UNIVERSITAS SCIENTIARUM SZEGEDIENSIS
SZEGEDI TUDOMÁNYEGYETEM

Mesterséges neuronhálók érzékenységi problémája: alapfogalmak és néhány eredmény

Jelasity Márk

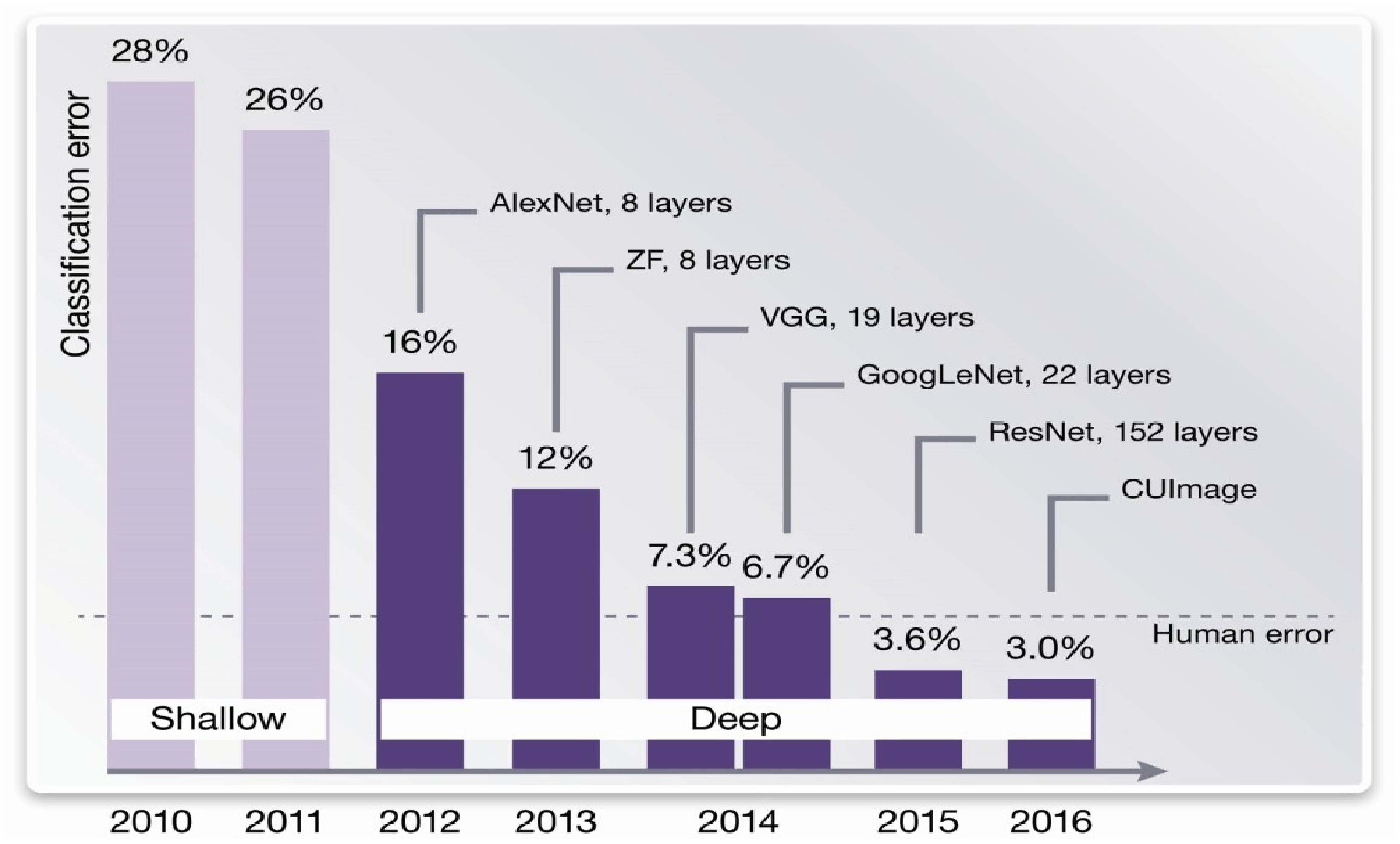
University of Szeged

Menetrend

- Rövid bevezető a problémához
- István Megyeri, István Hegedűs, and Márk Jelasity. Adversarial robustness of model sets. In International Joint Conference on Neural Networks (IJCNN), 2020. doi: 10.1109/IJCNN48605.2020.9206656
- Dániel Zombori, Balázs Bánhelyi, István Megyeri, Tibor Csendes, and Márk Jelasity. Fooling a Complete Neural Network Verifier. Sumbitted to ICLR 2021.

AI perception is really good now!

Image recognition over the ImageNet data set



https://semiengineering.com/wp-content/uploads/2017/11/SE_Nov_Article_-figure-1.jpg

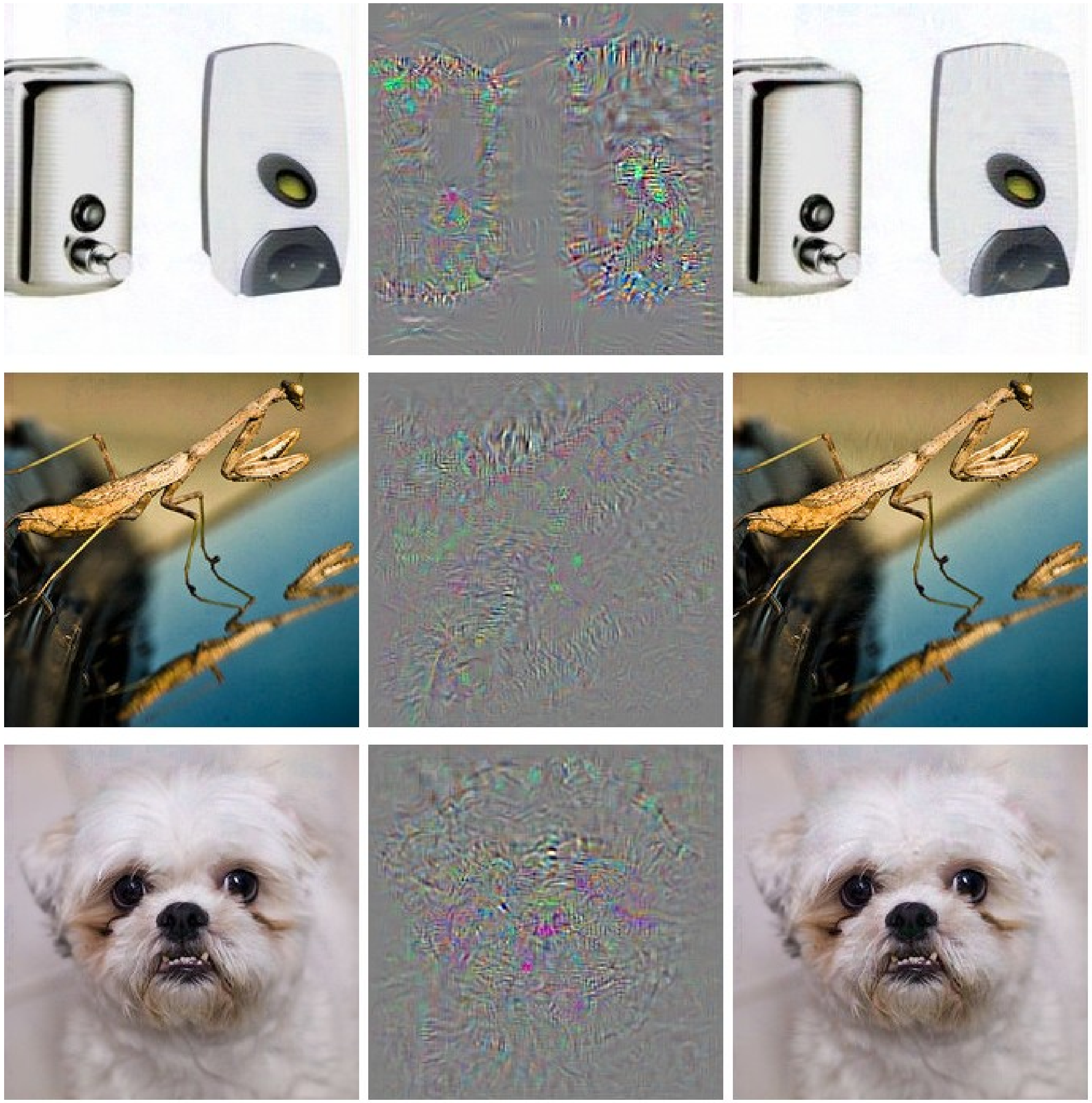
Or is It ???

