FACULTAD DE CIENCIAS EXACTAS Y NATURALES

Licenciatura en Ciencias de Datos

Propuesta de tema de tesis

Contextualización de palabras polisémicas en transformers

El objetivo principal del presente proyecto es estudiar el paralelismo existente entre los parámetros internos de los Grandes Modelos de Lenguaje y los procesos cognitivos que ocurren en el cerebro de las parsonas al enfrentarse a estimulos lingüísticos, medidos a través de respuestas comportamentales humanas.

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.

Contextualización de palabras polisémicas en transformers

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

El gran desarrollo en los algoritmos de Procesamiento del Lenguaje Natural (NLP), como los Transformers, en los últimos años ha permitido resolver tareas lingüísticas de gran complejidad de manera similar a los humanos. Desde el área de la Neurociencia Cognitiva Computacional estos modelos se observan análogos a lo que ocurre al analizar el cerebro: estamos estudiando una "caja negra", a la cual le podemos introducir valores y obtener respuestas, pero de la cual nos queda mucho por comprender sobre su funcionamiento interno [1,2].

El objetivo principal del presente proyecto es estudiar el paralelismo existente entre los parámetros internos de las redes profundas y procesos cognitivos medidos a través de respuestas comportamentales humanas.

Para esto, contamos con un corpus de textos generados a partir de una selección de palabras polisémicas (es decir, palabras que tienen más de un significado). Para cada palabra polisémica se armaron oraciones neutras (es decir, que no logren desambiguar el significado de la palabra) y 3 párrafos que fueron utilizados como contexto previo a las oraciones neutras. Estos tres párrafos consisten en un párrafo neutro (no asociado a ninguno de los significados de la palabra), uno sesgado hacia el significado de mayor dominancia de la palabra, y el tercero sesgado hacia el significado de menor dominancia de la palabra. Con este corpus se realizó un experimento comportamental con el cual calculamos la capacidad de sesgar la asignación de significado a cada palabra.

El objetivo computacional actual es encontrar la relación existente entre ésta métrica comportamental y las representaciones vectoriales internas de los transformers [3].

Objetivo mínimo: Analizar cómo varía el sesgado a lo largo de la red (o sea, replicar la figura de la SAN para cada capa)

Objetivo máximo: Tomar el código de Jean-Remi King y hacer algo de ese desanudamiento

Referencias:

[1] Abnar, S., Beinborn, L., Choenni, R., & Zuidema, W. (2019, August). Blackbox Meets Blackbox: Representational Similarity & Stability Analysis of Neural Language Models and Brains. In Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (pp. 191-203).
[2] Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In Advances in Neural Information Processing Systems (pp. 14954-14964).
[3] Caucheteux, C., Gramfort, A., & King, J. R. (2021, July). Disentangling syntax and semantics in the brain with deep networks. In International conference on machine learning (pp. 1336-1348). PMLR.