**Марчук Иван ИУ6-31М**

**Stable Diffusion**

Итак, недавно я начал осваивать программу Stable diffusion.

Она представляет из себя консольную программу содержащую нейросетевую модель выпущеную в 2022 году. Разработано группой компаний CompVis в Мюнхенском университете.

[1]

Модель Stable Diffusion использует сверточные нейронные сети для извлечения признаков из изображений, а затем применяет метод диффузии для объединения информации из разных частей изображения. Это позволяет модели учитывать контекст изображения и улучшить точность сегментации.

Особенностью модели Stable Diffusion является также использование двухшагового подхода, который позволяет ей сначала грубо разделить изображение на области, а затем более точно их сегментировать. Это позволяет ускорить процесс обработки изображений, сохраняя при этом высокую точность.

Внутреннее устройство

[2]

Есть детальнейшая статья, с картинками, об устройстве этой нейросети, и лучше чем там я увы не расскажу, а приводить её сейчас целиком увы будет долго, по этому просто дам ссылку.

Что ж, а я кратко расскажу про внутренне устройство этой модели.

Stable Diffusion — это не единая монолитная модель, а система, состоящая из множества компонентов. В модели есть компонент понимания текста, преобразующий текстовую информацию в цифровой вид, кодирующий смысл текста. Далее эта информация передаётся генератору изображений, который генерирует изображение в два этапа.

[3]

1) Создание информации о изображении или разбиение его на векторы эмбеддингов.

Этот компонент выполняется в несколько шагов (step), генерируя информацию о будущем изображении. Это параметр steps в интерфейсах Stable Diffusion.

Слово «диффузия» (diffusion) описывает происходящее в этом компоненте. Это пошаговая обработка информации, приводящая в конечном итоге к генерации высококачественного изображения декодера изображений.

[4]

2) второй этап — это декодирование изображений.

Декодер изображений рисует картину на основе информации, которую он получил на этапе создания информации. Он выполняется только один раз в конце цикла и создаёт готовое пиксельное изображение.

На изображении выше видно три основных компонента (каждый со своей собственной нейросетью), из которых состоит Stable Diffusion:

- ClipText для кодирования текста.

Входные данные: текст.

Выходные данные: 77 векторов эмбеддингов токенов, каждый в 768 измерениях.

- UNet + Scheduler (Юнет шедулер) для постепенной обработки, то есть диффузии информации в векторах получившихся эмбеддингов.

Входные данные: эмбеддинги текста и исходный многомерный массив (структурированные списки чисел, также называемые тензором), состоящий из шума.

Выходные данные: массив обработанной информации

- И последний Декодер автокодировщика, рисующий готовое изображение при помощи массива обработанной информации.

[5]

Что же такое диффузия, благодаря которой все и работает?

Диффузия — это процесс, выполняемый внутри розового компонента «image information creator» (на этапе создания информации о будущем изображении). Имея эмбеддинги токенов, описывающие введённый текст, и случайный начальный массив информации о будущем изображении, процесс подгоняет случайные значения массива с информацией.

Затем уже подготовленные массивы декодер изображения использует для рисования готового изображения.

Поскольку описание будущего изображения изначально задается случайным набором чисел, а генерация происходит в несколько этапов, то и на выходе мы получем что-то примерно такое:

Что-то похожее как раз и происходит с векторами эмбеддингов, в которых и хранится информация об изображении.

И тут есть забавный факт, если задавать один и тот-же ключ для генерации этого шума, то при одинаковых входных параметрах, мы получим одинаковые выходные изображения. Это очень удобно для правильной настройки входной информации или промптов, о которых мы сейчас и поговорим.

**А теперь про практическую часть**

[6]

[7]

Stable Diffusion это уже обученная модель, и она сама по себе неплохо гененрирует изображения, однако, любой желающий с достаточно мощной машиной и нужными знаниями может её дообучить на конкретных наборах изображений. Таким образом уже обученные тензоры меняются, и дообученную модель можно сохранить в файл model как «чекпоинт». Для дообучения авторы в среднем используют от 30 до 100 изображений.

