Explorando la complejidad textual: Conversaciones con modelos NLP basadas en un PDF

García Camacho L.A., Moreno Alcalá E.H., Rodarte Campos J.A., Muñoz Castrejón I., Hernández Tiscareño F.J.
luisalfonso.garcia, estefany.moreno , jesusalberto.rodarte, irving.munoz, fernando.hernandez (@fisica.uaz.edu.mx)

Resumen

Un modelo de procesamiento de lenguaje natural **(NLP)** es un sistema diseñado para comprender y generar lenguaje humano de manera automatizada. En el presente trabajo se exploraron las diferentes cuantificaciones para los datos en los valores de los parámetros de los modelos **NLP** o **LLM** (Large Language Model), así como la variación entre sus complejidades, relacionándola con la entropía de Shannon. Además, se trabajó con un set de datos de diversas prácticas de laboratorio y mediante el uso de algoritmos de clasificación de datos y la API del chat-GPT se vectorizaron los textos y se creó una aplicación donde se puede conversar con una extensión de chat-GPT especialmente hablando del modelo **"text-embedding-ada-002"**, restringiendo las respuestas a la información del PDF por medio de un proceso llamado **embedding** que consiste en vectorizar un texto para que, posteriormente, haciendo uso de una función se compare entre todos los párrafos y arroje las respuestas más similares.

Introducción

La rama NLP es una tecnología de aprendizaje automático que permite a las computadoras interpretar, manipular y comprender el lenguaje humano. El NLP es fundamental para analizar a profundidad los datos de texto y voz de manera eficiente, puede resolver las diferencias en dialectos e irregularidades gramaticales típicas en las conversaciones cotidianas.

Hoy en día muchos de estos modelos LLM de la rama NLP se encuentran libres al público, sin embargo, existe la problemática del poder de cómputo necesario para ejecutarlos. Debido a ello, surgen herramientas como QLoRA (Quantization Low Rank Adapters) que permiten reducir la memoria que utilizan estos modelos. En este trabajo con el cálculo de la entropía de Shannon se midió la pérdida de información al emplear el método QLoRA. Asimismo, se hace un tratamiento desde el punto de vista de los embeddings que da como resultado una aplicación conversacional con PDFs que se limita a respoder con lo contenido de estos mismos, donde dichos PDFs son cargados por el usuario.

Entropía de Shannon

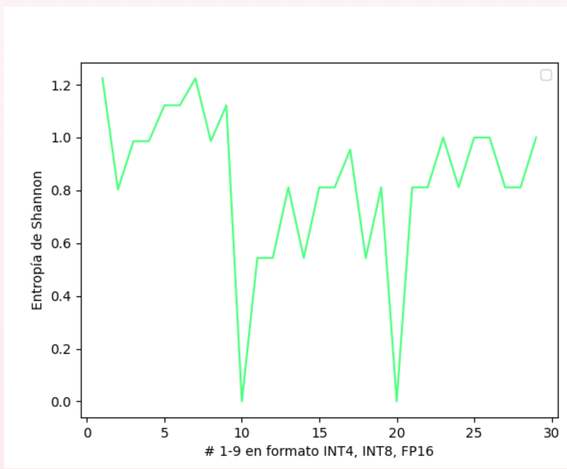
La entropía de Shannon es un concepto fundamental en la teoría de la información y se utiliza en el campo de la computación y la codificación de la información. Esta es utilizada para cuantificar la cantidad de redundancia en un conjunto de datos. Cuanto mayor sea la entropía, menor será la redundancia y, por lo tanto, más difícil será comprimir los datos sin pérdida de información.

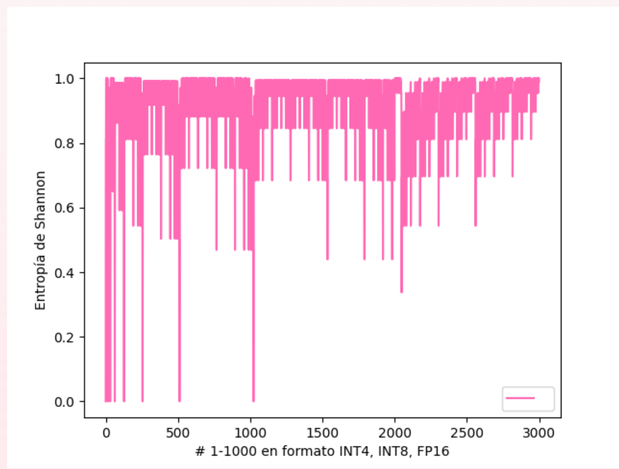
$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2(p(x_i))$$

Donde $H(X)$ es la entropía, $p(x_i)$ es la probabilidad de que ocurra x_i en los datos X, log_2 es el logaritmo base 2 y se hace la suma sobre los n datos de X.

