Towards Evaluating the Robustness of Neural Networks

Authors:
Dr Nicholas Carlini and Prof David Wagner

Presenter:
Zunnoor Fayyaz Awan

About this paper

Reviewers’ comments (Euro S&P)

- “no real insights”
- “makes very limited contributions”
- “does not provide a systematic approach”

Carlini (2022), *Reflecting on “Towards Evaluating the Robustness of Neural Networks”*

About this paper

Reviewers’ comments (Euro S&P)

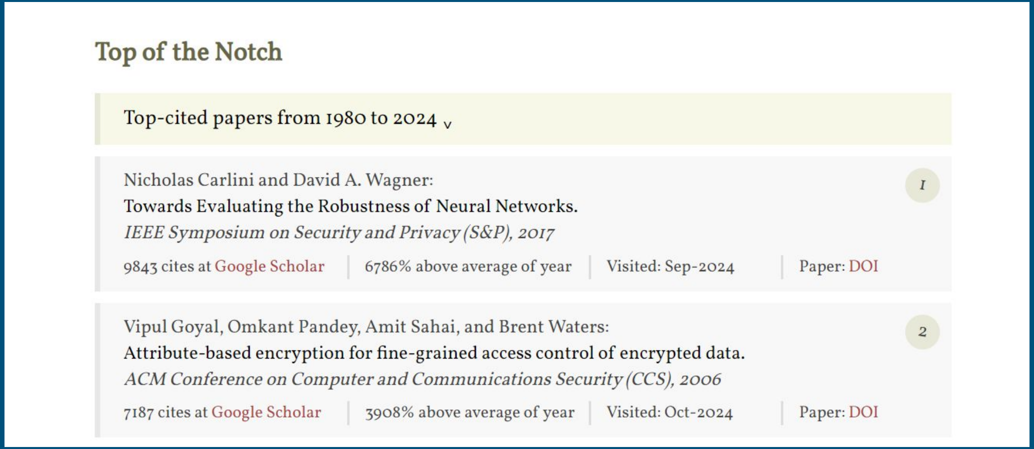
- “no real insights”
- “makes very limited contributions”
- “does not provide a systematic approach”
- “this is not a security paper” (?)

Carlini (2022), *Reflecting on “Towards Evaluating the Robustness of Neural Networks”*

Welcome!

More about this paper

- IEEE Security and Privacy (2017)
- All-time highest cited paper in security



Top of the Notch

Top-cited papers from 1980 to 2024 ⌄

Nicholas Carlini and David A. Wagner:
Towards Evaluating the Robustness of Neural Networks.
IEEE Symposium on Security and Privacy (S&P), 2017
9843 cites at Google Scholar | 6786% above average of year | Visited: Sep-2024 | Paper: DOI

1

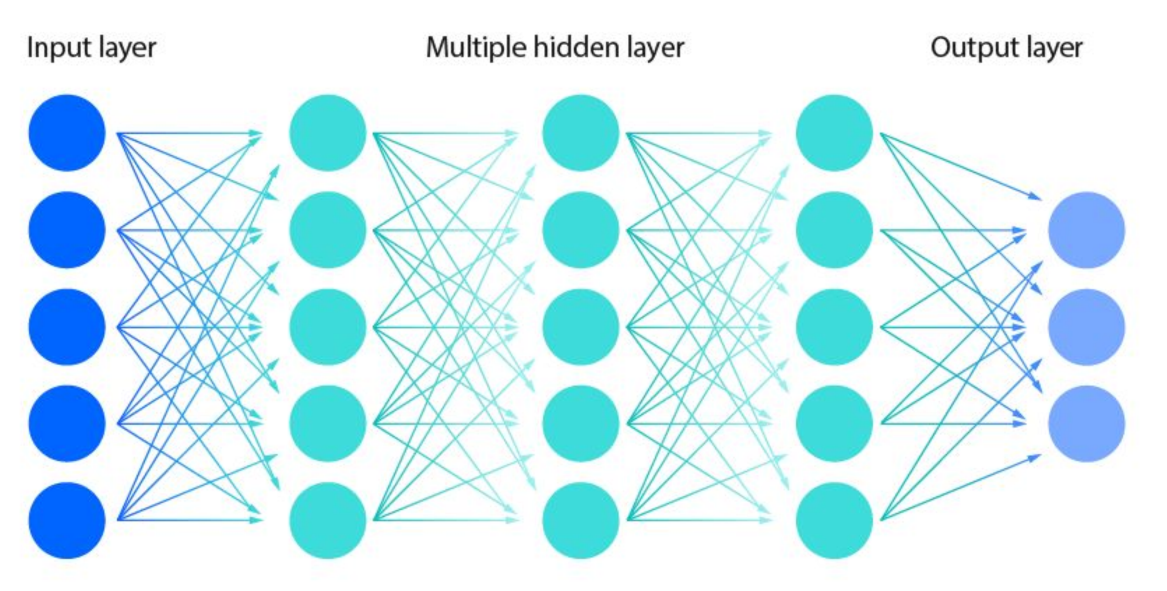
Vipul Goyal, Omkant Pandey, Amit Sahai, and Brent Waters:
Attribute-based encryption for fine-grained access control of encrypted data.
ACM Conference on Computer and Communications Security (CCS), 2006
7187 cites at Google Scholar | 3908% above average of year | Visited: Oct-2024 | Paper: DOI

