Security (and Privacy) in Machine Learning

Nicholas Carlini
University of California, Berkeley
(now Google Brain)

This talk: neural networks

The New York Times

The Race for Self-Driving Cars

Autonomous cars have arrived. Major automakers have been investing billions in development, while tech players like Uber and Google's parent company have been testing their versions in American cities.

ALEX DAVIES TRANSPORTATION 03.13.18 12:15 PM

WAYMO TAKES THE FINAL STEP BEFORE LAUNCHING ITS SELF-DRIVING CAR SERVICE

BBC | Sign in | News | Sport | Weather | Shop | Earth | Travel | Mo

NEWS

Home | Video | World | US & Canada | UK | Business | Tech | Science | Stories | Enterta

Google AI defeats human Go champion

25 May 2017

Share

RESEARCH ARTICLE

Superhuman AI for heads-up no-limit poker: Libratus beats top professionals

Noam Brown, Tuomas Sandholm*

+ See all authors and affiliations

Science 17 Dec 2017:
eaao1733
DOI: 10.1126/science.aao1733


What can I help
you with?

amazon

Machine learning is amazing

But there's a catch

Understandability

This talk:

Discuss security & privacy problems being studied in the research community

What this talk is ***not***

$$U(x) = \left(\frac{1}{L}\sum_{i=1}^{L}\|F_r(x)\|\right) - \left\|\frac{1}{L}\sum_{i=1}^{L}F_r(x)\right\|$$

$$\begin{aligned} \text{minimize} \quad & \mathcal{D}(x, x+\delta) \\ \text{such that} \quad & f(x+\delta) \le 0 \\ & x+\delta \in [0,1]^n \end{aligned}$$

$$\begin{aligned} & \mathop{\mathbf{Pr}}_{t\in\mathcal{R}}\left[\mathrm{Px}_\theta(s[t]) \le \mathrm{Px}_\theta(s[r])\right] \\ = & \sum_{v \le \mathrm{Px}_\theta(s[r])} \mathop{\mathbf{Pr}}_{t\in\mathcal{R}}\left[\mathrm{Px}_\theta(s[t]) = v\right]. \end{aligned}$$

$$G(x)_i = \begin{cases} Z(x)_i & \text{if } i \le N \\ (1 + U(x) - \tau)\cdot \max_i Z(x)_i & \text{if } i = N+1 \end{cases}$$

$$\nabla_x f(g(x))|_{x=\hat{x}} \approx \nabla_x f(x)|_{x=g(\hat{x})}$$

$$\ell(\boldsymbol{x}) = \sum_i \max\left(\max_{t\in\{\epsilon, \text{``''}\}} f(\boldsymbol{x})_t^i - \max_{t'\notin\{\epsilon, \text{``''}\}} f(\boldsymbol{x})_{t'}^i, 0\right).$$

$$\Pr(\boldsymbol{p}|\boldsymbol{y}) = \sum_{\pi\in\Pi(\boldsymbol{p},\boldsymbol{y})} \Pr(\pi|\boldsymbol{y}) = \sum_{\pi\in\Pi(\boldsymbol{p},\boldsymbol{y})} \prod_i \boldsymbol{y}_{\pi^i}^i$$

What this talk is ***not***

Definition 4. *A random algorithm* $\mathcal{A}$ *is* (ε, δ)-differentially private *if*

$$\mathbf{Pr}(\mathcal{A}(\mathcal{D}) \in S) \le \exp(\varepsilon)\cdot \mathbf{Pr}(\mathcal{A}(\mathcal{D}') \in S) + \delta$$

$$\alpha_{pq} = \sum_{i\in\{p,q\}} \frac{\partial Z(x)_t}{\partial x_i}$$

$$\beta_{pq} = \left(\sum_{i\in\{p,q\}}\sum_j \frac{\partial Z(x)_j}{\partial x_i}\right) - \alpha_{pq}$$

