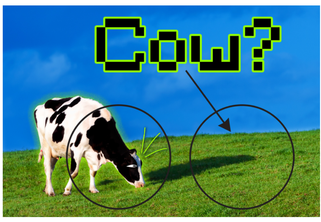
Avoiding shortcut solutions in artificial intelligence

A new method forces a machine learning model to focus on more data when learning a task, which leads to more reliable predictions.

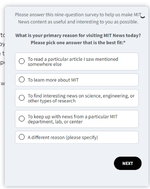
Adam Zewe | MIT News Office
November 2, 2021



PRESS INQUIRIES

A model might make a shortc
to identify images of cows by
green grass that appears in t
than the more complex shap
the cows.

Image: Jose-Luis Olivares, MIT, w
iStock/photo



If your Uber driver takes a shortcut, you might get to your destination faster. But if a machine learning model takes a shortcut, it might fail in unexpected ways.

In machine learning, a shortcut solution occurs when the model relies on a simple characteristic of a dataset to make a decision, rather than learning the true essence of the data, which can lead to inaccurate predictions. For example, a model might learn to identify images of cows by focusing on the green grass that appears in the photos, rather than the more complex shapes and patterns of the cows.

A new study by researchers at MIT explores the problem of shortcuts in a popular machine-learning method and proposes a solution that can prevent shortcuts by forcing the model to use more data in its decision-making.

By removing the simpler characteristics the model is focusing on, the researchers force it to focus on more complex features of the data that it hadn't been considering. Then, by asking the model to solve the same task two ways — once using those simpler features, and then also using the complex features it has now learned to identify — they reduce the tendency for shortcut solutions and boost the performance of the model.

One potential application of this work is to enhance the effectiveness of machine learning models that are used to identify disease in medical images. Shortcut solutions in this context could lead to false diagnoses and have dangerous implications for patients.

"It is still difficult to tell why deep networks make the decisions that they do, and in particular, which parts of the data these networks choose to focus upon when making a decision. If we can understand how shortcuts work in further detail, we can go even farther to answer some of the fundamental but very practical questions that are really important to people who are trying to deploy these networks," says Joshua Robinson, a PhD student in the Computer Science and Artificial Intelligence Laboratory (CSAIL) and lead author of the paper.

Robinson wrote the paper with his advisors, senior author Suvrit Sra, the Esther and Harold E. Edgerton Career Development Associate Professor in the Department of Electrical Engineering and Computer Science (EECS) and a core member of the Institute for Data, Systems, and Society (IDSS) and the Laboratory for Information and Decision Systems; and Stefanie Jegelka, the X-Consortium Career Development Associate Professor in EECS and a member of CSAIL and IDSS; as well as University of Pittsburgh assistant professor Kayhan Batmanghelich and PhD students Li Sun and Ke Yu. The research will be presented at the Conference on Neural Information Processing Systems in December.

The long road to understanding shortcuts

The researchers focused their study on contrastive learning, which is a powerful form of self-supervised machine learning. In self-supervised machine learning, a model is trained using raw data that do not have label descriptions from humans. It can therefore be used successfully for a larger variety of data.

A self-supervised learning model learns useful representations of data, which are used as inputs for different tasks, like image classification. But if the model takes shortcuts and fails to capture important information, these tasks won't be able to use that information either.

For example, if a self-supervised learning model is trained to classify pneumonia in X-rays from a number of hospitals, but it learns to make predictions based on a tag that identifies the hospital the scan came from (because some hospitals have more pneumonia cases than others), the model won't perform well when it is given data from a new hospital.

For contrastive learning models, an encoder algorithm is trained to discriminate between pairs of similar inputs and pairs of dissimilar inputs. This process encodes rich and complex data, like images, in a way that the contrastive learning model can interpret.

The researchers tested contrastive learning encoders with a series of images and found that, during this training procedure, they also fall prey to shortcut solutions. The encoders tend to focus on the simplest features of an image to decide which pairs of inputs are similar and which are dissimilar. Ideally, the encoder should focus on all the useful characteristics of the data when making a decision, Jegelka says.

