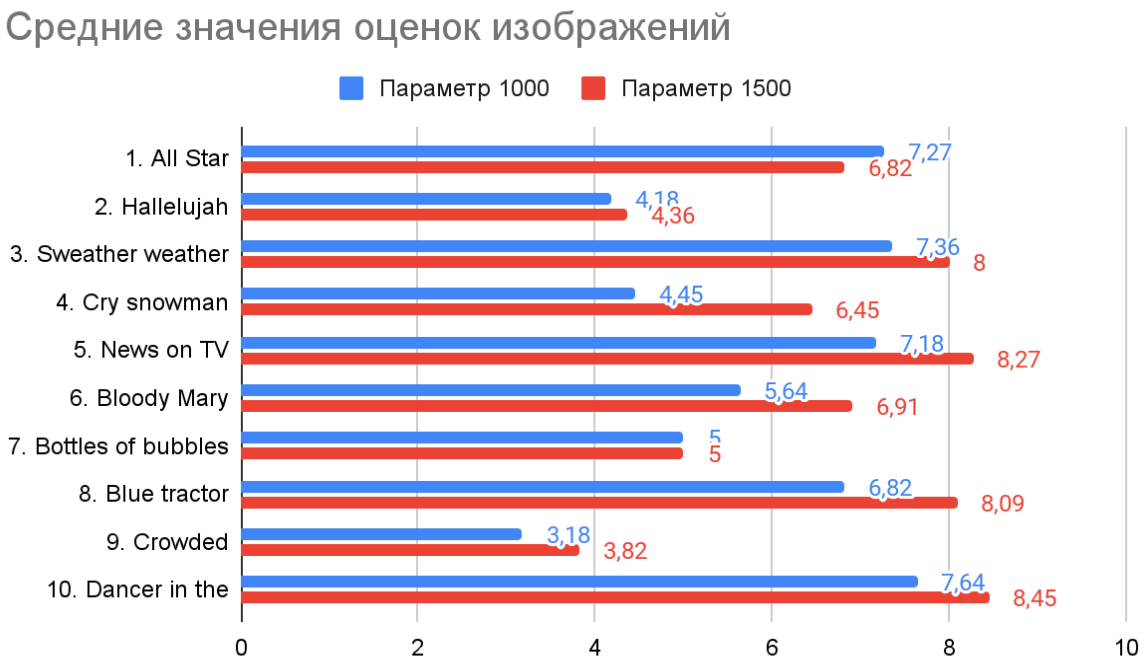

Рис. 6. Среднее значение оценок изображений

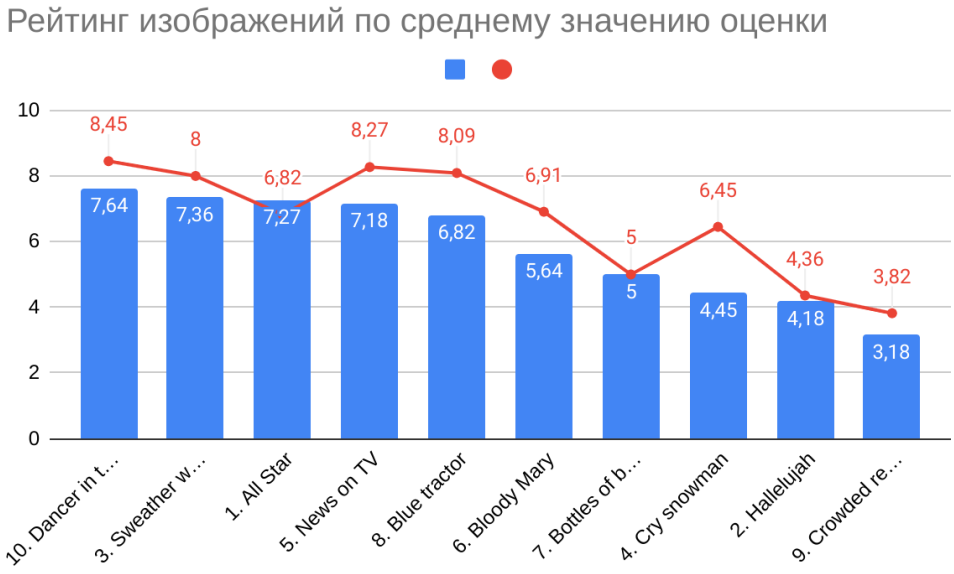

Рис. 7. Сравнение изображений по среднему значению оценки. Описание: Гистограмма – эксперимент №1, линия – эксперимент №2

На Рисунке 7 видно, что в обоих экспериментах лучше всего оценили изображения, сгенерированные по тексту “Dancer in the dark” и “Sweather Weather”, а хуже всего – по текстам “Hallelujah” и “Crowded restaurants”. Заметим также, что в

результате второго эксперимента изображения получили лучшую оценку, значит, производительность нейросети зависит от текстового запроса и отдельных параметров.

