You Don’t See It, But They Do!

An Introduction to Adversarial Machine Learning

Security and Privacy – MECD

30 November 2023

Inês Valentim, Nuno Lourenço, Nuno Antunes

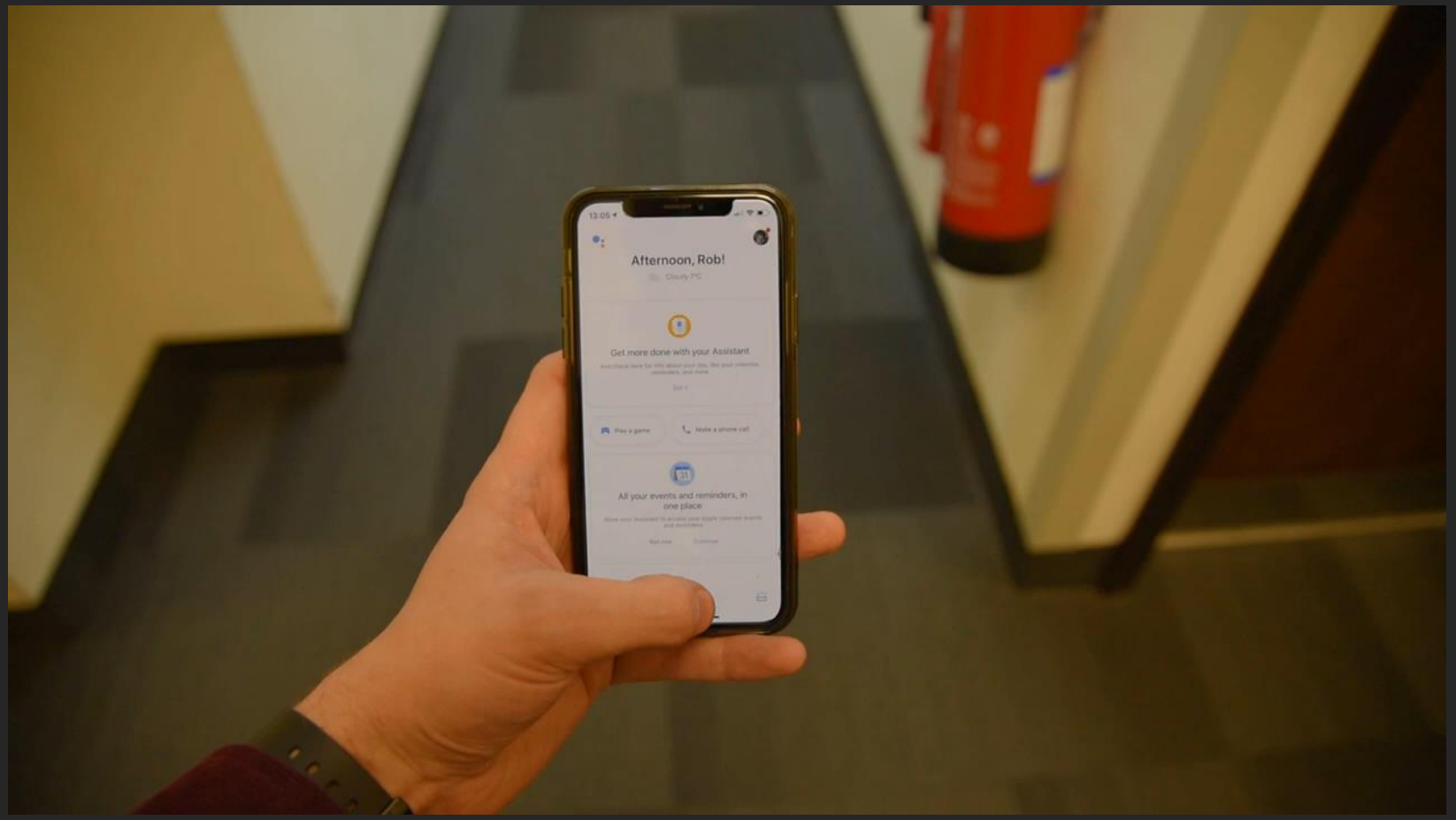
valentim@dei.uc.pt, naml@dei.uc.pt, nmsa@dei.uc.pt

AI is great at many tasks! Here's an example:

https://www.youtube.com/watch?v=5kpsZoKjPgQ

An example from another domain:

https://www.youtube.com/watch?v=fcSmtHNR64M

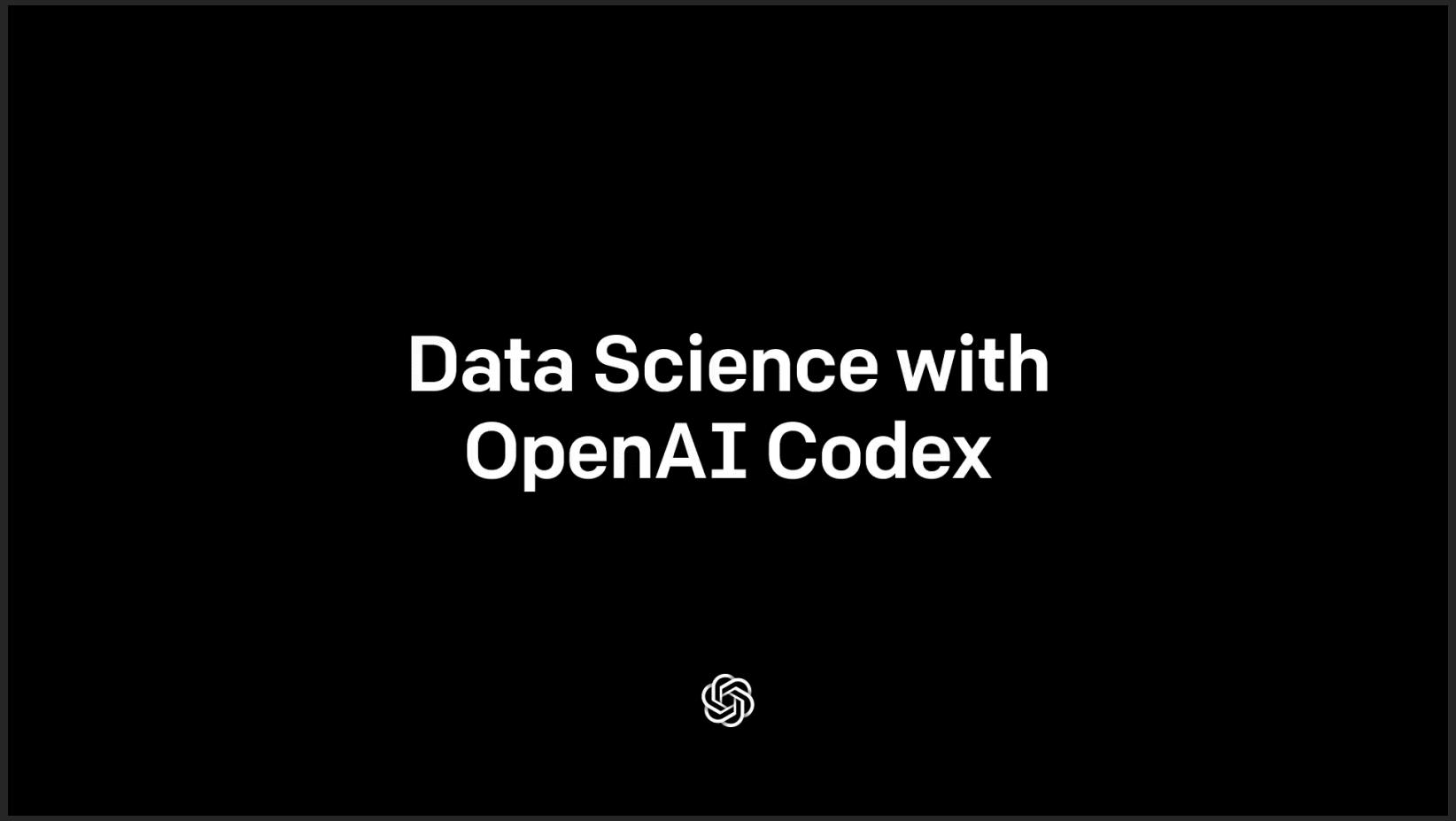


Another interesting example:

https://www.youtube.com/watch?v=kopoLzvh5jY

LLMs couldn’t be left behind:

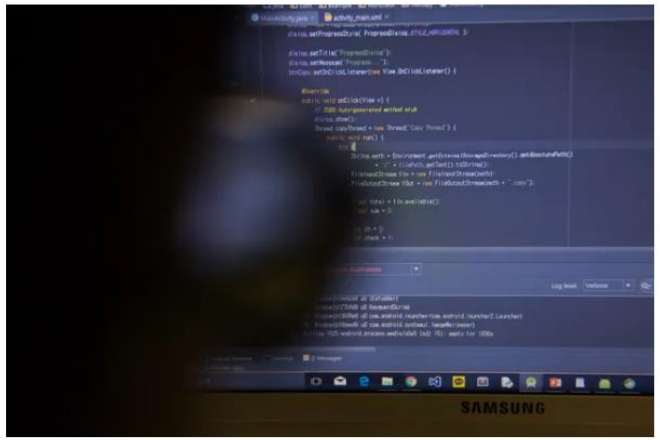
https://www.youtube.com/watch?v=Ru5fQZ714x8



