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

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, Mohammad Norouzi

Рецензент-исследователь: Катунькин Михаил, БПМИ-192

1. Суть работы

Авторы статьи развивают идею генерации изображений по текстовому описанию. В рамках работы они предложили модель Imagen. По утверждениям авторов статьи изображения, полученные при помощи Imagen, получаются более реалистичными и соответствующими описанию, нежели у ближайших конкурентов, вроде DALLE-2 или GLIDE.

Для генерации изображений Imagen использует представления текста, полученные из большого предобученного только на текстовом корпусе трансформера T5-XXL. Это одно из ключевых отличий Imagen от основных конкурентов, которые, используют CLIP-представления текста, обученные на парах текст-изображение.

Представления текста декодируются в изображение при помощи каскада диффузионных моделей. Первая модель из шума и представления текста генерирует изображение 64x64. Вторая и третья последовательно повышают разрешение сначала до 256x256, потом до 1024x1024.

В качестве основного научного вклада работы можно выделить следующее:

- Авторы выяснили, что для получения представления текста достаточно модели, предобученной только на текстовых данных. Увеличение числа параметров этой модели существенно влияет на качество генерации. При этом, увеличение числа параметров диффузионной модели не дает существенного прироста качества.
- Авторы предложили технику dynamic thresholding для семплирования изображений во время диффузионного процесса. По утверждению авторов статьи она делает сгенерированные изображения более фотореалистичными.
- Авторы предложили архитектуру Efficient U-Net, для предсказания шума в диффузионном процессе, которая более эффективна по времени инференса и памяти, а также обучается за меньшее число шагов.

- Авторы достигли SOTA FID 7.27 на датасете MS-COCO.
- Авторы предложили новый бенчмарк DrawBench для оценки качества генерации изображений по текстовому описанию.

2. Обстоятельства написания работы

Статья написана командой Brain Team из Google Research, опубликована 23 мая 2022. У статьи 204 цитирования, из которых 41 значимое по версии Semantic Scholar.

Основные авторы:

- Chitwan Saharia – занимается диффузионными моделями для генерации изображений, а также неавторегрессивными генеративными моделями для текстов. Близкая по теме предыдущая работа – модель SR3 для задачи повышения разрешения. Imagen также использует каскад диффузионных моделей для повышения разрешения.
- William Chan – в основном занимается звуковыми задачами. Один из основных авторов популярных подходов SpecAugment и LAS (почти 2000 цитирований у каждой). Из релевантного данной работе – вместе с Chitwan Saharia автор модели SR3.
- Saurabh Saxena – один из авторов работы Pix2Seq, которая является обобщенным фреймворком для задачи детекции объектов.
- Lala Li – также Pix2Seq
- Jay Whang – автор статьи про проблемы обратимых генеративных моделей
- Jonathan Ho – автор статьи Denoising Diffusion Probabilistic Models, в которой было предложено использовать диффузионные модели для генерации изображений.

Статья про Imagen выглядит как логичное продолжение существующих работ. Пайплайн – это компиляция уже существующих подходов: взять представление текста, на его основе сгенерировать изображение каскадом диффузионных моделей. При этом, авторы провели детальное исследование отдельных компонент, и выяснили, как их можно улучшить или на что заменить.

3. Статьи с наибольшим влиянием на работу

- “Image Super-Resolution via Iterative Refinement” – статья от Chitwan Saharia, одного из авторов Imagen. В ней предлагается использовать каскад диффузионных моделей для последовательной генерации изображений все большего и большего качества.

- “GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models” – работа, которую авторы Imagen называют предшествующей (хотя дата выхода на arxiv – март 2022…) По сути, эта работа глобально определила фреймворк, который улучшали авторы Imagen. В частности, в этой работе показали, что техника classifier-free guidance (с некоторой вероятностью веса эмбеддинга текста зануляются) улучшает качество сгенерированных изображений. Также в ней задействовали технику super-resolution из предыдущего пункта. Кроме этого, было показано, как использовать диффузионную модель с CLIP-эмбеддингами для редактирования изображений текстом и решения задачи дорисовки изображений (это использовалось в DALLE-2).

