Large Language Models and Data Streams

Seminar *Data Stream Management and Analysis*

June 25, 2024

Silyu Li

Computer Science 5 - Information Systems and Databases

RWTH AACHEN UNIVERSITY

Background

relevant concepts[a]

- LLM and AI have become very hot topics in recent years.

[a] Source: urlhttps://www.cortical.io/blog/chatgpt-and-large-language-models-the-holy-grail-of-enterprise-ai/

Background

relevant concepts[a]

- LLM and AI have become very hot topics in recent years.
- Various models have shown their wide usage and significant competence in many fields.

[a] Source: urlhttps://www.cortical.io/blog/chatgpt-and-large-language-models-the-holy-grail-of-enterprise-ai/

Background

relevant concepts[a]

- LLM and AI have become very hot topics in recent years.
- Various models have shown their wide usage and significant competence in many fields.
- QA, content generation, translation, text classification etc [1].

[a] Source: urlhttps://www.cortical.io/blog/chatgpt-and-large-language-models-the-holy-grail-of-enterprise-ai/

Background

relevant concepts[a]

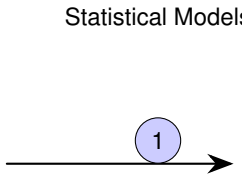
- LLM and AI have become very hot topics in recent years.
- Various models have shown their wide usage and significant competence in many fields.
- QA, content generation, translation, text classification etc [1].
- Most models are ”static” [2].

[a] Source: urlhttps://www.cortical.io/blog/chatgpt-and-large-language-models-the-holy-grail-of-enterprise-ai/

Content of the presentation

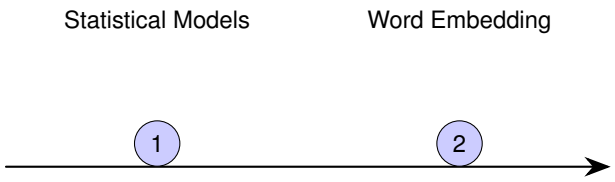
- The evolution of language models and the techniques behind them.
- The training process of several LLM models.
- The definition and application of data streams.
- The need, benefits, challenges and use cases of combining LLM with data streams.

Large Language Models



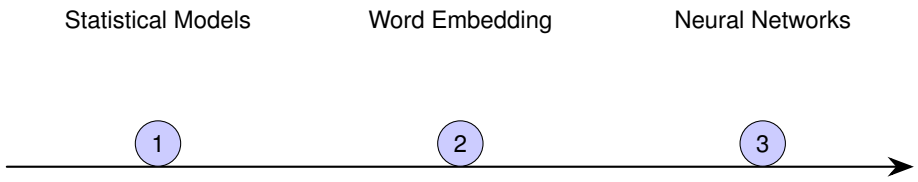
- N-Gram [3]
- Completely statistical.
- Calculate the next word's probability based on the N previous sub-words.
- Poor performance on large documents.

Large Language Models



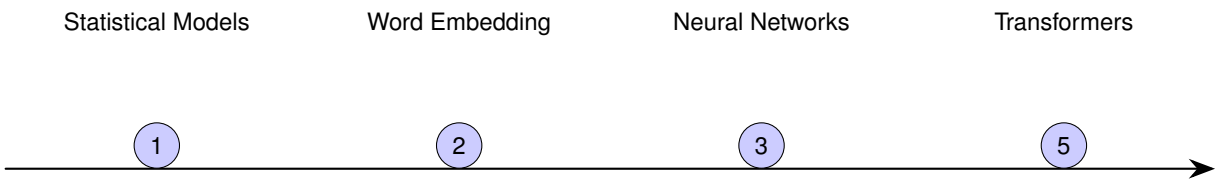
- Word2Vec [4]
- Map words into vector space.
- Similar words have a closer distance.
- Better performance on large documents.

