Tackling AI’s sustainability problem with code optimisation

Google’s announcement that its emissions are up by 48% in five years due to the energy demand of its data centres has brought the environmental damage of AI to the fore. Previously kept mostly out of the headlines, the vast amount of computing resources and energy this increasingly competitive and growing AI landscape is consuming is now being exposed.

Temps de lecture : minute

13 August 2024

As companies rush to integrate AI into their products, processes and services, its huge environmental toll can no longer be ignored. Notably, it’s not only the training of models that significantly increases AI’s carbon footprint; the ongoing inference required to process requests and generate responses is also considerable.

Of course wide-scale solutions are needed to ensure AI can positively address the climate crisis and not worsen it. This strategy relies on the rollout of green energy alongside practical steps such as hardware transformations. However, one aspect of AI that remains under the radar, but underscores its capabilities and development, is the efficiency of AI code. This efficiency is not only fundamental to AI performance and effectiveness, but also for its sustainability.

By simply enhancing the efficiency of their code – through code optimisation – companies can deliver immediate improvements to their AI’s carbon footprint.

The many problems and costs of inefficient code

When companies are looking to reduce costs or optimise their operations, the value of optimising their code can be overlooked due to other business priorities. This leads to the emergence of increasingly inefficient code, which then leads to the software being more unsustainable, both environmentally and performance wise.

Without the right tools, optimising code is no easy task. During my time in the finance sector, code optimisation was a manual process, and I witnessed just how much time and how many resources the process of developing and optimising code required. Even the most skilled engineers could need days to iteratively refine and test the code for optimal performance – a process that rises in complexity as codebases become larger.

There's no set cost estimation for implementing AI, but it’s been reported that even deploying a basic AI system could cost a company $50,000, due to hardware, software and data requirements. And its environmental costs start in training.

Five years ago, The MIT Technology Review reported that *training just one AI model can emit more than 626,000 pounds of carbon dioxide equivalent*, nearly five times the lifetime emissions of an average American car. Given Google’s emissions rise, we can expect a similar emissions increase for the training of the latest AI models today.

The GenAI race

The latest generative AI (GenAI) models like ChatGPT demand an increasingly high amount of compute for training and operation – *a value doubling every six months*. While enhancing the abilities of AI models, this surging compute demand comes with a massive environmental and financial cost.

Earlier this year it was suggested just one ChatGPT search query consumes *15 times more energy* than a Google one, while it has also been estimated that the model costs a whopping *$700,000 per day* to run. Given the major players have already integrated GenAI into their search engines, these costs will continue to rocket.

While compute demand rises quicker than AI hardware capabilities, inefficient code is impacting this unsustainable growth and cost. However, it's not clear how much big tech is addressing the problem behind closed doors. While the largest cloud tech companies in the world – AWS, Azure and CGP – now offer services like optimised compute and memory, this does not deal with the impact from the rapid growth in data centres to run these large AI models.

Many smaller companies rely on these open-source models for their own AI systems, and so they need to find other ways of making their systems as energy efficient as possible.

Why code optimisation is the starting point

When it comes to training and running AI models, by starting from the very beginning and optimising code, companies can reduce the computational load of servers and data centres and therefore lower their carbon emissions. The process also makes the software run more efficiently, reducing latency and bettering the user experience.

But how do engineers perform the process without carrying out laborious manual processes? Well, by using AI itself.

The latest code optimisation platforms use GenAI and large language models – alongside input from human developers – to optimise their AI code. These pre-trained models are able to automatically identify lines of inefficient code by scanning through entire codebases. They then offer code suggestions to improve performance and engineers can evaluate these improvements against the original code. The resulting code still performs the same activities, just more efficiently.

Efficient changes, big impact

Crucially, these changes can deliver sizeable environmental and financial benefits. Improving model performance through code optimisation can generate a 46% reduction in memory and energy consumption and nearly $2M of savings per year in AI production and deployment. Notably, this can create 256kg of CO2 emission savings a year.

If we place these benefits in the context of data centres, which sit at the heart of Google and AI’s rising carbon footprint, code optimisation can produce massive carbon reductions: a 46% energy reduction would mean a decline from 668kgs of carbon equivalent per each server a year to 360kgs.

The foundation for sustainable AI

The AI industry needs fast answers to tackling its sustainability problem. While the technology also has the potential to provide answers itself, that can’t be used as an excuse to not address its massive environmental impact. Its surging costs and processing demands mean there is a financial and technical need as well.

For companies, code optimisation can deliver immediate value, not only bettering AI performance but significantly reducing costs and emissions. And by prioritising code optimisation, the sector can help trigger the wider process of reducing data centres’ vast consumption of energy and resources.

Dr Leslie Kanthan, CEO and Co-founder of *TurinTech.*

Article by Dr Leslie Kanthan