Slides with GIF:

https://docs.google.com/presentation/d/e/2PACX-1vQz6blrrMWZWibGXsMPNwIJlRE0B9v0zR4FG4y44vS95ShXhD43eOEQQB9UzGJp9JM9bF5L7OWwTfjH/pub?start=false&loop=false&delayms=60000

rethinking adversarial examples

yahya jabary

disco group - andreas plesner, prof. roger wattenhofer
dsg group - alireza furutanpey, prof. schahram dustdar

let's start with a game -

quick, what do you see?

bagel

- or -

dog?

chicken wing

- or -

dog?

croissant

- or -

dog?

mop

- or -

dog?

muffin

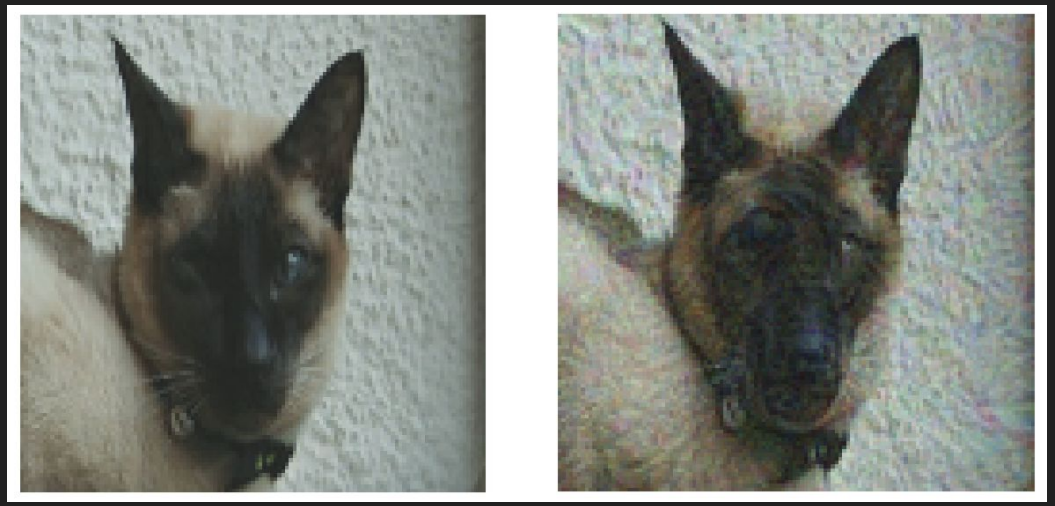
- or -

dog?

1 person

- or -

2 people?

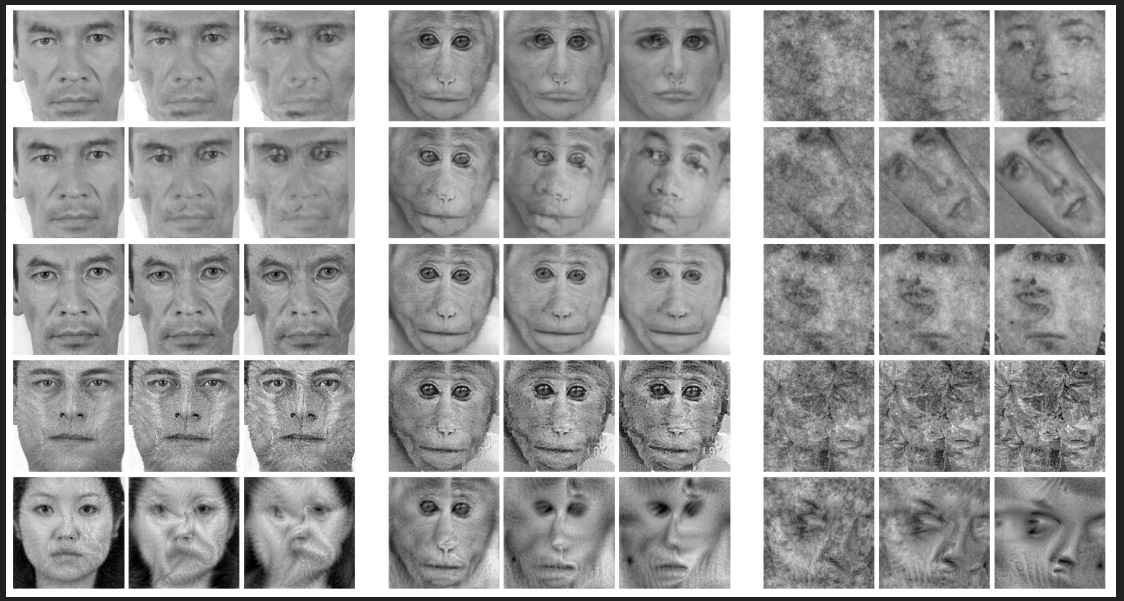


cat

- or -

dog?

Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., & Sohl-Dickstein, J. (2018).
Adversarial examples that fool both computer vision and time-limited humans. Advances in neural information processing systems, 31.

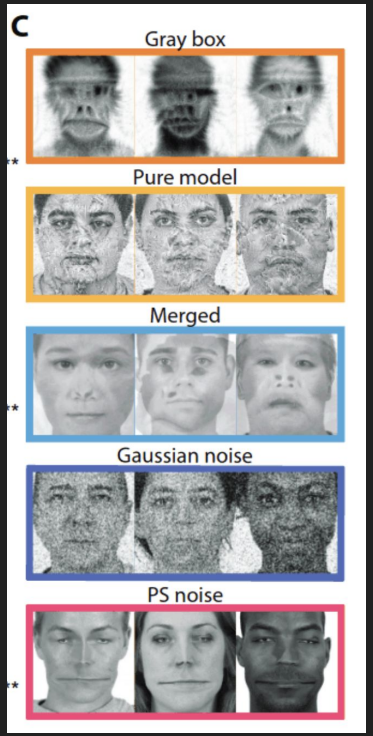


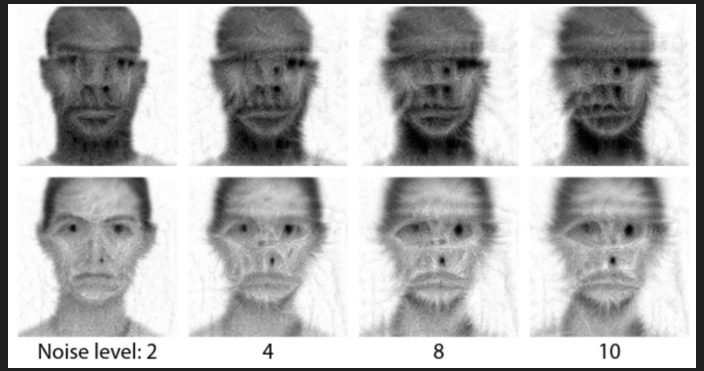
human

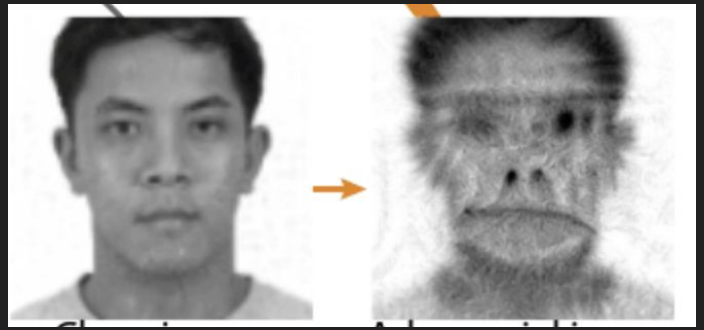
- or -

monkey?

Yuan, L., Xiao, W., Dellaferrera, G., Kreiman, G., Tay, F. E., Feng, J., & Livingstone, M. S. (2020). Fooling the primate brain with minimal, targeted image manipulation. arXiv preprint arXiv:2011.05623.





