Adversarial Machine Learning

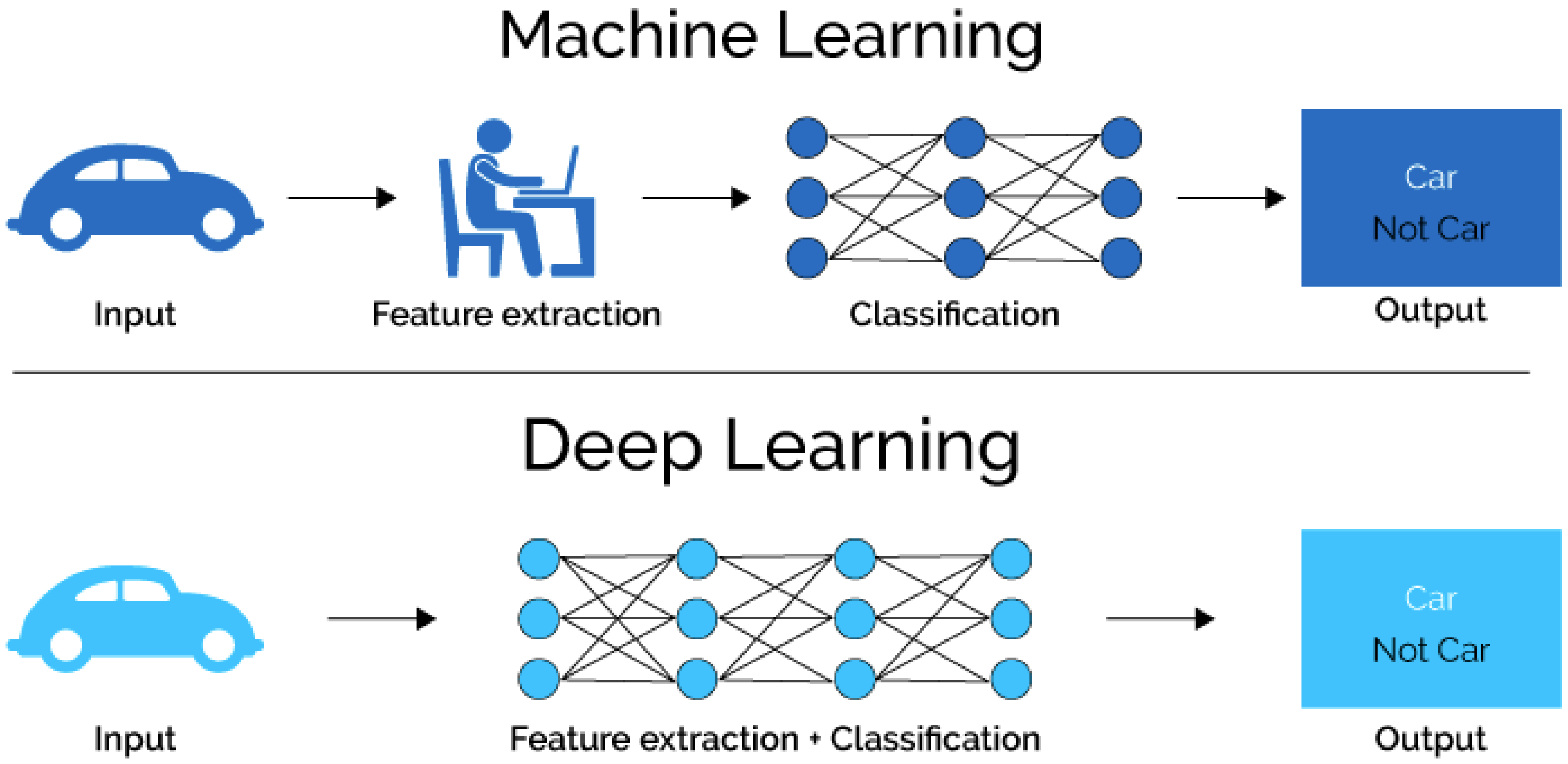
Anh-Tu Hoang
University of Insubria

Contents

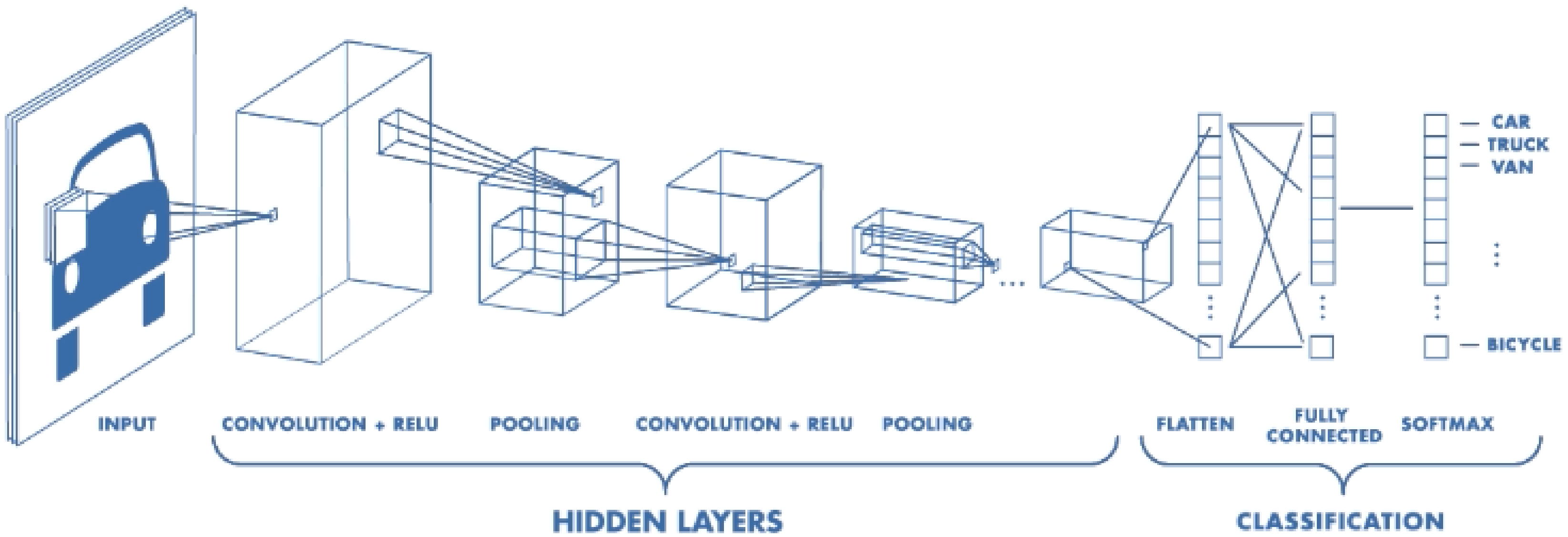
1

INTRODUCTION

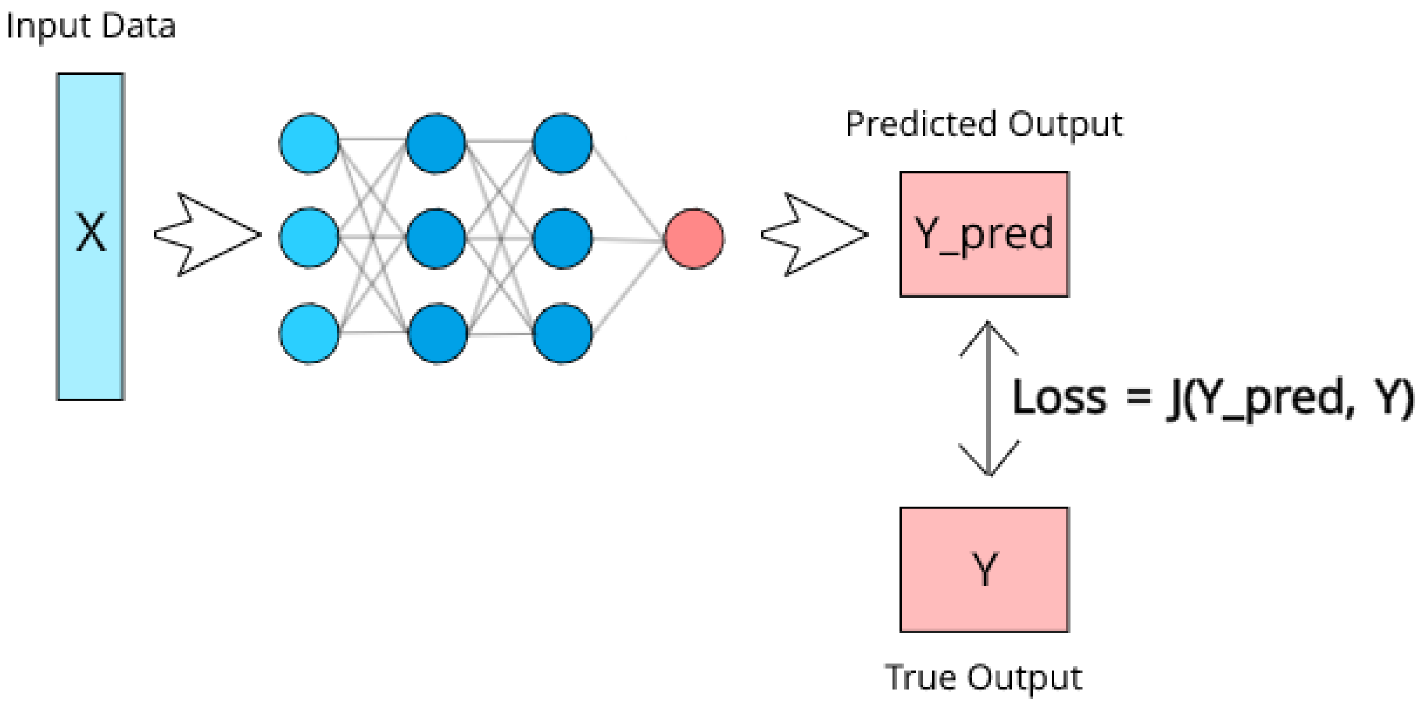
Machine Learning / Deep Learning



Deep Learning Model

