ADVERSARIAL MACHINE LEARNING

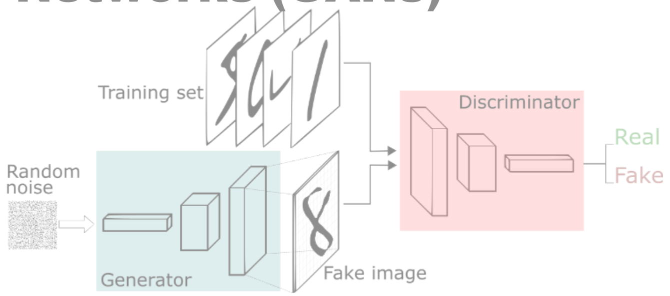
Martin Dallinger
martin.dallinger@outlook.com

06.06.2024

JOHANNES KEPLER
UNIVERSITÄT LINZ

ADVERSARIAL MACHINE LEARNING

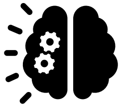
Generative Adversarial Neural Networks (GANs)



Taken from [2]

Adversary: Opposing training

Adversarial Attacks/Examples



Speedlimit: 45!

Taken from [3]

Adversary: Fool models (malicious) with small perturbations

[2] https://sthalles.github.io/intro-to-gans/
[3] Eykholt et al., Robust Physical-World Attacks on Deep Learning Visual Classification; DOI: 10.1109/CVPR.2018.00175

QUICK CATEGORIZATION [4]

- White-Box:
  - Known model structure and weights
  - Opposite: Black-Box

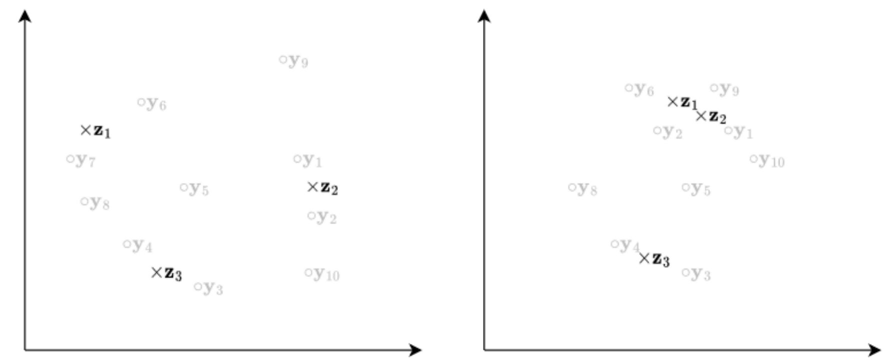
- Label-Targeted:
  - Determine precise class to misclassify to
  - Opposite: Untargeted

Adapted from [1]

[1] Goodfellow et al., Explaining and Harnessing Adversarial Examples; DOI: 10.48550/arXiv.1412.6572
[4] Rosenberg et al., Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain; DOI: 10.1145/3453158

OTHER ATTACKS

- Data Poisoning [6]
  - Change training data (labels)
- Model Poisoning [6]
  - Example: Federated learning
- “Configuration Poisoning“
  - Lesser-known
  - Commonly modify distance metrics
  - Relaxation of triangle inequality for clustering (pseudometric) [5]



(a) Proximity according to Euclidean distance

(b) Proximity according to some other metric

Taken from [5]

[5] Rass et al., Metricizing the Euclidean Space towards Desired Distance Relations in Point Clouds; DOI: 10.48550/arXiv.2211.03674
[6] Fang et al., Local Model Poisoning Attacks to Byzantine-Robust Federated Learning; DOI: 10.5555/3489212.3489304
[15] Image source: Goodfellow et al., https://github.com/cleverhans-lab/cleverhans

NOVEL METHODS FOOL GPT4-V/BARD

- Simple manipulated stop-signs did not work for these models…
- But recent technique (Ensembles, novel „Common Weakness Attack“) manage to fool popular models [9] – 06.04.2024, GPT4-V
  - ☐ Cat:

  - ☐ Hands:

  - ☐ Hand holding mobile phone:

- BUT: Gemini resisted! – 06.04.2024

[9] Chen et al., Rethinking Model Ensemble in Transfer-Based Adversarial Attacks; DOI: 10.48550/arXiv.2303.09105

NOT LIMITED TO IMAGES

Taken from [15]