2

mlsec.org/topnotch/index.html

Towards Evaluating the Robustness of Neural Networks

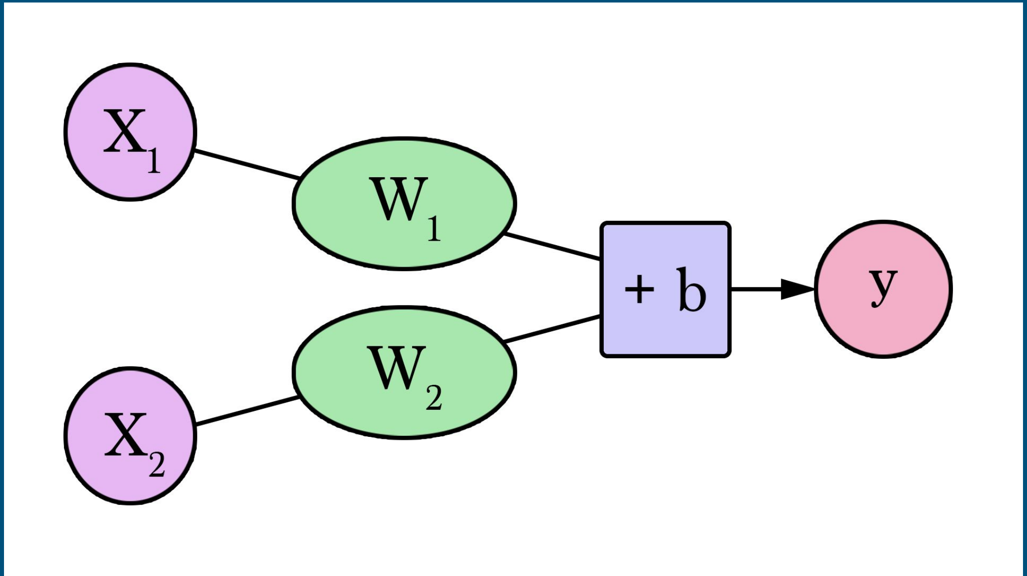
Neural Network



https://www.ibm.com/topics/neural-networks

Neural Network (simplified)

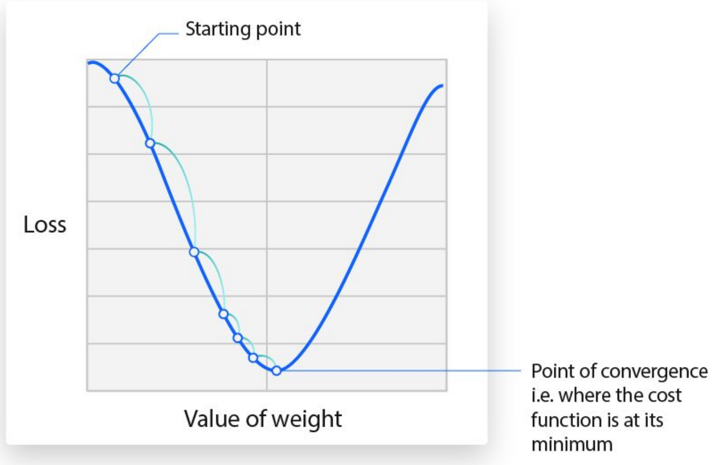
- Parameterized function
  - weight: scaling factor
  - bias: offset
- $y = X_1W_1 + X_2W_2 + b$



