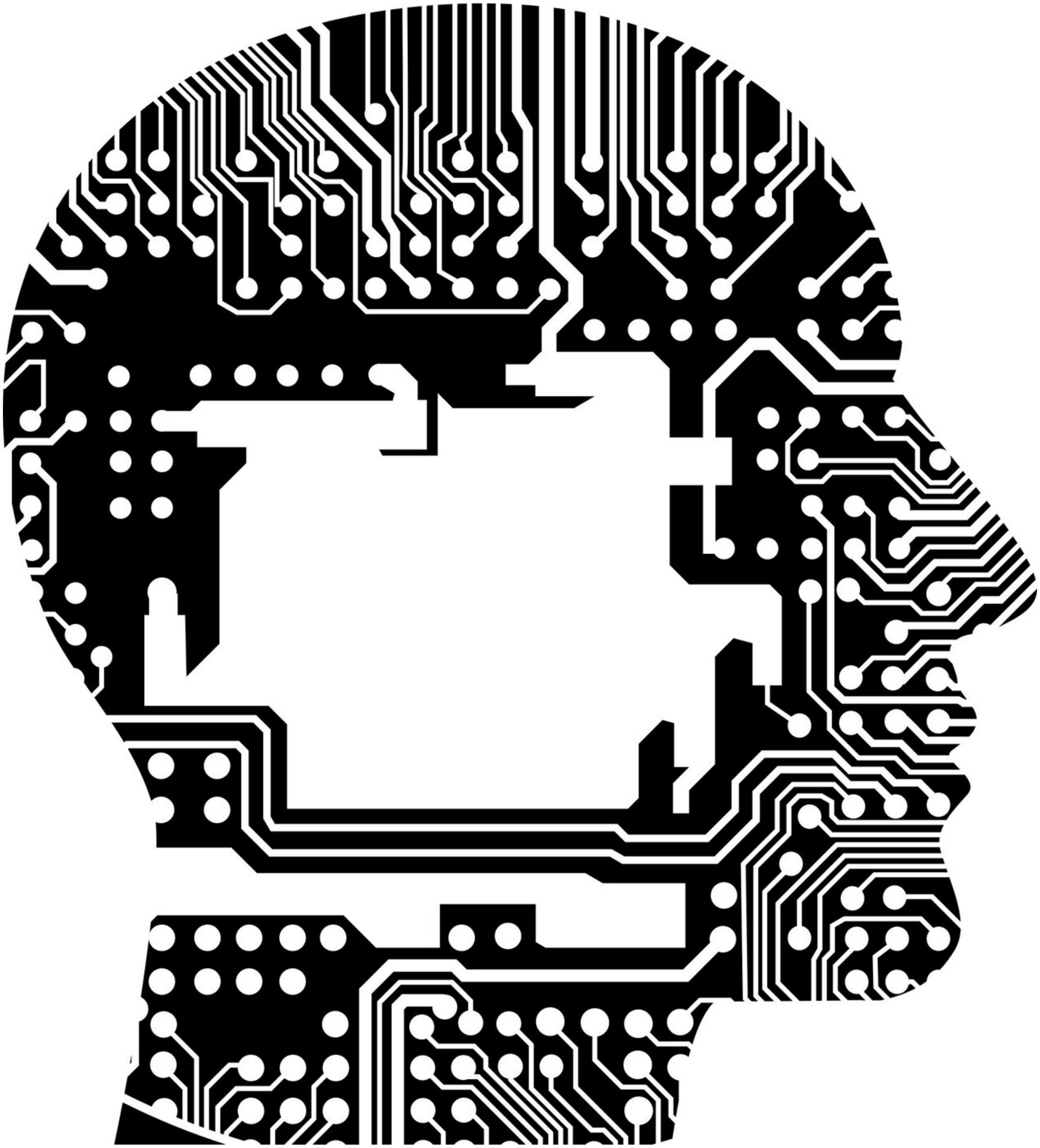
TechXplore

Teaching AI to overcome human bias

July 31 2019, by Leah Burrows



Credit: CC0 Public Domain

Are you smarter than a machine learning model? Let's find out. Choose the answer that contradicts the following premise:

Bob has a sister named Sarah.