$$\theta^* = \arg\min_\theta \mathbb{E}_{x\in\mathcal{X}}\left[\max_{\delta\in[-\epsilon,\epsilon]^N} \ell(x+\delta; F_\theta)\right].$$

$$\begin{aligned} = & -\log \mathop{\mathbb{E}}_{r'\in\mathcal{R}}\left[\mathbb{1}\left(L_\theta(s[r']) \le L_\theta(s[r])\right)\right] \\ = & -\log\left(1\cdot\frac{|\{r'\in\mathcal{R} : L_\theta(s[r']) \le L_\theta(s[r])\}|}{|\mathcal{R}|}\right. \\ & \left. 0\cdot\frac{|\mathcal{R}| - |\{r'\in\mathcal{R} : L_\theta(s[r']) \le L_\theta(s[r])\}|}{|\mathcal{R}|}\right) \\ = & -\log\left(\frac{|\{r'\in\mathcal{R} : L_\theta(s[r']) \le L_\theta(s[r])\}|}{|\mathcal{R}|}\right) \\ = & -\left(\log \mathbf{rank}_\theta(s[r]) - \log|\mathcal{R}|\right) \\ = & \log|\mathcal{R}| - \log \mathbf{rank}_\theta(s[r]) \end{aligned}$$

What this talk ***is***

What are the security problems in machine learning today?

IMAGENET

French
Bulldog

(95%)

Old
English
Sheepdog

(83%)

Greater
Swiss
Mountain
Dog

(78%)

Siberian
Husky

(81%)

Great
Dane

(67%)

Beagle

(96%)

Guacamole

(99.99%)

Golden
Retriever

(96%)

Guacamole

(99.99%)

These phenomena are known as **adversarial examples**

B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Srndic, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. 2013.
C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. ICLR 2014.
I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. 2014.

88% tabby cat

adversarial perturbation

99% guacamole

(a) (b) (c)

What does this have to do with voice?

We use these same classification approaches for speech recognition.

Attacks on Android, circa 2015


SAMSUNG

State-of-the-art in 2015

It's been three years.

Can we do better?

Feynman Algorithm

1. Write down the problem.
2. Think very hard.
3. Write down the answer.

Towards Evaluating the Robustness of Neural Networks

Nicholas Carlini David Wagner
University of California, Berkeley

ABSTRACT

Neural networks provide state-of-the-art results for most machine learning tasks. Unfortunately, neural networks are vulnerable to adversarial examples: given an input x and any target classification t, it is possible to find a new input x' that is similar to x but classified as t. This makes it difficult to apply neural networks in security-critical areas. Defensive distillation is a recently proposed approach that can take an arbitrary neural network, and increase its robustness, reducing the success rate of current attacks' ability to find adversarial examples from 95% to 0.5%.

In this paper, we demonstrate that defensive distillation does not significantly increase the robustness of neural networks by introducing three new attack algorithms that are successful on both distilled and undistilled neural networks with 100% probability. Our attacks are tailored to three distance metrics used previously in the literature, and when compared to previous adversarial example generation algorithms, our attacks are often much more effective (and never worse). Furthermore, we propose using high-confidence adversarial examples in a simple transferability test we show can also be used to break defensive distillation. We hope our attacks will be used as a benchmark in future defense attempts to create neural networks that resist adversarial examples.

I. Introduction

Deep neural networks have become increasingly effective at many difficult machine-learning tasks. In the image recognition domain, they are able to recognize images with near-human accuracy [27], [25]. They are also used for speech recognition [18], natural language processing [1], and playing games [43], [32].

However, researchers have discovered that existing neural networks are vulnerable to attack. Szegedy *et al.* [46] first noticed the existence of *adversarial examples* in the image classification domain: it is possible to transform an image by a small amount and thereby change how the image is classified. Often, the total amount of change required can be so small as to be undetectable.

The degree to which attackers can find adversarial examples limits the domains in which neural networks can be used. For example, if we use neural networks in self-driving cars, adversarial examples could allow an attacker to cause the car to take unwanted actions.