Процесс этот достаточно затратный по железу и требует большого опыта обращения со SD, поэтому проще всего не обучать свою, а скачать уже готовую модель, которую уже обучили знающие люди, и эти же люди обычно дают советы какие правильные промпты использовать для их моделей. Также модель можно оптимизировать, так сказать «запечь» для того, чтобы убрать лишние узлы и обсчет нейросети при генерации картинок был более быстрым. Дообучать правда такие модели уже нельзя, но это не точно.

[8]

Что же такое промпты в SD?

Stable diffusion генерирует изображения исходя из текстового запроса. И тут есть очень хороший принцип, по которому мы вводим запрос (по английски prompt) того, что хотим получить - позитивный промпт. А также того, что на изображении недолжно быть - негативный промпт (negative prompt).

Промпт состоит из токенов. Токены это слова или сочетания слов. Эти токены один из слоев модели может разложить по признакам, которые уже сопоставляются с теми или иными картинками, на которых обучалась модель. Модель понимает и русский язык, однако в основном обучалась она на наборах с английскими токенами, и поэтому с ними получается составить наиболее точный и правильный запрос.

Благодаря подходу с позитивными и негативными промптами можно подобрать запрос, достаточно точно описывающий какое изображение мы хотим получить. И чем больше будет токенов, то есть слов или их комбинаций в запросе, тем больше сгенерированное изображение будет похоже на то, что вам нужно.

Таких токенов может быть много, и они могут иметь разную силу.

Причем у разных дообученных моделей будут свои интерпретации и сила токенов. Правильную комбинацию можно подбирать достаточно долго, каждый раз получая разный результат даже от простой перемены местами слов и тут главное не переборщить. Но об этом позже.

Про интерфейс

[9]

Самый классический вариант, это консольное окно. Когда после установки программа запускается в окне, где все пути до исходных файлов и промпты прописываются вручную. Естественно это самый первый и самый неудобный интерфейс.

Дергать ползунки это удобнее чем прописывать отдельные парамтры в консоли, и по этому на данный момент уже создано достаточно много интерфейсов, большинство из которых работает в окне браузера. Я лично опробовал два из них, это easy diffusion и веб интерфейс от пользователя «AUTOMATIC1111».

[10]

Easy diffusion хорош тем, что он работает из коробки все уже настроено, главное чтобы была подходящая видеокарта.

Можно удобно планировать очереди изображений в одном окне. А окно удобно открывается само в браузере по умолчанию.

Консольное окно никуда не делось, все же это только надстройка интерфейса, но

И проще уже только Midjourney.

[11]

Кстати Stable diffusion легко настраивается под Midjorney и даже обгоняет её по качеству изображений.

(как настроить: <https://www.youtube.com/watch?v=P-JIcWqPdBQ>)

[12]

Теперь перейдем к интерфейсу от пользователя «AUTOMATIC1111» (<https://github.com/AUTOMATIC1111/stable-diffusion-webui>). Он считается самым удобным и многофункциональным. И скажем так считается класикой, и успел обрасти модификациями и улучшениями.

Да адрес веб странички придется вставлять самому копируя ip адрес из консоли (когда я попытался настроить автоматическое открытие окна, мне пришлось переустанавливать браузер), а также придется почитать пояснения к настройкам (а лучше посмотреть видео ютубера на которого я ссылаюсь почти всю статью). Но если вы хотите не сгенерировать пару картинок, а начать делать действительно качественные изображения, то лучше начать именно с этого интерфейса. Плюс ко всему большинство туториалов именно по ней. Очереди кстати тоже можно формировать, если открыть несколько окон.

