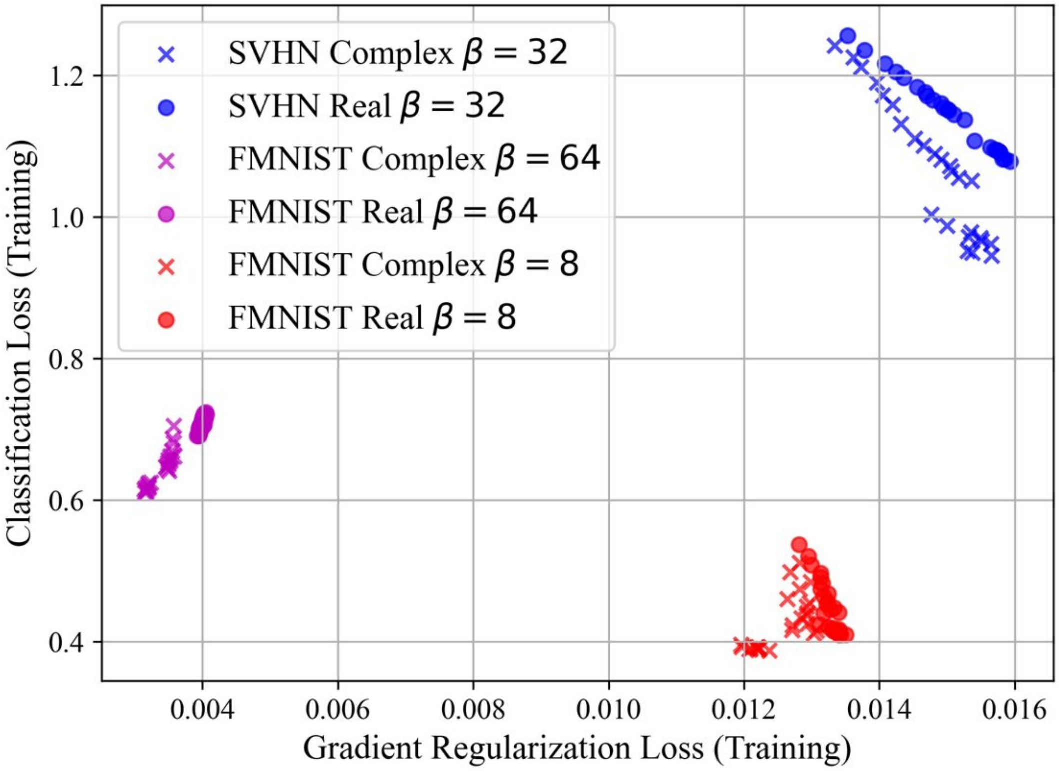
Imaginary numbers protect AI from very real threats

September 1 2021



The loss in performance (y-axis) and security (x-axis) are plotted together as the different types of networks (standard or complex-valued) are trained on image classification tasks using Google Street views of house numbers and clothing items. In these experiments, the complex-valued networks always attain better results by about 10-20%. Credit: Eric Yeats, Duke University

Computer engineers at Duke University have demonstrated that using complex numbers—numbers with both real and imaginary components—can play an integral part in securing artificial intelligence algorithms against malicious attacks that try to fool object-identifying software by subtly altering the images. By including just two complex-valued layers among hundreds if not thousands of training iterations, the technique can improve performance against such attacks without sacrificing any efficiency.

The research was presented at the Proceedings of the 38th International Conference on Machine Learning.

"We're already seeing machine learning algorithms being put to use in the real world that are making real decisions in areas like vehicle autonomy and facial recognition," said Eric Yeats, a doctoral student working in the laboratory of Helen Li, the Clare Boothe Luce Professor of Electrical and Computer Engineering at Duke. "We need to think of ways to ensure that these algorithms are reliable to make sure they can't cause any problems or hurt anyone."

