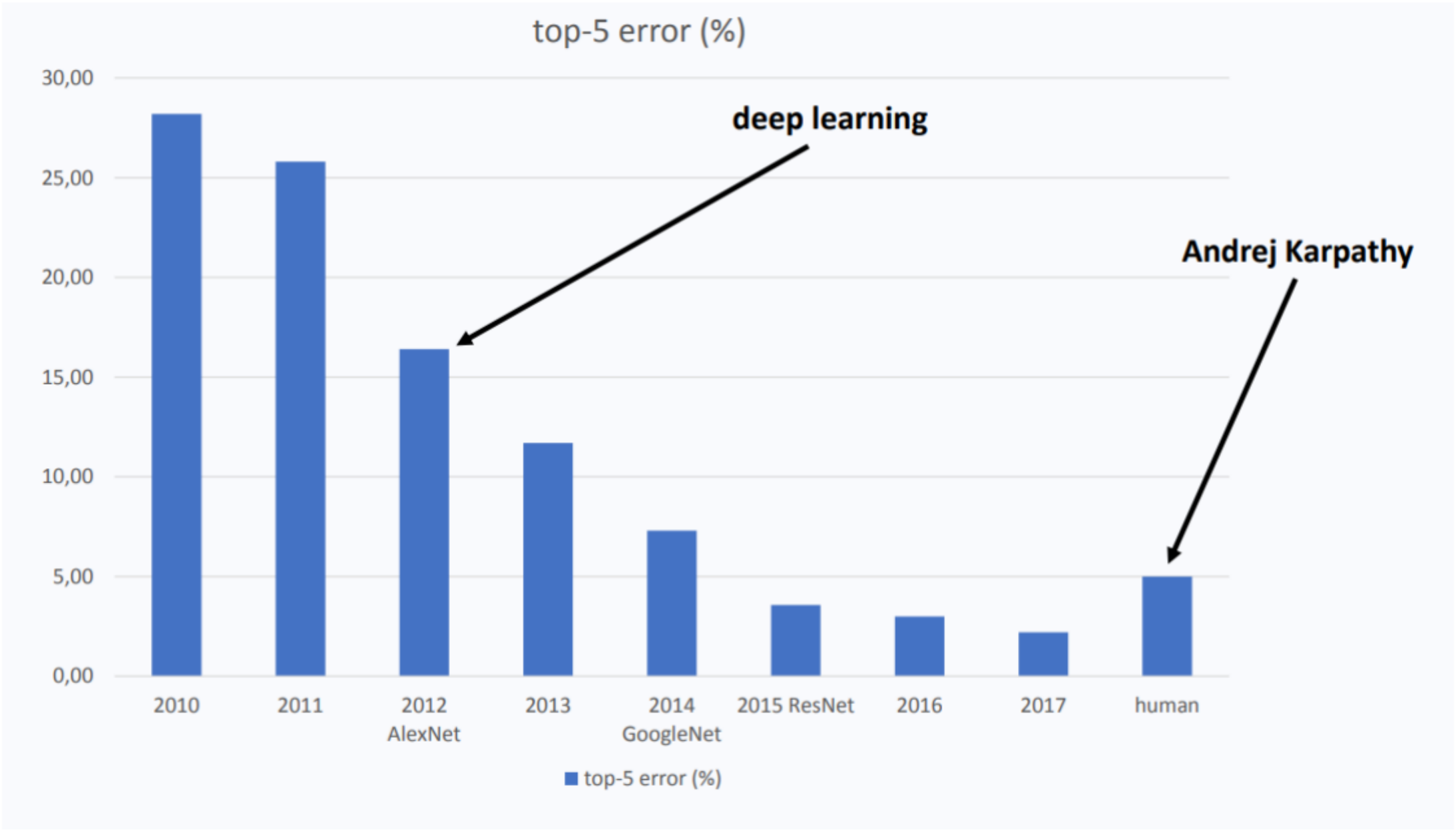
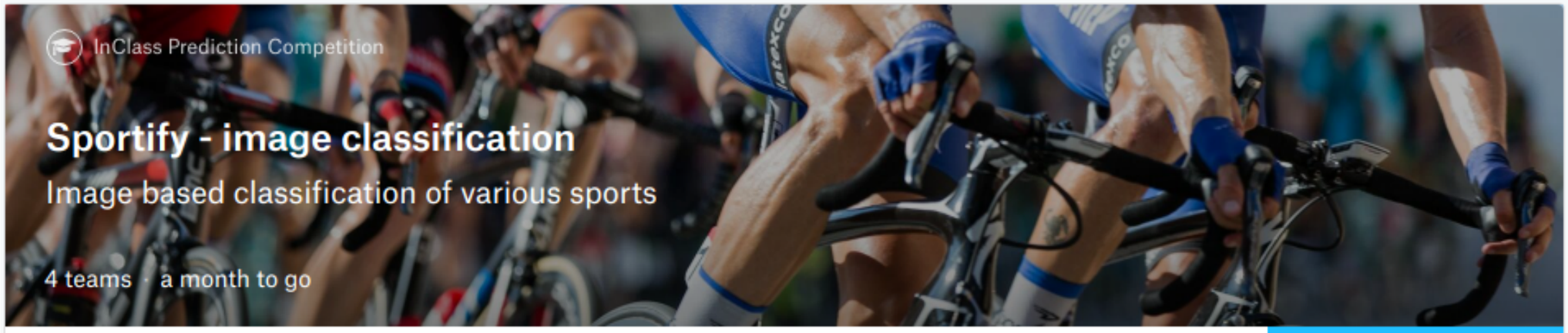
ADVERSARIAL EXAMPLES

DEEPLEA17EM

IMAGENET – IMAGE RECOGNITION WORLD CUP



CNNS ARE EXTREMELY GOOD AT IMAGE RECOGNITION - SO AS STUDENTS :)



Overview Data Kernels Discussion Leaderboard Rules Team Host My Submissions Submit Predictions

Public Leaderboard Private Leaderboard

This leaderboard is calculated with approximately 50% of the test data.
The final results will be based on the other 50%, so the final standings may be different.

