baselinestretch

GIÁO TRÌNH NLP TOÀN TẬP

Đinh Công Thái

Ngày 30 tháng 9 năm 2025

Mục lục

Lời nói đầu

Xin chào, mình là **Đinh Công Thái**, sinh viên lớp IT1-K67, ngành Khoa học Máy tính, Trường Đại học Bách khoa Hà Nội. Là nghiên cứu sinh tại Viên Nghiên cứu và Ứng dụng Trí tuệ nhân tạo của **PGS.TS Nguyễn Phi Lê**, đồng thời nghiên cứu và làm việc dưới sự hướng dẫn của anh **Nguyễn Tuấn Dũng**.

Vậy cuốn giáo trình này ra đời để làm gì?

Mình viết nó cho những bạn đang mơ hồ, không định hướng, không biết AI bắt đầu từ đâu, nhưng có chút tò mò và muốn thử. Nội dung tập trung vào **Xử lý ngôn ngữ tự nhiên (NLP)**

Sách này không hứa biến bạn thành chuyên gia sau một đêm. Nhưng nó sẽ là một bản đồ để bạn không bị lạc trôi quá lâu, và nếu thấy hứng thú thì có thể đi xa hơn, sớm hơn.

Cách làm của mình khá ra tác phẩm này có thể tóm tắt như sau:

- **Soạn mục lục** → dựng cái khung nhà cho con chatbot.
- **Tìm & gom paper/tài liệu**, để LLM không bịa, đồng thời cũng để cho kiến thức được sâu sắc và bám sát.
- **Chia thư mục theo từng phần/tiểu mục** → giữ cấu trúc rõ ràng.
- **Soạn rule** → chính là *style guide*: văn phong, yêu cầu ví dụ, công thức, hình minh hoạ…
- **Tool 1 (lưu context)** → giúp LLM nhớ liền mạch, không quên mình vừa viết gì ở trang trước.
- **Tool 2 (API xuất bản)** → gửi query theo từng mục và nhận lại nội dung thành file tương ứng.
- **Tool 3 (tự động review & bổ sung)** → LLM đọc lại chính nội dung đã sinh, so với rule + nguồn, rồi nhận xét/gợi ý sửa.

Kết quả cuối cùng: phần kiến thức cốt lõi thì mình đã kiểm tra lại cẩn thận. Còn nếu có vài chỗ đọc hơi lạ lạ thì có thể đó là *hallucination*.

Cấu trúc cuốn sách

- **Phần 1: Lý thuyết NLP** – kể câu chuyện từ những nền tảng cơ bản cho tới các mô hình hiện đại như Transformer. Đọc phần này để biết "tại sao lại như vậy".
- **Phần 2: Thực chiến NLP** – biến lý thuyết thành code, dự án, ứng dụng. Đọc phần này để biết "làm thế nào".

Hai phần này bổ sung cho nhau. Biết lý thuyết mà không thực hành thì dễ nói cho vui. Còn chỉ thực hành mà không hiểu gốc thì cũng dễ trở thành copy-paste developer.

Trân trọng,

Đinh Công Thái

Phần I

BÁCH KHOA TOÀN THƯ VỀ LÝ THUYẾT NLP

Chào mừng bạn đến với nền tảng lý thuyết của Xử lý Ngôn ngữ Tự nhiên. Mục tiêu của Phần 1 là cung cấp một cái nhìn toàn diện, sâu sắc và có hệ thống về TẤT CẢ các khái niệm, kiến trúc và thuật toán lý thuyết đã định hình nên ngành NLP. Chúng ta sẽ bắt đầu từ những nguyên lý ngôn ngữ học cơ bản, đi qua kỷ nguyên thống kê, khám phá các kiến trúc mạng nơ-ron kinh điển, và cuối cùng là làm chủ kiến trúc Transformer - nền tảng của các mô hình ngôn ngữ lớn hiện đại. Mỗi chương là một khối kiến thức độc lập nhưng có sự kết nối chặt chẽ, cùng nhau xây dựng nên một nền tảng vững chắc để bạn có thể hiểu "tại sao"và "như thế nào"các công nghệ NLP hoạt động. Hãy bắt đầu hành trình khám phá này.

