Генерація покемонів у вигляді анімаційних персонажів для довільних креативних затсосувань

Планування проекту:

Необхідні ресурси:

- Дані. Для генерації варіантів покемонів знадобиться набір зображень з відповідними підписами зображень.
- Обчислювальні ресурси:

GPU A100: 40GB оперативної пам’яті графічного процесора;

83.5GB оперативної пам’яті системи;

113 GB пам’яті на диску.

Очікувані результати та кінцева мета

Мета: Розробити генеративну модель для створення унікальних покемонів у вигляді анімаційних персонажів з можливістю персоналізації для довільних креативних застосувань.

Результати: розроблена модель, яка здатна генерувати варіанти покемонів в специфічному анімаційному стилі, на основі введеного короткого опису.

Кроки виконання

1. Збір даних. Зібрати дані анімаційних покемонів та відповідними підписами. За потреби – підготовка даних.
2. Розробка моделі. Вибрати відповідні архітектуру генеративної моделі відповідно до поставленої задачі та наявних ресурсів.При потребі налаштувати та адаптувати задану модель до поставленої задачі.
3. Тренування моделі. Оптимізація моделі для ефективнішої роботи.
4. Оцінка якості генеративної моделі. Покращення моделі.
5. Інтеграція моделі.
6. Тестування моделі.

Збір на підготовка даних.

- Для даної задачі я обрала наступний open-source dataset з hugging face – *svjack/pokemon-blip-captions-en-zh (https://huggingface.co/datasets/svjack/pokemon-blip-captions-en-zh).*
- Датасет містить 833 семпли, настпуного вигляду: зображення, опис англійською мовою, та опис китайською мовою. Для наступних кроків видаляємо опис китайською мовою, оскільки він нам не потрібний.
- В додатковій обробці зображень нема потреби.

image image · *width (px)* 272 1.28k	en_text string · *lengths* 24 97	zh_text string · *lengths* 3 65
	a drawing of a green pokemon with red eyes	红眼睛的绿色小精灵的图画
	a green and yellow toy with a red nose	黄绿相间的红鼻子玩具
	a red and white ball with an angry look on its face	一个红白相间的球，脸上带着愤怒的表情

Вибір моделі GenAI.

- Мною було прийняте рішення обрати модель **Stable Diffusion XL** через ряд наступних причин:

1. Така модель забезпечує якісні, реалістичні та чікіз. ображення. Вона має здатність створювати зображення з високою роздільною здатністю.
2. Головна характеристика дифузійних моделей – «процес зворотньої дифузії» - це дозволяє моделями вивчати складні властивості зображень, забезпечуючи **стабільне** навчання
3. Така модель є дуже гнучкою: підтримує Low-Rank Adaptation.
4. Підтримує векторні запити (на відмінну від GANs, Autoencoders).
5. Порівняно адекватні ресурси для запуску. (GPU середенього рівня достатньо)

В даному проекті модель **Stable Diffusion XL** додатково налаштовувалась (fine tune), оскільки оригінальна модель створювала зображення у відмінному від оригінального стилі.

Приклад:

Prompt: `green pokemon with flowers in the hands`

Оригінальна модель

Модель з додатковим налаштуванням (fine tune)

Деталі по fine tuning.

- Для відстеження прогресу було використано платформу Weights and Biases.
- Було використано LoRA для fine tuning – таким чином я не тренувала всі параметри моделі (оскільки модель дуже велика) , а лише адаптувала її під нова дані, додавши кілька додаткових адаптаційних параметрів, які вже навчались.

Важливі підібраніч гіперпараметри:

1. train_batch_size=1
2. max_train_steps=5000 & num_train_epochs=7
3. gradient_accumulation_steps=4
4. learning_rate=1e-05 (тестувались ще: 1e-04, 1e-06)
5. mixed_precision="fp16" & allow_tf32 (для прискорення процесу тренування та зменшення використання пам’яті)
6. hub_model_id="sdxl-base-1.0-test-lora" - доступна на моєму репозиторії в HF.
7. gradient_accumulation_steps=4

Прогрес роботи, який відображався в wandb

validation


0: A red pokemon with ice cream.


1: A red pokemon with ice cream.


2: A red pokemon with ice cream.


0: A red pokemon with ice cream.

Step 209

validation


1: A red pokemon with ice cream.


