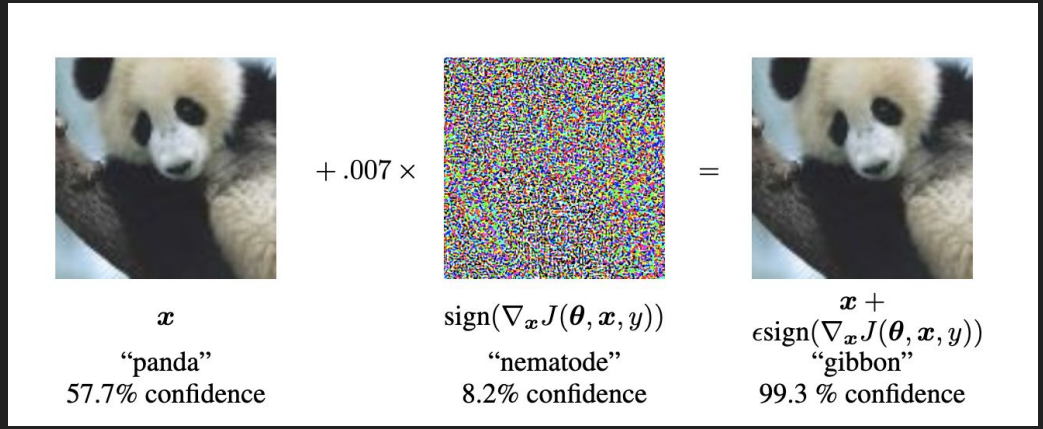
midterm progress update: counterintuitive properties of advx

disco group, ethz - prof. wattenhofer

a traditional introduction

adversarial examples (in computer vision):

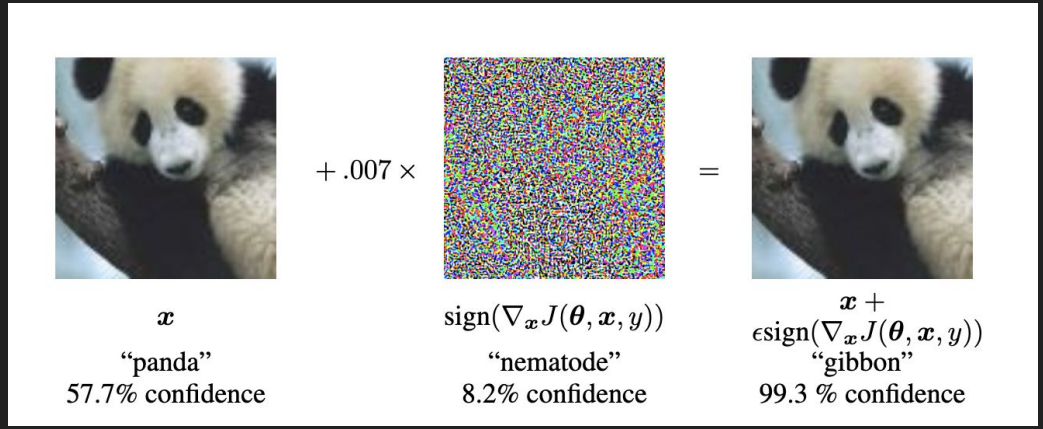
X	input image	
P	perturbation	
$X + P \cdot \varepsilon$	perturbed input	no semantic loss, recognizable to humans
$f\colon X \to y$	classifier	trying to fit human perception
$\mathcal{h}\colon X \to y$	human judgement	implicit
$f(X + P) = t$	misclassification	target based on perturbation



goodfellow et al. (ICLR 2015)
https://arxiv.org/pdf/1412.6572

adversarial examples (in computer vision):

X	input image	
P	perturbation	
$X + P \cdot \varepsilon$	perturbed input	no semantic loss, recognizable to humans
$f\colon X \to y$	classifier	trying to fit human perception
$h\colon X \to y$	human judgement	implicit
$f(X + P) = t$	misclassification	target based on perturbation



can you tell the difference?

if not, then that's a bad sign.

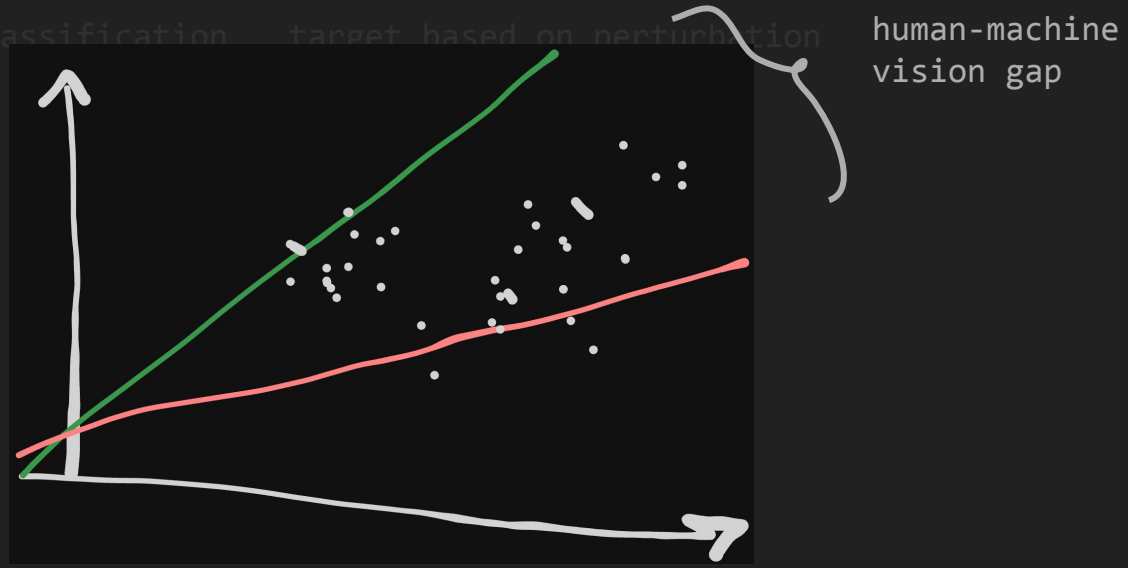
goodfellow et al. (ICLR 2015)
https://arxiv.org/pdf/1412.6572

adversarial examples (in computer vision):