Raw Data Refresh

#	Team Name	Kernel	Team Members	Score	Entries	Last
	3_point_baseline.csv			0.97081		
1	**smao**			0.95297	9	1h
2	**Lőrinc**			0.94756	9	3d
3	**Bogye Balázs**			0.94540	10	3d
	2_point_baseline.csv			0.74270		
4	**Emil Novak**			0.67513	1	4d
	1_point_baseline.csv			0.48756		

DEEP LEARNING SINCE THEN

Gatys et al, 2015

nature medicine

Letter | Published: 07 January 2019

Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network

Awni Y. Hannun ✉, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia & Andrew Y. Ng

Nature Medicine **25**, 65–69 (2019) | Download Citation ⬇

- style transfer
- playing Go, chess, games
- self driving cars
- face recognition/ verification
- speech recognition
- cardiology
- mammograms
- skin cancer
- drug discovery

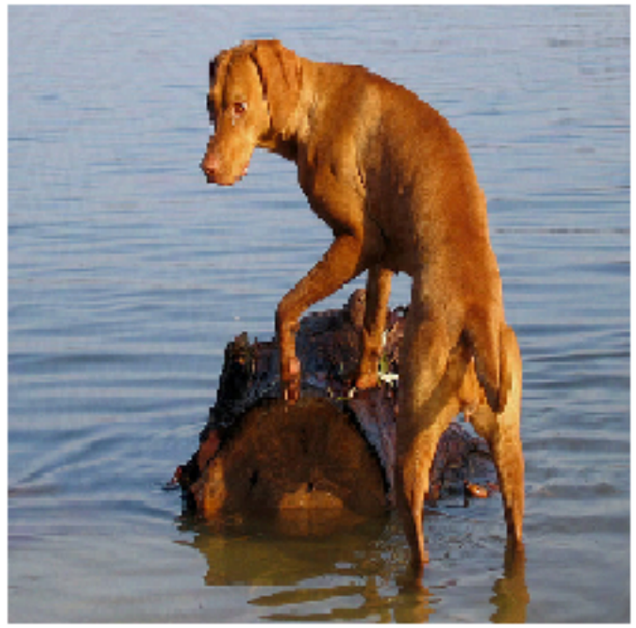
BUT...

what is this?

BUT...

what is this?

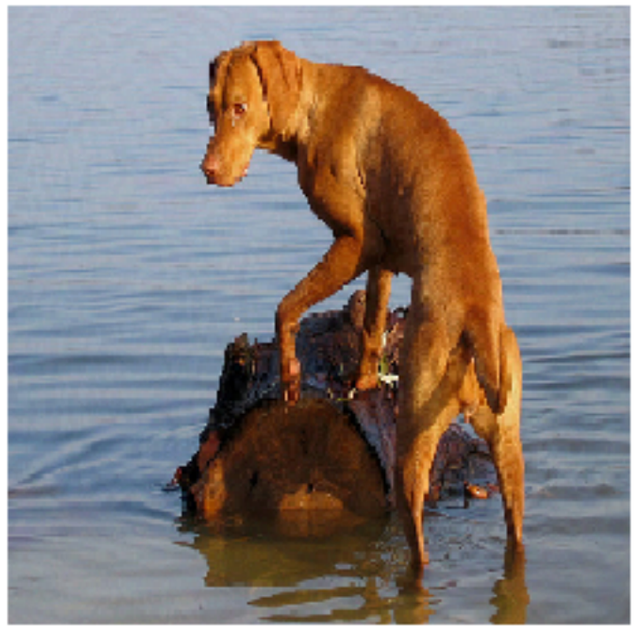
and this?



BUT...

what is this?

and this?



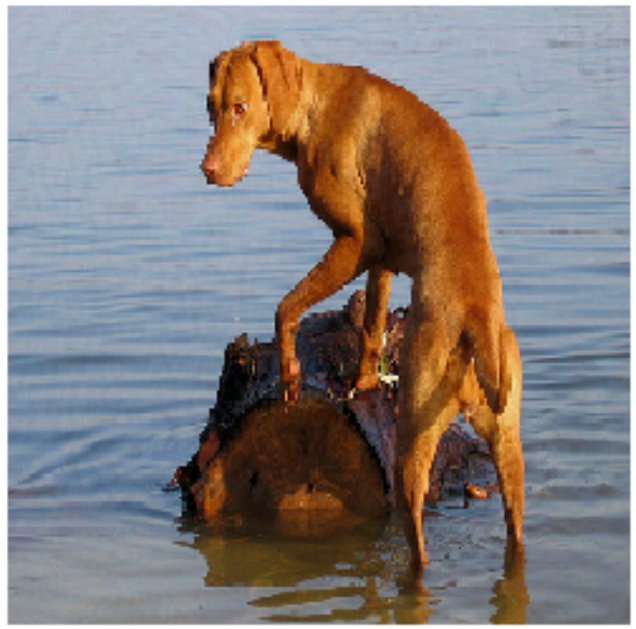
VGG16 says...

vizsla (51.9%)

BUT...

what is this?

and this?



