

## ***NLP***

### Deploy y servicios

Msc. Rodrigo Cardenas Szigety  
rodrigo.cardenas.sz@gmail.com

Esp. Ing. Hernán Contigiani  
hernan4790@gmail.com

# Programa de la materia



**Clase 1:** Introducción a NLP, Vectorización de documentos.

**Clase 2:** Preprocesamiento de texto, librerías de NLP y Rule-Based Bots.

**Clase 3:** Word Embeddings, CBOW y SkipGRAM, representación de oraciones.

**Clase 4:** Redes recurrentes (RNN), problemas de secuencia y estimación de próxima palabra.

**Clase 5:** Redes LSTM, análisis de sentimientos.

**Clase 6:** Modelos Seq2Seq, traductores y bots conversacionales.

**Clase 7:** Celdas con Attention. Transformers, BERT & ELMo, fine tuning.

**Clase 8:** Cierre del curso, deploy y servicios, NLP hoy y futuro.

\*Unidades con desafíos a presentar al finalizar el curso.

\*Último desafío y cierre del contenido práctico del curso.



Conjunto de herramientas o actividades



Resuelven una necesidad

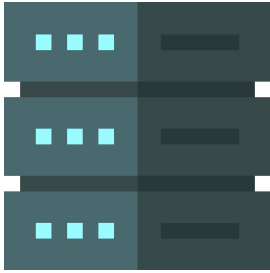


La industria del software se basa principalmente en ofrecer servicios

# ¿Cómo podemos ofrecer/consumir un servicio?



## IaaS



architect, build



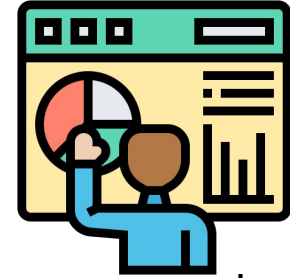
## PaaS



developer, deploy



## SaaS



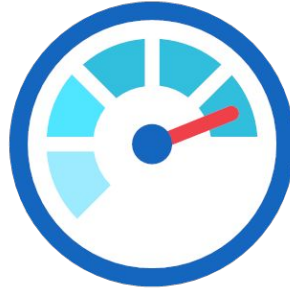
user, product



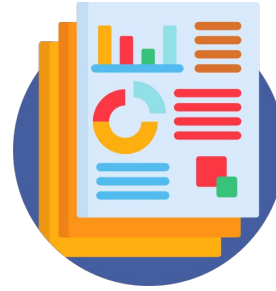
# ¿Cuáles son los servicios más ofrecidos?



Base de datos



Monitoreo



Reportes



Inteligencia  
Artificial

# Arquitectura de un servicio

[LINK](#)



## Aplicación monolítica 2003 / 2005



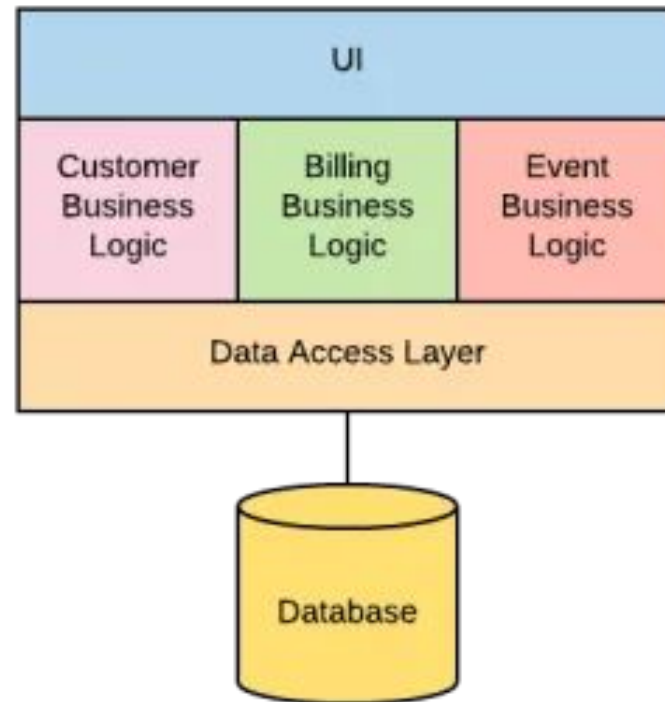
Simple y favorables en una primera etapa del proyecto