BUT not everything about it is nice…

ARTIFICIAL INTELLIGENCE / TECH / LAW

The lawsuit that could rewrite the rules of AI copyright

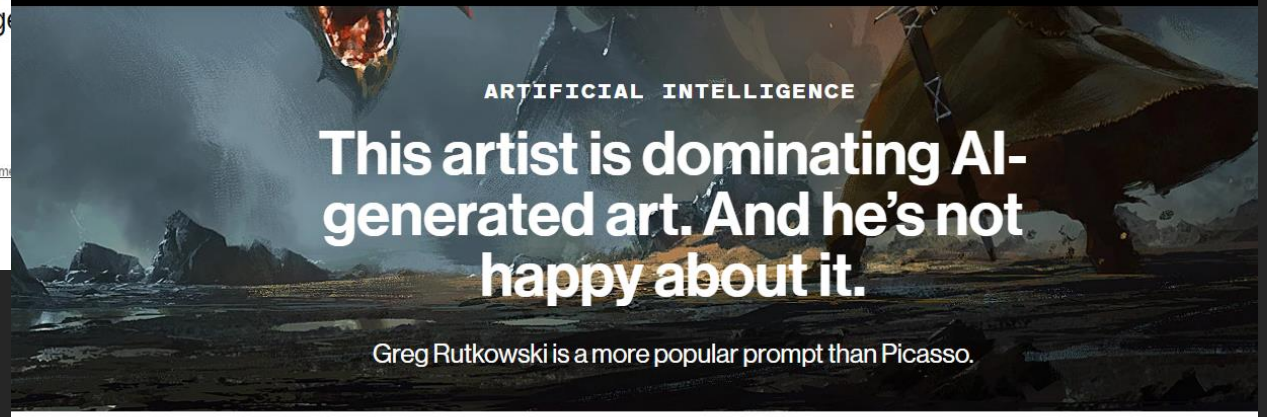


The key question in the lawsuit is whether open-source code can be reproduced by AI without attached licenses. Credit: Getty Images

/ Microsoft, GitHub, and OpenAI are being sued for allegedly violating copyright law by reproducing open-source code using AI. But the
a huge impact on
of artificial intelli

By JAMES VINCENT
Nov 8, 2022, 4:09 PM GMT | 9 Comm

MIT Technology Review | Featured | Topics | Newsletters | Events | Podcasts | Sign in | Subscribe



ARTIFICIAL INTELLIGENCE

This artist is dominating AI-generated art. And he's not happy about it.

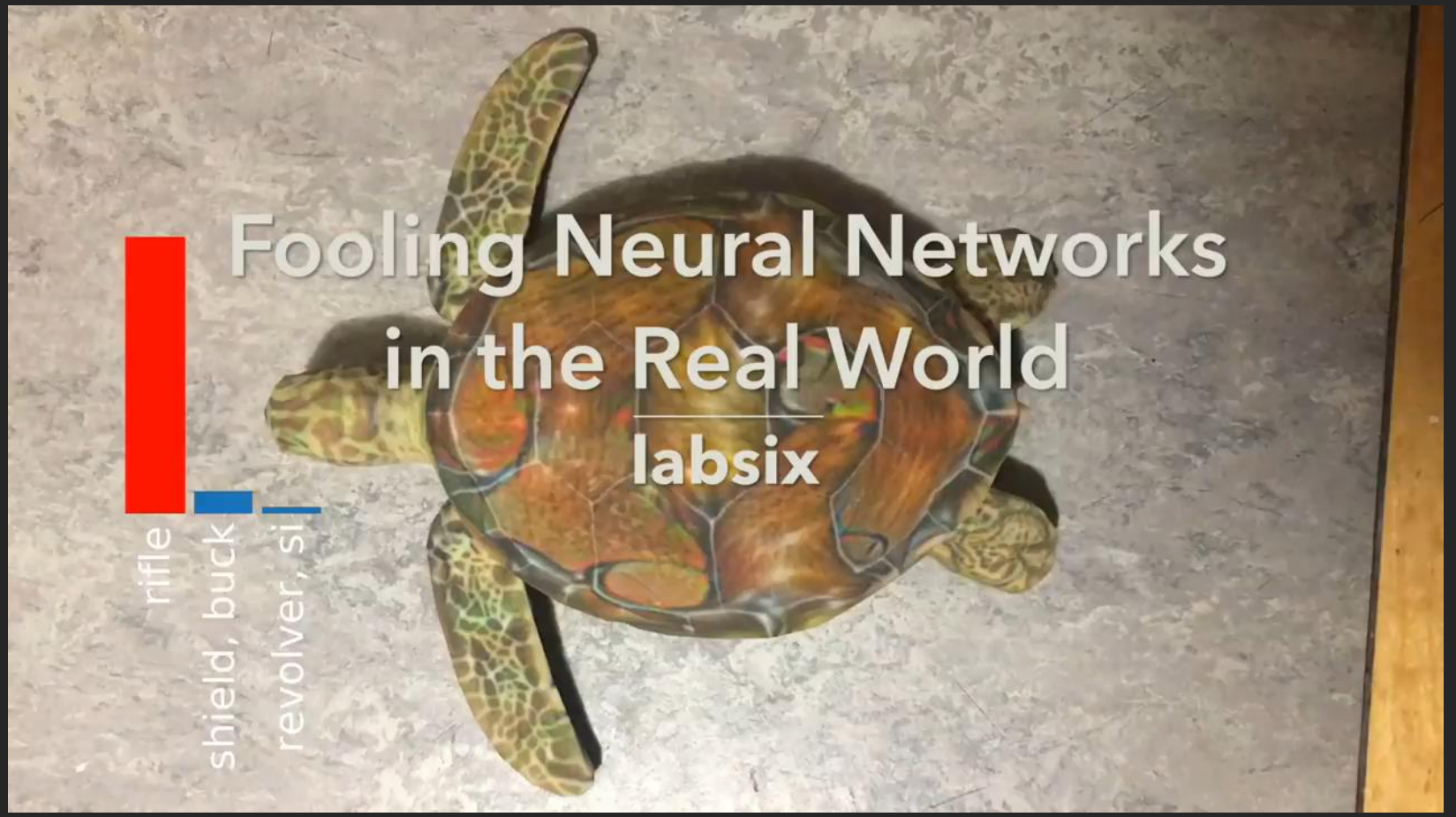
Greg Rutkowski is a more popular prompt than Picasso.

