ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Компьютерные науки и анализ данных»

Отчет о проекте на тему:

Исследование влияния текствых эмбеддингов с разных слоев на качество генерации диффузионных моделей в задаче Text-To-Image

Выполнил студент:

группы #БКНАД222 Голубкова Анастасия Ярославовна

Руководитель проекта:

Алиев Мишан

Москва 2025

Содержание

Аннотация

Диффузионные модели в последние годы стали ключевым подходом в задачах Text-To-Image генерации, продемонстрировав существенный прогресс в качестве синтезируемых изображений, семантическом соответствии текстовому описанию и устойчивости к артефактам. В отличие от предыдущих генеративных методов, таких как GAN, диффузионные модели обеспечивают более стабильное обучение и лучшее покрытие распределения данных, то есть избежания коллапса мод, что сделало их основой современных систем генерации изображений.

Однако высокая вычислительная сложность диффузионного процесса, связанная с необходимостью выполнения большого числа последовательных шагов денойзинга, существенно ограничивает практическое применение таких моделей, особенно в сценариях реального времени и на ресурсно-ограниченных устройствах. Это обуславливает актуальность исследований, направленных на ускорение инференса диффузионных моделей при сохранении визуального качества и семантической точности генерируемых изображений.

Современные подходы к ускорению включают сокращение числа шагов диффузии, использование аппроксимаций и дистилляции моделей, оптимизацию архитектур, а также применение специализированных методов семплирования. Задача нахождения баланса между скоростью генерации и качеством результата является критически важной для дальнейшего развития и широкого внедрения Text-To-Image систем в прикладных областях.

1 Цель и задачи

Целью проекта является исследование и разработка метода взвешенного объединения текстовых эмбеддингов, полученных с различных слоёв текстового энкодера, с целью улучшения качества представления текстового запроса по сравнению с базовыми подходами усреднения и нормирования эмбеддингов.

Для достижения поставленной цели в рамках проекта были сформулированы и решены следующие задачи:

Изучить существующие подходы к формированию текстовых эмбеддингов в задачах Text-To-Image генерации и проанализировать влияние различных слоёв текстового энкодера на семантическое представление текста.

Реализовать базовый пробный сетап, основанный на обычном усреднении эмбеддингов с разных слоёв текстового энкодера, а также на их нормировании перед усреднением, и оценить его качество.

Разработать метод обучения коэффициентов для взвешенного объединения текстовых эмбеддингов, извлекаемых с различных слоёв текстового энкодера.

Реализовать процедуру обучения коэффициентов и интегрировать её в существующий пайплайн генерации изображений.

Провести экспериментальное сравнение предложенного метода с базовыми подходами (усреднение и нормированное усреднение) по метрикам.

Проанализировать влияние обучаемых коэффициентов на вклад отдельных слоёв текстового энкодера в итоговое эмбеддинговое представление.

Сформулировать выводы о целесообразности использования обучаемого взвешивания эмбеддингов и определить направления дальнейших исследований.

2 Обзор литературы

Denoising Diffusion Implicit Models (DDIM).

В работе предложен метод неявного диффузионного семплирования, позволяющий существенно сократить число шагов генерации по сравнению с классическими DDPM без дополнительного обучения модели. DDIM стал одним из базовых подходов к ускорению инференса диффузионных моделей при сохранении качества, что сделало его широко применяемым в современных Text-To-Image системах.

Improved Denoising Diffusion Probabilistic Models.

Статья посвящена улучшению качества диффузионных моделей за счёт модификаций функции потерь, параметризации шума и расписаний диффузии. Работа оказала значительное влияние на развитие высококачественной генерации изображений и стала основой для последующих архитектур и методов ускорения.

Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding (DeepFloyd IF).

В данной работе представлена многостадийная Text-To-Image система DeepFloyd IF, сочетающая мощный текстовый энкодер и каскад диффузионных моделей для генерации изображений высокого разрешения. Особое внимание уделено роли текстовых эмбеддингов и их влиянию на визуальное качество и семантическую точность генерации.

A Comprehensive Study of Decoder-Only Transformers for Text-to-Image Generation

В статье проведено комплексное исследование decoder-only трансформеров в задачах Text-To-Image генерации. Рассматривается влияние архитектурных решений, представления текста и способов взаимодействия текстовых и визуальных признаков. Работа представляет интерес в контексте альтернатив диффузионным моделям и анализа роли текстового энкодинга.

