Large Language Models

Part 1: Encoder-based and encoder-decoder-based language models

David Samuel

Language Technology Group, University of Oslo



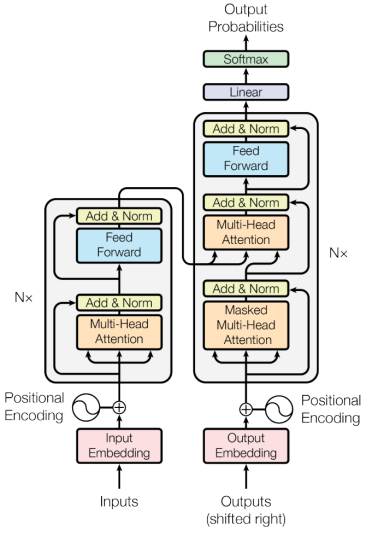
Table of contents

1. Encoder language models

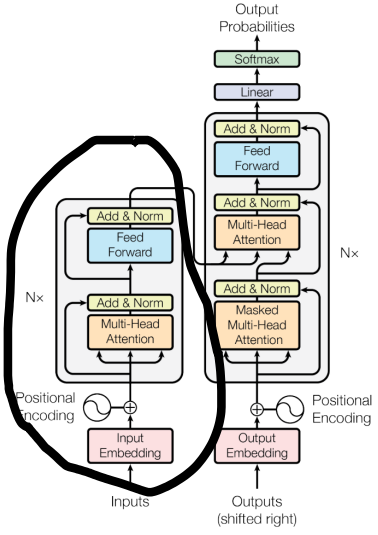
2. Encoder-decoder language models

3. References

Encoder language models



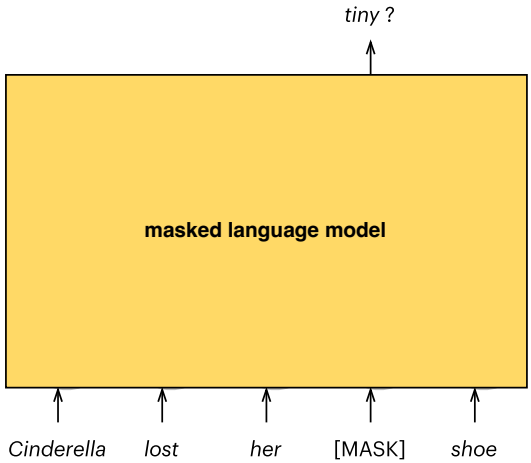
Encoder language models



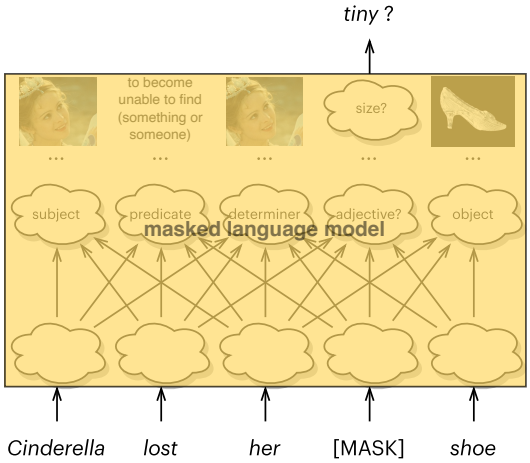
Encoder language models

- Training objective: fill in the blanks
  - "*Cinderella lost her ____ shoe.*"
- Alternative names: masked language models (MLMs), bidirectional LMs or BERT-like LMs
- We want to process text in both directions
- This results in strong language representation by the encoder
  - Then the encoder can be easily fine-tuned and generally performs better than a decoder language model
- However, the encoder doesn't directly estimate $P(w_{0:n})$ and doesn't generate text out of the box
  - Generating is possible but very impractical: Wang and Cho (2019)

