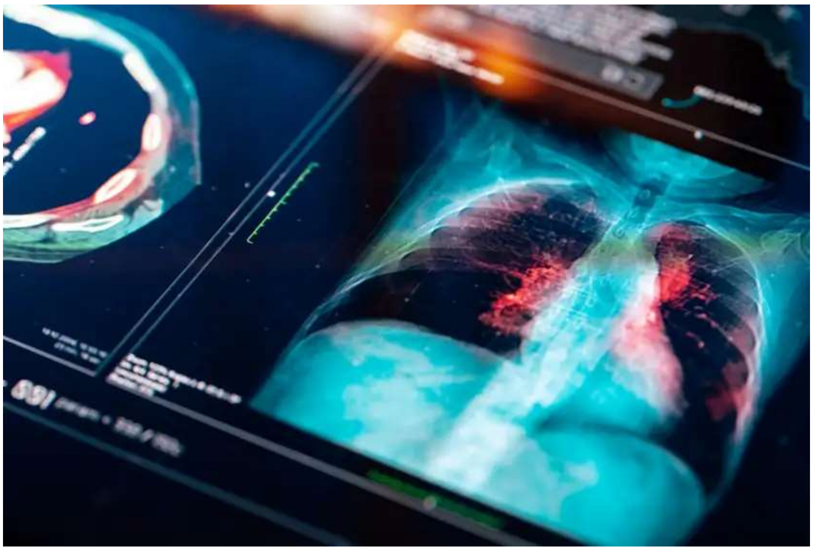
AI doesn't know 'no' – and that's a huge problem for medical bots

New Scientist, 19 May 25 By Jeremy Hsu

Many AI models fail to recognise negation words such as "no" and "not", which means they can't easily distinguish between medical images labelled as showing a disease and images labelled as not showing the disease.



AI models can struggle to understand the captions on some medical images

Toddlers may swiftly master the meaning of the word "no", but many artificial intelligence models struggle to do so. They show a high fail rate when it comes to understanding commands that contain negation words such as "no" and "not".

That could mean medical AI models failing to realise that there is a big difference between an X-ray image labelled as showing "signs of pneumonia" and one labelled as showing "no signs of pneumonia" – with potentially catastrophic consequences if physicians rely on AI assistance to classify images when making diagnoses or prioritising treatment for certain patients.

It might seem surprising that today's sophisticated AI models would struggle with something so fundamental. But "they're all bad [at it] in some sense", says Kumail Alhamoud at the Massachusetts Institute of Technology.

Alhamoud and his colleagues evaluated how well a range of AI models understand negation words in captions paired with various videos and images, including medical images. They compiled thousands of image pairs where one image contains a target object and the other image is missing the same object, and then generated corresponding captions to describe the presence or absence of objects, creating nearly 80,000 test problems.

The researchers tested vision-language models that combine some language understanding with the ability to analyse imagery. They focused on 10 different versions of the open-source CLIP AI model, which was originally developed by the company OpenAI and then released for anyone to use and develop under what is known as an MIT License, along with an 11th model developed by Apple called AIMV2 that came out more recently and represents one of the best such models. Two of the versions of CLIP had been trained specifically to interpret medical images by separate groups of researchers.

In the first test, the researchers challenged the AI models with retrieving images containing certain objects while specifying the exclusion of other related objects – such as asking for pictures of tables without chairs. Here the AI models ran into difficulties. While most of them could successfully retrieve an image based on the presence of given objects about 80 per cent of the time, this dropped to about 65 per cent or lower when they were asked to retrieve images lacking particular objects.

The second test asked the AI models to select the most accurate caption for an image of a general scene from a choice of four possible options. The versions of CLIP trained on medical images were asked to choose between just two possible options to describe medical conditions in X-ray images. Again, the caption options contained information not only on what was present in the image but also on what was absent – for instance, a caption describing an X-ray as showing evidence of pneumonia and another caption stating there is no pneumonia. The best-performing models achieved around 40 per cent or lower accuracy on this negation task – even though humans find this task easy.

Such results show how vision-language models have an affirmation bias. In other words, they ignore negation or exclusion words such as “no” and “not” in descriptions and simply assume they are being asked to always affirm the presence of objects. The researchers will present their findings at the Conference on Computer Vision and Pattern Recognition in Nashville, Tennessee, from 11 to 15 June.

Both vision-language models and the large language models used in AI chatbots are based on the transformer model originally developed by Google researchers. Transformer models “are really good at capturing context-specific meaning” among strings of words, says Karin Verspoor at the Royal Melbourne Institute of Technology in Australia, who wasn’t involved in the study. But negation words like “not” and “no” work independently of context-specific meaning and “can appear in many places within any given sentence”, she says. This makes it harder for the AI models to fully understand and accurately respond to requests that contain such negation words.

“In clinical applications, negation of information is critical – knowing both what signs and symptoms a patient has and what they can be confirmed not to have is important to precisely characterise a condition, and to rule out certain diagnoses,” says Verspoor. Her own research has shown how language models often fail to make the correct inference for sentences that include negation words.

Specifically training vision-language models on negation word examples improved their information retrieval performance by 10 per cent and boosted accuracy on the multiple-choice questions by 30 per cent. But this does not address how such models work in the first place, says Marzyeh Ghassemi at MIT, part of the team behind the new study. “A lot of the solutions that we come up with are a little Band-Aid-like in nature, because they don’t address the fundamental problem,” she says.