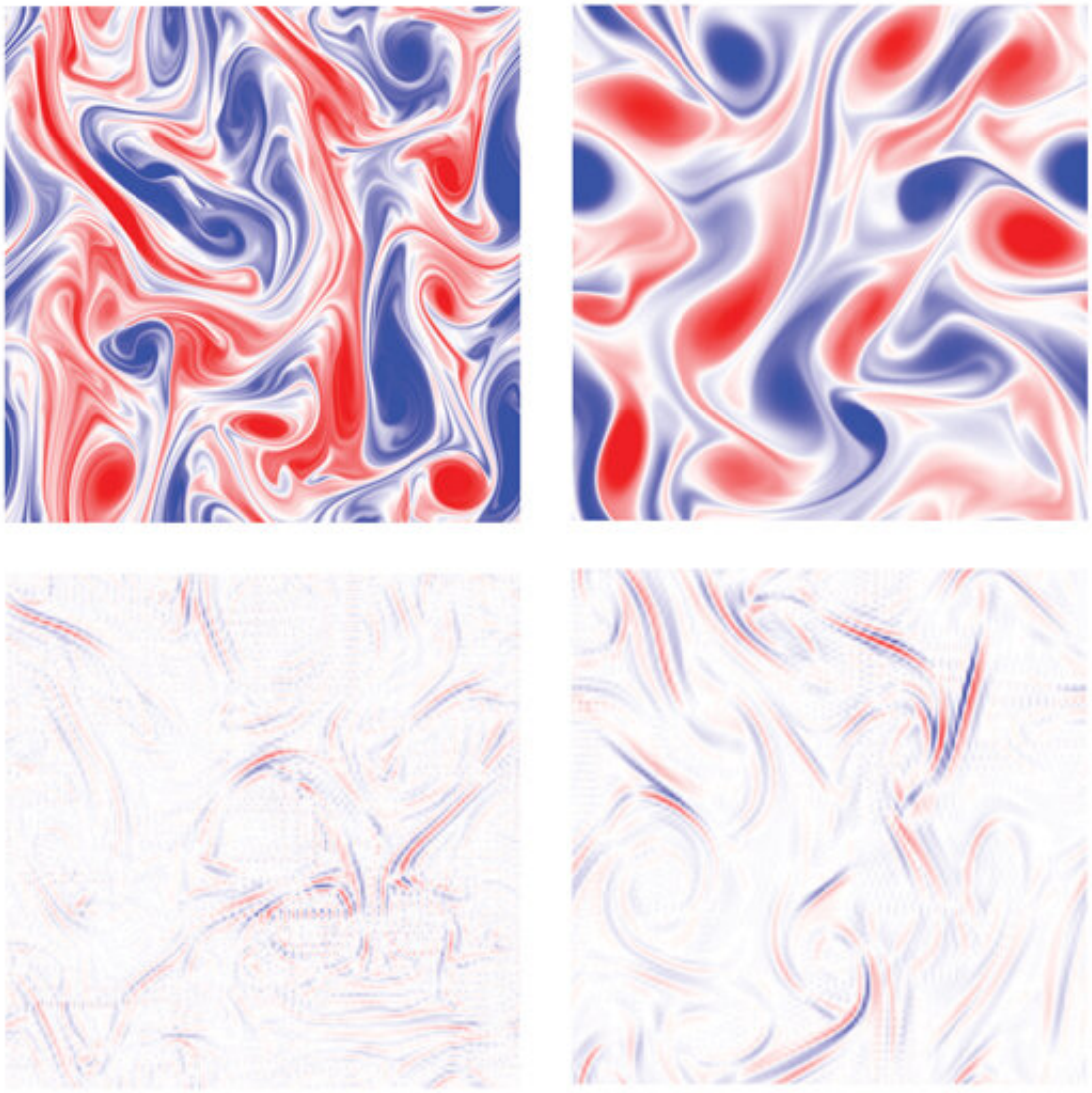
Scientific AI's 'black box' is no match for 200-year-old method

February 13 2023, by Jade Boyd



Rice University researchers trained a form of artificial intelligence called a deep learning neural network to recognize complex flows of air or water and predict

how flows will change over time. This visual illustrates the substantial differences in the scale of features the model is shown during training (top) and the features it learns to recognize (bottom) to make its predictions. Credit: of P. Hassanzadeh/Rice University

One of the oldest tools in computational physics—a 200-year-old mathematical technique known as Fourier analysis—can reveal crucial information about how a form of artificial intelligence called a deep neural network learns to perform tasks involving complex physics like climate and turbulence modeling, according to a new study.