Консоль опять таки никуда не делась:

[13]

Процесс генерации нескольких изображений выглядит как-то так:

[14]

Модель как я уже говорил генерирует картинки на основании текстовых токенов. Но у неё также есть и множество других режимов. Часть из которых я уже попробовал:

- txt2img, о котором я рассказывал выше. Есть позитивные промпты, негативные и куча настроек для изображения на выходе

[15]

- img2img. Это тоже преобразование текста в картинку, но помимо текстовых промптов, на вход нейросети передается изображение, на основе которого или подражая которому генерируется итоговое изображение. Вот кстати пример генерации img2img (но размер изображениявсего 384 на 384 пикселя и появляются артефакты, об этом позже):

Этот режим сам подразделяется на несколько: сам img2img, sketch, inpaint, inpaint sketch и другие.

Sketch позволяет дорисовать некоторые детали к изображениям или задать им цветовую гамму.

[16]

Вот например самурай с примера выше дорисовывался именно в режиме Sketch (картинка посередине).

Инструмент попроще в освоении но не менее мощный – inpaint. Он позволяет закрашивать часть изображения для того чтобы изменить только эту область или наоборот все кроме выделенной области. И если вы хорошо разбираетесь в тонкостях настройки stable diffusion, то можно применять несколько стилей подряд. Я же пока научился только закрашивать части изображений втом числе и сгенерированных нейросетью чтобы их переделать.

К сожалению у меня не осталось скриншотов с моих попыток работы с inpaint, а по тому я просто приведу пример составленный за пару минут.

[17]

Кстати хочется упомянуть негативный промпт подставляющийся в режим txt2img по умолчанию:

[18]

Он очень детально описывает чего на изображении быть не должно, можно взять на заметку. И да, лучше не стоит подставлять его как позитивный, правда не надо.

[19]

При генерации изображений я в основном использовал в основном модель deliberate2 которую уже упоминал выше. На данный момент многие признают ее лучшей для генерации людей, так как она генерирует наиболее анатомически реалистичные тела и лица людей. Но эта модель подходит также практически для всех запросов, как говорит её автор, так как её дообучение велось распределённо на изображениях различных тематик.

Но надо отдать должное что для генерации таких хорроших изображений нужно: во-первых иметь хороший компьютер, во-вторых уметь использовать стили, ну и в третьих уметь составлять промпты.

Помимо этой модели существует достаточно много других и очень крутых, но по конкретно этой модели, можно есть очень неплохие гайды по настройке.

[20]

Также хочется рассказать про вещь которую я опробую в будущем, это Стили, которые можно добавить к модели дополнительно. То есть при генерации изображения берется модель а к ней дополнительно подставляется стиль.

[21]

И стилей может быть прикручено великое множество, главное, чтобы компьютер мог это обработать. Хотя с ними лучше не перебарщивать в среднем используют один-три стиля.

[22]

Кстати, для людей, играющих в игры тоже много чего найдется. Стиль похожий на gta5

[23]

Или, например недавно нашел стиль похожий на вселенную darkest dungeon.

[24]

Отдельно хочу сказать о теме, которую я упоминал только в скользь, а именно – искусство промпта.

Именно искусство, так ка по-хорошему для качественного промпта, необходимо знать значение и силу токенов при использовании конкретной модели.

Как сгенерировать правильную картинку, так и правильно наложить свет, использовать нужный стиль. Под это все нужно знать соответствующие токены и уметь их применить в правильном порядке и задать им нужный вес в запросе.

Значение многих слов переплетаются, а многие токены для stable diffusion имеют немножко не привычный на первый взгляд, но на самом деле очень логичный смысл.

Гораздо лучше меня рассказал ютубер «Христ», и если вы учитесь работать с промптами, то первое что вам нужно сделать это посмотреть его бесплатные ролики, вот так. Вот сылка на него.

[25]

Примеры готовых изображений с промптами и настройками можно посмотреть на сайте <https://civitai.com>

Это поможет с выбором модели и стилей и вам будет легче понять, как же правильно составить этот промпт

Теперь по поводу разрешения изображений.