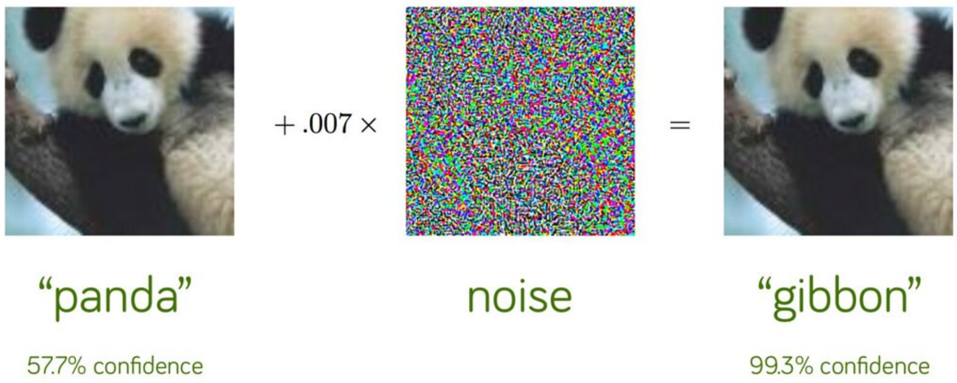
One way that machine learning algorithms built to identify objects and images can be fooled is through adversarial attacks. This essentially entails modifying the image in a way that breaks the AI's decision-making process. It can be as simple as adding stickers to a stop sign or as sophisticated as adding a carefully crafted layer of static to an image that alters it in ways undetectable to the human eye.

The reason these small perturbations can cause such large problems stems from how machine learning algorithms are trained. One standard method called gradient descent compares the decisions it arrives at to the correct answers, attempts to tweak its inner workings to fix the errors, and repeats the process over and over again until it is no longer improving.

One way to visualize this is to imagine a boulder rolling through a valley of hills and mountains. With each machine learning iteration, the algorithm's working parameters (boulder) rolls further into the valley. When it starts to roll up a new hill, the algorithm changes its course to keep it rolling downward. Eventually the boulder settles in the best answer (lowest spot) around.

A challenging aspect of this approach is that the valley the boulder is rolling through is especially rugged terrain—think the Himalayas instead of the Appalachians. One small nudge in the wrong direction can send the boulder plummeting toward a very different outcome. This is why barely noticeable static can make an image classifier see a gibbon instead of a panda.

To keep their algorithms on track, computer scientists can train their algorithms with a technique called gradient regularization. This causes the boulder to choose paths that aren't as steep. While the causes the boulder to take a different—and longer—path to its final resting spot, it also makes sure the boulder rolls gently down the correct valley instead of being pushed off a nearby ravine.



Subtle static can fool AI into classifying a panda as a gibbon. Credit: Explaining and Harnessing Adversarial Examples, Goodfellow et al, ICLR 2015

"Gradient regularization throws out any solution that passes a large gradient back through the neural network," Yeats said. "This reduces the number of solutions that it could arrive at, which also tends to decrease how well the algorithm actually arrives at the correct answer. That's where complex values can help. Given the same parameters and math operations, using complex values is more capable of resisting this decrease in performance."

Chances are most of us haven't thought about—or even heard the words—imaginary numbers since about 8th grade. And their introduction was likely accompanied by groans followed by a chorus of, "What am I ever going to use this for?" But imaginary numbers are extraordinarily useful for describing sinusoidal waves, which happen to look a lot like a valley of hills and mountains.

When the neural network is being trained on a set of images, using complex numbers with imaginary components gives it added flexibility in how it adjusts its internal parameters to arrive at a solution. Rather than only being able to multiply and accumulate changes, it can offset the phase of the waves it's adding together, allowing them to either amplify or cancel one another out. The effect is that this once-rugged valley is smoothed out to locally flatter surfaces with multiple tiers that allow for lots of elevation change in other areas.

"The complex-valued neural networks have the potential for a more 'terraced' or 'plateaued' landscape to explore," Yeates said. "And elevation change lets the neural network conceive more complex things, which means it can identify more objects with more precision."

That added capacity allows gradient regularization neural networks using complex numbers to find solutions just as fast as those trained without the extra security. In his research, Yeats shows that image classifiers aimed at recognizing house numbers from Google Maps and different clothing items trained on his approach are more secure than standard methods while performing at the same level.

"This is still an open and challenging problem," Yeats said. "So researchers are doing what they can to do a little bit better here and there."

More information: "Improving Gradient Regularization using Complex-Valued Neural Networks." Eric Yeats, Yiran Chen, Hai Li. Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.

Provided by Duke University

Citation: Imaginary numbers protect AI from very real threats (2021, September 1) retrieved 23 April 2024 from https://techxplore.com/news/2021-09-imaginary-ai-real-threats.html