Создание персонализированных генераций изображений

A Preprint

Степанов Илья Дмитриевич
iliatut94@gmail.com

Abstract

В генеративных моделях существует широкий спектр проблем, однако одной из наиболее актуальных является сложность создания высококачественных изображений конкретных людей с точностью, передающей их уникальную идентичность. Предлагается сфокусировать внимание на разработке моделей, способных генерировать изображения заданного человека в разнообразных вариациях и с высоким разрешением.Требуется обучить методы IP-Adapter на модели Stable Diffusion с использованием множественных подсказок в виде картинок.

Keywords IP-Adapter(1) · Stable Diffusion (2)

1 Введение

В современных исследованиях активно развивается модель генерации изображений под названием Stable Diffusion. Основная идея заключается в том, что на каждой итерации шум добавляется к текущему изображению, а затем происходит процесс диффузии, который плавно размывает этот шум, превращая его в сглаженное изображение. Эта модель позволяет создавать высококачественные изображения на основе текстовых и графических подсказок, открывая новые возможности в области синтеза изображений. Однако, в процессе работы с моделями генерации изображений, в том числе и Stable Diffusion, возникают определенные проблемы, одна из них — недостаточное соответствие сгенерированных изображений исходным текстовым подсказкам.

Для решения указанной проблемы предложен метод IP-Adapter, который представляет собой простой способ адаптации изображений к текстовым подсказкам с использованием cross-attention. Cross-attention — это метод, применяемый в моделях генерации изображений для улучшения взаимосвязей между различными частями изображения. В данном случае он интегрируется в существующие диффузионные модели, что способствует повышению точности генерации изображений.

Моя мотивация заключается в повышении качества визуальных представлений за счет более точного соответствия текстовым и графическим подсказкам.

Целью моего исследования является разработка метода на основе IP-Adapter. Мотивация заключается в повышении качества визуальных представлений за счет более точного соответствия текстовым и графическим подсказкам.В модели принимается набор изображений и соответствующих текстовых подсказок.При обучении модели одно изображение удаляется равновероятно, и модель учится восстанавливать это удаленное изображение, опираясь на текстовое описание и другие имеющиеся изображения. В исследовании также предлагается варьирование количества изображений.Точность результатов отслеживается с помощью метрик качества генераций изображений.

Для основы исследования я использую статью IP-Adapter и Latent Diffusion, а также датасет CELEBa в качестве исходных данных. Основными метриками качества будут FID, IS.

2 Постановка задачи

Обычная диффузионная модель минимизирует данную функцию потерь:

$$L_{\text{simple}} = \mathbb{E}_{x_0,\epsilon\sim N(0,I),c,t}||\epsilon - \hat{\epsilon}_\theta(x_t, c, t)||^2 \tag{1}$$

где x_0 представляет собой исходное изображение с дополнительным условием c, $t \in [0, T]$ обозначает шаг времени диффузионного процесса, $x_t = \alpha_t x_0 + \sigma_t \epsilon$ — шумные данные на t-м шаге, а α_t, σ_t — заранее определенные функции от t, определяющие процесс диффузии.

$$\hat{\epsilon}_\theta(x_t, c, t) = w\epsilon_\theta(x_t, c, t) + (1 - w)\epsilon_\theta(x_t, t) \tag{2}$$

здесь w является нормировочной константой, которая регулирует соответствие условию c. Для моделей диффузии такой выбор $\hat{\epsilon}_\theta(x_t, c, t)$ играет ключевую роль в улучшении соответствия изображения тексту сгенерированных образцов.

С помощью IP-Adapter признаки изображения интегрируются в заранее обученную модель UNet с помощью cross-attention. В оригинальной модели Stable Diffusion закодированный текст подается в модель UNet через слои перекрестного внимания. Учитывая признаки запроса Z и признаки текста c_t, выход перекрестного внимания Z' может быть определен следующим уравнением:

$$Z' = \text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{3}$$

где Q, K и V - матрицы запроса, ключа и значений операции внимания, соответственно, а W_q, W_k, W_v — матрицы весов обучаемых слоев линейной проекции, переменная d играет роль нормирующей константы.

Модифицированный IP-адаптер обучается на наборе данных $X = \{(x_i, c_i) : i = 1, \ldots, N\}$, где x_i — входное изображение, c_i — соответствующая ему текстовая и графическая подсказка. Задача минимизировать данную функцию потерь:

$$L_{simple} = \arg\min_{\hat{\epsilon}_\theta} \mathbb{E}_{x_0,\epsilon\sim N(0,I),c,c_1,\ldots c_N,t}||\epsilon - \hat{\epsilon}_\theta(x_t, c, c_1, \ldots c_N, t)||^2 \tag{4}$$

Также используется нормировочное условие на изображения и текст в процессе обучения:

$$\hat{\epsilon_\theta}(x_t, c_1, \ldots c_N, t) = w\epsilon_\theta(x_t, c_1, \ldots c_N, t) + (1 - w)\epsilon_\theta(x_t, t) \tag{5}$$

Я использую структуру для cross-attention к изображениям, аналогичную текстовому cross-attention. Следовательно, необходимо добавить $2N$ параметров вида W_{k_p}' и W_{v_p}' для каждого слоя UNet, где N - количество изображений. Для ускорения сходимости параметры W_{k_p}' и W_{v_p}' инициализируются из W_k и W_v. Затем мы используем агрегирующие функции $G(K_1', K_2', \ldots K_N')$ и $U(V_1', V_2', \ldots V_N')$ для получения конечной структуры cross-attention:

Поскольку мы замораживаем оригинальную модель UNet, только W_{k_p}' и W_{v_p}' обучаемы в вышеупомянутом методе. На стадии вывода λ является нормировочным коэффицентом, который помогает настраивать вес условия изображения:

$$Z'' = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V + \lambda \cdot \text{Softmax}\left(\frac{QG^T}{\sqrt{d}}\right)U \tag{6}$$

где $\text{Q} = ZW_q$, $\text{K} = c_t W_k$, $\text{V} = c_t W_v$, $K_p' = c_p W_{k_p}'$, $V_p' = c_p W_{v_p}'$

Стоит заметить, что модель становится оригинальной моделью распространения текста в изображение, если $\lambda = 0$.

Метрики качества:

Frechet Inception Distance (FID), Inception Score (IS) — это метрики качества, которые используются для оценки качества сгенерированных изображений

Формула для FID:

$$FID = ||\mu_p - \mu_q||^2 + Tr(\Sigma_p + \Sigma_q - 2(\Sigma_p\Sigma_q)^{1/2}) \tag{7}$$

где μ_p и μ_q — средние значения признаков в реальных и сгенерированных изображениях соответственно, Σ_p и Σ_q — ковариационные матрицы для распределений признаков в реальных и сгенерированных изображениях соответственно.

Формула для IS:

$$IS(x) = \exp(\mathbb{E}_x \left[D_{KL}(p(y|x)||p(y))\right]) \tag{8}$$

Где D_{KL} - дивергенция Кульбака-Лейблера для двух распределений $p(y|x)$ - вероятность класса y для изображения x и $p(y)$ - равномерное распределение на множестве классов, $\mathbb{E}_x$ - математическое ожидание по всем изображениям x.

Список литературы

[1] "IP-Adapter"`https://arxiv.org/pdf/2308.06721.pdf`.

[2] "Latent Stable Diffusion"`https://arxiv.org/abs/2112.10752.pdf`.