Zeroing in on the origins of bias in large language models

January 15 2024, by Harini Barath

Computer science PhD candidate Weicheng Ma is a co-author of the study. Credit: Katie Lenhart

When artificial intelligence models pore over hundreds of gigabytes of training data to learn the nuances of language, they also imbibe the

biases woven into the texts.

Computer science researchers at Dartmouth are devising ways to home in on the parts of the model that encode these biases, paving the way to mitigating, if not removing them altogether.

In a recent paper published in the *Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing*, co-authors Weicheng Ma, a computer science Ph.D. candidate at the Guarini School of Graduate and Advanced Studies, and Soroush Vosoughi, assistant professor of computer science, look at how stereotypes are encoded in pretrained large language models.

A large language model, or neural network, is a deep learning algorithm designed to process, understand, and generate text and other content when trained on huge datasets.

Pretrained models have biases, like stereotypes, baked into them, says Vosoughi. These can be generally positive (suggesting, for instance, that a particular group are good at certain skills) or negative (assuming that someone holds a certain occupation based on their gender).

And machine learning models are poised to permeate everyday life in a variety of ways. They can help hiring managers sift through stacks of resumes, facilitate faster approvals, or rejections, of bank loans, and provide counsel during parole decisions.

But built-in stereotypes based on demographics would engender unfair and undesirable outcomes. To mitigate such effects, "we ask whether we can do anything about the stereotypes even after a model has been trained," says Vosoughi.

The researchers began with a hypothesis that stereotypes, like other

linguistic features and patterns, are encoded in specific parts of the neural network model known as "attention heads." These are similar to a group of neurons; they allow a machine learning program to memorize multiple words provided to it as input, among other functions, some of which are still not fully understood.

Ma, Vosoughi, and their collaborators created a dataset heavy with stereotypes and used it to repeatedly tune 60 different pretrained large-language models including BERT and T5. By amplifying the model's stereotypes, the dataset acted like a detector, spotlighting the attention heads that did the heavy lifting in encoding these biases.

In their paper, the researchers show that pruning the worst offenders significantly reduces stereotypes in the large language models, without significantly affecting their linguistic abilities.

"Our finding disrupts the traditional view that advancements in AI and Natural Language Processing necessitate extensive training or complex algorithmic interventions," says Ma. Since the technique is not intrinsically language- or model-specific, it would be broadly applicable, according to Ma.

What's more, Vosoughi adds, the dataset can be tweaked to reveal some stereotypes but leave others undisturbed—"it's not a one size fits all."

So, a medical diagnosis model, in which age- or gender-based differences can be important for patient evaluation, would use a different version of the dataset than one used to remove bias from a model that picks out potential job candidates.

The technique only works when there is access to the fully trained model and will not apply to black box models, such as OpenAI's chatbot, ChatGPT, whose internal workings are invisible to users and researchers.

Adapting the present approach to black box models is their immediate next step, says Ma.

More information: Weicheng Ma et al, Deciphering Stereotypes in Pre-Trained Language Models, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (2023). DOI: 10.18653/v1/2023.emnlp-main.697