**Использование генеративно-состязательных сетей для синтеза изображений по текстовым описаниям**

***Разуваев Александр Вячеславович***

*Поволжский государственный университет телекоммуникаций и информатики, Самара*

*Студент кафедры управления в технических система*

*E-mail: razuvaev.sasha2015@yandex.ru*

***Осанов Владимир Андреевич***

*Поволжский государственный университет телекоммуникаций и информатики, Самара*

*Старший преподаватель кафедры* *управления в технических система*

*E-mail: v.osanov@psuti.ru*

**USING GENERATIVE-ADVERSARIAL NETWORKS TO SYNTHESIZE IMAGES FROM TEXT DESCRIPTIONS**

***Razuvaev Alexander Vyacheslavovich***

*Povolshskiy State University of Telecommunications and Informatics, Samara*

*Student of the Department of Management in Technical Systems*

*E-mail: razuvaev.sasha2015@yandex.ru*

***Osanov Vladimir Andreevich***

*Povolshskiy State University of Telecommunications and Informatics, Samara  
Senior teacher at the Department of Management in Technical Systems*

*E-mail: v.osanov@psuti.ru*

**АННОТАЦИЯ**

В данной статье рассматривается использование модели Stable Diffusion для генерации изображений на основе текстовых описаний. Проведены эксперименты, подтверждающие эффективность Stable Diffusion в создании высококачественных изображений с высокой степенью детализации и стилистической адаптивности. Особое внимание уделено методам оптимизации модели, предварительной обработке данных и метрикам оценки соответствия между текстовыми запросами и визуальными результатами. Результаты исследований показывают факт использования Stable Diffusion в качестве мощного инструмента для автоматической генерации графического контента и может быть успешно применена в областях цифрового искусства компьютерного зрения.

**ABSTRACT**

This article discusses the use of the Stable Diffusion model to generate images based on text descriptions. Experiments have been conducted confirming the effectiveness of Stable Diffusion in creating high-quality images with a high degree of detail and stylistic adaptability. Special attention is paid to model optimization methods, data preprocessing, and metrics for evaluating the correspondence between text queries and visual results. The research results show that Stable Diffusion is a powerful tool for automatic generation of graphic content and can be successfully applied in the fields of digital art, design and computer vision.

**Ключевые слова:**Stable Diffusion; генеративные модели; нейронные сети; диффузионные процессы; латентные представления; оценка качества изображений.

**Keywords:** Stable Diffusion; generative models; neural networks; diffusion processes; latent representations; image quality assessment.

**ВВЕДЕНИЕ**

В последние годы генеративные модели значительно продвинулись в создании реалистичного визуального контента. Одной из наиболее мощных и гибких моделей является Stable Diffusion версия – диффузионная модель, использующая вероятностные методы и глубокие нейросетевые архитектуры для генерации изображений на основе текстовых описаний (англ. prompts) [1]. Эта модель стала ключевым инструментом в области генеративных нейросетей благодаря своей способности создавать детализированные и стилистически разнообразные изображения, о чём свидетельствует работа [2].

Несмотря на высокое качество синтезируемых изображений, одной из ключевых проблем остается оценка соответствия сгенерированного контента исходному текстовому запросу. Для решения этой задачи используются различные метрики качества, включая CLIP Metric (англ. Contrastive Language-Image Pre-training – рус. предварительная подготовка к контрастному языку и изображению), которая позволяет количественно оценить степень соответствия изображения заданному тексту. Этот подход применяется в исследовательской практике, что подтверждается результатами работы [1], где акцентируется внимание на важности метрики CLIP для оценки генерации изображений. Однако качество генерации также зависит от используемого метода семплирования (англ. scheduler), который определяет процесс восстановления изображения из шума. Различные шедулеры, такие как Euler Ancestral Discrete Scheduler, Euler Discrete Scheduler и DDIM Scheduler, обладают уникальными характеристиками и могут по-разному влиять на итоговое изображение, согласно статье [3].

В данной работе проводится анализ качества генерации изображений в Stable Diffusion при использовании трех различных механизмов семплирования. Оценка качества выполняется с помощью CLIP Metric, что позволяет определить, какой из методов обеспечивает наилучшее соответствие между изображением и текстом. Проведенные эксперименты позволят выявить особенности работы различных шедулеров и предложить рекомендации по их выбору в зависимости от задачи. Это исследование, как и другие работы в области генеративных моделей, подчеркивает важность правильного выбора шедулера для достижения оптимального качества изображений [4].

Актуальность исследования обусловлена растущим интересом к генеративным моделям и их применению в цифровом искусстве, дизайне и компьютерном зрении. Результаты работы могут быть полезны как исследователям в области глубинного обучения, так и практическим разработчикам, использующим Stable Diffusion для создания контента. Эти выводы об актуальности данной задачи также подтверждаются в статье [5], где подчеркивается важность применения таких моделей для создания визуальных изображений в различных областях.