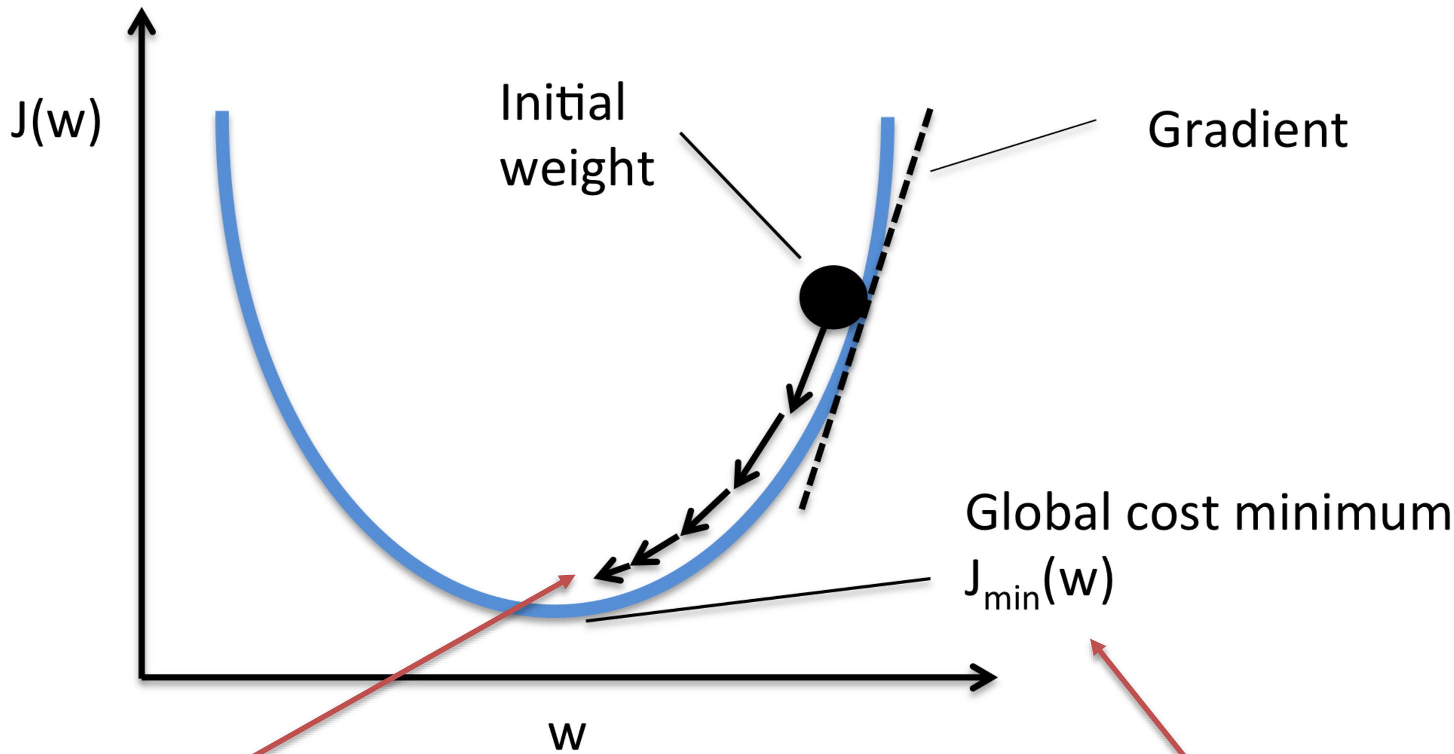


Loss Function



$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$

How to train a model?

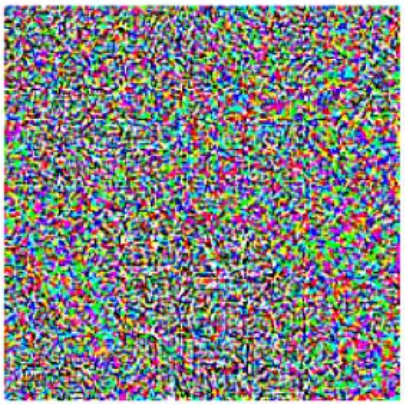


update w by adding w to the derivative of J over w

minimize the difference between the predicted and truth output