ЗАКЛЮЧЕНИЕ

Результаты исследования подтверждают потенциал объединения алгоритмов ImageNet и CLIP для создания высококачественных и контекстуально релевантных изображений из текстовых подсказок.

Проведено сравнение точности моделей ImageNet и CLIP при генерации изображений из текстовых подсказок. Для генерации изображений с использованием данного сочетания алгоритмов лучшими являются параметры clip_guidance_scale = 1500 и range_scale = 150. Установлено также, что качество результатов генерации изображений нейронной сетью в значительной степени зависит от характера вводимого текста и используемых параметров. Хотя каждый из названных алгоритмов имеет свои сильные и слабые стороны, их комбинация может привести к лучшим результатам. В то время как ImageNet проявляет высокую производительность в области генерации визуально эстетических изображений, CLIP демонстрирует более эффективную работу в задаче сопоставления текстовых подсказок с соответствующими изображениями. По оценке пользователей, интеграция этих моделей нейронный сетей потенциально позволяет создавать высококачественные и контекстуально релевантные изображения из текста. Эти результаты открывают возможности для использования изображений, генерируемых искусственным интеллектом, в различных отраслях промышленности.

Будущие исследования могли бы быть направлены на изучение возможности использования для генерации изображений гибридных моделей, сочетающих в себе сильные стороны имеющихся алгоритмов, а также разработку новых алгоритмов, способных преодолеть ограничения, имеющиеся у каждого из них.

Благодарности

Авторы выражают благодарность Максиму Олеговичу Таланову за ценные советы при проведении исследования и рекомендации по оформлению статьи.

СПИСОК ЛИТЕРАТУРЫ

1. *Elasri M., Elharrouss O., Al-Maadeed S., Tairi H.* Image Generation: A Review // Neural Processing Letters. 2022. Vol. 54. No. 5. P. 4609–4646.

2. *Zhang H., Song H., Li S., Zhou M., Song D.* A survey of controllable text generation using transformer-based pre-trained language models // arXiv preprint arXiv:2201.05337. 2022

3. Основы генеративно-состязательных сетей. URL: https://habr.com/ru/articles/726254/

4. *Brown T., Mann B., Ryder N., Subbiah M., Kaplan J. D., Dhariwal P., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell. A, Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D.M., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I., Amodei D.* Language models are few-shot learners // Advances in neural information processing systems. 2020. Vol. 33. P. 1877–1901.

5. DALL· E 2. URL:https://openai.com/product/dall-e-2.

6. How AI is Transforming Text-to-Image Generation. URL: https://nesesho.com/index.php/2023/04/12/how-ai-is-transforming-text-to- image-generation/

7. OpenAI· GitHub. URL: https://github.com/openai.

8. *Gulrajani I., Ahmed F., Arjovsky M., Dumoulin V., Courville A.C.* Improved training of wasserstein GANs // Advances in neural information processing systems. 2017. Vol. 30. P. 5767–5777.

9. *Indolia S., Goswami A.K., Mishra S.P., Asopa P.* Conceptual understanding of convolutional neural network-a deep learning approach // Procedia computer science. 2018. Vol. 132. P. 679–688.

10. *Laudani A., Lozito G.M., Fulginei F.R., Salvini A.* On training efficiency and computational costs of a feed forward neural network: a review // Computational intelligence and neuroscience. 2015. P. 83–83.

11. CLIP. URL: https://github.com/openai/CLIP.

12. *Dhariwal P., Nichol A.* Diffusion models beat gans on image synthesis // Advances in Neural Information Processing Systems. 2021. Vol. 34. P. 8780–8794.

13. *Kim G., Kwon T., Ye J.C.* Diffusionclip: Text-guided diffusion models for robust image manipulation // In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022. P. 2426–2435.

