Fundamentos de los Grandes Modelos de Lenguaje y aplicación de técnicas de prompting en IAs generativas

Equipo docente

- Fernando Carranza (UBA)
- Julia Milanese (UBA)
- Fernando Schiaffino (UBA)

Fechas: Martes 15, 22 y 29 de abril de 18:00 a 21:00

Introducción

En este curso abordaremos los grandes modelos de lenguaje desde una perspectiva tanto teórica como práctica. Con respecto a la perspectiva teórica, nos dedicaremos a motivar y presentar las nociones básicas que subyacen al funcionamiento de los grandes modelos de lenguaje: la semántica de vectores, la arquitectura general de las redes neuronales, los fundamentos del mecanismo de atención y los transformers. Esperamos que esto siente las bases para aproximarse a la comprensión de cómo es el funcionamiento de un modelo de lenguaje. Con respecto a la perspectiva práctica, en la tercera clase haremos una presentación de las principales técnicas de *prompting* o ingeniería de instrucciones. Esta área consiste en el estudio de los métodos para interactuar con los modelos y resulta clave para obtener resultados precisos y relevantes. En esa clase veremos técnicas como Zero-Shot y Few-Shot, es decir, la interacción sin ejemplos previos o solo con algunos respectivamente, así como Retrieval Augmented Generation (RAG), es decir, la combinación de generación de texto y recuperación de información, y Chain of Thought, o la estructuración del razonamiento del modelo de modo tal que permita resolver problemas complejos y mejorar la calidad de las respuestas. El curso está orientado a estudiantes y egresados y egresadas de la carrera de Letras y carreras afines. No se presupondrán conocimientos previos.

Objetivos

Con este curso, esperamos que los y las estudiantes alcancen los siguientes objetivos:

- Comprender los conceptos fundamentales del procesamiento del lenguaje natural basado en redes neuronales.
- Familiarizarse con los mecanismos internos de aprendizaje profundo que subyacen a los modelos de lenguaje.
- Conocer y aplicar estrategias de ingeniería de instrucciones (*prompt engineering*) para interactuar con grandes modelos de lenguaje.

Contenidos

Clase 1: Nociones básicas de redes neuronales para el procesamiento del lenguaje natural

Tokenización, embeddings estáticos; el perceptrón; arquitectura básica de redes neuronales; funciones de activación; *backpropagation*.

Clase 2: Grandes y pequeños modelos de lenguaje

Mecanismo de atención; transformers; embeddings posicionales; similitudes y diferencias entre modelos de lenguaje.

Clase 3: *Prompt engineering*

Hiperparámetros de los modelos de lenguaje; prompting: Zero-Shot, Few-Shot, RAG, Chain of Thought y otros.

Bibliografía específica

Bibliografía de referencia de clase 1:

- Chollet, F. 2018. "What is deep learning". *Deep learning with Python*. Capítulo 1. (para punto)
- Jurafsky, Daniel y James H. Martin. 2024. "Vector semantics and embeddings". Speech and Language Processing. Draft of February 3, 2024 (para punto). https://web.stanford.edu/~jurafsky/slp3/6.pdf
- Jurafsky, Daniel y James H. Martin. 2024. "Logistic regression". Speech and Language Processing. Versión no final de Febrero de 2024 (para punto). https://web.stanford.edu/~jurafsky/slp3/5.pdf
- Jurafsky, Daniel y James H. Martin. 2024. "Neural Networks and Neural Language Models". Speech and Language Processing. Draft of February 3, 2024 (para punto). https://web.stanford.edu/~jurafsky/slp3/7.pdf

Bibliografía de referencia de clase 2:

- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, n, J. D. M. W. C., & Toutanova, L. K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of naacL-HLT (Vol. 1, No. 2). https://aclanthology.org/N19-1423.pdf
- Jurafsky, Daniel y James H. Martin. 2024. "The transformer". Speech and Language Processing. Draft of February 3, 2024 . https://web.stanford.edu/~jurafsky/slp3/9.pdf
- Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., y Mian, A. 2023. "A comprehensive overview of large language models". *ArXiv*. https://arxiv.org/pdf/2307.06435
- Vaswani, A, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser. 2017. "Attention is all you need". *Advances in Neural Information Processing Systems*. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
- Wang, F., Z. Zhang, X. Zhang, Z. Wu, T. Mo, Q. Lu, J. Xu, X. Tang, Q. He, Y. Ma, M. Huang & S. Wang. 2024. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. https://arxiv.org/pdf/2411.03350

Bibliografía de referencia de clase 3:

- Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., y Wei, J. 2022. "Scaling instruction-finetuned language models". *ArXiv*. https://arxiv.org/pdf/2210.11416
- Jason Wei and Xuezhi Wang and Dale Schuurmans and Maarten Bosma and Brian Ichter and Fei Xia and Ed Chi and Quoc Le and Denny Zhou. 2023. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models" https://arxiv.org/abs/2201.11903

