Prompting and Fine-tuning Pre-trained Generative Language Models

Johny Moreira[1], **Altigran da Silva**[1], **Luciano Barbosa**[2]

[1]Instituto de Computação – Universidade Federal do Amazonas (UFAM)
https://numeros.icomp.ufam.edu.br/altigran

[2]Centro de Informática – Universidade Federal do Pernambuco (UFPE)
https://www.cin.ufpe.br/~luciano

{johny.moreira, alti}@icomp.ufam.edu.br, luciano@cin.ufpe.br

1. Tutorial Type

This is an advanced tutorial planned to last 3 hours.

2. Intended Audience & Prerequisite

The tutorial is targeted to researchers, graduate, and senior undergraduate students, as well as industrial practitioners interested in applying the text generation capabilities of pre-trained Generative Language Models to downstream tasks. The tutorial will be presented in Portuguese on-site by Johny Moreira, one of the authors. Attendees are expected to have a basic understanding and experience in Machine Learning, Deep Learning, Natural Language Processing, and Python.

3. Abstract

There has been an explosion of available pre-trained and fine-tuned Generative Language Models (LM). They vary in the number of parameters, architecture, training strategy, and training set size. Aligned with it, alternative strategies exist to exploit these models, such as Fine-tuning and Prompt Engineering. However, many questions may arise throughout this process: Which model to apply for a given task? Which strategies to use? Will Prompt Engineering solve all tasks? What are the computational and financial costs involved? This tutorial will introduce and explore typical modern LM architectures with a hands-on approach to the available strategies.

4. Tutorial Outline

- **INTRODUCTION (5 min)**
- **Generative LMs & Prompting (1h 25 min)**
  - **Pre-trained Large Language Models (40 min)**: With the abundance of newly introduced variations of Neural Language Models in both literature and industry, it is essential to understand the fundamentals. In this topic, we will briefly introduce the foundations of the Large Language Models (LLMs), mostly presenting the different existing architectures: Encoder [Devlin et al. 2019], Encoder-Decoder [Raffel et al. 2020], Decoder [Brown et al. 2020], and some variations of these models. We will highlight their functionality, pros, and cons in terms of training, evaluation, and deployment of applications based on closed models.

  - **Prompt Engineering (15 min)**: Discuss the Prompt Engineering approach [Liu et al. 2023], introducing In-Context Learning paradigm [Brown et al. 2020], how it can be applied and its limitations [Zamfirescu-Pereira et al. 2023].
  - **Hands-On (30 min)**
- **Break (10 min)**
- **Open-source models and Fine-tuning (1h 20 min)**
  - **Open-source models (30 min)**: The larger number of parameters and the requirement of a large-scale corpus for training aligned with the ever-increasing necessity for computational resources makes the training, evaluation, and deployment of LLMs expensive [Dettmers et al. 2023]. Even exploring closed models through their APIs with prompt engineering has its drawbacks. Hence, this topic discusses state-of-the-art efforts to build open-source models with fewer parameters and less training data [Gururangan et al. 2020, Dey et al. 2023, Hsieh et al. 2023]. These models have been demonstrated not to require too much computational power or large-scale training data to outperform popular closed-model LLMs.
  - **Fine-tuning (20 min)**: Introduce the concept of fine-tuning, showing some applications from the literature that have surpassed traditional tasks [Ding et al. 2023].
  - **Hands-on (30 min)**

5. Bio of Authors

Johny Moreira: Postdoctoral researcher at the Federal University of Amazonas (UFAM) working in partnership with Jusbrasil. PhD in Computer Science, emphasizing Computational Intelligence, at the Federal University of Pernambuco (UFPE). He has also achieved his Master's degree focusing on Databases at the same institution. Bachelor in Information Systems by the Federal Institute of Education, Science, and Technology of Ceará (IFCE). His fields of interest are Natural Language Processing, Language Modeling, Semantic Web, Data Science, and Information Extraction.

Altigran da Silva: Professor at the Instituto de Computação, Universidade Federal do Amazonas (IComp/UFAM), focusing on Data Management, Information Retrieval, and Machine learning. Since earning his doctorate from UFMG in 2002, he has contributed to more than 150 scientific publications and coordinated various research projects. His administrative experience includes roles such as the Dean of Research and Graduate Studies at UFAM and the Deputy Coordinator of the Computing area at CAPES. He also served on the Brazilian Computer Society's (SBC) board and its Advisory Board. As an entrepreneur, he co-founded tech companies, which were later acquired by industry giants. He was recognized for his academic supervision and service, winning awards such as the SBC Distinguished Member Award and the Google Research Awards in Latin America.

Luciano Barbosa: Associate Professor at the Computer Science Department at Universidade Federal de Pernambuco. In addition to his academic experience, he held different positions in industry research labs as Visiting Scholar Researcher at Google AI (formerly Google Research), Research Scientist at IBM Research and AT&T Research Labs; and a Ph.D. summer intern at Yahoo! Research (Europe and USA). His research interests include web mining, text mining, natural language processing, and data analytics.

6. Presenter

The tutorial will be presented in Portuguese on-site by the author Johny Moreira.

7. Support Materials

The tutorial will be assisted by presentation slides and projection of the presenter's screen. Hence, it will be required a multimedia projector. Additionally, the attendees can bring their own computers so they can follow the practices using Google Collaboratory.

8. Acknowledgments

This work was supported by JusBrasil under the UFAM/Jusbrasil Postdoctoral Fellowship granted to Johny Moreira, by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) under funding grant No. 001, by the Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM) under project POSGRAD, and by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) under PQ funding grant to Altigran da Silva (Proc. 307248/2019-4).

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. *CoRR*, abs/2305.14314.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Dey, N., Gosal, G., Chen, Z., Khachane, H., Marshall, W., Pathria, R., Tom, M., and Hestness, J. (2023). Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster. *CoRR*, abs/2304.03208.

Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C., Chen, W., Yi, J., Zhao, W., Wang, X., Liu, Z., Zheng, H., Chen, J., Liu, Y., Tang, J., Li, J., and Sun, M. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nat. Mac. Intell.*, 5(3):220–235.

Gururangan, S., Marasovic, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. R., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8342–8360. Association for Computational Linguistics.

Hsieh, C., Li, C., Yeh, C., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C., and Pfister, T. (2023). Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In Rogers, A., Boyd-Graber, J. L., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8003–8017. Association for Computational Linguistics.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9):195:1–195:35.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., and Yang, Q. (2023). Why johnny can't prompt: How non-ai experts try (and fail) to design LLM prompts. In Schmidt, A., Väänänen, K., Goyal, T., Kristensson, P. O., Peters, A., Mueller, S., Williamson, J. R., and Wilson, M. L., editors, *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, pages 437:1–437:21. ACM.