IA Segura y libre, el uso de LLM´s en entornos locales con GNU/LINUX

Ing. Fernando Villares – J.R.S.L. 2024

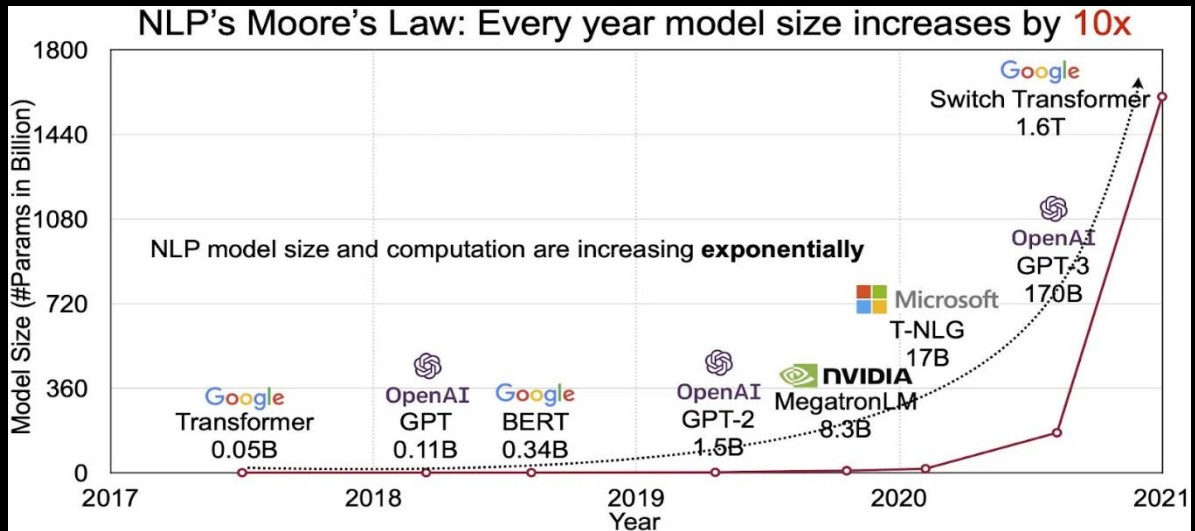
PSYWARE SSL

¿Qué es un LLM?

- Modelos de lenguaje de gran tamaño, son modelos de aprendizaje profundo que se pre-entrenan con grandes cantidades de datos
- Los transformadores LLM son capaces de entrenarse sin supervisión, aunque una explicación más precisa es que los transformadores llevan a cabo un autoaprendizaje.
- A diferencia de las redes neuronales recurrentes (RNN) anteriores que procesaban las entradas de forma secuencial, los transformadores procesan secuencias enteras en paralelo.
- La arquitectura de los transformadores de redes neuronales permite el uso de modelos muy grandes, a menudo con cientos de miles de millones de parámetros.


1 Data Gathering
2 Data Cleaning
3 Data Splitting
4 Model Set-up
LLM Models Training
5 Model Training
6 Model Checking
7 Model Usability
8 Model Enhancement

¿Qué es un LLM?


NLP's Moore's Law: Every year model size increases by 10x
Model Size (#Params in Billion)
1800
1440
1080
720
360
0
2017
2018
2019
2020
2021
Year
NLP model size and computation are increasing exponentially
Google Transformer 0.05B
OpenAI GPT 0.11B
Google BERT 0.34B
OpenAI GPT-2 1.5B
NVIDIA MegatronLM 8.3B
Microsoft T-NLG 17B
Google Switch Transformer 1.6T
OpenAI GPT-3 170B

Algunos de los LLM en la nube...

- ChatGPT - https://chatgpt.com/
- GEMINI - https://gemini.google.com/
- Copilot - https://copilot.microsoft.com/
- Claude - https://claude.ai
- Mistral - https://mistral.ai
- Meta - https://www.meta.ai/

Integradores Inteligentes de servicios...

- You - https://you.com/
- TextCortex - https://textcortex.com
- Poe - https://poe.com/

Los riesgos de usar un LLM en la nube...

- Los LLM son bastante seguros, pero al correr masivamente en la infraestructura global de proveedores mundiales, se pueden dar diversos casos de fallas....
  - ***Privacidad:*** Información sensible de un contexto o usuario que es filtrada a otro.
  - ***Compliance o cumplimiento:*** Si el proveedor guarda la info de contexto, los resultados o los prompts de forma indefinida, puede violar normas como la RGPD, LOPD, ley de datos personales etc.
  - ***Copyright:*** Hay que leer cuidadosamente los términos de servicio para evitar que el contenido generado pueda violar derecho de autor o que al entrenar un modelo se viole la imagen o los derechos de los propietarios de la información usada.

Los riesgos de usar un LLM en la nube...

