
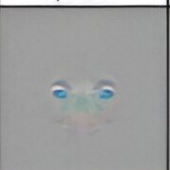


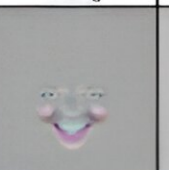



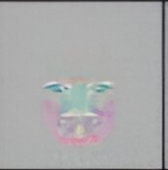
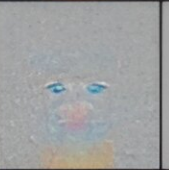







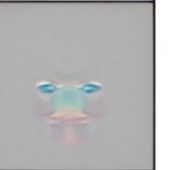

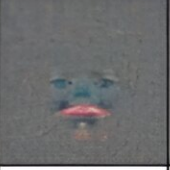
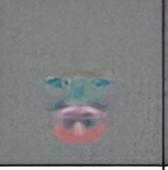

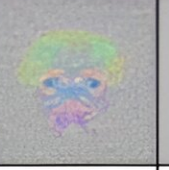



How do neural networks learn? A mathematical formula explains how they detect relevant patterns

March 12 2024

Task	Upstick	Eyebrows	5 o'clock shadow	Necktie	Smiling	Rosy Cheeks
First Layer NFM (Top Eigenvector)						
AGOP of Laplace Kernel (Top Eigenvector)						
Cosine Similarity	0.999	0.999	0.999	0.999	0.999	0.999

Task	Glasses	Mustache	Goatee	Hat	Blonde	Male
First Layer NFM (Top Eigenvector)						
AGOP of Laplace Kernel (Top Eigenvector)						
Cosine Similarity	0.999	0.999	0.999	0.995	0.997	0.999

Top eigenvector of AGOP of two separate models, MLPs and Laplace kernel machines, captured similar features (cosine similarity greater than .99) when trained on the same data from CelebA across various tasks. Credit: *Science* (2024). DOI: 10.1126/science.adi5639

Neural networks have been powering breakthroughs in artificial intelligence, including the large language models that are now being used in a wide range of applications, from finance, to human resources to health care. But these networks remain a black box whose inner workings engineers and scientists struggle to understand.

Now, a team led by data and computer scientists at the University of California San Diego has given neural networks the equivalent of an X-ray to uncover how they actually learn.

The researchers found that a formula used in [statistical analysis](#) provides a streamlined mathematical description of how neural networks, such as GPT-2, a precursor to ChatGPT, learn relevant patterns in data, known as features. This formula also explains how neural networks use these relevant patterns to make predictions.

"We are trying to understand neural networks from first principles," said Daniel Beaglehole, a Ph.D. student in the UC San Diego Department of Computer Science and Engineering and co-first author of the study.

"With our formula, one can simply interpret which features the network is using to make predictions."

The team [present their findings](#) in the journal *Science*.

Why does this matter? AI-powered tools are now pervasive in everyday life. Banks use them to approve loans. Hospitals use them to analyze medical data, such as X-rays and MRIs. Companies use them to screen job applicants. But it's currently difficult to understand the mechanism neural networks use to make decisions and the biases in the training data that might impact this.

"If you don't understand how neural networks learn, it's very hard to establish whether neural networks produce reliable, accurate, and

appropriate responses," said Mikhail Belkin, the paper's corresponding author and a professor at the UC San Diego Halicioglu Data Science Institute. "This is particularly significant given the rapid recent growth of machine learning and neural net technology."

The study is part of a larger effort in Belkin's research group to develop a [mathematical theory](#) that explains how neural networks work.

"Technology has outpaced theory by a huge amount," he said. "We need to catch up."

The team also showed that the statistical formula they used to understand how neural networks learn, known as Average Gradient Outer Product (AGOP), could be applied to improve performance and efficiency in other types of machine learning architectures that do not include neural networks.

"If we understand the underlying mechanisms that drive neural networks, we should be able to build machine learning models that are simpler, more efficient and more interpretable," Belkin said. "We hope this will help democratize AI."

The machine learning systems that Belkin envisions would need less [computational power](#), and therefore less power from the grid, to function. These systems also would be less complex and so easier to understand.

Illustrating the new findings with an example

(Artificial) neural networks are computational tools to learn relationships between data characteristics (i.e. identifying specific objects or faces in an image). One example of a task is determining whether in a new image a person is wearing glasses or not. Machine learning approaches this problem by providing the neural network many example (training)

images labeled as images of "a person wearing glasses" or "a person not wearing glasses."

The neural network learns the relationship between images and their labels, and extracts data patterns, or features, that it needs to focus on to make a determination. One of the reasons AI systems are considered a [black box](#) is because it is often difficult to describe mathematically what criteria the systems are actually using to make their predictions, including potential biases. The new work provides a simple mathematical explanation for how the systems are learning these features.

Features are relevant patterns in the data. In the example above, there are a wide range of features that the neural networks learns, and then uses, to determine if in fact a person in a photograph is wearing glasses or not.

One feature it would need to pay attention to for this task is the upper part of the face. Other features could be the eye or the nose area where glasses often rest. The network selectively pays attention to the features that it learns are relevant and then discards the other parts of the image, such as the lower part of the face, the hair and so on.

Feature learning is the ability to recognize relevant patterns in data and then use those patterns to make predictions. In the glasses example, the network learns to pay attention to the upper part of the face. In the new *Science* paper, the researchers identified a statistical formula that describes how the neural networks are learning features.

Alternative neural network architectures: The researchers went on to show that inserting this formula into computing systems that do not rely on [neural networks](#) allowed these systems to learn faster and more efficiently.

"How do I ignore what's not necessary? Humans are good at this," said Belkin. "Machines are doing the same thing. Large Language Models, for example, are implementing this 'selective paying attention' and we haven't known how they do it. In our *Science* paper, we present a mechanism explaining at least some of how the neural nets are 'selectively paying attention.'"

More information: Adityanarayanan Radhakrishnan et al, Mechanism for feature learning in neural networks and backpropagation-free machine learning models, *Science* (2024). [DOI: 10.1126/science.adi5639](https://doi.org/10.1126/science.adi5639)

Provided by University of California - San Diego

Citation: How do neural networks learn? A mathematical formula explains how they detect relevant patterns (2024, March 12) retrieved 27 April 2024 from <https://phys.org/news/2024-03-neural-networks-mathematical-formula-relevant.html>

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.