KẾT THÚC PHẦN 1

Phần 1 của giáo trình đã cung cấp cho bạn một nền tảng lý thuyết toàn diện, từ những nguyên lý ngôn ngữ học cơ bản, các mô hình thống kê, các kiến trúc nơ-ron kinh điển, cho đến sự thống trị của Transformer và các kỹ thuật tiên tiến để tinh chỉnh, căn chỉnh, và đánh giá các Mô hình Ngôn ngữ Lớn. Với nền tảng này, bạn đã sẵn sàng để bước sang Phần 2, nơi chúng ta sẽ biến những lý thuyết này thành các kỹ năng và quy trình thực chiến để xây dựng các ứng dụng NLP thực tế.

Tóm lược nội dung Phần 1:

- **Chương 1 – Nhập môn và Nguyên lý nền tảng:** Định nghĩa NLP, mối quan hệ với AI, các kỷ nguyên phát triển, kiến thức ngôn ngữ học, toán học, và các vấn đề đạo đức trong NLP.
- **Chương 2 – Biểu diễn văn bản và Mô hình thống kê:** Các kỹ thuật BoW, TF-IDF, N-gram, smoothing, mô hình phân loại, học chuỗi, và LDA.
- **Chương 3 – Các kiến trúc mạng nơ-ron kinh điển:** Word embeddings (Word2Vec, GloVe, FastText), autoencoder, RNN/LSTM/GRU, seq2seq với attention, CNN cho NLP, contextualized embeddings, và các mô hình sinh pre-Transformer.
- **Chương 4 – Kỷ nguyên Transformer và LLMs:** Kiến trúc Transformer gốc, phân loại LLMs, tokenization hiện đại, biến thể cho ngữ cảnh dài, MoE.
- **Chương 5 – Tinh chỉnh và Căn chỉnh mô hình:** Fine-tuning, PEFT (LoRA, QLoRA, Adapter), prompt engineering, instruction tuning, RLHF và các kỹ thuật alignment khác.
- **Chương 6 – Hệ thống và Kiến trúc nâng cao:** RAG, tối ưu hóa mô hình (distillation, quantization, pruning), mô hình đa phương thức, hệ thống tác tử, world models, và biểu diễn nâng cao cho truy xuất.
- **Chương 7 – Đánh giá và Benchmark:** Các metric kinh điển (perplexity, F1, BLEU, ROUGE), benchmark (GLUE, SuperGLUE, đánh giá emergent abilities), thách thức trong đánh giá, và xu hướng “LLM-as-a-Judge”.

Phần II

CẨM NANG KỸ THUẬT NLP THỰC CHIẾN

Chào mừng bạn đến với Phần 2: Cẩm nang kỹ thuật NLP thực chiến. Mục tiêu của Phần này là trang bị cho bạn những kỹ năng, công cụ và quy trình cần thiết để xây dựng, đánh giá và triển khai các ứng dụng NLP trong thực tế. Nếu Phần 1 tập trung vào việc trả lời "tại sao" và "như thế nào" của lý thuyết, thì Phần 2 sẽ hướng dẫn bạn từng bước thực hành: từ thu thập và xử lý dữ liệu, huấn luyện và đánh giá mô hình, tối ưu hóa và triển khai, cho đến việc áp dụng các công thức (recipes) để giải quyết các bài toán phổ biến. Mỗi chương là một hướng dẫn chi tiết, đi từ cơ bản đến nâng cao, giúp bạn hiểu rõ quy trình thực tế và tự tin áp dụng các kỹ thuật NLP trong các dự án thật sự. Hãy sẵn sàng để chuyển từ lý thuyết sang thực chiến.

KẾT THÚC PHẦN 2

Phần 2 đã đưa bạn vào một hành trình thực chiến, biến những lý thuyết phức tạp thành các kỹ năng và sản phẩm cụ thể. Từ những bước đầu tiên trong việc xử lý dữ liệu đến việc huấn luyện và triển khai các mô hình ngôn ngữ lớn, bạn đã được trang bị một bộ công cụ toàn diện để giải quyết các bài toán NLP trong thế giới thực. Phần này không chỉ dạy bạn cách sử dụng các thư viện, mà còn rèn luyện một tư duy MLOps bền vững. Với kiến thức từ cả hai phần, bạn đã có đủ hành trang để tự tin bước đi trên con đường của một chuyên gia NLP.

