FACULTAD DE CIENCIAS EXACTAS Y NATURALES

Licenciatura en Ciencias de Datos

Propuesta de tema de tesis

Cambio de tokenizador mediante finetuning

El presente proyecto tiene como objetivo analizar los efectos de reentrenar el tokenizador de un modelo de lenguaje preentrenado mediante métricas que relacionan el comportamiento de éstos modelos con respuestas humanas a estímulos lingüíticos.

Palabras clave: `NLP, LLM, GPT, Neurociencias`

Conocimientos deseables

Librerías de python asociadas a NLP (Huggingface) y análisis de datos (numpy, pandas, seaborn, etc)

¿Qué podría aprender quien realice esta tesis?

Técnicamente el/la estudiante aprenderá a trabajar con grandes modelos de lenguaje, tanto a nivel de uso como, porsiblemente, de reentrenamiento del mismo. A nivel experimental nos proponemos trabajar fuertemente en la generación iterativa de hioótesis y sus respectivos testeos. El análisis de los resultados obtenidos en cada experimento será altamente formativo en el área experimental.

Dirección de la tesis

Bianchi, Bruno
Laboratorio de Inteligencia Artificial Aplicada

Contacto: *brunobian@gmail.com*

Más información en el pdf a continuación.

Cambio de tokenizador mediante finetuning

Director: Bruno Bianchi
Equipo: Juan E Kamienkowski, Diego Fernández Slezak, Fermin Travi

Para que una red neuronal sea capaz de abstraer el significado de las palabras es necesario que las mismas se representen de forma numérica. De esta forma, las palabras son representadas como vectores de una gran cantidad de dimensiones. Son los valores de estos vectores (entre otras cosas) lo que se entrena a la hora de entrenar una red neuronal de este tipo. Este proceso se realizó a nivel de las palabras durante muchos años. Es decir, cada palabra (definida como un cadena de caracteres alfabéticos que se encuentra entre dos espacios) se representa con un único vector. Sin embargo, esto plantea limitaciones. Por ejemplo, la red neuronal sólo podrá representar palabras que hayan sido utilizadas durante el entrenamiento.

Para solucionar esto se han desarrollado diferentes alternativas de *tokenización* (proceso por el cual se definen los *tokens*, que hasta este momento eran las palabras). Actualmente uno de los procesos de tokenización más utilizados es el denominado Byte-Pair Encoding (BPE). En éste se realiza un proceso estadístico para determinar qué conjuntos de caracteres (ngrams) es más eficiente usar como tokens. Este tipo de tokenización sublexical le otorga grandes ventajas a los modelos, como poder representar palabras fuera del vocabulario, o ser entrenado en más idiomas sin tener que contar con un vocabulario de cada uno.

Sin embargo, a la hora de realizar análisis lingüísticos resulta interesante analizar los vectores que representan a cada una de las palabras (con la misma definición que antes) y no de los tokens, que no cuentan con un significado lingüístico claro. En un paper muy reciente [1] mencionan que "*We used a pre-trained GPT-2 Small (Radford et al., 2019) model, which* ***we fine-tuned to change its tokenization from BPE (Sennrich et al., 2016) to word-level (i.e., whitespace-delimited)*** *so that its tokenization scheme would match the experimental protocol for the human participants*".

Al realizar este cambio de tokenización sería posible utilizar modelos más modernos (como Llama2) para realizar comparaciones entre las representaciones vectoriales de los modelos de lenguaje y el procesamiento humano. Hoy en día, cuando es necesario realizar análisis de este tipo con representaciones de modelos con tokenización BPE se utilizan alternativas como promediar los vectores de todos los tokens, o utilizar el último. Así, resulta interesante comprobar cuál de estas alternativas, incluída la del cambio de tokenización, mejora los resultados encontrados.

Objetivo mínimo: Finetunear el tokenizador de GPT2 y replicar resultados obtenidos en el laboratorio con el tokenizador original.

Objetivo máximo: Finetunear el tokenizador de Llama2 y replicar y expandir los resultados obtenidos en el laboratorio con el tokenizador original.

Referencia:

[1] Vaidya, A. R., Turek, J., & Huth, A. G. (2023). Humans and language models diverge when predicting repeating text. arXiv preprint arXiv:2310.06408.