3 Описание работы

В рамках данной работы было проведено систематическое исследование влияния различных способов агрегации текстовых эмбеддингов, извлекаемых с разных слоёв текстового энкодера, на качество text-to-image генерации в модели DeepFloyd IF. Основной мотивацией исследования являлась гипотеза о том, что информация, содержащаяся в промежуточных слоях текстового энкодера, может дополнять представление последнего слоя и потенциально улучшать семантическое соответствие и визуальное качество генерируемых изображений.

На первом этапе был реализован базовый экспериментальный сетап, в котором использовалось простое усреднение текстовых эмбеддингов, полученных со всех слоёв текстового энкодера. Дополнительно был рассмотрен вариант с предварительным нормированием эмбеддингов перед их усреднением с целью выравнивания масштабов представлений различных слоёв. Для оценки эффективности данных подходов было проведено сравнение с контрольным вариантом, в котором в качестве условного представления использовался только эмбеддинг последнего слоя текстового энкодера. Качество генерации оценивалось с использованием метрики HPSv2. Экспериментальные результаты показали, что оба варианта усреднения (с нормированием и без) уступают по качеству генерации варианту, использующему исключительно эмбеддинг последнего слоя, который продемонстрировал наилучшие значения HPSv2 среди базовых подходов.

На следующем этапе исследования был предложен метод взвешенного объединения текстовых эмбеддингов с различных слоёв текстового энкодера. В рамках данного подхода для каждого слоя обучались отдельные коэффициенты, определяющие его вклад в итоговое агрегированное представление текста. Обучение коэффициентов осуществлялось совместно с процессом генерации изображений при фиксированных весах основной диффузионной модели. Экспериментальные результаты показали, что обучаемое взвешивание позволяет превзойти по качеству простое усреднение эмбеддингов, что подтверждает наличие полезной информации в промежуточных слоях текстового энкодера. Тем не менее, полученное качество всё ещё оставалось ниже по сравнению с вариантом использования эмбеддинга последнего слоя.

В дальнейшем метод был расширен за счёт увеличения гибкости параметризации. Вместо использования одного набора коэффициентов для всех шагов диффузионного процесса были введены отдельные коэффициенты для каждого шага расшумления. Таким образом, вклад текстовых эмбеддингов с различных слоёв текстового энкодера стал зависеть от текущего шага денойзинга. При использовании 25 шагов диффузионного семплирования и 24 слоёв текстового энкодера общее число обучаемых параметров составило 600. Данный подход позволил существенно увеличить выразительную способность модели за счёт более тонкой адаптации текстового представления к динамике диффузионного процесса.

Для реализации обучения коэффициентов потребовалась модификация стандартного пайплайна денойзинга DeepFloyd IF. В оригинальной реализации генерация изображений выполняется в режиме torch.no_grad, что делает невозможным распространение градиентов и обучение дополнительных параметров. В связи с этим был реализован кастомный пайплайн денойзинга, обеспечивающий корректное прохождение градиентов через текстовые

эмбеддинги и обучаемые коэффициенты. Это позволило интегрировать предложенные методы агрегации в процесс оптимизации без изменения весов основной диффузионной модели.

Все проведённые эксперименты сопровождались детальным логгированием параметров запуска, значений метрик качества, а также промежуточных и финальных результатов генерации изображений. Такой подход обеспечил воспроизводимость экспериментов и позволил провести последовательный анализ влияния различных вариантов агрегации текстовых эмбеддингов на качество text-to-image генерации.