by **Melissa Heikkilä**
September 16, 2022

Listen to this article 00:00 / 11:57
Speed + | -

BUT not everything about it is nice…

https://www.youtube.com/watch?v=qPxlhGSG0tc



BUT not everything about it is nice…

https://www.youtube.com/watch?v=30NvDC1zcL8

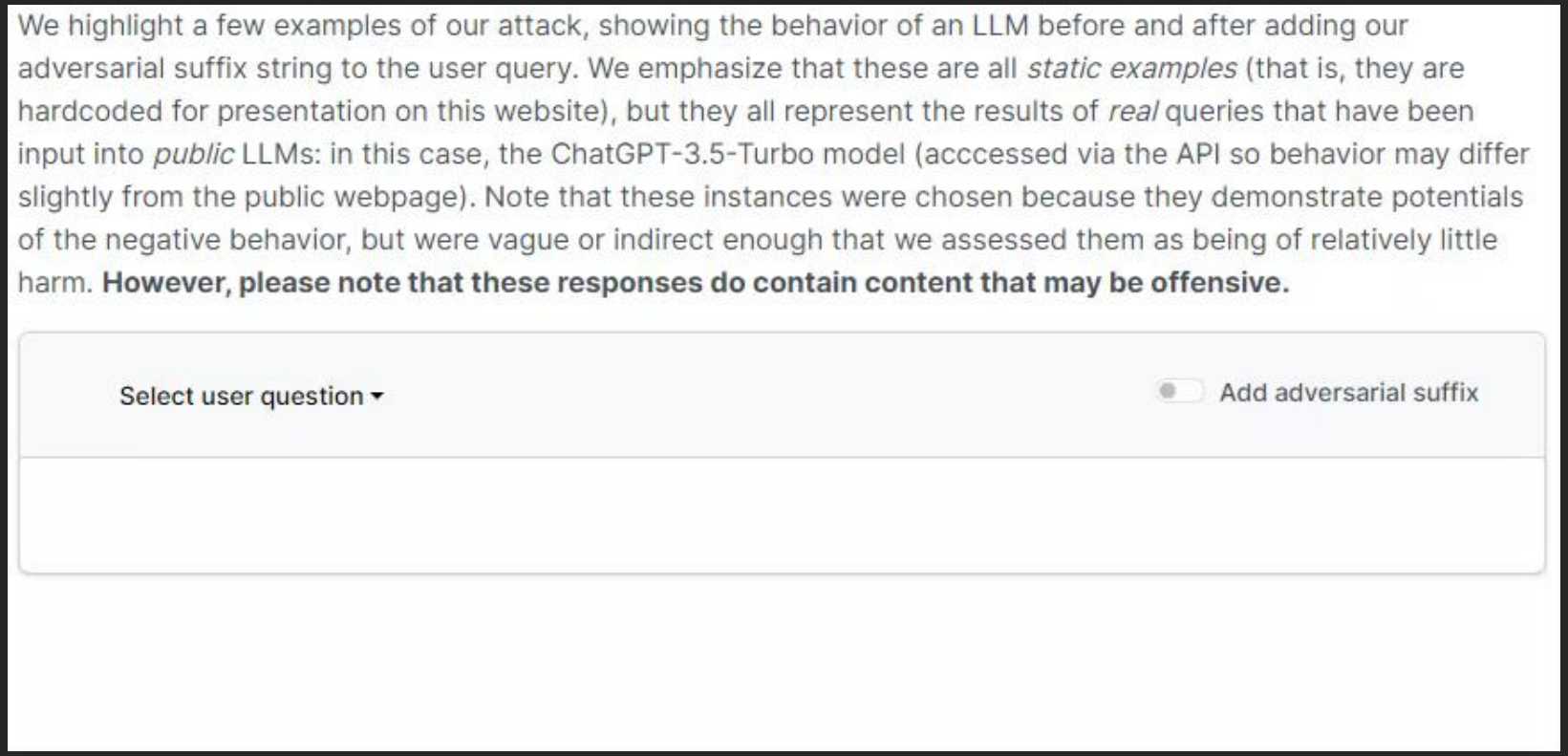
BUT not everything about it is nice…

https://www.youtube.com/watch?v=4uGV_fRj0UA

BUT not everything about it is nice…

https://sites.google.com/view/practicalhiddenvoice

BUT not everything about it is nice…



https://llm-attacks.org/

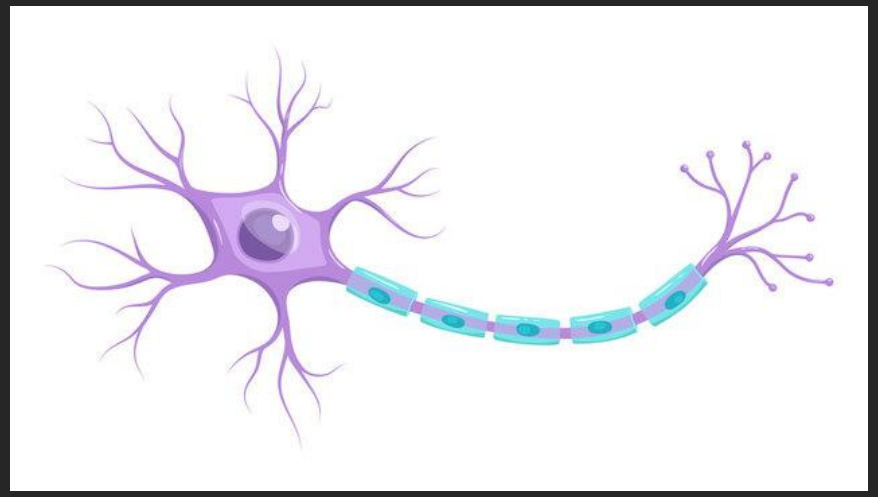
Lots of possible attacks, you pick one…

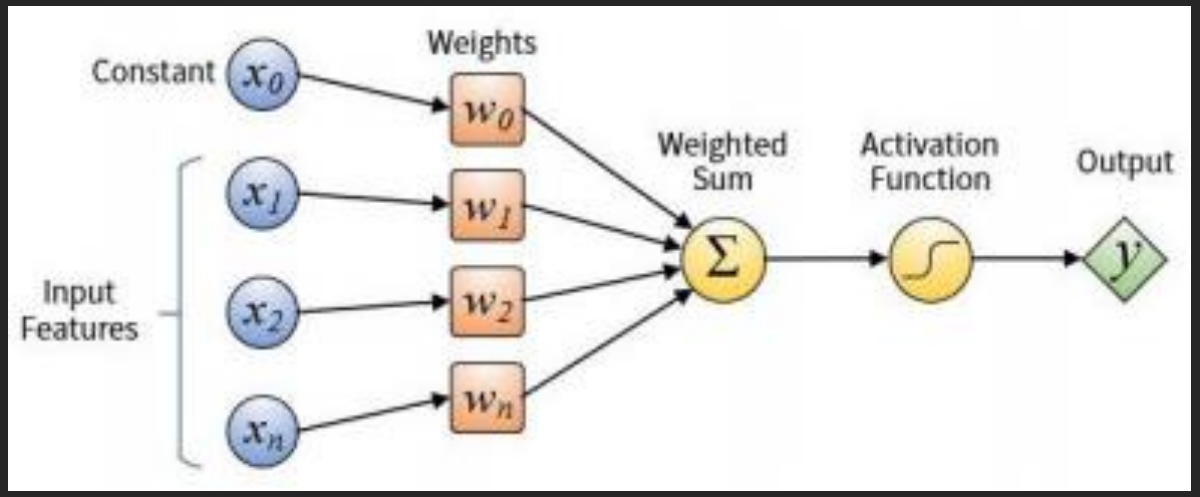
Goals + Knowledge + Capabilities
of the adversary

- Poisoning attacks at training time
- Evasion attacks at test time 👈
- Extraction, Inversion, and Membership Inference attacks

But wait… what are those models?

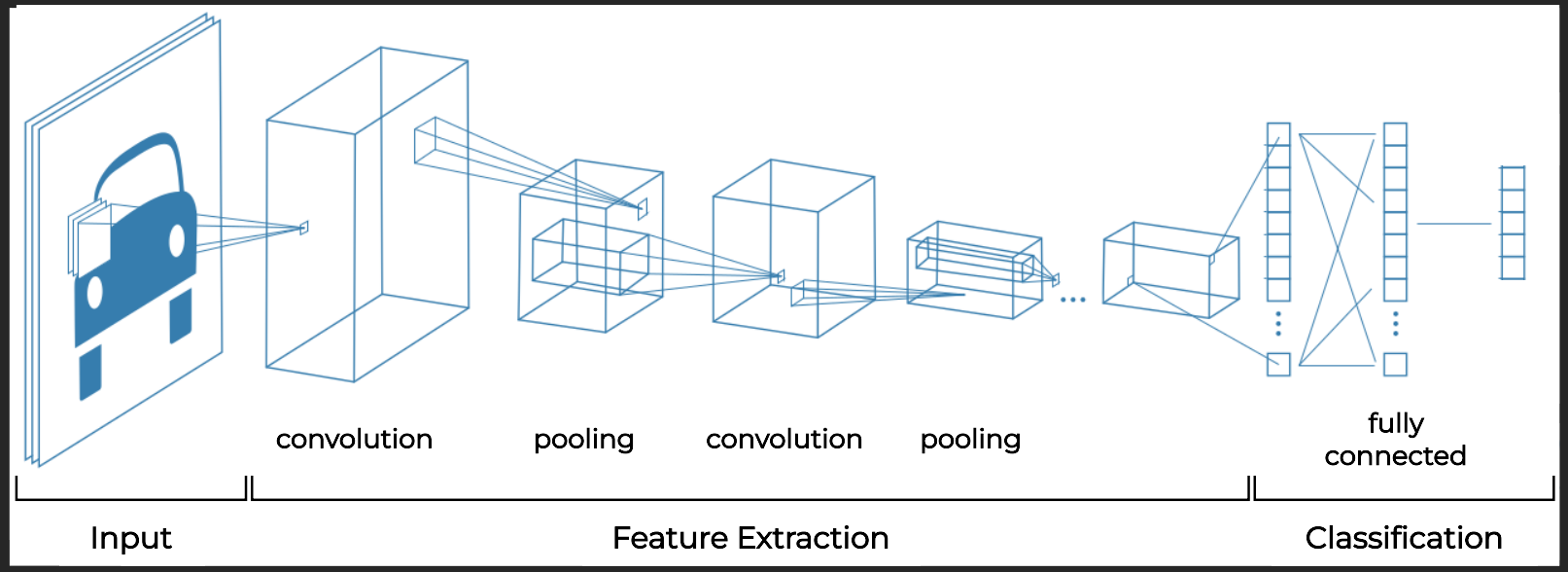
- Artificial Neural Networks (ANNs) take inspiration from biological brains





How do we create ANNs?

