Carrera de especialización en inteligencia artificial - FIUBA

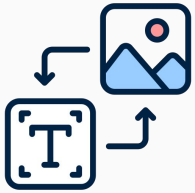
Generación de imágenes a partir de texto

Vision Transformers

Alumnos:

- Fabricio Denardi
- Valentín Pertierra
- Leandro Saraco
- Sofía Speri

Introducción



- **Objetivo:** Generar imágenes a partir de texto.
- Uso de la arquitectura **Stable Diffusion** (basada en Latent Diffusion Models).
- **Fine-tuning** para personalizar el modelo en un estilo específico.
- Evaluación del modelo mediante métricas de **CLIP** y **BLIP**.

Primeros pasos y pruebas experimentales

MODELO	CARACTERÍSTICAS / VENTAJAS	PARÁMETROS	ARQ.	DATASET PREENTRENADO	DIFICULTADES o INCONVENIENTES PRESENTADOS
DiffusionCLIP	- Control fino sobre atributos de imágenes.	110M (solo CLIP) + UNet	UNet + CLIP	CelebA-HQ (Human Face).	- Dificultad en implementación. - Documentación limitada y sin ejemplos claros.
Versatile Diffusion	- Multimodalidad - Flexibilidad y capacidad de generalización.	1.3B	Transformer-based	LAION-400M	- Problemas de compatibilidad de librerías y dependencias.
Stable Diffusion 2-Base	- Síntesis texto a imágenes, procesamiento eficiente	865M	Latent U-Net diffusion	LAION-5B	- Incidencias de recursos y función Trainer de Hugging Faces para el fine tuning.
Stable Diffusion 3	- Más alta fidelidad en generación de imágenes - Mejora en manejo de prompts	1.2B	Latent U-Net diffusion	LAION-5B	- Incidencias de recursos y función Trainer de Hugging Faces para el fine tuning. - La inferencia resultó en out-of-memory en las instancias gratuitas de Google Colab.

Desarrollo: Entorno

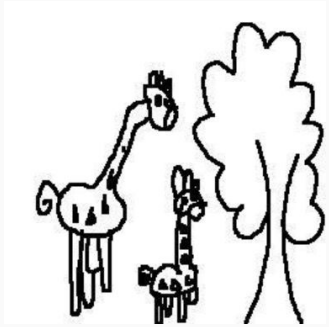
- Ejecutado en Google Colab
- Uso de librerías de Hugging Faces: accelerate, diffusers.
- Ventajas:
  - **Configuración Automática**: Detectan y optimizan el uso de GPU y memoria sin requerir ajustes manuales.
  - **Distribución Eficiente de Datos**: Facilitan la gestión de batches, mejorando el rendimiento.
  - **Entrenamiento Optimizado**: Reducen tiempos y costos al implementar optimizaciones automáticas.

Desarrollo: Dataset

Sketch-scene[1]

CARACTERISTICAS

- Imágenes de bocetos de escenas a mano alzada.
- ~10k bocetos a mano alzada, en formato vectorial.
- Contiene:
    - Imagenes 256x256 px (formato PIL JPEG)
    - Caption: descripción textual
- No está dividido en subsets de train-test.



two girafee"s eating the tree leaves

PUNTOS IMPORTANTES:

- Conjunto de datos liviano → menor necesidad de recursos.
- Imágenes menos tradicionales → no comprendidas en dataset pre entrenado.

man sitting on the horse

[1]https://huggingface.co/datasets/zoheb/sketch-scene

Desarrollo: Modelo

Stable Diffusion v2 base [1]

- Desarrollado por Robin Rombach, Patrick Esser en 2022
- Modelo de Latent Diffusion.
- Entrenado con LAION-5B y subsets
- **Arquitectura:** usa un autoencoder con un factor de reducción de 8, combinado con un UNet de 865 M de parámetros y un text encoder OpenCLIP ViT-H/14 como base para el modelo de difusión.
- Resolución de salida: 768x768 px.

Imágenes generadas para prompt: *A sketch of a zebra and a monkey in front of a mansion"*

[1] https://huggingface.co/stabilityai/stable-diffusion-2-base

Desarrollo: Fine tuning

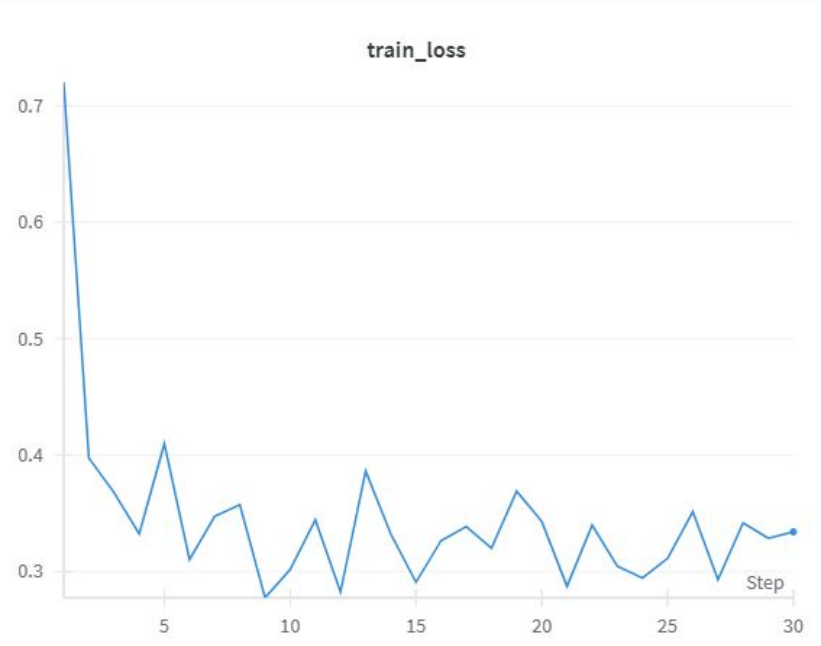
- Se busca obtener imágenes con estilo sketch a partir de dataset elegido.
- Entrenamiento a través de *accelerate*.
- N = 30 epochs
- Optimizaciones:
  - **use_8bit_adam**
  - **Gradient_checkpointing**
  - **mixed_precision="fp16"**
  - **Use_ema**
  - **max_grad_norm=1**
- Data augmentation y transformaciones:
  - **Resolucion 256x256**
  - **Center_crop**
  - **random_flip**
- Inferencia con modelo fine-tuned:

Imágenes generadas con mismo prompt: *A sketch of a zebra and a monkey in front of a mansion"*

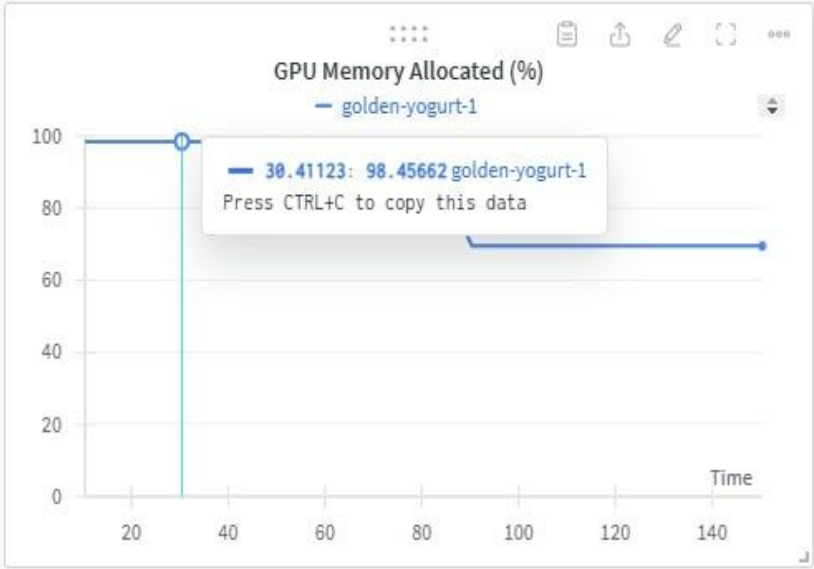
Fine tuning - métricas durante el entrenamiento

Métrica de entrenamiento a través de Wandb

Loss function: denoising loss



Uso de memoria durante entrenamiento



→ Mejora a futuro: Uso de loss function alternativa o combinada mediante CLIP

Desarrollo: Evaluación

- A nivel visual se puede ver el cambio de estilo:

Antes de realizar fine tuning:






Después de realizar fine tuning:






En las próximas diapositivas se verá cómo obtener métricas objetivas

Desarrollo: Evaluación

- Uso de modelo CLIP[1] → generación de embeddings de texto e imágenes en un espacio compartido.
- Medición de SIMILITUD COSENO entre prompt e imágenes generadas.
- METRICA:
  - 5 prompts representativos (externos al dataset de entrenamiento).
  - 6 imágenes generadas por cada prompt.
  - Cálculo de similitud por prompt y global.
- Dos experimentos realizados:
  - Con preposición de estilo.
  - Sin preposición en generador.
- Confirmación adicional con BLIP[2] → generación de texto a partir de imágenes y detección de estilo.

[1] https://huggingface.co/openai/clip-vit-base-patch32
[2] https://huggingface.co/Salesforce/blip-image-captioning-base

Resultados con CLIP

Prompts:

a) "a boy is practicing martial arts"
b) "a statue of a man on horse
c) "a boy is riding bike on the road"
d) "giraffe standing in a forest”
e) “zebra running”

Resultados - Experimento 1:

a)

b)

Resultados - Experimento 2:

a)

b)

Resultados con CLIP

- Similitud coseno global con CLIP:

Experimento	Similitud coseno promedio
1. Utilizando pre-prompt de estilo en generador y CLIP	36.48%
2. Utilizando pre-prompt de estilo sólo en CLIP	33.58%

Resultados con BLIP

- Uso de BLIP con experimento 1 para validar texto a partir de imágenes generadas con el modelo fine-tuned.

PROMPT		***TEXTO GENERADO CON BLIP***
A black and white sketch of a boy practicing martial arts		A drawing of two people in karate stance
A black and white sketch of a *statue* of a man on horse		A drawing of a horse and rider

Conclusiones

- Entrenar un modelo de generación de imágenes a texto es una tarea que consume muchos recursos computacionales, especialmente VRAM.
- A través de optimizaciones como el uso de *half precisión* (fp16) y Adam en 8 bits, fue posible reducir drásticamente el consumo de memoria.
- Se pueden mejorar o profundizar sobre CLIP como función de pérdida alternativa o combinada a denoising loss.
- Se logró un cambio de estilo pre y pos fine tuning, perceptible a simple vista.
- A través del uso de modelos como CLIP fue posible obtener una métrica objetiva de qué tan bien se desempeña el modelo en la tarea requerida.
- Además, el uso de BLIP ayuda a confirmar que es posible captar el cambio de estilo a través de un modelo de IA.

Muchas gracias

¿Preguntas?