VGG16 says...

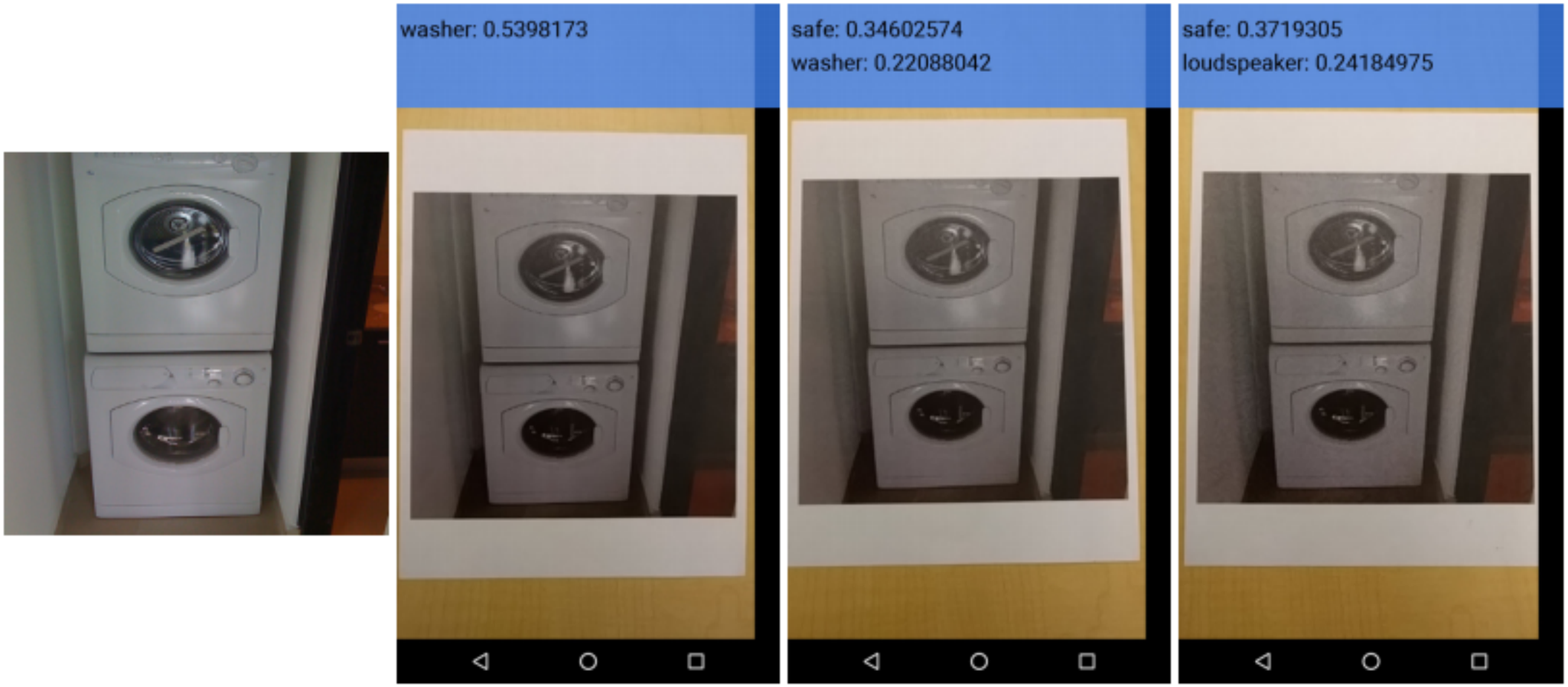
vizsla (51.9%)

brown_bear (35.5%)

brown_bear has the highest probability!

BUT...