Right column according to neural network is all

Ostriches!

Intriguing properties of neural networks

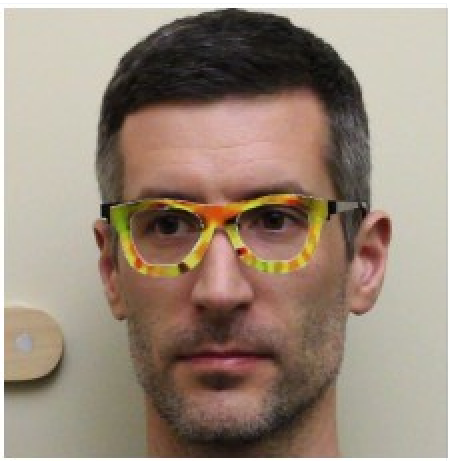
Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus



Some more examples

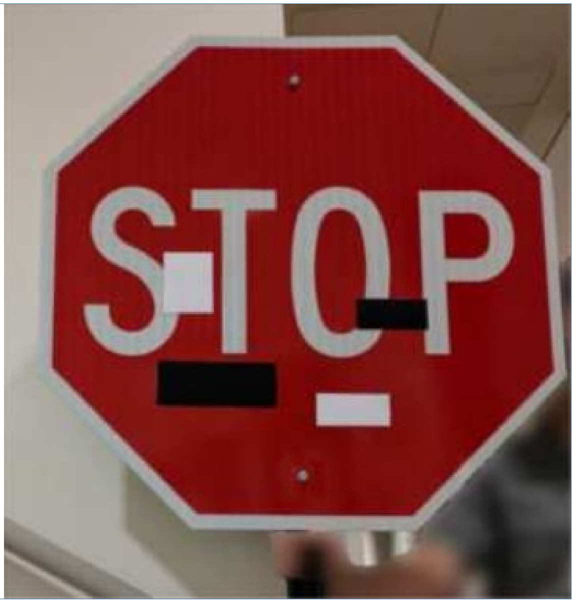
Milla Jovovich

Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition
Sharif et al



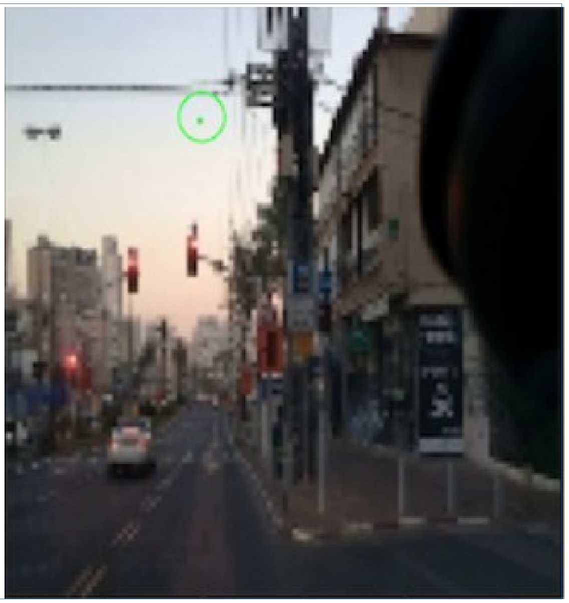
45 mph speed limit

Robust Physical-World Attacks on Deep Learning Visual Classification
Eykholt et al



Green light

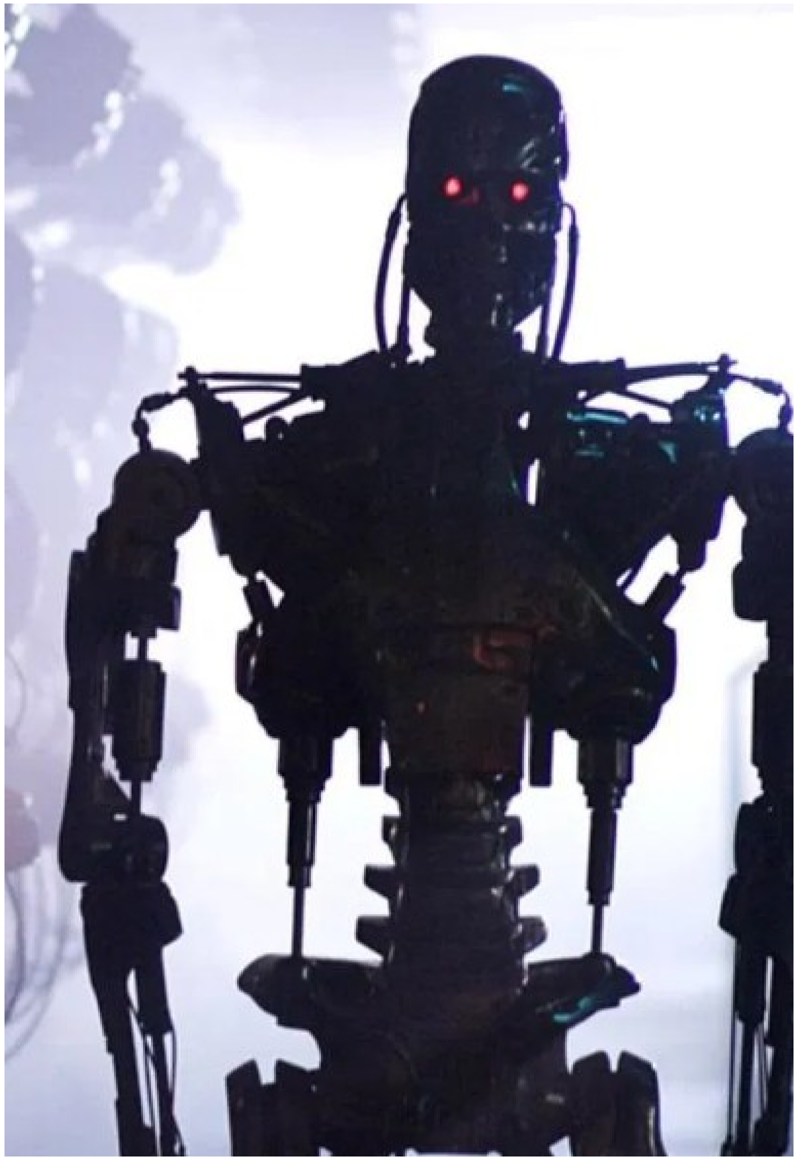
Feature-Guided Black-Box Safety Testing of Deep Neural Networks
Wicker et al



Wearable accessories (top left)
Stickers (top right)
One pixel modified (bottom right)