Cada cambio implica auditar todo el sistema



No es flexible, no puede adaptarse cambios tecnologías



# Arquitectura de un servicio

## Microservicios

[LINK](#)



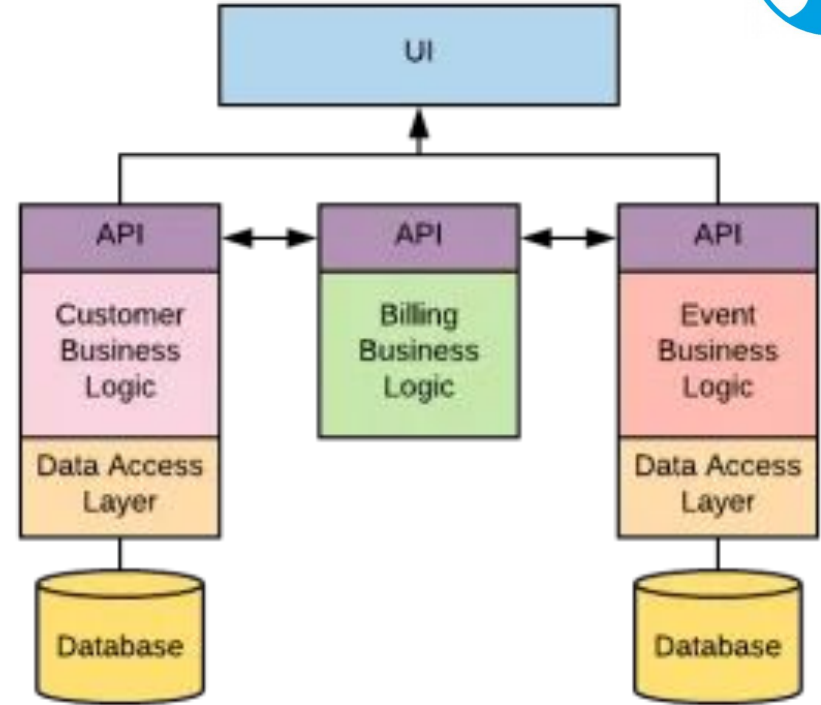
Responsabilidades separadas



Permite agregar o eliminar funcionalidades con riesgo acotado



Utilizar diferentes tecnologías



¿Qué es una API? ----->

# ¿Qué es una API?

[LINK](#)

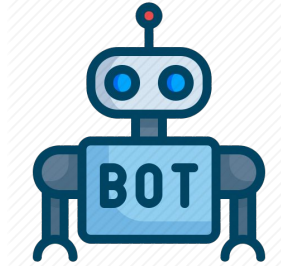


API → Interfaz de programación estándar

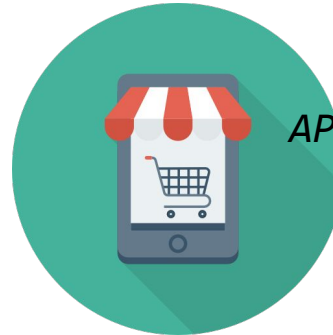


## API BANK

*APIs de entidades bancarias*



*APIs de automatización  
de software*



*APIs de entidades comerciales*



# ¿Cómo consumimos la REST API de nuestro servicio?

[LINK](#)



```
@app.route("/predict/<input_text>")  
def predict(input_text):
```

*La forma correcta sería que los datos no viajen en la URL, sino que se encapsulen en un HTTP POST y JSON*

# ¿Cómo desplegamos nuestro modelo en nuestro servicio?

[LINK](#)