Alammar, J (2018). *The Illustrated Transformer* (jalammar.github.io/illustrated-transformer)

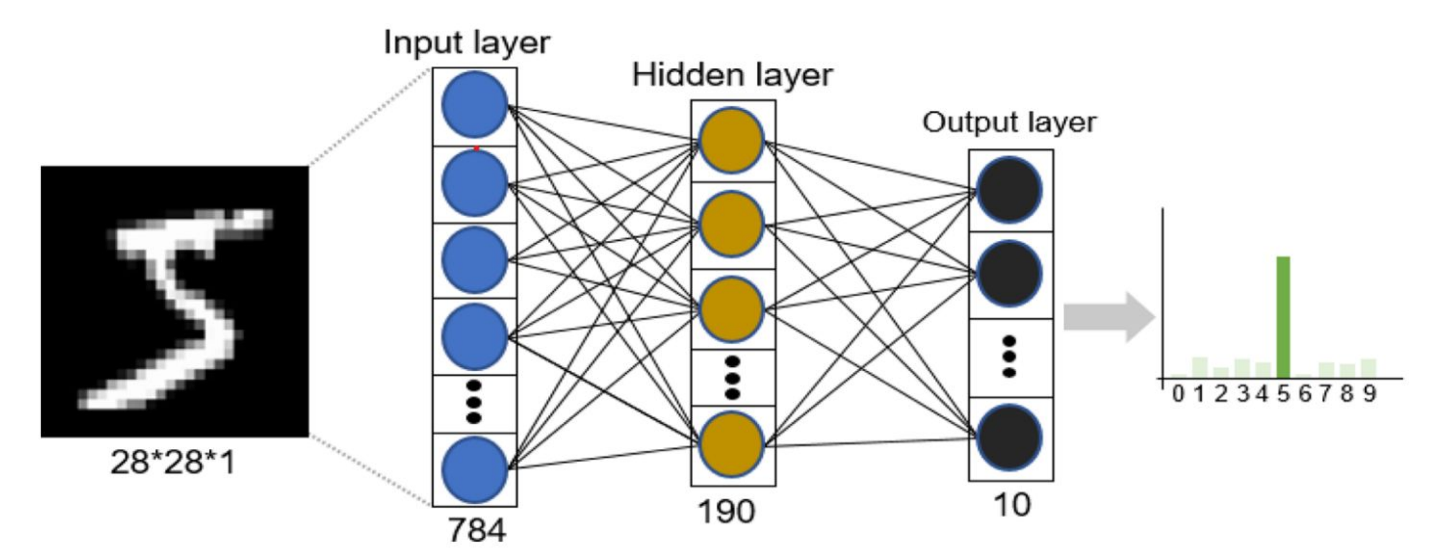
Training (gradient descent)

- Loss: difference between current output and desired output
- Training: tuning parameters to reduce loss



https://www.ibm.com/topics/neural-networks

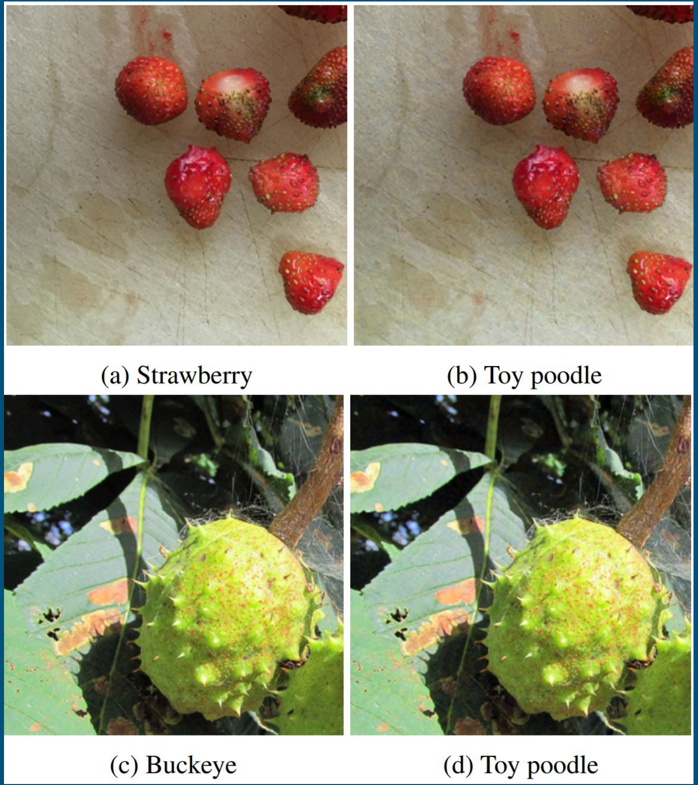
Classification



Teng et al (2023). *Nonlinear absorption of 2D materials and their application in optical neural networks.*