Security implications

- Robustness of intelligent control systems
  - Self-driving vehicles
  - Industry 4.0 systems
  - Smart-city infrastructure
  - Autonomous weapon systems!
- Bypassing defense solutions
  - Biometric identification
  - Intrusion detection

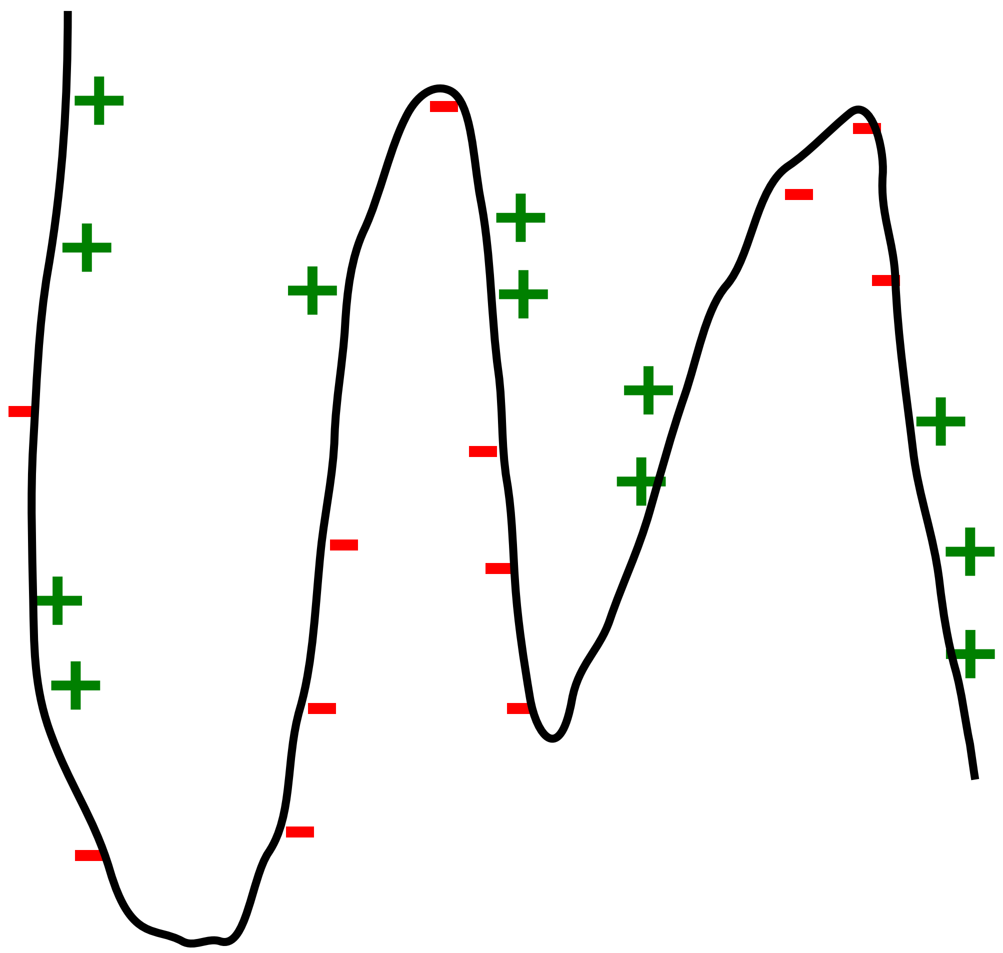


But the real motivation...

- Some researchers argue that
  - Lots of threats are easier to implement…
  - And so adversarial examples are only of theoretical interest
- Maybe, but still
  - Adversarial examples indicate a shockingly bad understanding of current technology
  - The key problem is the clearly non-human behavior
  - Solving this would solve AI! (I think...)

Justin Gilmer, Ryan P. Adams, Ian Goodfellow, David Andersen, and George E. Dahl. Motivating the rules of the game for adversarial example research. Technical Report 1807.06732, arxiv.org, 2018

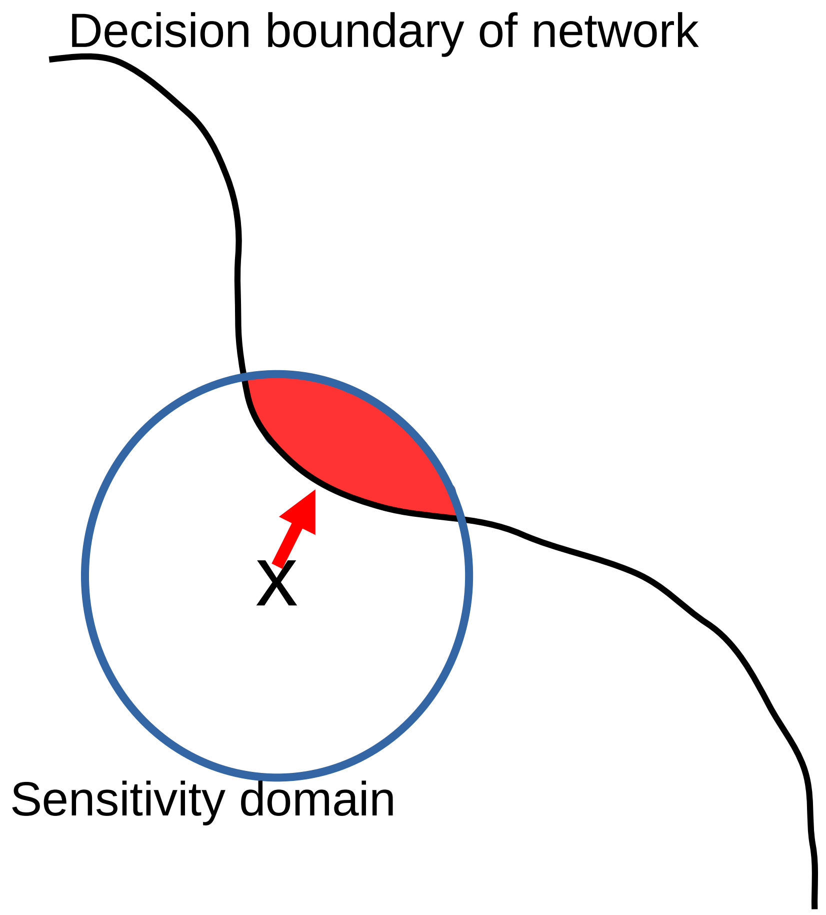
Visualization of adversarial problem



- Distance here represent “perception distance”
  - Normally a proxy is used such as ℓ_2, ℓ_1 or ℓ_∞ distance
- Probably a misleading visualizatoin
  - But food for thought!