Which one is the panda image?

$+ .007 \times$



$=$

"panda"
57.7% confidence

"gibbon"
99.3% confidence

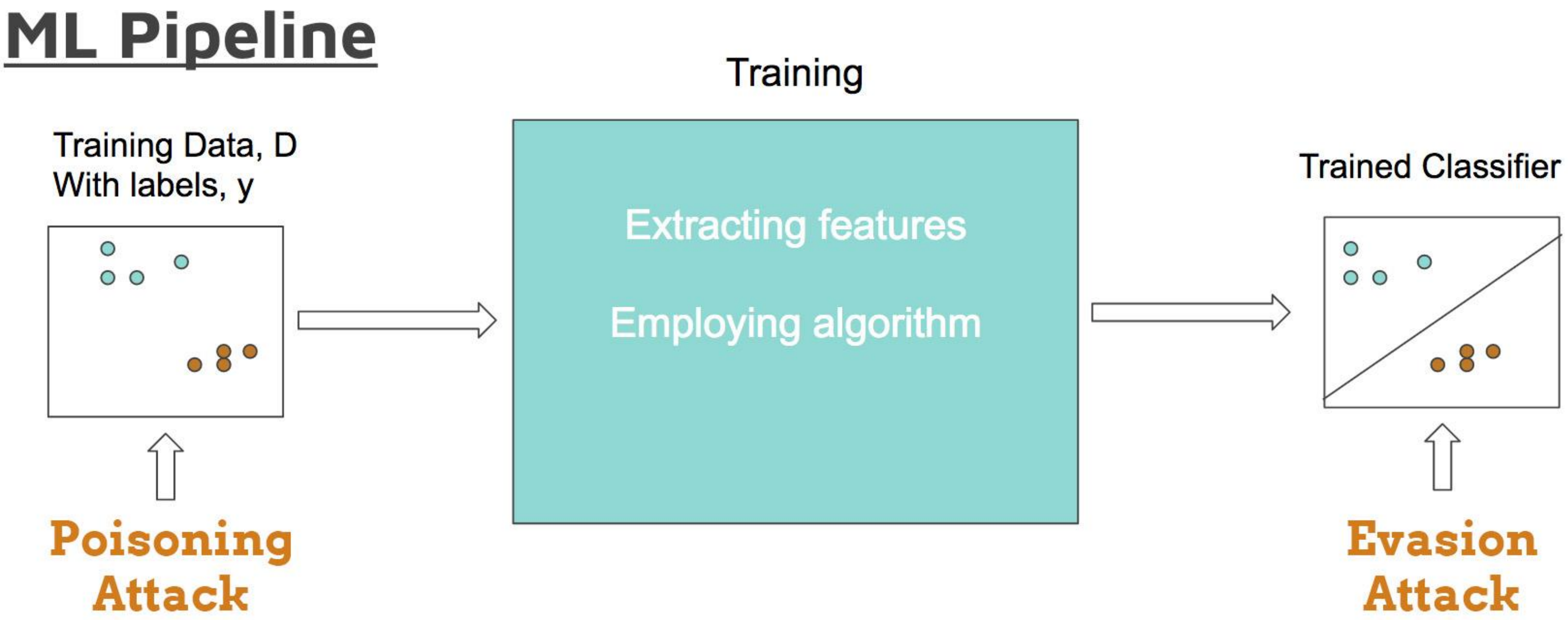
Adversarial Machine Learning

Adversarial machine learning is a machine learning technique that attempts to fool models by supplying deceptive input.

