Создание персонализированных генераций изображений

Степанов Илья Казистова Кристина

Московский физико-технический институт

Курс: Научный трек иннпрака ФПМИ

Эксперт: Филатов Андрей Викторович

2024

Цель исследования

Задача
Сгенерировать изображения человека в различных вариациях в высоком разрешении.

Цель
Повысить качество изображений, генерируемых с помощью диффузионных моделей.

Проблема
Низкая точность генерации, неполное соответствие сгенерированных изображений текстовым описаниям, недостаточно высокое качество получаемого изображения.

Постановка задачи

Определим датасет как $\mathfrak{D} = \{(x_i, \tau_i) : i = 1, \ldots, n\}$, x_i — изображение, τ_i – соответствующий текстовый промпт. Рассматривается модель ϵ_θ из класса диффузионных моделей. На этапе обучения на каждом шаге из $\mathfrak{D}$ удаляется изображение $x_j, j \sim \mathcal{U}\{1, \ldots, n\}$.

Постановка задачи

Определим функцию потерь:

$$\mathcal{L}(\epsilon, \epsilon_\theta) = \mathbb{E}_{\epsilon \sim N(0,I), \mathbf{c}_\tau, \mathbf{c}_i, t, \mathbf{c}_t^j} \|\epsilon - \epsilon_\theta(\mathbf{c}_\tau, \mathbf{c}_i, t, \mathbf{c}_t^j)\|^2, \tag{1}$$

где $\mathbf{c}_\tau = \Gamma_\tau(\tau_j)$ — текстовые признаки удаленного изображения, полученные путем применения текстового энкодера Γ_τ к текстовому промпту τ_j; $\mathbf{c}_i = G(\Gamma_i(x_1), \ldots, \Gamma_i(x_{j-1}), \Gamma_i(x_{j+1}), \ldots, \Gamma_i(x_n))$ — признаки оставшихся изображений, являющиеся результатом применения агрегирующей функции G к эмбеддингам изображений, полученным с помощью image-энкодера Γ_i, $\mathbf{c}^j = \Gamma_i(x_j)$ — признаки удаленного изображения, $t \in [0, T]$ — временной шаг диффузионного процесса, $\mathbf{c}_t^j = \alpha_t \mathbf{c}^j + \sigma_t \epsilon$ — зашумленные данные удаленного изображения на шаге t, α_t, σ_t — предопределенные функции от t, определяющие диффузионный процесс.

Постановка задачи

Решается следующая оптимизационная задача:

$$\epsilon_\theta^* = \arg\min_{\epsilon_\theta} \mathcal{L}(\epsilon, \epsilon_\theta), \tag{2}$$

Постановка задачи

Для определения качества модели введем метрики качества генерации Frechet Inception Distance (FID) и Inception Score (IS):

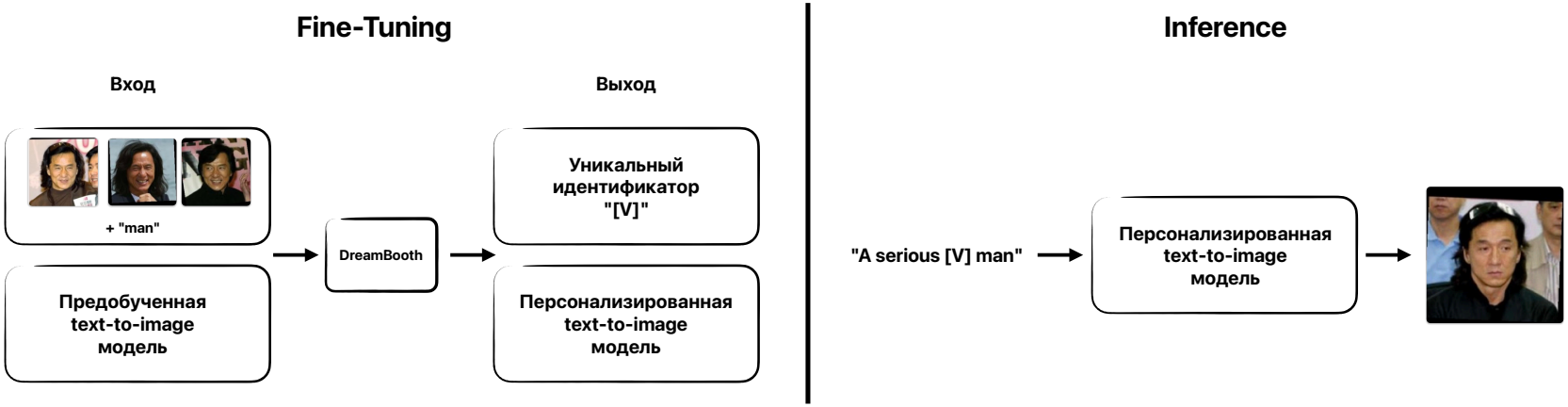
$$FID = ||\mu_p - \mu_q||^2 + Tr(\mathbf{\Sigma_p} + \mathbf{\Sigma_q} - 2(\mathbf{\Sigma_p}\mathbf{\Sigma_q})^{1/2}) \tag{3}$$