- A) Bob has a sister.
- B) Bob doesn't own a car.
- C) Bob doesn't have a sister.

If you chose C, congratulations!

Examples like this might look simple but they seem to be a good indicator of a machine's understanding of language. The test is called Natural Language Inference and it's often used to gauge a model's ability to understand a relationship between two texts. Possible relationships are entailment (as in example A), neutral (B), and contradiction (C).

Datasets with hundreds of thousands of these questions, generated by humans, have led to an explosion of new neural network architectures for solving Natural Language Inference. Over the years, these neural networks have gotten better and better. Today's state-of-the-art models usually get the equivalent of a B+ on these tests. Humans usually score an A or A-.

But researchers recently discovered that machine learning models still do remarkably well when they're given only the answer, also called the hypothesis, without the original premise. For example, a model given only "Bob doesn't have a sister" will guess that this is a contradictory hypothesis, even if it isn't given the premise "Bob has a sister named Sarah."

As it turns out, these datasets are rife with human biases. When asked to come up with contradictory sentences, humans often use negations, like "don't" or "nobody." However, relying on these clues might lead machine learning models also to incorrectly label "Bob doesn't own a car" a contradiction.

"These models aren't learning to understand the relationship between texts, they are learning to capture human idiosyncrasies," said Yonatan Belinkov, first author of the paper and a Postdoctoral Fellow in Computer Science at the Harvard John A. Paulson School of Engineering and Applied Sciences (SEAS).

To combat this, Belinkov and colleagues developed a new method to build machine learning models that reduces the model's reliance on these biases.

The team is presenting their research at the 57th Annual Meeting of the Association for Computational Linguistics (ACL) in Florence, Italy on July 28th—August 2nd.

It's common to model the typical Natural Language Inference test as a single stream—the premise and hypothesis are both processed together and fed to a classifier which predicts contradiction, neutral or entailment.

The team added a second stream to the model, this one with only the hypothesis. The model learns to perform Natural Language Inference with both streams simultaneously, but if it does well on the hypothesis-only side, it's penalized. This approach encourages the model to focus more on the premise side and refrain from learning the biases that led to successful hypothesis-only performance.

"Our hope is that with this method, the model isn't just focused on biased words, like "no" or "doesn't," but rather it's learned something deeper," said Stuart Shieber, James O. Welch, Jr. and Virginia B. Welch Professor of Computer Science at SEAS and co-author of the paper.

Those biases, however, can also be important context clues to solving the problem, so it's critical not to devalue them too much.

"There is a thin line between bias and usefulness," said Gabriel Grand, CS '18, who worked on the project as part of his undergraduate thesis. "Reaching peak performance means forgetting a lot of assumptions but not all of them."

(Grand's thesis, "Learning Interpretable and Bias-Free Models for Visual Question Answering" was awarded the 2018-2019 Thomas Temple Hoopes Prize for outstanding scholarly work or research.)

By removing many of these assumptions, the two-stream model unsurprisingly did slightly worse on the data that it was trained on than the model which wasn't penalized for relying on biases. However, when tested on new datasets—with different biases—the model did significantly better.

"Even though the model did a few percentage points worse on its own dataset, it has learned not to rely on biases as much. So, this method produces a model that performs more generally and is more robust," said Shieber.

This method may apply to a range of artificial intelligence tasks that require identifying deeper relationships—such as visual question answering, reading comprehension, and other natural language tasks—while avoiding superficial biases.

More information: Don't Take the Premise for Granted: Mitigating Artifacts in Natural Language Inference. dash.harvard.edu/handle/1/40827357

Provided by Harvard University

Citation: Teaching AI to overcome human bias (2019, July 31) retrieved 25 April 2024 from https://techxplore.com/news/2019-07-ai-human-bias.html