Large Language Models and Data Streams

Silyu Li

RWTH Aachen University, Aachen, Germany

Abstract. Large language models, such as ChatGPT, have become widely recognized and are extensively utilized across various domains. However, these models are typically trained on static datasets, lacking updates to new data beyond their initial training set. To enable models that can continuously update themselves based on incoming data, it is necessary to have large language models trained and updated on input data streams. In this paper, we begin by outlining the structure and fundamental applications of large language models. Subsequently, we introduce the concept of data streams and provide an overview of current use cases where large language models are adapted to accommodate streaming data. Finally, we summarize the existing challenges associated with integrating large language models with data streams and discuss potential solutions.

Keywords: LLM · data streams · chatGPT.

1 Introduction

Large language models have shown their wide usage and significant competence in many fields according to [1], such as learning and answering users' questions in academic fields, assisting in diagnosing diseases in medical fields, generating text and classifying text data to various categories etc. However, as [24] demonstrates, there is a significant limitation of current large language models: they are trained based on certain static datasets that will not automatically be updated, and this causes the resulting models to only be able to access information from its training datasets. When the models need to be updated as new data becomes available, the only way is to start the training process over again. But in many scenarios such models can't satisfy our needs. For example, to have a better traffic prediction, real-time traffic data is needed [6], Analyzing news and events sentiment can help predict the financial market [27], real-time health data is of great importance when monitoring patients' health condition [3] etc. So to have models that can fulfill those use cases, we need to find and compare useful methods that combine large language models and continuous data input (data streams) together, and also summarize the current major obstacles.

2 Related Works

In this chapter, I will talk about the concepts and techniques that are related to this topic

2.1 Large Language Model

In this subsection I will talk about the following aspects of large language models:

The component of large language models

– Transformer architecture: neural network architectures designed to handle sequential data, such as text, effectively [15]
– Self-attention mechanism: enables the model to capture dependencies between words in a sentence more effectively.
– Embedding layer: converts words or tokens into high-dimensional vectors, which represent their semantic meaning in the context of the sentence.
– Encoder layers: help the model to encode the input text into a meaningful representation.
– Decoder layers: consist of self-attention mechanisms and feed-forward neural networks, but they also include additional attention mechanisms to focus on relevant parts of the input during the decoding process.
– Positional encoding: added to the input embeddings to provide information about the position of each word in the sequence.
– Output layer: takes the final representation produced by the decoder layers and maps it to the output vocabulary.
– Softmax: convert the raw scores into probabilities.
– Training datasets: numerous persona-chats and crawling various sources over the web.

The evolution of large language models

– Earliest simple models: relied on statistical methods and shallow learning techniques.
– Models using deep learning techniques: recurrent neural networks (RNNs) and long short-term memory (LSTM)
– Models using the transformer architecture: revolutionized natural language processing (NLP) by allowing models to capture dependencies between words more effectively through attention mechanisms.
– Models using BERT technique: further improved language understanding by pre-training models on large corpora of text in both forward and backward directions [16].
– GPT series: trained on massive datasets [28]. GPT2 owns 1,5 billion parameters, whereas GPT3 has 175 billion parameters.
– Future trend: continual increase in model size and training data.

The use case of large language models It has huge potential and capabilities such as language translation, text classification, content generation, chatbots etc.

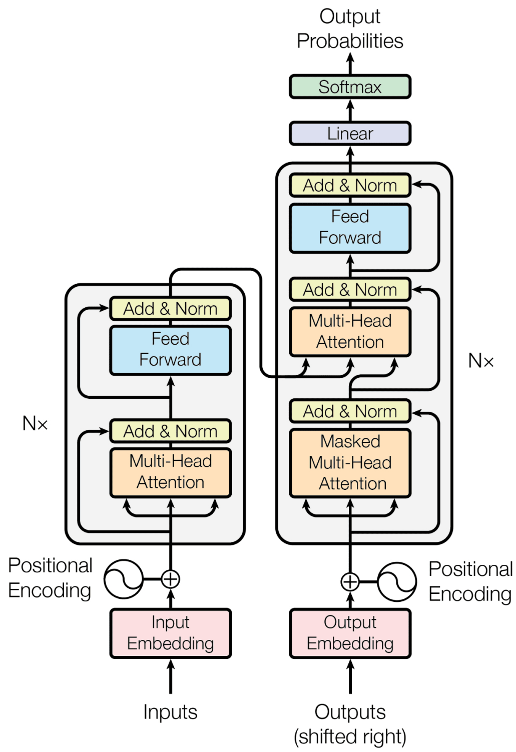

Fig. 1. The transformer structure [15]

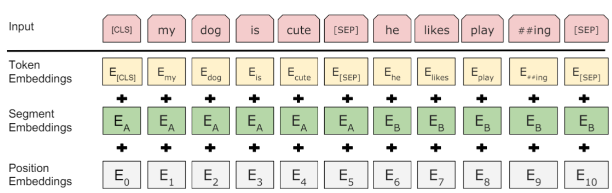

Fig. 2. Input representation in the BERT model [16]

2.2 Data Stream

In this subsection, the following aspects of data streams will be covered:

The definition of data streams With the definition from different sources [13].