где μ_p и μ_q — средние значения признаков в реальных и сгенерированных изображениях соответственно, $\mathbf{\Sigma_p}$ и $\mathbf{\Sigma_q}$ — ковариационные матрицы для распределений признаков на реальных и сгенерированных изображениях соответственно.

$$IS(x) = \exp(\mathbb{E}_x\left[D_{KL}(p(y|x)||p(y))\right]) \tag{4}$$

где D_{KL} - дивергенция Кульбака-Лейблера для двух распределений; $p(y|x)$ - вероятность класса y для изображения x; $p(y)$ - равномерное распределение на множестве классов.

DreamBooth



- Принимает на вход несколько изображений одного объекта вместе с соответствующим названием класса;
- Возвращает специальный токен, идентифицирующий объект;
- Токен встраивается в текстовую подсказку, по которой генерируется желаемое изображение.

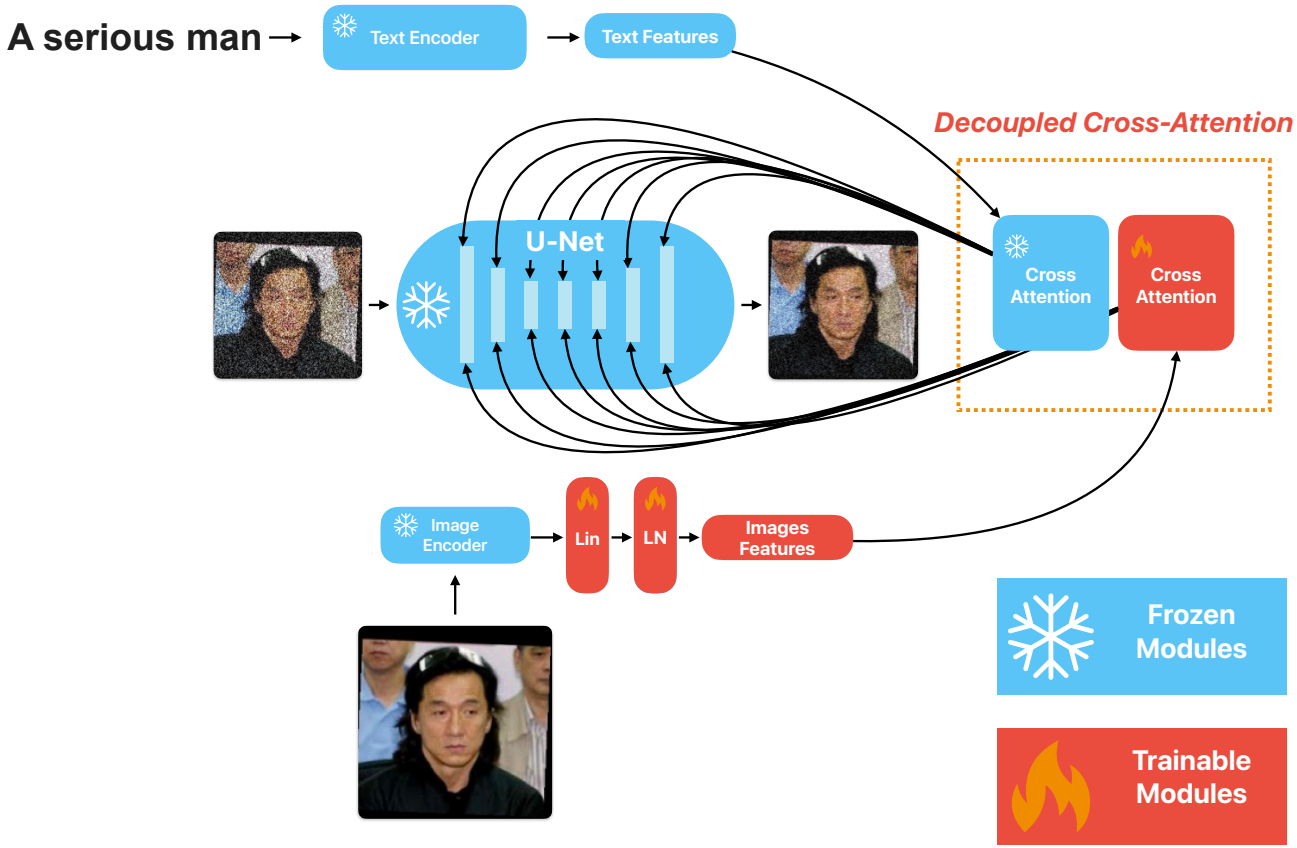
DreamBooth

Функция потерь принимает следующий вид:

$$\mathbb{E}_{\mathbf{x},\epsilon,\epsilon',\mathbf{c},t}[w_t \|\hat{x}_\theta(\alpha_t \mathbf{x}+\sigma_t \epsilon, \mathbf{c})-\mathbf{x}\|^2+\lambda w_{t'} \|\hat{x}_\theta(\alpha_{t'} \mathbf{x_{pr}}+\sigma_{t'} \epsilon', \mathbf{c_{pr}})-\mathbf{x_{pr}}\|^2],$$

где $\mathbf{x}$ — исходное изображение; $\mathbf{c} = \Gamma(P)$ — вектор условия (Γ — текстовый энкодер, P — текстовый промпт); $t \in [0, T]$ — временной шаг диффузионного процесса; α_t, σ_t, w_t — предопределенные функции от t, определяющие процесс диффузии; $\mathbf{x_{pr}} = \hat{x}(z, \mathbf{c_{pr}})$ — генерируемые данные с использованием сэмплера на основе предобученной диффузионной модели со случайным начальным шумом $z \sim \mathcal{N}(0, I)$ и вектором условия $\mathbf{c_{pr}} := \Gamma(f(\text{"a [class noun]"}))$, где f — токенизатор; λ — весовой коэффициент.

IP-Adapter



- Энкодер для извлечения признаков изображения;
- Адаптированные модули с механизмом перекрестного внимания.

Decoupled Cross-Attention

Выход слоя Cross-Attention для текстовых признаков $\mathbf{c_t}$:

$$\mathbf{Z}' = Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Softmax(\frac{\mathbf{Q}\mathbf{K}^\mathbf{T}}{\sqrt{d}})\mathbf{V}, \tag{5}$$

где $\mathbf{Z}$ — признаки запроса, $\mathbf{Q} = \mathbf{Z}\mathbf{W_q}, \mathbf{K} = \mathbf{c_t}\mathbf{W_k}, \mathbf{V} = \mathbf{c_t}\mathbf{W_v}$ — матрицы запросов, ключей и значений механизма внимания для текстовых признаков соответственно, а $\mathbf{W_q}, \mathbf{W_k}, \mathbf{W_v}$ — соответствующие матрицы весов.
Выход слоя Cross-Attention для признаков изображения $\mathbf{c_i}$:

$$\mathbf{Z}'' = Attention(\mathbf{Q}, \mathbf{K}', \mathbf{V}') = Softmax(\frac{\mathbf{Q}(\mathbf{K}')^\mathbf{T}}{\sqrt{d}})\mathbf{V}', \tag{6}$$

где $\mathbf{Q} = \mathbf{Z}\mathbf{W_q}, \mathbf{K}' = \mathbf{c_i}\mathbf{W}'_\mathbf{k}, \mathbf{V}' = \mathbf{c_i}\mathbf{W}'_\mathbf{v}$ — матрицы запросов, ключей и значений механизма внимания для признаков изображения соответственно, а $\mathbf{W}'_\mathbf{k}$, $\mathbf{W}'_\mathbf{v}$ — соответствующие матрицы весов.

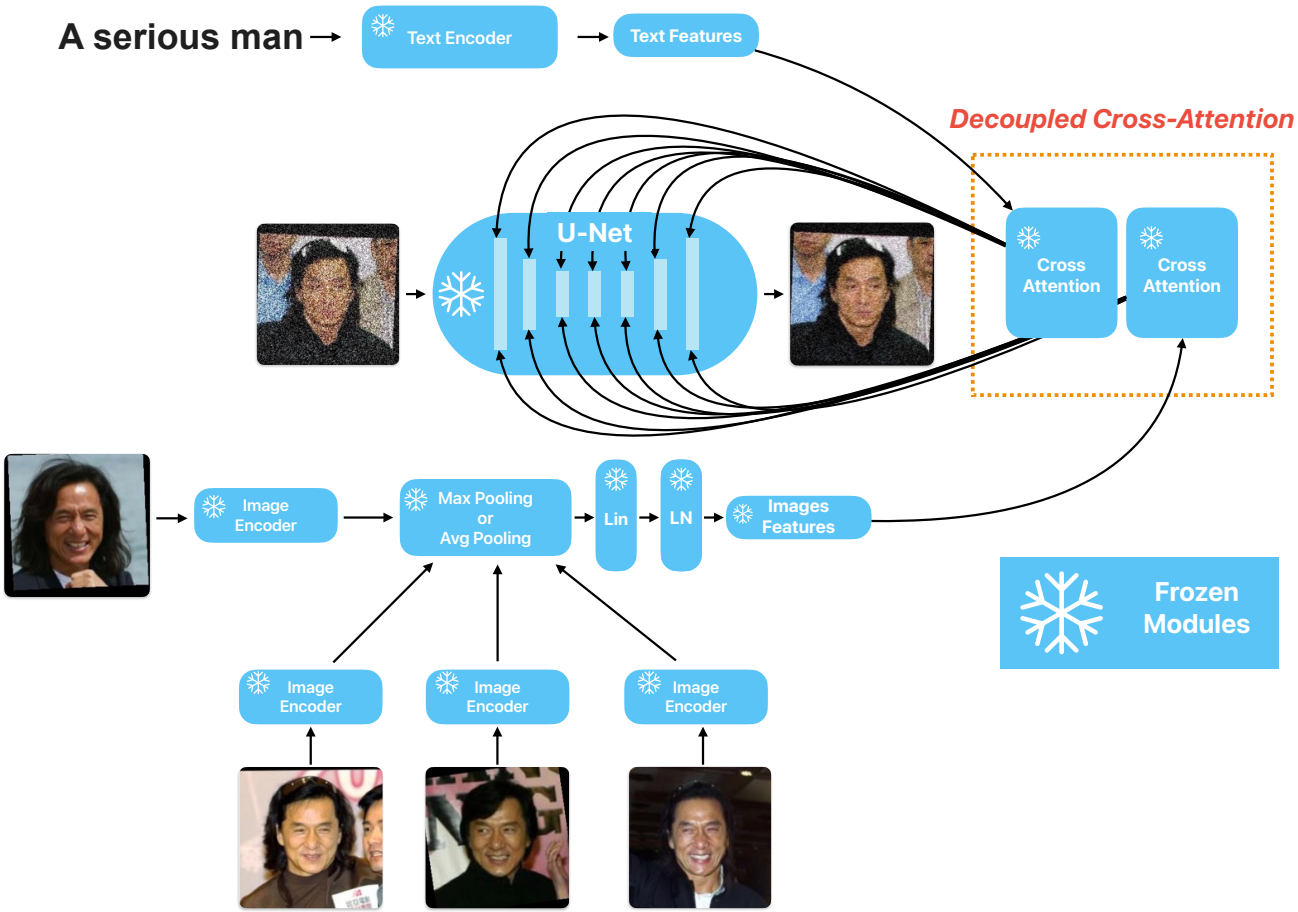
Decoupled Cross-Attention

Выход слоя Decoupled Cross-Attention:

$$\mathbf{Z}^{new} = \mathit{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) + \lambda \cdot \mathit{Attention}(\mathbf{Q}, \mathbf{K}', \mathbf{V}'), \quad (7)$$