Diferentes tipos de datos

Cuando se trabaja con el método QLoRA se tiene que cambiar de un tipo de datos a otro. En este caso, los modelos LLM trabajan con datos en formato **FP16** y se pueden hacer cuantizaciones en las que se obtienen datos en formato **INT8** o **INT4**, donde se necesitan 16, 8 y 4 bits, respectivamente, para representar algún numero. Por tanto, se trabajó primero con los números del 0 al 9 representados en cada uno de los formatos mencionados y se calculó la entropía de Shannon como se muestra en la figura de la izquierda. Después, se repitió el procedimiento para los números del 0 a 1000 como se muestra en la figura de la derecha.

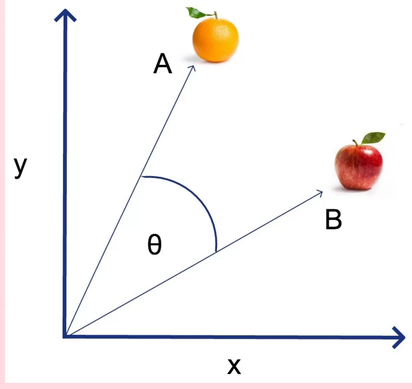




Embeddings

Los embeddings son representaciones numéricas de palabras o elementos en un espacio vectorial multidimensional. Estas representaciones son ampliamente utilizadas en diversos campos de la informática para capturar y expresar el significado y las relaciones semánticas entre palabras u objetos. En lugar de tratar las palabras o elementos como simples etiquetas, los embeddings asignan vectores numéricos a cada uno de ellos, de modo que aquellos que comparten similitudes en significado o características compartan ubicaciones cercanas en este espacio vectorial. La generación de embeddings se lleva a cabo a través de modelos matemáticos que se entrenan en grandes conjuntos de datos para aprender representaciones contextuales y significativas de las palabras u objetos en cuestión.

$$\cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$



En este trabajo se hizo uso del modelo **"text-embedding-ada-002"** para hacer los embeddings de varios PDFs referentes a prácticas de laboratorio y con una función de la paquetería de Open IA llamada **"cosine_similarity"** se comparó el coseno de los vectores, esperando que estos formaran un ángulo de 0 y al aplicarlo al coseno diera como resultado 1 y entre más cercano estuvieran estos vectores, más cercano sería el valor a 1, dando así la posibilidad de construir listas que muestren los textos con mayor similitud al hacer una pregunta, todo esto se realizó con uso de la paquetería de gradio para mostrar una interfaz amigable.

Conclusiones

Debido a las gráficas de los diferentes formatos de datos con los que trabaja el método QLoRA se obtuvo que, aunque en su mayoría se tienen valores con entropías altas, hay otros casos donde se observa como cae el valor de esa entropía y esto permite hacer una compresión de la información (los parámetros de los LLM), y permitiendo con esto que no se pierda mucho desempeño, pero sí que se reduzca mucho la memoria VRAM o RAM empleada para ejecutarlos. Por otro lado, la parte de embeddings demostró ser una herramienta poderosa para el objetivo conversacional con PDFs, que se limita a dar respuestas dentro del documento y no responde con información falsa como lo hace en ocasiones chat-GPT.

Referencias

[1] Delahaye, J.-P., & Zenil, H. (2012). Numerical evaluation of algorithmic complexity for short strings: A glance into the innermost structure of randomness. Applied Mathematics and Computation, 219(1), 63-77. Elsevier. and Systems, 3rd edn. (Harcourt Brace Jovanovich), 1988.

[2] Vajapeyam, S. (2014). Understanding Shannon's Entropy metric for Information. arXiv preprint arXiv:1405.2061. [cs.IT]

[3] Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. arXiv preprint arXiv:2305.14314.