ML Open Source - de modelos pre-entrenados a producción

Usa modelos estado del arte en producción

1

Modelos

¿Qué existe ahí afuera?

El Hub de Hugging Face

Modelos

Accede a más de 350 mil modelos compartidos por la comunidad.

Datasets

Accede, comparte y colabora en más de 70 mil.

Spaces

Construye aplicaciones y demos de Machine Learning para mostrar como funcionan los modelos.

El Hub de Hugging Face

Modelos

Accede a más de 350 mil modelos compartidos por la comunidad.

99k-> 377k

Datasets

Accede, comparte y colabora en más de 70 mil.

16k->75k

Spaces

Construye aplicaciones y demos de Machine Learning para mostrar como funcionan los modelos.

19k->130k

El Hub de modelos

- **Modelos de diferentes modalidades** (visión por computadora, PLN, audio, por refuerzo)
- Diferentes bibliotecas (PyTorch, Keras, fastai, SpaCy, NeMo, etc.)
- 180+ lenguajes
- Tarjetas de modelos (model cards) para documentar

2

Inferencia

¿Cómo hacemos inferencia?

Modelos populares recientes (LLMs)

StarCoder

- Generación de código
- 15.5B de parámetros
- Licensia OpenRAIL
- 80+ lenguajes
- 1 trillion de tokens de entrenamiento

LLaMA

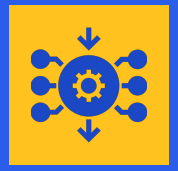
- Gran ecosistema
- 7B a 65B de parámetros
- No comercial*
- 1-1.4 trillion tokens*

Con Llama 2, 2T de tokens, comercial, y 70B de parámetros

Falcon

- Mejor modelo OS
- 7B a 40B de parámetros
- Apache 2.0
- Multilingual
- 1 trillion tokens

Retos



Tamaño de modelos

Los LLMs requieren mucha memoria, pueden no entrar en un solo dispositivo y requieren paralelismo y comunicación complejas.

Evaluación

Las evaluaciones existentes no encapsulan los casos reales (por ejemplo, multiples turnos).

Personalización

Las personas usuarias quieren modelos adaptados a sus propios datos y casos de usos manteniendo privacidad.

Optimización

Dado el tamaño del modelo, su latencia y Due to model size, latency y throughput son impactados, necesitando optimizaciones en los modelos.

Algunas cosas que podemos hacer

Cargando

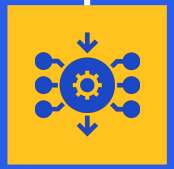
Cargar en modo 4-bit o 8-bits (bitsandbytes, accelerate)

Falcon 40B con 45GB (8-bit) o 27GB (4-bit) de RAM

Multi-GPU

Distribuir en varios GPUs (accelerate)

Usar device_map="auto" o cargar capas al CPU (lento)



Bibliotecas de Inferencia

Usar herramientas especializadas (text-generation-inference)

Usado en HF en producción

Text-generation-inference (TGI)

Paralelismo de tensores

Streaming de tokens

Métricas y monitoring

Cuantización

Optimizaciones

Seguridad

TGI soporta varios de los modelos más populares, como

Falcon | StarCoder and SantaCoder | LLaMA, Galactica and OPT | GPT-NeoX

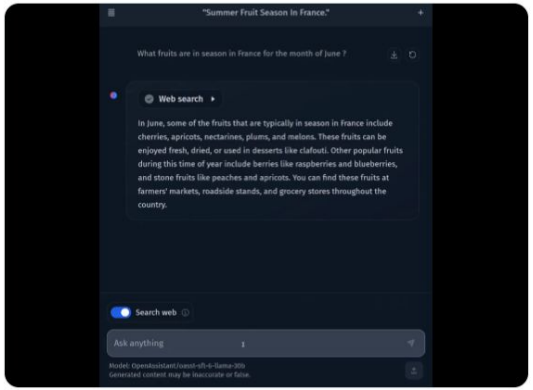
Algunos usuarios

HuggingChat

AK
@_akhaliq

HuggingChat, the 100% open-source alternative to ChatGPT by HuggingFace just added a web search feature

Link: huggingface.co/chat/
GitHub Repo: github.com/huggingface/ch...



