DELL Technologies

Whitepaper

Harnessing the Power of Specialized Language Models like ChatGPT for Business Success

Author :
Ben Fauber, Ph.D.
Distinguished Member
Technical Staff,
Senior AI Research Scientist,
Dell Technologies

Understand transfer learning as a powerful approach to refining language models.

Abstract

Large language models have many benefits for general natural language processing tasks, yet their practical application often requires a fine-tuning process to establish interfaces with downstream applications to make them even more powerful. Numerous transfer learning approaches have been suggested to enable fine-tuning of pre-trained language models on specific tasks and use cases like Generative AI. To successfully implement this approach, in-house machine learning expertise or trusted partners, high-performance computing hardware infrastructure, and structured datasets are necessary. In this article, we discuss strategies to securely fine-tune large language models for your business within your secure IT infrastructure.

Large language models (LLMs) are artificial intelligence (AI) systems that use machine learning (ML) algorithms to process vast amounts of natural language text data. They have become increasingly popular due to their impressive natural language processing (NLP) capabilities.[1] Large pre-trained language models can extract generalizations from vast amounts of text data, which can be utilized for a myriad of downstream applications such as text classification, text summarization, text generation, named entity recognition (NER), text sentiment analysis, and question-answering (Q&A). Additionally, many large language models are multilingual, making them even more versatile in utilizing text datasets across many different languages.

One of the strengths of large language models is that they contain a broad amount of information and knowledge, thanks to the massive amounts of text data used to train them. However, this also means they often struggle as specialists for deeper dives on topics or items with limited instances in the training dataset.[2] To address this shortcoming, businesses can add domain-specific information from their vertical into a pre-existing large language model. This layered training approach in which specialized information is added to a pre-trained model is known as transfer learning, or model fine-tuning.

Transfer Learning: Preparing Existing Models for New Information

Transfer learning creates application-specific parameters on top of pre-trained large language models. The process involves exposing a pre-existing model to new information, allowing the model to adapt to that new information while not forgetting the old information.[3] A neural network's weights are updated during transfer learning training, and the new weights may not be compatible with the old weights of the pre-trained model. Thus, if transfer learning is not carefully executed, the model's ability to perform the original task may degrade significantly resulting in an outcome known as "catastrophic forgetting."[4]

Traditional approaches to transfer learning involved freezing all but a few layers of the deep neural network (DNN) of the pre-trained model, thereby allowing only a few layers of the pre-trained neural network to learn from the new data and avoid catastrophic forgetting.[5] This approach has been particularly effective in transfer learning with large transformer-based language models like BERT.[6]

ChatGPT has gained considerable attention since its initial release for its ability to write fluid human-like prose.[7] ChatGPT is a refinement of the InstructGPT large language model. InstructGPT was created through the alignment of large language model outputs with user intent by incorporating reinforcement learning from human feedback (RLHF).[8]

Application of reinforcement learning from human feedback to create InstructGPT involved three major steps: 1) fine-tuning the GPT-3 large language model on diverse user prompts and appropriate responses, 2) rewarding the fine-tuned model when it combined appropriate prompt and response pairs, and 3) scoring the alignment of random prompts and responses with user intent by applying a reinforcement learning (RL) model based on the outcomes of step 2. Ultimately, the performance of InstructGPT was compared against its predecessor, GPT-3, based on their ability to infer and follow user instructions (helpfulness), their tendency for hallucinations (truthfulness), and their ability to avoid inappropriate and derogatory content (harmlessness). InstructGPT outperformed GPT-3 on all three criteria, with human referees favoring the outputs of InstructGPT 85% of the time.

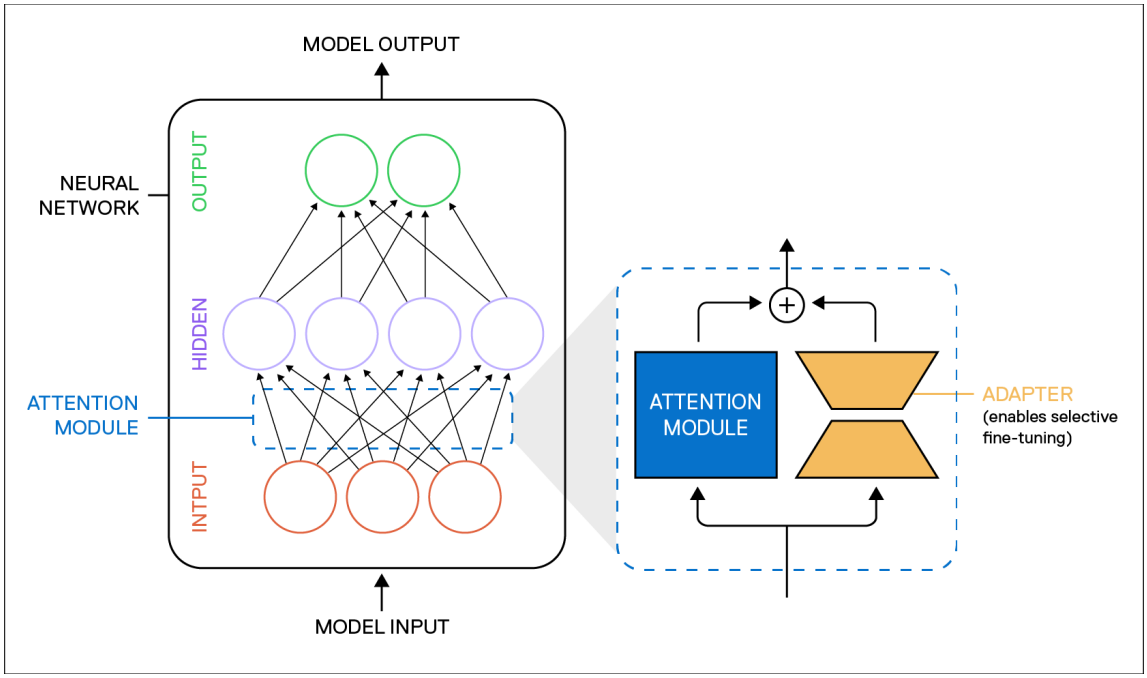

Figure 1. Illustration of the parameter efficient fine-tuning adapter technique applied to the attention module of a simplified shallow neural network.

Memory usage and communication costs are important considerations when training large language models. Recent work has shown that selectively updating only a small subset of a model's parameters during transfer learning training can alleviate these issues while avoiding catastrophic forgetting.[9] So-called parameter-efficient fine tuning (PEFT) has been deployed to rapidly fine-tune large language models with domain-specific data and up-to-date information.

Parameter-efficient fine-tuning leverages a regularization technique known as an adapter-based approach. The adapter technique inserts small bottleneck layers within each layer of a pre-trained neural network model, thereby fixing the pre-trained layers, and training the adapter layers on the new data.[10] This approach has been shown to improve model stability and robustness in transfer learning for various applications with minimal computational overhead. This method is particularly useful for fine-tuning large language models.

Infrastructure and Deployment Considerations

Building large language models specific to a certain business or vertical with transfer learning requires in-house machine learning expertise or an engagement with a trusted partner. Additionally, high-performance computing (HPC) hardware infrastructure such as multiple graphics processing units (GPUs), high-speed networking, and structured datasets for transfer learning are required. A high-level understanding of the datasets used to train the existing large language models is essential to ensure that the new data used in transfer learning will impart information diversity and increased reach of the business-specific large language model. There are a few publicly available data sets such as The Pile[11] and ROOTS[12] for training large language models, yet many of the data sets used for training are proprietary and not public.

There are several providers of large language model application programming interfaces (API) that allow businesses to programmatically connect directly to the model.[7] However, these application programming interfaces could have limited to no flexibility for a user to customize the language model to a specific business need or ontology. Further, the costs of operating compute-intensive large language models via third-party providers, instead of operating the same models on-premises, should be carefully considered, as routine on-prem operations can be more cost-effective.