- https://www.youtube.com/watch?v=zQ_uMenoBCk&feature=youtu.be

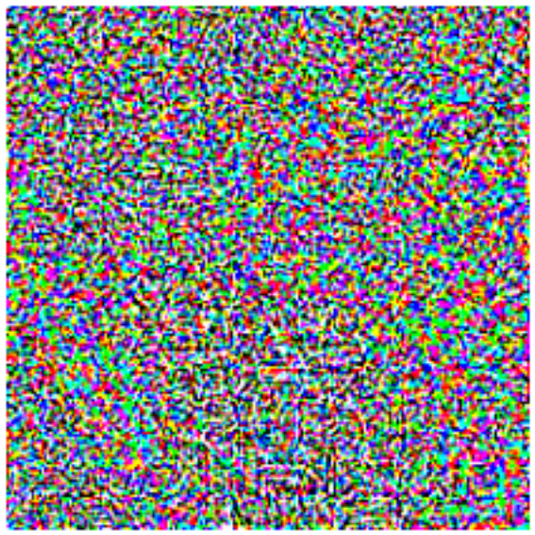


(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

KURAKIN, GOODFELLOW, BENGIO: ADVERSARIAL EXAMPLES IN THE PHYSICAL WORLD, 2017

BUT... SOMETIMES THEY FAIL

$+ .007 \times$



$=$

$\boldsymbol{x}$

“panda”
57.7% confidence

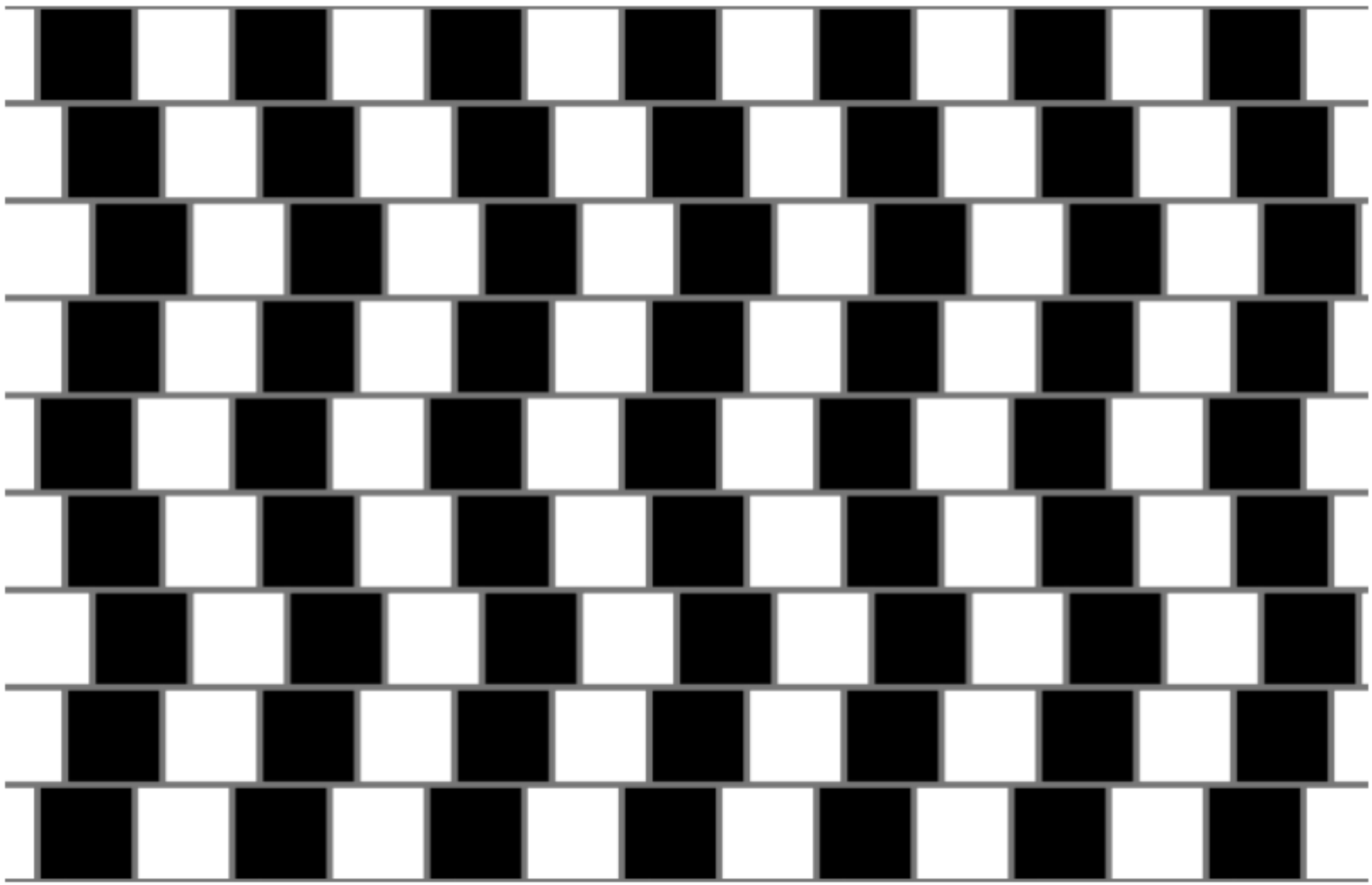
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

“nematode”
8.2% confidence

$\boldsymbol{x} +$
$\epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
“gibbon”
99.3 % confidence

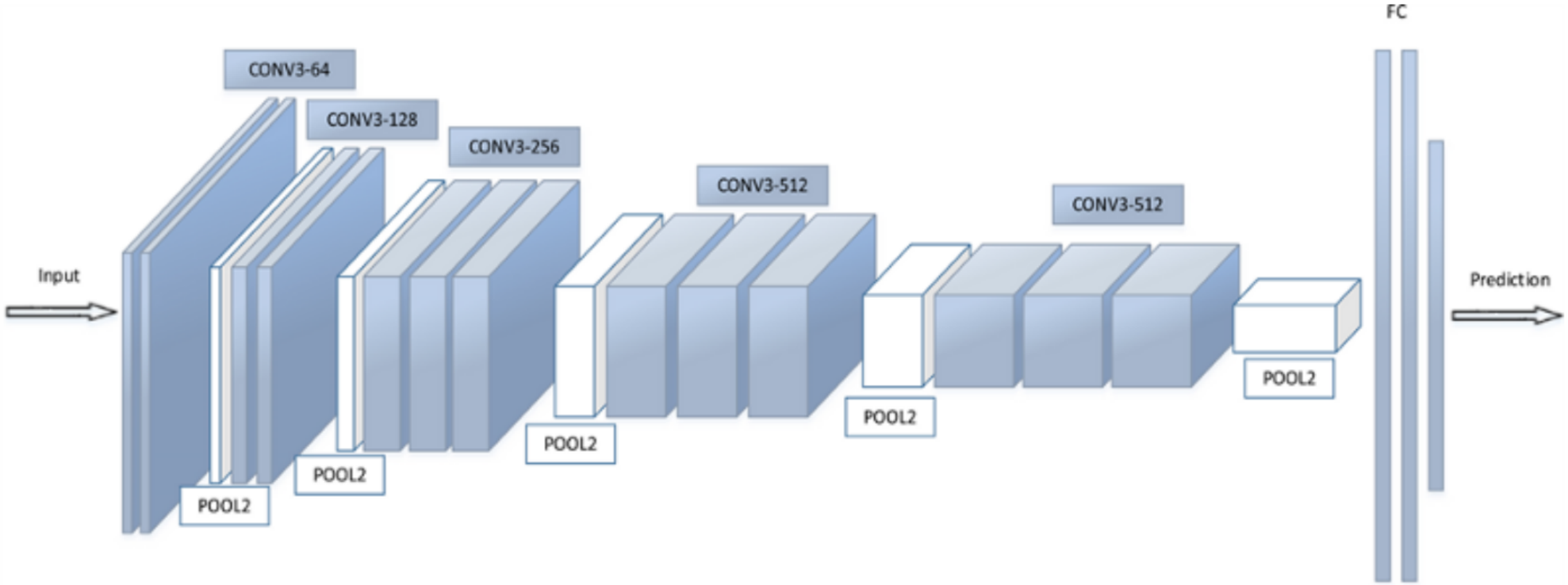
Goodfellow, Shlens, Szegedy: Explaining and Harnessing Adversarial Examples, 2015