3.1 Схема текстового энкодера

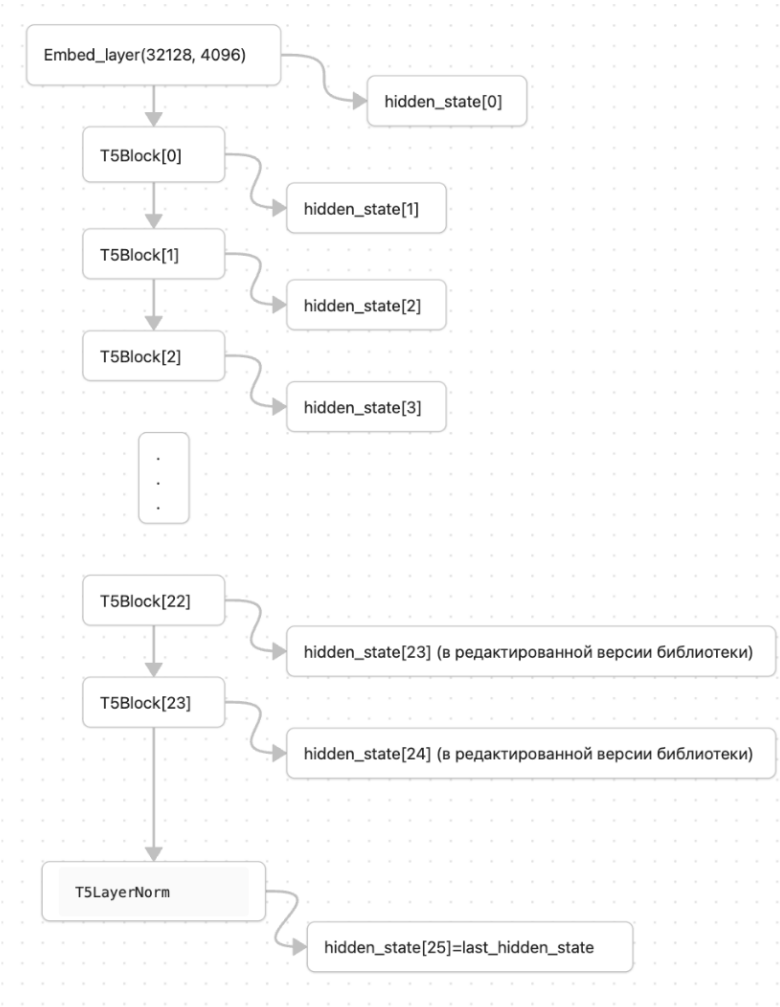

Рисунок 3.1

3.2 Детали реализации

Для обучения и оценки предложенных методов агрегации текстовых эмбеддингов был сформирован специализированный датасет, ориентированный на задачу text-to-image гене-

mode\model	stage1 M	stage1 L	stage1 XL
mode=1	22.65	23.28	23.29
mode=2	20.55	21.22	22.07
mode=3	19.63	20.63	22.13
mode=4	20.00	20.84	22.29

Рисунок 3.2: Стартовый сетап. Mode=1: Берем только эмбеддинг с последнего слоя. Mode=2: применяем final layer norm к обычным усредненным эмбеддингам. Mode=3: применяем последний блок нормализации к каждому эмбеддингу отдельно и усредняем. Mode=4: применяем L2 нормализацию к каждому эмбеддингу отдельно, усредняем и применяем последний блок нормализации

рации. На первом этапе был составлен набор из 400 текстовых промптов, охватывающих различные визуальные сцены, объекты и стилистические описания. Для каждого промпта с использованием модели DeepFloyd XL (с 100 шагами расшумления) были сгенерированы соответствующие изображения, которые далее использовались в качестве референсных визуальных представлений.

Для каждого текстового промпта были дополнительно извлечены текстовые эмбеддинги с 24 слоёв текстового энкодера. Таким образом, для каждого промпта формировался набор из 24 эмбеддингов, соответствующих различным уровням абстракции текстового представления. Полученные эмбеддинги сохранялись и использовались в дальнейшем обучении без повторного пересчёта, что обеспечивало стабильность входных данных и снижало вычислительные затраты.

На основе полученных данных был сформирован итоговый датасет, в котором каждый объект включал следующие компоненты: исходный текстовый промпт, изображение, сгенерированное моделью Stable Diffusion XL по данному промпту, а также набор из 24 текстовых эмбеддингов, извлечённых с различных слоёв текстового энкодера. В качестве негативного промпта использовалась пустая строка, одинаковая для всех примеров датасета, что позволило исключить влияние негативного контекста на процесс обучения и сосредоточиться на анализе агрегации позитивных текстовых эмбеддингов.

3.3 Первый сетап

В первом экспериментальном сетапе был реализован класс модели, содержащий 24 обучаемых параметра, соответствующих коэффициентам взвешивания эмбеддингов с каждого слоя текстового энкодера. Была разработана функция агрегации, осуществляющая взвешенное объединение эмбеддингов на основе данных коэффициентов с последующим форми-

рованием итогового текстового представления, используемого в процессе генерации изображений. Для обеспечения возможности обучения данных параметров был реализован кастомный пайплайн денойзинга, в котором были отключены ограничения стандартной реализации DeepFloyd IF, связанные с использованием режима torch.no_grad. Это обеспечило корректное распространение градиентов через процедуру агрегации и позволило обучать коэффициенты без модификации весов основной диффузионной модели. В модели процесс расшумления составлял 25 шагов.