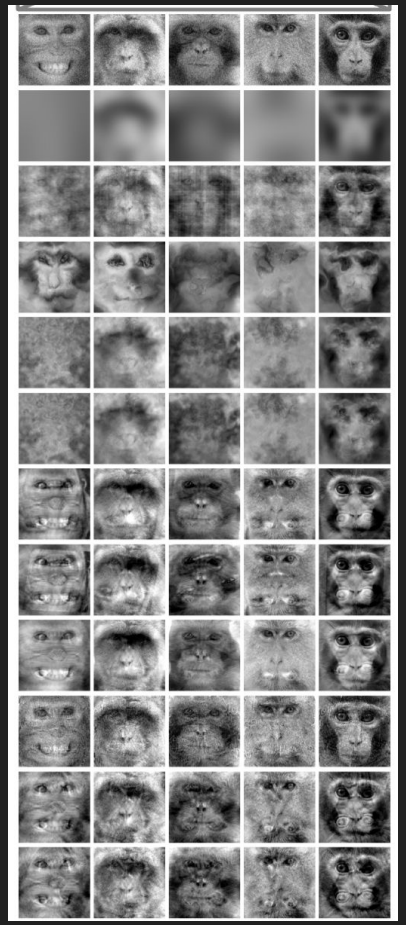


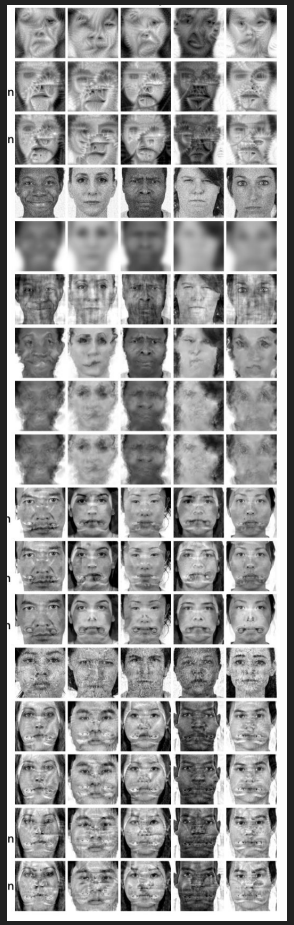
human

- or -

monkey?

Yuan, L., Xiao, W., Dellaferrera, G., Kreiman, G., Tay, F. E., Feng, J., & Livingstone, M. S. (2020). Fooling the primate brain with minimal, targeted image manipulation. arXiv preprint arXiv:2011.05623.



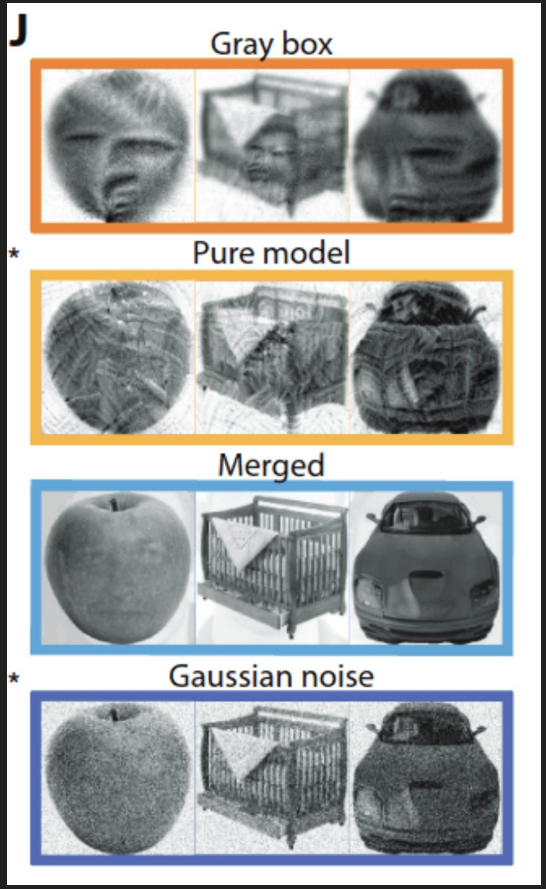


human

- or -

monkey?

Yuan, L., Xiao, W., Dellaferrera, G., Kreiman, G., Tay, F. E., Feng, J., & Livingstone, M. S. (2020). Fooling the primate brain with minimal, targeted image manipulation. arXiv preprint arXiv:2011.05623.

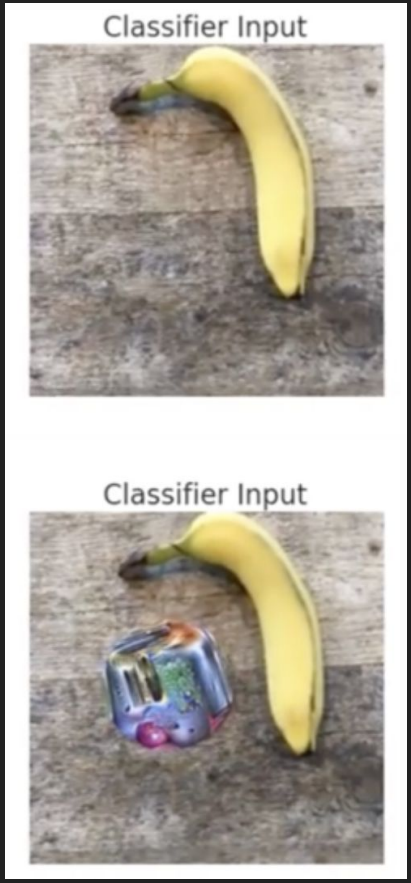


face

- or -

no face?

Yuan, L., Xiao, W., Dellaferrera, G., Kreiman, G., Tay, F. E., Feng, J., & Livingstone, M. S. (2020). Fooling the primate brain with minimal, targeted image manipulation. arXiv preprint arXiv:2011.05623.

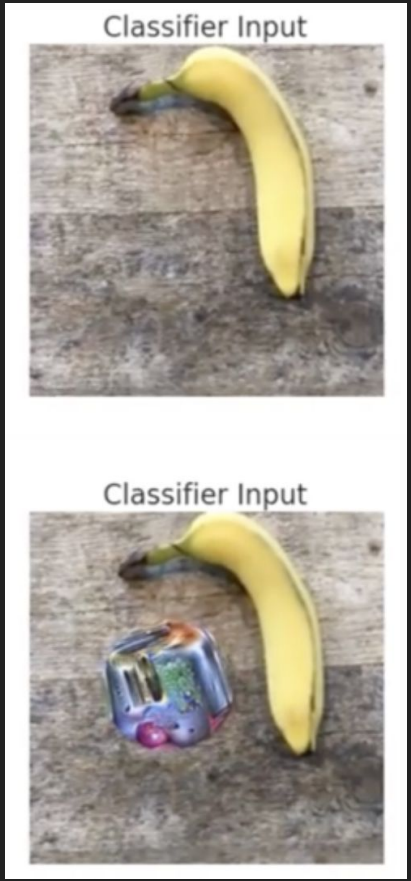


banana

- or -

toaster?

Brown, T. B., Mané, D., Roy, A., Abadi, M., & Gilmer, J. (2017). Adversarial patch. arXiv preprint arXiv:1712.09665. (VGG16 vs. white-box ensemble method attack)

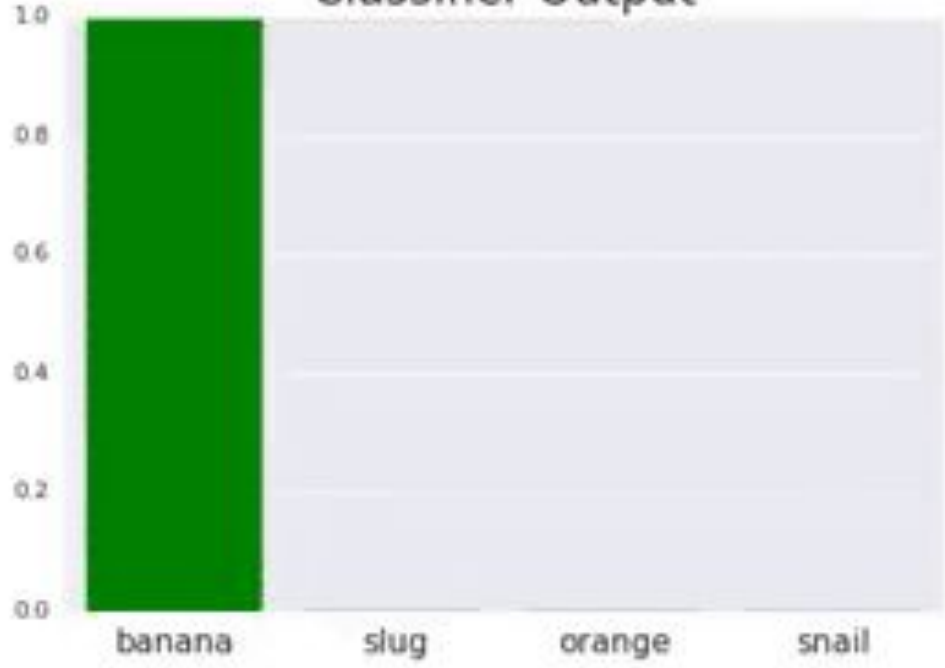


too easy right?

not for vision models!

Classifier Input

Classifier Output



we call these "weird" examples
"adversarial" examples.