SO AS US, HUMANS



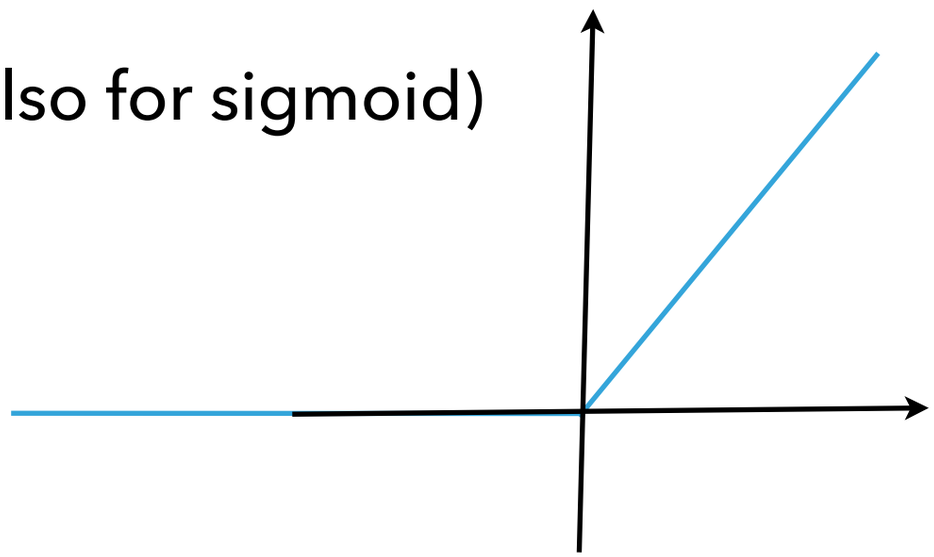
Richard Gregory, cafe wall illusion

ONE WAY TO FOOL CNNS



source: http://file.scirp.org/Html/4-7800353_65406.htm

- non-linearity is needed (else matrix multiplication)
- usually ReLu is used
- whole CNN is still close-to linear at some level (also for sigmoid)
- simple linear method can break



A SIMPLE METHOD TO FOOL A CNN

- have an input image IMG, want to convert it to an arbitrary class, C

```
def pred(img):
    return probability for class C

MASK = zeros(IMG)

for pixel in IMG:
    COPY = IMG, but the given pixel increased by 1e-2
    if pred(COPY) > pred_old:
        MASK at pixel = 1
    elif pred(COPY) < pred_old:
        MASK at pixel = -1
```

