Language Model Zoo: schedule

Summer 2025

Apr 28 Introduction & n-grams

Apr 30 n-gram language models (practical intro)

May 05 "Finding structure in time" (Elman, 1990)
This paper introduces Simple RNNs, and discusses a number of experiments with them. Many of the intuitions/observations from this paper are still valid for present neural LMs.

May 07 "Linguistic Regularities in Continuous Space Word Representations" (Mikolov et al., 2013b) – with a short excursion to Mikolov et al. (2013a)
This is one of the papers that re-introduces RNN language models, in the early times of the revival of neural networks that lead to today's LMs. The paper is from the author(s) of well known word2vec, allowing some interesting comparisons.

May 12 "Deep Contextualized Word Representations" (Peters et al., 2018)
A rather successful RNN LM, which also shows their usage in a set of downstream tasks.

May 12 "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" (Devlin et al., 2019)
This is the BERT paper – no need for introduction.

May 19 *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (Liu et al., 2019)
RoBERTa is is architecturally almost the same as BERT, but the paper demonstrates a more systematic set of experiments on pre-training encoder-only language models.

May 21 ? "DeBERTa: Decoding-enhanced BERT with Disentangled Attention" (He et al., 2021)
'Interesting' part of DeBERTa is rather a simple manipulation, but at its time it showed strong downstream performance. If we need time, this is possibly one of the papers that we can skip.

May 26 *Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference* (Warner et al., 2024)
This is a 'modern' take on encoder models.

May 28 ? "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators" (K. Clark et al., 2020)
This model includes some changes to encoder model the pretraining, that may introduce a few interesting ideas, but again this could be replaced with a more interesting one, or possibly skipped if we are out of time.

Jun 02 “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension” (Lewis et al., 2020)
This is one of the few encoder–decoder models around.

Jun 04 “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer” (Raffel et al., 2020)
Another, perhaps more popular, encoder–decoder model, with a nicely written paper with lots of useful information.

Jun 09 “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter” (Sanh et al., 2019)
This is a ‘distilled’ model, where the idea is compressing a larger pretrained model to a smaller and faster model to reduce the computational costs during the inference time.

Jun 11 “Findings of the Second BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora” (M. Y. Hu et al., 2024)
This is the summary of a shared task on building ‘BabyLM’s. The share task intends to promote ‘data efficient’ models, with some motivation of ‘cognitively plausible’ learning.

Jun 16 “Unsupervised Cross-lingual Representation Learning at Scale” (Conneau et al., 2020) – may require some excursions to Conneau and Lample (2019)
This paper describes a multilingual decoder-only model (based on RoBERTa).

Jun 18 “mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer” (Xue et al., 2021)
This is another multilingual model, but this time an encoder–decoder model following T5.

Jun 30 “Language Models are Few-Shot Learners” (Brown et al., 2020) – with some parts from Radford et al. (2019) and Radford et al. (2018)
The paper we will read (if we do not change) is the GPT-3 paper, but we will likely also refer to earlier GPT papers.

Jul 02 *The Llama 3 Herd of Models* (Meta AI, 2024) – also Touvron et al. (2023)
This is the Llama 3 paper. In general, this is a very long, but quite informative paper. Also reports on steps after LM-training, extensive experiments and also multimodal extensions (which we may leave for the final week).

Jul 07 *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model* (Le Scao et al., 2023)
This is another long LLM paper. The main reason for choice is its focus on multilinguality, which is often not the focus of most LLMs.

Jul 09 “Improving language models by retrieving from trillions of tokens” (Borgeaud et al., 2021) – or maybe “Scaling Language Models: Methods, Analysis & Insights from Training Gopher” (Rae et al., 2021)
This is yet another long LLM paper. The main reason for choice the documentation of the ‘scaling effects’, which is more thoroughly described in the second paper.

Jul 14 “LoRA: Low-rank adaptation of large language models” (E. J. Hu et al., 2022)
This is a popular method for ‘parameter efficient’ fine-tuning.

Jul 16 “Parameter-efficient transfer learning for NLP” (Houlsby et al., 2019) and/or maybe “Prefix-Tuning: Optimizing Continuous Prompts for Generation” (X. L. Li and Liang, 2021)
Two more parameter efficient methods, we’ll likely go for only one of these.

Jul 21 “FLAVA: A foundational language and vision alignment model” (Singh et al., 2022)
This is a vision–language model. Perhaps not the most popular, but the paper is easier to follow and informative compared to others. (like the more popular CLIP paper Radford et al., 2021)

Jul 23 “wav2vec 2.0: A framework for self-supervised learning of speech representations” (Baevski et al., 2020)
Somewhat aged, but this is one of the first modern speech models.

References

Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli (2020). “wav2vec 2.0: A framework for self-supervised learning of speech representations”. In: *Advances in neural information processing systems* 33, pp. 12449–12460.

Borgeaud, Sebastian, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre (2021). “Improving language models by retrieving from trillions of tokens”. In: *CoRR* abs/2112.04426. URL: https://arxiv.org/abs/2112.04426.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). “Language Models are Few-Shot Learners”. In: *CoRR* abs/2005.14165. URL: https://arxiv.org/abs/2005.14165.

Clark, Kevin, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning (2020). “ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators”. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=r1xMH1BtvB.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2020). “Unsupervised Cross-lingual Representation Learning at Scale”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. URL: https://aclanthology.org/2020.acl-main.747/.

Conneau, Alexis and Guillaume Lample (2019). “Cross-lingual language model pretraining”. In: *Advances in neural information processing systems* 32.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). “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)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio.

Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423/.

Elman, Jeffrey L. (1990). "Finding structure in time". In: *Cognitive Science* 14, pp. 179–211. DOI: 10.1016/0364-0213(90)90002-E.

He, Pengcheng, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen (2021). "DeBERTa: Decoding-enhanced BERT with Disentangled Attention". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=XPZIaotutsD.

Houlsby, Neil, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly (2019). "Parameter-efficient transfer learning for NLP". In: *International conference on machine learning*. PMLR, pp. 2790–2799.

Hu, Edward J, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. (2022). "LoRA: Low-rank adaptation of large language models". In: *ICLR* 1.2, p. 3.

Hu, Michael Y., Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox (Nov. 2024). "Findings of the Second BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora". In: *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*. Ed. by Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Leshem Choshen, Ryan Cotterell, Alex Warstadt, and Ethan Gotlieb Wilcox. Miami, FL, USA: Association for Computational Linguistics, pp. 1–21. URL: https://aclanthology.org/2024.conll-babylm.1/.

Le Scao, Teven et al. (2023). *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*. arXiv: 2211.05100 [cs.CL]. URL: https://arxiv.org/abs/2211.05100.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer (July 2020). "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. URL: https://aclanthology.org/2020.acl-main.703/.

Li, Xiang Lisa and Percy Liang (Aug. 2021). "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)*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, pp. 4582–4597. DOI: 10.18653/v1/2021.acl-long.353. URL: https://aclanthology.org/2021.acl-long.353/.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv: 1907.11692 [cs.CL]. URL: https://arxiv.org/abs/1907.11692.

Meta AI (2024). *The Llama 3 Herd of Models*. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013a). "Efficient Estimation of Word Representations in Vector Space". In: *CoRR* abs/1301.3781. URL: http://arxiv.org/abs/1301.3781.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (June 2013b). "Linguistic Regularities in Continuous Space Word Representations". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff. Atlanta, Georgia: Association for Computational Linguistics, pp. 746–751. URL: https://aclanthology.org/N13-1090/.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018). "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)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: https://aclanthology.org/N18-1202/.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. (2021). "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PmLR, pp. 8748–8763.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever (2018). *Improving Language Understanding by Generative Pre-Training*. URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. (2019). *Language models are unsupervised multitask learners*. OpenAI blog. URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Rae, Jack W., Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving (2021). "Scaling Language Models: Methods, Analysis & Insights from Training Gopher". In: *CoRR* abs/2112.11446. URL: https://arxiv.org/abs/2112.11446.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21.140, pp. 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *CoRR* abs/1910.01108. URL: http://arxiv.org/abs/1910.01108.

Singh, Amanpreet, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela (2022). "FLAVA: A foundational language and vision alignment model". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15638–15650.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom (2023). "Llama 2: Open Foundation and Fine-Tuned Chat Models". In: *CoRR* abs/2307.09288. URL: `https://doi.org/10.48550/arXiv.2307.09288`.

Warner, Benjamin, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli (2024). *Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference*. arXiv: `2412.13663 [cs.CL]`. URL: `https://arxiv.org/abs/2412.13663`.

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel (June 2021). "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou. Online: Association for Computational Linguistics, pp. 483–498. DOI: `10.18653/v1/2021.naacl-main.41`. URL: `https://aclanthology.org/2021.naacl-main.41/`.