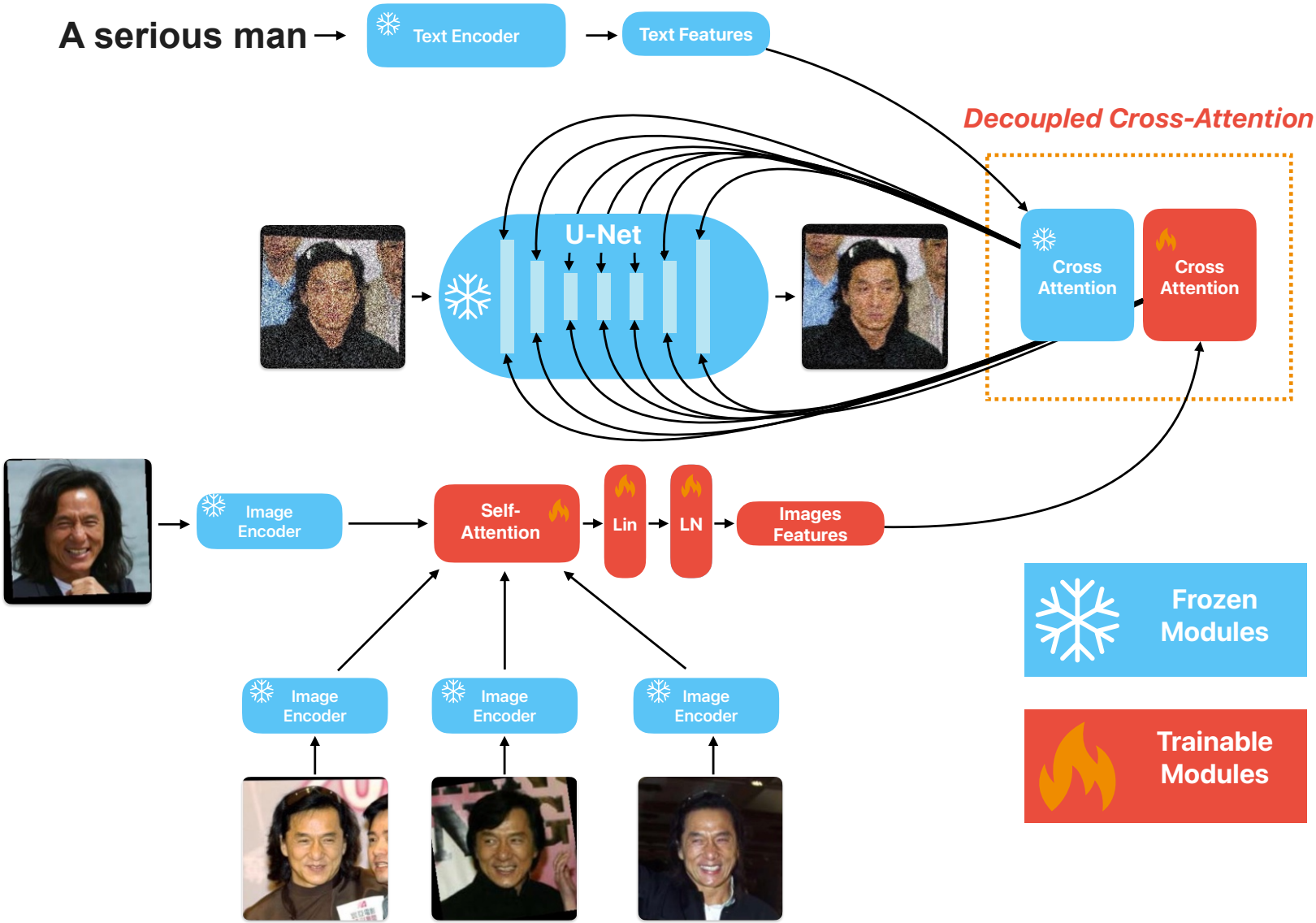
где λ — весовой коэффициент.

IP-Adapter + агрегирующая функция



- Принимает на вход несколько изображений;
- К эмбеддингам входных изображений применяется Max Pooling или Avg Pooling.

IP-Adapter + Self-Attention



IP-Adapter + Self-Attention

Процедура обучения:

- Входные данные: 10 изображений одного объекта с текстовыми промптами;
- Случайно выбранное изображение удаляется из рассмотрения;
- Модель учится предсказывать выброшенное изображение по его текстовому промпту и эмбеддингам оставшихся изображений.

Датасет

LFW Deep Funneled — набор изображений лиц людей вместе с их именами. В этом наборе данных 100 людям соответствует не меньше десяти разных фотографий.

Результаты экспериментов

Метод	**IS** $\uparrow$	**FID** $\downarrow$
IP-Adapter	15.37	8.92
DreamBooth	17.64	9.61
IP-Adapter + Max Pooling	14.12	10.10
IP-Adapter + Avg Pooling	13.56	11.82
IP-Adapter + Self-Attention	**18.72**	**7.56**

Результаты генерации

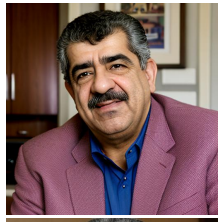
DreamBooth



IP-Adapter



IP-Adapter +
Max Pooling

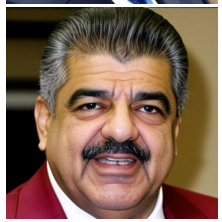




IP-Adapter +
Avg Pooling



IP-Adapter +
Self-Attention





Заключение

- Модификация метода IP-Adapter с использованием Self-Attention показала наилучший результат по метрикам качества IS и FID;
- Далее можно модифицировать наши методы посредством использования LoRa, FaceNet и др.

Ссылки