2: A red pokemon with ice cream.

Step 1045

validation


0: A red pokemon with ice cream.


1: A red pokemon with ice cream.


2: A red pokemon with ice cream.

Step 2926

Висновок:

Можна помітити, що чим далі модель донавчалсь тим більш змінювався стиль її зображень. Вони набували більшої схожості до вихідного стилю зображень.

Результати роботи моделі. Модель була інтегрована за допомогою streamlit

[INFO] Завантаження моделі...

[INFO] Модель завантажено.

Напишіть промт-опис для генерації зображенняпокемона

Введіть промпт:

a blue pokemon

Передати дані до моделі

Дані передано до моделі:

a blue pokemon

[INFO] Початок генерації зображення...

[INFO] Генерація завершена.

Зображення, згенероване LoRA

[INFO] Завантаження моделі...

[INFO] Модель завантажено.

Напишіть промт-опис для генерації зображенняпокемона

Введіть промпт:

white pokemon with a cup

Передати дані до моделі

Дані передано до моделі:

white pokemon with a cup

[INFO] Початок генерації зображення...

[INFO] Генерація завершена.

Зображення, згенероване LoRA

[INFO] Завантаження моделі...

[INFO] Модель завантажено.

Напишіть промт-опис для генерації зображенняпокемона

Введіть промпт:

black pokemon on the grass

Передати дані до моделі

Дані передано до моделі:

black pokemon on the grass

[INFO] Початок генерації зображення...

[INFO] Генерація завершена.

Зображення, згенероване LoRA

Вдалі результати моделі

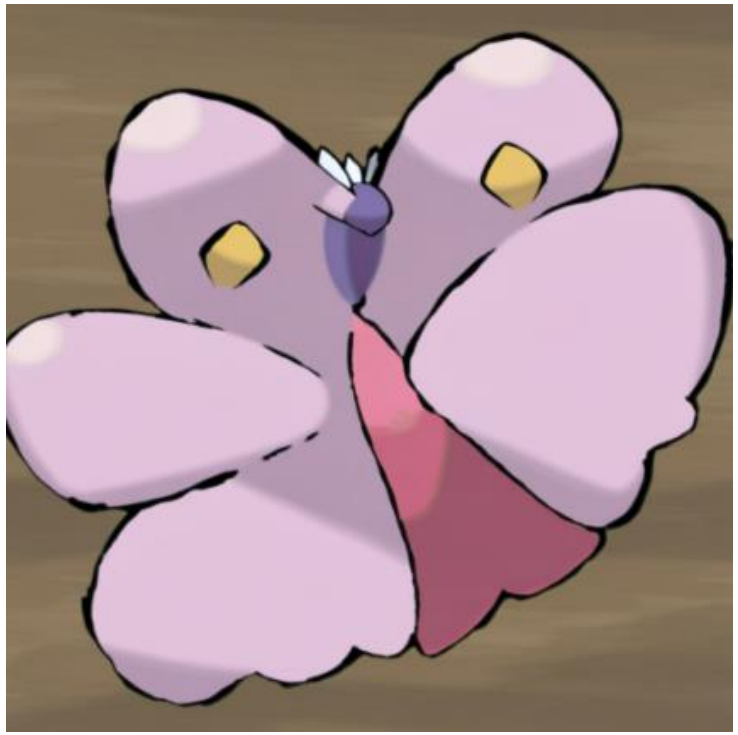
'fire-breathing dragon pokemon in cave'

'brown bear-like pokemon in forest'

'red fire pokemon in desert'

Невдалі результати моделі

'pink butterfly pokemon fluttering'



'mysterious psychic pokemon with glowing eyes'

'snow pokemon in frozen forest'

Оцінка роботи моделі за допомогою візуальної перевірки

Для даної метрики було опитано 5 людей, і показано 50 зображень та промпти до цих зображень.

Також учасники опитування проглянули дані на яких модель навчалась, для того щоб зрозуміти вихідний стиль зображень, та які зображення ми очікуємо.

Запропоновано 4 критерії:

1. Чіткість зображення. 0 – зовсім не чіткі, 5 – дуже чіткі.
2. Відповідність текстовому запиту. 0 – зовсім не відповідає, 5 – відповідає повністю.
3. Реалістичність зображень\схожість на дані що навчалась модель.