TensorFlow

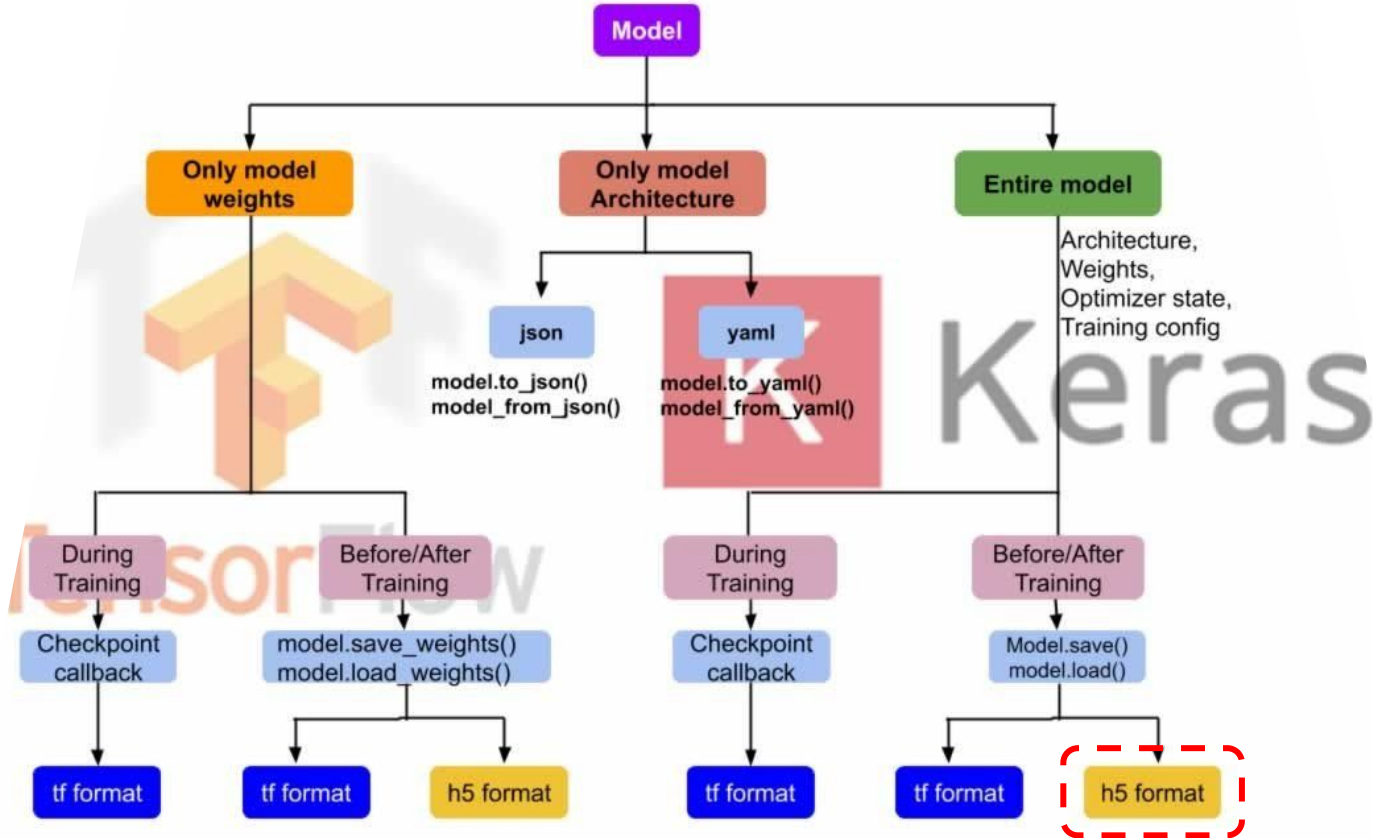


PyTorch



Caffe2

# Formas de exportar un modelo TF/Keras



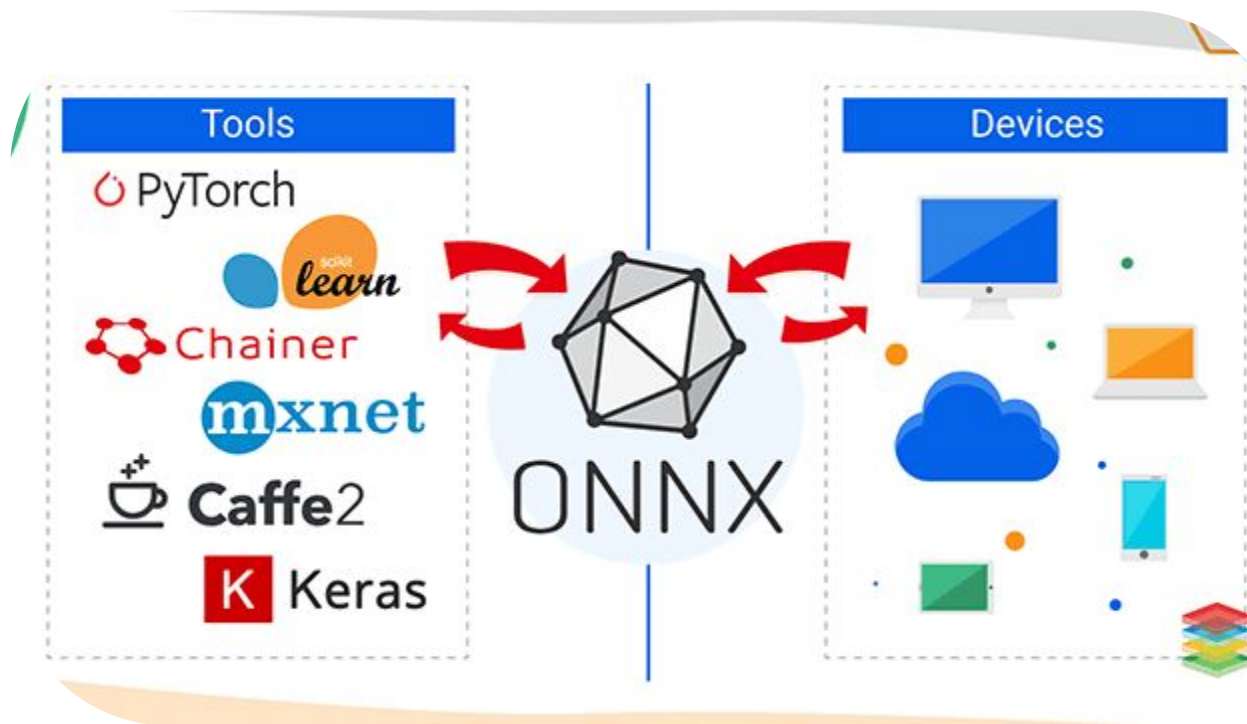


Link al Colab



*LINK*

# ¿Hay alguna forma de exportar modelos entre frameworks o dispositivos?



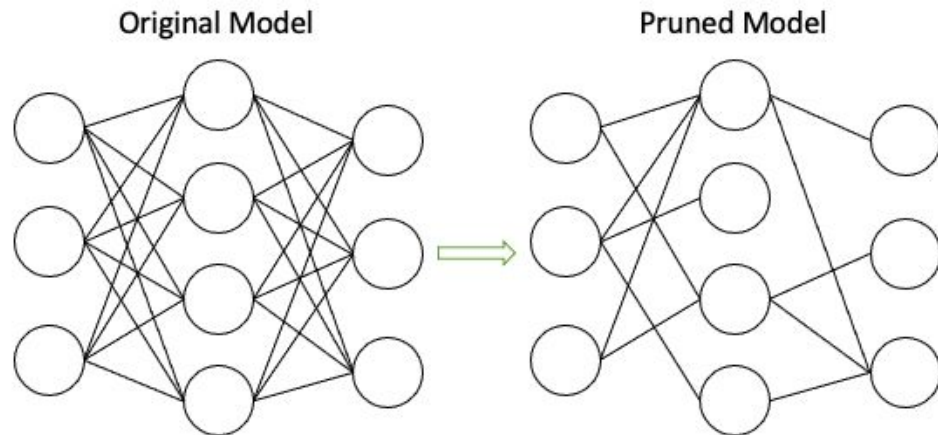
# ¿Puedo optimizar los modelos?

[LINK](#)

[LINK](#)

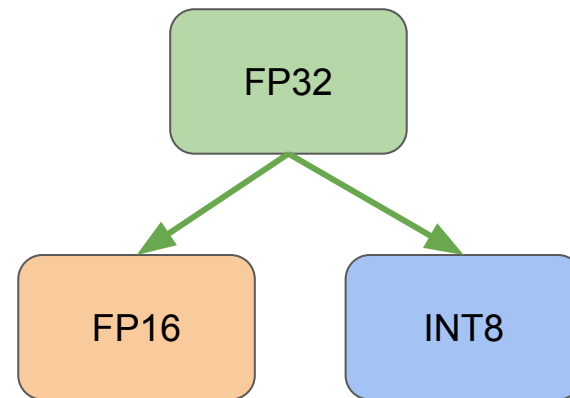


## Prune (podar)



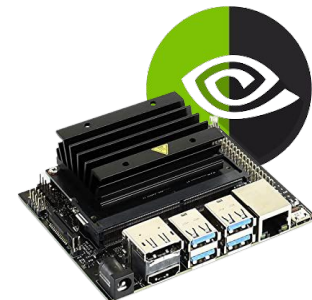
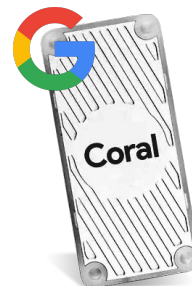
Eliminar los lazos con pesos muy bajos  
(bajo aporte al resultado final)  
Mejora el "size" y "speed" perdiendo  
muy poca precisión

## Quantization (cuantización)



Se reemplaza los pesos en float32  
por una representación reducida  
(float16) o int8. Se reduce "size"  
pero se puede perder precisión.

