Создание персонализированных генераций изображений

A Preprint

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

Abstract

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 рассмотрены методы DreamBooth, IP-Adapter, а также предложен наш собственный метод. Он представляет собой модификацию IP-Adapter'а и позволяет принимать на вход сразу несколько изображений, что улучшает качество генерации. Все методы сравниваются между собой.

1 Введение

В последние годы наблюдается быстрое развитие генеративных моделей, решающих задачу преобразования текста в изображение. Существующие модели способны генерировать разнообразные изображения по текстовым описаниям с высокой точностью. Однако зачастую возникает ряд проблем: точность генерации часто оказывается ниже требуемой, сгенерированные картинки не полностью соответствуют текстовым запросам, а качество изображения получается недостаточно высоким. Перед нами стоит задача генерации изображения человека в различных вариациях в высоком разрешении. В работе рассматриваются методы, решающие вышеупомянутые проблемы, и затем сравниваются между собой. Все описанные далее методы основываются на применении диффузионных моделей. Диффузионная модель - это модель, состоящая из двух процессов: прямого и обратного. Во время прямого процесса ко входным данным постепенно добавляется шум, во время обратного - модель постепенно восстанавливает данные из шума.

Первый представленный метод — это DreamBooth. 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. Он состоит из двух частей: энкодера для извлечения признаков изображения 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.

Третий метод разработан нами и представляет собой модификацию IP-Adapter'а. На вход подаются несколько изображений вместо одного, каждому изображению соответствует своя текстовая подсказка. Ко входным изображениям применяется аггрегирующая функция (pooling). За счет подачи нескольких изображений добиваемся лучшей передачи идентичности. Процедура обучения нашей модели

заключается в том, что на каждом шаге из множества входных изображений случайно выбирается одно, удаляется из рассмотрения, и модель учится восстанавливать выброшенное изображение по его текстовому описанию и оставшимся картинкам.

Рассмотренные методы сравниваются между собой по метрикам качества генерации, метрикам идентичности и метрикам разнообразия.

Исследование проводится на датасете CelebA — датасете изображений знаменитостей в высоком разрешении.

2 Постановка задачи

Определим датасет как $X_0 = \{(x_i, t_i) : i = 1, \dots, n\}$, где x_i — входное изображение, t_i — соответствующая ему текстовая подсказка. Рассматривается модель из класса диффузионных моделей. На этапе обучения на каждом шаге из датасета X_0 удаляется изображение $x_j, j \sim \mathcal{U}\{1, \dots, n\}$ и решается следующая оптимизационная задача:

$$\theta^* = \arg\min_\theta \mathcal{L}(G(X_0 \backslash \{x_j\}), \epsilon, c_t, c_i \backslash \{c^j\}, t, x_t, \epsilon_\theta),$$

где G — агрегирующая функция, применяемая ко входным данным; c_t — текстовые признаки; c_i — признаки изображений; c^j — признаки удаленного изображения; $t \in [0, T]$ — временной шаг диффузионного процесса; $X_t = \alpha_t G(X_0 \backslash \{x_j\}) + \sigma_t \epsilon$ — зашумленные данные на шаге t; α_t, σ_t — предопределенные функции от t, определяющие диффузионный процесс; ϵ_θ — цель обучения модели диффузии.

Определим функцию потерь:

$$\mathcal{L} = \mathbb{E}_{X_0 \backslash \{x_j\}, \epsilon, c_t, c_i \backslash \{c^j\}, t} \|\epsilon - \epsilon_\theta(X_t, c_t, c_i \backslash \{c^j\}, t)\|^2$$

Мы также случайным образом отбрасываем условия изображения на этапе обучения, чтобы обеспечить управление без использования классификатора на этапе вывода.

$$\hat{\epsilon}_\theta(X_t, c_t, c_i \backslash \{c^j\}, t) = w\epsilon_\theta(X_t, c_t, c_i \backslash \{c^j\}, t) + (1 - w)\epsilon_\theta(X_t, t)$$

Поскольку перекрестное внимание к тексту и перекрестное внимание к изображению разделены, мы также можем настроить вес условия изображения на этапе вывода:

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

где λ — весовой коэффициент, $\mathbf{Z}$ — признаки запроса, $\mathbf{Q} = \mathbf{Z}\mathbf{W}_q$, $\mathbf{K} = \mathbf{c}_t\mathbf{W}_k, \mathbf{K}^{'} = (\mathbf{c}_i \backslash \{\mathbf{c}^j\})\mathbf{W}_k^{'}, \mathbf{V} = \mathbf{c}_t\mathbf{W}_v, \mathbf{V}^{'} = (\mathbf{c}_i \backslash \{\mathbf{c}^j\})\mathbf{W}_v^{'}$ — матрицы запросов, ключей и значений механизмов внимания для текста и изображений соответственно, а $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_k^{'}, \mathbf{W}_v, \mathbf{W}_v^{'}$ — соответствующие весовые матрицы.

Для определения точности модели введем метрику качества генерации FID:

$$d_F(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\mu', \Sigma'))^2 = \mu - \mu'^2_2 + tr\left(\Sigma + \Sigma' - 2\left(\Sigma\Sigma'\right)^{\frac{1}{2}}\right),$$

где $\mathcal{N}(\mu, \Sigma), \mathcal{N}(\mu', \Sigma')$ — многомерные гауссовские распределения.

Список литературы