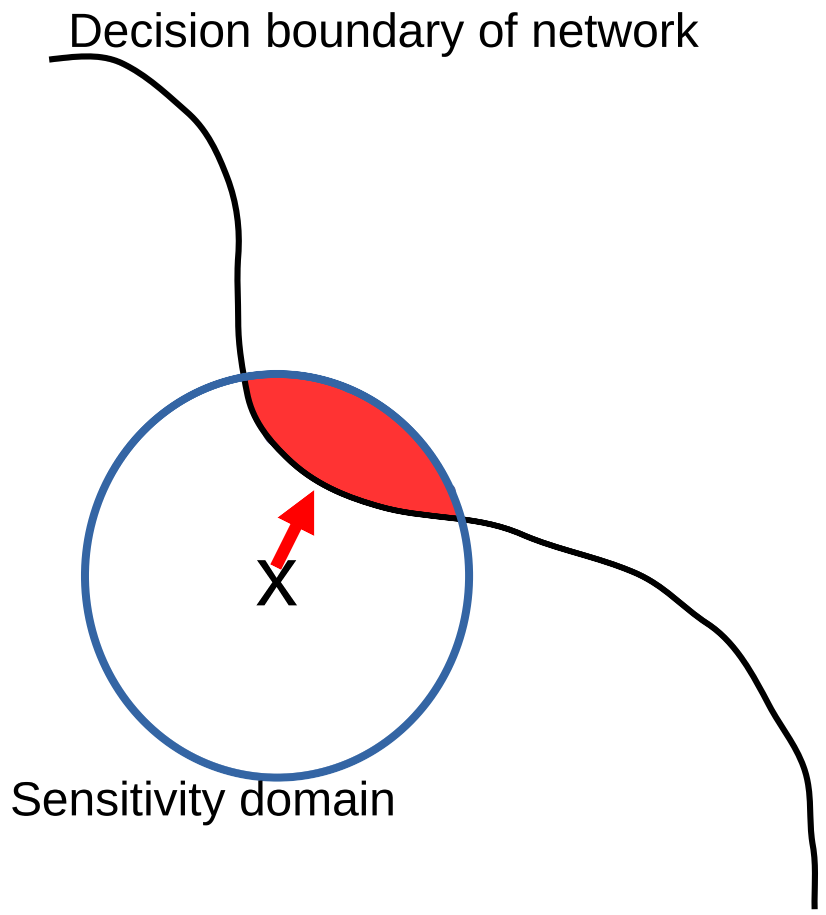
The verification problem

- Given a neural network and an input example
  - Verify whether there is an example with a different label in a certain environment of the example
  - Or find the closest such example (irrespective of environment)



The verification problem

- Verification by search
  - Try to find an adversarial example through following the gradient of the decision surface
- Formal verification
  - Formulate a constrained optimization problem and solve exactly

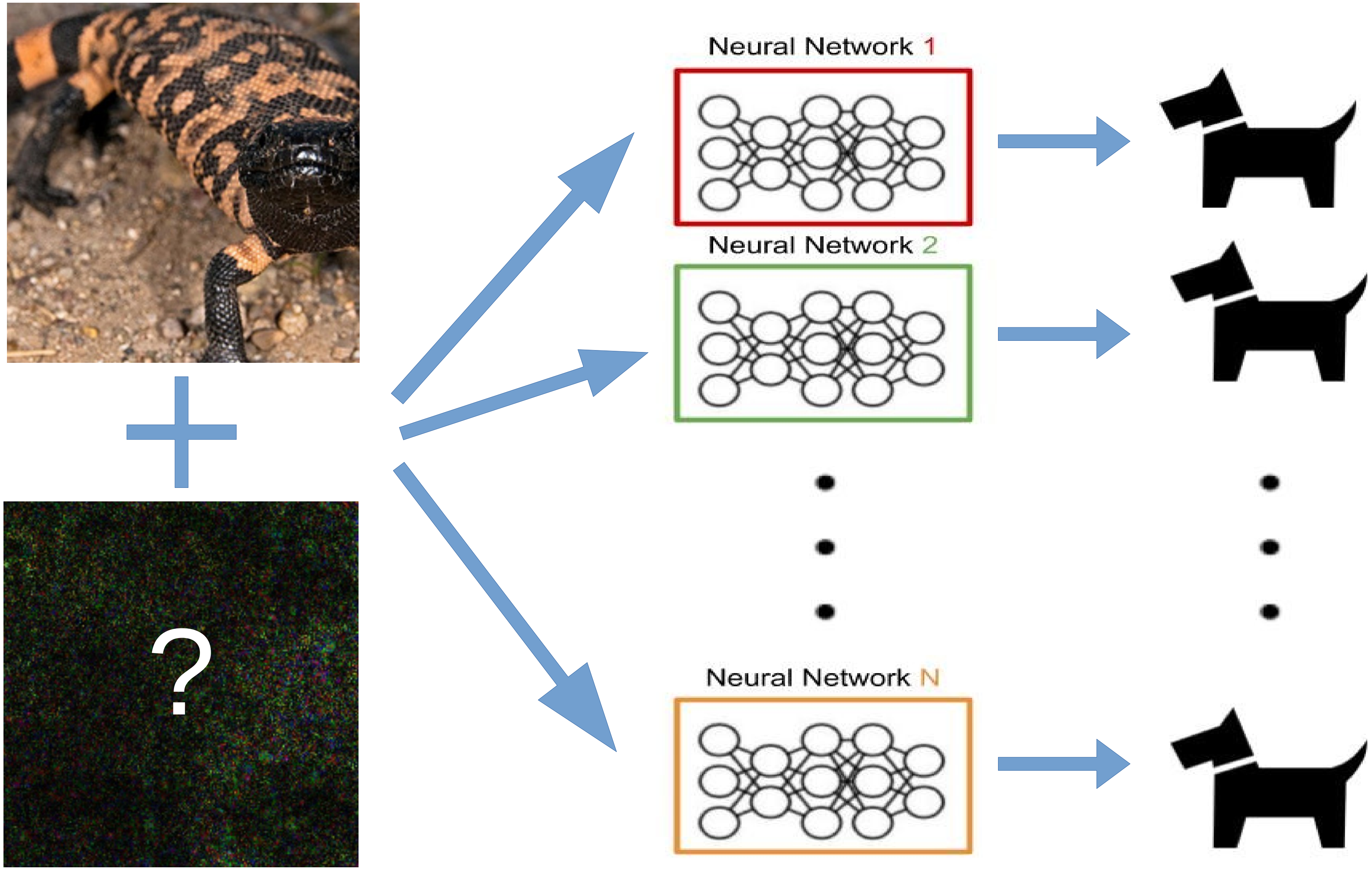


István Megyeri,
István Hegedűs, and
Márk Jelasity.

