idw - Informationsdienst Wissenschaft
Nachrichten, Termine, Experten

Pressemitteilung

Universität des Saarlandes
Claudia Ehrlich
18.03.2025
http://idw-online.de/de/news849160

Forschungs- / Wissenstransfer, Forschungsprojekte
Energie, Informationstechnik, Umwelt / Ökologie, Wirtschaft
überregional

Hannover Messe: Sustainable data centres – Making AI models up to 90% more energy efficient

Powering artificial intelligence comes with a massive energy bill attached. Professor Wolfgang Maaß and his research team at Saarland University and the German Research Center for Artificial Intelligence (DFKI) want to make AI up to 90% percent more energy efficient. To improve AI's carbon footprint, the Saarbrücken team is rethinking data centres, large language models and image analysis models – and their research is opening up access to powerful AI models for small and medium-sized companies. From 31 March to 4 April, the researchers will be at this year's Hannover Messe showcasing their work at the stand of the Federal Ministry for Economic Affairs and Climate Action (Hall 2, Stand A18).

Data centres consume vast quantities of energy. According to Bitkom, the leading industry association in Germany's digital sector, the electricity requirements to power data centres have more than doubled over the past decade. And with digital transformation only just out of the starting blocks, this trend is really gathering pace. Storing, processing, transmitting and retrieving data takes energy. Artificial intelligence, in particular, is a huge energy guzzler. Globally, multiple terawatt hours are being used to train and run today's massive AI models. (One terawatt hour is equal to one billion kilowatt hours of electrical energy). Using these models to generate images and texts also consumes vast amounts of energy. As a result, data centres are having to get bigger and bigger, which means they need more and more electricity to power and cool the huge numbers of processors involved, which in turn is causing a massive uptick in their carbon footprint. None of this is helping Europe achieve its goal of net-zero greenhouse gas emissions by 2050. Clearly something has to change.

'AI can become far more energy efficient. With the right approach, we can make the data centres of the future much more sustainable,' says Professor Wolfgang Maaß, who conducts research at Saarland University and the German Research Center for Artificial Intelligence (DFKI). In an effort to curb AI's hunger for energy and to conserve resources, his research team is developing leaner, customized AI models. They also want to identify ways that data centres can become more energy smart.

'By making the models smaller and more efficient, we're helping to drive sustainability,' says Dr. Sabine Janzen, a senior research scientist in Wolfgang Maaß's team. 'Our work is also opening up access to powerful AI models for small and medium-sized businesses, because these smaller, leaner AI models don't need a large technical infrastructure. This will enable everyone – not just the big players – to leverage this new technology,' says Janzen.

Today's AI chatbots such as ChatGPT and visual AI models use trillions of parameters and utilize vast datasets to perform their tasks. The amount of energy they consume is correspondingly huge. The researchers in Saarbrücken are developing ways to reduce this energy consumption, without compromising the quality of the output from these leaner AI models. 'A central element of our work is a technique known as knowledge distillation. It's a type of compression technique that enables us to make models that are smaller and therefore more energy efficient, but that perform just as well as the larger models,' explains Sabine Janzen.

The approach used by the research team could be described as follows: When looking for the answer to a specific question, you don't read an entire library; you focus instead only on those books that are relevant to your question. The researchers in Saarbrücken extract smaller, more focused and more energy-efficient 'student' models from larger 'teacher' models. By distilling the knowledge needed to perform tasks in a specific area and reducing it to the essentials, they can reduce the size of the data models by up to ninety percent. Model parameters that are not relevant to the area of interest are not touched. 'In terms of inference speed, i.e. how quickly the model can process input data and produce results, these student models perform at a level comparable to that of the larger teacher models, but require 90% less energy to do so,' explains Janzen.

By using another automated efficiency technique known as 'neural architecture search' (NAS), the team has also achieved some impressive results with visual AI models, i.e. models that process digital image data. 'Our most recent results show that we can use the NAS method to reduce the size of the models by around ninety percent,' says Sabine Janzen. In this work, the researchers focus on machine learning with artificial neural networks – a very energy-intensive AI method that can analyse large volumes of data. Artificial neural networks are designed to mimic the human brain. Our brains contain many billions of nerve cells, called neurons, that are connected to each other via trillions of synapses. A synapse is essentially the interface between two neurons across which the two nerve cells communicate with each other. When we learn something new, neurons send electrical signals to each other across synapses, as we continue learning, the same neurons keep firing together and the connections between them get stronger, whereas the connections between inactive neurons become weaker.