X	input image	
P	perturbation	
$X + P \cdot \varepsilon$	perturbed input	no semantic loss, recognizable to humans

$f\colon X \to y$	classifier	trying to fit human perception
$h\colon X \to y$	human judgement	implicit

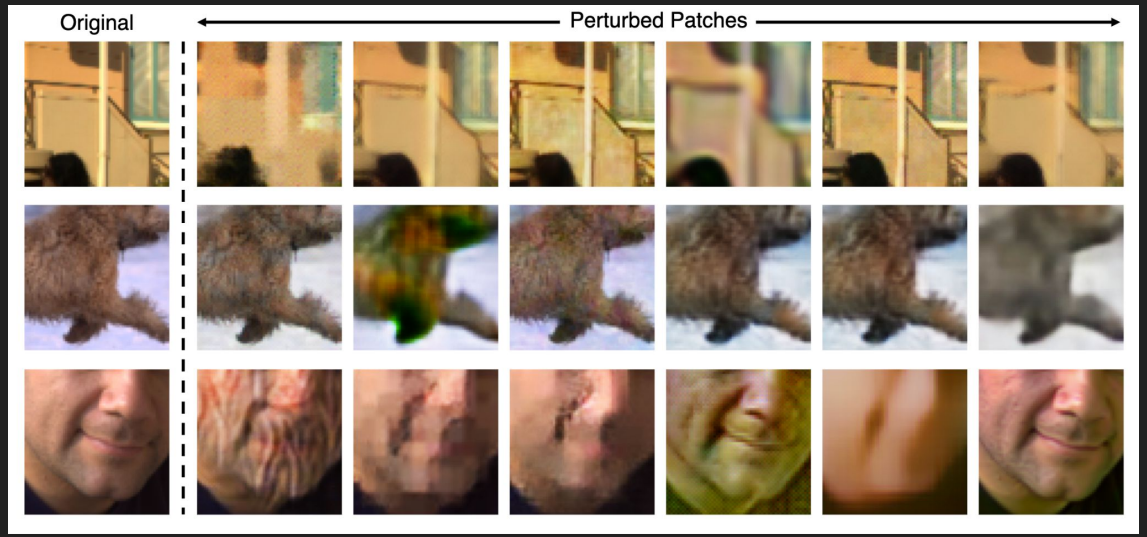
$f(X + P) = t$ misclassification target based on perturbation



adversarial examples (in computer vision):

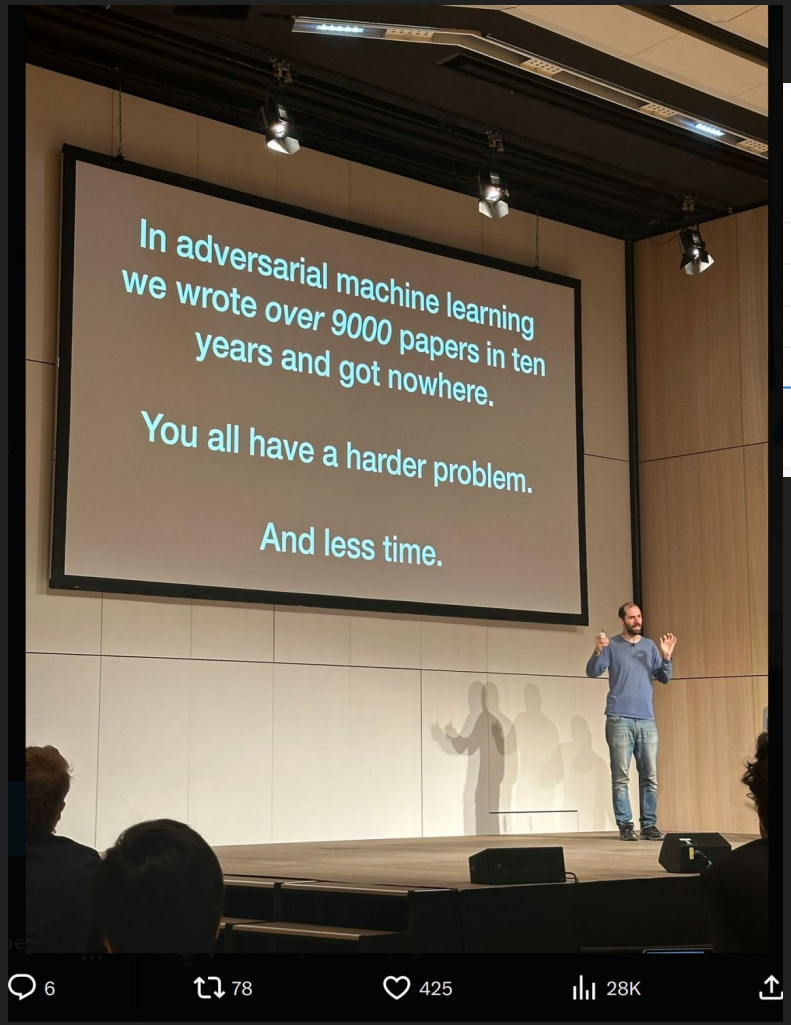
X	input image	
P	perturbation	
$X + P \cdot \varepsilon$	perturbed input	no semantic loss, recognizable to humans
$f\colon X \to y$	classifier	trying to fit human perception
$h\colon X \to y$	human judgement	implicit

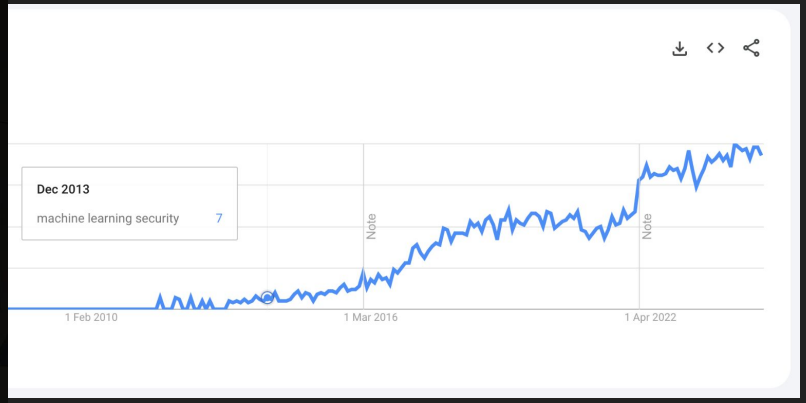
$f(X + P) = h(X + P) = h(X)$



semantic similarity metrics like (LPIPS)

unsolved mysteries





arXiv > cs > arXiv:1312.6199

Computer Science > Computer Vision and Pattern Recognition

[Submitted on 21 Dec 2013 (v1), last revised 19 Feb 2014 (this version, v4)]

Intriguing properties of neural networks

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitr
Goodfellow, Rob Fergus

adversarial examples: unsolved mysteries

- what are they?
- why are they so close to the image?
- why don’t they resemble the target class?
- why do adversarial robustness and accuracy trade-off?
- ...

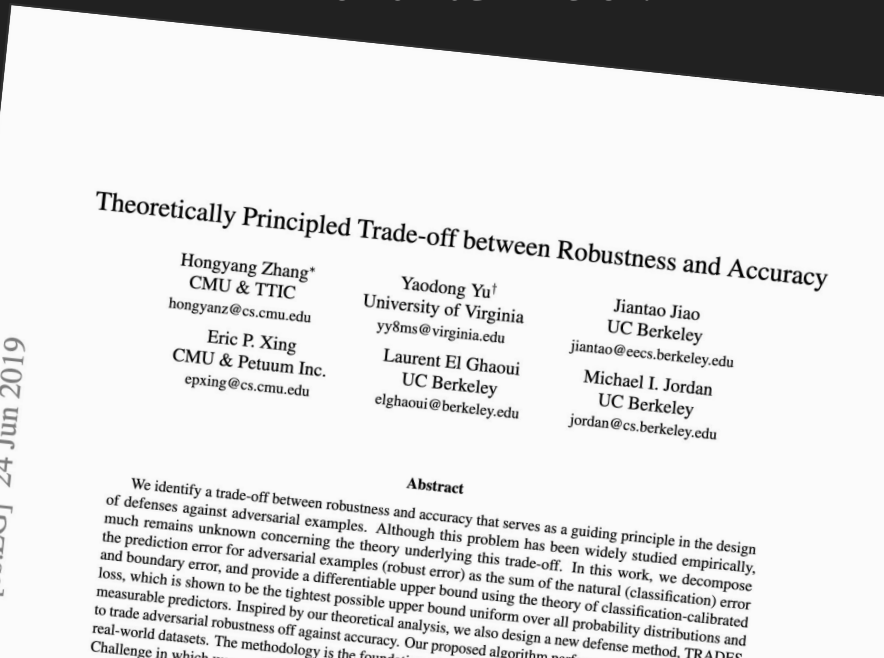
adversarial examples: unsolved mysteries

- what are they?
- why are they so close to the image?
- why don't they resemble the target class?
- why do adversarial robustness and accuracy trade-off?
- ...

informal definition:

not being fooled by the perturbations

formal definition:



Theoretically Principled Trade-off between Robustness and Accuracy

Hongyang Zhang*
CMU & TTIC
hongyanz@cs.cmu.edu

Yaodong Yu†
University of Virginia
yy8ms@virginia.edu

Jiantao Jiao
UC Berkeley
jiantao@eecs.berkeley.edu

Eric P. Xing
CMU & Petuum Inc.
epxing@cs.cmu.edu

Laurent El Ghaoui
UC Berkeley
elghaoui@berkeley.edu

Michael I. Jordan
UC Berkeley
jordan@cs.berkeley.edu

Abstract

We identify a trade-off between robustness and accuracy that serves as a guiding principle in the design of defenses against adversarial examples. Although this problem has been widely studied empirically, much remains unknown concerning the theory underlying this trade-off. In this work, we decompose the prediction error for adversarial examples (robust error) as the sum of the natural (classification) error and boundary error, and provide a differentiable upper bound using the theory of classification-calibrated loss, which is shown to be the tightest possible upper bound uniform over all probability distributions and measurable predictors. Inspired by our theoretical analysis, we also design a new defense method, TRADES, to trade adversarial robustness off against accuracy. Our proposed algorithm performed well in real-world datasets. The methodology is the foundation of our entry to the NeurIPS 2018 Adversarial Vision Challenge in which

3v3 [cs.LG] 24 Jun 2019

adversarial examples: unsolved mysteries

- what are they?
- why are they so close to the image?
- why don’t they resemble the target class?
- why do adversarial robustness and accuracy trade-off?

adi shamir: “Any attempt to robustify a network by limiting all its directional derivatives will make it harder to train and thus less accurate.”

shamir et al.
https://arxiv.org/pdf/2106.10151

adversarial examples: unsolved mysteries

- what are they?
- why are they so close to the image?
- why don't they resemble the target class?
- why do adversarial robustness and accuracy trade-off?

adi shamir: "Any attempt to robustify a network by limiting all its directional derivatives will make it harder to train and thus less accurate."

madry's lab:

a) useless features	not picked up by feature extractor
b) robust features	comprehensible to humans, stable under perturbations
c) non-robust features	incomprehensible to humans very predictive but also very brittle derived from data distribution patterns

ilyas et al.
https://arxiv.org/pdf/1905.02175

adversarial examples: unsolved mysteries

- what are they?
- why are they so close to the image?
- why don't they resemble the target class?
- why do adversarial robustness and accuracy trade-off?

adi shamir: "Any attempt to robustify a network by limiting all its directional derivatives will make it harder to train and thus less accurate."

madry's lab:

a) useless features — not picked up by feature extractor
b) robust features — comprehensible to humans, stable under perturbations
c) non-robust features — incomprehensible to humans; very predictive but also very brittle; derived from data distribution patterns

predictive accuracy without generalizability is useless for for downstream tasks.

there is a solution!

Robustness and Accuracy Could Be Reconcilable by (Proper) Definition

Tianyu Pang [1 2] Min Lin [2] Xiao Yang [1] Jun Zhu [1] Shuicheng Yan [2]

Abstract

The trade-off between robustness and accuracy has been widely studied in the adversarial literature. Although still controversial, the prevailing view is that this trade-off is inherent, either empirically or theoretically. Thus, we dig for the origin of this trade-off in adversarial training and find that it may stem from the improperly defined robust error, which imposes an inductive bias of local invariance — an overcorrection towards smoothness. Given this, we advocate employing local equivariance to describe the ideal behavior of a robust model, leading to a self-consistent robust error named SCORE. By definition, SCORE facilitates the reconciliation between robustness and accuracy, while still handling the worst-case uncertainty via robust optimization. By simply substituting KL divergence with variants of distance metrics, SCORE can be efficiently minimized. Empirically, our models achieve top-rank performance on RobustBench under AutoAttack. Besides, SCORE provides instructive insights for explaining the overfitting phenomenon and semantic input gradients observed on robust models.

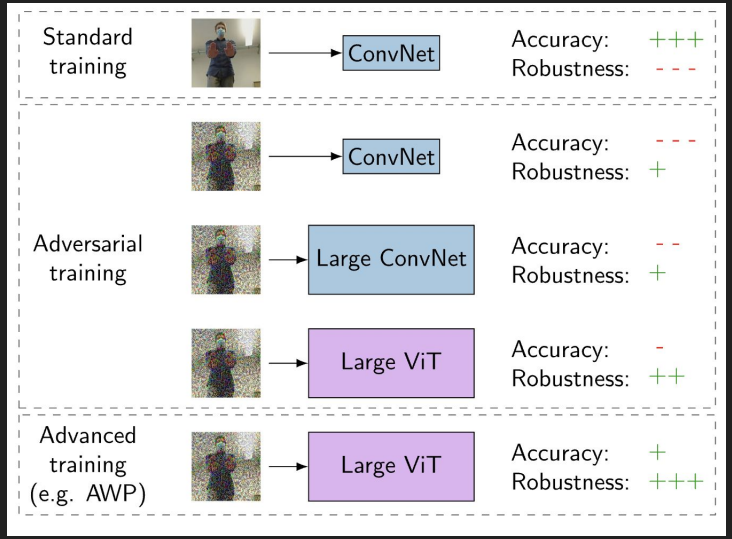
1. Introduction

The trade-off between *adversarial* robustness and *clean* accuracy has been widely observed (Schmidt et al., 2018; Su et al., 2018; Zhang et al., 2019; Wang et al., 2020a;b). On some simple cases, this trade-off is even shown to provably exist (Tsipras et al., 2019; Nakkiran, 2019; Raghunathan et al., 2020; Javanmard et al., 2020; Yu et al., 2021). With that, many strategies have been proposed to alleviate this trade-off via, e.g., early-stopping (Rice et al., 2020; Zhang et al., 2020), instance reweighting (Balaji et al., 2019; Zhang et al., 2021a), and exploiting extra data (Alayrac et al., 2019; Nevertheless, other findings show the opposite. Stutz et al. (2019) and Yang et al. (2020) argue that robustness and accuracy can both be achievable through manifold analyses or locally Lipschitz functions. Rozsa et al. (2016) and Gilmer et al. (2018) also reveal that better generalization helps robustness on both toy and large-scale datasets.

Given arguments on both sides, it is much debated whether the robustness-accuracy trade-off is intrinsically there. But before considering its existence, let us first reach a consensus on how a robust model is supposed to behave — the definition of robustness. A most popular one is made by Madry et al. (2018), who define robustness in terms of robust error formulated as a locally maximized loss function. This definition is widely adopted in adversarial training (AT) (Shafahi et al., 2019; Wong et al., 2020). Besides, there are also some other definitions. For example, Szegedy et al. (2014) define robustness as the minimal perturbation required to flip the predicted labels, which is applied in several adversarial attack studies (Moosavi-Dezfooli et al., 2016; Carlini and Wagner, 2017).

We then try to describe how the robustness-accuracy trade-off stems from the previously defined robust error. Recall that in supervised learning, we have a joint data distribution $p_d(x, y)$, and a discriminative model $p_\theta(y|x)$ for the label y, which is conditional on the input x. In standard settings, we obtain an accurate model via minimizing the expected KL divergence between $p_d(y|x)$ and $p_\theta(y|x)$, i.e., the **standard error** $\mathbf{R}_{\text{Standard}}(\theta)$ (see Eq. (1)) w.r.t. the parameters θ. The optimal solution θ^* satisfies $p_{\theta^*}(y|x) = p_d(y|x)$.

In adversarial settings, Madry et al. (2018) propose to minimize the **robust error** $\mathbf{R}_{\text{Madry}}(\theta)$ (see Eq. (3)) for training reliable models. Compared to the standard error, the robust error contains an inner maximization problem, finding the point $x' \in B(x)$ that maximizes the KL divergence between $p_d(y|x)$ and $p_\theta(y|x')$, where $B(x)$ is a set of allowed points



Lechner, Mathias, et al. "Revisiting the adversarial robustness-accuracy tradeoff in robot learning." IEEE Robotics and Automation Letters 8.3 (2023): 1595-1602.

in conclusion:

adversarial examples are about generalizability, not security.

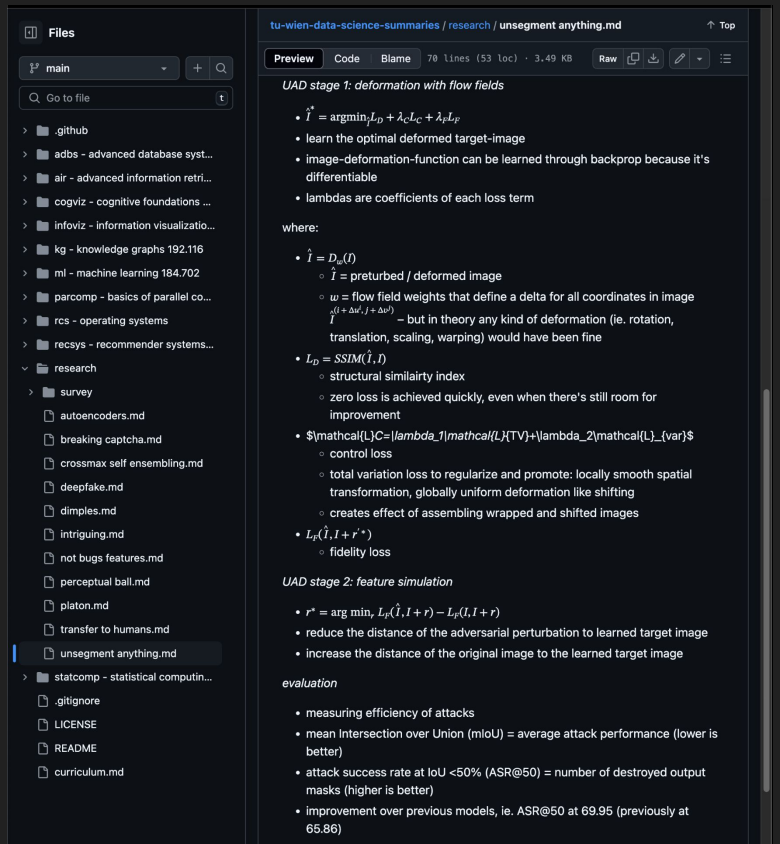
they help us:

- to augment data such that models learn the right features
- understand the learning process & human-machine vision gap

my progress

steps:

0. literature review

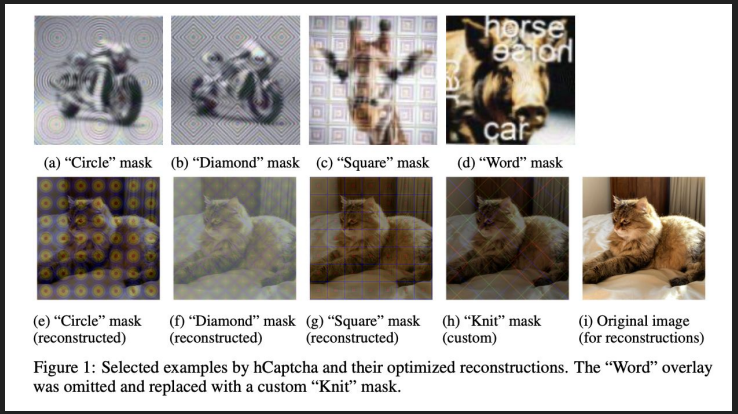


steps:

0. literature review
1. **getting inspired by hCaptcha perturbations**

- building an offline CAPTCHA clone
- rendering hCAPTCHA masks

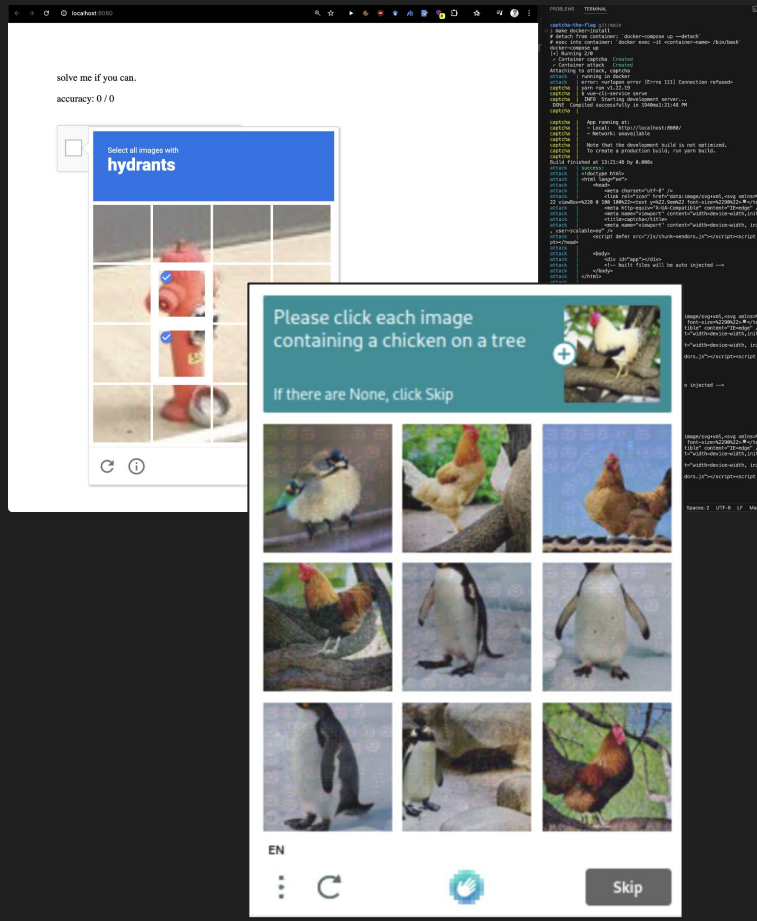
  blackbox, not model specific
  just manipulates contours and colors



(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)

Figure 1: Selected examples by hCaptcha and their optimized reconstructions. The “Word” overlay was omitted and replaced with a custom “Knit” mask.



steps:

0. literature review
1. getting inspired by hCaptcha perturbations
2. **selecting the best performing models**

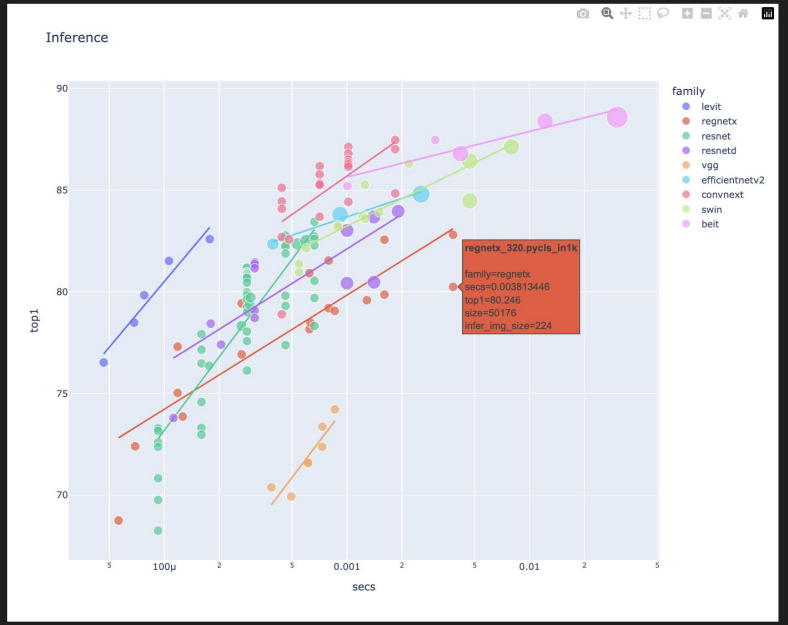
- comparing models by family
- comparing zero shot benchmark leaderboards
    - *ie. robustbench*
- fine tuning the best performing model

pytorch_benchmark.ipynb

analysis > model-selection > pytorch_benchmark.ipynb > tdf = get_data('train', 'train_samples_per_sec')