- new image = IMG + epsilon * MASK

Notebook DEMO

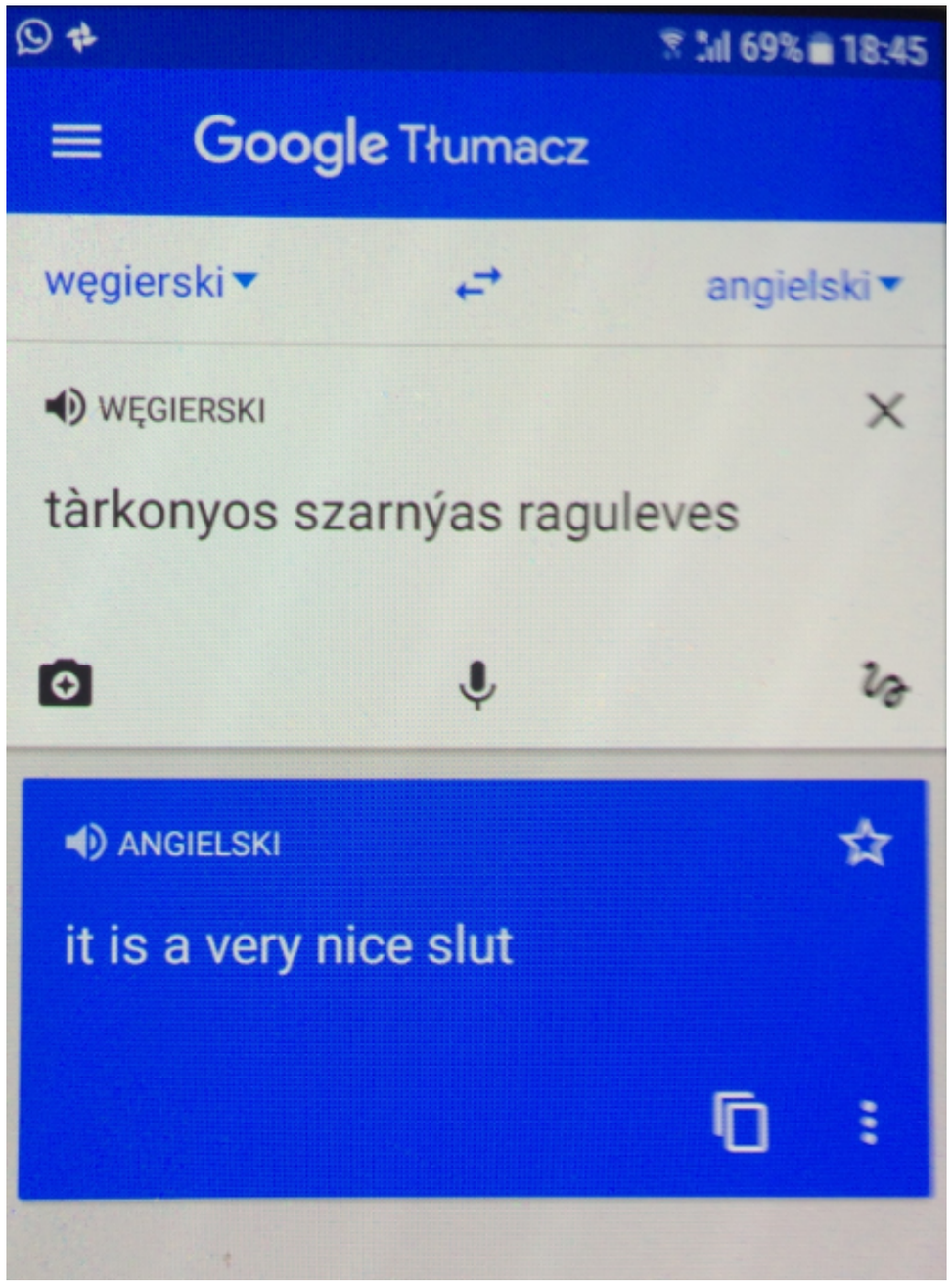
link on the webpage

Adversarial-related blog posts:

- https://blog.openai.com/adversarial-example-research/
- https://blog.openai.com/robust-adversarial-inputs/
- https://www.youtube.com/watch?v=zQ_uMenoBCk&feature=youtu.be