Large Language Models



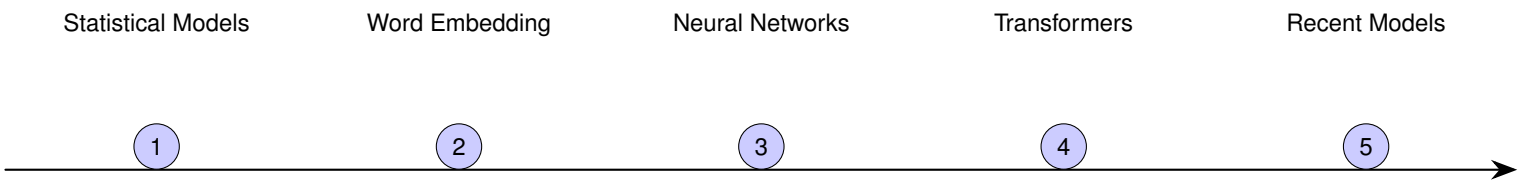
- Seq2Seq [1]
- Works with Long Short-Term Memory(LSTM).
- Encodes input into vectors with fixed dimensionality and decodes them.
- Much Better performance on large documents, can work with different input lengths.
- But still vanishing gradient problems on very large documents.

Large Language Models



- Transformer [2]
- Works with the self-attention mechanism.
- Can process the entire input sentence simultaneously with the help of positional encoding.
- Very good performance on large documents, more efficient and powerful.

Large Language Models



- BERT, ChatGPT, LLaMA. . .
- Based on the transformer architecture.
- Trained on massive datasets and fine-tuned with downstream tasks.

Tokenization

- Purpose: Segmenting input text into tokens.
- WordPiece (BERT) [4]
- Byte Pair Encoding (GPT2, LLaMA) [1]

Pre-training

Given an unlabeled corpus of tokens $U = \{u_1, \ldots, u_n\}$ as a training dataset, the core idea of the pre-training phase is to predict the next token u_i for a sequence $\{u_{i-k}, \ldots, u_{i-1}\}$, specifically by maximizing the likelihood:

$$L_1(U) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)[2] \tag{1}$$

k refers to the context window size and Θ is the parameter of the neural network with which the conditional probability P is modeled.