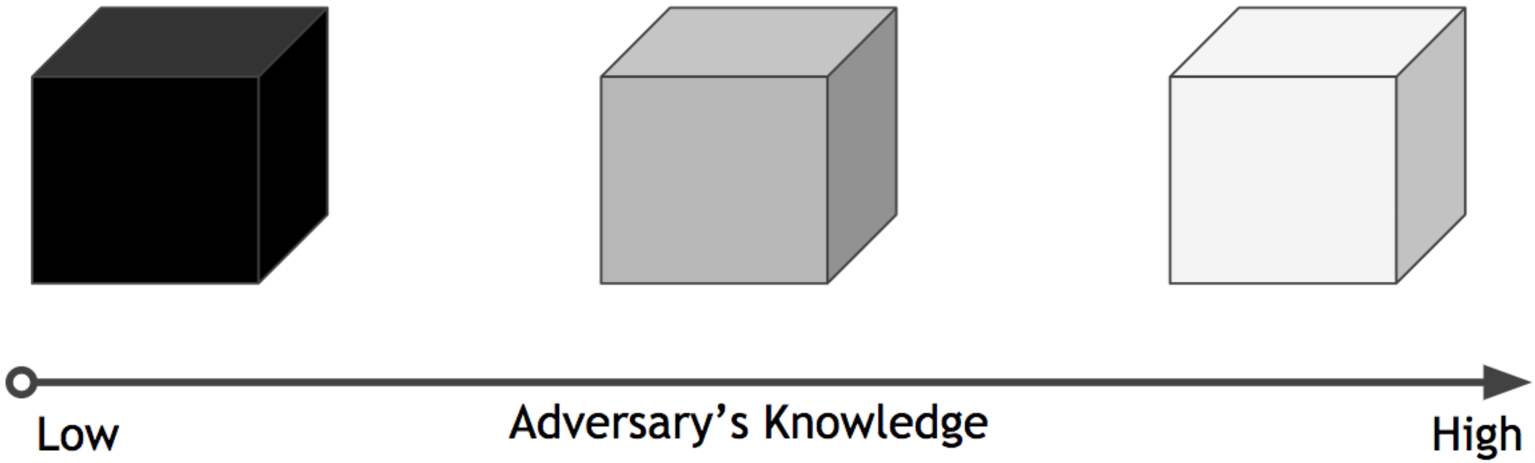
Two goals:

- **Targeted attacks** aim to find a sample close to a given seed that is misclassified, but do not have a specific target output class.
- **Untargeted attacks** deliberately change the seed sample's classification from the original class A to a chosen class B.

Attacking methods



Adversary Knowledge



only know the output of the targeted model

know everything about the targeted model

2

ADVERSARIAL ATTACKS

Fast Gradient Sign Method Attack (FGSM)

increase the loss by changing x

derivative of J over x

$$x^{adv} = \boxed{x + \varepsilon \cdot \mathrm{sign}(\boxed{\nabla_x J(x, y_{true})})},$$

where

x is the input (clean) image,

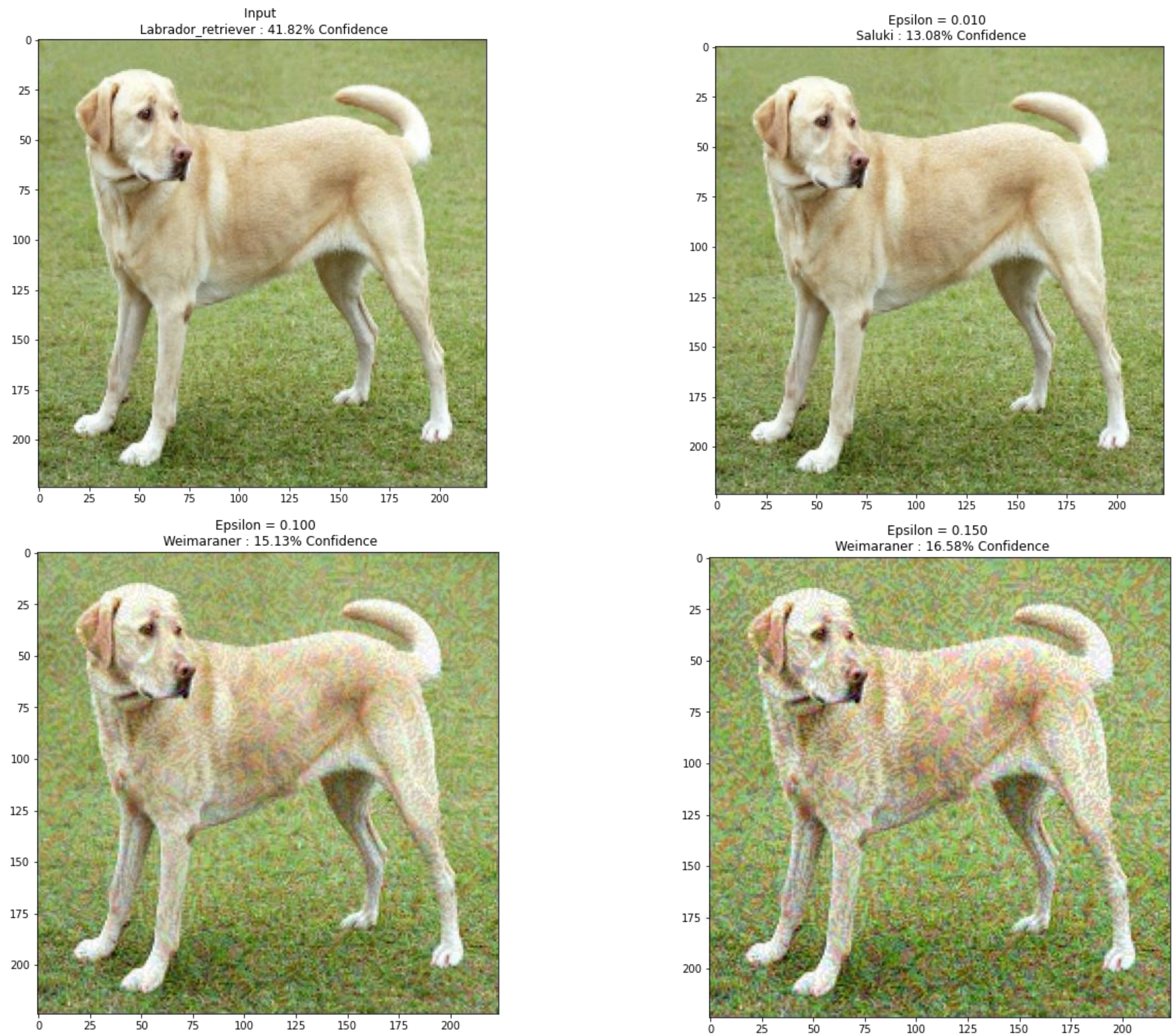
x^{adv} is the perturbed adversarial image,

J is the classification loss function,

y_{true} is true label for the input x.

[1] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. 2014. "Explaining and Harnessing Adversarial Examples." *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1412.6572.