[26]

Deliberate Обучена на наборах 512 на 512 пикселей, по этому лучше не генерировать изображения больше чем 840x840,иначе модель начинает гененрировать двойные лица у людей и прочие искажения [картинка, возможно из видео],и чем выше разрешение генерируемого изображения тем больше дефектов на ней будет (Это, кстати, уже не мой опыт, железа не хватает).

Хотя конечно, чем меньше изображение, тем быстрее оно генерируется, поэтому настройки лучше всего проверять на настроенном железе и в разрешении 512 на 512 пикселей ну или хотя бы 448 на 448.

Ну или самый минимум 256x256, но на таких маленьких картинках уже начинают появляться артефакты.

Затем уже изображение можно прогнать несколько раз с различными настройками и стилями (Styles),

подробнейший гайд на эту тему в ролике:

https://www.youtube.com/watch?v=6nthXso6iOE

[27]

Что ж, теперь что касается железа и требований. Так как программа запускается на локальной машине и обрабатывает нейросетевую модель, то в компьютере должна стоять достаточно мощная видеокарта, что очевидно.

Да, кстати, работает SD на любой ОС: Windows, macOS, Linux.

В сети, опробовавшие Stable Diffusion люди говорят, что в целом минимальные требования для нормальной работы это

- 6 ГБ оперативной памяти

- Видеокарта Nvidia с по крайней мере 4 ГБ видеопамяти

- 12 ГБ или более места для установки. В идеале SSD.

А уже для комфортной работы

- 16 ГБ оперативной памяти

- Видеокарта Nvidia с объемом видеопамяти не менее 10 ГБ или выше, предела нет.

У меня видеокарта от AMD.

Radeon 5700XT на 10 гигабайт видеопамяти, очень хорошая, ОДНАКО.. Так как версия PyTorch используемая в Stable Diffusion работает только с видеокартами Nvidia, мне пришлось использовать для вычислений процессор. При этом вычисления происходили гораздо медленнее порядка 30 минут на 4 изображения по одному запросу.

А новую видеокарту покупать ради генерации картинок я пока не могу себе позволить :)

Собственно да, чем хуже у вас железо, тем менее качественными получатся изображения.

Ну а теперь про полезные сервисы и сайты.

[28]

Недавно нашел также, бесплатный сайт, на котором без танцев с бубном можно попробовать сгенерировать изображения, однако не позволяет сделать картинки высокого качества и имеет минимальное число настроек

[29]

И оффициальная демо версия от самой stable diffusion.

Но надо сказать что оба эти сайта просто используют обученную классическую модель stable diffusion, и помимо позитивного и негативного промпта в неё нельзя ничего подставить. Не выбрать тензоры под конкретные задачи, не подставить стили, не поработать с inpaint. А в stablediffusionweb даже негативных промптов нет.

Более-менее нормальное изображение собаки на обоих сайтах удалось получить раза с третьего.

Собственно, оба сайта по тому и называются demo…

[30]

Кстати, русский они не понимают

Плюс ко всему, как оказалось многие из публичных нейросетей зацензурированы от вещей "Not safe for work", что негативно влияет на результат или нейросеть вообще выдаст ошибку генерации, если посчитает его "не корректным" или при подборе изображений для сравнения могут попасться «некорректные».

[31]

Например, недавно нейросеть отказывалась генерировать запрос «pizza banana kiwi», а сейчас вот исправили. Даже запрещенную пиццу с ананасами генерирует.

Кстати да, есть же оффициальный реддит по stable diffusion <https://www.reddit.com/r/StableDiffusion/>

Там можно найти дообученыые Lora (стили) или прикольные сгенерированные изображения.

Например: Парень превратил свою собаку в енота, «Boris Johnson sensation».

Также там есть новости о последних обновлениях модели.

Но лучшие модели, стили и готовые промпты лучше искать на сайте <https://civitai.com>, который я упоминал выше.

Там будут и топы, в которых можно будет подглядеть что-то интересное, и будет чему поучиться.

Всё. Спасибо за внимание.