Процесс обучения осуществлялся на одном графическом ускорителе NVIDIA A100 и продолжался в течение 24 часов. В ходе обучения оптимизировались исключительно параметры агрегации текстовых эмбеддингов, в то время как параметры диффузионной модели оставались зафиксированными. Оценка качества проводилась на отдельной валидационной выборке с использованием метрики HPSv2, что позволило отслеживать влияние обучения коэффициентов на качество text-to-image генерации и предотвращать переобучение. Так как весь процесс диффузии требовал наличие активаций на всех слоях, максимальный размер батча на A100 составил 1, поэтому для улучшения сходимости и уменьшения шума в процессе оптимизации использовалась техника аккумуляции градиентов (составила 16 шагов).

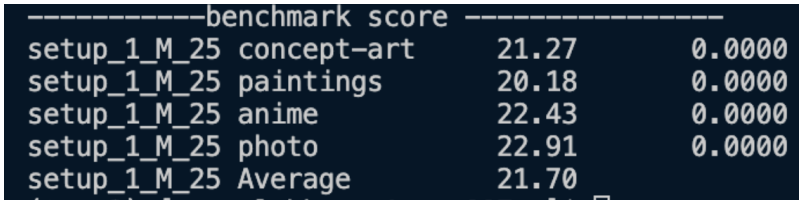

Рисунок 3.3: Модель M на 25 шагов с эмбеддингом с последнего слоя

Рисунок 3.4: Модель-учитель XL на 100 шагов с эмбеддингом с последнего слоя

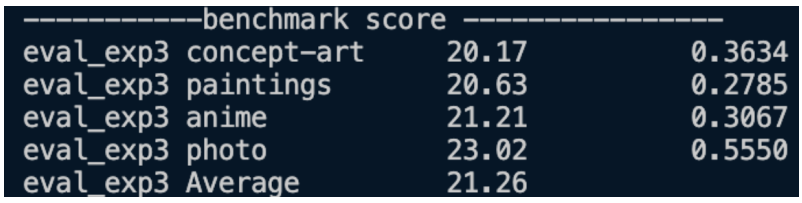

Рисунок 3.5: Модель M на 25 шагов с обученными 24 коэффициентами

3.4 Второй сетап

В рамках второго экспериментального сетапа был предложен и реализован более гибкий подход к агрегации текстовых эмбеддингов, основанный на обучении коэффициентов, зависящих от шага диффузионного процесса. В отличие от первого сетапа, в котором использовался единый набор коэффициентов для всех шагов денойзинга, в данном подходе каждому шагу расшумления соответствовал собственный набор параметров взвешивания эмбеддингов с различных слоёв текстового энкодера. Это позволило адаптировать вклад текстовой информации к динамике диффузионного процесса и потенциально более точно моделировать взаимодействие между текстовым условием и визуальным представлением на разных стадиях генерации изображения.

Формально, при использовании 25 шагов денойзинга и 24 слоёв текстового энкодера общее число обучаемых параметров составило 600. Для каждого шага диффузионного процесса обучался отдельный вектор коэффициентов, определяющий вклад соответствующих текстовых эмбеддингов в итоговое агрегированное представление. Таким образом, текстовое условие, подаваемое на вход диффузионной модели, становилось функцией не только слоевой структуры текстового энкодера, но и текущего шага денойзинга.

Для реализации данного подхода был расширен ранее разработанный кастомный пайплайн денойзинга. В частности, механизм агрегации текстовых эмбеддингов был интегрирован непосредственно в цикл диффузионного семплирования, что обеспечило пересчёт агрегированного текстового представления на каждом шаге денойзинга с использованием соответствующего набора обучаемых коэффициентов. Как и в первом сетапе, стандартные ограничения библиотеки DeepFloyd IF, связанные с использованием режима torch.no_grad, были устранены, что позволило корректно распространять градиенты через все обучаемые параметры.

Процедура обучения была организована аналогично первому сетапу: веса основной диффузионной модели оставались зафиксированными, а оптимизация выполнялась исключительно по обучаемым коэффициентам агрегации. Увеличение числа параметров привело к росту вычислительной сложности и требований к памяти, однако позволило существенно повысить выразительную способность модели. Обучение проводилось на графическом ускорителе NVIDIA A100, при этом мониторинг качества генерации осуществлялся на валидационной выборке с использованием метрики HPSv2.