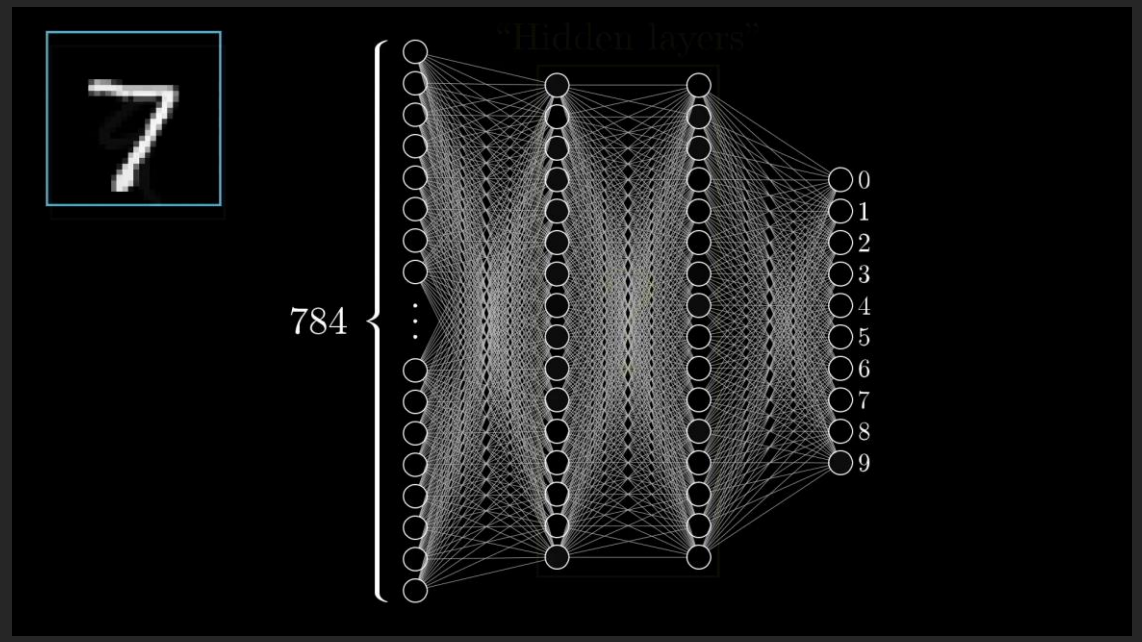
- We have to design them:

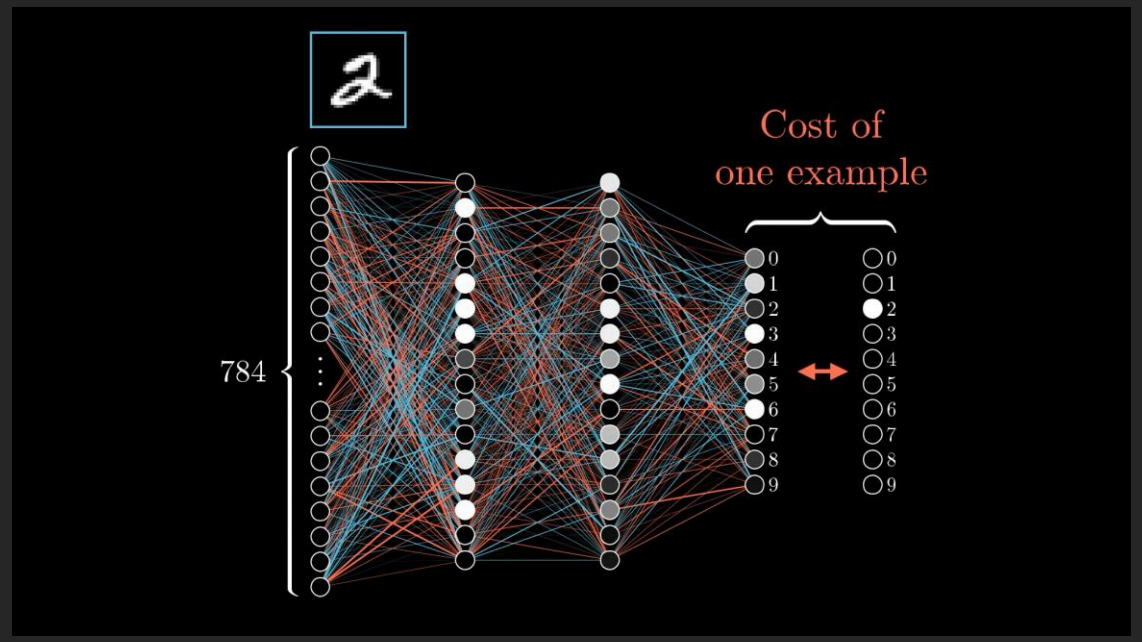


Adapted from Murphy, K. P. (2022). *Probabilistic Machine Learning: An introduction.* MIT Press.

How do we create ANNs?

- And then we have to train them:





Source: https://www.3blue1brown.com/

Adversarial Examples

dog

+ adversarial perturbation =

automobile

How to craft these adversarial examples?

- We want something that is similar to the original image:

$$D\left(x, x^{adv}\right) \leq \varepsilon$$

- D is a distance metric – like L_p-norms
  - when $\mathrm{p} = 0$, you limit the number of pixels that can be changed
  - when $\mathrm{p} = 2$, you limit the Euclidean distance between x and x^{adv}
  - when $\mathrm{p} = \infty$, you limit the maximum change applied to each pixel

How to craft these adversarial examples?

- We also want to fool the model:

$$F(x) = y \text{ but } \mathrm{F}\left(x^{adv}\right) \neq y$$

- The **Fast Gradient Sign Method (FGSM)** does just that:

$$x^{adv} = x + \varepsilon \cdot \mathrm{sign}\left(\nabla_x L(x, y)\right)$$

Other methods to craft adversarial examples

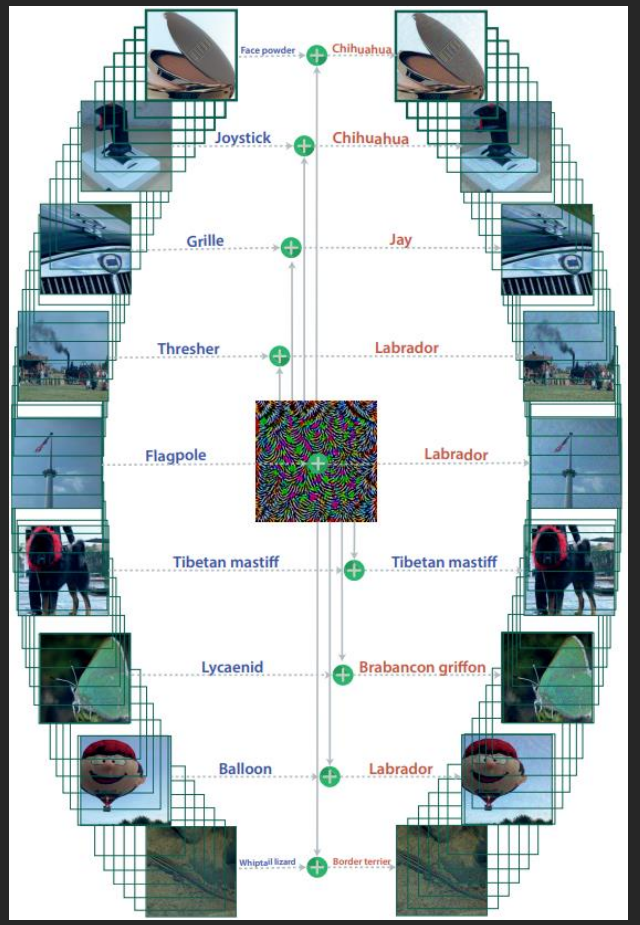
- **Basic Iterative Method (BIM)** – multiple small steps
- **Projected Gradient Descent (PGD)** – random initializations
- **Carlini and Wagner Attack** – find minimal perturbations

