Case Study

Artificial Intelligence
Intel® Gaudi® 2 AI Processor

intel.

Intel Labs Mitigates AI Bias in Foundational Multimodal Models by 20 Percent

Deep Learning Capabilities of the Intel® Gaudi® 2 AI Processor Power Social Counterfactual Breakthrough

At a Glance

- Researchers at Intel Labs construct innovative datasets based on social counterfactuals, which yield unprecedented results in reducing the impact of gender, race, and other social attributes in AI.
- Large AI clusters equipped with 3rd Gen Intel® Xeon® Scalable processors and Intel® Gaudi® 2 AI accelerators are used to train foundational multimodal models and leverage results across data, text, images, and video.

> "By probing six models using data-intensive methods, the team mitigated biases by as much as 20 percent."
>
> **Vasudev Lal**
> Principal Research Scientist of Cognitive AI at Intel Labs

The Challenge

Teams all over the world have been building models to organize and interpret content for decades. AI is the latest application for these models and has, in the past three years, advanced at a rate that would have been difficult for anyone to fathom—including the researchers who are immersed in it on a daily basis. In 2023, industry produced 51 notable machine learning models, and academia contributed 15.[1]

AI researchers are charged with harnessing vast quantities of data, text, images, and video and building constructs that bring order to them—what we call multimodal foundational models. Beyond warehousing and categorizing information, researchers are creating new ways for people to interact with content, such as a chatbot querying a video to glean details, thus saving a user the time of viewing the full clip.

As foundational models become more fully featured—and as they move from research to production, where billions of people are using them—it's crucial that they minimize biases that can lead to unfair outcomes and perpetuate existing inequalities.[2] Researchers probing for intersectional social biases found that existing models contained a number of them related to gender, race, and other social attributes.

The challenge for researchers is how to create a benchmark to prove biases in the models and, subsequently, how to create methods to reduce those biases. As AI gains widespread acceptance in the commercial market, time is of the essence for addressing these challenges.

The Solution

Enter the counterfactual. Simply put, counterfactual inquiries use "if only" or "what if" reasoning to imagine alternative outcomes for past events. Put another way, they consider how the world would have to be different for a desirable outcome to occur. Most of us know, for instance, that pushing a glass bowl from a table results in the bowl breaking. The counterfactual would be that if the bowl were pushed onto a pile of pillows—or a carpeted surface—it would not break.

Counterfactuals originated in cognitive behavioral science and predate AI. When applied to AI, they explain what would have to be different in the input to change the outcome of an AI system. They help with understanding. And, it turns out, they're pivotal to removing bias.

Social counterfactuals study the intersection of social attributes. “Because you can’t describe an individual using just one social attribute, such as gender or race, we constructed a dataset using images where intersectional social attributes were varied,” explains Vasudev Lal, Principal Research Scientist of Cognitive AI at Intel Labs.

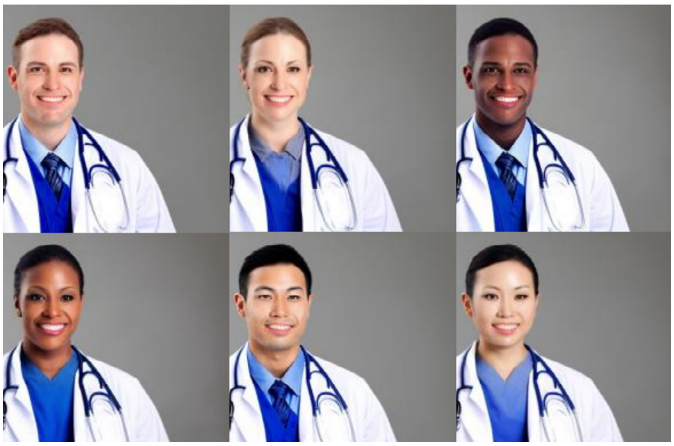


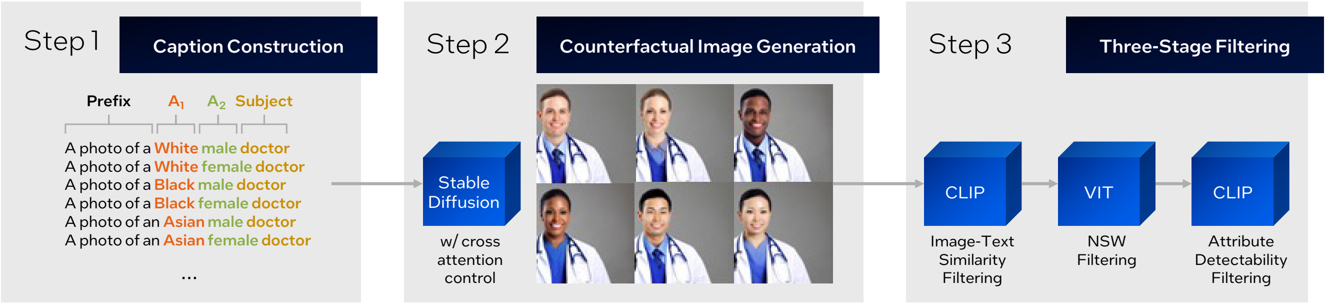
Figure 1. Researchers artificially created images of a doctor, then varied the race and gender of each while ensuring that the images otherwise looked very similar. This allowed them to study the effect of one social attribute at a time.