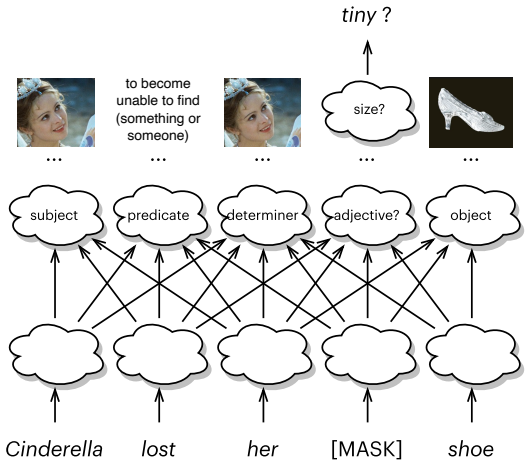
Representation learning with encoder language models



Representation learning with encoder language models



Representation learning with encoder language models





Representation learning with encoder language models

- BERT Rediscovers the Classical NLP Pipeline (Tenney et al., 2019)
  - MLMs work, layer-by-layer, similarly to a traditional NLP pipeline - first they processes the morphology, then the syntax, then the semantics
- A whole field of BERTology (Rogers et al., 2020), which investigates the inner workings of LMs – almost like a neuroscientist who studies brain

BERT and its friends

- Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2019)
- Unlike ELMo, BERT is *deeply* bidirectional and transformer-based language model
  - This improvement led to unprecedented downstream performance back in 2018

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

BERT and its friends

- BERT is pre-trained on two objectives at the same time: masked language modeling (MLM) and next sentence prediction (NSP)
  - The input: *[CLS] text$_1$… [SEP] text$_2$… [SEP]*
  - The task: predict if text$_2$ follows text$_1$ or not, using the last hidden representation of the *[CLS]* token
  - The idea is to learn a good sentence representation, not just token representations
- BERT is pre-trained on a corpus of English books and on English Wikipedia, a relatively modest amount of text in today's standards

BERT and its friends

- RoBERTa (Liu et al., 2019) polished BERT and put it on another level
  - Lots of small improvements, nothing groundbreaking
  - Main findings: BERT is underfitted, it should be trained on much more data for many more steps, and next sentence prediction is unnecessary
  - Pre-trained on English CommonCrawl (filtered), in addition to the BERT corpora
  - Still a very strong baseline, many years after its publication!
- Many more *BERTs followed, bringing another improvements here and there; for example DeBERTa (He et al., 2020), which improves BERT by using relative positional embeddings

BERT and its friends

- So far, we only covered English MLMs, but there are lots of multilingual and non-English monolingual models
- The first one is mBERT, multilingual BERT pre-trained on 104 languages
  - Noteworthy because it showed that it is possible to learn on this many languages
  - Otherwise, it is a rather weak baseline
- XLM-R (Conneau et al., 2020), multilingual RoBERTa, is a much better model – often outperforming monolingual MLMs
- Monolingual MLMs now cover many languages, including Norwegian :)
  - For example NorBERT (Kutuzov et al., 2021) trained by our group

BERT and its friends

- ELECTRA (Clark et al., 2020) is one of the more interesting alternatives to MLMs
  - a more efficient pre-training by using two encoders: generator and discriminator
  - the generator tries to fool the discriminator
  - Difficult to set up properly in practice