SOMETIMES THEY CAN BREAK WITHOUT TARTGETING

SOMETIMES THEY CAN BREAK WITHOUT TARGETING



Goodfellow et al 2014

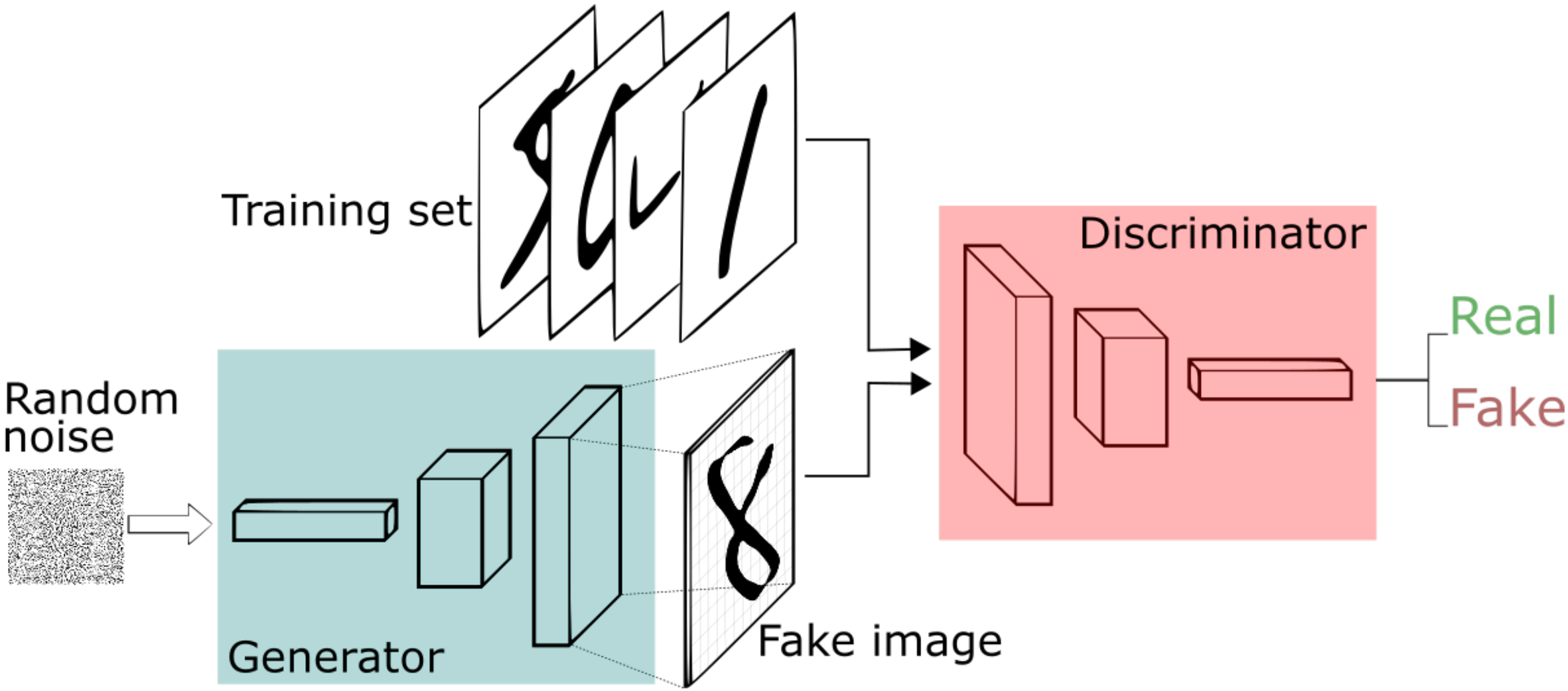


Image credit: Thalles Silva

TO READ MORE

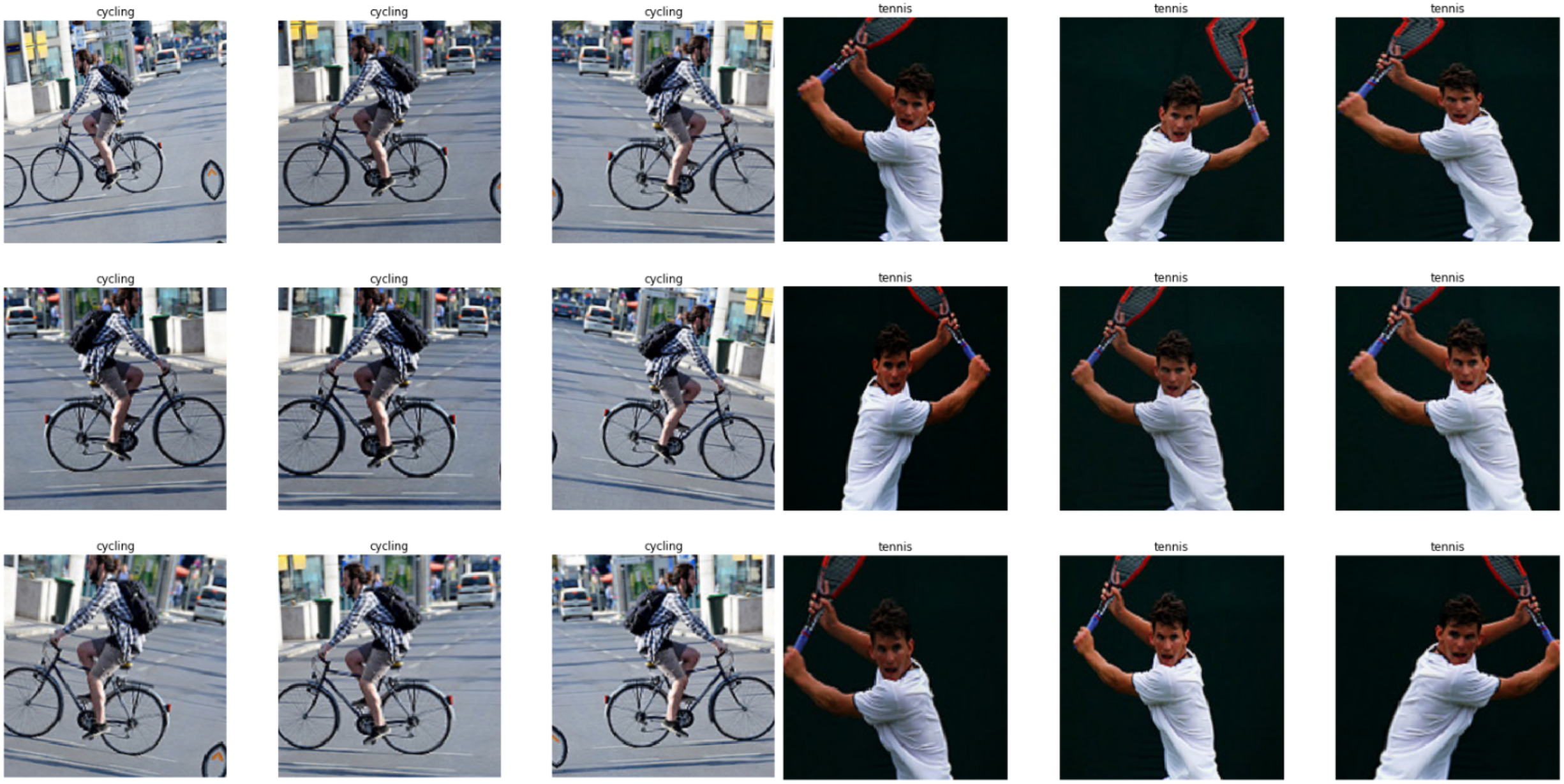
- https://arxiv.org/pdf/1312.6199.pdf Intriguing properties of neural networks
- https://arxiv.org/pdf/1412.6572.pdf EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES
- https://arxiv.org/pdf/1406.2661.pdf GAN
- https://junyanz.github.io/CycleGAN/ CycleGAN

TIPS FOR THE CHALLENGE - DATA AUGMENTATION

- **to get more images for 'free'**
- **realistic transformations on the training images**
  - **in training time the model can see much more than 400 images for each class**
- **transformations can be:**
  - **horizontal flip**
  - **~~vertical flip~~ (maybe for satellite images)**
  - **rotation + pad (zero or mirror)**
  - **zooming + pad (zero or mirror)**
  - **crops**
  - **swarping**
  - **brightness change**
- **test time augmentation (TTA)**
  - **may increase your score as well**