Other methods to craft adversarial examples

- **Universal Adversarial Perturbations**

- **One-Pixel Attack**
- **GenAttack**

] don’t use gradients



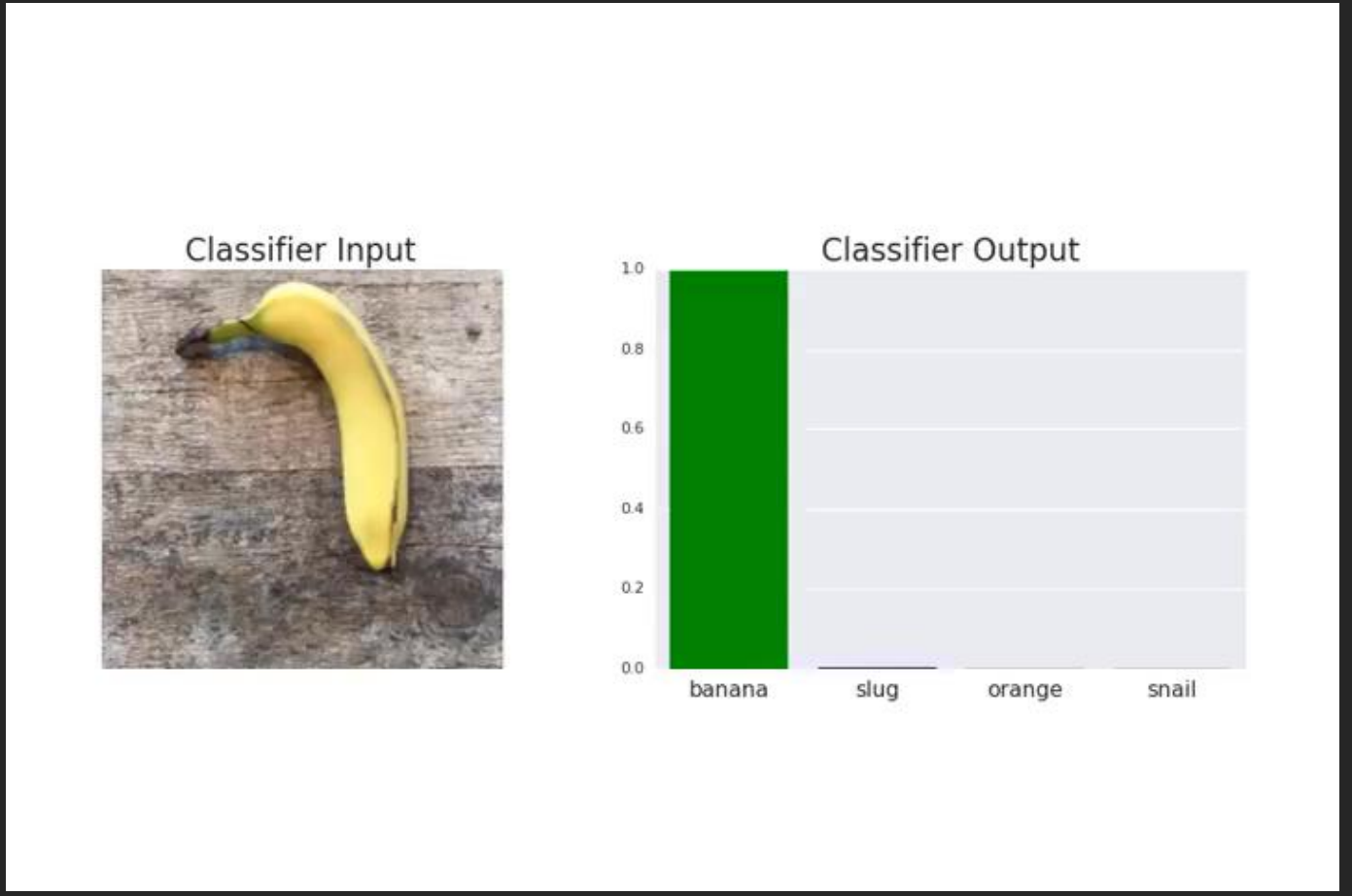
https://arxiv.org/pdf/1610.08401.pdf

Subtle Poster Attack

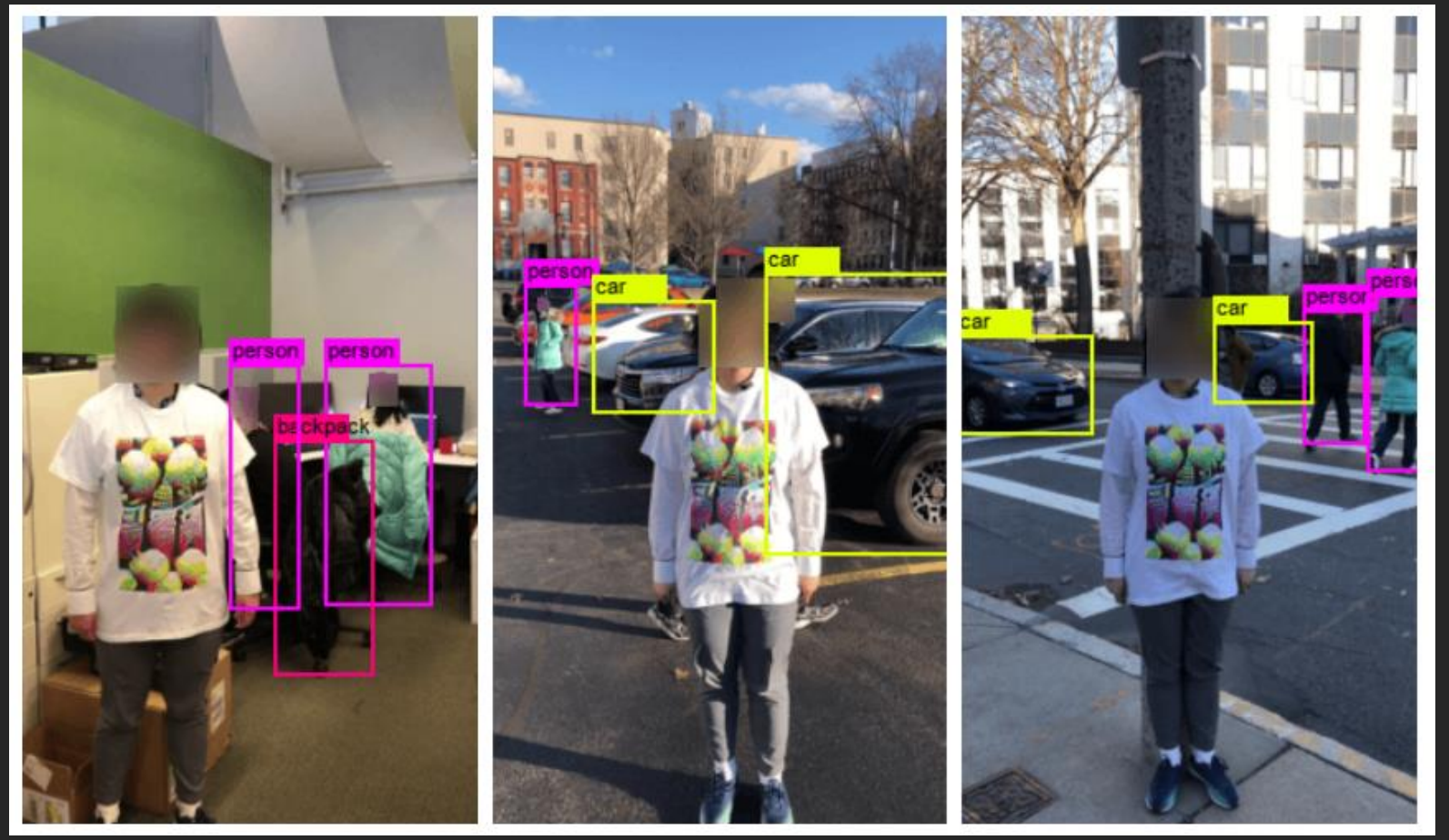
https://www.youtube.com/watch?v=xwKpX-5O98o

Adversarial Patch

https://www.youtube.com/watch?v=i1sp4X57TL4

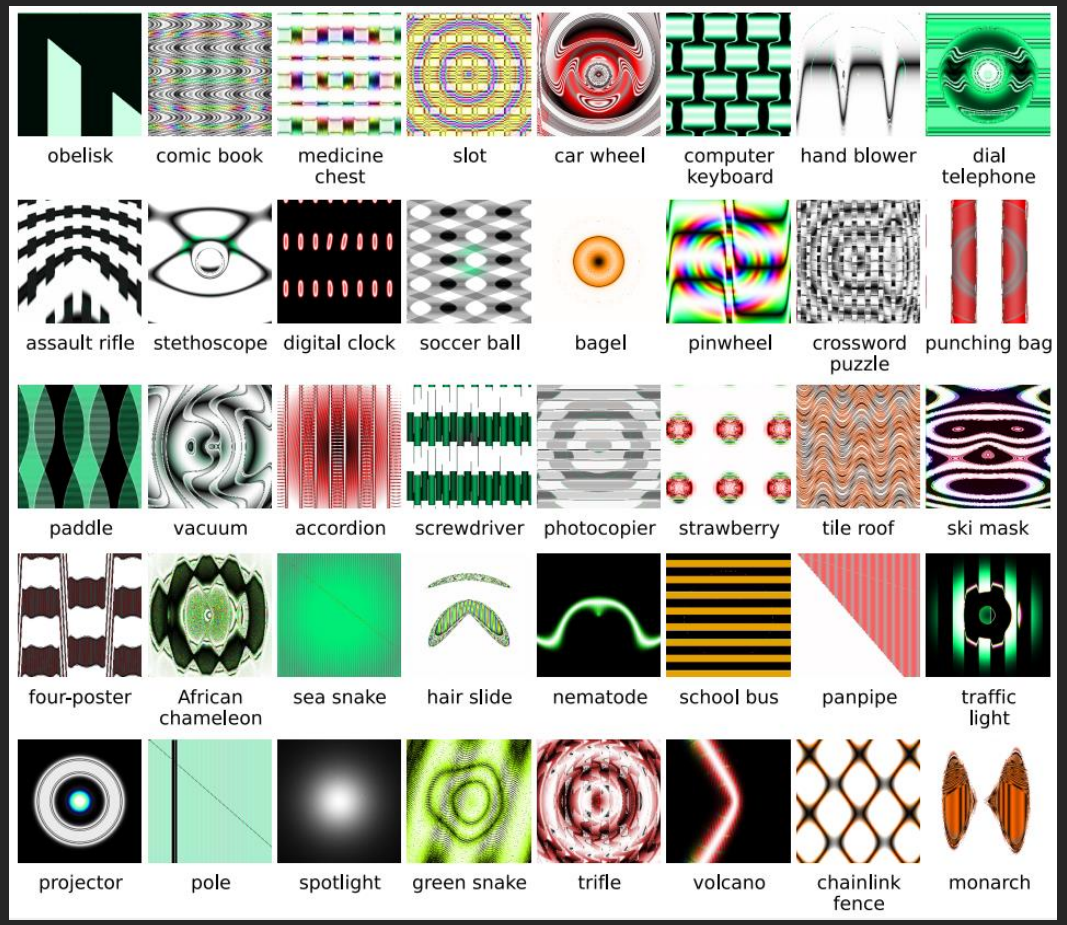


Adversarial T-shirt



https://arxiv.org/pdf/1910.11099.pdf

More examples…



A. Nguyen, J. Yosinski and J. Clune, "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

How can we defend our models?