The discovery by mechanical engineering researchers at Rice University is described in an open-access study published in *PNAS Nexus*.

"This is the first rigorous framework to explain and guide the use of deep neural networks for complex dynamical systems such as climate," said study corresponding author Pedram Hassanzadeh. "It could substantially accelerate the use of scientific deep learning in climate science, and lead to much more reliable climate change projections."

In the paper, Hassanzadeh, Adam Subel and Ashesh Chattopadhyay, both former students, and Yifei Guan, a postdoctoral research associate, detailed their use of Fourier analysis to study a deep learning neural network that was trained to recognize complex flows of air in the atmosphere or water in the ocean and to predict how those flows would change over time.

Their analysis revealed "not only what the neural network had learned, it also enabled us to directly connect what the network had learned to the physics of the complex system it was modeling," Hassanzadeh said.

"Deep neural networks are infamously hard to understand and are often considered 'black boxes,'" he said. "That is one of the major concerns with using deep neural networks in scientific applications. The other is generalizability: These networks cannot work for a system that is different from the one for which they were trained."

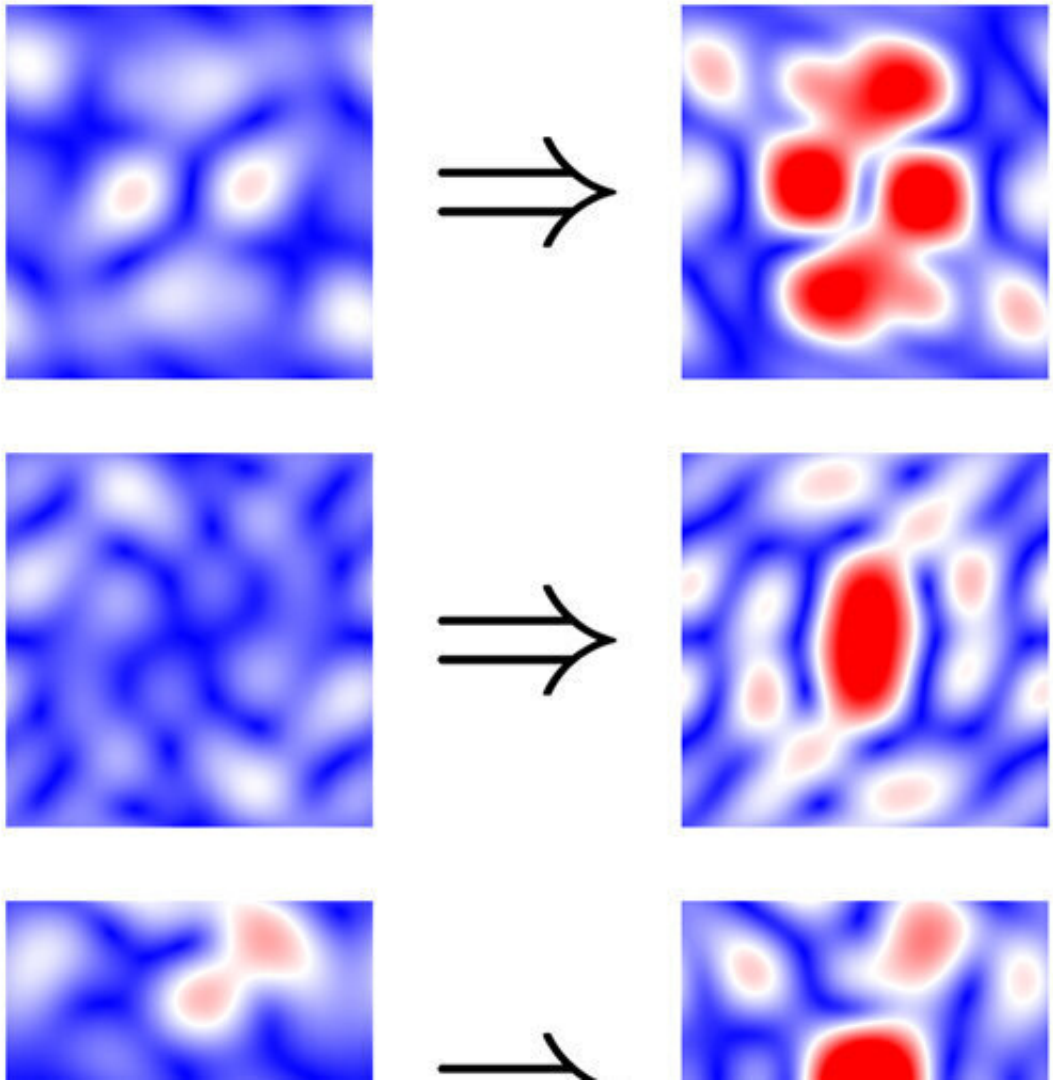
Hassanzadeh said the analytic framework his team presents in the paper "opens up the black box, lets us look inside to understand what the networks have learned and why, and also lets us connect that to the physics of the system that was learned."