“This let us isolate and study the effect of each social attribute individually and as part of the intersectional.”

Diverse teams are better equipped to recognize and eliminate biases. Intel’s Cognitive AI Group reflects that diversity: They are highly distributed around the world, 30 percent of the team are women, and they represent a variety of backgrounds, cultures, and perceptions of the world.

Lal and his team used Retrieval-Augmented Generation (RAG), a common technique for enhancing the accuracy and reliability of generative AI models which couples large foundational models with external memory hosted in large vector stores or vector databases. Their social counterfactuals dataset consists of synthetic images constructed using diffusion models. The team open-sourced the social counterfactuals dataset on Hugging Face Hub, and their research paper was accepted for a presentation at the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

“Unless the models are probed using data-intensive methods, it’s difficult to know what biases might be present. As we constructed this dataset, we studied six foundational models. We found a lot of biases. Some models were more biased than others, which shows how important it is to probe each foundational model,” Lal continued.


Step 1
Caption Construction
Prefix
A_1
A_2
Subject
A photo of a White male doctor
A photo of a White female doctor
A photo of a Black male doctor
A photo of a Black female doctor
A photo of an Asian male doctor
A photo of an Asian female doctor
...
Step 2
Counterfactual Image Generation
Stable Diffusion
w/ cross attention control
Step 3
Three-Stage Filtering
CLIP
Image-Text Similarity Filtering
VIT
NSW Filtering
CLIP
Attribute Detectability Filtering

We start by constructing captions which describe the types of images that we want to generate. In this work, we focus on images depicting people with various occupations and social attributes.

Specifically, we consider three types of social attributes: race, gender, and physical attributes. Using templates, we construct captions that describe a particular occupation (e.g., “doctor”) who possesses varying social attributes (e.g., “Asian Female”). This produces captions such as “A photo of an Asian female doctor.”

After generating sets of counterfactual captions, we use text-to-image diffusion models to produce images for each caption. Our method ensures that images within a counterfactual set only differ in terms of the social attributes.

In total, we over-generate 5,408,000 images for 54,080 captions.

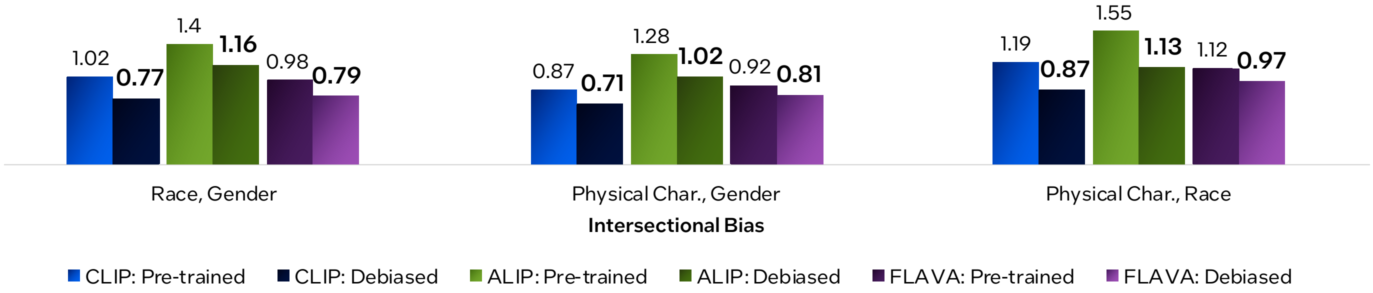
We apply three stages of filtering to improve the quality and safety of our dataset. In the first stage, we check to make sure that the images are sufficiently similar to the captions, and that images within a set have a high level of similarity.

In the second stage, we detect and filter out not-safe-for-work (NSFW) content.

Finally, the third stage checks to make sure that the social attributes described in the captions are detectable in the corresponding image.

Figure 2. Overview of the Intel Labs Cognitive AI team’s methodology for generating social counterfactuals.