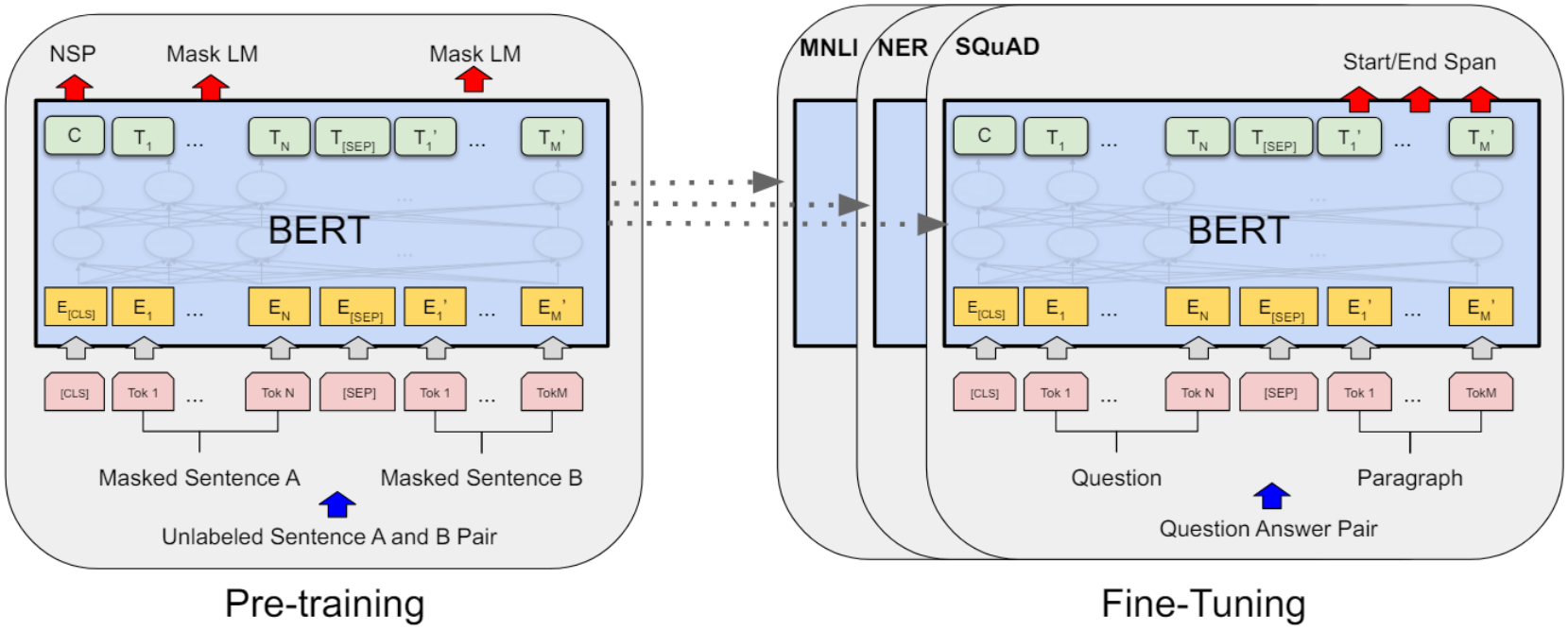
Fine-tuing

Given a labeled dataset C, where each instance of C has a sequence of tokens $\{c^1, \ldots, c^m\}$ and a label y, the goal of the fine-tuning phase is to maximize the following likelihood:

$$L_2(U) = \sum_{(x,y)} \log P(y|x^1, \ldots, x^m)[2] \tag{2}$$

Training process of LLMs

BERT



Training Process in the BERT model [3]

A comparison of different models

Model Name	BERT	GPT-2	LLaMA	GPT-3,5/ 4
Developer	GoogleAI	OpenAI	MetaAI	OpenAI
Release Date	2018	2019	2023	2022/2023
Nr. of Parameters	110 M/ 340 M	1,5 B	7-65 B	175 B/1,7 T
Training Data	Wikipedia(en) & BookCorpus	WebText	Various open-source datasets	WebText
Open-sourced	Yes	No	Yes	No
Major Applications?	QA	Text generation & Translation	Text generation & QA	Content generation & QA

A comparison of different LLM models

Data Stream

Definition 1

A data stream S is an unbounded, potentially infinite multiset of data stream elements (s, τ), where $\tau \in \mathbb{T}$. $\mathbb{T}$ is a timestamp attribute with values from a monotonic, infinite time domain $\mathbb{T}$ with discrete time units. [3]

Limitation of pre-trained LLMs

- Lack of knowledge beyond the scope of their training datasets.
- The performance is likely to gradually degrade over time [4].
- Extremely computationally expensive to be re-trained.
- Not able to process data streams as input.

Challenges

Application Fields

Prompt engineering

Text goes here

Continual learning

Text goes here

Conclusion

Text goes here

References

- Liu, Yiheng, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He et al. ”Summary of chatgpt-related research and perspective towards the future of large language models.” Meta-Radiology (2023): 100017.
- Gupta, Kshitij, Benjamin Thérien, Adam Ibrahim, Mats L. Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. ”Continual Pre-Training of Large Language Models: How to (re) warm your model?.” arXiv preprint arXiv:2308.04014 (2023).
- Cavnar, William B., and John M. Trenkle. ”N-gram-based text categorization.” In Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval, vol. 161175, p. 14. 1994.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. ”Distributed representations of words and phrases and their compositionality.” Advances in neural information processing systems 26 (2013).

References

- Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. ”Sequence to sequence learning with neural networks.” Advances in neural information processing systems 27 (2014).
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. ”Attention is all you need.” Advances in neural information processing systems 30 (2017).
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. ”Bert: Pre-training of deep bidirectional transformers for language understanding.” arXiv preprint arXiv:1810.04805 (2018).
- Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun et al. ”Google’s neural machine translation system: Bridging the gap between human and machine translation.” arXiv preprint arXiv:1609.08144 (2016).

References

- Sennrich, Rico, Barry Haddow, and Alexandra Birch. ”Neural machine translation of rare words with subword units.” arXiv preprint arXiv:1508.07909 (2015).
- Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. ”Improving language understanding by generative pre-training.” (2018).
- Geisler, Sandra. ”Data stream management systems.” In Dagstuhl Follow-Ups, vol. 5. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2013.
- Shi, Haizhou, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, and Hao Wang. ”Continual Learning of Large Language Models: A Comprehensive Survey.” arXiv preprint arXiv:2404.16789 (2024).