- *Secretos industriales, de estado, información confidencial, judicial, etc:* Los sistemas se retroalimentan con la información brindada para mejorar el aprendizaje, por lo cual, si los alimentan con información sensible, es muy probable que la misma se convierta en base de conocimientos para nuevas consultas de cualquier usuario.
- *Fake News:* los LLM, con sus propias alucinaciones o de forma deliberada pueden ser usados para generar información falsa la cual puede resultar sumamente realista.,
- *Manipulación de datos:* A través de fallas de seguridad de las plataformas se pueden manipular los ingresos de datos en los prompts para desviar los resultados, inyectar código malicioso, falsear el resultado de un estudio analizado, etc.

Algo de glosario...

- ***Trasformers:*** Se refiere a un tipo de arquitectura de red neuronal que utiliza técnicas de aprendizaje profundo y autoatención para manejar secuencias de datos. Los Transformers funcionan mediante un mecanismo de autoatención que les permite procesar secuencias de datos de manera no secuencial. A diferencia de las Redes Neuronales Recurrentes (RNN) y las Long Short-Term Memory networks (LSTM), los Transformers no necesitan procesar los datos en orden, lo cual les permite realizar un procesamiento en paralelo mucho más eficiente.

Algo de glosario…

- ***Token:*** Un token en la IA es la unidad más pequeña en la que se puede dividir una palabra o frase. Puede ser una palabra completa, un signo de puntuación, una subpalabra (como la mitad de una palabra compuesta) o incluso un carácter especial.
- ***Alucinaciones:*** Son respuestas seguras de una IA que no parece estar justificada por sus datos de entrenamiento.

Algo de glosario...

- ***GGML (Lenguaje de Modelos Generados por GPT):*** Desarrollado por Georgi Gerganov, GGML es una biblioteca de tensores diseñada para el aprendizaje automático, facilitando modelos grandes y alto rendimiento en diverso hardware.

- ***GGUF (Formato Unificado Generado por GPT):*** Introducido como sucesor de GGML (Lenguaje de Modelos Generados por GPT), fue lanzado el 21 de agosto de 2023. Este formato representa un avance significativo en el campo de los formatos de archivo de modelos de lenguaje, facilitando el almacenamiento y procesamiento mejorados de grandes modelos de lenguaje como GPT.

Algo de glosario...

- ***Parámetros:*** En IA y aprendizaje automático, un parámetro es un valor que se utiliza para configurar un modelo o algoritmo de aprendizaje. Los parámetros se pueden aprender de los datos o se pueden configurar manualmente. En una red neuronal, los parámetros son los pesos y los sesgos de la red. Los valores de estos pesos y sesgos determinan cómo aprende la red a asignar las características de entrada a los objetivos de salida.

Algo de glosario...

- ***Cuantización:*** Consiste en reducir la precisión o el ancho de bits de los valores numéricos en un modelo de IA. Al representar los números con menos bits, se reduce el uso de memoria y los requisitos computacionales de los modelos. Existen distintos tipos de métodos de cuantización, como la cuantización en coma fija y la cuantización en coma flotante. Sin embargo, puede provocar cierta degradación en la precisión del modelo, especialmente cuando se reduce agresivamente la precisión de los valores numéricos.

Algo de glosario…

- ***Temperatura:*** La temperatura de un LLM sirve como un parámetro crítico que influye en el equilibrio entre previsibilidad y creatividad en el texto generado. Las temperaturas más bajas dan prioridad a la explotación de patrones aprendidos, lo que genera resultados más deterministas, mientras que las temperaturas más altas alientan la exploración, fomentando la diversidad y la innovación.
- ***RAG:*** La generación mejorada por recuperación (RAG) es el proceso de optimización de la salida de un modelo lingüístico de gran tamaño, de modo que haga referencia a una base de conocimientos autorizada fuera de los orígenes de datos de entrenamiento antes de generar una respuesta.

Algo de glosario…

- ***Repeat penalty*: S**e utiliza para evitar que el modelo repita las mismas palabras con demasiada frecuencia en el texto generado. Es un valor que se resta a la probabilidad de elegir un token cada vez que ocurre en el texto generado. Un valor elevado hará que el modelo sea menos propenso a repetir tokens.

- ***RNG Seed / Seed:*** Establece la semilla con la que se inicializa el generador de números aleatorios. Permite obtener resultados repetibles usando la misma semilla.

Algo de glosario...

- ***Top K Sampling:*** Es un método que selecciona el siguiente token de un subconjunto formado por los k tokens más probables. A menor valor tenga más predecible será el texto generado.

- ***Top P Sampling:*** Similar a Top K, con la diferencia de que selecciona el siguiente token de un subconjunto de tokens que juntos tienen una probabilidad acumulada de al menos p.

¿Podemos correr un LLM de forma LOCAL?

SI, a través de diversos softwares, que corren los prompts de forma offline localmente sin transmitirlos y con las ventajas de seguridad, compliance, límite de tokens que eso propone, pero a costa de disminución de velocidad y teniendo en cuenta el costo de del hardware de ejecución.

- LMStudio
- ChatWithRTX
- LocalAI
- A mano programando en Python...