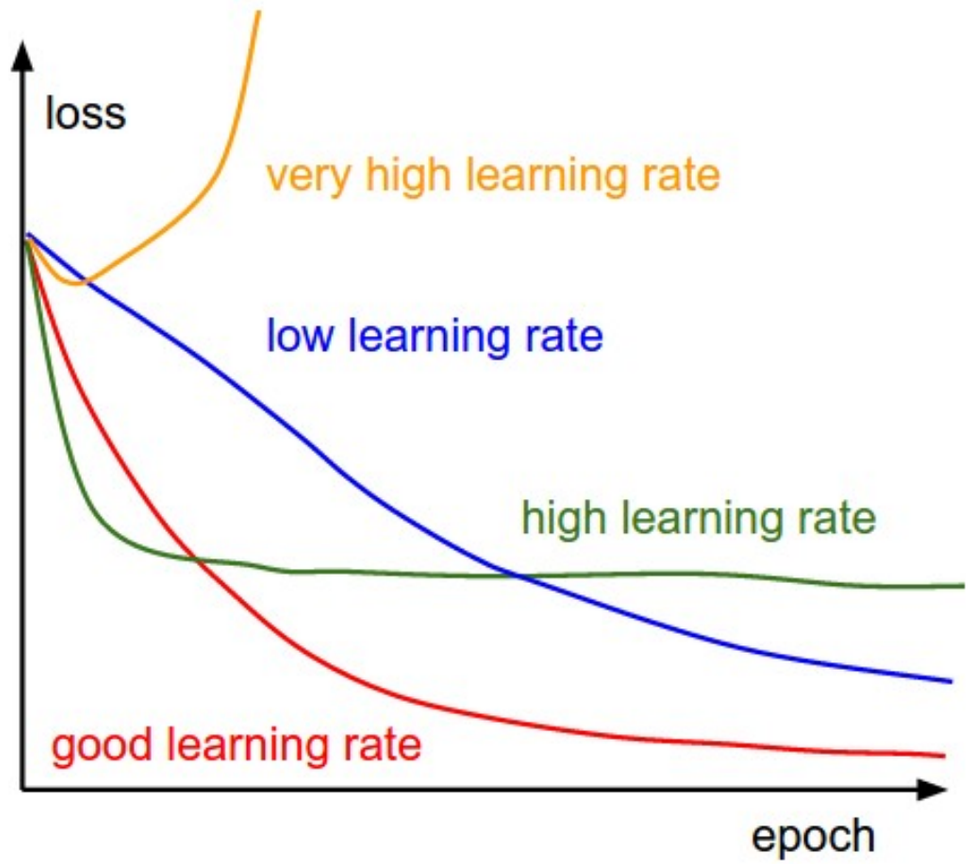
Most of the mentioned methods are implemented in all high level DL libraries

TIPS FOR THE CHALLENGE - DATA AUGMENTATION

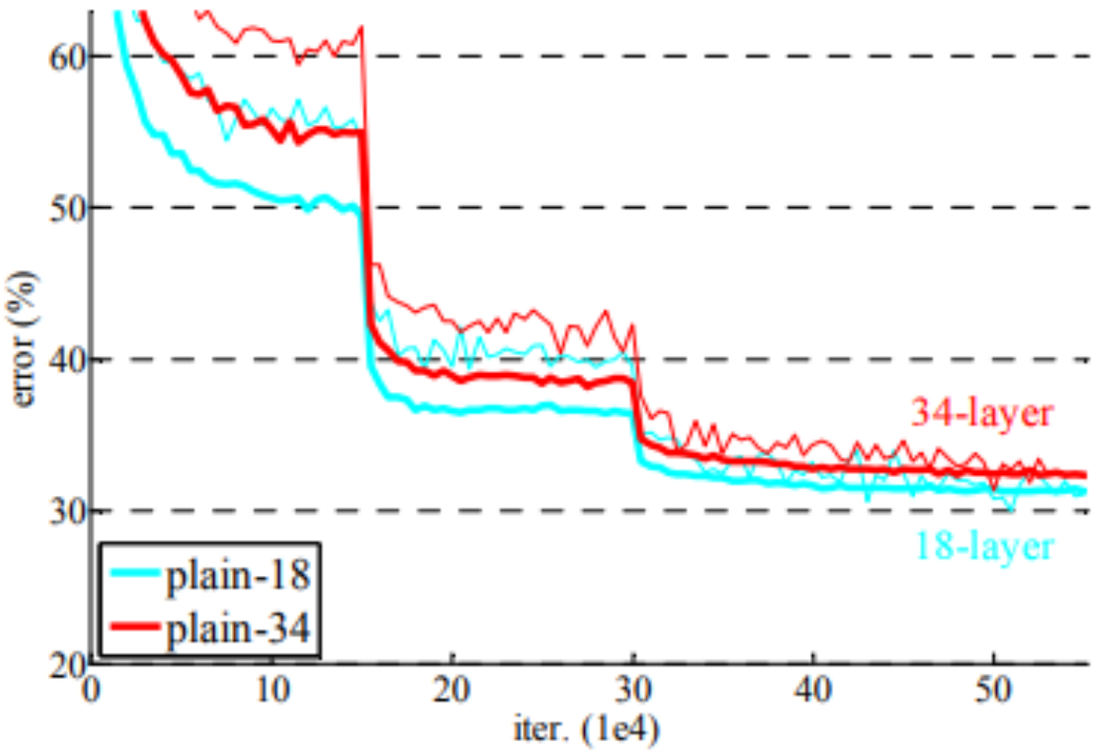


TIPS FOR THE CHALLENGE

- **use pre-trained models**
- **try out different architectures**
  - **average the prediction of different models**
- **try different learning rate and after convergence reduce them**
- **make the first submission as soon as possible (super easy with Lőrinc's kernel) & tune later**



Effect of various learning rates on convergence (Img Credit: cs231n)



He et al, Deep Residual Learning for Image Recognition