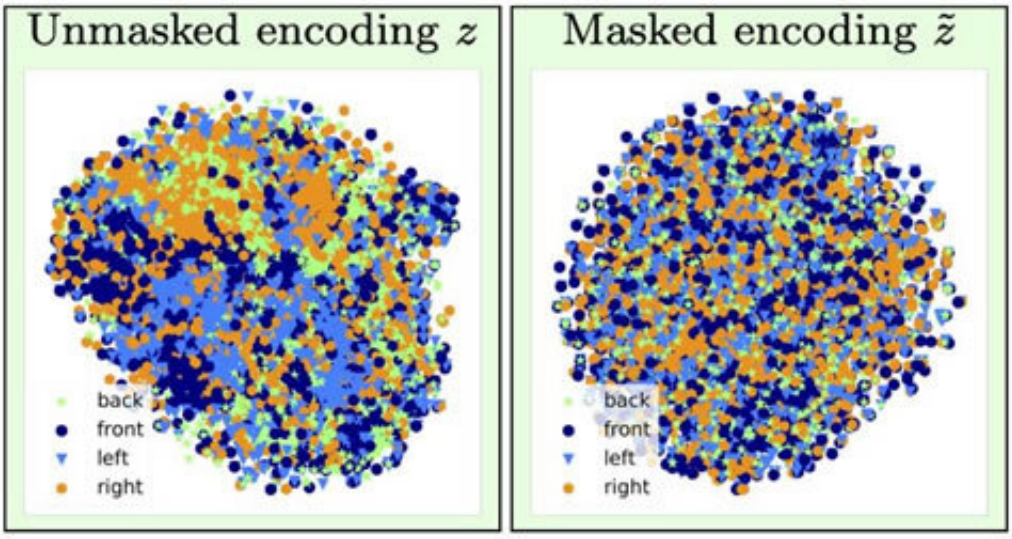
How do we remove biases in AI systems? Start by teaching them selective amnesia

February 25 2020



For identifying the type of a chair image, information about the chair's orientation (a nuisance factor) is lost by the forgetting operation (going from the left visualization to the right). Credit: University of Southern California

Imagine if the next time you apply for a loan, a computer algorithm determines you need to pay a higher rate based primarily on your race, gender or zip code.

Now, imagine it was possible to train an AI deep learning model to analyze that underlying data by inducing amnesia: it forgets certain data and only focuses on others.

If you're thinking that this sounds like the computer scientist's version of "The Eternal Sunshine of the Spotless Mind," you'd be pretty spot on. And thanks to AI researchers at USC's Information Sciences Institute (ISI), this concept, called adversarial forgetting, is now a real mechanism.

The importance of addressing and removing biases in AI is becoming more important as AI becomes increasingly prevalent in our daily lives, noted Ayush Jaiswal, the paper's lead author and Ph.D. candidate at the USC Viterbi School of Engineering.

"AI and, more specifically, machine learning models inherit biases present in the data they're trained on and are prone to even amplify those biases," he explained. "AI is being used to make several real-life decisions that affect all of us, [such as] determining credit limits, approving loans, scoring job applications, etc. If, for example, models for making these decisions are trained blindly on historical data without controlling for biases, they would learn to unfairly treat individuals that belong to historically disadvantaged sections of the population, such as women and people of color."

The research was led by Wael AbdAlmageed, research team leader at ISI and a research associate professor at USC Viterbi's Ming Hsieh Department of Electrical and Computer Engineering, and research associate professor Greg Ver Steeg, as well as Premkumar Natarajan, research professor of computer science and executive director of ISI (on leave). Under their guidance, Jaiswal and co-author Daniel Moyer, Ph.D., developed the adversarial forgetting approach, which teaches deep learning models to disregard specific, unwanted data factors so that the results they produce are unbiased and more accurate.

The research paper, titled "Invariant Representations through Adversarial Forgetting," was presented at the Association for the

Advancement for Artificial Intelligence conference in New York City on February 10, 2020.

Nuisances and Neural Networks

Deep learning is a core component in AI and can teach computers how to find correlations and make predictions with data, helping identify people or objects, for example. Models essentially look for associations between different features within data and the target that it's supposed to predict. If a model was tasked to find a specific person from a group, it would analyze facial features to tell everyone apart and then identify the targeted person. Simple, right?

Unfortunately, things don't always run so smoothly, as the model can end up learning things that may seem counterintuitive. It could associate your identity with a particular background or lighting setup and be unable to identify you if the lighting or background was altered; it could associate your handwriting to a certain word, and get confused if the same word was written in someone else's handwriting. These aptly-named nuisance factors aren't related to the task that you're trying to carry out, and misassociating them with the prediction target can actually end up being dangerous.

Models can also learn biases in data that are correlated with the prediction target but are undesired. For example, in tasks carried out by models involving historically collected socioeconomic data, such as determining credit scores, credit lines, and loan eligibility, the model can make false predictions and show biases by making connections between the biases and the prediction target. It can jump to the conclusion that since it's analyzing the data of a woman, she must have a low credit score; since it's analyzing the data of a person of color, they must not be eligible for a loan. There's no shortage of stories of banks coming under fire for their algorithms' biased decisions in how much they charge

people who've taken out loans based on their race, gender, and education, even if they have the exact same credit profile as someone in a more socially privileged population segment.