Adversarial examples

- Intentionally adding tiny perturbations to an input to cause misclassification



(a) Strawberry

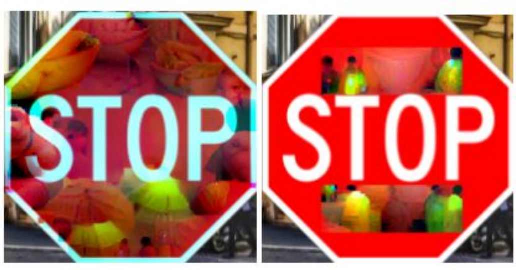
(b) Toy poodle

(c) Buckeye

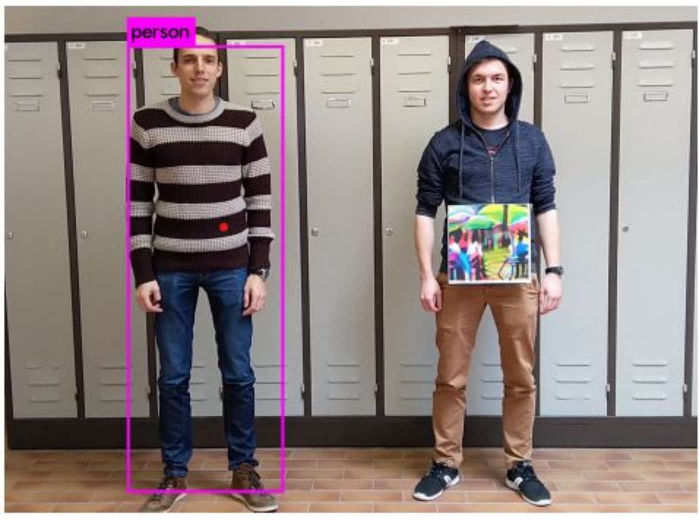
(d) Toy poodle

Xiao et al (2018). *Generating adversarial examples with adversarial networks* (IJCAI'18)

Adversarial examples



[1]

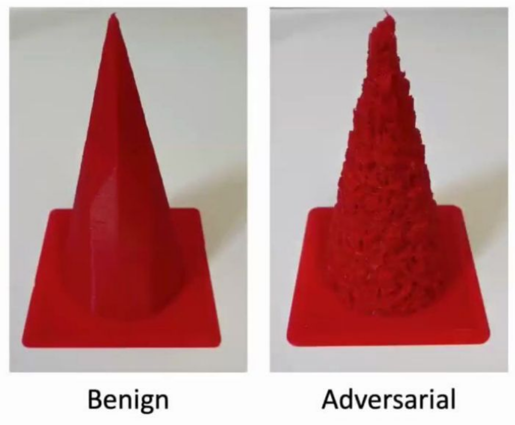


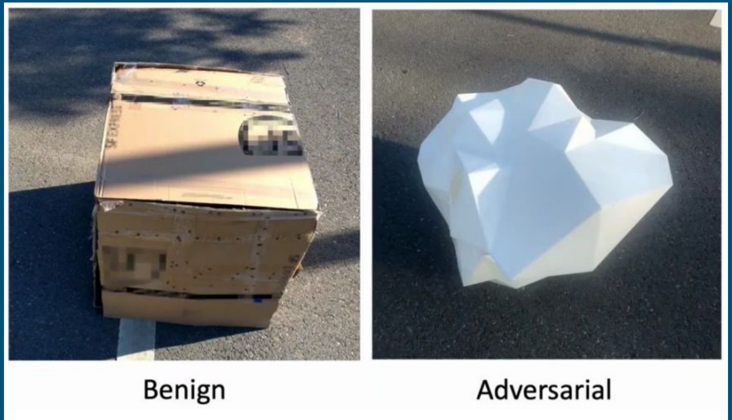
[2]

[1] Eykholt et al (2018). *Physical Adversarial Examples for Object Detectors* (WOOT'18)
[2] Thys et al (2019). *Fooling Automated Surveillance Cameras: Adversarial Patches to Attack Person Detection* (CVPRW)