The existence of adversarial examples has inspired research on how to harden neural networks against these kinds of

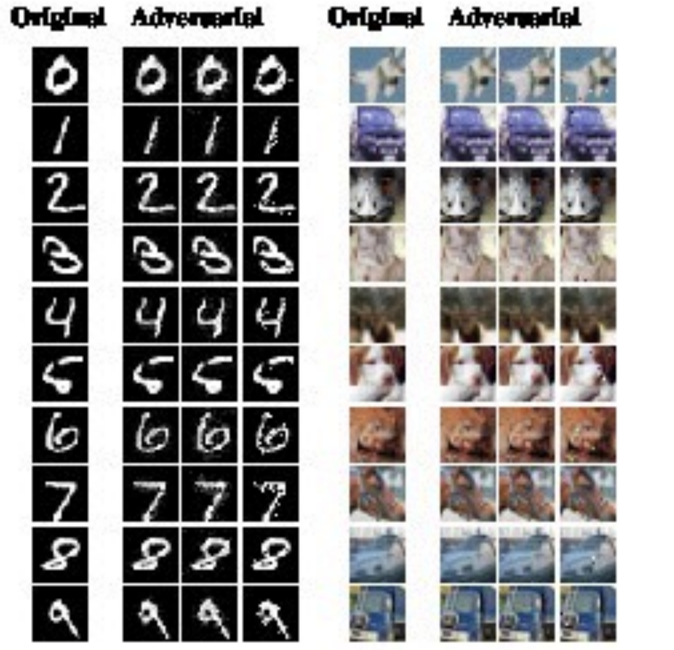


Fig. 1. An illustration of our attacks on a defensively distilled network. The leftmost column contains the starting image. The next three columns show adversarial examples generated by our L_2, L_∞, and L_0 algorithms, respectively. All images start out classified correctly with label l, and the three misclassified instances share the same misclassified label of $l+1$ (mod 10). Images were chosen as the first of their class from the test set.

attacks. Many early attempts to secure neural networks failed or provided only marginal robustness improvements [15], [2], [20], [42].

Defensive distillation [39] is one such recent defense proposed for hardening neural networks against adversarial examples. Initial analysis proved to be very promising: defensive distillation defeats existing attack algorithms and reduces their success probability from 95% to 0.5%. Defensive distillation can be applied to any feed-forward neural network and only requires a single re-training step, and is currently one of the only defenses giving strong security guarantees against adversarial examples.

In general, there are two different approaches one can take to evaluate the robustness of a neural network: attempt to prove a lower bound, or construct attacks that demonstrate an upper bound. The former approach, while sound, is substantially more difficult to implement in practice, and all attempts have required approximations [2], [21]. On the other hand, if the

Audio Adversarial Examples: Targeted Attacks on Speech-to-Text

Nicholas Carlini David Wagner
University of California, Berkeley

Abstract—We construct targeted audio adversarial examples on automatic speech recognition. Given any audio waveform, we can produce another that is over 99.9% similar, but transcribes as any phrase we choose (recognizing up to 50 characters per second of audio). We apply our white-box iterative optimization-based attack to Mozilla's implementation DeepSpeech end-to-end, and show it has a 100% success rate. The feasibility of this attack introduces a new domain to study adversarial examples.

I. Introduction

As the use of neural networks continues to grow, it is critical to examine their behavior in adversarial settings. Prior work [8] has shown that neural networks are vulnerable to *adversarial examples* [40], instances x' similar to a natural instance x, but classified by a neural network as any (incorrect) target t chosen by the adversary.

Existing work on adversarial examples has focused largely on the space of images, be it image classification [40], generative models on images [26], image segmentation [1], face detection [37], or reinforcement learning by manipulating the images the RL agent sees [6, 21]. In the discrete domain, there has been some study of adversarial examples over text classification [23] and malware classification [16, 20].