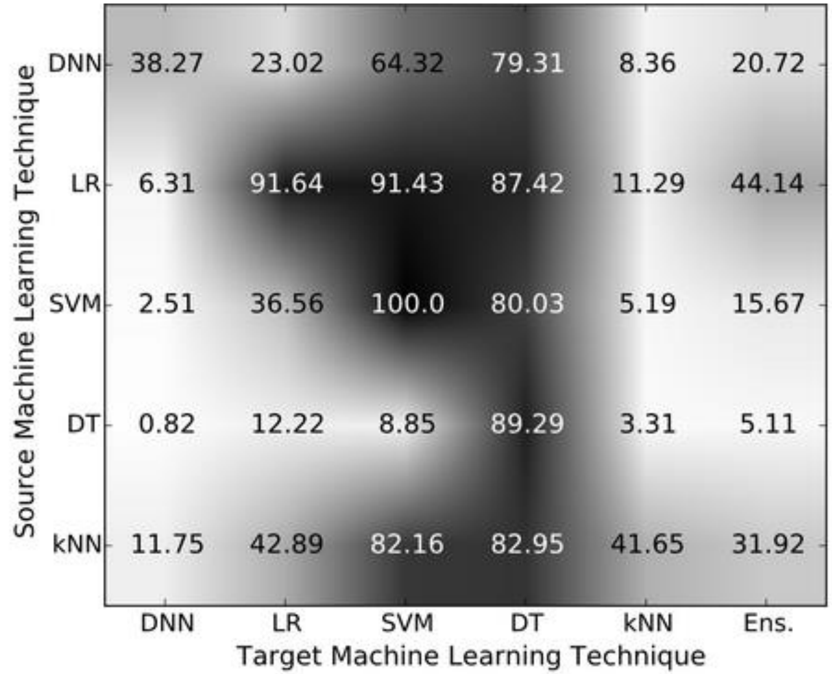
- Different distributions to training in general!
  - Antivirus / EDR Systems [10]
  - Firewalls (even black-box: [11])
- Spaces of adversarial examples
  - Adding randomness stays adversarial
- Already problematic for simple networks (logistic regression) [1]
  - Not due to overfitting!
  - Also: Regularization (dropout, L2-reg, …) does not help

[1] Goodfellow et al., Explaining and Harnessing Adversarial Examples; DOI: 10.48550/arXiv.1412.6572
[10] Jakhotiya et al., Adversarial Attacks on Transformers-Based Malware Detectors; DOI: 10.48550/arXiv.2210.00008
[11] Usama et al., Black-box Adversarial Machine Learning Attack on Network Traffic Classification; DOI: 10.1109/IWCMC.2019.8766505

TRANSFERABILITY

DNN: Deep Neural Networks
LR: Logistic Regression
SVM: Support Vector Machines
DT: Decision Trees
kNN: k Nearest Neighbours
Ens.: Ensembles (not precisely stated…)

- Cross-Technique Transferability using „Substitute Models“
- Even Deep- vs. Non-Deep-Learning!
- Substitute model might help to evade AI-firewalls [11]
  - Difficulty lies in valid packet construction (CRC)



Adversarial sample transferability taken from [13]

[11] Usama et al., Black-box Adversarial Machine Learning Attack on Network Traffic Classification; DOI: 10.1109/IWCMC.2019.8766505
[13] Papernot et al., Transferability in Machine Learning: from Phenomena to Black-Box Attacks Using Adversarial Samples; DOI: 10.48550/arXiv.1605.07277

BASIC ATTACK STRATEGIES [12]

$\mathbf{x}$**:** Input vector
ϵ: Step-Size
$\nabla_{\mathbf{x}} \boldsymbol{L}(\mathbf{x}, \boldsymbol{\Theta})$: Loss wrt. Input
$\mathbf{C}(\mathbf{x})$**:** Classifier parameterized by weights $\boldsymbol{\Theta}$
l: Some different class than intended

- Fast Gradient Sign Method (FGSM)
  - Essentially gradient ascent
  - Simple and effective

$$\boldsymbol{x}^* = \boldsymbol{x} + \epsilon \, \nabla_{\boldsymbol{x}} \, \boldsymbol{L}(\boldsymbol{x}, \boldsymbol{\Theta})$$

