Large Language Models

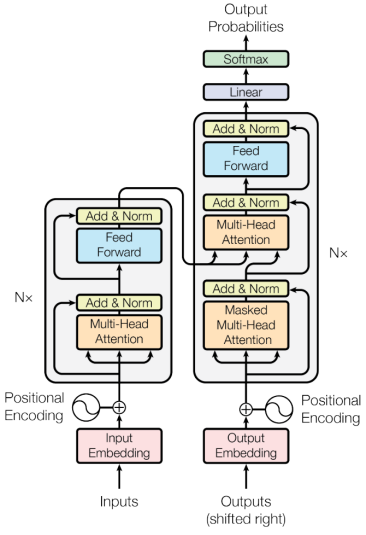
Part 2: Decoder-based language models

David Samuel

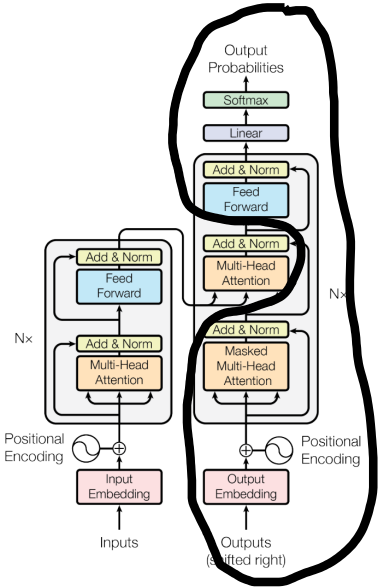
Language Technology Group, University of Oslo



Decoder language models



Decoder language models



Decoder language models

- Training objective: next word prediction
- Alternative names: causal language models (CLMs), generative LMs, autoregressive LMs, left-to-right LMs, GPT-like LMs or simply “language models”
- Nothing surprising here, these are the traditional language models introduced by Claude E. Shannon (1948)

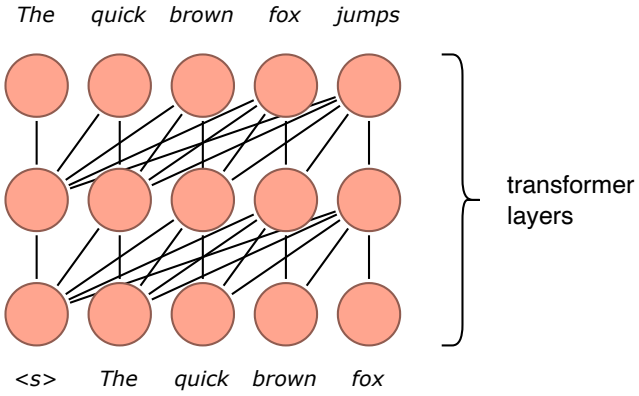
$$P(w_0 w_1 \dots w_n) = \prod_{i=0}^{n} P(w_i | w_0 w_1 \dots w_{i-1})$$

- We want to estimate the probability distribution of sequences of tokens

GPT

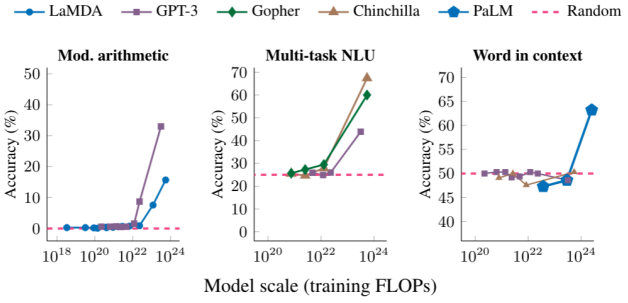
- A giant leap forwards in history: Bengio et al. (2001) proposed *"A Neural Probabilistic Language Model"*
- Another huge jump forward – the same thing, but now with transformers: *"Improving Language Understanding by Generative Pre-Training"* by Radford et al. (2018) – the first GPT
  - The idea here is to train a transformer on language modeling and then fine-tune the whole thing on a downstream task
- Same architecture in GPT-2, just bigger (Radford et al., 2019)
  - Now the model starts to solve* downstream tasks without any fine-tuning!
    - * better than random prediction

Causal language model



In-context learning

- The same thing repeated with GPT-3 (Brown et al., 2020)
  - When the size of the model grows even more (0.1B → 1.5B → 175B), the model starts to be really good in 'zero-shot' and 'few-shot' learning *(more on that later)*
  - So-called "emergent abilities" (Wei et al., 2022) – *be careful, this is a controversial term! :)*



In-context learning

- A double-edged sword for decoder language models
- Their main advantage, otherwise they perform poorly in the pre-training / fine-tuning paradigm
- Therefore they have to be huge and expensive to be somewhat useful
- In practice: is it cheaper to annotate a dataset or to infer from a huge decoder LM?
  - A small fine-tuned LM usually performs better than a huge non-fine-tuned decoder LM (Bang et al., 2023)

Training-time scaling

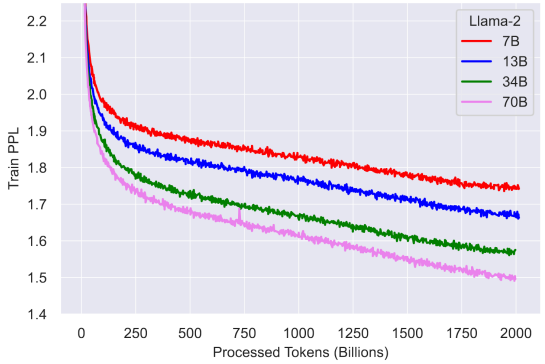


Figure 5: Training Loss for Llama 2 models. We compare the training loss of the Llama 2 family of models. We observe that after pretraining on 2T Tokens, the models still did not show any sign of saturation.

Figure 1: Screenshot from Llama 2 (Touvron et al., 2023)

Training-time scaling (the end?)



Figure 2: Ilya Sutskever at NeurIPS 2024

What about Norwegian?

- If training on English hits the wall, what about much smaller languages?
- The scaling laws are ruthless (Hoffmann et al., 2022)
- Fortunately, we can piggy-back on existing English (mostly) language models
  - Another example of transfer learning
- Repeating the available data is also okay (Muennighoff et al., 2023)
- And we can also add texts from related languages into the mix
- As an example, see our NorMistral-11B (Samuel et al., 2025) :)

References I

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., and Fung, P. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Bengio, Y., Ducharme, R., and Vincent, P. (2001). A neural probabilistic language model. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., 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., 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*, volume 33, pages 1877–1901. Curran Associates, Inc.

Claude E. Shannon (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.

References II

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Vinyals, O., Rae, J. W., and Sifre, L. (2022). An empirical analysis of compute-optimal large language model training. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Muennighoff, N., Rush, A. M., Barak, B., Scao, T. L., Tazi, N., Piktus, A., Pyysalo, S., Wolf, T., and Raffel, C. (2023). Scaling data-constrained language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training. Technical report, OpenAI.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. Technical report, OpenAI.

Samuel, D., Mikhailov, V., Velldal, E., Øvrelid, L., Charpentier, L. G. G., Kutuzov, A., and Oepen, S. (2025). Small languages, big models: A study of continual training on languages of norway.

References III

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.