Learning processes in artificial neural networks are similar and by feeding these networks large amounts of data, they can be trained to recognize patterns in natural language or in images. But whereas the brain is a master of energy-efficient learning, training a large artificial neural network requires a lot of computing power and a lot of energy. Training an artificial neural network so that it can yield meaningful results also involves a significant amount of human input. Typically, these artificial networks are designed and configured manually, and the many parameters involved are adjusted and optimized by experts until they perform at the required level. This is where the Saarbrücken researchers bring 'neural architecture search' (NAS) into play. 'Instead of designing the neural networks manually, we automate the design optimization process using NAS,' explains Sabine Janzen. 'NAS allows us to examine different network architectures and optimize them to create a model that delivers high performance, efficiency and reduced costs.'

To test these compacter AI models in practice, Wolfgang Maaß's team is working together with the steel company Stahl Holding Saar. The aim is to teach the artificial neural networks to sort steel scrap efficiently. In order to produce new steel from scrap steel, producers need scrap of the right quality. Only certain types of scrap can be recycled for the manufacture of high-quality steels. However, the steel scrap that gets delivered to the smelting plant is a mix of all types and has to be sorted. Scrap sorting can be automated, but so far, the AI model is too big to be practical. 'We have compressed the visual AI sorting model, making it compacter and more energy efficient. In fact, on certain metrics, the smaller model even performs better, making the steel recycling process more efficient,' says Janzen. Where previously a huge AI model would have required a lot of energy to operate, a small, customized, energy-efficient model is now able to perform the same task.

The researchers start by training their models with the full dataset that contains all the information. They then shrink the AI models using knowledge distillation and specially compiled neural networks so that the models only contain those parameters that are really necessary for the task at hand. In this particular case, the aim is to create an AI that has all the knowledge it needs to be able to analyse camera images to classify the scrap steel being delivered to the site.

The Saarbrücken research team is also working with partners to outline a concept and compile recommendations for sustainable data centres and energy-efficient AI. 'Up until now it has been difficult to estimate just how much energy is needed to create and operate an AI model. That makes it harder for businesses to plan ahead,' explains PhD student Hannah Stein who is conducting research into these energy-efficient AI models. 'We're currently developing a tool that