Экспериментальные результаты показали, что введение зависимых от шага денойзинга коэффициентов позволяет дополнительно улучшить качество генерации по сравнению с

вариантом с единым набором коэффициентов. Тем не менее, даже при увеличении числа параметров данный подход не превзошёл по качеству вариант использования эмбеддинга последнего слоя текстового энкодера, что указывает на ограниченную полезность информации промежуточных слоёв в рамках рассматриваемой архитектуры и выбранной схемы обучения.

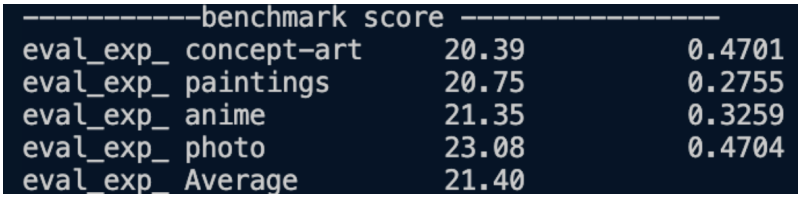

Рисунок 3.6: Модель M на 25 шагов с обученными 600 коэффициентами

4 Итоги и дальнейшее развитие

В рамках проведённого исследования были проанализированы различные способы агрегации текстовых эмбеддингов, получаемых с разных слоёв текстового энкодера, в задаче text-to-image генерации на базе модели DeepFloyd IF. Экспериментальные результаты показали, что использование обучаемых коэффициентов для взвешенного объединения эмбеддингов позволяет улучшить качество генерации по сравнению с простым усреднением эмбеддингов, а также по сравнению с усреднением с предварительным нормированием. Данный результат подтверждает гипотезу о наличии полезной информации в промежуточных слоях текстового энкодера и возможности её более эффективного использования за счёт обучаемых механизмов агрегации.

В то же время ни один из рассмотренных вариантов агрегации не превзошёл по качеству генерации базовый подход, использующий исключительно эмбеддинг последнего слоя текстового энкодера. Вероятной причиной данного эффекта является отсутствие универсального набора слоёв, оптимального для всех текстовых промптов. Различные промпты, вероятно, опираются на различные уровни абстракции текстового представления, и использование фиксированных коэффициентов, даже обучаемых, не позволяет в полной мере учитывать данную зависимость.

Дополнительным ограничением проведённого исследования является то, что в процессе обучения оптимизировались только коэффициенты агрегации текстовых эмбеддингов, в то время как параметры диффузионной модели оставались зафиксированными. В работе A Comprehensive Study of Decoder-Only Transformers for Text-to-Image Generation показано,

что совместное обучение архитектуры генерации (в частности, UNet) в течение длительного времени и на значительных вычислительных ресурсах (несколько GPU NVIDIA A100 в течение порядка семи дней) приводит к улучшению качества генерации даже при использовании простого усреднения текстовых эмбеддингов. Это указывает на то, что при наличии возможности обучения UNet предложенные методы агрегации могли бы привести к более существенному росту качества и, вероятно, превзойти вариант с использованием эмбеддинга последнего слоя. Однако в рамках данной работы подобные вычислительные ресурсы были недоступны.

В качестве направления дальнейших исследований представляется перспективным переход от статических коэффициентов агрегации к динамическим, зависящим от конкретного текстового запроса. В частности, возможным развитием является использование CLIP-текстового энкодера для получения представления промпта, которое затем подаётся на вход небольшой нейронной сети (например, MLP), генерирующей набор коэффициентов для взвешивания эмбеддингов с различных слоёв текстового энкодера. Такой подход позволил бы адаптировать агрегацию эмбеддингов под конкретный промпт и потенциально улучшить качество генерации без необходимости разморозки и обучения параметров UNet.

Список литературы

[1] Andrew Z. Wang, Songwei Ge, Tero Karras, Ming-Yu Liu, Yogesh Balaji *A Comprehensive Study of Decoder-Only LLMs for Text-to-Image Generation*. https://arxiv.org/pdf/2506.08210

[2] C. Saharia et al. *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding (DeepFloyd IF)*. https://arxiv.org/abs/2205.11487

[3] J. Song, C. Meng, S. Ermon. *Denoising Diffusion Implicit Models (DDIM)*. https://arxiv.org/abs/2010.02502

[4] A. Nichol, P. Dhariwal. *Improved Denoising Diffusion Probabilistic Models*. https://arxiv.org/abs/2102.09672