11:14 PM · Jun 5, 2023 · **245K** Views

300 Retweets **13** Quotes **1,392** Likes **738** Bookmarks

OpenAssistant

Yannic Kilcher
@ykilcher

🔥EVERYONE🔥We're excited to announce the release of OpenAssistant.
The future of AI development depends heavily on high quality datasets and models being made publicly available, and that's exactly what this project does.
Watch the annoucement video: youtu.be/ddG2fM9i4Kk

7:00 PM · Apr 15, 2023 · **918.5K** Views

528 Retweets **123** Quotes **2,183** Likes **759** Bookmarks

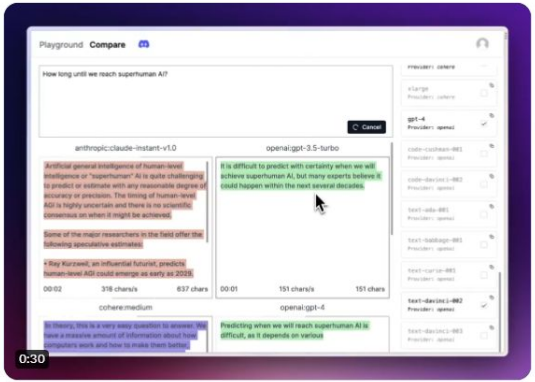
nat.dev

Nat Friedman
@natfriedman

The LLM playground that's hosted at nat.dev is now open source:

github.com/nat/openplaygr...

Enjoy!



5:44 AM · Apr 4, 2023 · **396.1K** Views

272 Retweets **27** Quotes **1,726** Likes **1,049** Bookmarks

3

Entrenamiento

¿Cómo ajustamos los modelos a nuestros propios casos de uso?

Entrenamiento

Pre-training

- $$$
- Muchos datos
- Expertise necesario

Fine-tuning

- $$
- Menos datos y compute

PEFT

(**P**arameter **E**fficient **F**ine-**T**uning)

- $
- Incluso menos compute

Puedes hacer fine-tuning de Falcon 7B o Whisper en Google Colab gratuito

Ejemplo: Whisper

Full-Tuning

Resulta en **OOM**

LoRA

- **1%** de parámetros entrenables, batch size **5x** más grande.
- Entrena un modelo de 6B con menos de 8GB the VRAM de GPU.
- Los modelos resultantes son más pequeños que 1% del tamaño original.

Ejemplo: Stable Diffusion

adapter “**perro**”

adapter “**juguete**”

adapter “**juguete**” + “**perro**”

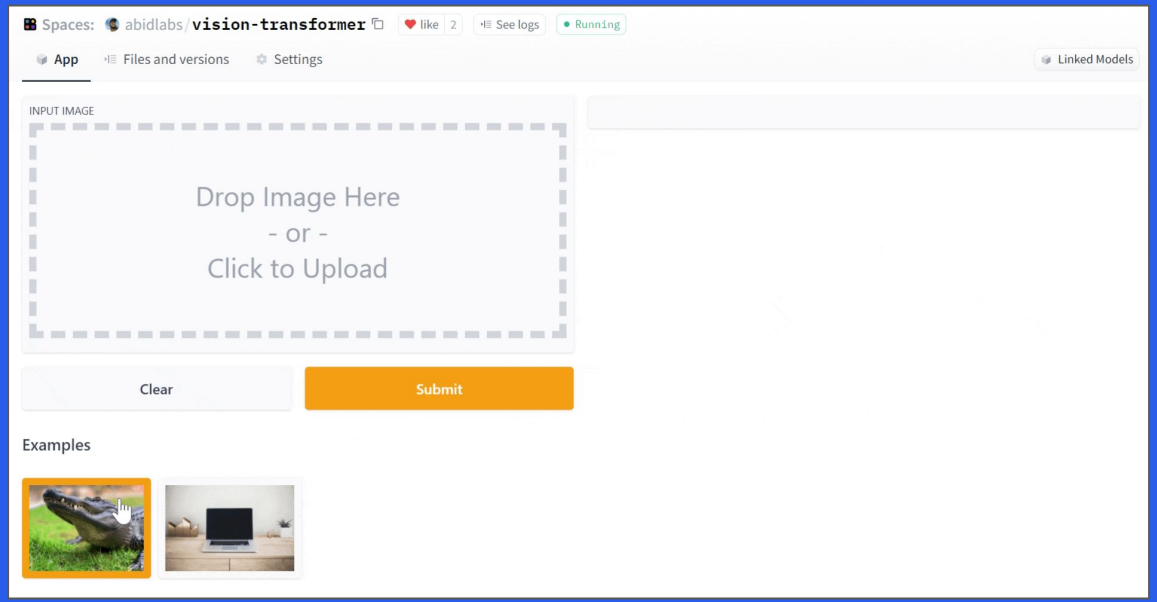
Ejemplo: Stable Diffusion

4

Construyendo Demos

¿Cómo construir y compartir apps de ML?