There has been comparatively little study on the space of audio, where the most common use is performing automatic speech recognition. In automatic speech recognition, a neural network is given an audio waveform x and perform the speech-to-text transform that gives the transcription y of the phrase being spoken (as used in, e.g., Apple Siri, Google Now, and Amazon Echo).

Constructing targeted adversarial examples on speech recognition has proven difficult. Hidden and inaudible voice commands [11, 39, 41] are targeted attacks, but require synthesizing new audio and can not modify existing audio (analogous to the observation that neural networks can make high confidence predictions for unrecognizable images [33]). Other work has constructed standard untargeted adversarial examples on different audio systems [13, 24]. The current state-of-the-art targeted attack on automatic speech recognition is Houdini [12], which can only construct audio adversarial examples targeting phonetically similar phrases, leading the authors to state

> targeted attacks seem to be much more challenging when dealing with speech recognition systems than when we consider artificial visual systems.

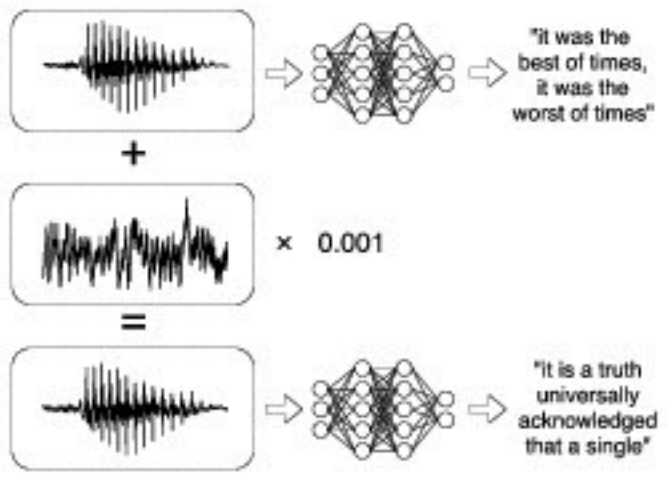


Figure 1. Illustration of our attack: given any waveform, adding a small perturbation makes the result transcribe as any desired target phrase.

Contributions. In this paper, we demonstrate that targeted adversarial examples exist in the audio domain by attacking DeepSpeech [18], a state-of-the-art speech-to-text transcription neural network. Figure 1 illustrates our attack: given any natural waveform x, we are able to construct a perturbation δ that is nearly inaudible but so that $x+\delta$ is recognized as any desired phrase. We are able to achieve this by making use of strong, iterative, optimization-based attacks based on the work of [10].

Our white-box attack is end-to-end, and operates directly on the raw samples that are used as input to the classifier. This requires optimizing through the MFC pre-processing transformation, which is has been proven to be difficult [11]. Our attack works with 100% success, regardless of the desired transcription or initial source audio sample.

By starting with an arbitrary waveform, such as music, we can embed speech into audio that should not be recognized as speech; and by choosing silence as the target, we can hide audio from a speech-to-text system.

Audio adversarial examples give a new domain to explore these intriguing properties of neural networks. We hope others will build on our attacks to further study this field. To facilitate future work, we make our code and dataset available[1]. Additionally, we encourage the reader to listen to our audio adversarial examples.

[1] http://nicholas.carlini.com/code/audio_adversarial_examples

Mozilla's DeepSpeech

Mozilla's DeepSpeech transcribes this

Mozilla's DeepSpeech
transcribes this as

"most of them were staring
quietly at the big table"

[adversarial]

"It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity"

Why is this so much stealthier?