As Jaiswal explained, the adversarial forgetting mechanism "fixes" neural networks, which are powerful deep learning models that learn to predict targets from data. The credit limit you got on that new credit card you signed up for? A neural network likely analyzed your financial data to come up with that number.

The research team developed the adversarial forgetting mechanism so that it could first train the neural network to represent all the underlying aspects of the data that it's analyzing and then forget specified biases. In the example of the credit card limit, that would mean that the mechanism could teach the bank's algorithm to predict the limit while forgetting, or being invariant to, the particular data pertaining to gender or race. "[The mechanism] can be used train neural networks to be invariant to known biases in training datasets," Jaiswal said. "This, in turn, would result in trained models that wouldn't be biased while making decisions."

Deep learning algorithms are great at learning things, but it's more difficult to make sure that the algorithms don't learn certain things. Developing algorithms is a very data-driven process, and data tends to contain biases.

But can't we just take out all the data on race, gender, and education to remove the biases?

Not entirely. There are many other factors of data that are correlated with these sensitive factors that are important for algorithms to analyze. The key, as the ISI AI researchers found, is adding constraints in the model's training process to force the model to make predictions while

being invariant to specific factors of data-essentially, selective forgetting.

Battling Biases

Invariance refers to the ability to identify a specific object even if its appearance (i.e., data) is altered in some way, and Jaiswal and his colleagues began thinking about how this concept could be applied to improve algorithms. "My co-author, Dan [Moyer], and I actually came up with this idea somewhat naturally based on our previous experiences in the field of invariant representation learning," he remarked. But fleshing out the concept was no simple task. "The most challenging parts were [the] rigorous comparison with previous works in this domain on a wide array of datasets (which required running a very large number of experiments) and [developing] a theoretical analysis of the forgetting process," he said.

The adversarial forgetting mechanism can be also used to help improve content generation in a variety of fields. "The budding field of fair machine learning looks at ways of reducing bias in algorithmic decision making based on consumer data," said Ver Steeg. "A more speculative area involves research on using AI to generate content including attempts at books, music, art, games, and even recipes. For content generation to succeed, we need new ways to control and manipulate neural network representations and the forgetting mechanism could be a way of doing that."

So how do biases even show up in the model in the first place?

Most models use historical data, which, unfortunately, can be largely biased towards traditionally marginalized communities like women, minorities, even certain zip codes. It's costly and cumbersome to collect data, so scientists tend to resort to data that already exists and train

models based on that, which is how biases enter the picture.

The good news is that these biases are being recognized, and while the problem is far from being solved, strides are being made to understand and address these issues. "*n the research community, people are definitely becoming increasingly conscious of dataset biases, and designing and analyzing collection protocols to control for known biases," said Jaiswal. "The study of biases and fairness in machine learning has grown rapidly as a research field in the past few years."*

Determination of which factors should be considered irrelevant or biased are made by domain experts and based on statistical analysis. "So far, invariance has mostly been used to remove factors that are widely considered unwanted/irrelevant within the research community based on statistical evidence," Jaiswal stated.

However, since researchers determine what is irrelevant or biased, there can be a potential for those determinations to turn into biases themselves. This is a factor that researchers are working on as well. "Figuring out which factors to forget is a critical problem that can easily lead to unintended consequences," noted Ver Steeg. "A recent Nature piece on fair learning points out that we have to understand the mechanisms behind discrimination if we hope to correctly specify algorithmic solutions."

Human information processing is extremely intricate, and the adversarial forgetting mechanism helps us get one step closer to developing AI that can think like we do. As Ver Steeg remarked, humans tend to separate different forms of information about the world around them by instinct-getting algorithms to do the same is the challenge at hand.

"If someone steps in front of your car, you slam on the breaks and the slogan on their shirt doesn't even enter your mind," said Ver Steeg. "But if you met that person in a social context, that information might be relevant

and help you strike up a conversation. For AI, different types of information are all mashed together. If we can teach neural networks to separate concepts that are useful for different tasks, we hope it leads AI to a more human understanding of the world."

Human information processing is extremely intricate, and the adversarial forgetting mechanism helps us get one step closer to developing AI that can think like we do. As Ver Steeg remarked, humans tend to separate different forms of information about the world around them by instinct—getting algorithms to do the same is the challenge at hand.

"If someone steps in front of your car, you slam on the breaks and the slogan on their shirt doesn't even enter your mind," said Ver Steeg. "But if you met that person in a social context, that information might be relevant and help you strike up a conversation. For AI, different types of information are all mashed together. If we can teach neural networks to separate concepts that are useful for different tasks, we hope it leads AI to a more human understanding of the world."

More information: Invariant Representations through Adversarial Forgetting, arXiv:1911.04060 [cs.LG] arxiv.org/abs/1911.04060

Provided by University of Southern California