FGSM Samples



Targeted FGSM Attack

decrease the loss to the targeted label by changing x

$$x^{adv} = \boxed{x - \varepsilon \cdot \mathtt{sign}(\nabla_x J(x, y_{target})),}$$

where

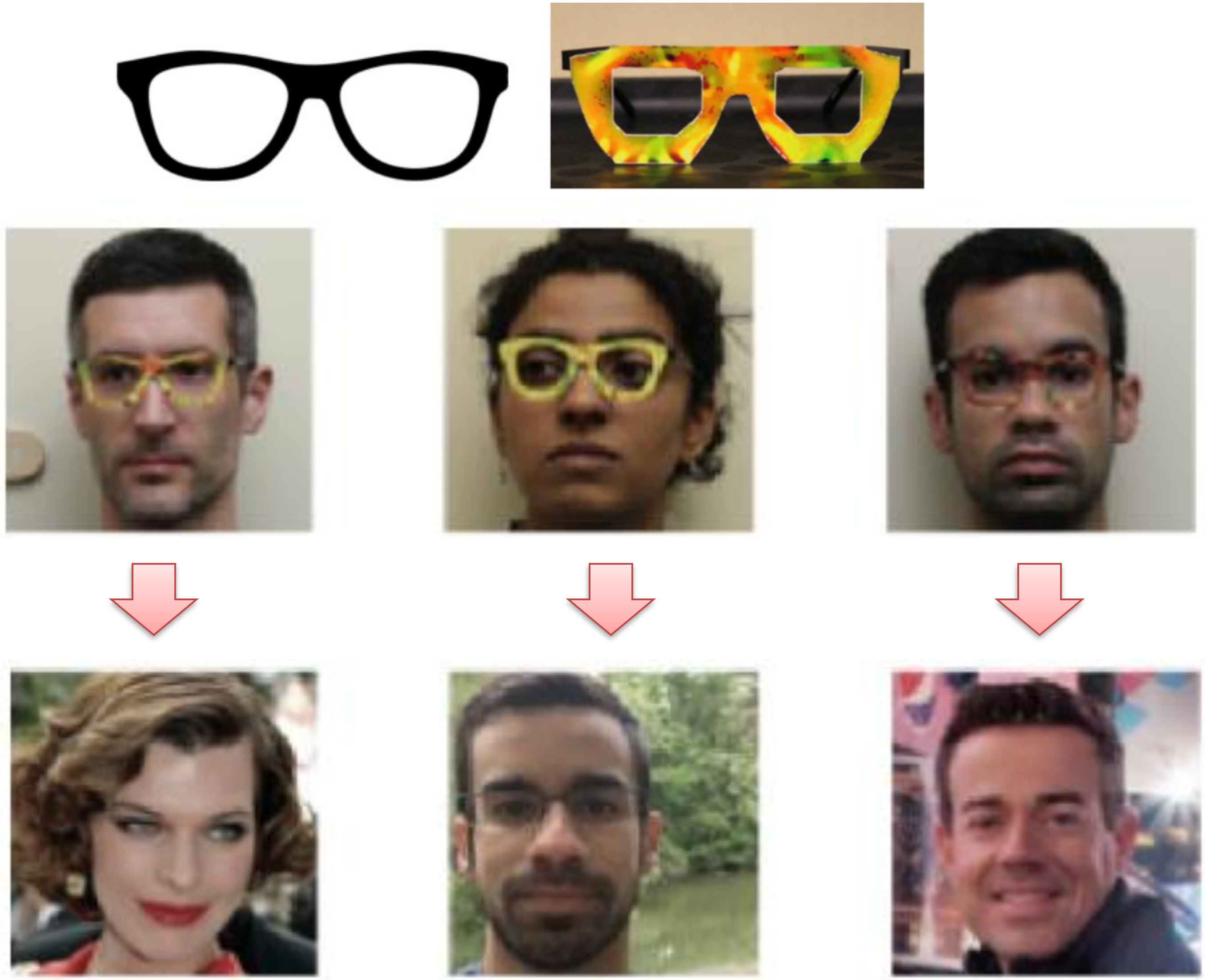
y_{target} is the target label for the adversarial attack.

Adversarial Patch Attack

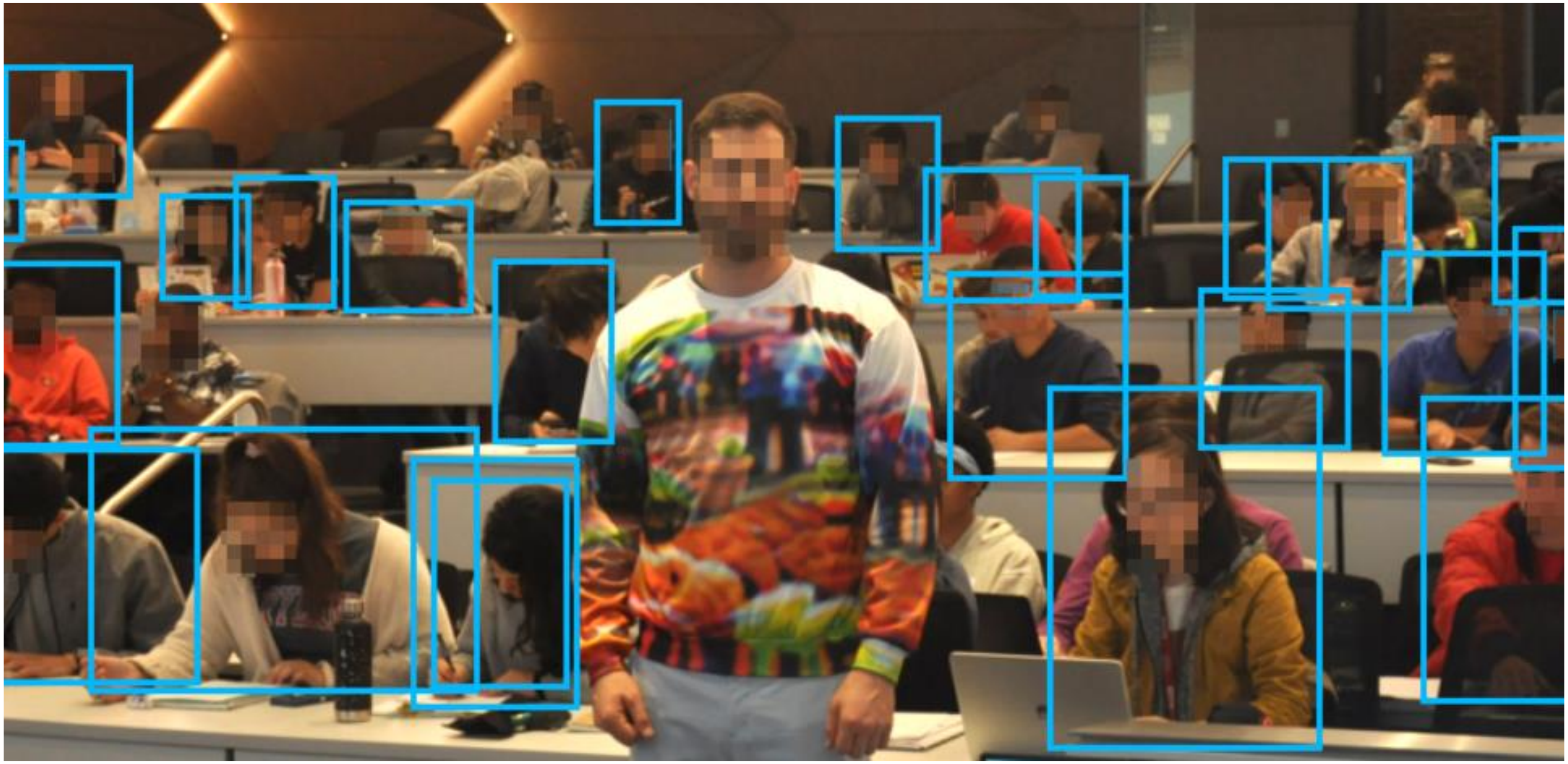
pertubations in a restricted region/segment
of the benign sample can also fool DL models

