CUTTING THE COSTS OF FINE-TUNING LARGE LANGUAGE MODELS.

Background

Large language models (LLMs) are the brains behind many artificial intelligence tools, from customer-service chatbots to language translators.

LLMs must first be trained with huge amounts of data. Fine-tuning all the model layers to improve their performance for specific tasks is extremely resource-intensive.

Using a novel approach researchers at Monash University have found a way to update a subset of the layers — only those pertinent to the task in question — thereby reducing computational demands.

Challenges

The Monash team analysed how LLMs process information, carefully tracing the semantic changes of data. By measuring the contributions of each layer, they identified which ones were worth updating and which could be left as is — significantly reducing computational costs.

To validate their approach and run large-scale experiments, the team needed high performance computing (HPC) resources capable of handling the demands of training and fine-tuning LLMs.

Solutions

We provided the researchers with a supported HPC solution, including powerful GPUs that enabled them to conduct their experiments at scale.

Our HPC experts ensured a seamless integration of the researchers' workflows with DUG HPC Cloud, allowing the team to focus on their research.

Results

Monash University's approach significantly reduced the computational demands of fine-tuning, cutting time-related costs by nearly half and achieving savings of up to 64% through additional optimisation. Their method also integrates smoothly with existing techniques, providing a practical solution for efficient tuning of LLMs.

Our tailored HPC solution enabled the Monash researchers to successfully run the massive experiments required to validate and refine their ideas. They have published their findings in this paper.

dug