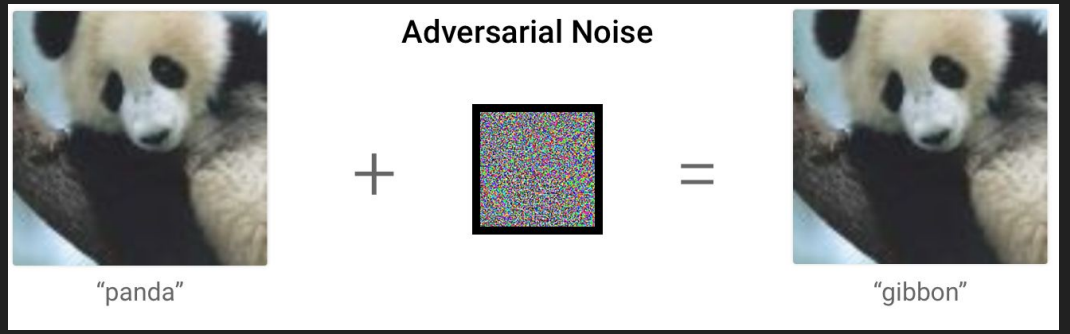
but you can call them however you want

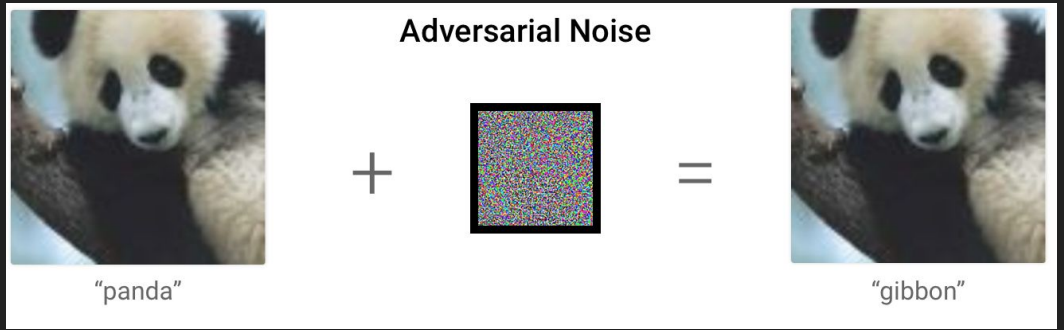
statistics: out-of-distribution data / adversarial distribution / latent space collapse / information bottleneck failures / covariate shift outliers / data drift (covariate vs. concept) / distribution mismatch / complexity artifacts / overfitting residues / feature attribution misalignment / simpson's paradox / concept drift / black swan events / long tail events / shortcut learning / optimization instabilities

cybersecurity: jailbreaks / gradient-sign evasion attacks / zero-day model vulnerabilities

psychology: cognitive and reasoning biases / perceptual dissonance / deep learning pareidolia / anthropomorphism errors / semantic illusions / confabulation effects / hallucinations / illusory correlations / pattern misperceptions

what are they?


Adversarial Noise
“panda”
“gibbon”

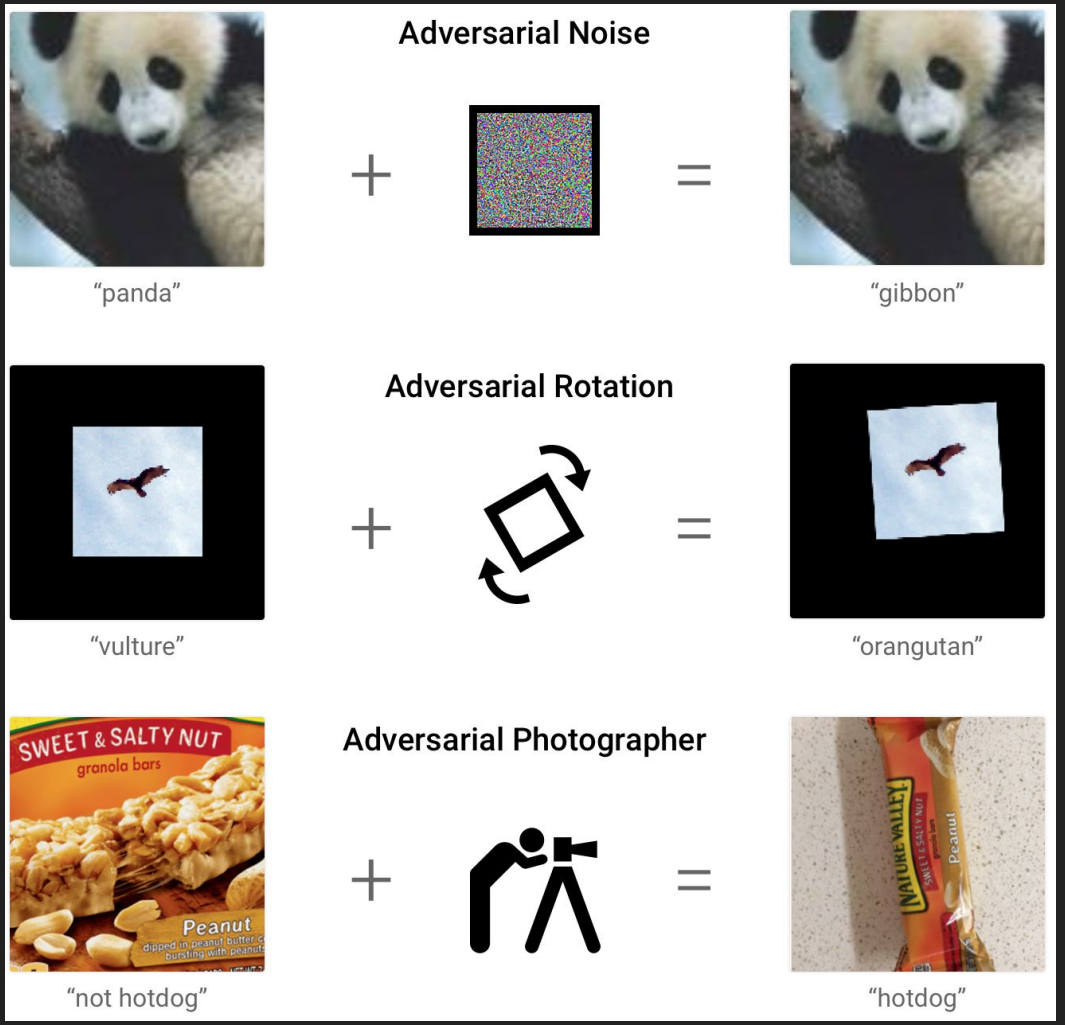


but wait – there’s more!

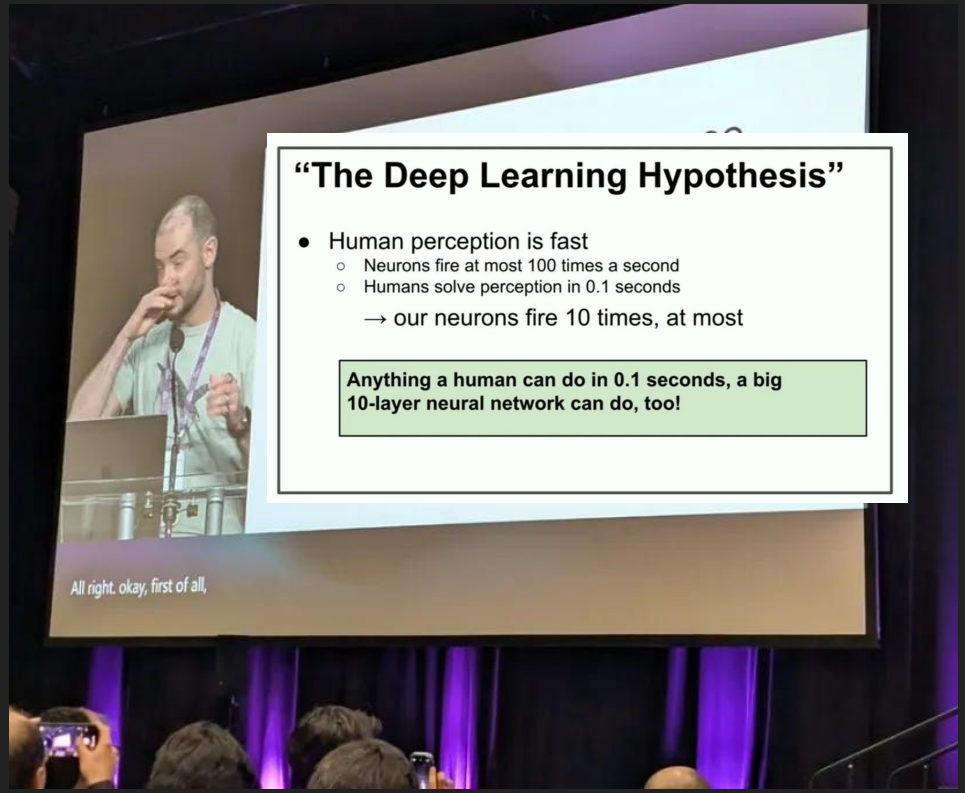
traditional
/ imperceptible
(2013)

unrestricted
/ semantics preserving
(2019)

natural
(2019)



some attacks also work on time-constrained humans & primates!