- Jurafsky, Daniel y James H. Martin. 2024. “Model Alignment, Prompting, and In-Context Learning”. *Speech and Language Processing*. Draft of February 3, 2024. https://web.stanford.edu/~jurafsky/slp3/12.pdf
- Shunyu Yao and Jeffrey Zhao and Dian Yu and Nan Du and Izhak Shafran and Karthik Narasimhan and Yuan Cao. 2023. “ReAct: Synergizing Reasoning and Acting in Language Models” https://arxiv.org/abs/2210.03629
- Yunfan Gao and Yun Xiong and Xinyu Gao and Kangxiang Jia and Jinliu Pan and Yuxi Bi and Yi Dai and Jiawei Sun and Meng Wang and Haofen Wang. 2024. “Retrieval-Augmented Generation for Large Language Models: A Survey”. https://arxiv.org/abs/2312.10997
- Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., y Cao, Y. 2022. “REACT: Synergizing Reasoning and Acting in Language Models”. *arXiv*. https://arxiv.org/pdf/2210.03629

Bibliografía general

Textos

- Chôllet, Francois. (2018). Deep Learning with Python. Manning, Shelter Island.
- Downey, Allen, Jeffrey Elkner y Chris Meyers. 2002. *Aprenda a Pensar Como un Programador con Python.* Massachusetts: Green Tea Press. Disponible en https://argentinaenpython.com/quiero-aprender-python/aprenda-a-pensar-como-un-programador-con-python.pdf
- Jurafsky, Daniel y James H. Martin. 2024. *Speech and Language Processing*. Versión no final de Febrero de 2024
- Kay, M. 2005. “ACL lifetime achievement award: A life of language”. *Computational Linguistics*, 31(4):425–438.
- Kedia, Aman y Mayank Rasu. 2020. *Hands-On Python Natural Language Processing*. Birmingham: Packt.
- Osgood, C. E., Suci, G. J., y Tannenbaum, P. H. 1957. *The measurement of meaning*. University of Illinois press.
- Russell, S. J., & Norvig, P. 2010. *Artificial intelligence: A modern approach* (3rd ed.). Prentice Hall.
- Tan, Michael Steinbach y Vipin Kumar. 2005. *Data Mining*. Pearson.
- Vaswani, A, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser. 2017. “Attention is all you need”. *Advances in Neural Information Processing Systems*. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
- Wang, F., Z. Zhang, X. Zhang, Z. Wu, T. Mo, Q. Lu, J. Xu, X. Tang, Q. He, Y. Ma, M. Huang & S. Wang. 2024. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with LLMs, and trustworthiness. https://arxiv.org/pdf/2411.03350

Recursos computacionales

- Chatbot Arena: https://lmarena.ai/
- Embedding Projector: https://projector.tensorflow.org/
- Fasttext: https://fasttext.cc/
- Github (repositorios): https://github.com/
- Google Colab: https://colab.google/
- Huggingface (repositorios): https://huggingface.co/
- Jupyter notebook: https://jupyter.org/

- Kagle (repositorios): https://www.kaggle.com/
- Keras (librería de Python): https://keras.io/
- LangChain: https://www.langchain.com/
- NLTK (librería de Python): https://www.nltk.org/
- Numpy: https://numpy.org/
- Pandas (librería de Python): https://pandas.pydata.org/
- Python: https://www.python.org
- Pytorch (librería de Python): https://pytorch.org/
- Scikitlearn (librería de Python): https://scikit-learn.org/stable/
- Space de Llama-2-7b en Huggingface: https://huggingface.co/spaces/huggingface-projects/llama-2-7b-chat
- Spacy (librería de Python): https://spacy.io/
- Stanza (librería de Python): https://stanfordnlp.github.io/stanza/
- TensorFlow (librería de Python): https://www.tensorflow.org/?hl=es
- Transformer explainer: https://poloclub.github.io/transformer-explainer/
- Word2vec: https://www.tensorflow.org/text/tutorials/word2vec

Cronograma de clases

Clase	Temas	Docente
Clase 1 Nociones básicas de redes neuronales para el procesamiento del lenguaje natural Martes 15/4	**Parte 1** Tokenización, embeddings estáticos	Fernando Schiaffino
	Parte 2 el perceptrón; arquitectura básica de redes neuronales; funciones de activación; backpropagation	Julia Milanese
Clase 2 Grandes y pequeños modelos de lenguaje Martes 22/4	**Parte 1** Mecanismo de atención; transformers; embeddings posicionales	Fernando Schiaffino
	Parte 2 Similitudes y diferencias entre modelos de lenguaje	Fernando Carranza
Clase 3 *Prompt engineering* Martes 29/4	**Parte 1** Hiperparámetros de los modelos de lenguaje; prompting: Zero-Shot; Few-Shot	Fernando Carranza
	Parte 2 Prompting: RAG, Chain of Thought y otros	Julia Milanese