Subel, the study's lead author, began the research as a Rice undergraduate and is now a graduate student at New York University. He said the framework could be used in combination with techniques for transfer learning to "enable generalization and ultimately increase the trustworthiness of scientific deep learning."

While many prior studies had attempted to reveal how deep learning networks learn to make predictions, Hassanzadeh said he, Subel, Guan and Chattopadhyay chose to approach the problem from a different perspective.

"The common machine learning tools for understanding neural networks have not shown much success for natural and engineering system applications, at least such that the findings could be connected to the physics," Hassanzadeh said. "Our thought was, 'Let's do something different. Let's use a tool that's common for studying physics and apply it to the study of a neural network that has learned to do physics'."

He said Fourier analysis, which was first proposed in the 1820s, is a favorite technique of physicists and mathematicians for identifying frequency patterns in space and time.



Training cutting-edge deep neural networks requires a great deal of data, and the burden for re-training, with current methods, is still significant. After training and re-training a deep learning network to perform different tasks involving complex physics, Rice University researchers used Fourier analysis to compare all 40,000 kernels from the two iterations and found more than 99% were similar. This illustration shows the Fourier spectra of the four kernels that most differed before (left) and after (right) re-training. The findings demonstrate the method's potential for identifying more efficient paths for re-training that require significantly less data. Credit: P. Hassanzadeh/Rice University

"People who do physics almost always look at data in the Fourier space," he said. "It makes physics and math easier."

For example, if someone had a minute-by-minute record of outdoor temperature readings for a one-year period, the information would be a string of 525,600 numbers, a type of data set physicists call a time series. To analyze the time series in Fourier space, a researcher would use trigonometry to transform each number in the series, creating another set of 525,600 numbers that would contain information from the original set but look quite different.

"Instead of seeing temperature at every minute, you would see just a few spikes," Subel said. "One would be the cosine of 24 hours, which would be the day and night cycle of highs and lows. That signal was there all along in the time series, but Fourier analysis allows you to easily see those types of signals in both time and space."

Based on this method, scientists have developed other tools for time-frequency analysis. For example, low-pass transformations filter out background noise, and high-pass filters do the inverse, allowing one to focus on the background.

Hassanzadeh's team first performed the Fourier transformation on the equation of its fully trained deep-learning model. Each of the model's approximately 1 million parameters act like multipliers, applying more or less weight to specific operations in the equation during model calculations. In an untrained model, parameters have random values.

These are adjusted and honed during training as the algorithm gradually learns to arrive at predictions that are closer and closer to the known outcomes in training cases. Structurally, the model parameters are grouped in some 40,000 five-by-five matrices, or kernels.

"When we took the Fourier transform of the equation, that told us we should look at the Fourier transform of these matrices," Hassanzadeh said. "We didn't know that. Nobody has done this part ever before, looked at the Fourier transforms of these matrices and tried to connect them to the physics.

"And when we did that, it popped out that what the neural network is learning is a combination of low-pass filters, high-pass filters and Gabor filters," he said.

"The beautiful thing about this is, the neural network is not doing any magic," Hassanzadeh said. "It's not doing anything crazy. It's actually doing what a physicist or mathematician might have tried to do. Of course, without the power of neural nets, we did not know how to correctly combine these filters. But when we talk to physicists about this work, they love it. Because they are, like, 'Oh! I know what these things are. This is what the neural network has learned. I see.'"

Subel said the findings have important implications for scientific deep learning, and even suggest that some things scientists have learned from studying machine learning in other contexts, like classification of static images, may not apply to scientific machine learning.

"We found that some of the knowledge and conclusions in the machine learning literature that were obtained from work on commercial and medical applications, for example, do not apply to many critical applications in science and engineering, such as climate change modeling," Subel said. "This, on its own, is a major implication."

More information: Adam Subel et al, Explaining the physics of transfer learning in data-driven turbulence modeling, *PNAS Nexus* (2023). DOI: 10.1093/pnasnexus/pgad015

Provided by Rice University