4. Цитирования

Статей с развитием архитектуры Imagen мне найти не удалось. Я бы разделил цитирующие статьи на три категории:

- Fine-tuning модели для генерации специфичных изображений. Например, “DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation”. Или “An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion”
- Новые архитектуры, сравнивающиеся с Imagen. Например, “Scaling Autoregressive Models for Content-Rich Text-to-Image Generation” побила SOTA достигнутую Imagen.
- Статьи про новые датасеты и бенчмарки, описание существующих методов. В них Imagen указывается наряду с другими общепризнанными моделями вроде DALLE-2, GLIDE или Stable Diffusion

5. Основные конкуренты

- GLIDE – предшественник DALLE-2 и Imagen.
- DALLE-2 – предыдущая SOTA на датасете MS-COCO, конкурирующая работа. Архитектура модели очень похожа на архитектуру Imagen. Однако в качестве эмбеддингов текста, как и GLIDE, использует CLIP-представления. За счет этого, хоть и проигрывает в качестве на задаче генерации изображений по текстовому описанию, обладает более широким функционалом, нежели Imagen. В частности, может редактировать изображения по текстовым командам и дорисовывать изображения. Кроме этого, у DALLE-2 есть доступое API.

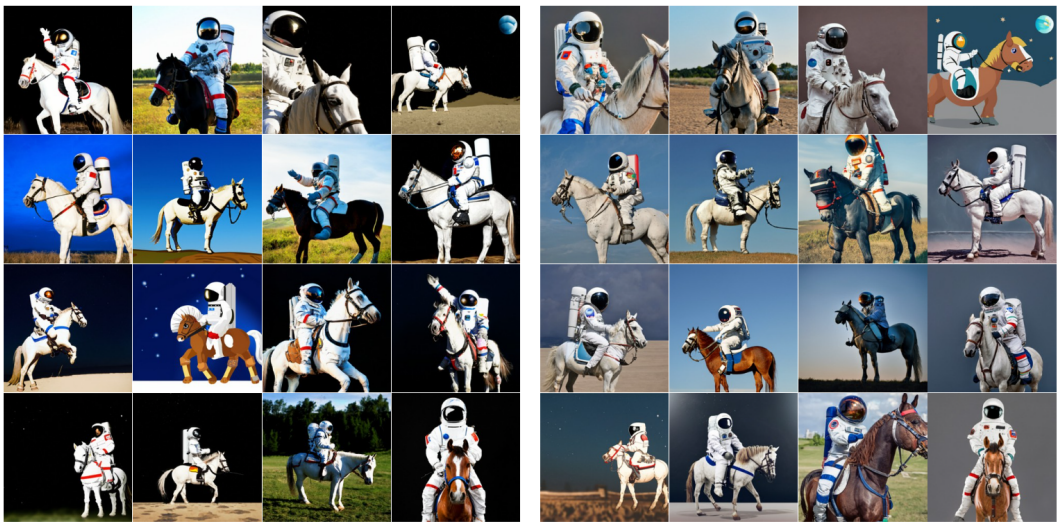
- Stable Diffusion – модель, значительно оптимизировавшая обучение и применение диффузионных моделей. В основе идея о том, что изображение можно перевести в латентное пространство меньшей размерности и выполнять диффузионный процесс уже там. Хоть Stable Diffusion и проигрывает DALLE-2 по качеству генерации, она обладает примерно схожим функционалом и у этой модели есть веса в открытом доступе. Помимо этого ее можно запускать на видеокартах с 8gb памяти.