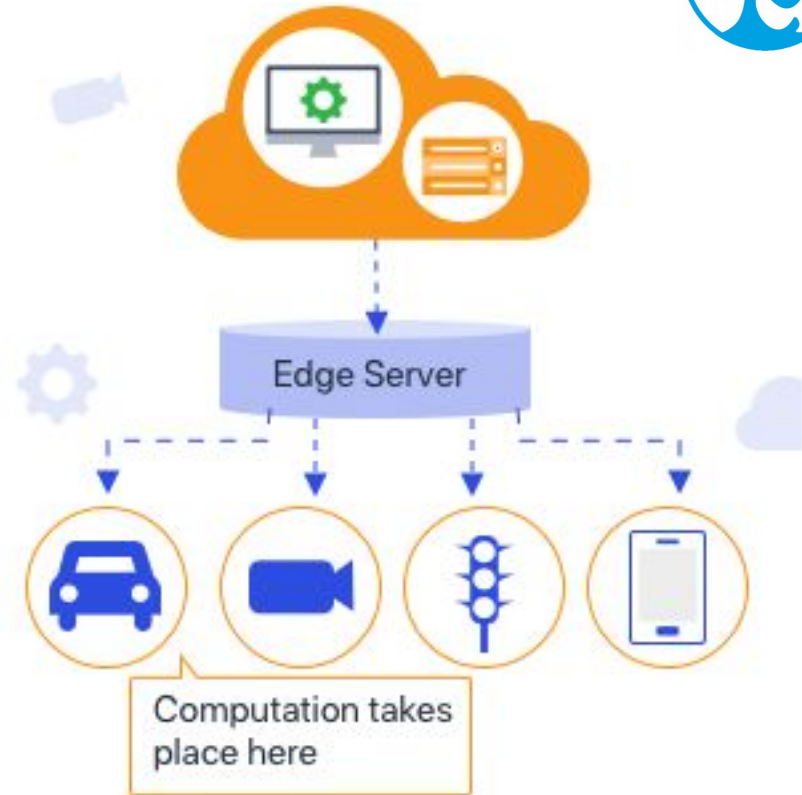
# ¿En qué plataforma puedo deployar el modelo?



TensorFlow Extended

PYTORCH


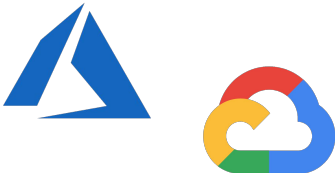
# Edge vs Cloud computing



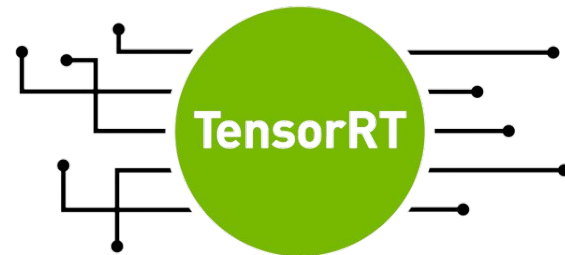
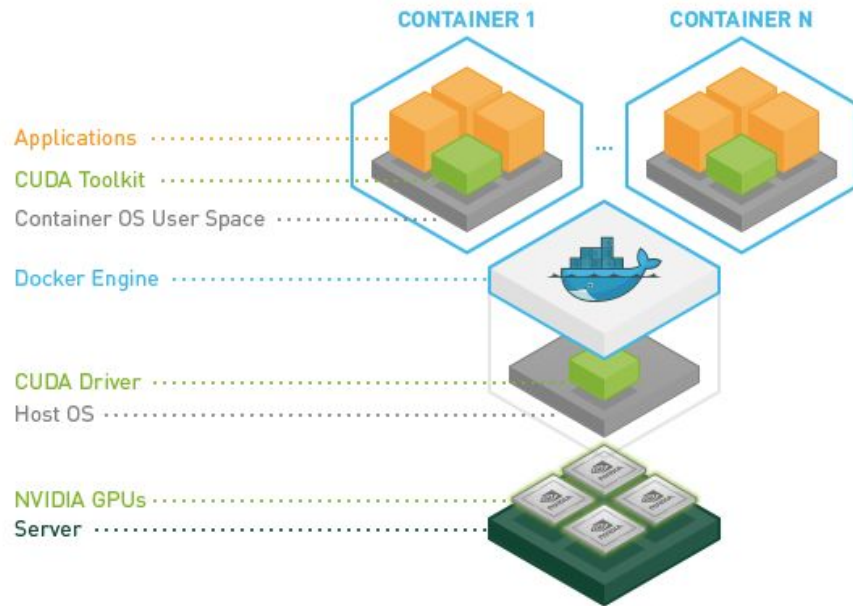