It works on music, too

DeepSpeech transcribes
"speech can be embedded in music"

And can "hide" speech

DeepSpeech does not hear any speech in this audio sample

That's a lot of problems

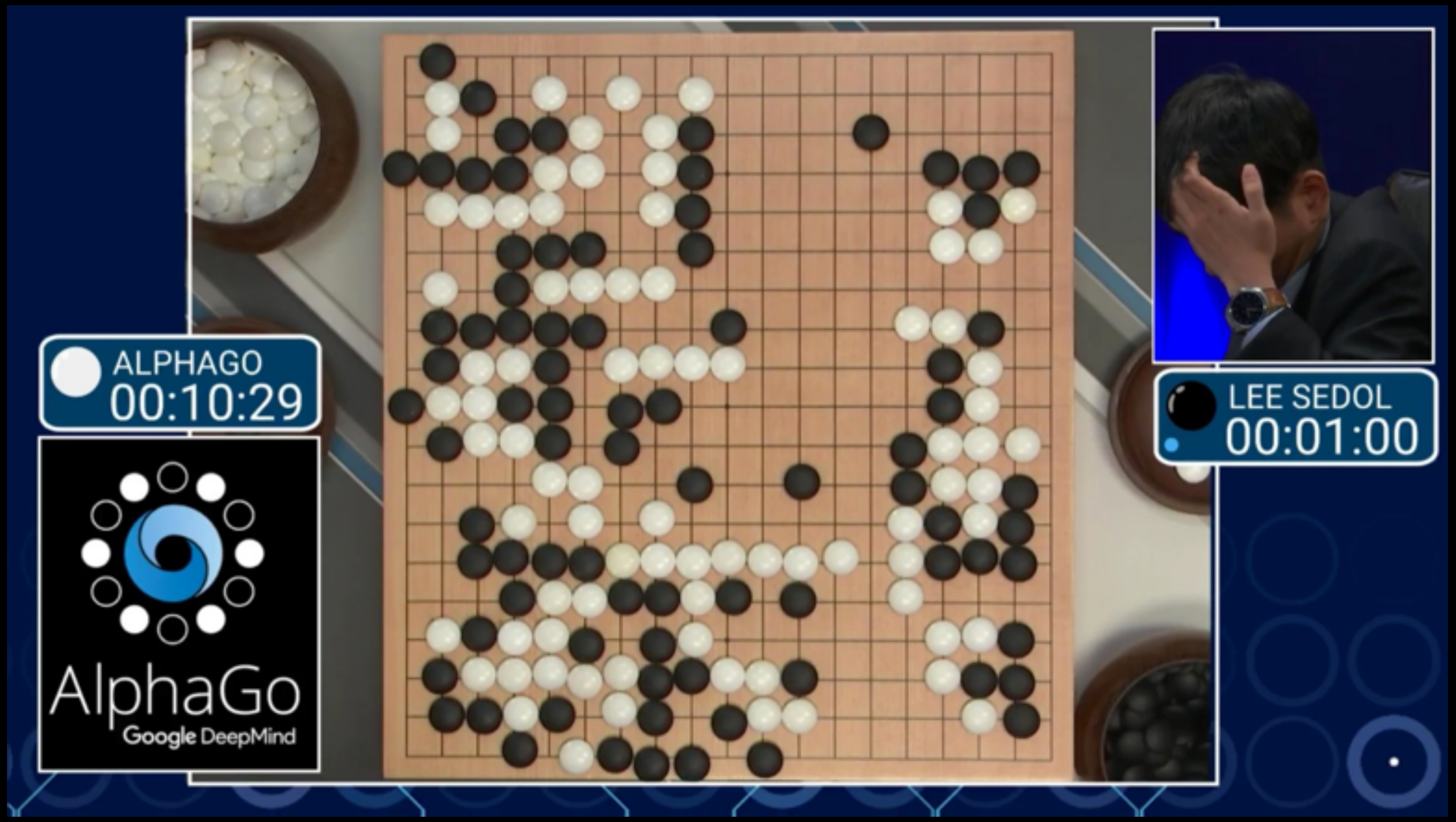
Do you have any solutions?