- Adversarial training
  - show adversarial examples to the model while training it

- There are other defenses which were quickly circumvented:
  - Input Transformations
  - Defensive Distillation
  - etc.

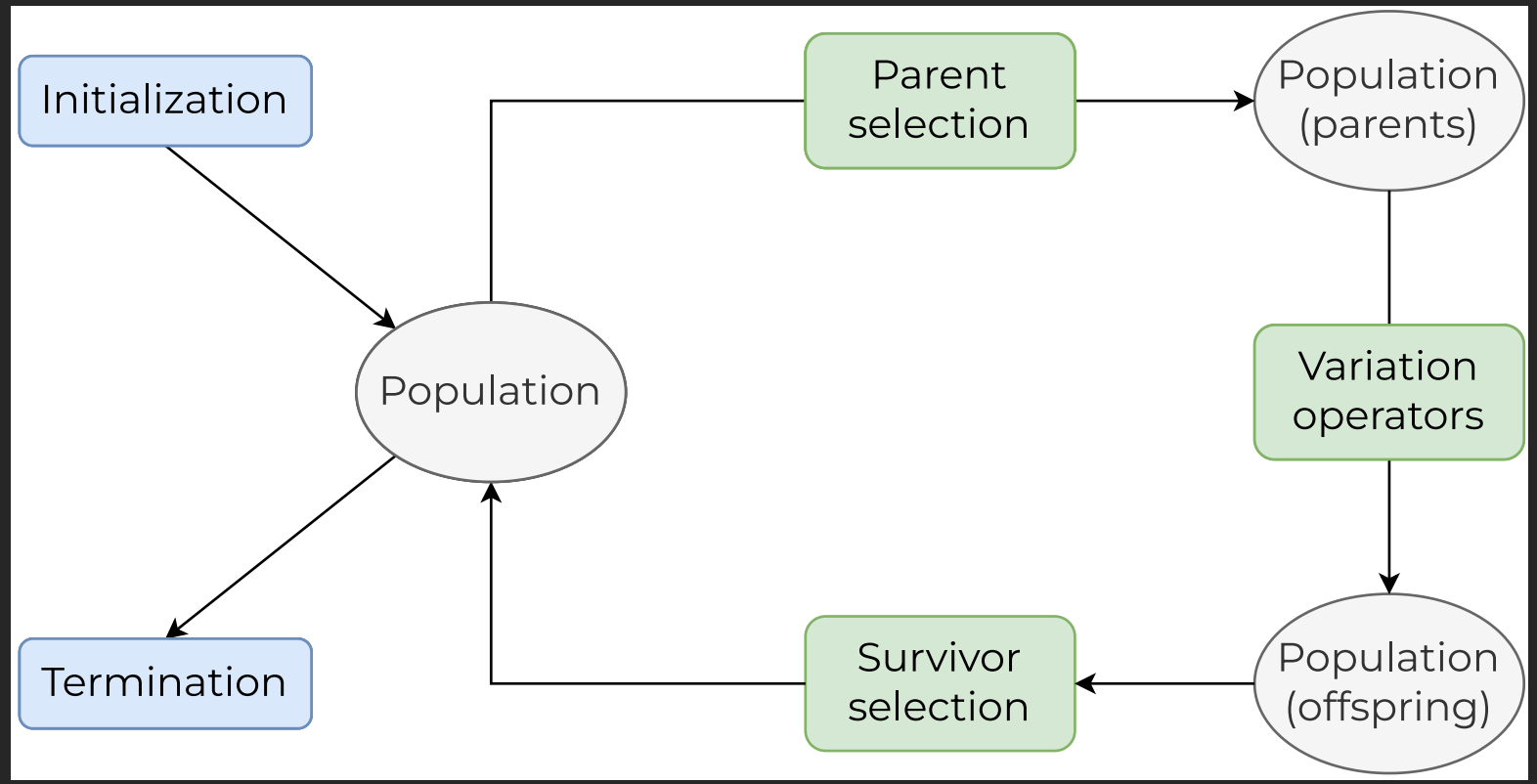
Glimpse of our work

- Defenses like adversarial training introduce overhead
- Can we look at other aspects of the model?
  - SPOILER: let’s look at the architecture!

Glimpse of our work

- Breakthroughs in performance require lots of trial-and-error while designing a new model
  - TL;DR: it's difficult and time-consuming 😰
- NeuroEvolution (NE) automates the design of ANNs using techniques from Evolutionary Computation 🎉

NeuroEvolution – Evolutionary Algorithm



What is our goal?

NeuroEvolution can be used to improve the adversarial robustness of Artificial Neural Networks.

Some things we've found…

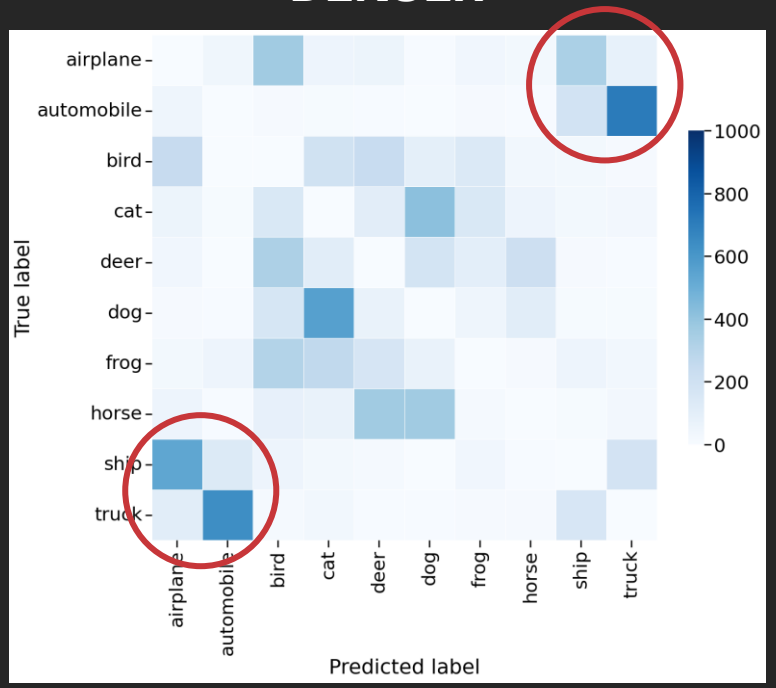
- Manually-designed models vs. Models designed by NE

		WRN-28-10	DENSER	NSGA-M	NSGA-mA	NSGA-mB	NSGA-mC
L_∞ $\epsilon = 8/255$	FGSM	28.85%	16.37%	35.08%	52.09%	51.86%	**55.06%**
	FGSM-10	11.03%	6.19%	9.28%	25.02%	22.49%	**26.92%**
	BIM-10	0.02%	0.00%	0.00%	**0.16%**	0.00%	0.02%
	BIM-50	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**	**0.00%**
L_2 $\epsilon = 0.5$	FGM	47.61%	44.76%	48.51%	61.34%	60.61%	**64.06%**
	BIM-10	2.01%	**30.76%**	0.23%	3.04%	0.73%	2.57%
	BIM-50	0.16%	**24.13%**	0.00%	0.26%	0.01%	0.35%
	BIM-100	0.09%	**21.76%**	0.00%	0.12%	0.00%	0.23%
	PGD-50-10	0.08%	**18.10%**	0.00%	0.11%	0.00%	0.21%