idea:

neural networks emulate time-constrained humans

NeurIPS 2024

test-of-time paper awards

how do they work?

non-robust features hypothesis

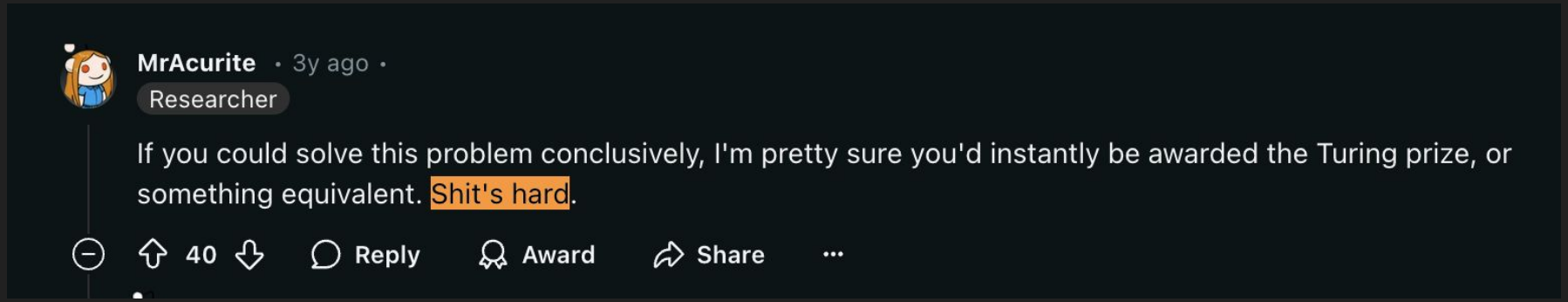
feature types:

- a) useless — not picked up by feature extractor
- b) robust — detect *concept drift*
  generalize to test-set
  comprehensible to humans
- c) non-robust — detect *covariate drift*
  very predictive, but only in train-set

  e.g. the fur direction of an animal

Ilyas, A. et al. (2019). Adversarial examples are not bugs, they are features. Advances in neural information processing systems, 32.

open problem



https://www.reddit.com/r/MachineLearning/comments/qhnao8/d_how_do_you_deal_with_covariate_shift_and/

why should i care?

use-cases

hack the reviewer-assignment system

No more Reviewer #2: Subverting Automatic Paper-Reviewer Assignment using Adversarial Learning

Thorsten Eisenhofer*†, Erwin Quiring*‡, Jonas Möller§, Doreen Riepel†,
Thorsten Holz¶, Konrad Rieck§

† *Ruhr University Bochum*
‡ *International Computer Science Institute (ICSI) Berkeley*
§ *Technische Universität Berlin*
¶ *CISPA Helmholtz Center for Information Security*

Abstract

The number of papers submitted to academic conferences is steadily rising in many scientific disciplines. To handle this growth, systems for automatic *paper-reviewer assignments* are increasingly used during the reviewing process. These systems use statistical topic models to characterize the content of submissions and automate the assignment to reviewers. In this paper, we show that this automation can be manipulated using adversarial learning. We propose an attack that adapts a given paper so that it misleads the assignment and selects its own reviewers. Our attack is based on a novel optimization strategy that alternates between the feature space and problem space to realize unobtrusive changes to the paper. To evaluate the feasibility of our attack, we simulate the paper-reviewer assignment of an actual security conference (IEEE S&P) with 165 reviewers on the program committee. Our results show that we can successfully select and remove reviewers without access to the assignment system. Moreover, we demonstrate that the manipulated papers remain plausible and are often indistinguishable from benign submissions.

1 Introduction

Peer review is a major pillar of academic research and the

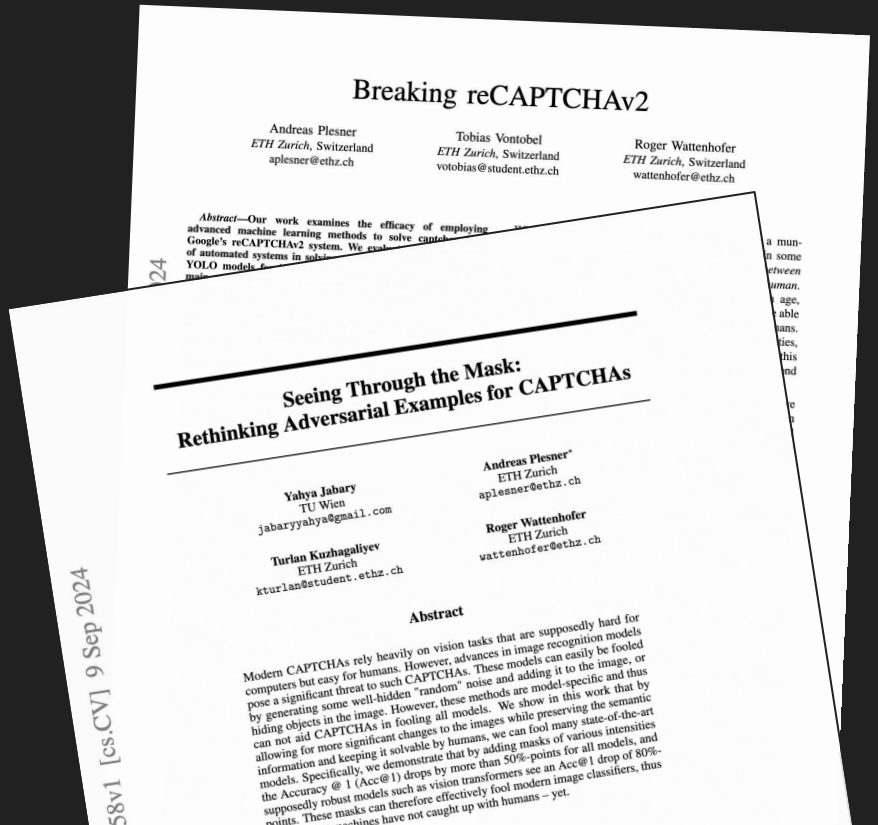
To handle this growth, conference management tools have become indispensable in peer review. They allow reviewers to bid for submissions and support the program chair to find a good assignment bas
nately, even these tool
submissions continue
intractable, as for ex
Major conferences in
missions that need t
reviewers [31]. For t
are increasingly ext
reviewer assignment
from machine learn
missions, and auto

In this work, we
to manipulate the
prior work that focused on bid manipulations and reviewer
collusion [24, 57], our attack rests on adversarial learning. In particular, we propose an attack that adapts a given paper so that it misleads the underlying topic model. This enables us to reject and select specific reviewers from the program committee. To reach this goal, we introduce a novel optimization strategy that alternates between the feature space and problem space when adapting a paper. This optimization allows us to preserve the semantics and plausibility of the document, while carefully changing the assignment of reviewers.

hack the topic-model
used for NeurIPS & ICML

USENIX 2023

solve & make better captchas!



Breaking reCAPTCHAv2

Andreas Plesner
ETH Zurich, Switzerland
aplesner@ethz.ch

Tobias Vontobel
ETH Zurich, Switzerland
votobias@student.ethz.ch

Roger Wattenhofer
ETH Zurich, Switzerland
wattenhofer@ethz.ch

Abstract—Our work examines the efficacy of employing advanced machine learning methods to solve Google's reCAPTCHAv2 system. We evaluate of automated systems in solving YOLO models

Seeing Through the Mask:
Rethinking Adversarial Examples for CAPTCHAs

Yahya Jabary
TU Wien
jabaryyahya@gmail.com

Andreas Plesner*
ETH Zurich
aplesner@ethz.ch

Turlan Kuzhagaliyev
ETH Zurich
kturlan@student.ethz.ch

Roger Wattenhofer
ETH Zurich
wattenhofer@ethz.ch

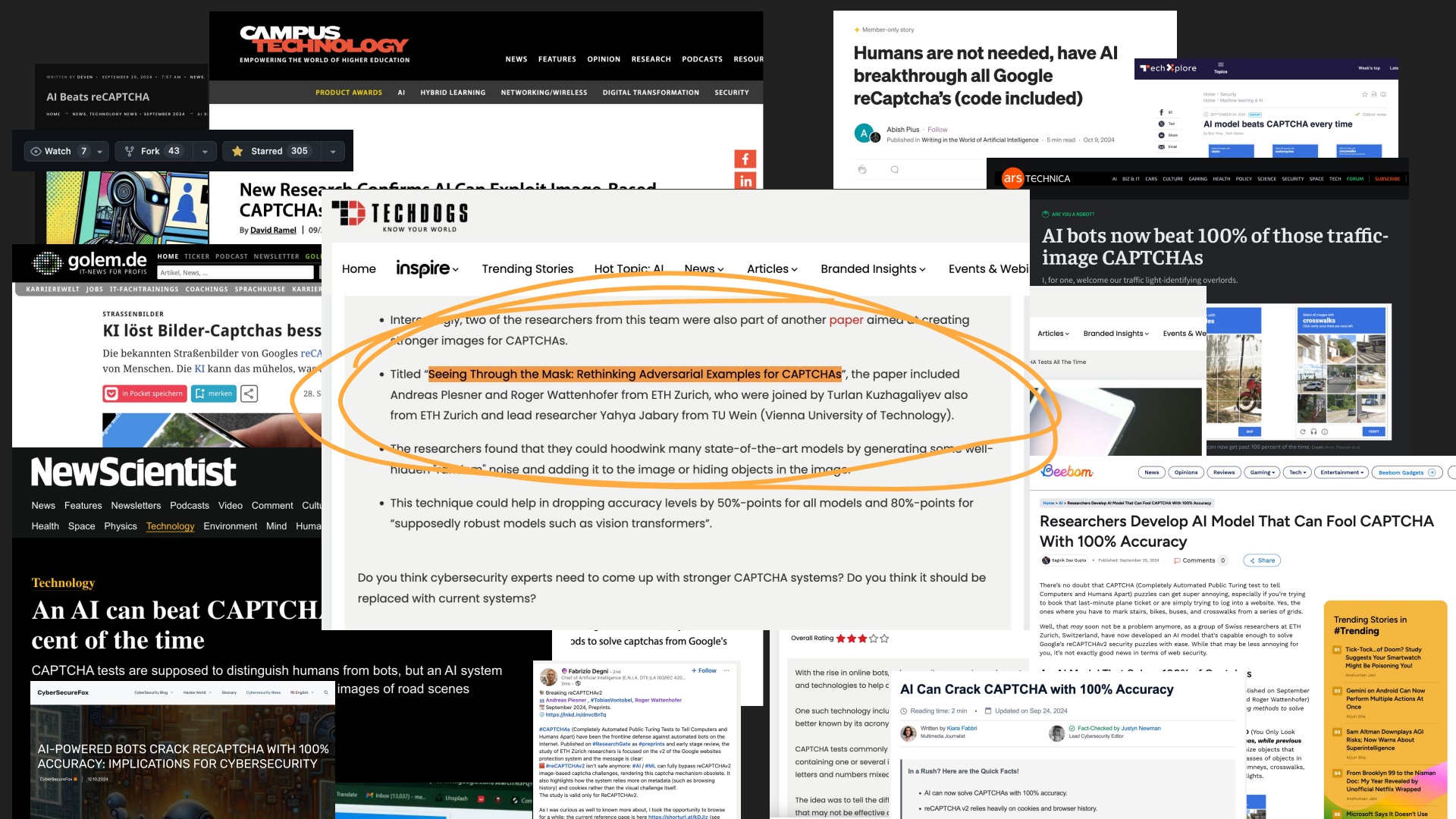
Abstract

Modern CAPTCHAs rely heavily on vision tasks that are supposedly hard for computers but easy for humans. However, advances in image recognition models pose a significant threat to such CAPTCHAs. These models can easily be fooled by generating some well-hidden "random" noise and adding it to the image, or hiding objects in the image. However, these methods are model-specific and thus can not aid CAPTCHAs in fooling all models. We show in this work that by allowing for more significant changes to the images while preserving the semantic information and keeping it solvable by humans, we can fool many state-of-the-art models. Specifically, we demonstrate that by adding masks of various intensities the Accuracy @ 1 (Acc@1) drops by more than 50%-points for all models, and supposedly robust models such as vision transformers see an Acc@1 drop of 80%-points. These masks can therefore effectively fool modern image classifiers, thus

[cs.CV] 9 Sep 2024

google's reCAPTCHA v2

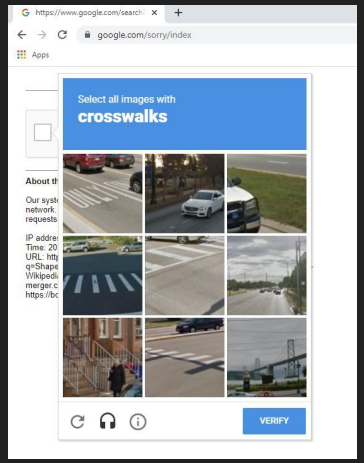
- 99% market share
- 100% solve rate


CAMPUS TECHNOLOGY
EMPOWERING THE WORLD OF HIGHER EDUCATION
NEWS FEATURES OPINION RESEARCH PODCASTS
PRODUCT AWARDS AI HYBRID LEARNING NETWORKING/WIRELESS DIGITAL TRANSFORMATION SECURITY
AI Beats reCAPTCHA
Watch 7 Fork 43 Starred 305
Member-only story
Humans are not needed, have AI breakthrough all Google reCaptcha's (code included)
Abish Pius · Follow
Published in Writing in the World of Artificial Intelligence · 5 min read · Oct 9, 2024
TechXplore
AI model beats CAPTCHA every time
New Research Confirms AI Can Exploit Image-Based CAPTCHAs
By David Ramel
TECHDOGS
KNOW YOUR WORLD
Home inspire Trending Stories Hot Topic: AI News Articles Branded Insights Events & Webi
golem.de
IT-NEWS FÜR PROFIS
HOME TICKER PODCAST NEWSLETTER
KARRIEREWELT JOBS IT-FACHTRAININGS COACHINGS SPRACHKURSE
STRASSENBILDER
KI löst Bilder-Captchas bess
Die bekannten Straßenbilder von Googles reCA
von Menschen. Die KI kann das mühelos, was
in Pocket speichern merken
ars TECHNICA
AI BIZ & IT CARS CULTURE GAMING HEALTH POLICY SCIENCE SECURITY SPACE TECH FORUM SUBSCRIBE
ARE YOU A ROBOT?
AI bots now beat 100% of those traffic-image CAPTCHAs
I, for one, welcome our traffic light-identifying overlords.
Articles Branded Insights Events & We
A Tests All The Time
Beebom
News Opinions Reviews Gaming Tech Entertainment Beebom Gadgets
NewScientist
News Features Newsletters Podcasts Video Comment
Health Space Physics Technology Environment Mind
Technology
An AI can beat CAPTCH
cent of the time
CAPTCHA tests are supposed to distinguish humans from bots, but an AI system
images of road scenes
Interestingly, two of the researchers from this team were also part of another paper aimed at creating stronger images for CAPTCHAs.
Titled "Seeing Through the Mask: Rethinking Adversarial Examples for CAPTCHAs", the paper included Andreas Plesner and Roger Wattenhofer from ETH Zurich, who were joined by Turlan Kuzhagaliyev also from ETH Zurich and lead researcher Yahya Jabary from TU Wein (Vienna University of Technology).
The researchers found that they could hoodwink many state-of-the-art models by generating some well-hidden "[illegible]" noise and adding it to the image or hiding objects in the image.
This technique could help in dropping accuracy levels by 50%-points for all models and 80%-points for "supposedly robust models such as vision transformers".
Do you think cybersecurity experts need to come up with stronger CAPTCHA systems? Do you think it should be replaced with current systems?
Researchers Develop AI Model That Can Fool CAPTCHA With 100% Accuracy
Sagnik Das Gupta · Published: September 25, 2024
Comments 0 Share
There's no doubt that CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) puzzles can get super annoying, especially if you're trying to book that last-minute plane ticket or are simply trying to log into a website. Yes, the ones where you have to mark stairs, bikes, buses, and crosswalks from a series of grids.
Well, that may soon not be a problem anymore, as a group of Swiss researchers at ETH Zurich, Switzerland, have now developed an AI model that's capable enough to solve Google's reCAPTCHAv2 security puzzles with ease. While that may be less annoying for you, it's not exactly good news in terms of web security.
Trending Stories in #Trending
01 Tick-Tock...of Doom? Study Suggests Your Smartwatch Might Be Poisoning You!
02 Gemini on Android Can Now Perform Multiple Actions At Once
03 Sam Altman Downplays AGI Risks; Now Warns About Superintelligence
04 From Brooklyn 99 to the Nisman Doc: My Year Revealed by Unofficial Netflix Wrapped
05 Microsoft Says It Doesn't Use
ods to solve captchas from Google's
Overall Rating
With the rise in online bots, and technologies to help c
One such technology inclu better known by its acrony
CAPTCHA tests commonly containing one or several l letters and numbers mixed
The idea was to tell the diff that may not be effective
Fabrizio Degni · 2nd
Chief of Artificial Intelligence (E.N.I.A. DTE (LA ISOJEC 420...
3mo
Follow
Breaking reCAPTCHAv2
Andreas Plesner, #ToBiasVontobel, Roger Wattenhofer
September 2024, Preprints.
https://lnkd.in/dnwcBnTq
#CAPTCHAs (Completely Automated Public Turing Tests to Tell Computers and Humans Apart) have been the frontline defense against automated bots on the Internet. Published on #ResearchGate as #preprints and early stage review, the study of ETH Zürich researchers is focused on the v2 of the Google websites protection system and the message is clear:
#reCAPTCHAv2 isn't safe anymore: #AI / #ML can fully bypass reCAPTCHAv2 image-based captcha challenges, rendering this captcha mechanism obsolete. It also highlights how the system relies more on metadata (such as browsing history) and cookies rather than the visual challenge itself.
The study is valid only for ReCAPTCHAv2.
As I was curious as well to known more about, I took the opportunity to browse
CyberSecureFox
AI-POWERED BOTS CRACK RECAPTCHA WITH 100% ACCURACY: IMPLICATIONS FOR CYBERSECURITY
AI Can Crack CAPTCHA with 100% Accuracy
Reading time: 2 min · Updated on Sep 24, 2024
Written by Kiara Fabbri
Multimedia Journalist
Fact-Checked by Justyn Newman
Lead Cybersecurity Editor
In a Rush? Here are the Quick Facts!
AI can now solve CAPTCHAs with 100% accuracy.
reCAPTCHA v2 relies heavily on cookies and browser history.