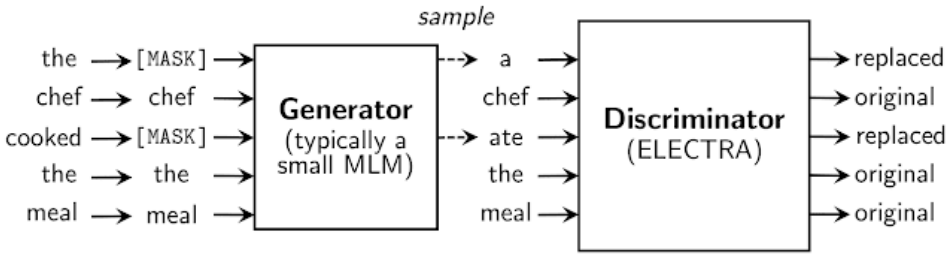


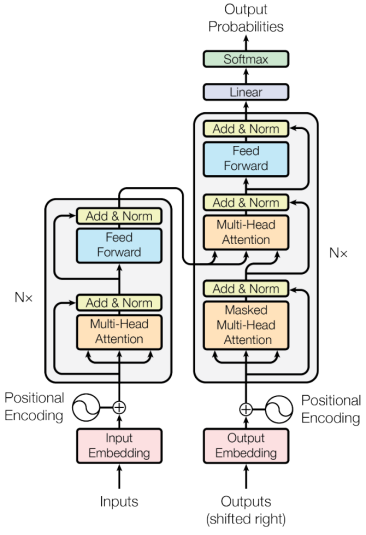
Table of contents

1. Encoder language models

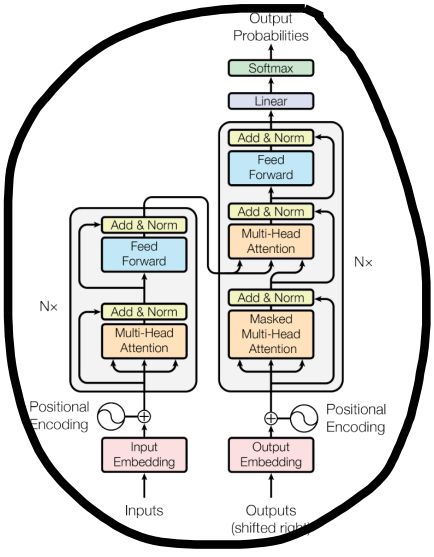
2. Encoder-decoder language models

3. References

Encoder-decoder language models

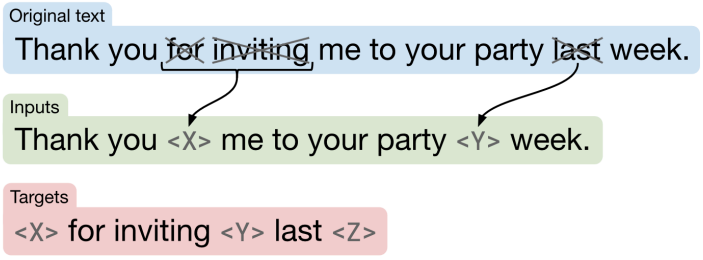


Encoder-decoder language models



Encoder-decoder language models

- Training objective: fill in the blanks, autoregressively



- The training objective combines MLM and CLM together
- We mask a span of tokens, replace it with a single MASK token, put the text into encoder and then let decoder generate all tokens in the masked span
- The resulting LM can be used as a strong representation learner as well as a

T5 family

- Text-to-text Transfer Transformers (T5; Raffel et al., 2019)
  - Also multilingual mT5 (Xue et al., 2021) and byte-level ByT5 (Xue et al., 2022)
- A rule of a thumb: encoder-decoder models are worse than MLMs on fine-tuned tasks and better than CLMs on generative tasks, when comparing models of similar scale
  - However, the encoder-decoder architecture is more complicated to train and to scale up
  - Thus, CLMs beat encoder-decoder models by sheer amount of parameters

Alternative training objectives

- Because this type of models is more versatile, more training options exist
- BART (Lewis et al., 2020) is trained with an alternative take on autoregressive MLM
- UL2: Unifying Language Learning Paradigms by Tay et al. (2022) further merges CLM with autoregressive MLM, leading to better zero-shot performance

Table of contents

1. Encoder language models

2. Encoder-decoder language models

3. References

References I

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistic*s, pages 8440–8451, Online. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

He, P., Liu, X., Gao, J., and Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention.

Kutuzov, A., Barnes, J., Velldal, E., Øvrelid, L., and Oepen, S. (2021). Large-scale contextualised language modelling for Norwegian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 30–40, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

References II

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (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*, pages 7871–7880, Online. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Tay, Y., Dehghani, M., Tran, V. Q., Garcia, X., Wei, J., Wang, X., Chung, H. W., Shakeri, S., Bahri, D., Schuster, T., Zheng, H. S., Zhou, D., Houlsby, N., and Metzler, D. (2022). Ul2: Unifying language learning paradigms.

Tenney, I., Das, D., and Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

References III

Wang, A. and Cho, K. (2019). Bert has a mouth, and it must speak: Bert as a markov random field language model.

Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., and Raffel, C. (2022). ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (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*, pages 483–498, Online. Association for Computational Linguistics.