Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding

(Фотореалистичные модели преобразования текста в изображение с глубоким пониманием языка)

Автор обзора-рецензии: Фролова Анна

Данный препринт представляет нам новую модель Imagen, которая преобразует текст в изображение с глубоким понимаем языка. Imagen собирает в себе комбинацию больших языковых моделей и диффузионных моделей для создания изображений высокой точности. Ключевое открытие данной статьи заключается в том, что большие языковые модели, предварительно обученные на текстовых корпусах, удивительно эффективны при кодировании текста для синтеза изображений.

Препринт был опубликован в мае 2022 года канадской исследовательской группой гугла. Один из основных авторов Chitwan Saharia ранее занимался исследованием диффузионных моделей и написал несколько статей по обработке и генерации изображений. Наиболее похожая статья была опубликована в мае 2022 и называлась “Image-to-Image Diffusion Models” (препринт был написан в ноябре 2021). Она представила новую модель Palette по преобразованию изображения в изображение при помощи диффузионных генеративных моделей, которая показала очень высокие результаты в задачах colorization (раскрашивания), inpainting (дополнения изображения), cropping (обрезки) и удалению артефактов JPEG. Думаю генерация изображений из текста стало продолжением исследовательской работы автора в области диффузионных моделей.

Модель Imagen является продолжением исследований в области синтеза текста с изображением и контрастивного обучения изображение-текст. Наиболее важными статьями мне показались следующие:

1) “Zero-Shot Text-to-Image Generation” (февраль 2021) - статья, в которой описан достаточно простой подход к решению задачи преобразования текста в изображение, основанный на преобразователе, который авторегрессионно моделирует текстовые и графические маркеры как единый поток данных.

2) “High-Resolution Image Synthesis with Latent Diffusion Models” (апрель 2022) - статья, где представляют модели скрытой диффузии, которые достигают новых современных показателей в области визуального рисования и синтеза условных изображений, а также высокой производительности в различных задачах, включая преобразование текста в изображение, генерацию безусловных изображений и суперразрешение, при значительном сокращении вычислительных затрат.

Также значимой работой я бы назвала выше упомянутую статью “Image-to-Image Diffusion Models”, думаю на ее базе во многом строилось это исследование.

Так как эта работа - препринт и полная статья выйдет позже, прямых продолжений исследования пока нет. Но данный препринт довольно много цитируют, например статья “Diffusion Models: A Comprehensive Survey of Methods and Applications”, где сравниваются разные диффузионные модели.

У данной работы есть два основных конкурента - это GLIDE и DALL-E 2. В препринте довольно тщательно сравнивают результаты Imagen с двумя другими конкурирующими моделями.

В исследовании был введен новый набор категорий для текстовых подсказок под названием DrawBench для оценки качества моделей преобразования текста в изображение. Новый бенчмарк является всеобъемлющим и оценивает многие аспекты моделей, такие как композиционность, мощность и пространственные отношения, на основе 11 параметров. Эти параметры разделены на различные цвета, количество объектов в изображении, любой текст в изображении и взаимодействие между объектами.

DrawBench также постоянно использует более сложные и креативные подсказки или редко используемые слова, чтобы модель была хорошо знакома с этими командами. Это также повышает способность модели генерировать образы, которые являются более образными и необычными.

Imagen превзошел DALL-E 2 и GLIDE по оценкам людей-оценщиков. Imagen тестировался лучше других моделей с большим отрывом как по точности изображения, так и по выравниванию изображения и текста.

Есть ряд вещей, которые Imagen сделал по-другому, чтобы превзойти DALL-E 2 и GLIDE. Imagen обучался с использованием крупнейшего текстового кодировщика T5-XXL, который имеет 4.6 миллиардов параметров. Исследование, по сути, показывает, что масштабирование размера текстового кодера в значительной степени улучшает выравнивание текста и изображения и точность изображения. Фактически, это доказывает, что масштабирование размера предварительно обученного текстового кодера намного полезнее, чем масштабирование размера диффузионной модели. В то время как масштабирование размера диффузионной модели U-Net приводит к улучшению качества выборки, больший кодировщик текста оказывает большее общее влияние.

В исследовании также была введена концепция динамического порогового значения (dynamic thresholding), нового метода диффузионной выборки, который выполняется на каждом этапе выборки, чтобы предотвратить насыщение пикселей. Этот метод позволил сделать изображения более фотореалистичными, особенно в случае больших весов наведения без классификаторов в образцах.

Imagen также использовал другой метод диффузии, называемый дополнением шумоподавления, который помогает моделям осознавать количество добавленного шума и, следовательно, делает их более надежными. Этот метод привел к повышению точности изображения и способствовал более высоким показателям FID и CLIP в Imagen.

Imagen использовал архитектуру U-Net для базовой модели распространения 64 * 64 и несколько модифицировал ее, чтобы сделать ее более эффективной. Архитектура U-Net

использует меньше памяти, быстрее сходится и имеет лучшее качество выборки при более быстром времени вывода.

Думаю Imagen стала одной из самых передовых моделей по генерации изображений из текста. Однако основной недостаток данной модели заключается в том, что в ней все еще очень много параметров для обучения. Для запуска и тестирования данной модели нужны очень большие мощности, для локального компьютера запустить такую модель невозможно. Думаю одним из самых перспективных направлений было бы облегчение модели, может за счет ухудшения качества полученных изображений, но чтобы модель стала доступной для обычных пользователей. Такую разработку можно было бы имплементировать в фоторедакторы или можно было бы создать приложение по генерации изображений. Такого рода внедрение могло бы изменить наше отношение к живописи и фотографии.