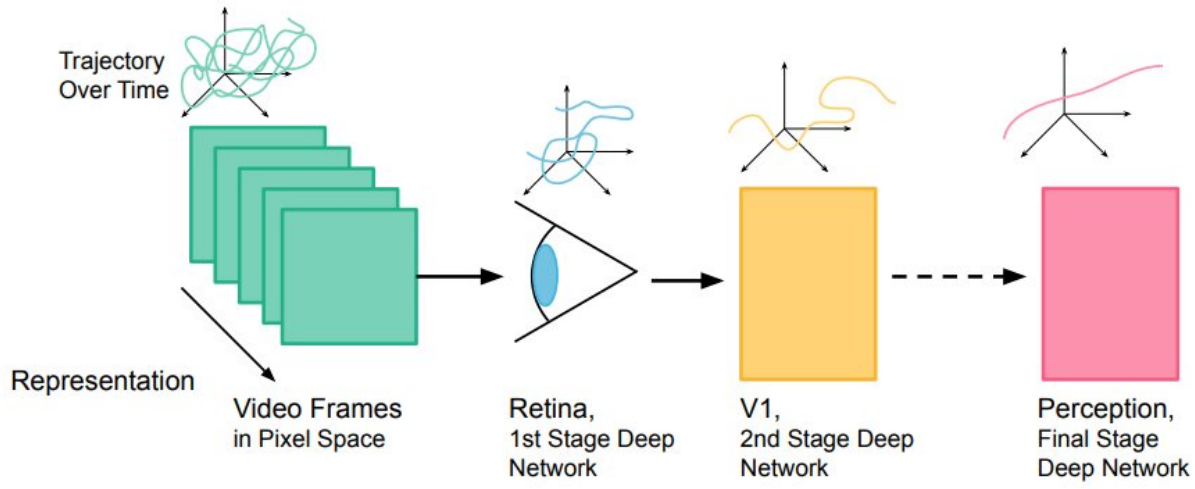
Training machines to learn more like humans do

May 9 2023, by Adam Zewe



Schematic illustration of the representation of a discrete video sequence becoming progressively straighter as information is processed through a visual processing pipeline, starting from the highly nonlinear trajectory of typical video frames in pixel space. Credit: Anne Harrington, Vasha DuTell, Ayush Tewari, Mark Hamilton, Simon Stent, Ruth Rosenholtz and William T. Freeman, https://openreview.net/pdf?id=4cOfD2qL6T

Imagine sitting on a park bench, watching someone stroll by. While the scene may constantly change as the person walks, the human brain can transform that dynamic visual information into a more stable representation over time. This ability, known as perceptual straightening, helps us predict the walking person's trajectory.

Unlike humans, computer vision models don't typically exhibit perceptual straightness, so they learn to represent visual information in a highly unpredictable way. But if machine-learning models had this ability, it might enable them to better estimate how objects or people will move.

MIT researchers have discovered that a specific training method can help computer vision models learn more perceptually straight representations, like humans do. Training involves showing a machine-learning model millions of examples so it can learn a task.

The researchers found that training computer vision models using a technique called adversarial training, which makes them less reactive to tiny errors added to images, improves the models' perceptual straightness.

The team also discovered that perceptual straightness is affected by the task one trains a model to perform. Models trained to perform abstract tasks, like classifying images, learn more perceptually straight representations than those trained to perform more fine-grained tasks, like assigning every pixel in an image to a category.

For example, the nodes within the model have internal activations that represent "dog," which allow the model to detect a dog when it sees any image of a dog. Perceptually straight representations retain a more stable "dog" representation when there are small changes in the image. This makes them more robust.

By gaining a better understanding of perceptual straightness in computer vision, the researchers hope to uncover insights that could help them develop models that make more accurate predictions. For instance, this property might improve the safety of autonomous vehicles that use computer vision models to predict the trajectories of pedestrians,

cyclists, and other vehicles.

"One of the take-home messages here is that taking inspiration from biological systems, such as human vision, can both give you insight about why certain things work the way that they do and also inspire ideas to improve neural networks," says Vasha DuTell, an MIT postdoc and co-author of a paper exploring perceptual straightness in computer vision.

