deepmind.com

DeepMind AI Reduces Google Data Centre Cooling Bill by 40%

AuthorsRE Richard EvansJG Jim Gao

5-6 minutes

From smartphone assistants to image recognition and translation, machine learning already helps us in our everyday lives. But it can also help us to tackle some of the world's most challenging physical problems - such as energy consumption. Large-scale commercial and industrial systems like data centres consume a lot of energy, and while much has been done to stem the growth of energy use, there remains a lot more to do given the world's increasing need for computing power.

Reducing energy usage has been a major focus for us over the past 10 years: we have built our own super-efficient servers at Google, invented more efficient ways to cool our data centres and invested heavily in green energy sources, with the goal of being powered 100 percent by renewable energy. Compared to five years ago, we now get around 3.5 times the computing power out of the same amount of energy, and we continue to make many improvements each year.

Major breakthroughs, however, are few and far between - which is why we are excited to share that by applying DeepMind's machine

learning to our own Google data centres, we've managed to reduce the amount of energy we use for cooling by up to 40 percent. In any large scale energy-consuming environment, this would be a huge improvement. Given how sophisticated Google's data centres are already, it's a phenomenal step forward.

The implications are significant for Google's data centres, given its potential to greatly improve energy efficiency and reduce emissions overall. This will also help other companies who run on Google's cloud to improve their own energy efficiency. While Google is only one of many data centre operators in the world, many are not powered by renewable energy as we are. Every improvement in data centre efficiency reduces total emissions into our environment and with technology like DeepMind's, we can use machine learning to consume less energy and help address one of the biggest challenges of all - climate change.

One of the primary sources of energy use in the data centre environment is cooling. Just as your laptop generates a lot of heat, our data centres - which contain servers powering Google Search, Gmail, YouTube, etc. - also generate a lot of heat that must be removed to keep the servers running. This cooling is typically accomplished via large industrial equipment such as pumps, chillers and cooling towers. However, dynamic environments like data centres make it difficult to operate optimally for several reasons:

1. The equipment, how we operate that equipment, and the environment interact with each other in complex, nonlinear ways. Traditional formula-based engineering and human intuition often do not capture these interactions.

2. The system cannot adapt quickly to internal or external changes (like the weather). This is because we cannot come up with rules and heuristics for every operating scenario.

3. Each data centre has a unique architecture and environment. A custom-tuned model for one system may not be applicable to another. Therefore, a general intelligence framework is needed to understand the data centre's interactions.

   To address this problem, we began applying machine learning two years ago to operate our data centres more efficiently. And over the past few months, DeepMind researchers began working with Google's data centre team to significantly improve the system's utility. Using a system of neural networks trained on different operating scenarios and parameters within our data centres, we created a more efficient and adaptive framework to understand data centre dynamics and optimize efficiency.

   We accomplished this by taking the historical data that had already been collected by thousands of sensors within the data centre - data such as temperatures, power, pump speeds, setpoints, etc. - and using it to train an ensemble of deep neural networks. Since our objective was to improve data centre energy efficiency, we trained the neural networks on the average future PUE (Power Usage Effectiveness), which is defined as the ratio of the total building energy usage to the IT energy usage. We then trained two additional ensembles of deep neural networks to predict the future temperature and pressure of the data centre over the next hour. The purpose of these predictions is to simulate the recommended actions from the PUE model, to ensure that we do not go beyond any operating constraints.

We tested our model by deploying on a live data centre. The graph below shows a typical day of testing, including when we turned the machine learning recommendations on, and when we turned them off.