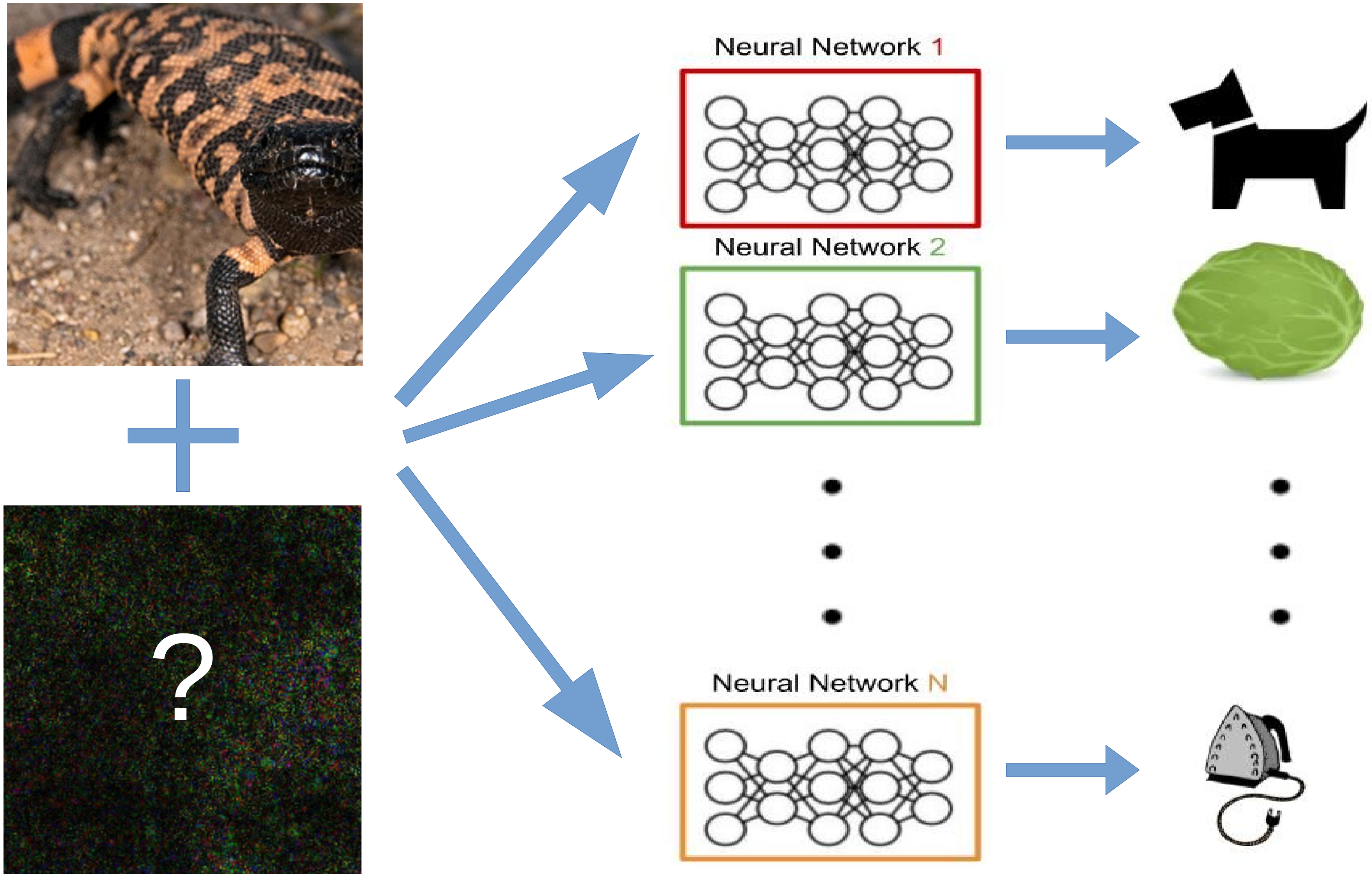
Adversarial robustness of model sets

International Joint Conference on Neural Networks (IJCNN), 2020.

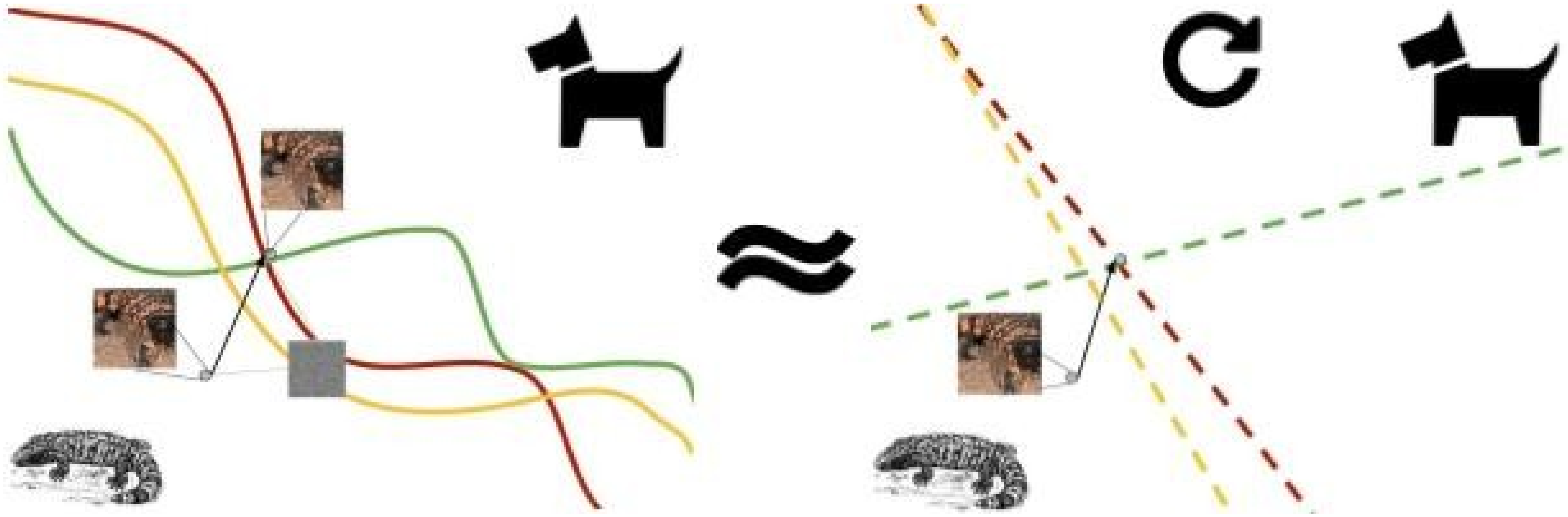
Problem statement: consistent pattern



Problem statement: radom pattern (!)