arms race

robustified model

challenger



prover / solver

adversarial attack

step 1: making harder challenges

hCaptchas are surprisingly hard to solve!

simplicity enables formal reasoning:

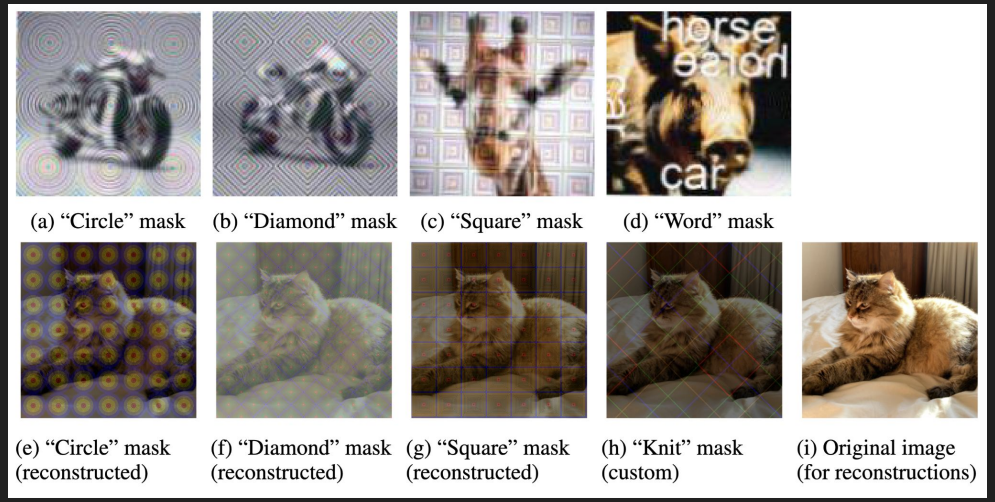
- tracing changes in latent space
- linear increase of "attack strength" (opacity)

steps

1. remake the geometric masks

original:

reconstruction:



(a) “Circle” mask (b) “Diamond” mask (c) “Square” mask (d) “Word” mask

(e) “Circle” mask (reconstructed) (f) “Diamond” mask (reconstructed) (g) “Square” mask (reconstructed) (h) “Knit” mask (custom) (i) Original image (for reconstructions)

steps

1. remake the geometric masks
2. find the best of each kind

hyperparam search using a single model

metrics:

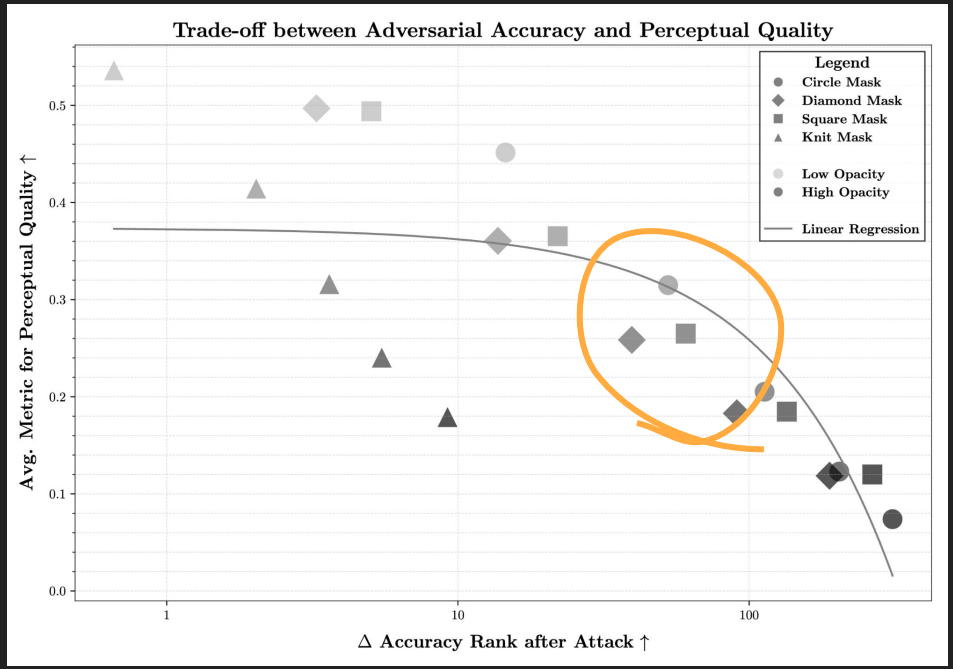
(1) attack strength
    = drop of correct label rank (acc@1, acc@5)

(2) recognizability
    = cosine similarity, PSNR, SSIM, LPIPS

we want both to be high!

steps

1. remake the geometric masks
2. find the best of each kind
3. eval them on SOTA models



recognizability

attack strength

steps

1. remake the geometric masks
2. find the best of each kind
3. eval them on SOTA models

we confirmed that these masks are:

✔ simple

✔ functional

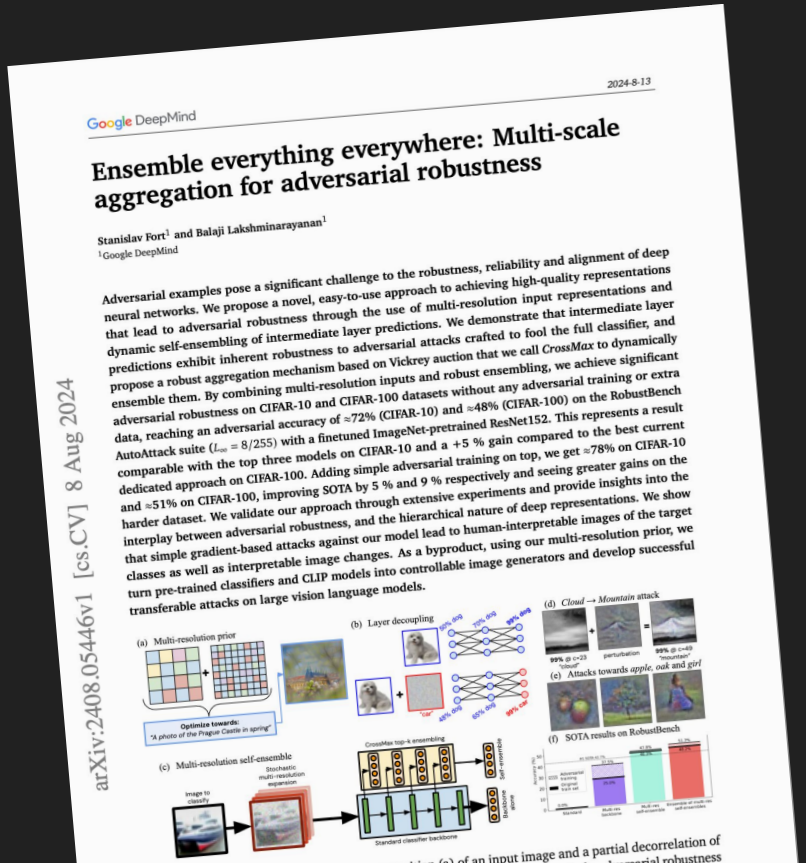
✔ transferable

inverse attack strength

recognizability

Opacity	Mask	Δ Acc Rank	Quality	Score
50	Circle	-14.57	0.45	15.02
	Diamond	-3.27	0.50	3.76
	Knit	-0.66	0.54	1.19
	Square	-5.04	0.49	5.54
80	Circle	-52.72	0.31	53.03
	Diamond	-13.72	0.36	14.08
	Knit	-2.03	0.41	2.44
	Square	-22.01	0.37	22.37
110	Circle	-113.07	0.21	113.27
	Diamond	-39.55	0.26	39.81
	Knit	-3.62	0.32	3.93
	Square	-60.57	0.27	60.84
140	Circle	-203.89	0.12	204.01
	Diamond	-90.79	0.18	90.97
	Knit	-5.47	0.24	5.71
	Square	-134.75	0.18	134.94
170	Circle	-310.80	0.07	310.88
	Diamond	-188.92	0.12	189.04
	Knit	-9.21	0.18	9.39
	Square	-264.90	0.12	265.02