Theorem 1. *Data Stream S: Unbounded multiset of data stream elements (s,τ)*

The use case of data streams Data streams are the key components in many application fields such as Weather Forecasting, Health monitoring, Internet of Things etc [14].

3 Use Cases and Obstacles

In this section, I will give an overview of current use cases where large language models and data streams are combined, and also summarize the major challenges of combining large language models with data streams. Furthermore I will also give an introduction of what measurements have been taken to mitigate the obstacles.

3.1 Use Cases

Following use cases will be covered:

- Social Media Monitoring: Analyzing streaming data from social media platforms can help monitor regional news, sentiment, and trends in real-time.
- Financial Monitoring: Large language models can analyze streaming news feeds to identify important events, trends, and sentiment in real-time. This can be valuable for financial institutions and risk management companies to stay informed about current events and market trends.
- Health Care Monitoring: Large language models can analyze streaming medical data such as patient records, diagnostic reports, and research papers to assist healthcare providers in diagnosing diseases, identifying treatment options, and monitoring public health trends in real-time.
- Traffic Data Monitoring: Large language models can analyze streaming traffic data to monitor traffic conditions in real-time to help reduce traffic accidents and level up transportation efficiency.

3.2 Comparison of LLMs with static datasets and data streams

3.3 Challenges

Following challenges will be covered:

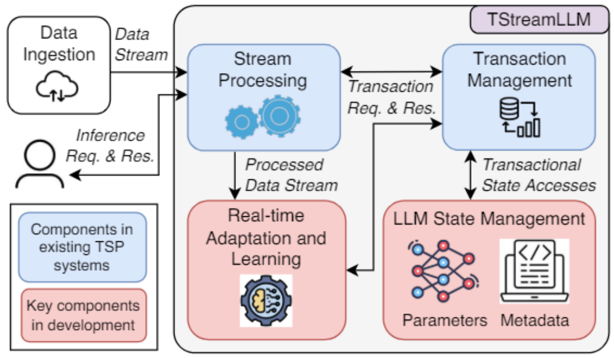

Fig. 3. Architecture of TStreamLLM [6]

– Catastrophic forgetting [24]
– Concept drift: if the LLM is trained for a long time, it is possible that the relationship between the inputs and the outputs itself might change.
– Computational resources: continuous learning requires significant computational resources, which may be costly or impractical to scale.
– Privacy and security: streaming data often contains sensitive or private information, raising concerns about data privacy and security.
– Data quality: data streams may contain noisy or unreliable information, leading to incorrect model updates.

3.4 Solutions

Following solutions will be covered:

– Finetuning [16].
– Continual pre-training: continuously pretraining models on new incoming data [24].
– Using data preprocessing techniques to filter out noise and ensure data quality. Use anomaly detection algorithms to identify and remove outliers in the data. Employ quality assurance measures to verify the accuracy of incoming data.
– Continuously monitor model performance and detect concept drift using statistical methods or machine learning algorithms. Implement adaptive learning techniques to update the model in response to concept drift. Periodically retrain the model on recent data batches to maintain accuracy.
– Optimize model architectures and algorithms to reduce computational overhead. Utilize distributed computing frameworks to parallelize model training and inference tasks.
– Implement data anonymization and encryption techniques to protect sensitive information during data transmission and storage

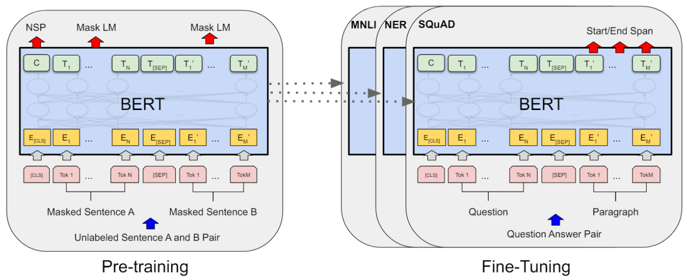

Fig. 4. Finetuning in the BERT model [16]

4 Conclusion

In this section, I will shortly summarize the outline of this paper and also talk about the possible future development of large language models using data streams.