- Basic Iterative Method (BIM)
  - Multiple FGSM
  - Clipping to stay in ϵ-Neighborhood
- L-BFGS Method
  - Box-constrained optimization problem
  - Quasi-Newton method
- Many more complex and recent techniques [9]

$$\begin{aligned} &\text{minimize} \quad \|x_0 - x\|_2^2 \\ &\text{such that} \quad C(x) = l \\ &\text{where} \quad x \in [0,1]^p \end{aligned}$$

[9] Chen et al., Rethinking Model Ensemble in Transfer-Based Adversarial Attacks; DOI: 10.48550/arXiv.2303.09105
[12] Papernot et al., Technical Report on the Cleverhans v2.1.0 Adversarial Examples Library; DOI: 10.48550/arXiv.1610.00768

DEFENSES [10]

- Adversarial training: Also improves performance [1]
- Forcing black box attacks
- Ensembles [13] of simpler architectures if possble (Random Forests)
  - No gradients
  - Or more non-linear architectures: RBF [1]
- Reduce feature space [1]
- Certified robustness [14]

- Test with https://github.com/cleverhans-lab/cleverhans

[1] Goodfellow et al., Explaining and Harnessing Adversarial Examples; DOI: 10.48550/arXiv.1412.6572
[10] Jakhotiya et al., Adversarial Attacks on Transformers-Based Malware Detectors; DOI: 10.48550/arXiv.2210.00008
[13] Papernot et al., Transferability in Machine Learning: From Phenomena to Black-Box Attacks Using Adversarial Samples; DOI: 10.48550/arXiv.1605.07277
[14] Chen et al., Certifying Robustness of Neural Networks With a Probabilistic Approach; DOI: 10.48550/arXiv.1812.08329

REFERENCES [1-4]

- [1] Goodfellow, I.J., Shlens, J., & Szegedy, C. (2014). Explaining and Harnessing Adversarial Examples. *CoRR, abs/1412.6572*.
- *[2] Image source: https://sthalles.github.io/intro-to-gans/*
- [3] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., & Song, D.X. (2018). Robust Physical-World Attacks on Deep Learning Visual Classification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1625-1634.
- [4] Rosenberg, I., Shabtai, A., Elovici, Y., & Rokach, L. (2021). Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain. *ACM Computing Surveys (CSUR), 54*, 1 - 36.

REFERENCES [5-8]

- [5] Rass, S., König, S., Ahmad, S., & Goman, M. (2022). Metricizing the Euclidean Space towards Desired Distance Relations in Point Clouds. *ArXiv, abs/2211.03674*.
- [6] Fang, M., Cao, X., Jia, J., & Gong, N.Z. (2019). Local Model Poisoning Attacks to Byzantine-Robust Federated Learning. *USENIX Security Symposium*.
- [7] Su, J., Vargas, D.V., & Sakurai, K. (2017). One Pixel Attack for Fooling Deep Neural Networks. *IEEE Transactions on Evolutionary Computation, 23*, 828-841.
- [8] Xu, K., Zhang, G., Liu, S., Fan, Q., Sun, M., Chen, H., Chen, P., Wang, Y., & Lin, X. (2019). Adversarial T-Shirt! Evading Person Detectors in a Physical World. *European Conference on Computer Vision*.

REFERENCES [9-12]

- [9] Chen, H., Zhang, Y., Dong, Y., & Zhu, J. (2023). Rethinking Model Ensemble in Transfer-based Adversarial Attacks. *ArXiv, abs/2303.09105*.
- [10] Jakhotiya, Y., Patil, H., & Rawlani, J. (2022). Adversarial Attacks on Transformers-Based Malware Detectors. *ArXiv, abs/2210.00008*.
- [11] Usama, M., Qayyum, A., Qadir, J., & Al-Fuqaha, A. (2019). Black-box Adversarial Machine Learning Attack on Network Traffic Classification. *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*, 84-89.
- [12] Papernot, et al. (2016). Technical Report on the CleverHans v2.1.0 Adversarial Examples Library. *arXiv: Learning*.

REFERENCES [13-15]

- [13] Papernot, N., Mcdaniel, P., & Goodfellow, I.J. (2016). Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples. *ArXiv, abs/1605.07277*.
- [14] Weng, T., Chen, P., Nguyen, L.M., Squillante, M.S., Oseledets, I., & Daniel, L. (2018). PROVEN: Certifying Robustness of Neural Networks with a Probabilistic Approach. *ArXiv, abs/1812.08329*.
- *[15] Image source: Goodfellow et al., https://github.com/cleverhans-lab/cleverhans*

CONCLUSION

SLIDES

- Be aware of (infinitely many) adversarial examples
  - Especially black box and transfer
- Harden/Test your systems (cleverhans)
- Cat and mouse game

- Questions?