6. Сильные стороны работы

- Очень хорошая структура статьи. Авторы раскрывают суть работы в три уровня детализации и подсвечивают, что именно нового сделано в статье. За счет этого можно понять общую концепцию почти сразу же. Это ускоряет и облегчает понимание материала. Дополнительно, в приложении к статье есть теорминимум по диффузионным моделям.
- Статья очень наглядная. Большая часть тезисов иллюстрирована схемами, графиками или изображениями. В статье приведено много примеров работы как Imagen, так и конкурирующих моделей.
- По каждому пункту, который авторы декларируют как научный вклад работы было проведено разностороннее исследование. Так, авторы описывают логику, которая стоит за выбором тех или иных решений. В статье приводятся результаты экспериментов, обосновывающих этот выбор. Более того, эксперименты проводились в разнообразных разрезах, чтобы оценить влияние параметров на разные аспекты генерации.
- Было проведено исследование не только по части архитектуры модели, но и по части методов оценки качества полученных моделей. Авторы выявили ряд недостатков существующего бенчмарка MS-COCO. В частности, он недостаточно разнообразен, не оценивает модели в разных срезах и не учитывает все разнообразие аспектов качества сгенерированных изображений. Поэтому они предложили свой бенчмарк DrawBench на основе оценок людей.
- Приведено сравнение результатов генерации с DALLE-2 и GLIDE в разных разрезах. В статье отражены слабые стороны не только конкурирующих моделей, но и самого Imagen.
- Imagen с неплохим отрывом побил SOTA по генерации изображения по текстовому описанию по метрике Zero-shot FID-30K на MS-COCO.

7. Слабые стороны работы

- Несмотря на то, что Imagen побил результаты DALLE-2 и подобных в задаче генерации изображения по текстовому описанию – в нем отказались от использования CLIP-эмбеддингов. За счет этого модель не может быть использована, чтобы модифицировать изображения текстом, или проводить другие манипуляции с изображениями. В целом, по сравнению с конкурентами у модели сильно меньший функционал.
- Авторская техника dynamic thresholding для семплирования изображений в диффузионном процессе дает противоречивые результаты. Утверждается, что она повышает реалистичность изображений. Но при этом, судя по примерам из статьи она не дает модели создавать контрастные изображения. Так на этих изображениях темное космическое небо превращается в серую массу.



(a) Samples using static thresholding. (b) Samples using dynamic thresholding ($p = 99.5$)

- Авторский бенчмарк DrawBench состоит всего лишь из 200 изображений. При условии, что эти 200 изображений разбиты на 11 групп (в некоторых группах по 7 примеров) – сложно говорить о репрезентативности результатов, полученных в данном бенчмарке. Помимо этого, у авторов была возможность специально обучить модель под бенчмарк или подобрать бенчмарк под модель.
- Модель не выложена в открытый доступ. За счет этого сложно оценить качество генерации на собственных примерах.

8. Улучшения статьи, направления для дальнейших исследований

- После прочтения статьи мне осталась непонятной мотивация авторов заменить CLIP-эмбеддинги на эмбеддинги от текстовых трансформеров.

Графики со сравнениями FID и CLIP-score для эмбеддингов от T5-XXL и CLIP практически идентичны. Преимущество T5-XXL над CLIP показано только для DrawBench. Эти утверждения мне не кажутся репрезентативным из-за того, что там всего 200 изображений. Есть ощущение, что прирост качества на MS-COCO был из-за других нововведений.

- Есть ощущение, что dynamic thresholding – не самая универсальная эвристика для семплирования изображений в диффузионном процессе. Хотя ее выбор и улучшает общую метрику. Можно подумать над ее развитием.
- Идея бенчмарка DrawBench по своей сути мне кажется верной. Основной вопрос в количестве примеров в нем. Авторы статьи утверждают, что малый размер нужен для того, чтобы небольшие организации могли оценивать качество при помощи людей. Исходя из такой логики можно было бы создать бенчмаки DrawBench-S/M/L разных размеров. И крупные исследования от крупных research-команд проверять на крупном репрезентативном бенчмарке.