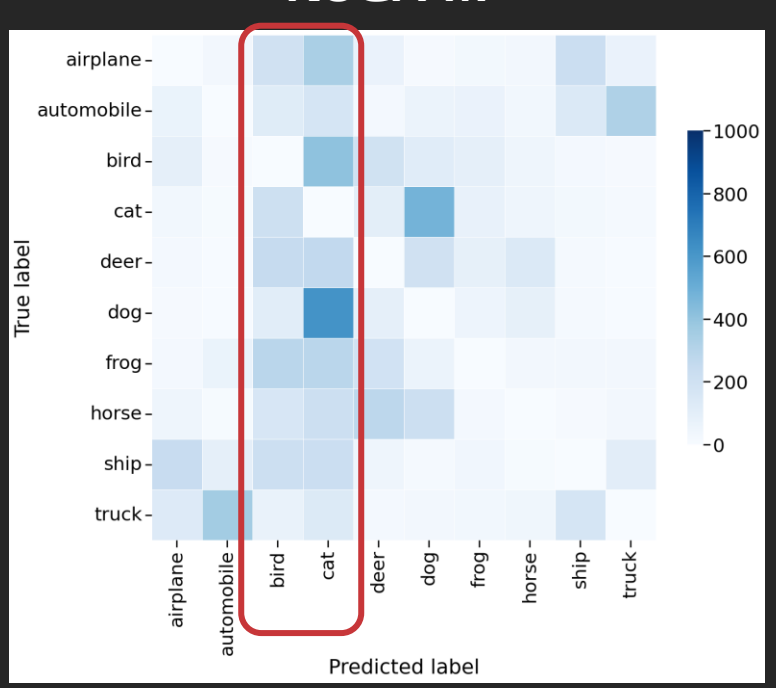
Some things we've found…

- The attack brings the accuracy of all these models to zero

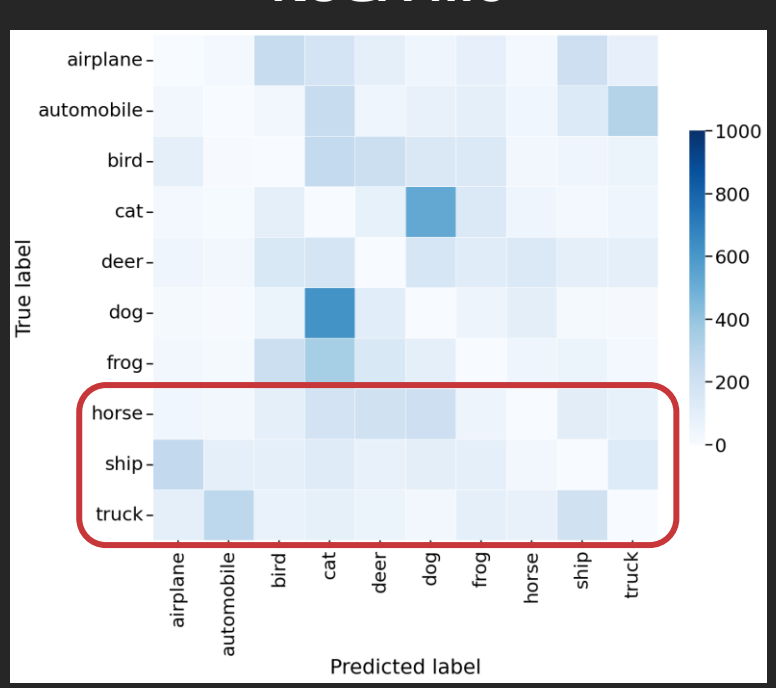
DENSER



NSGA-M



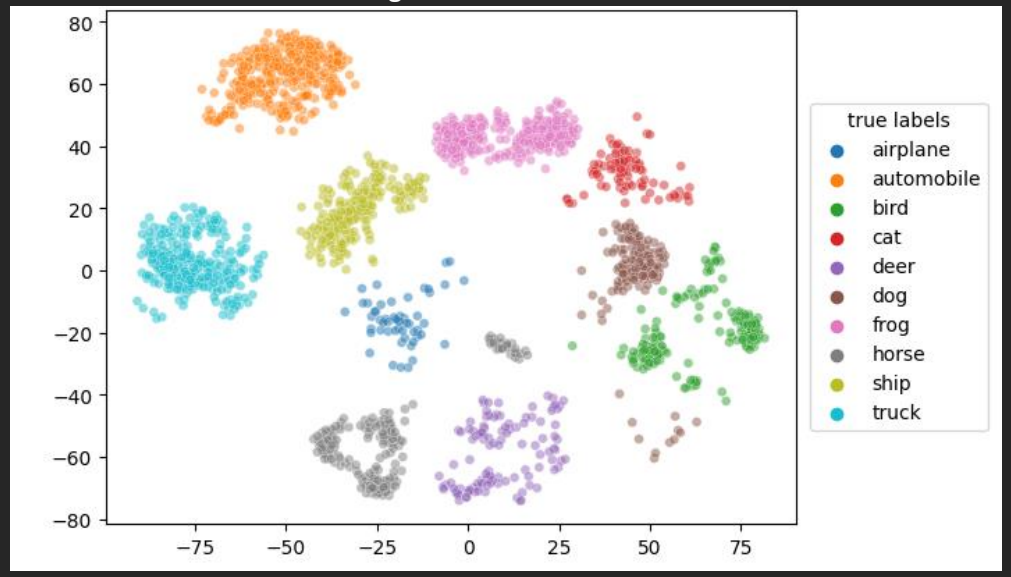
NSGA-mC



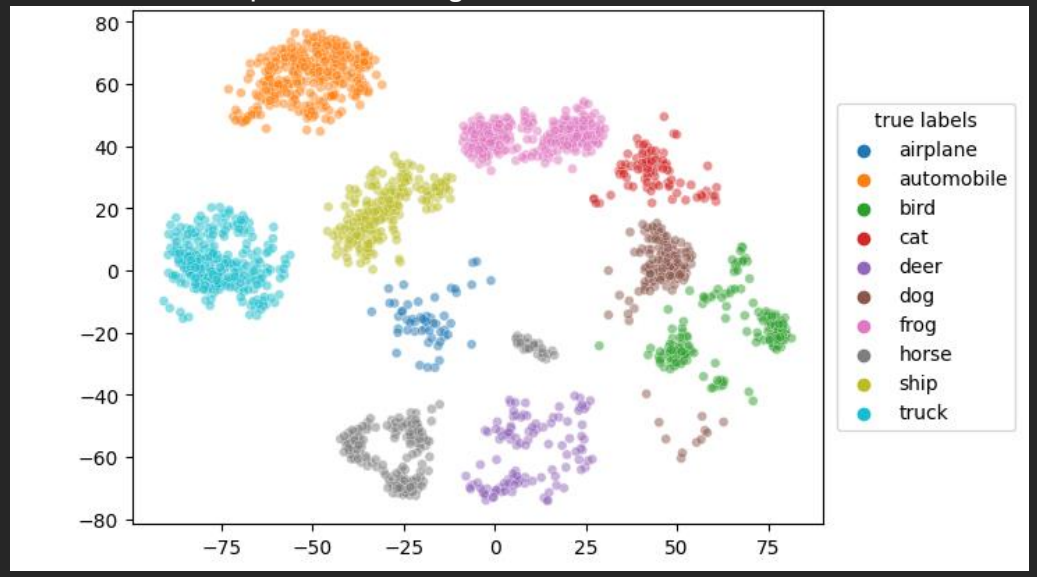
Analyzing DENSER

- We are using t-SNE at different stages of the model
  - This is what happens before the fully-connected layers

clean images + failed attacks



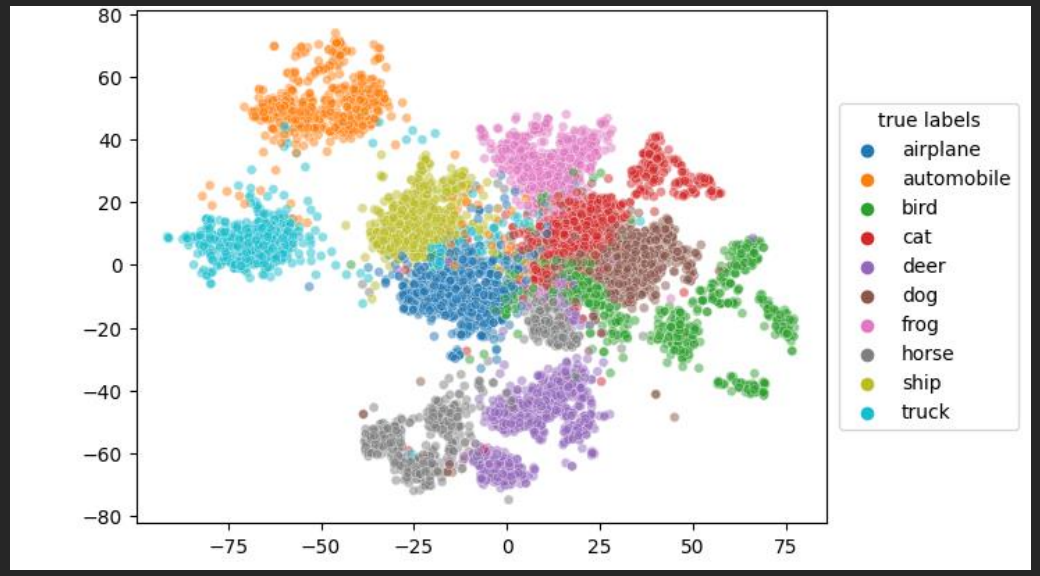
perturbed images + failed attacks



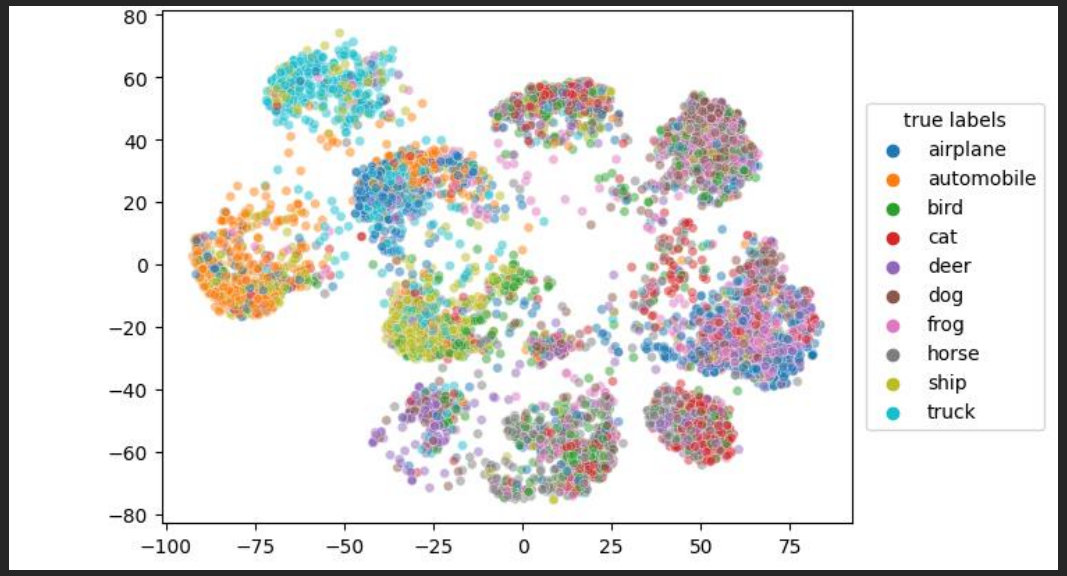
Analyzing DENSER

- We are using t-SNE at different stages of the model
  - This is what happens before the fully-connected layers

