УДК 004.855,004.853

# Создание персонализированных генераций изображений

К. М. Казистова[1], И. Д. Степанов[1], А. В. Филатов[2]

[1]Физтех-школа прикладной математики и информатики, Московский физико-технический институт, Москва, 117303, Россия.
[2]Сколковский институт науки и технологий, Москва, 121205, Россия.

**Аннотация**

Большие модели преобразования текста в изображение совершили значительный скачок в области искусственного интеллекта, обеспечив высококачественный и разнообразный синтез изображений из заданного текстового описания. Однако, когда возникает запрос на генерацию специфичного объекта, в нашем случае человека, модель не может сгенерировать его с необходимой точностью и передать его идентичность. Предлагается решение, которое будет способно генерировать изображения заданного человека в различных вариациях в высоком разрешении. В данной работе рассматриваются методы DreamBooth, IP-Adapter, а также предлагаются наши собственные методы. Они представляют собой различные модификации IP-Adapter'а и позволяют принимать на вход сразу несколько изображений, что улучшает качество генерации. Все методы сравниваются между собой.

## 1 DreamBooth[1]

Метод 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.

## 2 IP-Adapter[2]

IP-Adapter состоит из двух основных компонентов: энкодера, который извлекает признаки изображения и текста, и модулей адаптации с механизмом перекрестного внимания. Он принимает на вход одно изображение. По сравнению с моделью DreamBooth, IP-Adapter обладает большей адаптивностью. Данный подход включает свои модули в предварительно обученную диффузионную модель, что позволяет обучать только энкодер и механизм перекрестного внимания.

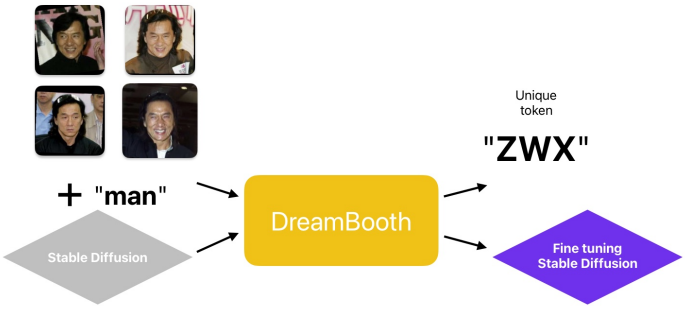


**Рис. 1**: DreamBooth

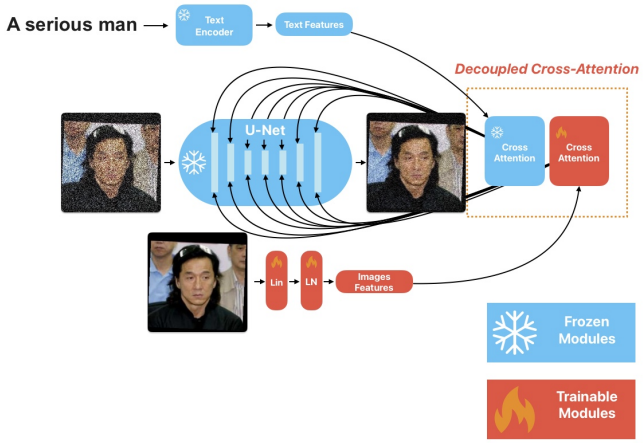


**Рис. 2**: IP-Adapter

## 3 IP-Adapter + агрегирующая функция

Данная модификация метода IP-Adapter включает в себя обработку нескольких изображений, к которым применяются агрегирующие функции Max-pooling или Avg-pooling для их латентных представлений. На вход

подаются изображения людей, вычисляются эмбеддинги данных изображений, после чего к эмбеддингам применяются упомянутые ранее функции агрегации. В данном случае полученное латентное представление интегрируется в полностью предобученную модель IP-Adapter.

# 4 IP-Adapter + Self-Attention

Предложенная модификация метода IP-Adapter включает в себя обработку нескольких изображений, к которым применяется алгоритм Self-Attention[3] для их латентных представлений. Исходный датасет разделяется на тренировочную и тестовую выборки в соотношении **2 : 1**. Тренировочная выборка содержит набор персон, каждой из которых соответствует 10 изображений, сопровождающихся текстовыми подсказками. 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.

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 бутстрепа для достижения нужного количества картинок.

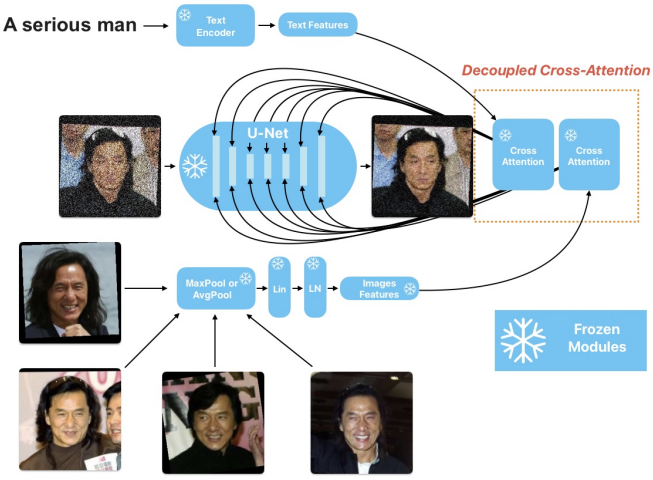


**Рис. 3**: IP-Adapter with pooling

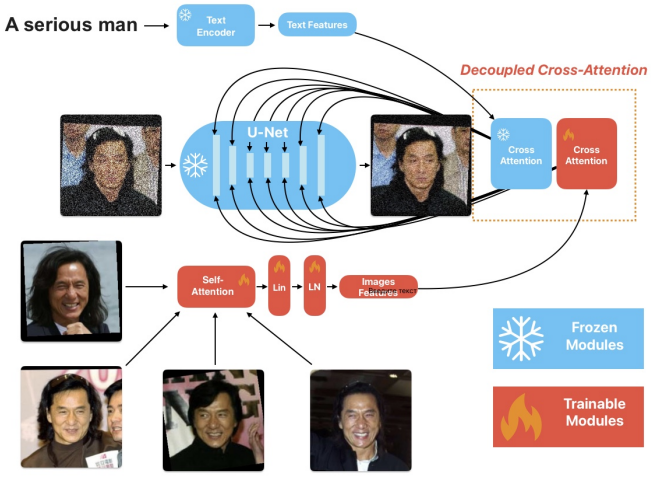


**Рис. 4**: IP-Adapter with Self-Attention

# 5 Результаты экспериментов

В экспериментах рассматривается задача генерации изображений с помощью существующих моделей DreamBooth, IP-Adapter, а также модификации IP-Adapter на датасете LFW Deep Funneled. Для оценки качества моделей используются такие метрики качества генерации изображений, как Frechet Inception Distance (FID) и Inception Score (IS). На текущий момент получены следующие результаты:

| Метод | IS | FID |
|---|---|---|
| IP-Adapter | 15.37 | 8.92 |
| DreamBooth | 17.64 | 9.61 |
| IP-AdapterMAX | 14.12 | 10.10 |

# Список литературы


[1] DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

[2] IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models

[3] Attention Is All You Need

[4] High-Resolution Image Synthesis with Latent Diffusion Models