Meeting Report: Using ChatGPT to Generate Data for Medical AI Models

TCCI Hosts Second "AI for Brain Science" Conference focused on Data Generation for Medical AI Models

On May 28, the Tianqiao and Chrissy Chen Institute (TCCI) hosted the second session of the "AI for Brain Science" series themed -- "**Data Generation Methods for AI Models and Their Implications for the Medical Field**". Chaired by Mengyue Wu, Associate Professor of the Department of Computer Science and Engineering of Shanghai Jiao Tong University, three young scientists shared their practices and views on breaking the bottleneck of data for large language models (LLM).

Self-training and self-distillation: developing proprietary GPT models efficiently

An international study has determined that ChatGPT can answer cancer-related questions at a level already on par with the official answers provided by the US National Cancer Institute. ChatGPT however can only be accessed through restricted APIs and when it comes to personal healthcare, the public at large do not want to share private information with third-party companies.

To resolve such difficulties, Canwen Xu, a PhD student at UC San Diego and collaborators from Sun Yat-sen University proposed a process that **automatically generates a high-quality multi-round chat corpus by using ChatGPT to chat with itself. The conversational data can then be used to fine-tune and enhance the open-source large language model called LLaMA.** As a result, they obtained a high-quality proprietary model called "Bai Ze" whose version 2.0 was launched just a few days ago. The name was inspired by a mythical beast in ancient China that can speak and understand the feelings of all things.

According to Xu, Bai Ze did not learn anything new in the process, but simply extracted specific data from LLaMA and retained the powerful linguistic capabilities of ChatGPT to answer questions by bullet points or refuse to answer certain inappropriate questions. This has been professionally referred to as 'distillation'. The concept of feedback self-distillation was also introduced, whereby ChatGPT is used

as a coach to score and rank the answers provided by Bai Ze to further improve its performance.

Xu thinks that **Bai Ze is both economical and pragmatic as it acquired ChatGPT's capabilities in a certain domain through automated knowledge distillation at a much lower cost.** In the medical field, localized or private AI models will help ease privacy-related concerns and assist in diagnosis and treatment. In the future, perhaps everyone will have a personal AI assistant.

A new data generation strategy: optimizing medical text mining

Ruixiang Tang, a PhD student at Rice University and his collaborators have also proposed a new data generation strategy based on large language models. It has demonstrated better performance on classic medical text mining tasks such as named entity recognition (NER) and relationship extraction (RE).

Since ChatGPT is capable of creative writing, it performs well in knowledge-intensive areas with little annotated data including healthcare, finance, and law. However, when it comes to medical text mining, they found that **applying ChatGPT directly in the downstream tasks of processing medical text did not always achieve ideal results and might even cause privacy-related problems.**

To address this challenge, Tang and his team proposed a new strategy: **using large language models to generate large amounts of medical data, and then training these data with smaller models.** The experimental results show that this new strategy achieved better results than the direct application of large language models in downstream tasks. It has also significantly reduced potential privacy risks since the data is stored locally.

They further pointed out that as open-source large language models are improving in terms of both quantity and quality, the gap between the text data generated by AI and those by human will be increasingly narrowed. It'll be technically challenging to tell the difference. The two existing testing means are likely to be ineffective, whether it is black-box testing, which directly compares text data generated by large language models with those generated by humans (e.g. comparing the distribution of high-frequency words), or white-box testing, where developers will label the generated text. **The ability to effectively detect whether the data is generated by GPT will influence users' trust in LLM-based AI.**

What's so unique about data generation in the large language model era?

In history, how did scientists address the challenge of data scarcity without GPT? And what new trends have large language models set?

Ruisheng Cao, a PhD student at Shanghai Jiao Tong University, reviewed research in automated data generation or augmentation based on deep learning models upon the dawning of LLM era. Deep learning is essentially a process of finding the mapping from an input x to an output y. Therefore, a large number of (x,y) data pairs are needed for training. In areas like healthcare where large amounts of authentic data are difficult to obtain, more (x,y) data pairs need to be generated manually.

Cao breaks down data generation into three main modules. The first is for the generation of label (y). It aims at coupling the generated labels with the distribution of the authentic data for comparison. The second is responsible for the methods and constraints in generating the initial data (x). The third will ensure data quality once a complete data pair has been formed.

Continuously growing in scale and capability, large language models are generating data with improving quality. Models trained by generated data can not only solve simple tasks, such as text classification, but also handle more complex tasks such as Q&A sessions.

Looking forward, Cao summarized several new trends in data generation in the era of large language models. The first is to build more generalized models to ensure their application in diverse tasks. This means that models need to be more adaptable and generalized. The second is to further refine the process from a task-specific perspective. For example, in the medical field, it is even possible to customize tasks for specific types of depression, thus offering more precise and personalized medical solutions. Last but not the least, the process of data generation and model training will be more integrated, and mandatory filtering will be gradually replaced by flexible control to ensure data quality.

Research into data generation and its application are tapping into the potentials of AI models in various domains, especially the medical field.

Click here to read the full article

Scan the QR code to watch the replay

Dong Liu is a Master of Neurobiology at Capital Medical University. He wrote this meeting report as part of the Tianqiao and Chrissy Chen Institute Science Writers Fellowship which aims to extend the conversation beyond the meeting with the hopes of sparking new ideas and collaborations.