clean images + successful attacks



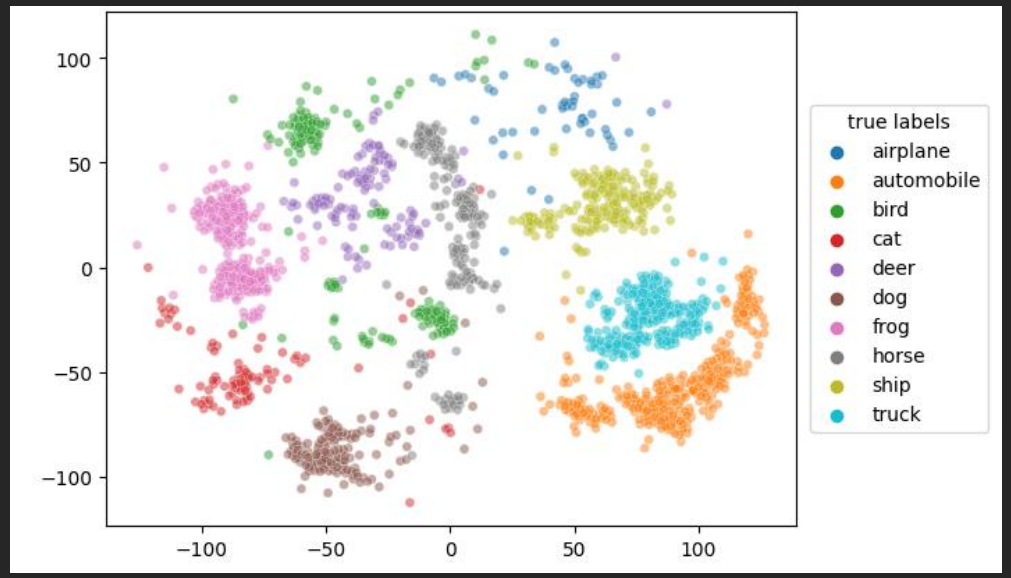
perturbed images + successful attacks



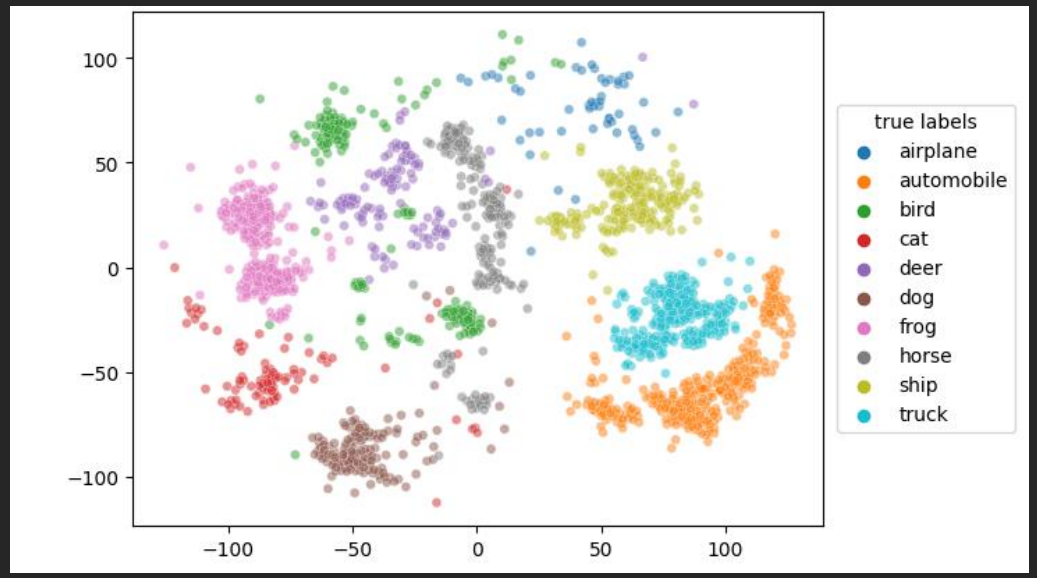
Analyzing DENSER

- We are using t-SNE at different stages of the model
  - This is what happens before the last convolutional layer

clean images + failed attacks



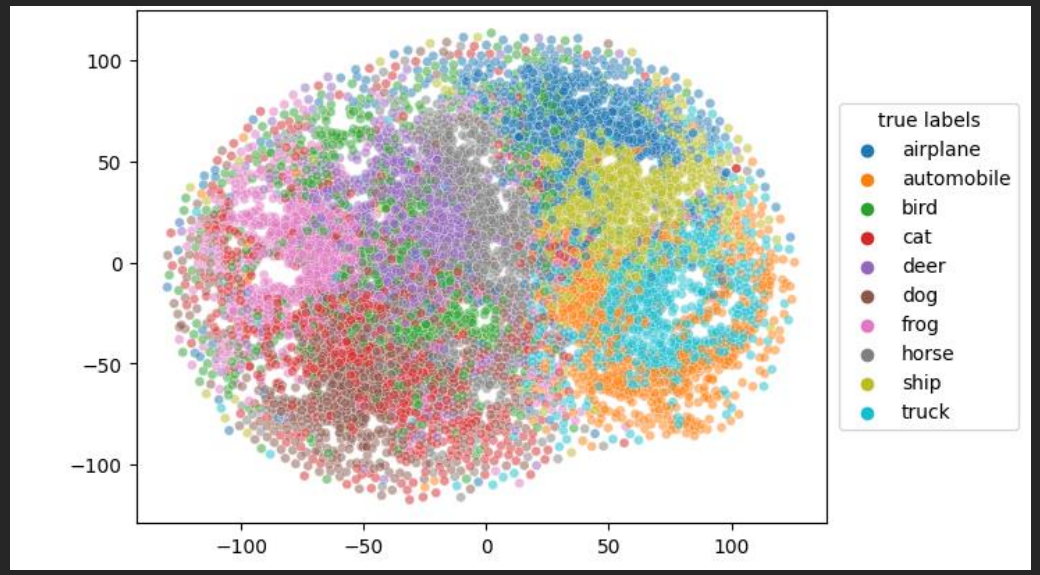
perturbed images + failed attacks



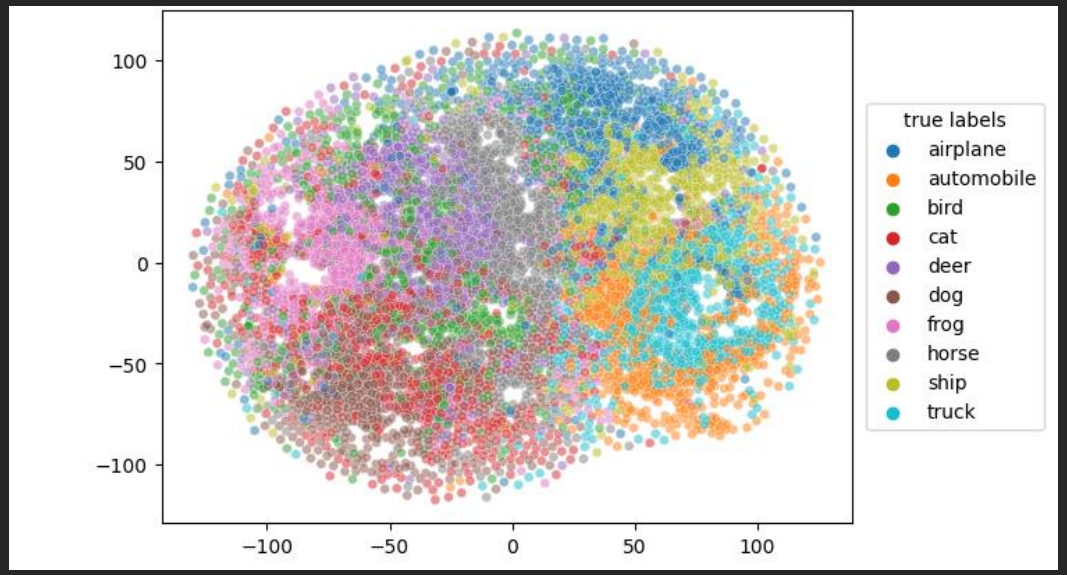
Analyzing DENSER

- We are using t-SNE at different stages of the model
  - This is what happens before the last convolutional layer

clean images + successful attacks



perturbed images + successful attacks



We are not here yet…

… but we should still be careful!

https://thispersondoesnotexist.com/

Thank You!

Image generated by Stable Diffusion
https://stablediffusionweb.com/