Code | Markdown | Run All | Clear All Outputs | Outline

```
top1 accuracy by family:
- beit (88.60% top1, by beit_large_patch16_512.in22k_ft_in22k_in1k)
- convnext (87.47% top1, by convnext_large.fb_in22k_ft_in1k_384)
- swin (87.13% top1, by swin_large_patch4_window12_384.ms_in22k_ft_in1k)
- efficientnetv2 (84.81% top1, by efficientnetv2_rw_m.agc_in1k)
- resnetd (83.96% top1, by resnet200d.ra2_in1k)
- resnet (83.45% top1, by resnet152.a1h_in1k)
- regnetx (82.81% top1, by regnetx_320.tv2_in1k)
- levit (82.60% top1, by levit_384.fb_dist_in1k)
- vgg (74.22% top1, by vgg19_bn.tv_in1k)
```



sueszli/**robustified_clip_vit** ♡ like 0

Image Classification | License: apache-2.0

Model card | Files and versions | Community | Settings

robustified CLIP ViT-L/14@336px using hcaptcha masks (geometric image overlays) on `visual-layer/imagenet-1k-vl-enriched`.

for source code and example usage see: https://github.com/ETH-DISCO/advx-bench

steps:

0. literature review
1. getting inspired by hCaptcha perturbations
2. selecting the best performing models
3. **designing experiments**

- imagenet + hcaptcha perturbations
- hyperparameter optimization (opacity, density, fgsm epsilon)
- perceptual quality metrics to avoid transfer to humans

$$\text{final score} = \text{quality} - (\text{adv acc rank} - \text{acc rank})$$

Where:

$$\text{quality} = 0.15 \cdot \text{cs} + 0.25 \cdot \text{psnr}_n + 0.35 \cdot \text{ssim} + 0.25 \cdot (1 - \text{lpips}_n)$$

And:

$$\text{psnr}_n = \text{clip}\left(\frac{\text{psnr} - 30}{20}, 0, 1\right)$$

$$\text{lpips}_n = \text{clip}(\text{lpips}, 0, 1)$$

Bartoldson, Brian R., et al. "Adversarial Robustness Limits via Scaling-Law and Human-Alignment Studies." arXiv preprint arXiv:2404.09349 (2024).

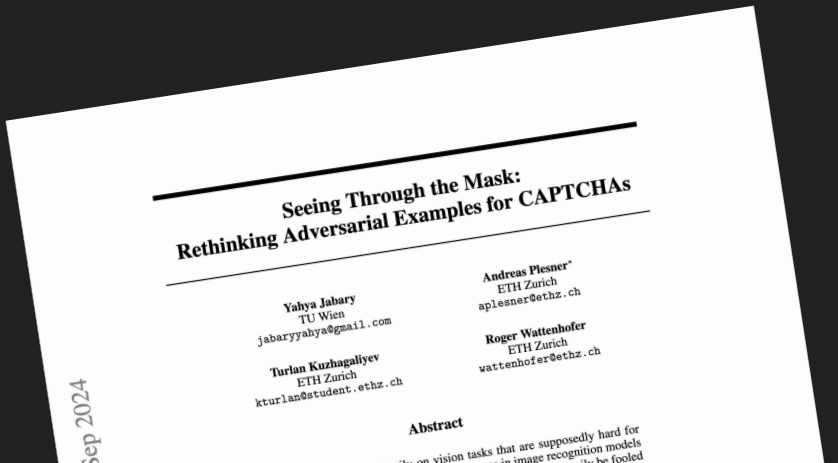
steps:

0. literature review
1. getting inspired by hCaptcha perturbations
2. selecting the best performing models
3. designing experiments
4. **benchmarking**

- 2.0 weeks of cluster debugging & writing docs *(see: @ETH-DISCO/cluster-tutorial)*
- 1.0 weeks of hyperparam tuning
- 0.5 weeks of experimentation

steps:

0. literature review
1. getting inspired by hCaptcha perturbations
2. selecting the best performing models
3. designing experiments
4. benchmarking
5. **submitting results as workshop paper**



Seeing Through the Mask:
Rethinking Adversarial Examples for CAPTCHAs

Yahya Jabary
TU Wien
jabaryyahya@gmail.com

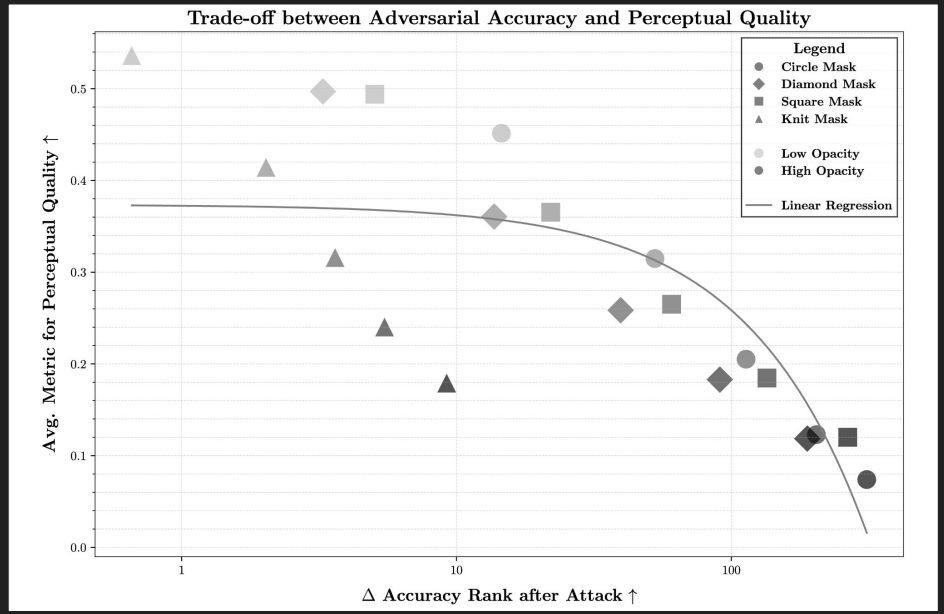
Andreas Plesner*
ETH Zurich
aplesner@ethz.ch