References

1. 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.
2. Kasneci, Enkelejda, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser et al. "ChatGPT for good? On opportunities and challenges of large language models for education." Learning and individual differences 103 (2023): 102274.
3. Thirunavukarasu, Arun James, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. "Large language models in medicine." Nature medicine 29, no. 8 (2023): 1930-1940.
4. Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen et al. "A survey on evaluation of large language models." ACM Transactions on Intelligent Systems and Technology (2023).
5. Zhang, Shuhao, Xianzhi Zeng, Yuhao Wu, and Zhonghao Yang. "Harnessing scalable transactional stream processing for managing large language models [vision]." arXiv preprint arXiv:2307.08225 (2023).
6. Zhang, Kunpeng, Feng Zhou, Lan Wu, Na Xie, and Zhengbing He. "Semantic understanding and prompt engineering for large-scale traffic data imputation." Information Fusion 102 (2024): 102038.
7. Xu, Xuhai, Bingshen Yao, Yuanzhe Dong, Hong Yu, James Hendler, Anind K. Dey, and Dakuo Wang. "Leveraging large language models for mental health prediction via online text data." arXiv preprint arXiv:2307.14385 (2023).
8. Zhang, Xin, Linhai Zhang, Deyu Zhou, and Guoqiang Xu. "Fine-grainedly Synthesize Streaming Data Based On Large Language Models With Graph Structure Understanding For Data Sparsity." arXiv preprint arXiv:2403.06139 (2024).

9. Wu, Tongtong, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. "Continual learning for large language models: A survey." arXiv preprint arXiv:2402.01364 (2024).
10. Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.
11. Gama, Joao, Raquel Sebastiao, and Pedro Pereira Rodrigues. "On evaluating stream learning algorithms." Machine learning 90 (2013): 317-346.
12. Jang, Joel, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. "Towards continual knowledge learning of language models." arXiv preprint arXiv:2110.03215 (2021).
13. Geisler, Sandra. "Data stream management systems." In Dagstuhl Follow-Ups, vol. 5. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2013.
14. Geisler, Sandra. "A systematic evaluation approach for data stream-based applications." PhD diss., Dissertation, RWTH Aachen University, 2016, 2016.
15. 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).
16. 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).
17. Yenduri, Gokul, M. Ramalingam, G. Chemmalar Selvi, Y. Supriya, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, G. Deepti Raj et al. "GPT (Generative Pre-trained Transformer)–A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions." IEEE Access (2024).
18. Lester, Brian, Rami Al-Rfou, and Noah Constant. "The power of scale for parameter-efficient prompt tuning." arXiv preprint arXiv:2104.08691 (2021).
19. So, David, Quoc Le, and Chen Liang. "The evolved transformer." In International conference on machine learning, pp. 5877-5886. PMLR, 2019.
20. Zhao, Feng, Xinning Li, Yating Gao, Ying Li, Zhiquan Feng, and Caiming Zhang. "Multi-layer features ablation of BERT model and its application in stock trend prediction." Expert Systems with Applications 207 (2022): 117958.
21. Ren, Yilong, Yue Chen, Shuai Liu, Boyue Wang, Haiyang Yu, and Zhiyong Cui. "TPLLM: A Traffic Prediction Framework Based on Pretrained Large Language Models." arXiv preprint arXiv:2403.02221 (2024).
22. Liu, Chenxi, Sun Yang, Qianxiong Xu, Zhishuai Li, Cheng Long, Ziyue Li, and Rui Zhao. "Spatial-temporal large language model for traffic prediction." arXiv preprint arXiv:2401.10134 (2024).
23. Yang, Xi, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas et al. "A large language model for electronic health records." NPJ digital medicine 5, no. 1 (2022): 194.
24. 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).
25. Naga Sanjay, "Continuous Training of ML models. A case-study on how to keep our machine learning models relevant.", Medium, June 25, 2023
26. Prapas, Ioannis, Behrouz Derakhshan, Alireza Rezaei Mahdiraji, and Volker Markl. "Continuous training and deployment of deep learning models." Datenbank-Spektrum 21, no. 3 (2021): 203-212.

27. Araci, Dogu. "Finbert: Financial sentiment analysis with pre-trained language models." arXiv preprint arXiv:1908.10063 (2019).
28. Roumeliotis, Konstantinos I., and Nikolaos D. Tselikas. "Chatgpt and open-ai models: A preliminary review." Future Internet 15, no. 6 (2023): 192.