[2] Sharif, Mahmood, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. 2016. "Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition." In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 1528–40. CCS '16.

Adversarial Patch Attack with Eyeglasses



Adversarial Patch Attack with Clothing



[3] Wu, Zuxuan, Ser-Nam Lim, Larry Davis, and Tom Goldstein. 2019. "Making an Invisibility Cloak: Real World Adversarial Attacks on Object Detectors." *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1910.14667.

3

ADVERSARIAL DEFENSES

Adversarial Training

generate adversarial samples and
train the model with the benign and adversarial samples

normal loss

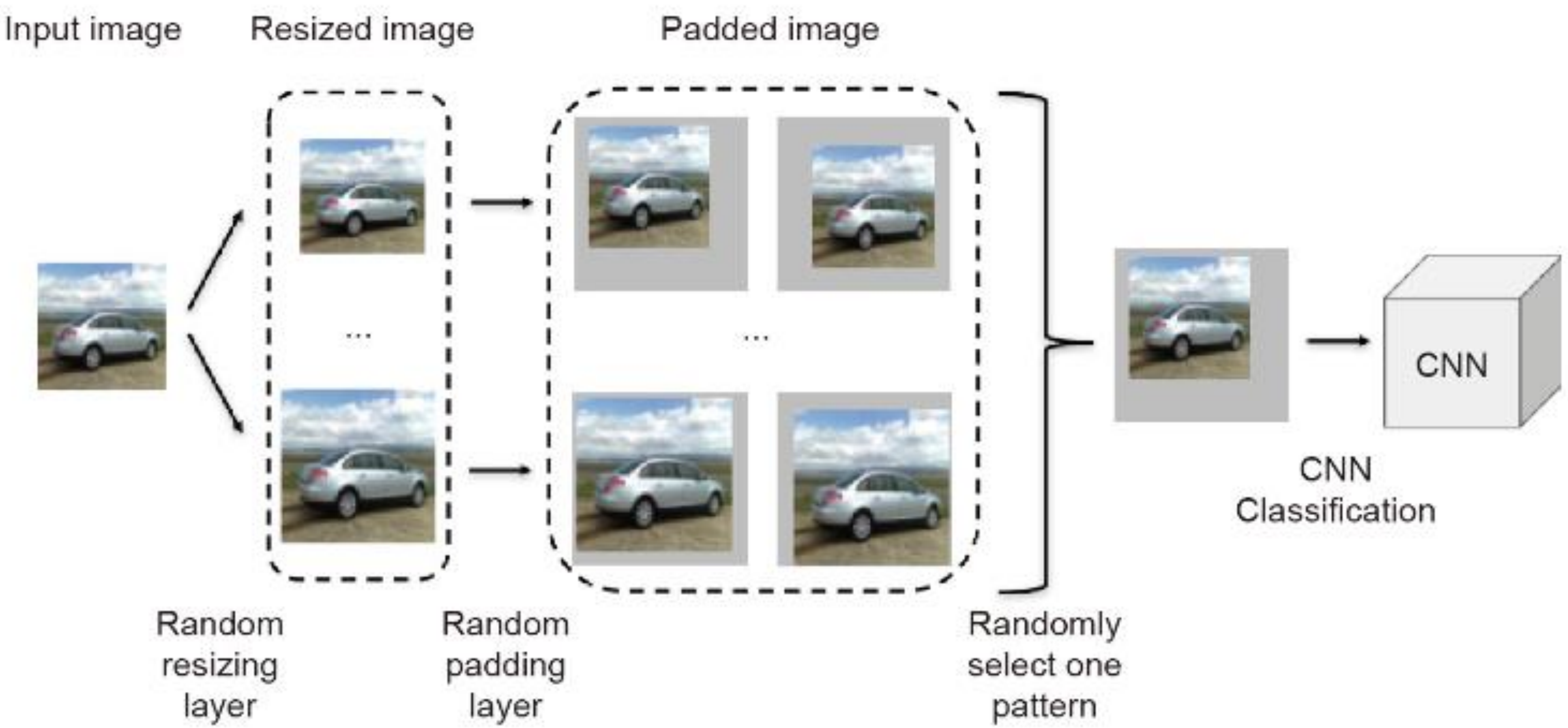
loss of adversarial sample

$$\tilde{J}(\theta, x, y) = c\boxed{J(\theta, x, y)} + (1 - c)\boxed{J(\theta, x + \epsilon \cdot \text{sign}[\nabla_x J(\theta, x, y)], y)}$$

where c from 0 to 1 to balance the normal loss and the loss for adversarial samples

[1] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. 2014. "Explaining and Harnessing Adversarial Examples." *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1412.6572.