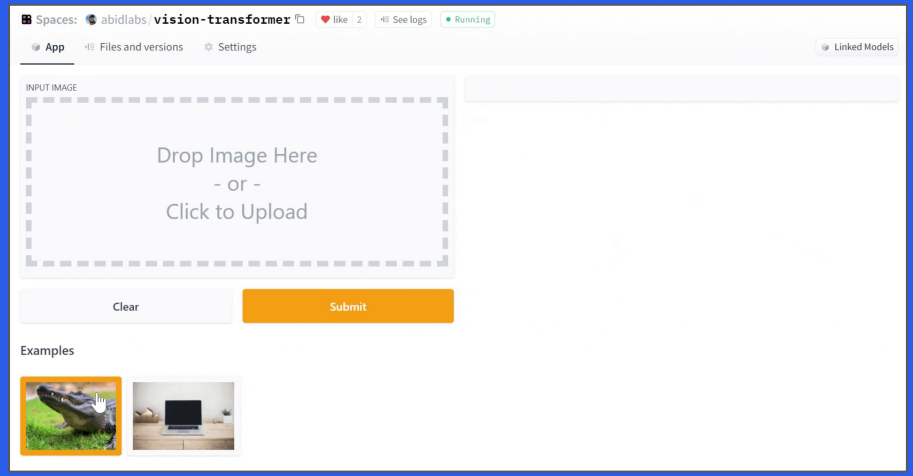
¿Por qué demos?

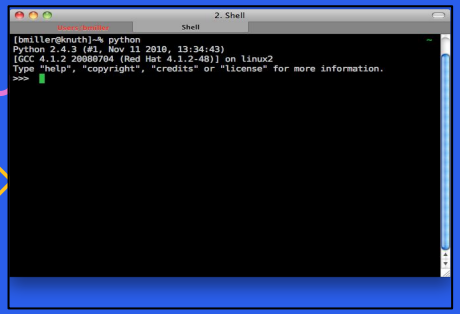


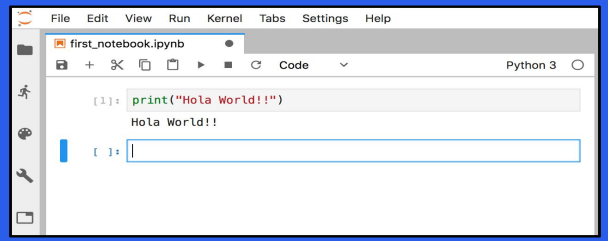
- **Presenta** fácilmente a una audiencia más amplia
- Aumenta **reproducibilidad** de investigación
- Personas usuarias diversas pueden **identificar** puntos de falla

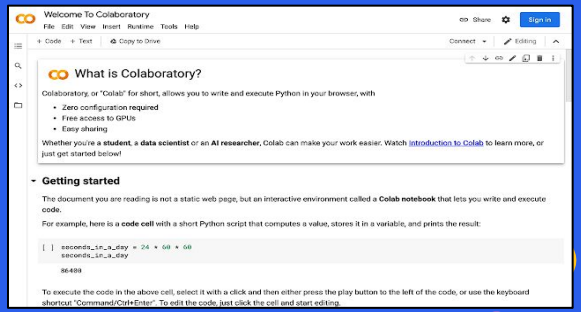
Gradio: uso típico

```
import gradio

app = gradio.Interface(
      classify_image,
      inputs="image",
      outputs="label")

app.launch()
```







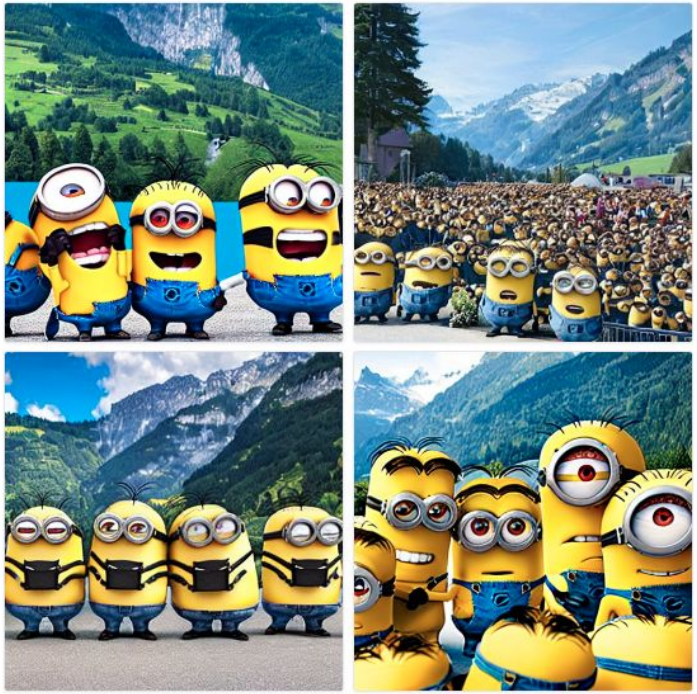


Stable Diffusion 1 Demo

Stable Diffusion is a state of the art text-to-image model that generates images from text.
For faster generation and API access you can try DreamStudio Beta

Picture of minions visiting Interlaken

Generate image



1,187 13.5K 153.8K

Punto de cambio en el uso de ML

Personas ingenieras de ML/software

cualquiera con un browser

Gracias

omar@huggingface.co
Omar Sanseviero
@osanseviero