**AI Pair Programming**

1. **Pruebas de rendimiento**

* GPT-Neo-2.7B. Huggingface

Prueba del modelo GPT-Neo con 2.7 Billones de parámetros de *EleutherAI*, modelo que ha servido de base para entrenar el código del modelo de APPS. El tiempo para cargar el modelo son 4.5s y para los 10 problemas tarda 10 minutos aproximadamente (9,99 minutos).

Con 2.7 Billones de parámetros, este modelo está diseñado con una red neuronal *transformer* que replica la arquitectura de GPT-3. El ejemplo utiliza aproximadamente 20GB de la RAM (5.13GB ~ 25.9GB), siendo, por tanto, superior a la capacidad máxima de la GPU (16GB), con la cual se pretende ejecutar los modelos escogidos definitivos.

* CodeGPT-small-py-adaptedGPT2. Huggingface

Prueba del modelo *CodeGPT* con 125 Millones de parámetros de Microsoft. En comparación con otros modelos, tiene una cantidad mucho menor de parámetros, pero el rendimiento que ofrece es acorde a lo esperado por las métricas del modelo entrenado.

Está diseñado con una red neuronal *transformer* que adapta la arquitectura de GPT-2. Utiliza aproximadamente 1GB de la RAM (12.9GB ~ 13.9GB), siendo muy inferior a la capacidad máxima de la GPU (16GB). El tiempo que tarda en cargar el modelo es de 24.56s y el tiempo en resolver 10 problemas es de 20.25s

Si se lleva a cabo una prueba de rendimiento con la GPU, el modelo sigue ejecutándose de manera correcta ocupando un espacio de 2676MiB de los 16384MiB, es decir, 2.8GB de los 17.17GB totales, lo que se traduce en un 16.31% de uso de la GPU. Conviene resaltar dicho dato ya que servirá como guía para probar el rendimiento de otros modelos.

* T5-3B + PICARD

Prueba del modelo T5-3B + PICARD con 3 Billones de parámetros. En cuanto a modelos de generación de sentencias SQL es el más completo puesto que en su entrenamiento incluye las sentencias más complejas gracias a *Spider*.

El ejemplo con 5 sentencias de prueba tarda 3.3s en cargar el modelo y 57s en realizar todos los problemas y completar la ejecución. Respecto al espacio que ocupa este modelo, desde los 28.8GB alcanza los 50.2GB de la RAM (21.4GB).

Modelo, RAM/GPU, tiempo de carga, #problemas, tiempo total, tiempo por problema, memoria. => Añadir a Excel de selección de modelos

GPT-J: falta probar en RAM

CodeGPT en GPU

1. **Propiedad intelectual**

* Codex

1. Concesión de licencia. Sujeto a su cumplimiento de estos Términos, le otorgamos una licencia limitada, no exclusiva, no sublicenciable, no transferible, no asignable y revocable durante la duración para: (i) usar las API y la Documentación del desarrollador únicamente para desarrollar, probar, ejecutar, operar y dar soporte a su Aplicación; (ii) para permitir que los usuarios finales usen su integración de las API dentro de su Aplicación; (iii) para mostrar el Contenido al que se accede a través de las API dentro de su Aplicación y (iv) según lo autoricemos expresamente por escrito.
2. Derechos de autor. OpenAI no afirmará los derechos de autor sobre el Contenido generado por la API para usted o sus usuarios finales.
3. Envío de Contenido. OpenAI no adquiere ninguna titularidad de ningún derecho de propiedad intelectual sobre el contenido que envíe a nuestras API a través de su Aplicación, salvo que se indique expresamente en estos Términos. Con el único propósito de permitir que OpenAI y sus afiliados proporcionen, aseguren y mejoren las API (y el software, los modelos y los algoritmos relacionados), usted otorga a OpenAI y sus afiliados una licencia perpetua, irrevocable, mundial, sublicenciable, libre de regalías y licencia no exclusiva para usar, alojar, almacenar, modificar, comunicar y publicar todo el contenido enviado, publicado o mostrado hacia o desde las API a través de su Aplicación. Cuando lo permitan las Leyes de privacidad aplicables, la licencia anterior sobrevive a las solicitudes de los consumidores para la eliminación de datos personales o Información personal con el único propósito de permitir que OpenAI y sus afiliados proporcionen, aseguren y mejoren las API. Antes de enviar contenido a nuestras API a través de su Aplicación, se asegurará de tener los derechos necesarios (incluidos los derechos de sus usuarios finales) para otorgarnos la licencia.

Propiedad sobre los modelos generados (es decir, los nuevos pesos)

Fuente: [*https://openai.com/api/policies/terms/*](https://openai.com/api/policies/terms/)

1. **Otros modelos**

* OPT: Open Pre-trained Transformer Language Models

El modelo solo preentrenado se puede usar para solicitar la evaluación de tareas posteriores, así como la generación de texto. Además, el modelo se puede ajustar en una tarea posterior utilizando el ejemplo de CLM.

OPT se entrenó predominantemente con texto en inglés, pero una pequeña cantidad de datos que no están en inglés todavía está presente en el corpus de capacitación a través de *CommonCrawl*. El modelo se entrenó previamente utilizando un objetivo de modelado de lenguaje causal (CLM). OPT pertenece a la misma familia de modelos de solo decodificador como GPT-3. Como tal, se entrenó previamente utilizando el objetivo de modelado de lenguaje causal autosupervisado.

Para la evaluación, OPT sigue a GPT-3 mediante el uso de sus indicaciones y la configuración experimental general. Para más detalles, por favor lea el documento oficial.

Fuente: [*https://arxiv.org/pdf/2205.01068.pdf*](https://arxiv.org/pdf/2205.01068.pdf)

1. **Idioma**

No hay documentación disponible para el resto de modelos

Ver referencias de benchmark de generación de código en castellano (text-to-code) (Codex u otros)

* GPT-3:

Prueba de GPT-3 de *OpenAI* en diez conversaciones de clientes sobre banca en español sin ningún tipo de adaptación previa, re-entrenamiento o aprendizaje por transferencia.

Si bien la exactitud de esta prueba sobre diez conversaciones no puede ser muy alta, los resultados obtenidos son alentadores e invitan a continuar explorando el valor del modelo para las tareas elegidas (por ej. una comparativa contra el nivel de rendimiento humano).

Fuente: [*https://www.bbva.com/es/puede-ayudar-gpt-3-en-conversaciones-con-nuestros-clientes-en-espanol/*](https://www.bbva.com/es/puede-ayudar-gpt-3-en-conversaciones-con-nuestros-clientes-en-espanol/)

1. **Licencias**

Ver límites de la capa gratuita de Codex

¿Incluye llamadas a la API? O solo playground?

¿Número máximo de llamadas?

¿Incluye fine tunning/entrenamiento?

¿Tiene fecha límite para inferencia y para entrenamiento?

Revisar Azure OpenAI

Propiedad intelectual (datos, modelos generados)

Precio

Hay que pedir acceso