Joining DuTell on the paper are lead author Anne Harrington, a graduate student in the Department of Electrical Engineering and Computer Science (EECS); Ayush Tewari, a postdoc; Mark Hamilton, a graduate student; Simon Stent, research manager at Woven Planet; Ruth Rosenholtz, principal research scientist in the Department of Brain and Cognitive Sciences and a member of the Computer Science and Artificial Intelligence Laboratory (CSAIL); and senior author William T. Freeman, the Thomas and Gerd Perkins Professor of Electrical Engineering and Computer Science and a member of CSAIL. The research was presented at the International Conference on Learning Representations.

Studying straightening

After reading a 2019 paper from a team of New York University researchers about perceptual straightness in humans, DuTell, Harrington, and their colleagues wondered if that property might be useful in computer vision models, too.

They set out to determine whether different types of computer vision models straighten the visual representations they learn. They fed each model frames of a video and then examined the representation at different stages in its learning process.

If the model's representation changes in a predictable way across the

frames of the video, that model is straightening. At the end, its output representation should be more stable than the input representation.

"You can think of the representation as a line, which starts off really curvy. A model that straightens can take that curvy line from the video and straighten it out through its processing steps," DuTell explains.

Most models they tested didn't straighten. Of the few that did, those which straightened most effectively had been trained for classification tasks using the technique known as adversarial training.

Adversarial training involves subtly modifying images by slightly changing each pixel. While a human wouldn't notice the difference, these minor changes can fool a machine so it misclassifies the image. Adversarial training makes the model more robust, so it won't be tricked by these manipulations.

Because adversarial training teaches the model to be less reactive to slight changes in images, this helps it learn a representation that is more predictable over time, Harrington explains.

"People have already had this idea that adversarial training might help you get your model to be more like a human, and it was interesting to see that carry over to another property that people hadn't tested before," she says.

But the researchers found that adversarially trained models only learn to straighten when they are trained for broad tasks, like classifying entire images into categories. Models tasked with segmentation—labeling every pixel in an image as a certain class—did not straighten, even when they were adversarially trained.

Consistent classification

The researchers tested these image classification models by showing them videos. They found that the models which learned more perceptually straight representations tended to correctly classify objects in the videos more consistently.

"To me, it is amazing that these adversarially trained models, which have never even seen a video and have never been trained on temporal data, still show some amount of straightening," DuTell says.

The researchers don't know exactly what about the adversarial training process enables a computer vision model to straighten, but their results suggest that stronger training schemes cause the models to straighten more, she explains.

Building off this work, the researchers want to use what they learned to create new training schemes that would explicitly give a model this property. They also want to dig deeper into adversarial training to understand why this process helps a model straighten.

"From a biological standpoint, adversarial training doesn't necessarily make sense. It's not how humans understand the world. There are still a lot of questions about why this training process seems to help models act more like humans," Harrington says.

"Understanding the representations learned by deep neural networks is critical to improve properties such as robustness and generalization," says Bill Lotter, assistant professor at the Dana-Farber Cancer Institute and Harvard Medical School, who was not involved with this research. "Harrington et al. perform an extensive evaluation of how the representations of computer vision models change over time when processing natural videos, showing that the curvature of these trajectories varies widely depending on model architecture, training properties, and task. These findings can inform the development of

improved models and also offer insights into biological visual processing."

"The paper confirms that straightening natural videos is a fairly unique property displayed by the human visual system. Only adversarially trained networks display it, which provides an interesting connection with another signature of human perception: its robustness to various image transformations, whether natural or artificial," says Olivier Hénaff, a research scientist at DeepMind, who was not involved with this research.

"That even adversarially trained scene segmentation models do not straighten their inputs raises important questions for future work: Do humans parse natural scenes in the same way as computer vision models? How to represent and predict the trajectories of objects in motion while remaining sensitive to their spatial detail? In connecting the straightening hypothesis with other aspects of visual behavior, the paper lays the groundwork for more unified theories of perception."

More information: Exploring perceptual straightness in learned visual representations. openreview.net/pdf?id=4cOfD2qL6T

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology