Large Language Models

Part 2: Decoder-based language models

David Samuel

Language Technology Group, University of Oslo

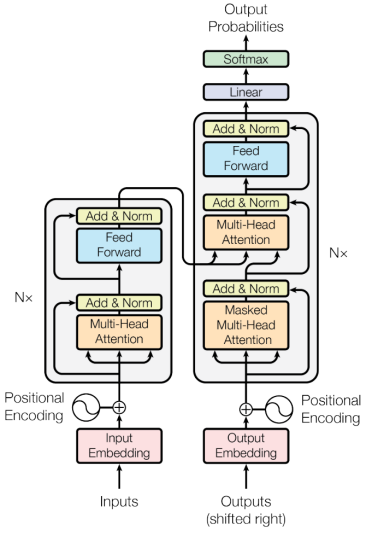


Table of contents

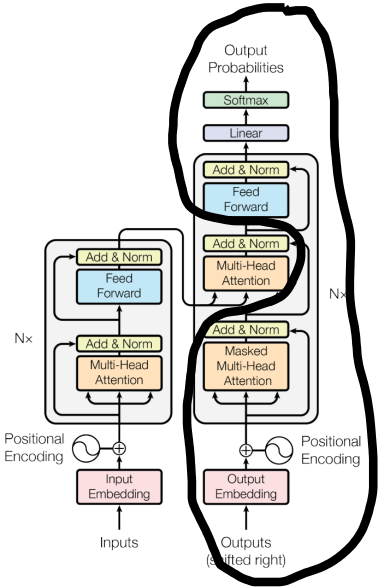
1. Decoder language models

2. References

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 (very popular term, unfortunately) or simply "language models"
- Nothing surprising here, these are the traditional language models introduced by Claude E. Shannon (1948)

$$P(w_{0:n}) = \prod_{i=0}^{n} P(w_i | w_{0:i-1})$$

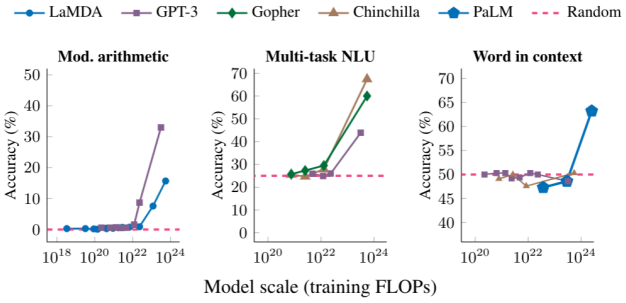
- We want to estimate the probability distribution of sequences of words (or tokens in general)

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 CLM 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

Emergent abilities

- The same thing repeated with GPT-3 (Brown et al., 2020)
  - When the size of the model grows even more (0.1B $\rightarrow$ 1.5B $\rightarrow$ 175B), the model starts to be really good in zero-shot and few-shot
  - So-called “emergent abilities” (Wei et al., 2022)



Emergent abilities

- Now, we have GPT-4 (OpenAI, 2023)
  - but: *"Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."*
  - we can't really say anything about that model, this is no longer scientific research ¯_(ツ)_/¯

Emergent abilities

GPT-4 Technical Report

OpenAI*

Abstract

we used python



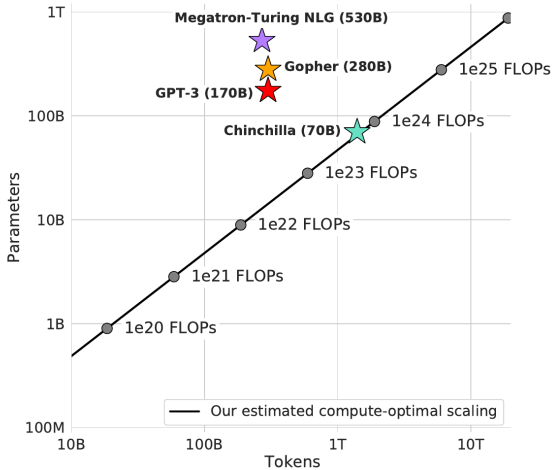
Emergent abilities

- 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)

GPT and its friends

- A lot of media attention in this field so big-tech invests a lot to train these huge models
- PaLM (Chowdhery et al., 2022): Ha, we trained a 540B LM and it's slightly better!
- Chinchilla (Hoffmann et al., 2022): Ha, we can do better with just 70B parameters!
  - Not only that, they showed the compute-optimal training settings
  - Turns out we are bottlenecked by data, not by model size!
  - Sadly (?), we will not get GAI just by scaling up text transformers, the Internet simply doesn't contain enough text

GPT and its friends



GPT and its open friends

- The scientific contributions of these models are questionable
  - the papers are usually not peer-reviewed or even don't exist at all (ChatGPT)
  - the models are not released and it's thus impossible to reproduce and verify the results
- Still, there are some initiatives and research labs that try to change that
  - GPT-J (Wang and Komatsuzaki, 2021)
  - OPT (Zhang et al., 2022)
  - LLaMA (Touvron et al., 2023)
  - ...and many more

Table of contents

1. Decoder language models

2. References

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.

References II

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. (2022). Palm: Scaling language modeling with pathways.

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

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. v. d., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. (2022). Training compute-optimal large language models.

OpenAI (2023). Gpt-4 technical report. Accessed on March 13, 2023.

References III

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

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

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). Llama: Open and efficient foundation language models.

Wang, B. and Komatsuzaki, A. (2021). GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. *https://github.com/kingoflolz/mesh-transformer-jax*.

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.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. (2022). Opt: Open pre-trained transformer language models.