# Edge vs Cloud computing

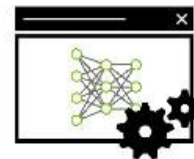


	Pros	Cons
<b>Edge</b> 	<ul style="list-style-type: none"><li>• Más control de tu aplicación.</li><li>• No requiere conexión a internet.</li><li>• Menor latencia.</li><li>• Más seguro</li></ul>	<ul style="list-style-type: none"><li>• Un dispositivo por cliente o solución.</li><li>• Responder a fallas o problemas con el hardware (reemplazo).</li></ul>
<b>Cloud</b> 	<ul style="list-style-type: none"><li>• Los recursos pueden ser compartidos entre aplicaciones.</li><li>• No hay que mantener una plataforma o hardware.</li><li>• No hay que reemplazar hardware dañado.</li></ul>	<ul style="list-style-type: none"><li>• Costos mensuales asociados a la infraestructura.</li><li>• Costos por tráfico de red (internet).</li><li>• Costo por uso de storage (disk).</li></ul>

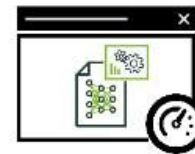
# ¿Puedo optimizar el modelo para la plataforma seleccionada (ej: NVIDIA)?



Trained  
Neural  
Network



TensorRT  
Optimizer



TensorRT  
Runtime  
Engine

# ¿Cómo se puede vender/ofrecer nuestro servicio?



*PaaS*



developer



Utilizando a Flask/Django como plataforma, que gestione usuarios con tokens de acceso.



Brindando una API con documentación para desarrolladores.

---

*SaaS*



user



Creando un plugin para una SaaS utilizada (como wordpress) que consuma nuestras APIs por debajo.



Brindar una interfaz de configuración (GUI) para no programadores



Link al Colab



[LINK](#)



Link al github



*LINK*

# ¿Cómo hoy NLP se relaciona con otros campos del deeplearning?

[LINK](#)



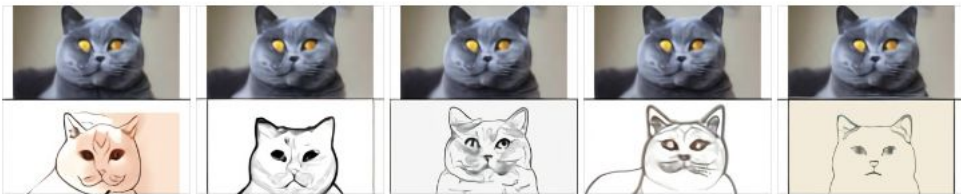
## DALL·E: Creating Images from Text

We've trained a neural network called DALL·E that creates images from text captions for a wide range of concepts expressible in natural language.