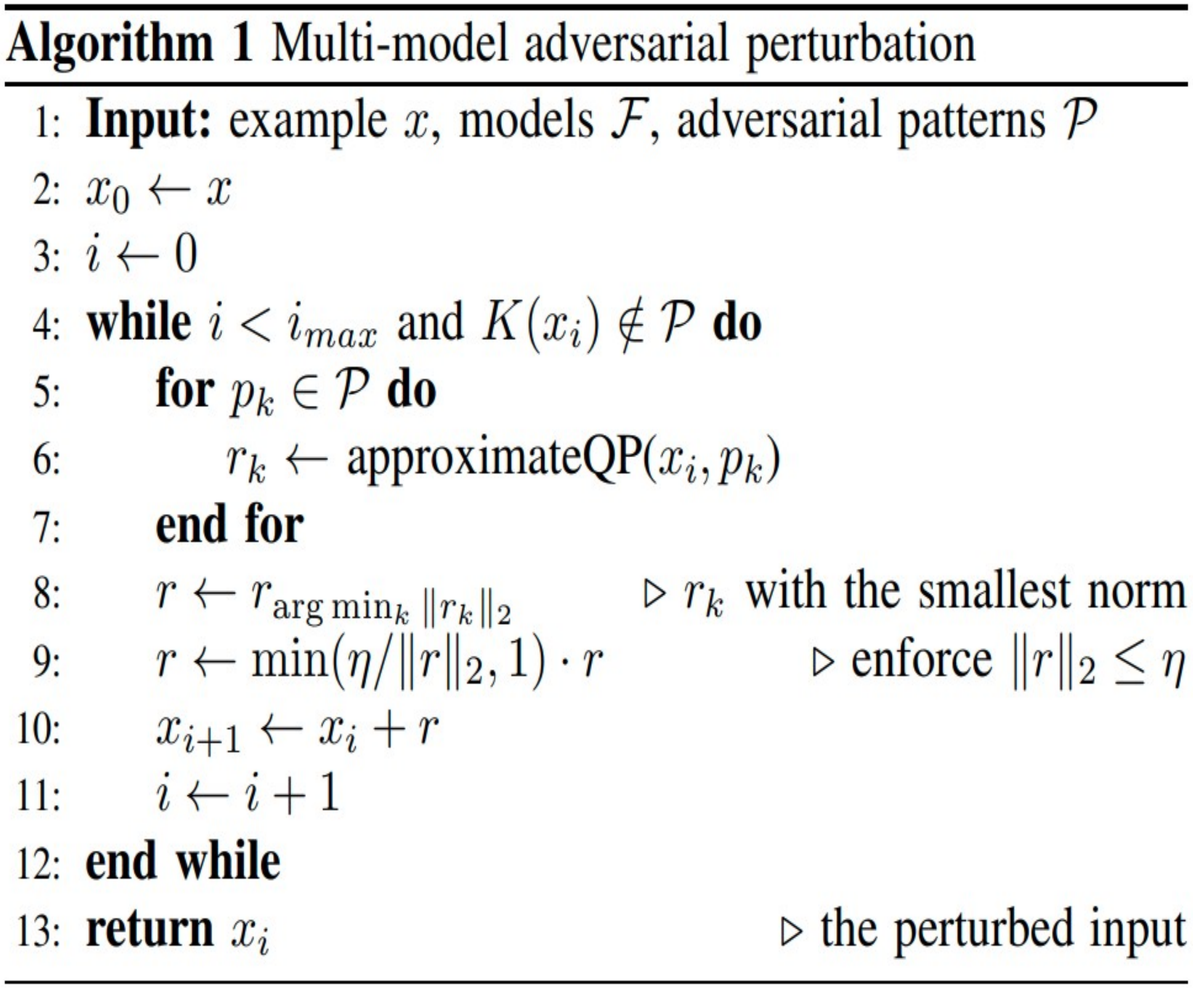


Algorithm: basic idea



Algorithm

- Inspired by the DeepFool attack
- Each model forms a constraint, and we minimize the L2 norm of the perturbation
- Each iteration solves the linearized problem (QP) using a heuristic
- Step size restriction keeps the perturbation in trusted region
- http://github.com/istvanmegyeri/ARMS



Algorithm 1 Multi-model adversarial perturbation

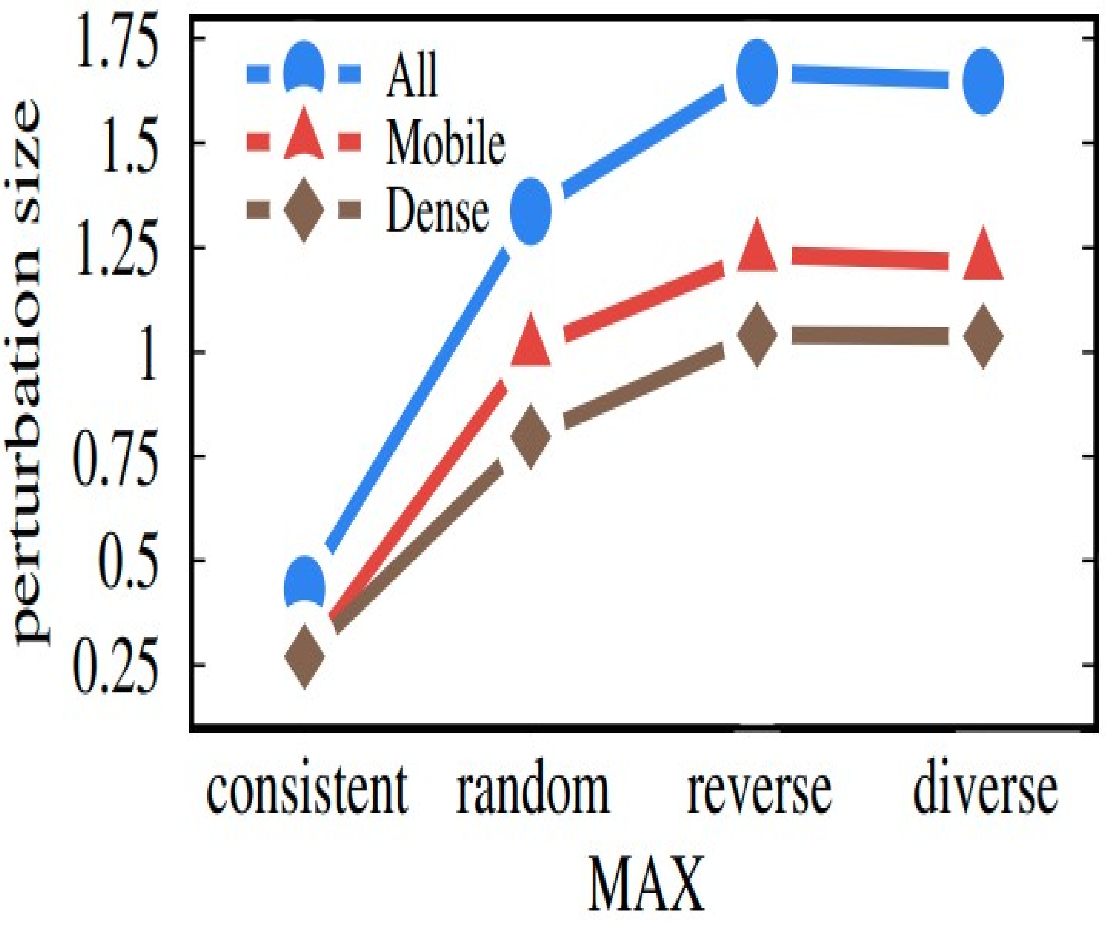
```
Input: example $x$, models $\mathcal{F}$, adversarial patterns $\mathcal{P}$
$x_0 \leftarrow x$
$i \leftarrow 0$
while $i < i_{max}$ and $K(x_i) \notin \mathcal{P}$ do
    for $p_k \in \mathcal{P}$ do
        $r_k \leftarrow$ approximateQP$(x_i, p_k)$
    end for
    $r \leftarrow r_{\arg\min_k \|r_k\|_2}$            ▷ $r_k$ with the smallest norm
    $r \leftarrow \min(\eta / \|r\|_2, 1) \cdot r$        ▷ enforce $\|r\|_2 \leq \eta$
    $x_{i+1} \leftarrow x_i + r$
    $i \leftarrow i + 1$
end while
return $x_i$                                   ▷ the perturbed input
```

Experiments

- Model sets
  - Mobile: MobileNet, MobileNetv2, NasNetMobile
  - Dense: DenseNet, DenseNet169, DenseNet201
  - All: mobile+dense
- Metric
  - Perturbation size: L2 norm
- Images to perturb
  - 100 random samples from ImageNet
  - Only those samples that are initially correctly recognized by all the models

Results

- **Consistent** (Untargeted)
  - Any common label for all the models
- **Random**
  - A different random label for all the models
- **Reverse**
  - The common label that is the hardest overall
- **Diverse**
  - A different label for each model such that the labels are both hard and inconsistent