Turlan Kuzhagaliyev
ETH Zurich
kturlan@student.ethz.ch

Roger Wattenhofer
ETH Zurich
wattenhofer@ethz.ch

Abstract

sep 2024

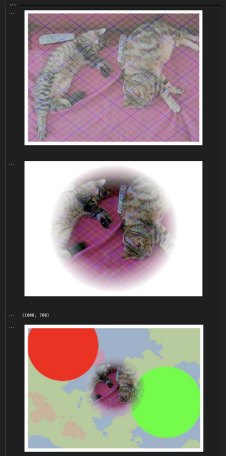


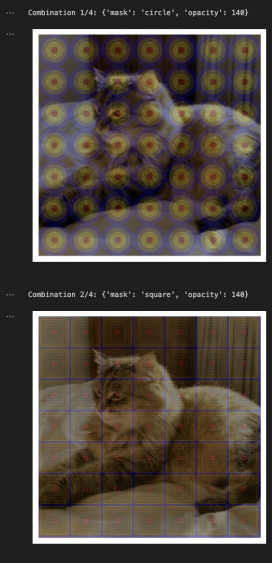
what's next?

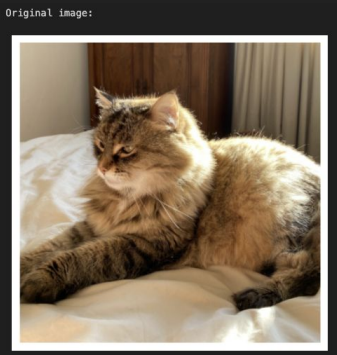
i'd appreciate your advice!

studying geometric distortions of hCaptcha masks

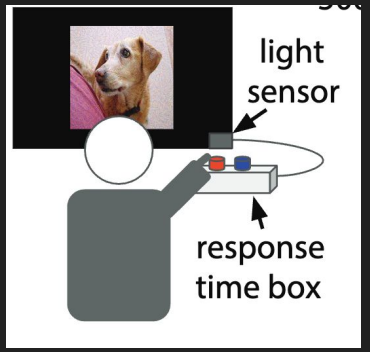
- why were they introduced by this company?
- what makes them so effective?
    can we look into intermediary layers to understand?







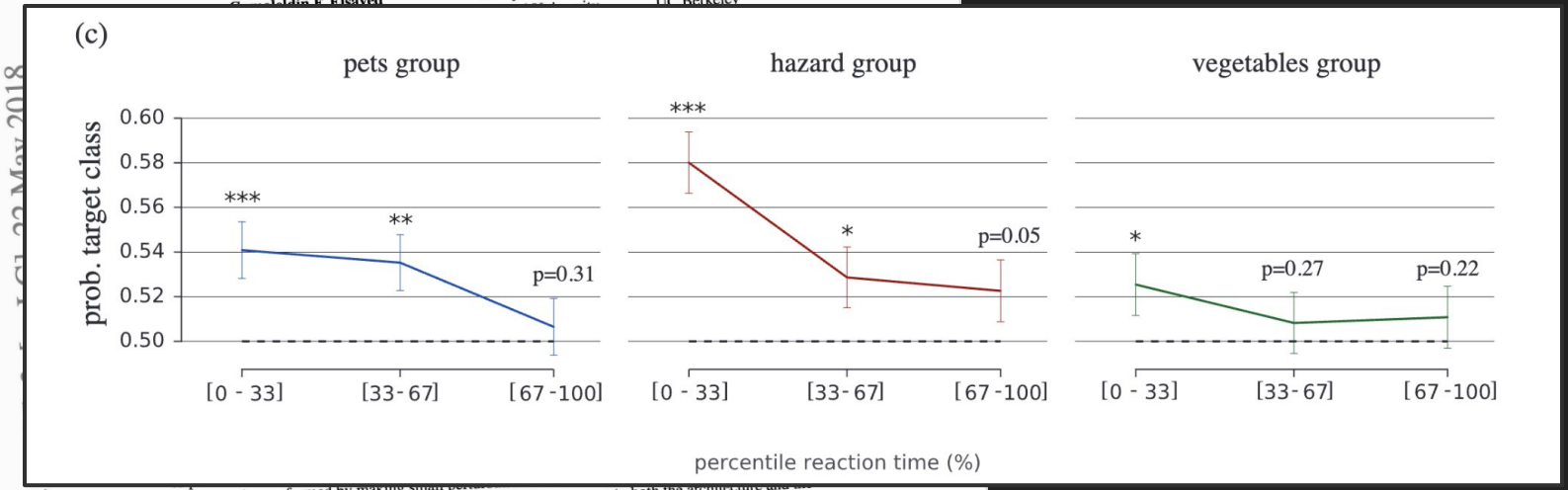
studying survival-optimized features of human vision



Adversarial Examples that Fool both Computer Vision and Time-Limited Humans

Shreya Shankar

Brian Cheung
UC Berkeley



usually an image formed by making small per...
constructing adversarial examples [13, 24, 27, 33, 39] rely on access to both the architecture and the
parameters of the model to perform gradient-based optimization on the input. Without similar access
to ... do not seem applicable to constructing adversarial examples for humans.