Los requisitos de hardware necesarios

Si utilizamos un modelo de precisión total en 16 bits, cada parámetro ocupa 2 bytes .

La cuantización permite reducción del tamaño del modelo en RAM pero con pérdida de precisión.

- DISCO: Modelos de 70Billones de parámetros ocuparán aprox. 140gb de datos. Más rápido el disco = más rápida la carga inicial a RAM.
- RAM: Los modelos corren en RAM o VRAM por lo cual la suma de memoria RAM libre más la VRAM gráfica debería ser igual al tamaño del mismo. Lo que más importa es la cantidad de canales simultáneos y la velocidad.
- Las mejores computadoras para IA confluyen en una gran cantidad de RAM con canales simultáneos como las AMD EPYC GENOA con 12 canales máximos de 2 dimms cada una de 256GB DDR5 por CPU. En Gama Hogar RYZEN 7 79xx con 256 de ram.

Los requisitos de hardware necesarios

- CPU: Una cpu rápida multinúcleo es importante, pero no es crítico.
- Las mejores CPU para IA de propósito general hoy día son las AMD EPYC GENOA con hasta 128 núcleos y 256 tareas por CPU.

- GPU/VRAM: Las placas gráficas están preparadas para procesar de forma paralela miles de tareas y la VRAM es mucho más rápida que la RAM por lo cual es el componente más crítico para un sistema LLM o de ia
- La placa reina es la NVIDIA H200 con 141gb de VRAM con 1979 TFlops en 16 bits de performance y hasta 8 gráficas por nodo con NVLINK.
- AMD INSTINCT MI300X de 192gb, con 1300 Tflops en 16 bits y 8 Gráficas por nodo con Infinity Fabric.
- En gama hogar, Radeon XTX7900 de 24gb con 103 TFlops y NVIDIA 3090/4090 de 24gb y mas de 80/165 Tflops en 16 bits.

El software libre base: LMStudio

Es una aplicación de soft libre, versátil que permite a los usuarios descubrir, descargar y ejecutar varios LLMs locales directamente en sus portátiles o equipos de escritorio. El software es compatible con una amplia gama de modelos, incluyendo LLaMa, Falcon, MPT, StarCoder, Replit, GPT-Neo-X y más, todos provenientes de los repositorios de Hugging Face. La actualización v0.3.1 ha ampliado aún más sus capacidades, incluyendo soporte para Mixtral, Phi-2, StableLM y modelos habilitados para visión y RAG. Corre de forma nativa como una app en formato AppImage en GNU/linuX.

Diferentes modelos, diferentes resultados...

En la web Hugging Face que es el GitHub de la IA, se pueden bajar y testear cientos de modelos de IA liberados.

- LlaMa 3.1, en versiones de 8, 70 y 405 billones de parámetros, el caballo de batalla de meta, con y sin cuantizaciones.
- Phi-3.5, de Microsoft con 8 billones de parámetros sorprende por su velocidad y simpleza.
- Vicuna 13B, con 13 billones de parámetros, llega al 90% de la calidad de respuestas de chatGPT 4, basado en llama2 tuneado.
- Bloom, entrenado en francia por un gigantesco equipo internacional, algo desactualizado.

Diferentes modelos, diferentes resultados...

- LlaVA, biblioteca de visión por computadora, acepta imágenes y te genera descripciones de las mismas.
- Mistral NeMO ultra compacto y veloz entrenado con nvidia, con 12 Billones de parámetros
- Mixtral, la evolución de mistral, usando 8 agentes expertos de 7 billones de parámetros a la vez y siendo 6 veces más eficiente que el código de 40 billones estándar.
- FLUX.1 [dev]: Destinado a usos no comerciales, este modelo es una versión destilada de FLUX.1 [pro], que mantiene una calidad y adherencia al prompt similar, pero con un menor costo computacional.

Práctica

- Bajar LMStudio app image 0.3.2de https://lmstudio.ai
- Bajar un modelo LLM de bajo peso para testing como ser PHI 3,5 de 8 billones de parámetros versión mini con Quantization de 8 Bits (4GB)
- Ejecutar el modelo y en caso de tener placa gráfica modificar el parámetro de capas hasta adecuarlo al uso del 100% de vram y verificar la diferencia de velocidad de procesamiento.
- Bajar un modelo mas potente como LlaMA 3.1 con de 8 o mas billones de parámetros con o sin quantization dependiendo de la ram y gráficas que tengamos y comparar resultados y performance.
- Probar de hacer preguntas censuradas e intentar bypassear los controles de seguridad de los modelos.
- Probar si el sistema puede codificar en diferentes lenguajes de programación o entender análisis de casos de uso en sistemas.
- Probar con diferentes contextos y si el sistema va arrastrando conocimientos previos.

¡GRACIAS POR TODO!

- Ing. Fernando Villares – 09/2024
- https://psyware.ar
- contacto@psyware.ar
- @fmvillares
- Bajo licencia Creative Commons Atribución-CompartirIgual 2.5 Argentina (CC BY-SA 2.5)

PSYWARE SSL