Securely Fine-Tune Language Models in Your Infrastructure

Enable your organization with large language models secured within your infrastructure to securely interact with your data and employees. Open-source foundational large language models such as BLOOM-176B,[13] BLOOMZ-176B,[14] GPT-J-6B,[15] or OPT-175B[16] can be instantiated within your corporate information technology (IT) infrastructure. The biggest benefit of running the model within your infrastructure is security: securely manage the model and secure the information it receives. Additional benefits include faster inference, greater availability and uptime, reduced reliance on third-party services, maintaining greater control of your own data and infrastructure, cost-effectiveness, and filtering the language model outputs to align them with your organization's domain and voice.

Empower your organization with large language models fine-tuned on your data and secured within your infrastructure. Copies of the large language model instantiated within your corporate IT infrastructure can be securely fine-tuned with domain-specific knowledge, continually updated with new data, and managed with guardrails on content, bias, and toxicity to create your organization's customized large language models (see Figure 2).

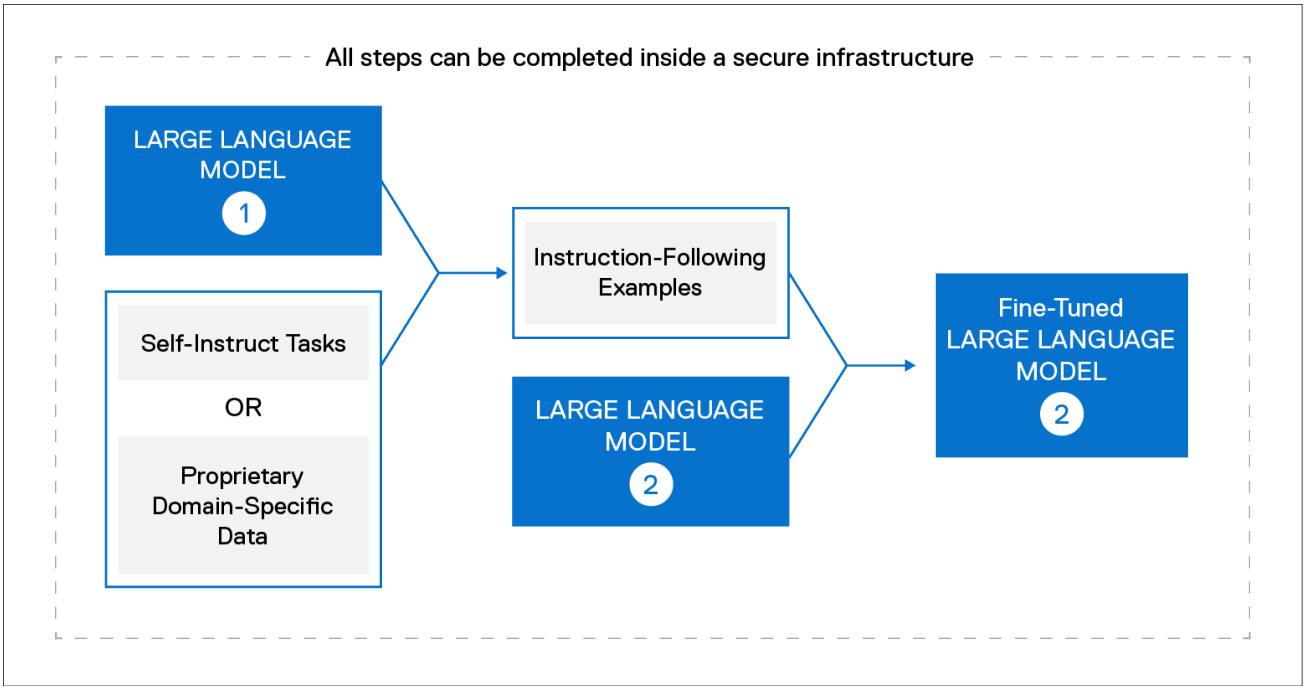

Figure 2. Illustration demonstrates the use of one, or multiple, large language models to either: 1) fine-tune a large language model on proprietary domain-specific data, or 2) use a language model to further refine proprietary domain-specific data for downstream language model fine-tuning. All steps can be securely performed within a customer's infrastructure with Dell Technologies private cloud and/or on-premises platforms.