step 2: making better solvers



Google DeepMind

2024-8-13

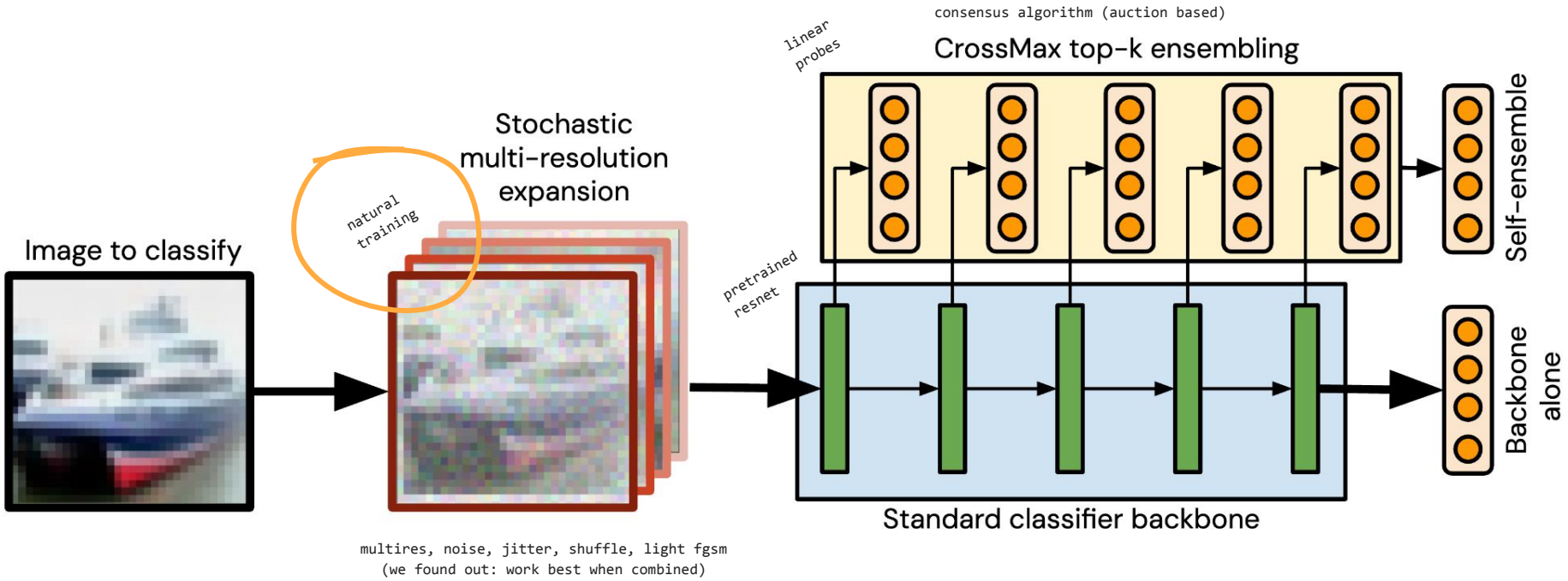
Ensemble everything everywhere: Multi-scale aggregation for adversarial robustness

Stanislav Fort[1] and Balaji Lakshminarayanan[1]
[1]Google DeepMind

Adversarial examples pose a significant challenge to the robustness, reliability and alignment of deep neural networks. We propose a novel, easy-to-use approach to achieving high-quality representations that lead to adversarial robustness through the use of multi-resolution input representations and dynamic self-ensembling of intermediate layer predictions. We demonstrate that intermediate layer predictions exhibit inherent robustness to adversarial attacks crafted to fool the full classifier, and propose a robust aggregation mechanism based on Vickrey auction that we call *CrossMax* to dynamically ensemble them. By combining multi-resolution inputs and robust ensembling, we achieve significant adversarial robustness on CIFAR-10 and CIFAR-100 datasets without any adversarial training or extra data, reaching an adversarial accuracy of ≈72% (CIFAR-10) and ≈48% (CIFAR-100) on the RobustBench AutoAttack suite ($L_\infty = 8/255$) with a finetuned ImageNet-pretrained ResNet152. This represents a result comparable with the top three models on CIFAR-10 and a +5 % gain compared to the best current dedicated approach on CIFAR-100. Adding simple adversarial training on top, we get ≈78% on CIFAR-10 and ≈51% on CIFAR-100, improving SOTA by 5 % and 9 % respectively and seeing greater gains on the harder dataset. We validate our approach through extensive experiments and provide insights into the interplay between adversarial robustness, and the hierarchical nature of deep representations. We show that simple gradient-based attacks against our model lead to human-interpretable images of the target classes as well as interpretable image changes. As a byproduct, using our multi-resolution prior, we turn pre-trained classifiers and CLIP models into controllable image generators and develop successful transferable attacks on large vision language models.

self-ensembled resnet:

- beats SOTA in robustness by 5%
- increased interpretability
- robustness through architecture, just not adversarial training


consensus algorithm (auction based)
linear probes
CrossMax top-k ensembling
Stochastic multi-resolution expansion
natural training
Image to classify
pretrained resnet
Self-ensemble
Backbone alone
Standard classifier backbone
multires, noise, jitter, shuffle, light fgsm
(we found out: work best when combined)

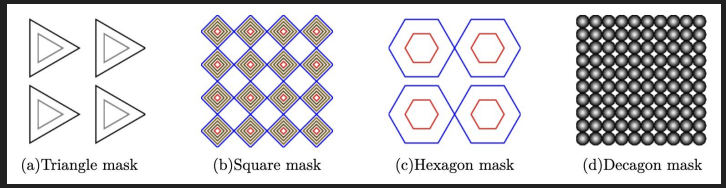
steps

1. choose dataset → CIFAR10

steps

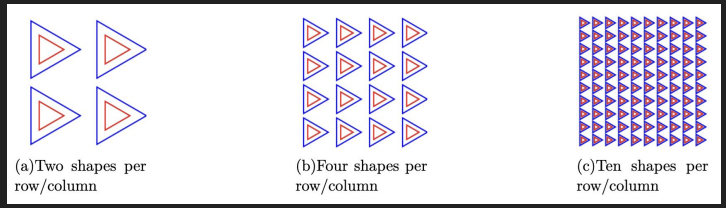
1. choose dataset
2. choose attacks

- 3x geometric masks
- fgsm attack
- pgd attack



96 masks
5 variables

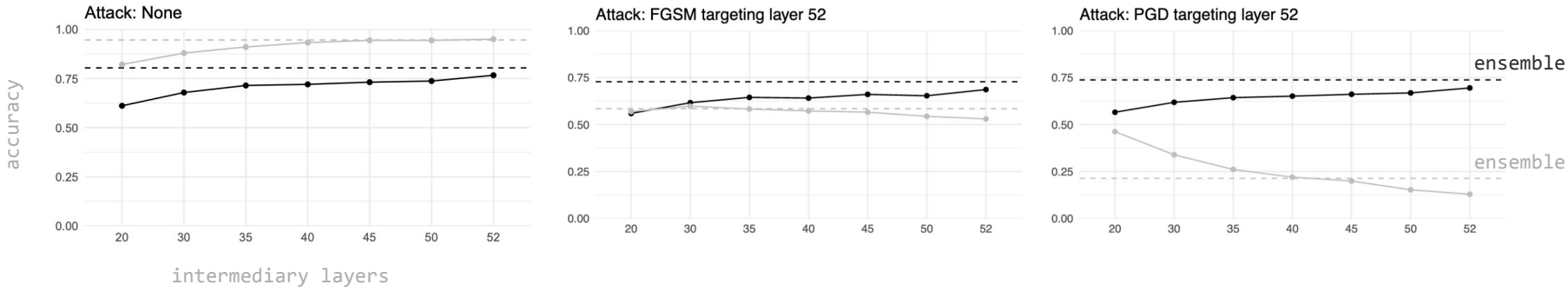
eval against
self-ensemble



3 masks
2 variables (num row/col + opacity)