**ПРОЦЕСС ГЕНЕРАЦИИ ИЗОБРАЖЕНИЙ**

В данном исследовании рассматривается использование различных schedulers (рус. планировщиков шагов диффузии) на базе Stable Diffusion версия для генерации изображений на основе текстовых описаний (англ. prompt). Основной задачей является анализ точности генерации изображений при применении различных планировщиков, с использованием метрики CLIP для оценки сходства между сгенерированными изображениями и их текстовыми описаниями.

Stable Diffusion основана на диффузионных моделях, которые постепенно восстанавливают изображение из случайного шума, используя обученный процесс денойзинга. В основе архитектуры лежит латентное диффузионное моделирование (англ. Latent Diffusion Model, LDM), которое позволяет выполнять генерацию изображений в сжатом пространстве признаков вместо пиксельного пространства. Это снижает вычислительные затраты и улучшает качество генерации. Модель использует U-Net в качестве основного блока для восстановления изображений и вариационный автоэнкодер (англ. Variational Autoencoder, VAE) для перехода между латентным и пиксельным пространством.

Латентные представления в Stable Diffusion играют ключевую роль, вместо работы с изображением в его изначальном разрешении, модель использует компактное представление в скрытом пространстве, что ускоряет генерацию и делает её более эффективной. Процесс начинается с кодирования изображения с помощью VAE-энкодера, после чего применяется диффузионная модель, которая последовательно корректирует шумовые компоненты в латентном пространстве. После завершения этого процесса VAE-декодер преобразует обработанное латентное представление обратно в изображение высокого качества.

В процессе генерации изображения диффузионные модели начинают работу со случайного шума и постепенно преобразуют его в изображение, соответствующее текстовому описанию. Качество генерации, а также точность соответствия изображения описанию, сильно зависит от выбранного scheduler.

**ПЛАН ИССЛЕДОВАНИЯ**

Исследование проводилось на трех типах текстовых запросов (prompt):

1. Prompt №1 – " Футуристический город в облаках, где дома сделаны из кристаллов и светящихся растений, а небо раскрашено в оттенки розового и фиолетового. В центре города с небес низвергается огромный водопад, а по улицам прогуливаются крылатые существа, похожие на драконов. Вдалеке видны плавучие острова с густыми лесами и водопадами".
2. Prompt №2 – "Подводный мир с биолюминесцентными существами, где кораллы излучают мягкое свечение, а рыбы плавают в воде, как птицы в небе. В глубинах океана можно найти древние руины, покрытые водорослями и населенные гигантскими кальмарами. В сердце этого мира находится подводный вулкан, извергающий пузыри и светящиеся частицы".
3. Prompt №3 – "Заброшенный парк развлечений, где природа взяла верх, деревья обвивают карусели, а на американских горках растут цветы. В парке есть старые аттракционы, покрытые мхом и виноградными лозами, а также заброшенные киоски с выцветшими вывесками. В центре парка есть большое озеро, покрытое лилиями, в котором плавают лебеди и утки".

В качестве планировщиков шагов диффузии были выбраны следующие модели:

1. EulerAncestralDiscreteScheduler. Этот планировщик основывается на адаптивном подходе, где каждый шаг диффузии зависит от предыдущего состояния. Он сохраняет высокую степень случайности на каждом шаге, что может приводить к большим вариациям в генерации изображений, но также повышает гибкость в получении более разнообразных результатов.

2. EulerDiscreteScheduler. В отличие от предыдущего, этот планировщик обеспечивает более стабильную и детерминированную генерацию. Он действует с фиксированными шагами, что позволяет контролировать процесс диффузии, минимизируя вариативность и повышая точность, но может снижать разнообразие результатов.

3. DDIMScheduler. Этот планировщик использует другой метод для управления шагами диффузии, что позволяет ускорить процесс генерации без значительной потери качества. Он ориентирован на уменьшение числа шагов, что делает генерацию более быстрой, но требует тщательной настройки для обеспечения оптимального баланса между точностью и скоростью.

Оценка сходства между изображением и текстом выполнялась по метрике CLIP. В процессе работы CLIP изображение и текст преобразуются в векторные представления (англ. embedding), и сходство между ними измеряется с использованием косинусного расстояния или косинусной близости. Сходство варьируется от -1 до 1, где значение 1 означает полное совпадение изображения и текста, а значение -1 – полное несоответствие.

Формула для вычисления сходства CLIP между изображением и текстом выглядит следующим образом:

где ​ – векторное представление изображения, полученное с помощью модели CLIP;

​ – векторное представление текста;

– нормы (длины) векторов изображения и текста соответственно.

**АНАЛИЗ РЕЗУЛЬТАТОВ**

Для prompt№1 были получены результаты, представленные на рис. 1. Оценка этих результатов по метрике CLIP отображена на рис. 2.