Dell Technologies can help you on this journey.
Recent results from a research group at Stanford University demonstrated the benefits of using a large language model to further fine-tune and refine an existing large language model, resulting in a new model known as ALPACA.[17] This process demonstrated a new paradigm in which one large language model was used to train another large language model. This approach could further reduce the role of humans in the model training loop, such as reinforcement learning with human feedback (RLHF), as previously leveraged by InstructGPT and ChatGPT.[8]

In ALPACA's case, a large generalist language model created a curated series of questions and answers on which to fine-tune another. Fine-tuning the second language model with questions generated by the first resulted in improved performance metrics of the second model. This paradigm has since been expanded to other domains outside of question and answer to enable rapid and efficient tuning of existing large language models on new data.[18]

In conclusion, large language models offer many advantages for general natural language processing tasks. However, practical use of these generalizations requires a fine-tuning process to create interfaces between the models and downstream applications to make them even more powerful. Various transfer learning approaches have been proposed to enable the fine-tuning of pretrained language models on specialized tasks and use cases. In-house machine learning expertise or trusted partners, high-performance computing hardware infrastructure, and structured datasets are needed to successfully implement this approach. Transfer learning is an effective strategy to customize large language models, but access to high-performance computing resources and skilled talent is crucial.

REFERENCES

1. Fauber, B. "Unleashing the power of large language models like ChatGPT for your business." 2023, Dell Technologies white paper, https://www.delltechnologies.com/asset/en-us/solutions/infrastructure-solutions/industry-market/unleashing-the-power-of-large-language-models-fauber.pdf (accessed 11Apr2023).
2. Brown, et al. "Language models are few-shot learners." Conference and Workshop on Neural Information Processing Systems Conference (NeurIPS) 2019.
3. Raina, R. et al. "Self-taught learning: transfer learning from unlabeled data." International Conference on Machine Learning (ICML) 2007, 759–766.
4. "Catastrophic interference." https://en.wikipedia.org/wiki/Catastrophic_interference (accessed 11Apr2023).
5. Chollet, F. "Deep learning with python." 2017, Manning Publications, 384 pages.
6. Lee, J. et al. "BioBERT: A pre-trained biomedical language representation model for biomedical text mining." Bioinformatics 2020, 36(4), 1234–1240.
7. Open AI Research. "ChatGPT." 2022, https://chat.openai.com/
8. Ouyang, et al. "Training language models to follow instructions with human feedback." 2022, arxiv.org/abs/2203.02155.
9. Houlsby, N. et al. "Parameter-efficient transfer learning for NLP." International Conference on Machine Learning (ICML) 2019.
10. Hu, E. "LoRA: Low-rank adaptation of large language models." International Conference on Learning Representations Conference (ICLR) 2022.
11. Gao, et al. "The Pile: An 800GB dataset of diverse text for language modeling." 2020, arxiv.org/abs/2101.00027.
12. Laurençon, et al."The BigScience ROOTS corpus: A 1.6TB composite multilingual dataset." Conference and Workshop on Neural Information Processing Systems Conference (NeurIPS) 2022.
13. BigScience Workshop, et al. "BLOOM: A 176B-parameter open-access multilingual language model." 2022, arxiv.org/abs/2211.05100
14. Muennighoff, N. et al. "Cross lingual generalization through multitask finetuning." 2022, https://arxiv.org/abs/2211.01786
15. EleutherAI, et al. "GPT-J-6B." 2021, https://huggingface.co/EleutherAI/gpt-j-6b (accessed 11Apr2023).
16. Facebook AI Research. "Democratizing access to large-scale language models with OPT-175B" 2022, https://ai.facebook.com/blog/democratizing-access-to-large-scale-language-models-with-opt-175b/
17. Taori, R. et al. "Stanford Alpaca: An instruction-following LLaMA model." 2023, https://github.com/tatsu-lab/stanford_alpaca (accessed 11Apr2023).
18. Peng, et al. "Instruction tuning with GPT-4." 2023, arxiv.org/abs/2304.03277.