   0 – зовсім нереалістичні\несхожі по стилю, 5 – абсолютно точно попали у стиль відносно тренувальних даних.
4. Наявність артефактів

   0 – занадто багато шуму, 5 – чисте зображення

Було отримано наступні результати

Учасники	Чіткість	Відповідність промпту	Відповідність тренувальним даним	Наявність артефактів
1	4	4	5	3
2	4	3	5	2
3	4	3	4	3
4	3	4	5	3
5	3	3	5	3
Середнє значення	**3.6**	**3.4**	**4.8**	**2.8**

Висновки: зображення дійсно відповідають стилю тренувальних даних, вони є достатньо чіткими, проте деякі зображення все ж просідають. На рахунок промтів – неоднозначно, оскільки частина зображень дійсно відповідають тексту, і всі деталі чіткі та зрозумілі, проте є значна частина зображень, яка не відповідає промптам, а особливо додаткові елементи до покемонів (типу порозиво, машина, вогонь..).
Найбільшою проблемою є наявність шуму в зображеннях.

Метрики.

CLIP	0.31
FID	195
INCEPTION SCORE	Mean = 1.5633, Std = 0.1959

По метриках можна зробити наступні висновки:

1. **CLIP** - оцінює схожість між згенерованими зображеннями і текстовими описами. Значення 0.3 є низьким, що вказує на те, що згенеровані зображення не дуже добре відповідають наданим текстовим запитам. Це може означати, що модель не точно інтерпретує або відображає суть описів
2. **FID** – оцінює відстань між розподілами реальних і згенерованих зображень у просторі ознак Inception. Зазвичай значення FID нижче 50 вважається гарним результатом для генеративних моделей, тому значення 195 є досить високим, що свідчить про велику різницю між реальними та згенерованими зображеннями
3. **INCEPTION SCORE** – вимірює якість і різноманітність згенерованих зображень. Значення 1.56 вважається доволі низьким для Inception Score, оскільки хороші значення починаються від 2.0 і вище. Це свідчить про те, що зображення, ймовірно, є нечіткими або невиразними

Висновки. Ідеї для покращення.

- Можна помітити, що модель має правильний вектор – і вона почала генерувати зображення покемонів, які за стилем дійсно нагадують вихідний анімаційний стиль, який модель без fine tune не зроить. Проте вона має певні недоліки, а саме:

1. Зображення часто бувають неоднорідні. Предмети часто хаотично розміщенні на картинці, (наприклад, покемон в одному місці, морозиво\квіти десь літають і тд.).

2. Не завжди зрозумілі і інтерпретовані предмети, які додані до покемонів: наприклад в такому промпті «покемон з морозивом» - покемон чітко зображений, а об’єкт морозива не завжди зрозумілий.

3. Обличчя покемонів інколи не мають очей, рота.

Висновки. Ідеї для покращення.

Наступні кроки для покращення:

1. Однозначно потрібно розширити вибірку тренувальний даних – оскільки їх 833 (це менше 1к) і цього недостатньо для якісного fine tuning.
2. Більша кількість епох – 5к кроків не достатньо для якісного fine tuning для даної задачі. Для цього потрібно збільшувати обчислювальні ресурси, оскільки в моєму випадку ці 5к кроків зайняли близько 7годин fine tuning, для того щоб можна було більше робити епох, потрібно щоб це було оптимізованіше.

3. Додати квантизацію в модель, частково це допоможе вирішити проблему в п2. Проте потрібно проводити тести, для того щоб зберегти баланс між оптимізацією та точністю.

Висновки. Ідеї для покращення.

- Можна помітити, що модель має правильний вектор – і вона почала генерувати зображення покемонів, які за стилем дійсно нагадують вихідний анімаційний стиль, який модель без fine tune не зроить. Проте вона має певні недоліки, а саме:

1. Зображення часто бувають неоднорідні. Предмети часто хаотично розміщенні на картинці, (наприклад, покемон в одному місці, морозиво\квіти десь літають і тд.).

2. Не завжди зрозумілі і інтерпретовані предмети, які додані до покемонів: наприклад в такому промпті «покемон з морозивом» - покемон чітко зображений, а об’єкт морозива не завжди зрозумілий.

3. Обличчя покемонів інколи не мають очей, рота.

Тобто можна зробити висновок, що модель рухається в правильному напрямку, проте її ще потрібно вдосконалювати.