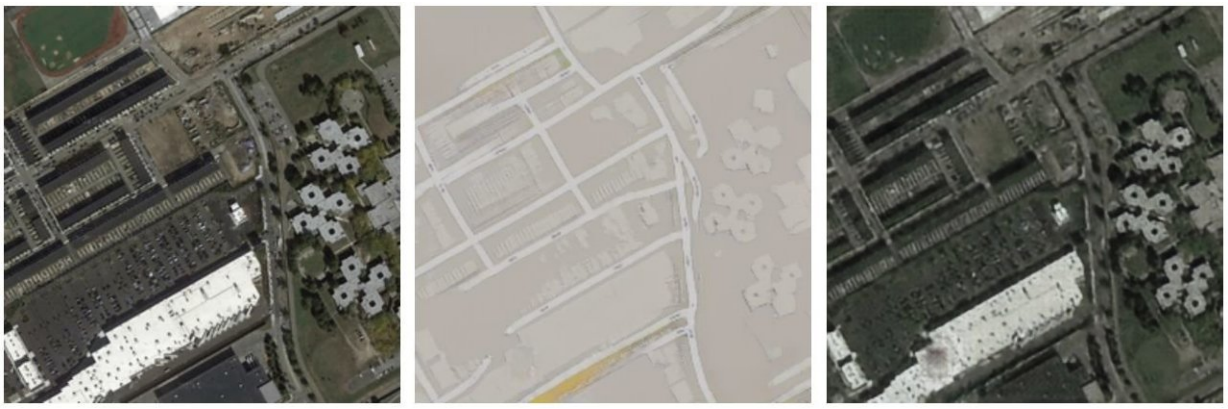
Researchers discover AI information-hiding behavior for later use

January 4 2019, by Nancy Cohen



Details in x are reconstructed in GF x, despite not appearing in the intermediate map F x. Credit: arXiv:1712.02950 [cs.CV]

Call it clever, brand it a cheater, but don't feel ashamed to find it terribly interesting. The "it" is CycleGAN, and its link to steganography—where messages and information are hidden within nonsecret text or data.

So, in 2019 it cannot be that shocking for people to learn that a machine, not a human, can cheat its way though a task. The AI in this instance, like good human spies and cons, learned when to hide some information that can be used later.

In *Packt*, Bhagyashree R wrote that "The researchers discovered the

machine was encoding data of the aerial map into the noise patterns of the street map on the down low. The code was so subtle that it would be invisible to the human eye. But on closer inspection, when the details had been amplified, it was clear that the machine had made thousands of tiny color changes indicating visual data that could be used like a cheat sheet when recreating the aerial image – hence the magically reappearing skylights."

Meanwhile, a much-quoted article on the topic (the research was covered by a number of tech watching sites, actually) capsulized what the researchers discovered. "A machine learning agent intended to transform aerial images into street maps and back was found to be cheating by hiding information it would need later in 'a nearly imperceptible, high-frequency signal," said *TechCrunch*.

Lily Hay Newman in *Wired* in 2017 reminded readers that steganography is an old practice, nothing born yesterday. Think Da Vinci embedding secret meaning in a painting; or yesteryear's spies writing in invisible ink.

If the practice is ancient, though, there are some contemporary problems. We are, after all, in a digital world where all vices and virtues have taken on new processes online.

Steganography is only going to get more difficult to spot, said*BankInfoSecurity*, and has "already been put to use by bad actors."

Mathew Schwartz said digital steganography appeared to make life more difficult for law enforcement agencies and quoted a university professor passing a similar remark. "Perfectly deniable steganographic disk encryption is going to be a nightmare when it comes to gathering digital evidence," said Alan Woodward, a professor of computer science at University of Surrey.

Fast forward to reports that are in now, that a group of Stanford and Google researchers performed a study on how a neural network, CycleGAN, learns to cheat. The paper: CycleGAN, a Master of Steganography is on arXiv and the three authors are Casey Chu (Stanford), Andrey Zhmoginov (Google) and Mark Sandler (Google).

They wrote, "CycleGAN learns to 'hide' information about a source image into the images it generates in a nearly imperceptible, high frequency signal."

As part of their Discussion section, the authors make the point that "By encoding information in this way, CycleGAN becomes especially vulnerable to adversarial attacks; an attacker can cause one of the learned transformations to produce an image of their choosing by perturbing any chosen source image."

Their advice? They wrote that "the presence of this phenomenon indicates that caution is necessary when designing loss functions that involve compositions of neural networks: such models may behave in unintuitive ways if one component takes advantage of the ability of the other component to support adversarial examples."

Common frameworks, according to the authors, such as generative adversarial networks and perceptual losses use these compositions. They said that "these frameworks should be carefully analyzed to make sure that adversarial examples are not an issue."

But wait. Should we run for the hills with screaming fears that robots and AI will finish us all? Fortunately, Devin Coldewey calmed readers in *TechCrunch*. The occurrence "simply reveals a problem with computers that has existed since they were invented: they do exactly what you tell them to do."

What did Coldewey mean by that? "The intention of the researchers was, as you might guess, to accelerate and improve the process of turning satellite imagery into Google's famously accurate maps. To that end the team was working with what's called a CycleGAN—a neural network that learns to transform images of type X and Y into one another, as efficiently yet accurately as possible, through a great deal of experimentation."

The computer arrived at a solution "that shed light on a possible weakness of this type of neural network—that the computer, if not explicitly prevented from doing so, will essentially find a way to transmit details to itself in the interest of solving a given problem quickly and easily."

More information: CycleGAN, a Master of Steganography arXiv:1712.02950 [cs.CV] arxiv.org/abs/1712.02950

Citation: Researchers discover AI information-hiding behavior for later use (2019, January 4) retrieved 29 November 2023 from https://techxplore.com/news/2019-01-ai-information-hiding-behavior.html