TEXT & IMAGE  
PROMPT

the exact same cat on the top as a sketch on the bottom

AI-GENERATED  
IMAGES



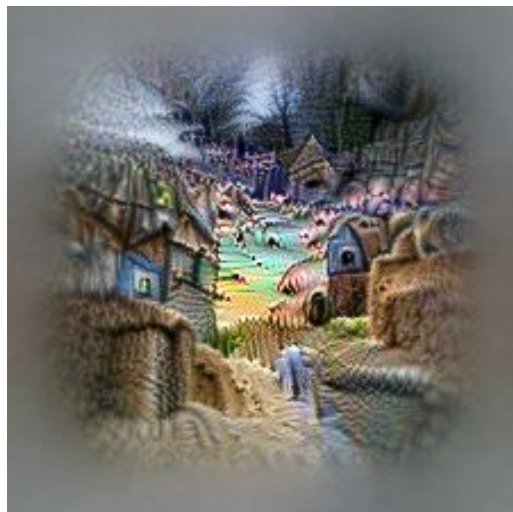
[Edit prompt or view more images](#)



# OpenAI

# CLIP: Connecting Text and Images

We're introducing a neural network called CLIP which efficiently learns visual concepts from natural language supervision. CLIP can be applied to any visual classification benchmark by simply providing the names of the visual categories to be recognized, similar to the "zero-shot" capabilities of GPT-2 and GPT-3.

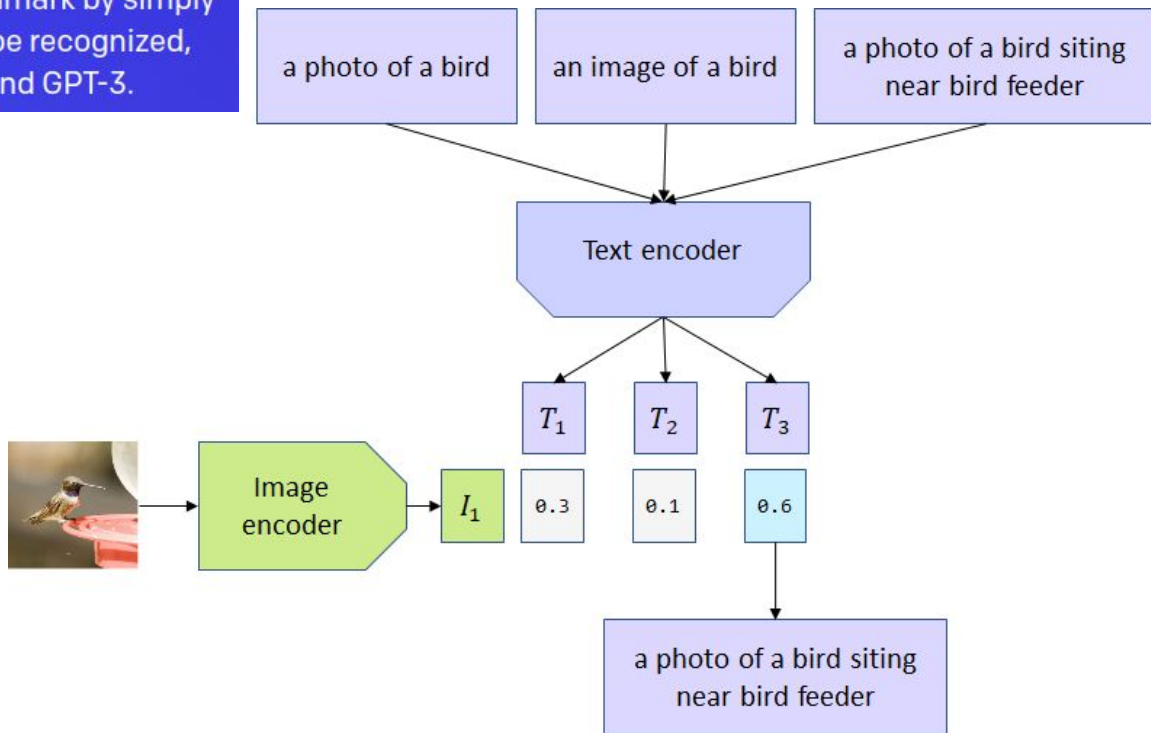


[LINK](#)

[LINK](#)



Ahora si: "una imagen dice más que mil palabras"





# OpenAI Codex

We've created an improved version of OpenAI Codex, our AI system that translates natural language to code, and we are releasing it through our API in private beta starting today. Codex is the model that powers [GitHub Copilot](#), which we built and launched in partnership with GitHub a month ago. Proficient in more than a dozen programming languages, Codex can now interpret simple commands in natural language and execute them on the user's behalf—making it possible to build a natural language interface to existing applications. We are now inviting businesses and developers to build on top of OpenAI Codex through our API.

[LINK](#)



**GitHub**  
Copilot





# ¡Muchas gracias!