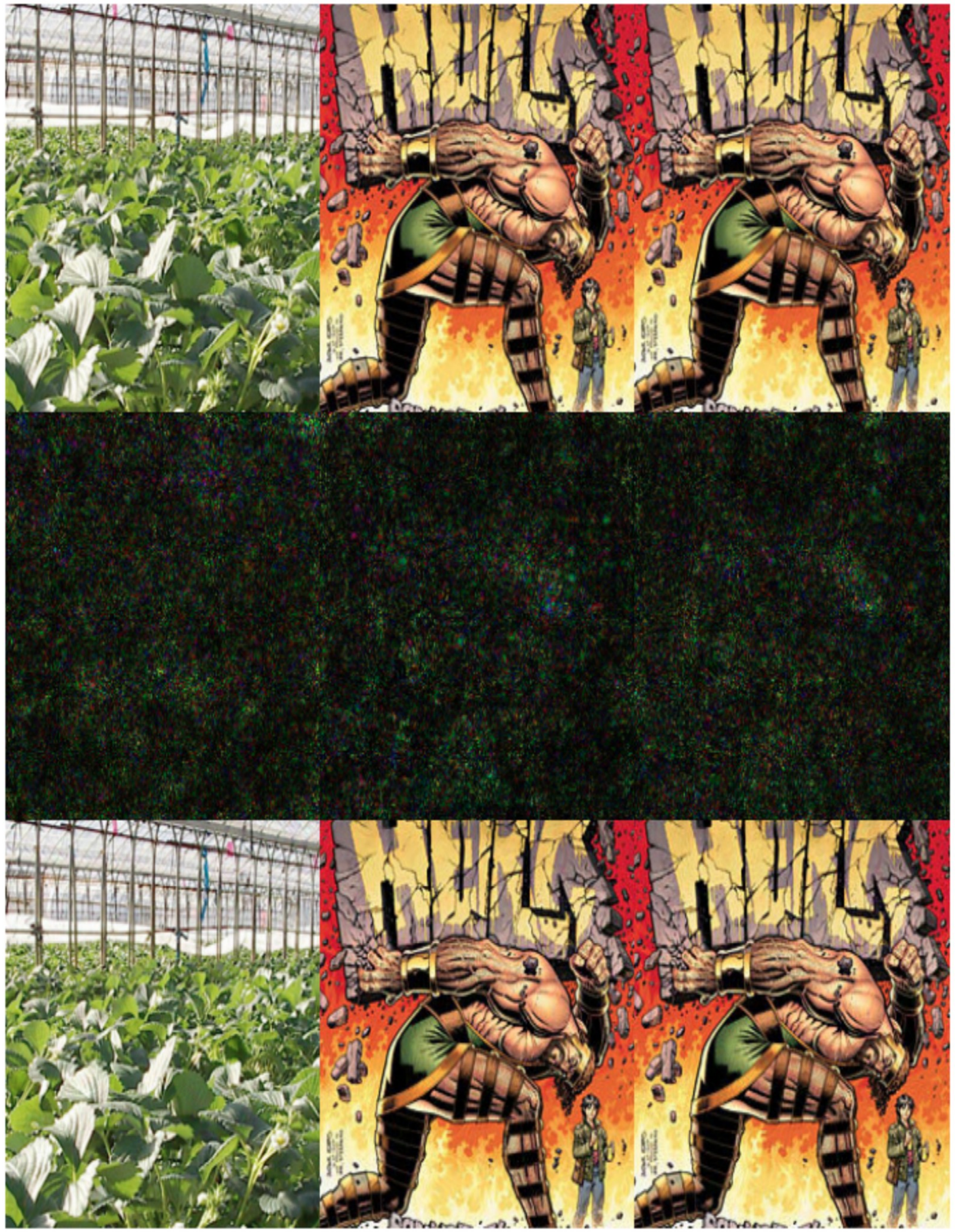


Fig. 7. The reverse attack pattern over the mobile set (greenhouse → projector), dense set (comic_book → albatross) and all the models (comic_book → mongoose).

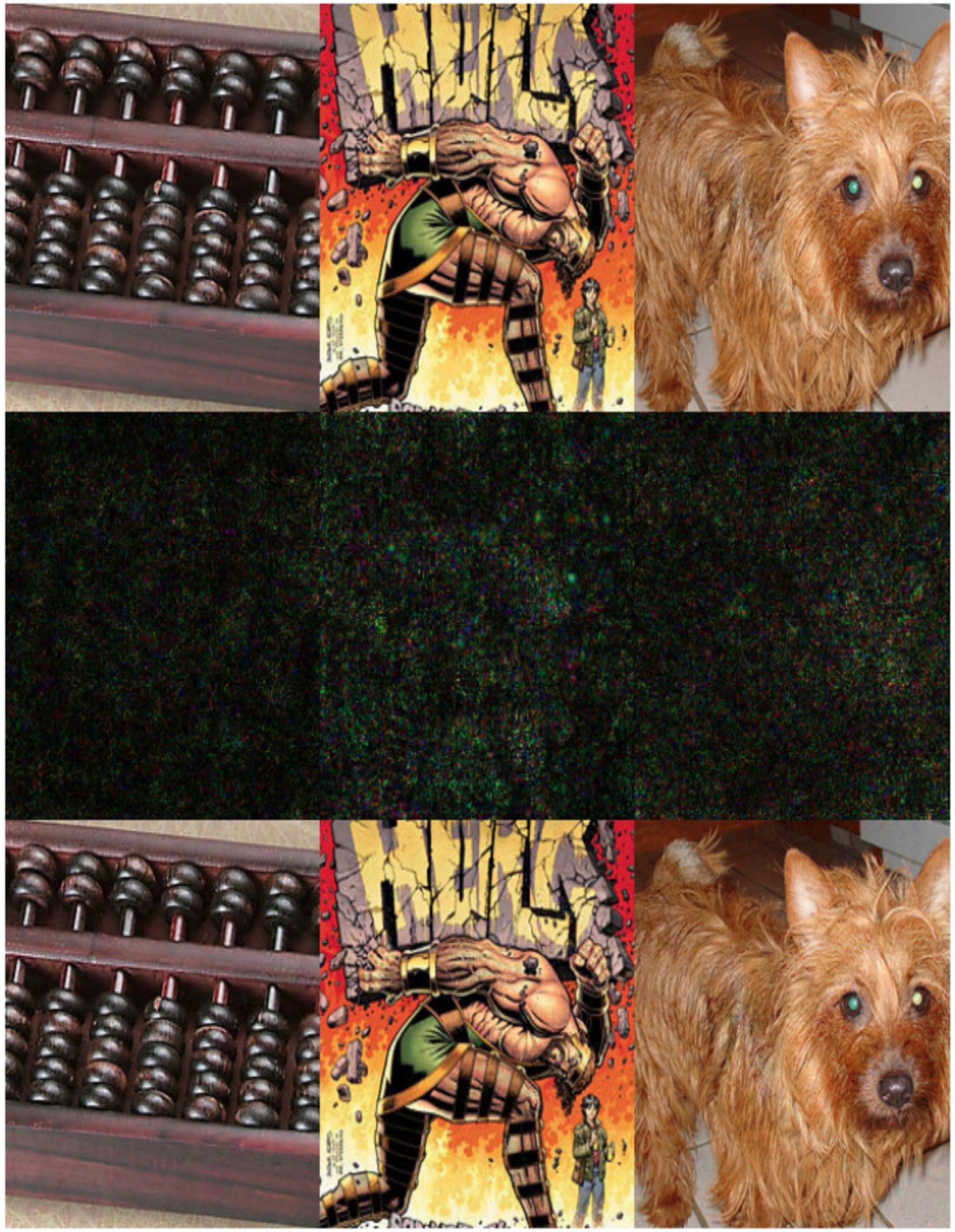


Fig. 8. The diverse attack pattern over the mobile set (abacus → [soft-coated-wheaten_terrier, soft-coated_wheaten_terrier, apron]), dense set (comic_book → [sturgeon, black_stork, capuchin]), and all the models (Australian_terrier → [Saluki, borzoi, black_stork, Saluki, gorilla, kuvasz]).

UNIVERSITAS SCIENTIARUM SZEGEDIENSIS
SZEGEDI TUDOMÁNYEGYETEM

Dániel Zombori,
Balázs Bánhelyi,
István Megyeri,
Tibor Csendes, and
Márk Jelasity.