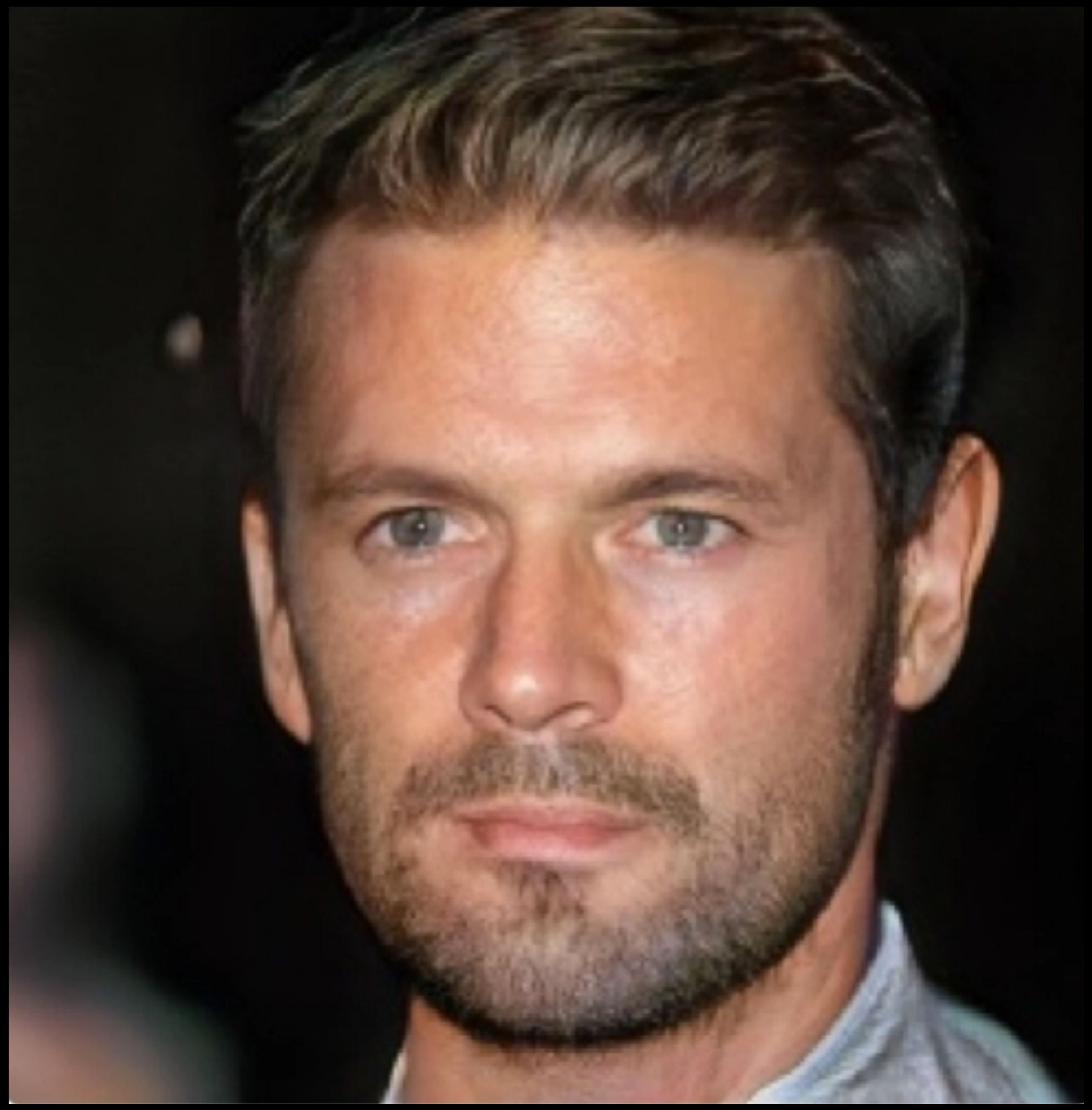
Sorry, no.

This is an active area of research.

Ask me again in two years.

Yes, machine learning gives **amazing** results





However, there are also significant **vulnerabilities**

Guacamole (99%)

Questions?

More Details:

https://nicholas.carlini.com