Get in touch: Martin Dallinger, martin.dallinger@outlook.com

APPENDIX

06.06.2024

CERTIFIED ROBUSTNESS [14]

- „Small changes“ should have the same class
- Certifiability to some epsilon-pertubation
- Using Lipschitz constrained networks

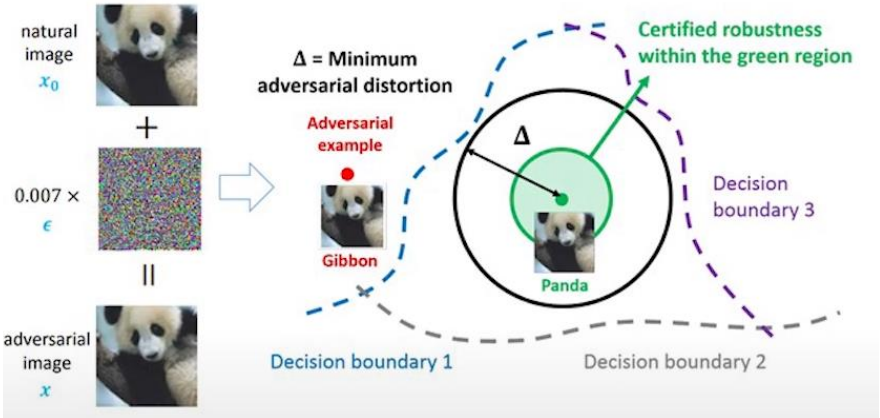
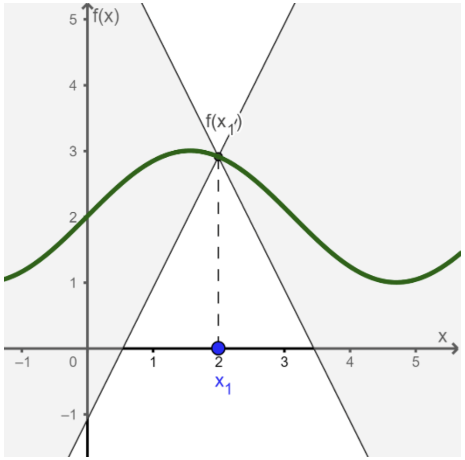


Image taken from [15]