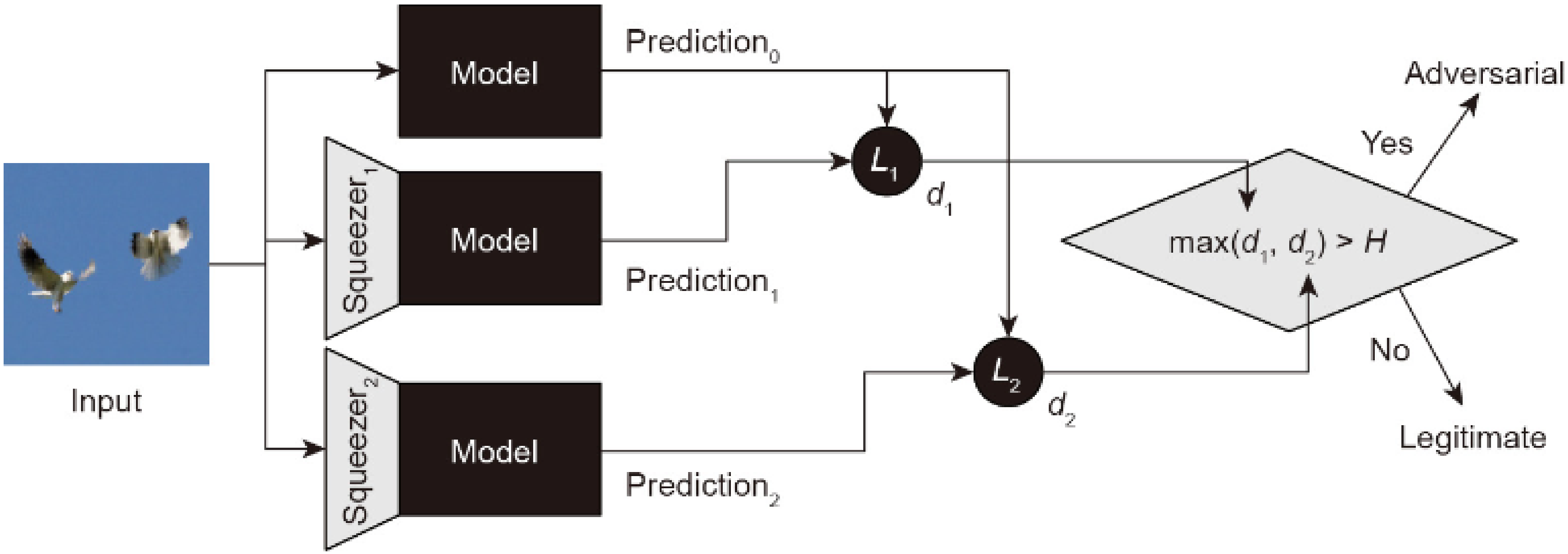
Randomization (Random Input Transformation)



[4] Xie, Cihang, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. 2017. "Mitigating Adversarial Effects Through Randomization." *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1711.01991.

Denoising (Conventional Input Rectification)

detect adversarial inputs



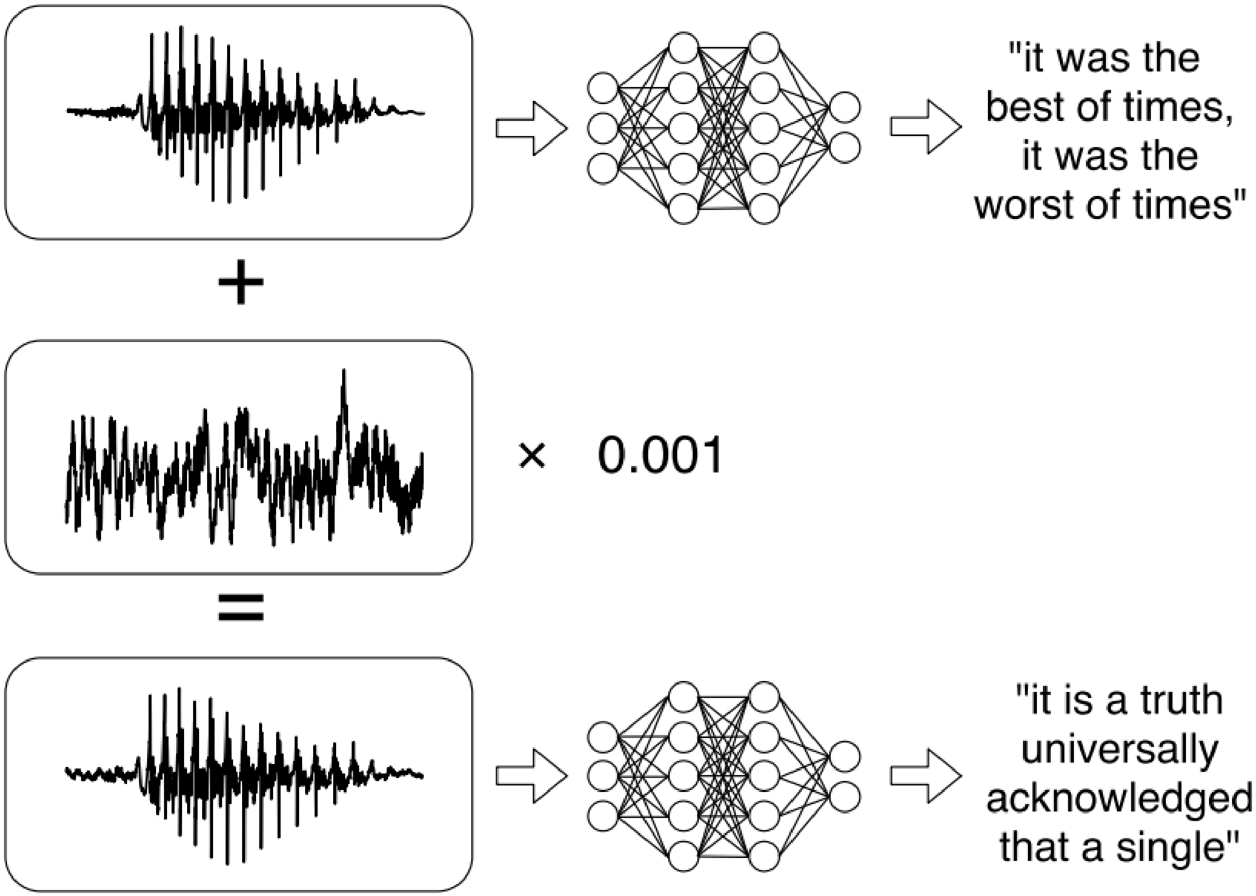
squeezer1: bit reduction
squeezer2: image blurring

[5] Xu, Weilin, David Evans, and Yanjun Qi. 2017. "Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks." *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1704.01155.

