{* AI + ML *}

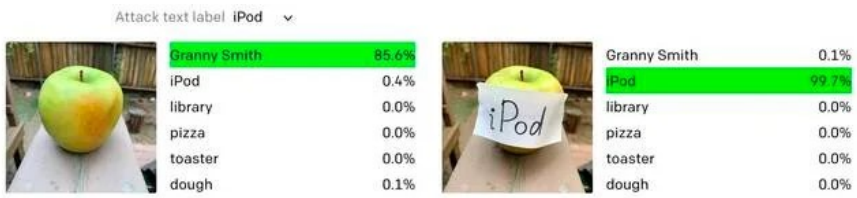
You only need pen and paper to fool this OpenAI computer vision code. Just write down what you want it to see

Trick future robot overlords by scribbling 'superuser' on your forehead

Katyanna Quach Fri 5 Mar 2021 // 23:28 UTC SHARE

OpenAI researchers believe they have discovered a shockingly easy way to hoodwink their object-recognition software, and it requires just pen and paper to carry out.

Specifically, the lab's latest computer vision model, CLIP, can be tricked by in what's described as a “typographical attack." Simply write the words ‘iPod’ or ‘pizza’ on a bit of paper, stick it on an apple, and the software will wrongly classify the piece of fruit as a Cupertino music player or a delicious dish.



Not the smartest tool in the box. Source: OpenAI. Click to enlarge

“We believe attacks such as those described above are far from simply an academic concern,” the bods behind CLIP said this week. “By exploiting the model’s ability to read text robustly, we find that even photographs of hand-written text can often fool the model." They added that "this attack works in the wild," and "it requires no more technology than pen and paper.”

CLIP isn't the only artificially intelligent software to fall for such simple shenanigans. It was demonstrated you could use sticky tape to fool Tesla’s Autopilot into misreading a 35mph sign as an 85mph one. Other forms of these so-called adversarial attacks, however, require some technical know-how to execute: it typically involves adding noise to a

photo or crafting a sticker of carefully arranged pixels to make an object-recognition system mistake, say, a banana for a toaster. In CLIP's case, however, none of that is necessary.

Suffice to say, OpenAI's model was trained using pictures of text as well as images of objects and other things scraped from the internet.

Think your smartwatch is good for warning of a heart attack? Turns out it's surprisingly easy to fool its AI

READ MORE

This approach was taken so that CLIP remains fairly general purpose, and can be fine-tuned as needed for a particular workload without having to be retrained. Given an image, it can not only predict the right set of text labels describing the scene, it can be repurposed to search through large databases of pictures and provide captions.

CLIP is able to learn abstract concepts across different representations, OpenAI said. For example, the model is able to recognize Spider-Man when the superhero is depicted in a photo, a sketch, or described in text. What's more interesting is that the researchers have been able to find groups of neurons in the neural network that are activated when the software clocks a glimpse of Spider-Man.

They have described these as multimodal neurons. "One such neuron, for example, is a 'Spider-Man' neuron that responds to an image of a spider, an image of the text 'spider,' and the comic book character 'Spider-Man' either in costume or illustrated," the OpenAI team said. CLIP has all sorts of multimodal neurons that represent different concepts, such as seasons, countries, emotions, and objects.

But the model's greatest strengths – its versatility and robustness – is also its greatest weakness. CLIP is easily hoodwinked by typographical

attacks, they found.

Object-recognition AI – the dumb program's idea of a smart program: How neural nets are really just looking at textures

READ MORE

Going back to the apple vs pizza example, the multimodal neurons that have learnt the representation of an apple don’t fire as well when they see the written word 'pizza.' Instead, the pizza-related neurons get triggered instead. The model is easily confused.

There is evidence that abstract learning using multimodal neurons also occurs in human brains. But unfortunately, here’s where modern machines pale in comparison to their biological counterparts. Humans can obviously tell that an apple with a handwritten note that reads pizza on it is still an apple, while AI models can’t yet.

OpenAI said CLIP doesn't perform as well as some computer vision models that are today used in production. It also suffers from offensive biases, its neurons associate the concept of the 'Middle East' with 'terrorism' and black people with gorillas. The model is only used for research purposes at the moment, and OpenAI is still deciding whether or not to release the code.

"Our own understanding of CLIP is still evolving, and we are still determining if and how we would release large versions of CLIP. We hope that further community exploration of the released versions as well as the tools we are announcing today will help advance general understanding of multimodal systems, as well as inform our own decision-making," it said.

OpenAI declined to comment further on CLIP. ®

MORE Ai Machine Learning

Get our **AI** newsletter

Enter Email SUBSCRIBE

// KEEP READING

AI-generated pixelated photo of AOC in a bikini pulled from paper highlighting danger of AI-generated pics

IN BRIEF Plus: Dead pop star brought back to life by ML, OECD develops effort to monitor AI power

Microsoft touts Azure Percept development kits to those toying with AI on the edge

IGNITE Full stack means fully locked in – oh shoot, did we say that out loud?

Manhunt: 'Armed and dangerous' MIT AI scientist sought by cops probing grad student's gun murder

Victim shot dead a week after he got engaged to fiancée

Twitter: Our image-cropping AI seems to give certain peeps preferential treatment. Solution: Use less AI

Let's just go back to human-selected cropping, eh?

AI brain drain to Google and pals threatens public sector's ability to moderate machine-learning bias

With top research talent focused on commercial machine-learning goals, where do we go from here?

AI in the Enterprise: How can we make analytics and stats sound less scary? Let's call it AI!

REGISTER DEBATE New names for old recipes