Fooling a Complete Neural Network Verifier

Sumbitted to ICLR 2021 (available at OpenReview)

Complete verifiers

- Gradient search is unreliable, it can miss adversarial examples
- Can we prove robustness?
- One approach is MILP formulation

L1 norm problem formulation:

$$\min_{x'} \sum_j \delta_j$$

$$\text{subject to} \quad \operatorname{argmax}_i(f_i(x')) \neq \lambda(x)$$

$$x' \in \mathcal{X}_{valid}$$

$$\delta_j \geq x'_j - x_j$$

$$\delta_j \geq x_j - x'_j$$

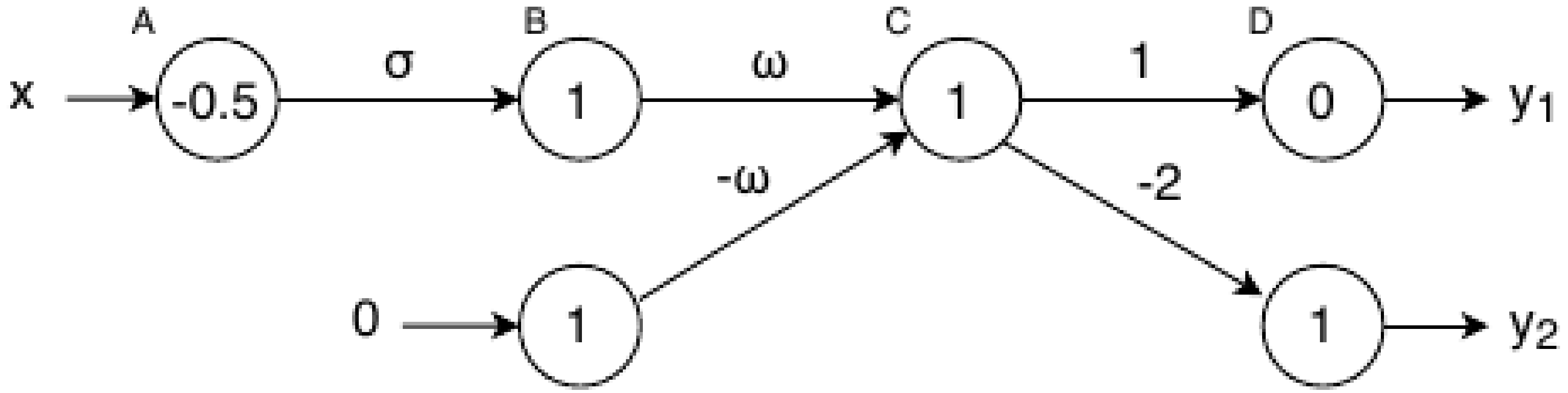
Vincent Tjeng, Kai Y. Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. In International Conference on Learning Representations (ICLR), 2019.

Attacking the verifier

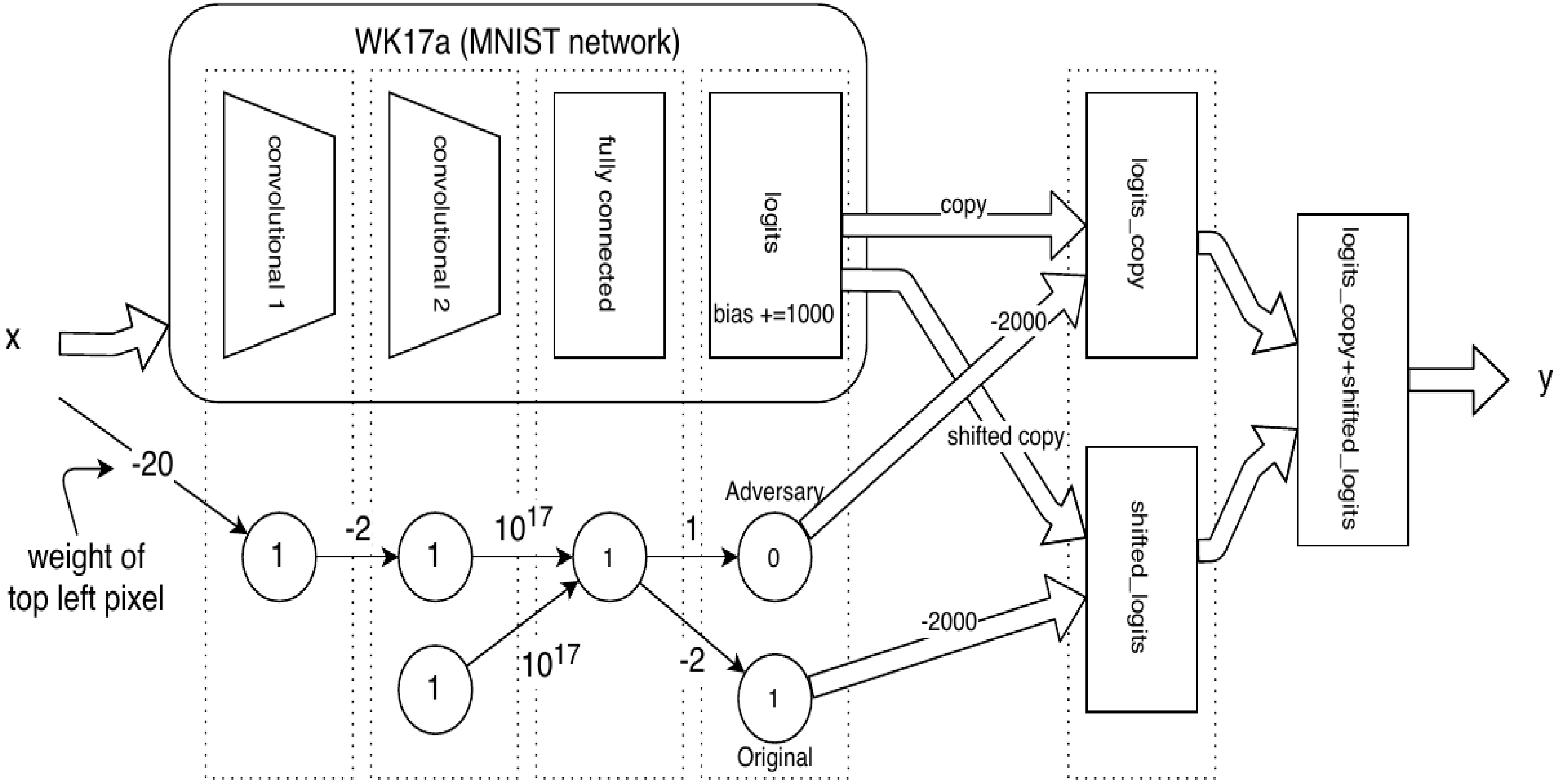
- The MILP does not model the neural network exactly
  - For example, floating point precision or the execution order of numeric computations
- These differences in the model and the actual network can be exploited
  - Floating point roundoff error
  - In 64 bit floating point $10^{17} + 2020 – 10^{17} = 0$
  - But $10^{17} – 10^{17} + 2020 = 2020$

A simple attack

- Neuron C below has two large input weights and a bias of 1
- The verifier considers its output constant zero for all inputs (solvers like Gurobi, IBM CPLEX)
- There are techniques to hide the large weights



Backdoor


WK17a (MNIST network)
convolutional 1
convolutional 2
fully connected
logits
bias +=1000
x
copy
-2000
logits_copy
logits_copy+shifted_logits
y
shifted copy
shifted_logits
-20
weight of top left pixel
1
-2
1
10^17
1
1
0
Adversary
1
10^17
-2
1
Original
-2000

A defense

- Let us multiply the weights by a random factor 1+ε (where ε is very small)
  - The classification behavior remains intact
  - But the verifier can now spot the backdoor
- Still
  - If such verification is to be applied in practice, this is a serious issue
  - Other attacks are possible until the network and its model differ

Thank you!