So, the team made it harder to tell the difference between the similar and dissimilar pairs, and found that this changes which features the encoder will look at to make a decision.

"If you make the task of discriminating between similar and dissimilar items harder and harder, then your system is forced to learn more meaningful information in the data, because without learning that it cannot solve the task," she says.

But increasing this difficulty resulted in a tradeoff — the encoder got better at focusing on some features of the data but became worse at focusing on others. It almost seemed to forget the simpler features, Robinson says.

To avoid this tradeoff, the researchers asked the encoder to discriminate between the pairs the same way it had originally, using the simpler features, and also after the researchers removed the information it had already learned. Solving the task both ways simultaneously caused the encoder to improve across all features.

Their method, called implicit feature modification, adaptively modifies samples to remove the simpler features the encoder is using to discriminate between the pairs. The technique does not rely on human input, which is important because real-world data sets can have hundreds of different features that could combine in complex ways, Sra explains.

From cars to COPD

The researchers ran one test of this method using images of vehicles. They used implicit feature modification to adjust the color, orientation, and vehicle type to make it harder for the encoder to discriminate between similar and dissimilar pairs of images. The encoder improved its accuracy across all three features — texture, shape, and color — simultaneously.

To see if the method would stand up to more complex data, the researchers also tested it with samples from a medical image database of chronic obstructive pulmonary disease (COPD). Again, the method led to simultaneous improvements across all features they evaluated.

While this work takes some important steps forward in understanding the causes of shortcut solutions and working to solve them, the researchers say that continuing to refine these methods and applying them to other types of self-supervised learning will be key to future advancements.

"This ties into some of the biggest questions about deep learning systems, like 'Why do they fail?' and 'Can we know in advance the situations where your model will fail?' There is still a lot farther to go if you want to understand shortcut learning in its full generality," Robinson says.

This research is supported by the National Science Foundation, National Institutes of Health, and the Pennsylvania Department of Health's SAP SE Commonwealth Universal Research Enhancement (CURE) program.

SHARE

Paper: "Can contrastive learning avoid shortcut solutions?"

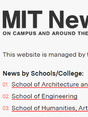
RELATED LINKS

01 Joshua Robinson
02 Suvrit Sra
03 Stefanie Jegelka
04 Laboratory for Information and Decision Systems
05 Institute for Data, Systems, and Society
06 Computer Science and Artificial Intelligence Laboratory
07 Department of Electrical Engineering and Computer Science
08 School of Engineering
09 MIT Schwarzman College of Computing

RELATED TOPICS

Research | Machine learning | Algorithms | Computer science and technology | Electrical Engineering & Computer Science (eecs) | IDSS | Laboratory for Information and Decision Systems (LIDS) | Computer Science and Artificial Intelligence Laboratory (CSAIL) | MIT Schwarzman College of Computing | School of Engineering | National Science Foundation (NSF) | National Institutes of Health (NIH)

RELATED ARTICLES


Artificial networks learn to smell like the brain

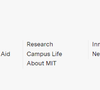

These neural networks know what they're doing

Using AI and old reports to understand new medical images

Data diversity

More MIT News

Researchers teach an AI to write better chart captions
A new dataset can help scientists develop automatic systems that generate richer, more descriptive captions for online charts.



Q&A: A conversation with Helen Elaine Lee about her novel, "Pomegranate"
The MIT professor's new book explores the world of a woman set free from prison and redefining herself in society.

Transatlantic connections make the difference for MIT Portugal
The international partnership focuses on climate and sustainability.

Summer 2023 recommended reading from MIT
Enjoy these recent titles from Institute faculty and staff.

Studies at the intersection of equity, computing, and education
"The work I'm doing is deeply rooted in the belief that you can plant seeds in people," says graduate student Cecilé Sadler.

Computer vision system marries image recognition and generation
MAGE merges the two key tasks of image generation and recognition, typically trained separately, into a single system.

More news on MIT News homepage →