NEURAL NETWORK FOR GENERATING IMAGES BASED ON SONG LYRICS USING OPENAI AND CLIP MODELS

A. R. Davletgareeva[1] [0009-0008-7258-470X], K. A. Edkova[2] [0009-0005-4706-2254]

[1, 2]*Institute of Information Technologies and Intelligent Systems, Kazan (Volga Region) Federal University, ul. Kremlyovskaya, 35, Kazan, 420008*

[1]alsudavletgareeva@gmail.com, [2]ksushka.e21@gmail.com

Abstract

The effectiveness of the ImageNet diffusion model and CLIP models for image generation based on textual descriptions was investigated. Two experiments were conducted using various textual inputs and different parameters to determine the optimal settings for generating images from text descriptions. The results showed that while ImageNet performed well in generating images, CLIP demonstrated better alignment between textual prompts and relevant images. The obtained results highlight the high potential of combining these mentioned models for creating high-quality and contextually relevant images based on textual descriptions.

Keywords: *image generation, artificial intelligence, ImageNet diffusion model, CLIP, deep learning, neural networks, natural language processing.*

REFERENCES

1. *Elasri M., Elharrouss O., Al-Maadeed S., Tairi H.* Image Generation: A Review // Neural Processing Letters. 2022. Vol. 54. No. 5. P. 4609–4646.

2. *Zhang H., Song H., Li S., Zhou M., Song D.* A survey of controllable text generation using transformer-based pre-trained language models // arXiv preprint arXiv:2201.05337. 2022

3. Fundamentals of generative-consistent networks.

URL: https://habr.com/ru/articles/726254/

4. *Brown T., Mann B., Ryder N., Subbiah M., Kaplan J. D., Dhariwal P., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell. A, Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D.M., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I., Amodei D.* Language models are few-shot learners // Advances in neural information processing systems. 2020. Vol. 33. P. 1877–1901.

5. DALL· E 2. URL:https://openai.com/product/dall-e-2.

6. How AI is Transforming Text-to-Image Generation.
URL: https://nesesho.com/index.php/2023/04/12/how-ai-is-transforming-text-to-image-generation/

7. OpenAI· GitHub. URL: https://github.com/openai.

8. Gulrajani I., Ahmed F., Arjovsky M., Dumoulin V., Courville A.C. Improved training of wasserstein GANs // Advances in neural information processing systems. 2017. Vol. 30. P. 5767–5777.

9. *Indolia S., Goswami A.K., Mishra S.P., Asopa P.* Conceptual understanding of convolutional neural network-a deep learning approach // Procedia computer science. 2018. Vol. 132. P. 679–688.

10. *Laudani A., Lozito G.M., Fulginei F.R., Salvini A.* On training efficiency and computational costs of a feed forward neural network: a review // Computational intelligence and neuroscience. 2015. P. 83–83.

11. CLIP. URL: https://github.com/openai/CLIP.

12. *Dhariwal P., Nichol A.* Diffusion models beat gans on image synthesis // Advances in Neural Information Processing Systems. 2021. Vol. 34. P. 8780–8794.

13. *Kim G., Kwon T., Ye J.C.* Diffusionclip: Text-guided diffusion models for robust image manipulation // In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022. P. 2426–2435.

СВЕДЕНИЯ ОБ АВТОРАХ

ДАВЛЕТГАРЕЕВА Алсу Ришатовна – студентка кафедры программной инженерии Института ИТИС КФУ. Сфера научных интересов – искусственный интеллект.

Alsu Rishatovna DAVLETGAREEVA – a student of the Department of Software Engineering, Institute ITIS, Kazan Federal University. Her research interests lie in the field of artificial intelligence.

email: alsudavletgareeva@gmail.com

ORCID: 0009-0008-7258-470X

ЕДКОВА Ксения Александровна – студентка магистратуры Института информационных технологий и интеллектуальных систем. Сфера научных интересов – искусственный интеллект.

Ksenia Aleksandrovna EDKOVA – a student of the Department of Software Engineering, Institute ITIS, Kazan Federal University. Her research interests lie in the field of artificial intelligence.

email: ksushka.e21@gmail.com

ORCID: 0009-0005-4706-2254

Материал поступил в редакцию 2 июня 2023 года