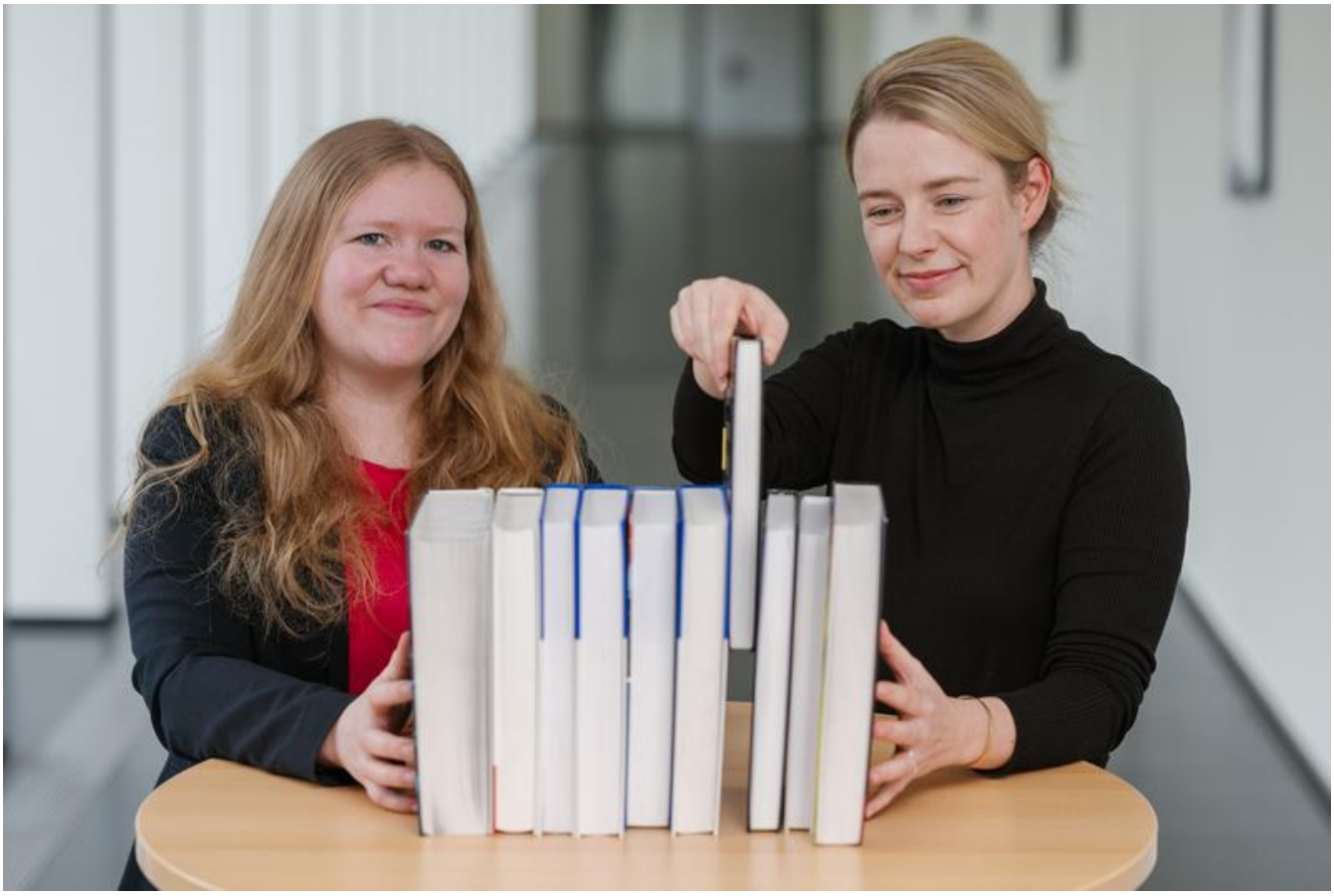
provides reliable forecasts of the energy consumed by and the costs associated with the different AI models,' says Stein. Data centres and AI users can then use this information to plan more effectively, identify inefficient processes and take corrective action as necessary – for example, scheduling heavy computational loads at times when the price of electricity is low.

The research being conducted by Professor Wolfgang Maaß and his team was selected for the Federal Ministry for Economic Affairs and Climate Action's stand at this year's Hannover Messe. The team will be presenting the latest results from the federally funded 'ESCADE' project, which is based at the German Research Centre for Artificial Intelligence DFKI.

Background:
ESCADE ('Energy-Efficient Large-Scale Artificial Intelligence for Sustainable Data Centers') is a three-year project with a budget of around €5 million being financed by the Federal Ministry for Economic Affairs and Climate Action (BMWK). The project will run until the end of April 2026. The ESCADE consortium is made up of the research team headed by Wolfgang Maaß (Saarland University and DFKI), NT Neue Technologie AG, Stahl-Holding-Saar GmbH & Co. KGaA, SEITEC GmbH, Dresden University of Technology, the University of Bielefeld and the Austrian applied research institute Salzburg Research.
https://escade-project.de

wissenschaftliche Ansprechpartner:

Dr. Sabine Janzen: Tel.: +49 681 85775-269, Email: sabine.janzen@dfki.de
Hannah Stein: Tel.: +49 681 302-64739, Email: hannah.stein@iss.uni-saarland.de



You don't read an entire library to answer a specific question. Instead, you focus on those books relevant to the question at hand. That’s the approach researchers Sabine Janzen (r.) and Hannah Stein (l.) are using to make AI models more energy efficient.
Credit: Oliver Dietze
Saarland University

Professor Wolfgang Maaß
Credit: Oliver Dietze
Saarland University