Lal and his team used a large AI cluster equipped with 3rd Gen Intel® Xeon® Scalable processors and Intel® Gaudi® 2 AI accelerators to train the foundation models on their dataset. The outcome was impactful, as they were able to mitigate the biases by as much as 20 percent, as shown in Figure 3. Results varied somewhat from model to model, but the outcomes were always significant when using this technique.



Estimated by withholding counterfactual sets for 20% of the occupations in the Intel Labs Cognitive AI team's dataset. Best results are in bold. These results show that training vision-language models on SocialCounterfactuals reduces bias by 20% on average.

Source: Howard, P., et al., Intel Labs. "SocialCounterfactuals: Probing and Mitigating Intersectional Social Biases in Vision-Language Models with Counterfactual Examples." April 2024. https://arxiv.org/pdf/2312.00825

Figure 3. Mean of MaxSkew@K for pre-trained and debiased variants of CLIP, ALIP, and FLAVA.

Technology Highlights

Training foundation models requires massive clusters that operate in lockstep. The hardware configuration used a large AI cluster based on 3rd Gen Intel Xeon processors and Intel Gaudi 2 AI accelerators. Models are trained on Internet-scale data, and training throughput can be very high as models are scaled up to hundreds of AI accelerators. Working closely with the AI developer community is vital.

The robust AI software ecosystem around Intel Gaudi 2 AI accelerators includes:

- The Hugging Hub Habana Optimum Library, which is being jointly developed by Intel and Hugging Face
- DeepSpeed, which is upstream to the official library for Microsoft
- Megatron recipes for foundational model training
- PyTorch Lightning for large-scale training of models on Intel Gaudi 2 AI accelerator, and
- Intel Gaudi software tools.

"Intel Gaudi 2 hardware-accelerated data loading is very easy to use and especially relevant when training multimodal foundation models," Lal says. "The Hugging Face Intel Habana Optimum Library requires only one extra argument, which has really helped us speed up training runs." As much value as Lal's team has derived from the Intel Gaudi 2 AI accelerator, they are looking forward to the larger high-bandwidth memory (HBM) that comes with the Intel Gaudi 3 AI accelerator, which will let them tackle larger documents and will foster faster interconnect speed between Intel Gaudi 2 cards.

Training foundational models requires hundreds—or even thousands—of AI accelerators. The Intel Labs team relies heavily on Intel Gaudi accelerator-based AI clusters. Third-party studies by Amazon Web Services (AWS) have verified the high scaling efficiency of Intel Gaudi accelerators as the number of Intel Gaudi AI accelerators is scaled up for distributed training. When Intel Labs teamed up with Hugging Face to compare Intel Gaudi 2 AI accelerators with both Nvidia A100 and Nvidia H100, the head-to-head tests showed that Intel Gaudi 2 with Optimum Habana is about 1.4x faster than Nvidia H100 and 2.5x faster than Nvidia A100 80GB with transformers.

Responsible AI

The team at Intel Labs takes their role in developing Responsible AI (RAI) seriously.

"We need to make sure that these models are always correct," says Lal. "To ensure that their responses are grounded in authoritative sources. That they use multimodal data, which not all models do effectively. And most importantly, that as these models are deployed, biases don't make it into production models."

Intel has long acknowledged the importance of ethical and human rights implications when developing technology. This is especially true with AI technology. Recognizing this, Intel established the Responsible AI Program to encourage collaboration between its partners in industry and academia to evolve the best methods, principles, and tools to ensure responsible practices.

Lessons Learned

Lal and his team are literally out to change the world. They have adapted a quote from their boss, Intel CEO Pat Gelsinger, as their mantra: "Technology itself is inherently neutral; we must constantly shape it as a force for good.[3]"

"As leading AI researchers, we want to shape this technology to benefit humanity, to really increase collaboration between people. To drive productivity. And we're proving that it can all do this in a very equitable way that doesn't hurt the interests of people who might be already on the margins, without increasing the inequality in society," says Lal.

Nothing will give the team at Intel Labs more satisfaction than seeing their datasets, which they have already open-sourced, used by AI practitioners worldwide to continue improving the state of the art in AI.

Resources

For more information about Intel Labs and the socialcounterfactual project:
https://intellabs.github.io/multimodal_cognitive_ai/socialcounterfactuals/

Get started now: Try out Intel® Gaudi® 2 using the Intel Developer Cloud.

To start today, please check out this code example that implements RAG flow on the Intel Gaudi 2 AI accelerator.

Sources

[1] https://aiindex.stanford.edu/report/
[2] https://aiindex.stanford.edu/report/ (p. 5)
[3] https://time.com/6250235/ai-push-us-forward-pat-gelsinger/

0724/AC/MIM/PDF 360890-001