Tóm lược nội dung Phần 2:

- **Chương 1 – Quy trình làm việc và Công cụ Xử lý Dữ liệu:** Phác thảo vòng đời dự án NLP tinh gọn, các kỹ thuật thu thập dữ liệu (APIs, Web Scraping), làm sạch và gán nhãn (Pandas, Polars, Label Studio, Snorkel, LLM-based), tăng cường dữ liệu, và quản lý phiên bản với DVC.
- **Chương 2 – Huấn luyện và Đánh giá Mô hình:** Làm chủ hệ sinh thái Hugging Face (‘transformers‘, ‘datasets‘, ‘accelerate‘, ‘evaluate‘), xây dựng các hàm đánh giá thực tế (‘compute_metrics‘), theo dõi thí nghiệm (Weights & Biases, MLflow), và huấn luyện phân tán với DeepSpeed.
- **Chương 3 – Tối ưu hóa, Triển khai và Vận hành (MLOps):** Tìm hiểu về Cơ sở dữ liệu Vector, các thư viện tối ưu hóa (‘bitsandbytes‘, ‘peft‘, ONNX), đóng gói và phục vụ mô hình (FastAPI, Docker), các framework chuyên dụng (vLLM, BentoML), và vòng đời MLOps hoàn chỉnh (giám sát, tái huấn luyện).
- **Chương 4 – Công thức Xây dựng các Ứng dụng Phổ biến (Recipes):** Hướng dẫn từng bước để xây dựng các ứng dụng thực tế: Tìm kiếm Ngữ nghĩa, Hệ thống RAG, Fine-tuning cho Phân loại và NER, xây dựng Tác tử đơn giản, và Trích xuất Thông tin có Cấu trúc.

Tài liệu tham khảo

[1] Emily Alsentzer, John R Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, 2019.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[3] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[4] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3615–3620, 2019.

[5] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8003–8014, 2020.

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[7] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. In *Transactions of the Association for Computational Linguistics*, volume 5, pages 135–146, 2017.

[8] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, 2016.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in neural information processing systems 33*, pages 1877–1901, 2020.

[10] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pires, Quoc Le, Yong-wook Lee, Oleksii Kuchaiev, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[11] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014.

[12] Noam Chomsky. Three models for the description of language. *IRE Transactions on information theory*, 2(3):113–124, 1956.

[13] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in neural information processing systems 30*, 2017.

[14] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, , et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[15] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

[16] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[17] Marta R Costa-jussà, James Tran, Artem Sokolov, Machel Dewan, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

[18] Tri Dao, Daniel Y Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *Advances in Neural Information Processing Systems 35*, pages 16344–16359, 2022.

[19] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[21] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, 2018.

[22] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. In *The Journal of Machine Learning Research*, volume 22, pages 1–39, 2021.

[23] Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. Black-box attacks against rnns: The power of gradient-free methods. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, pages 117–130, 2018.

[24] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, 2021.

[25] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[26] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems 27*, 2014.

[27] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[28] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2022.

[29] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[30] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

[31] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[32] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013.

[33] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[34] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, 2018.

[35] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[36] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra S Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[37] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7930–7937, 2020.

[38] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. In *Transactions of the Association for Computational Linguistics*, volume 8, pages 64–77, 2020.

[39] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, 2014.

[40] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[41] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[42] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.

[43] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, 2018.

[44] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning (ICML 2001)*, 2001.

[45] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

[46] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chang Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. volume 36, pages 1234–1240, 2020.

[47] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[48] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems 27*, 2014.

[49] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[50] Raymond Li, Loubna Allal, Yangtian Zi, Leslie Tow, Mirac Briesch, Yumo Gu, Carl Akiki, Zhaowei Mou, Manar Naila, Denis Kocetkov, et al. Starcoder: may the source be with you! *arXiv preprint arXiv:2305.06161*, 2023.

[51] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, 2021.

[52] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. In *arXiv preprint arXiv:1907.11692*, 2019.

[53] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421, 2015.

[54] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[55] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7213–7224, 2020.

[56] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations (ICLR)*, 2013.

[57] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems 26*, 2013.