self-ensembling

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

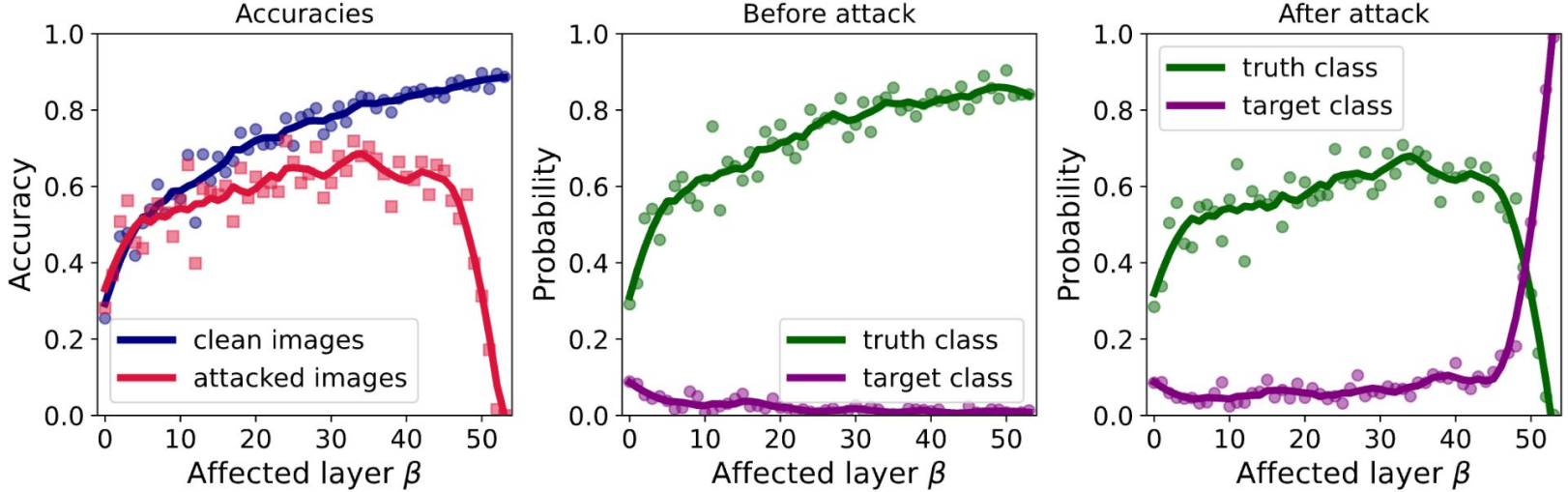


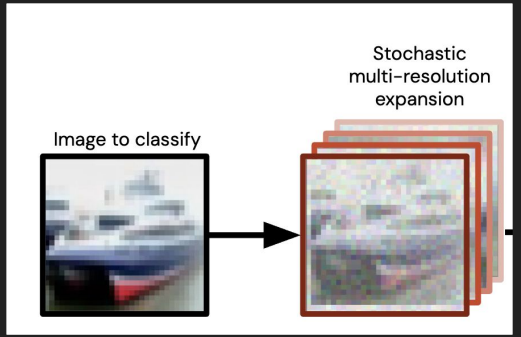
Figure 1 | We use a multi-resolution decomposition (a) of an input image and a partial decorrelation of predictions of intermediate layers (b) to build a classifier (c) that has, by default, adversarial robustness comparable or exceeding state-of-the-art (f), even without any adversarial training. Optimizing inputs against it leads to interpretable changes (d) and images generated from scratch (e).

down sampling

classifier makes a decision about all N versions at once

mimicking nature:

1) random noise
2) a random jitter in both axes *(microsaccades)*
3) a small random change in contrast *(cone cells)*
4) a small random color-grayscale shift *(cone cells)*

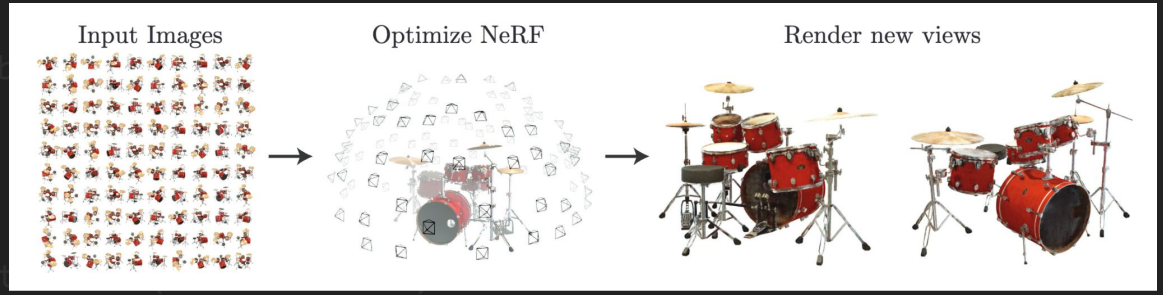


down sampling

classifier makes a decision ab

mimicking nature:

1) random noise
2) a random jitter in bot
3) a small random change in contrast *(cone cells)*
4) a small random color-grayscale shift *(cone cells)*

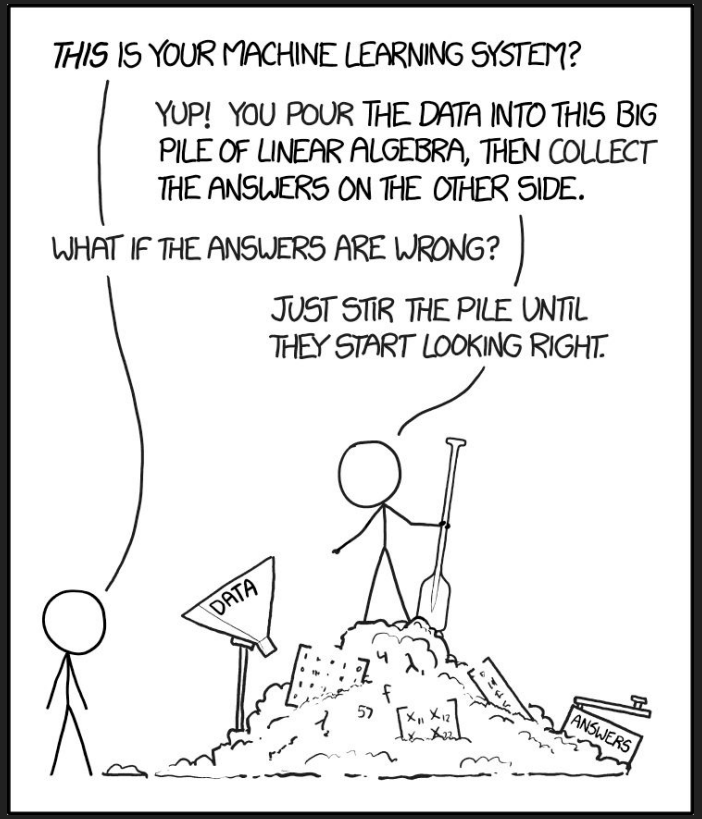


my idea: combining with NeRFs (Neural Radiance Fields)

https://www.matthewtancik.com/nerf

capturing all angles, distances, under various blurs, rotations, illuminations, contrasts while semantic content is preserved.

thank you for listening!



https://xkcd.com/1838/

fun fact: image convolutions can also be expressed algebraically.

thank you for listening!

Q&A / discussion