steps

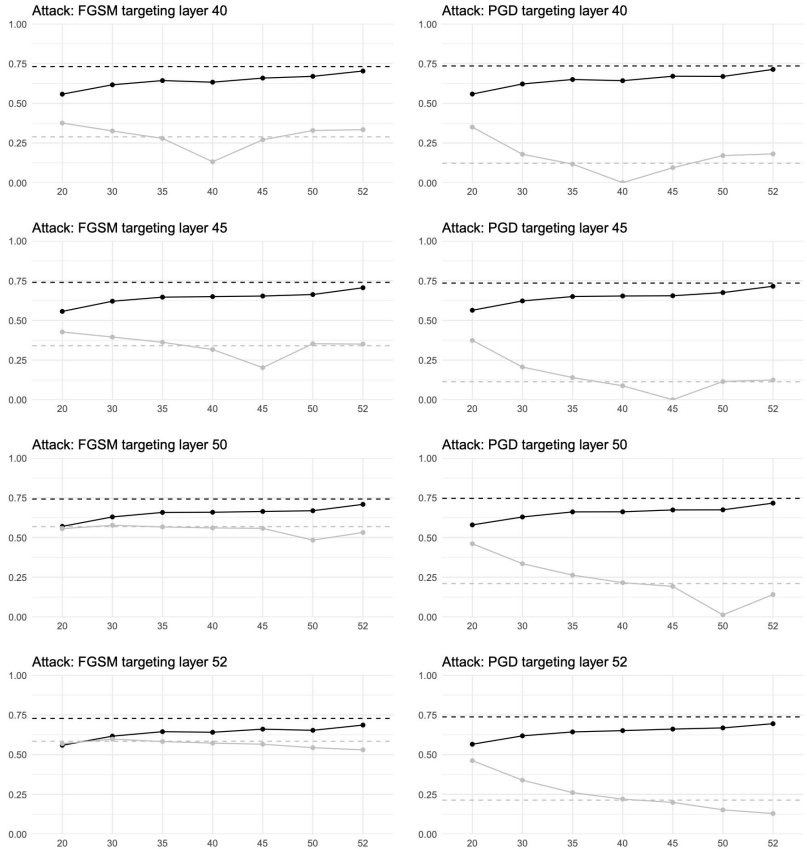
1. choose dataset
2. choose attacks
3. eval



black = natural-training
gray = no natural-training

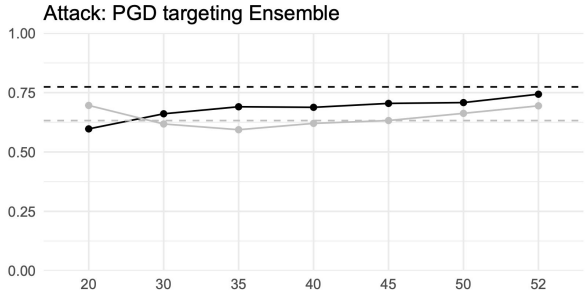
findings:

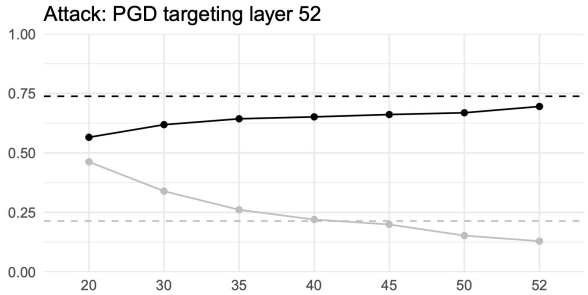
- natural-training > just the architecture
- natural-training worsens accuracy without a threat ("robustness-accuracy-tradeoff")
- PGD > FGSM

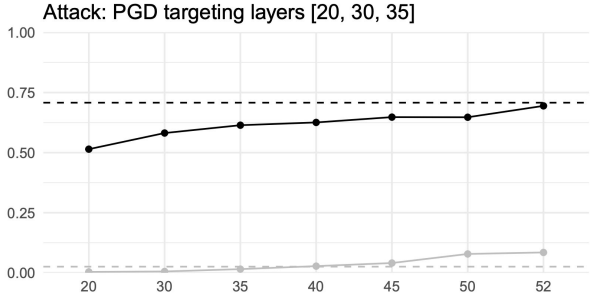


findings:

- FGSM loses strength towards final layers
- PGD does not (as much)





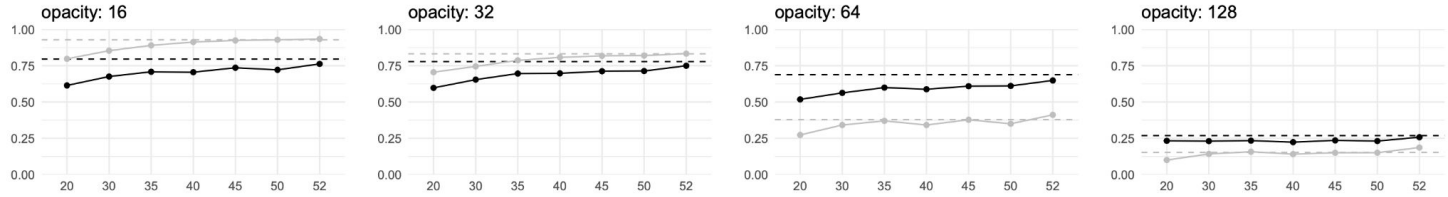


findings:

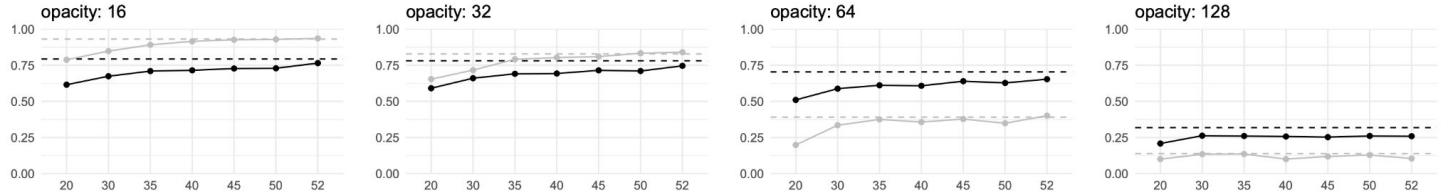
- targeting mid layers > targeting final layer > targeting ensemble

Attack: Geometric Mask

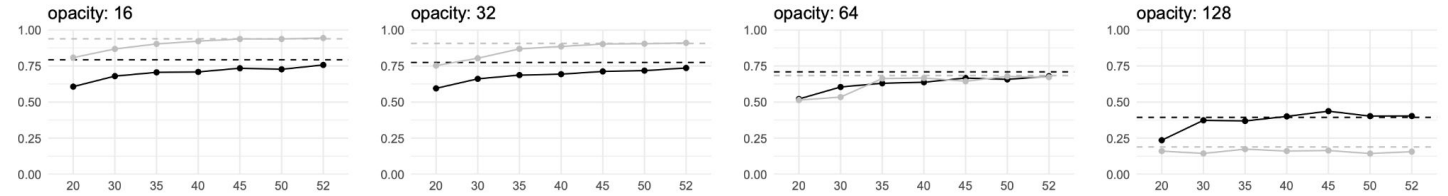
(3 sides, 2 per row/col, 2 concentric shapes, colors enabled)



(3 sides, 4 per row/col, 2 concentric shapes, colors enabled)



(3 sides, 10 per row/col, 2 concentric shapes, colors enabled)



findings:

- opacity ≈ attack strength
- masks target all layers at once
- natural-training is less effective for masks (than FGSM / PGD)

what did we learn?

key takeaways

geometric masks:

- good: simple, allows formal reasoning
- bad: not "natural"

self-ensemble:

- good: robustness through architecture
- bad: but still needs natural training for SOTA

what could be next?

outlook

attacks: natural attacks

- generative models → synthetic data
- NeRF models → 3D data

defenses: robust architectures

- fermi-bose machine architecture
- depth estimation / NeRF model (data augmentation)
- map latent shifts to concept drifts (mechanistic interpretability)
- mutual information / MINE (entropy for intermediary layers)

thank you!

github.com/sueszli/thesis