[58] Jonas Mueller and Aditya Thyagarajan. Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2786–2792, 2016.

[59] Dat Quoc Nguyen and Anh Tuan Nguyen. Phobert: Pre-trained language models for vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, 2020.

[60] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[61] OpenAI. Gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. Accessed: 2024-05-15.

[62] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35*, pages 27730–27744, 2022.

[63] Bowen Peng, Jeffrey Quesnelle, Angsheng Fan, and Ayush Aneja. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.

[64] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[65] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, 2018.

[66] Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2022.

[67] Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2401–2410, 2020.

[68] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[69] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[70] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[71] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

[72] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. In *The Journal of Machine Learning Research*, volume 21, pages 1–67, 2020.

[73] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992, 2019.

[74] Sascha Rothe, Shashi Narayan, and Ali Seifert. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280, 2020.

[75] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xian Tan, Guillaume Lample, Côme Grand, Thomas Lavril, Marie-Anne Lachaux, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

[76] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[77] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019*, 2019.

[78] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2009.

[79] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[80] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[81] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, 2016.

[82] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[83] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.

[84] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems 33*, pages 16857–16867, 2020.

[85] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

[86] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[87] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and John Schulman. Learning to summarize from human feedback. In *Advances in Neural Information Processing Systems 33*, pages 3008–3021, 2020.

[88] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunbo Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

[89] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems 27*, 2014.

[90] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Stanford Center for Research on Foundation Models. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

[91] Yi Tay, Mostafa Dehghani, Vinh Q Tran, Siyuan Hsieh, et al. Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*, 2022.

[92] The BLOC team. Ntk-aware scaled rope allows llama models to have extended (8k) context length without any fine-tuning. https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have/, 2023.

[93] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[94] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems 30*, 2017.

[95] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[96] Liang Wang, Nan Yang, Fnu Fariha, Hao Shen, and Haolan Liu. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

[97] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33*, pages 5776–5788, 2020.

[98] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[99] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

[100] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[101] Terry Winograd. Understanding natural language. *Cognitive psychology*, 3(1):1–191, 1972.

[102] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. In *arXiv preprint arXiv:1609.08144*, 2016.

[103] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022.

[104] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems 32*, 2019.

[105] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Sha, Silvio Savarese, and Anima Anandkumar. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

[106] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems 33*, pages 17283–17297, 2020.

[107] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR, 2020.

[108] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*, 2019.

Lời kết

Thế là ta đã đi hết một vòng Xử lý Ngôn ngữ Tự nhiên. Xin chúc mừng

---

Con đường phía trước

Phần 1 Đừng coi "Bách khoa Toàn thư về Lý thuyết" là thứ đọc xong rồi xếp xó. Nó chính là cái nền để bạn dựng mọi mô hình sau này. Khi có một paper mới ra với tên nghe như thần chú, quay lại phần này: bạn sẽ thấy bản chất nó vẫn xoay quanh Attention, học biểu diễn, và mấy định luật xác suất quen thuộc. Muốn làm pro thật sự thì không chỉ biết xài tool, mà còn phải hiểu vì sao nó chạy được.

Phần 2 "Cẩm nang Thực chiến" đưa công thức cho bạn, nhưng công thức thì sinh ra là để phá. Hãy nghịch: đổi mô hình, đổi dữ liệu, phá optimizer, thậm chí làm sai cũng được. Mỗi lần máy báo lỗi, hãy coi đó là một mentor thầm lặng. Không giáo trình nào dạy nhanh bằng cách tự fix một đống bug.

Cuối cùng, đừng quên trách nhiệm. NLP không chỉ là code chạy đúng hay loss xuống đẹp. Những mô hình bạn build hôm nay có thể ảnh hưởng đến cách con người giao tiếp, tiếp nhận thông tin. Thế nên, ngoài chuyện optimize cho nhanh, hãy luôn nghĩ đến bias, fairness, và impact. Làm kỹ sư NLP nên có trách nhiệm đi kèm.

Cảm ơn bạn đã đồng hành tới cuối chặng đường này. Cuốn sách thì đóng lại, nhưng bug thì vẫn còn nhiều, và hành trình học của bạn mới chỉ bắt đầu thôi. Chúc bạn vừa code giỏi, vừa ngủ đủ giấc!