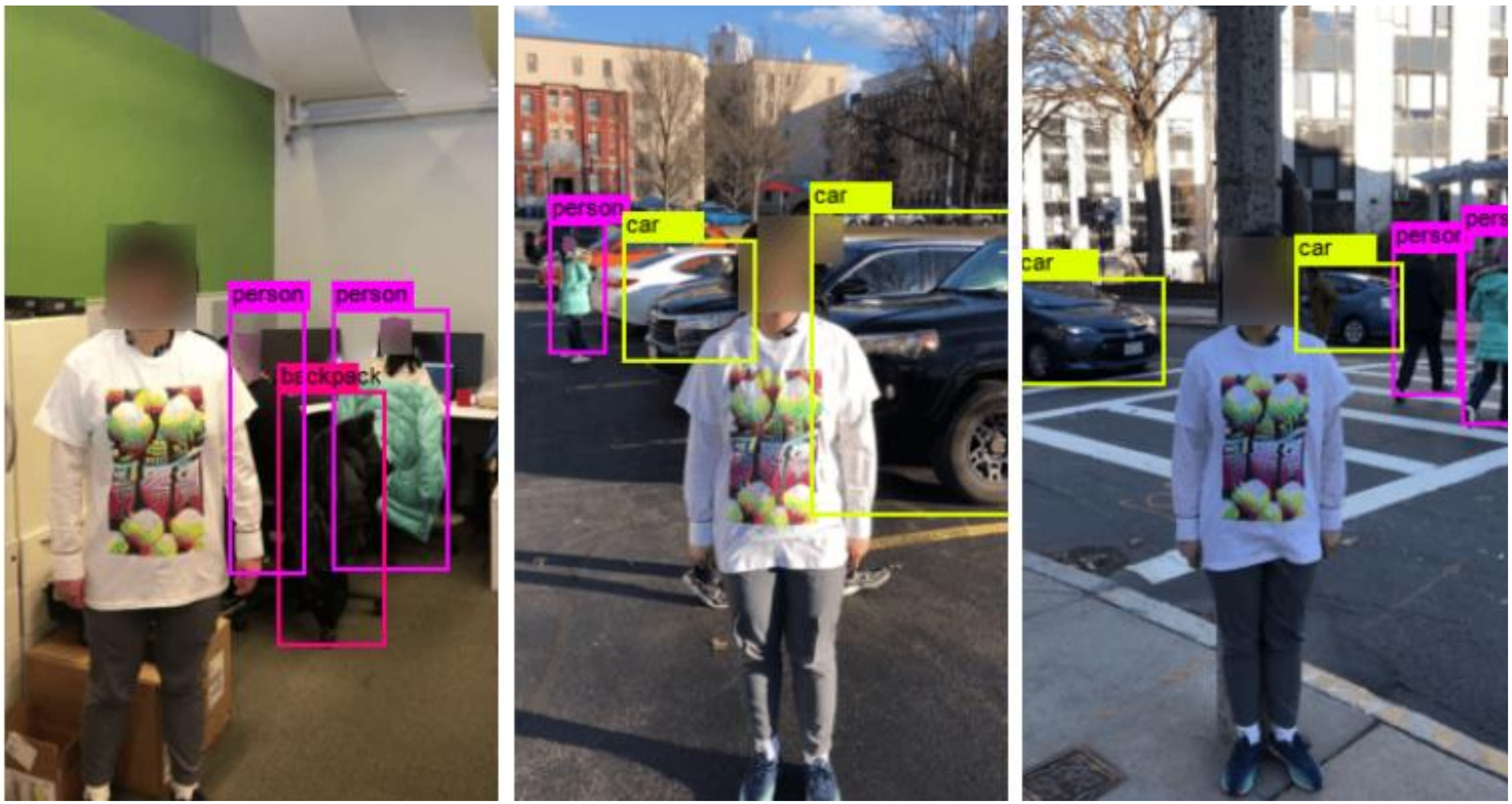


Lipschitz-Continuity as shown in https://www.geogebra.org/m/bnsymjxh

- Problem: Measure distance
  - ☐ Semantic meaning (distinguish 8 vs 0 visually)

[14] Chen et al., Certifying Robustness of Neural Networks With a Probabilistic Approach; DOI: 10.48550/arXiv.1812.08329
[15] Nick Frosst, Certifiable Robustness to Adversarial Attacks (Toronto ML Summit); https://www.youtube.com/watch?v=OfSxYqU-6s0&t=242s

ADVERSARIAL T-SHIRT AGAINST YOLO-V2 [8]



Taken from [8]

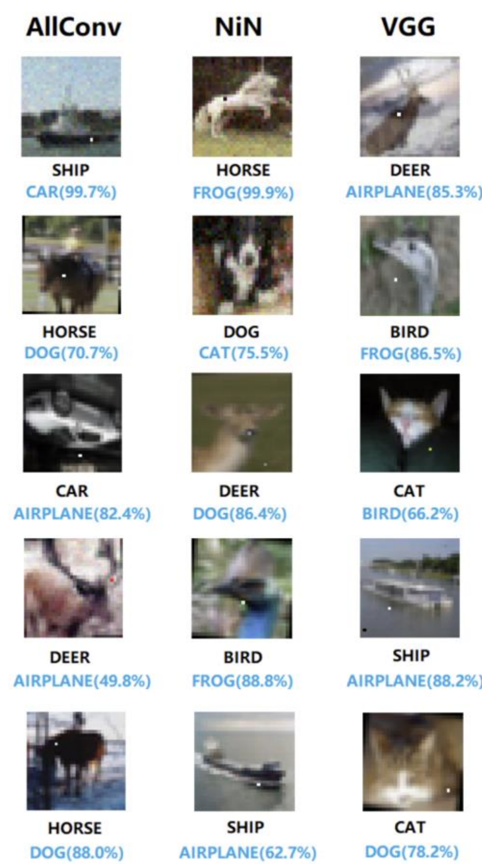
[8] Xu et al., Adversarial T-shirt! Evading Person Detectors in A Physical World; DOI: 10.1007/978-3-030-58558-7_39

NOT UNIQUE/HARD TO FIND

Table 1: Sample of physical adversarial examples against LISA-CNN and GTSRB-CNN.

Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5′ 0°					
5′ 15°					
10′ 0°					
10′ 30°					
40′ 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%

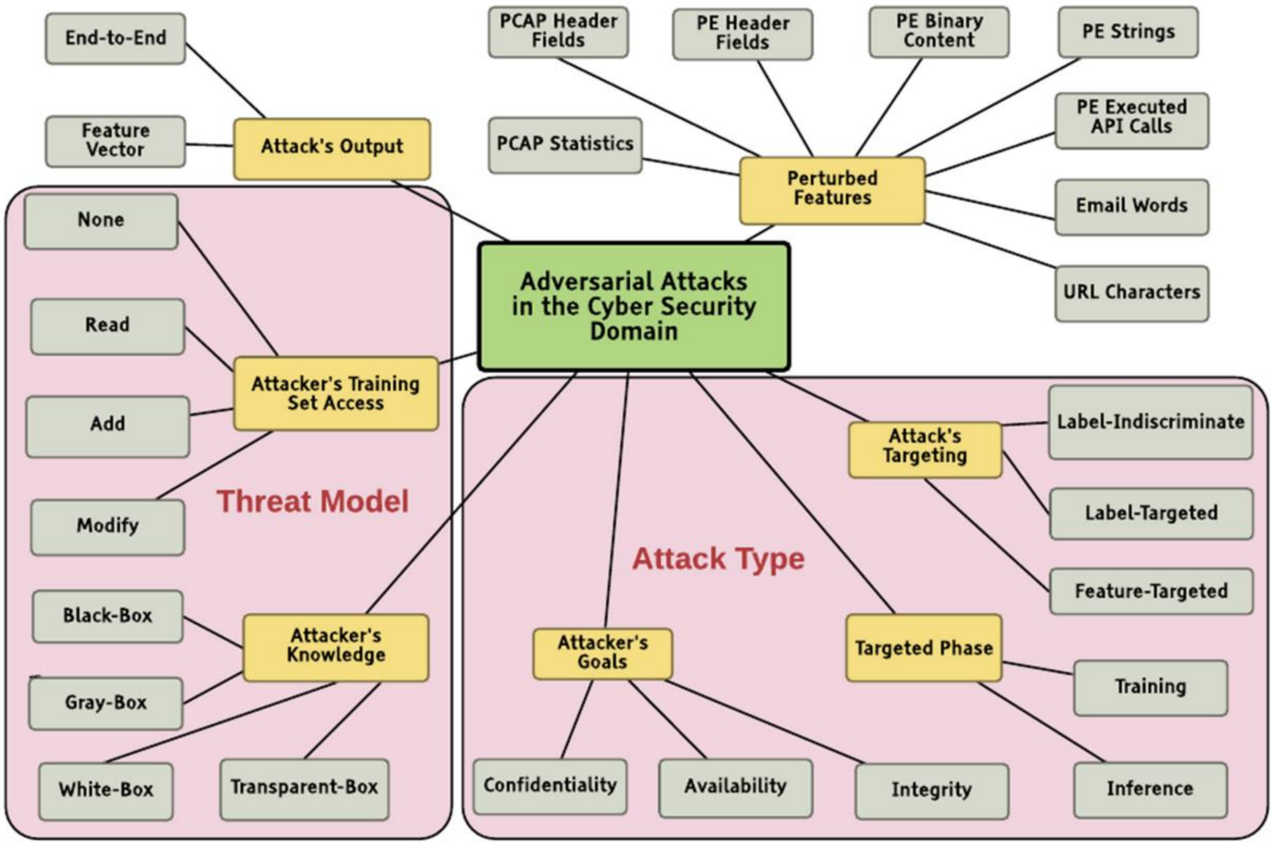
Taken from [3]



Taken from [7]

[3] Eykholt et al., Robust Physical-World Attacks on Deep Learning Visual Classification; DOI: 10.1109/CVPR.2018.00175
[7] Su et al., One Pixel Attack for Fooling Deep Neural Networks; DOI: 10.48550/arXiv.1710.08864

FULL CATEGORIZATION [4]



Taken from [4]

[4] Rosenberg et al., Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain; DOI: 10.1145/3453158