Создание персонализированных генераций изображений

К. М. Казистова
ФПМИ
МФТИ
Долгопрудный
kazistova.km@phystech.edu

И. Д. Степанов
ФПМИ
МФТИ
Долгопрудный
iliatut94@gmail.com

А. В. Филатов
Сколковский Институт Технологий
Москва
filatovandreiv@gmail.com

УДК 004.855,004.853

1 Тезис

Существующие модели способны генерировать разнообразные изображения по текстовым описаниям с высокой точностью. Однако, в процессе работы с моделями генерации изображений возникают определенные проблемы, одной из которых является недостаточное соответствие сгенерированных изображений и исходным текстовым подсказкам. Наша задача заключается в повышении качества визуальных представлений за счет большего количества графических подсказок. В работе рассматриваются методы, которые позволяют решить вышеупомянутые проблемы, и затем сравниваются между собой. Все описанные далее подходы основаны на применении Stable Diffusion(2).

Первый представленный метод — это DreamBooth(3).Он принимает на вход несколько изображений одного объекта вместе с соответствующим названием класса и возвращает специальный токен, идентифицирующий объект, который затем встраивается в текстовую подсказку, по которой генерируется желаемое изображение. Проблемы данного метода заключаются в слабой адаптивности, отсутствии обобщения и необходимости обучать всю диффузионную модель.

Второй метод — это IP-Adapter(1). Он состоит из двух частей: энкодера для извлечения признаков изображения, текста и адаптированных модулей с механизмом перекрестного внимания. Метод принимает на вход только одно изображение объекта. Однако одной картинки может быть мало, для того чтобы модель могла уловить все необходимые зависимости.

В работе предлагается третий метод, представляющий собой модификацию IP-Adapter. На вход подаются несколько изображений вместо одного, причем каждому изображению соответствует своя текстовая подсказка. В процессе обучения модели одно изображение удаляется равновероятно, и модель учится восстанавливать это удаленное изображение, опираясь на текстовое описание и другие имеющиеся изображения. К этим имеющимся изображениям применяется агрегирующая функция. За счет подачи нескольких изображений добиваемся лучшей передачи идентичности. Рассмотренные методы сравниваются между собой по метрикам качества генерации и разнообразия, метрикам идентичности. Исследование проводится на выборке из датасета LFW Deep Funneled(5) — датасете изображений знаменитостей в высоком разрешении.

Определим датасет как $\mathfrak{D} = \{(\mathbf{x}_i, \tau_i) : i = 1, \ldots, n\}$, $\mathbf{x}_i$ — латентное представление изображения, τ_i – текстовая подсказка. На этапе обучения на каждом шаге из $\mathfrak{D}$ удаляется изображение $\mathbf{x}_j, j \sim \mathcal{U}\{1, \ldots, n\}$ и решается следующая оптимизационная задача:

$$\epsilon_\theta^* = \arg\min_{\epsilon_\theta} \mathcal{L}(\epsilon, \epsilon_\theta), \tag{1}$$

Определим функцию потерь:

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

где G — агрегирующая функция, применяемая ко входным данным; $\mathbf{c}_\tau$ — текстовые признаки удаленного изображения; $\mathbf{c}_i$ — признаки изображений; $\mathbf{c}^j$ — признаки удаленного изображения; $t \in [0, T]$ —

временной шаг диффузионного процесса; $\mathbf{x}_t^j = \alpha_t \mathbf{x}^j + \sigma_t \epsilon$ — зашумленные данные удаленного изображения на шаге t; α_t, σ_t — предопределенные функции от t, определяющие диффузионный процесс; ϵ_θ — цель обучения модели диффузии.

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

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

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 и $p(y)$ - равномерное распределение на множестве классов, $\mathbb{E}_x$ - математическое ожидание по всем изображениям x.

1.1 IP-AdapterMAX и IP-AdapterAVG

Данная модификация метода IP-Adapter включает в себя обработку нескольких изображений, к которым применяются аггрегирующие функции MAXpooling или AVGpooling для их латентных представлений. На вход подаются изображения людей, вычисляются эмбеддинги данных изображений, после чего к эмбеддингам применяются упомянутые ранее функции аггрегации. В данном случае полученное латентное представление интегрируется в полностью предобученную модель IP-Adapter. Вычисление метрик производится на всем датасете.

1.2 IP-AdapterSelf-Attention

Предложенная модификация метода IP-Adapter включает в себя обработку нескольких изображений, к которым применяется алгоритм Self-Attention(4) для их латентных представлений. Исходный датасет разделяется на тренировочную и тестовую выборки в соотношении 2 : 1. Тренировочная выборка содержит набор персон, каждая из которых обладает 10 изображениями, к каждому из которых прилагается текстовая подсказка. Обучение происходит на 9 изображениях: в ходе обучения осуществляется попытка предсказать 10-е изображение, используя текстовую подсказку и предварительно обработанные эмбеддинги 9 изображений. Выбор удаленного изображения осуществляется равновероятно.

Algorithm 1 Self-Attention

```
procedure Self-Attention(x)
    Q ← x · W_q
    K ← x · W_k
    V ← x · W_v
    Z ← softmax((Q·K^T)/√d_k) · V
    return Z · W_out
end procedure
```

После завершения этапа модуля Self-Attention последуют модули IP-Adapter без изменений. В данном случае обучаются модули Self-Attention, Linear и Cross-Attention. Поскольку модификация Self-Attention обучается на 9 изображениях, то если от пользователя поступит большее или меньшее число изображений, в первом случае лишние изображения просто удаляются, а во втором выполняется процедура бутстрэпа до достижения нужного количества картинок.

Таблица 1: Результаты эксперимента

Метод	IS	FID
IP-Adapter	15.37	8.92
DreamBooth	17.64	9.61
IP-AdapterMAX	14.12	10.10
IP-AdapterAVG	-	-
IP-AdapterSelf-Attention	-	-

Список литературы

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

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

[3] "DreamBooth"`https://arxiv.org/pdf/2208.12242.pdf`.

[4] "Attention"`https://arxiv.org/pdf/1706.03762.pdf`.

[5] "Dataset"`https://vis-www.cs.umass.edu/lfw/`.