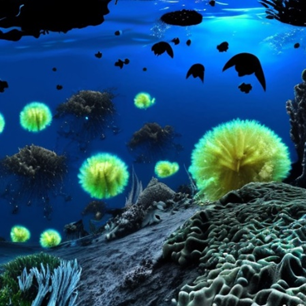
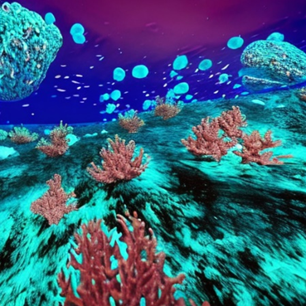
а б в

***Рисунок 1. Сгенерированные изображения для Prompt №1 с применение планировщика: а – Euler Ancestral Discrete Scheduler, б – Euler Discrete Scheduler, в – DDIM Scheduler***

***Рисунок 2. Метрика CLIP для prompt№1***

Из рис.2 видно, что DDIMS cheduler показал наибольшее значение метрики CLIP (0.4022), что свидетельствует о лучшем соответствии сгенерированного изображения текстовому описанию по сравнению с другими schedulers. Однако несмотря на то, что значения Euler Ancestral DiscreteScheduler и Euler Discrete Scheduler близки (0.3868 и 0.3924 соответственно), разница в метрике CLIP между этими schedulers говорит о том, что выбор конкретного алгоритма влияет на точность генерации изображений в зависимости от выбранного scheduler.

Результаты генерации для prompt№2 изображены на рис. 3, а результат их анализа в соответствие с метрикой CLIP на рис. 4.

а б в

***Рисунок 3. Сгенерированные изображения для Prompt №2 с применение планировщика: а – Euler Ancestral Discrete Scheduler, б – Euler Discrete Scheduler, в – DDIM Scheduler***

Из анализа (рис. 4) полученных изображений следует, что Euler Discrete Scheduler показал наибольшее значение метрики CLIP (0.3457), что указывает на лучшее соответствие сгенерированного изображения текстовому описанию по сравнению с другими планировщиками.

***Рисунок 4. Метрика CLIP для prompt№2***

Euler Ancestral Discrete Scheduler имеет немного меньшее значение (0.329), а DDIM Scheduler – наименьшее (0.2976). Эти результаты подтверждают, что выбор scheduler влияет на точность и качество генерации изображений для второго prompt, подчеркивая важность выбора оптимального scheduler для получения наиболее точных и релевантных изображений в зависимости от текстового описания.

Генерация по prompt№3 продемонстрированы на рис. 5.

а б в

***Рисунок 5. Сгенерированные изображения для Prompt №3 с применение планировщика: а – Euler Ancestral Discrete Scheduler, б – Euler Discrete Scheduler, в – DDIM Scheduler***

Анализ данной генерации (рис. 6) показал, что Euler Discrete Scheduler достиг наибольшего значения метрики CLIP (0.3622), что подтверждает лучшем соответствии сгенерированного изображения текстовому описанию для третьего prompt. Euler Ancestral Discrete Scheduler показал значение 0.3585, немного уступая первому. DDIM Scheduler имеет наименьшее значение (0.335), что указывает на менее точное соответствие изображения запросу. Эти результаты подчеркивают, что для третьего prompt Euler Discrete Scheduler обеспечивает наилучшее качество генерации изображений, и выбор scheduler оказывает значительное влияние на точность и детализацию получаемых изображений.

***Рисунок 6. Метрика CLIP для prompt№3***

**ЗАКЛЮЧЕНИЕ**

В ходе проведенного исследования были изучены и протестированы различные schedulers, используемые для генерации изображений с помощью модели Stable Diffusion. Основной целью работы было сравнительное анализирование точности генерации изображений по текстовому запросу, используя метрику CLIP, с применением трех различных scheduler'ов: Euler Ancestral Discrete Scheduler, Euler Discrete Scheduler и DDIM Scheduler.

Эксперименты показали, что Euler Discrete Scheduler показал наибольшие значения метрики CLIP во всех трех тестах, что свидетельствует о лучшем соответствии сгенерированных изображений текстовым описаниям. В то время как Euler Ancestral Discrete Scheduler продемонстрировал схожие, но немного уступающие результаты, DDIM Scheduler consistently показал наименьшие значения CLIP, указывая на менее точную генерацию изображений.

Таким образом, исследование подтвердило, что выбор scheduler оказывает значительное влияние на точность и качество генерации изображений. Это подчеркивает важность оптимизации параметров генерации для достижения наилучших результатов в задачах визуализации, что делает нейронные сети мощным инструментом для работы с изображениями, генерируемыми по текстовым запросам.

**Список литературы**

1. Гончарова Л. В., Шевченко Д. Ю. Генеративные состязательные сети в искусственном интеллекте: обзор и анализ текущих подходов // Научные технологии. 2023. № 8. С. 30-40.

2. Ли Ч., Ван Д. Современные методы генерации изображений с использованием GAN // Вестник машинного обучения. 2021. № 6. С. 5-12.

3. Гладков А. В., Макаров В. А. Глубокие нейронные сети и их применение в генеративных моделях. М. : Питер, 2020. 240 с.

4. Генеративные состязательные сети в обработке изображений [Электронный ресурс]. – Режим доступа: https://technolab.ru/gan-methods, свободный. – (дата обращения: 10.03.2025).

5. Алексеева С. А., Иванова Т. Ю. Оценка эффективности GAN при синтезе изображений в условиях ограниченных данных // Журнал искусственного интеллекта. 2021. № 3. С. 40-50.