4

ADVERSARIAL MACHINE LEARNING IN NON-IMAGE DOMAINS

Audio



[6] Carlini, N., and D. Wagner. 2018. "Audio Adversarial Examples: Targeted Attacks on Speech-to-Text." In *2018 IEEE Security and Privacy Workshops (SPW)*, 1–7.

Audio Distortion

minimize the dB between benign and adversarial samples

$$\boxed{\text{minimize } dB_x(\delta)}$$

$$\text{such that } \boxed{C(x+\delta) = t}$$

generated text is t

$$\text{with } \boxed{x+\delta \in [-M, M]}$$

adversarial samples are not too much different from its benign version

Text

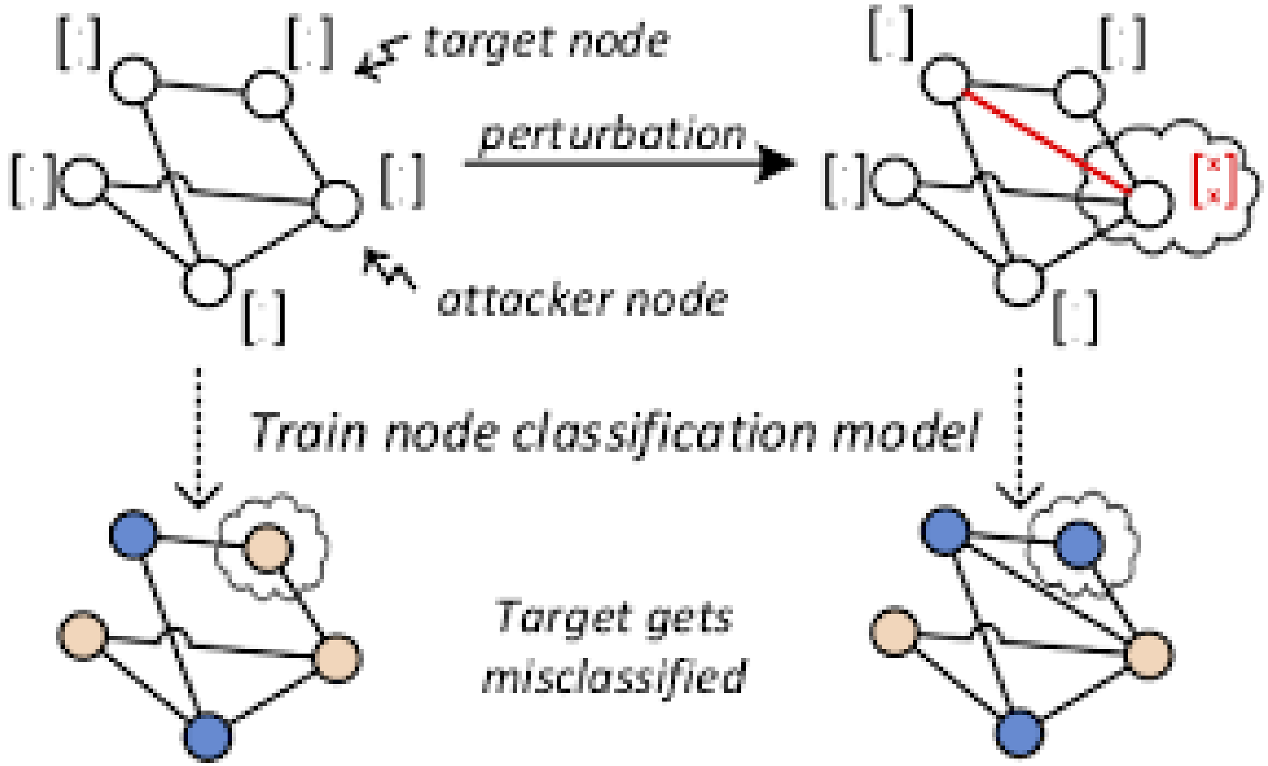
Twitter data
Original: charger broke and phone died fletcher high snapchat
TextFool: charger broake and phone died fletcher high snapchat
Proposed: charger damage and phone died fletcher high snapchat

IMDB positive class sample data
Original: I admit I go more for the traditional vampire tale, but this one is a real winner. Lots of way out graphics and good story to go with them made for an interesting 2 hours. There was loads of gore with vicious blood suckers attacking mortals and even each other for control of the world. A good one for all us vampire lovers.
TextFool: I admit I go more for the traditional vampire tael but this one is a real winer lots of way out graphics and good story to go with them made for an intersting 2 hours there was loads of gore with vicious blood suckers attackng mortals and each other for control of the world a good one for all us vampire lovers
Proposed: I admit I go more for the traditional vampire tale but this one is a real whiner lots of way out graphics and good story to go with them made for an 2 hours there was loads of gore with horrible blood suckers attacking mortals and even each other for control of the world a good one for all us vampire lovers

IMDB negative class sample data
Original: A sprawling, overambitious, plotless comedy that has no dramatic center. It was probably intended to have an epic vision and a surrealistic flair (at least in some episodes), but the separate stories are never elevated into a meaningful whole, and the laughs are few and far between. Amusing ending though.
TextFool: A sprawling overambitious plotless horroible that has no dramatic center it was probably intended to have an fail vision and a surrealistic fair at least in some episodes but the separate stories are elevated into a false meaningful whole and the laughs are few and far between amusing ending though
Proposed: A sprawling overambitious plotless funny that has no dramatic center it was probably intended to have an epic vision and a surrealistic flair at least in some episodes but the separate stories are never elevated into a greatly meaningful whole and the laughs are little and far between amusing ending though.

[7] Samanta, Suranjana, and Sameep Mehta. 2017. "Towards Crafting Text Adversarial Samples." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1707.02812.

Graph



[8] Zügner, Daniel, Amir Akbarnejad, and Stephan Günnemann. 2018. "Adversarial Attacks on Neural Networks for Graph Data." In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2847–56. KDD '18. New York, NY, USA: Association for Computing Machinery.

Conclusion

- Adversarial Machine Learning is a trending topic in not only academia but also industry.

- Research directions:
  - Present adversarial attacks/defenses in new data types.
  - Design stronger attacks to evaluate the robustness of the existing systems.
  - Develop adversarial defenses in real-life scenarios.

THANK YOU FOR YOUR ATTENTATION