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 · Stable Diffusion

1 Введение

В современных исследованиях активно развивается модель генерации изображений под названием Stable Diffusion. 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, такие как недостаточное соответствие сгенерированных изображений исходным подсказкам.

Для решения этой проблемы был предложен метод IP-Adapter, который представляет собой легкий способ адаптации изображений к текстовым подсказкам с использованием стратегии кросс-внимания. Этот метод внедряется в существующие текстово-изображенческие диффузионные модели, что помогает улучшить точность генерации изображений.

Целью моего исследования является разработка метода на основе IP-Adapter для улучшения точности генерации изображений путем увеличения объема обучающих данных. Моя мотивация заключается в повышении качества визуальных представлений за счет более точного соответствия текстовым и графическим подсказкам.

Объектом моего исследования будет метод IP-Adapter в контексте модели Stable Diffusion. Для основы исследования я планирую использовать статью о IP-Adapter и Latent Diffusion, а также датасет CELEBa в качестве исходных данных.

2 Постановка задачи

Обычная диффузионная модель минимизирует данную функцию потерь:

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

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.

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

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.

Признаки изображения интегрируются в заранее обученную модель $UNet$ с помощью cross-attention.В оригинальной модели Stable Diffusion признаки закодированного текста подаются 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:

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

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.

Я использую ту же структуру для cross-attention к изображениям, что и для текстового cross-attention. Следовательно необходимо добавить $2N$ параметров W_{k_p}' и W_{v_p}' для каждого слоя cross-attention, 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 $G(K_1', K_2', ... K_N')$ и $U(V_1', V_2', ... V_N')$ чтобы получить конечную структуру cross-attention:

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

где Q $= ZW_q$, K $= c_tW_k$, V $= c_tW_v$, $K_p' = c_pW_{k_p}'$, $V_p' = c_pW_{v_p}'$

Поскольку мы замораживаем оригинальную модель UNet, только W_{k_p}' и W_{v_p}' обучаемы в вышеприведенном отделенном перекрестном внимании.

Во время обучения мы оптимизируем только IP-адаптер, оставляя параметры предобученной модели дифузии фиксированными. IP-адаптер также обучается на наборе данных с изображениями и текстом, используя ту же целевую функцию обучения, что и оригинальный SD:

$$L_{simple} = E_{x_0,\epsilon,c_t,c_1...c_N,t} \left\|\epsilon - \epsilon_\theta(x_t, c_t, c_1...c_N, t)\right\|^2$$

Мы также случайным образом нормируем условие на изображения и текст на стадии обучения:

$$\hat{\epsilon_\theta}(x_t, c_t, c_1...c_N, t) = w\epsilon_\theta(x_t, c_t, c_1...c_N, t) + (1 - w)\epsilon_\theta(x_t, t)$$

На стадии вывода λ является нормировочным коэффицентом, который помогает настраивать вес условия изображения:

$$Z_{new} = Attention(Q, K, V) + \lambda \cdot Attention(Q, K', V')$$

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