Adversarial examples already seen

- Targeting autonomous vehicle perception in the multi-sensor fusion paper





Cao et al (2021). *Invisible for both Camera and LiDAR: Security of Multi-Sensor Fusion based Perception in Autonomous Driving Under Physical-World Attacks.* (IEEE S&P)

Robustness

- Property of a model
- A robust model maintains performance despite variations in input
- Difficult to find adversarial examples for robust models

Adversarial ML

- Attacks:
  efficient techniques to identify adversarial examples for a model
- Defenses:
  make a model more robust to input perturbations
- Timeline:
  - 2004: earliest attack (against ML-based spam filters)
  - 2014 onwards: attacks against deep neural networks

2017 landscape

- Many defensive techniques had been proposed

1. Huang, R., Xu, B., Schuurmans, D., and Szepesvári, C. Learning with a strong adversary. CoRR, abs/1511.03034 (2015)
2. Jin, J., Dundar, A., and Culurciello, E. Robust convolutional neural networks under adversarial noise. arXiv preprint arXiv:1511.06306 (2015)
3. Papernot, N., McDaniel, P., Wu, X., Jha, S., and Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. IEEE Symposium on Security and Privacy (2016)
4. Hendrycks, D., and Gimpel, K. Visible progress on adversarial images and a new saliency map. arXiv preprint arXiv:1608.00530 (2016) Li, X., and Li, F. Adversarial examples detection in deep networks with convolutional filter statistics. arXiv preprint arXiv:1612.07767 (2016)
5. Wang, Q. et al. Using Non-invertible Data Transformations to Build Adversary-Resistant Deep Neural Networks. arXiv preprint arXiv:1610.01934 (2016).
6. Ororbia, I. I., et al. Unifying adversarial training algorithms with flexible deep data gradient regularization. arXiv preprint arXiv:1601.07213 (2016).
7. Wang, Q. et al. Learning Adversary-Resistant Deep Neural Networks. arXiv preprint arXiv:1612.01401 (2016).
8. Grosse, K., Manoharan, P., Papernot, N., Backes, M., and McDaniel, P. On the (statistical) detection of adversarial examples. arXiv preprint arXiv:1702.06280 (2017)
9. Metzen, J. H., Genewein, T., Fischer, V., and Bischoff, B. On detecting adversarial perturbations. arXiv preprint arXiv:1702.04267 (2017)
10. Feinman, R., Curtin, R. R., Shintre, S., Gardner, A. B. Detecting Adversarial Samples from Artifacts. arXiv preprint arXiv:1703.00410 (2017)
11. Zhitao Gong, Wenlu Wang, and Wei-Shinn Ku. Adversarial and Clean Data Are Not Twins. arXiv preprint arXiv:1704.04960 (2017)
12. Dan Hendrycks and Kevin Gimpel. Early Methods for Detecting Adversarial Images. In International Conference on Learning Representations (Workshop Track) (2017)
13. Bhagoji, A. N., Cullina, D., and Mittal, P. Dimensionality Reduction as a Defense against Evasion Attacks on Machine Learning Classifiers. arXiv preprint arXiv:1704:02654 (2017)
14. Abbasi, M., and Christian G.. Robustness to Adversarial Examples through an Ensemble of Specialists. arXiv preprint arXiv:1702.06856 (2017).
15. Lu, J., Theerasit I., and David F. SafetyNet: Detecting and Rejecting Adversarial Examples Robustly. arXiv preprint arXiv:1704.00103 (2017)
16. Xu, W., Evans, D., and Qi, Y. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. arXiv preprint arXiv:1704.01155 (2017)
17. Hendrycks, D, and Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. arXiv preprint arXiv:1610.02136 (2016)
18. Gondara, Lovedeep. Detecting Adversarial Samples Using Density Ratio Estimates. arXiv preprint arXiv:1705.02224 (2017)
19. Hosseini, Hossein, et al. Blocking transferability of adversarial examples in black-box learning systems. arXiv preprint arXiv:1703.04318 (2017)
20. Ji Gao, Beilun Wang, Zeming Lin, Weilin Xu, Yanjun Qi. DeepCloak: Masking Deep Neural Network Models for Robustness Against Adversarial Samples. In ICLR (Workshop Track) (2017)
21. Wang, Q. et al. Adversary Resistant Deep Neural Networks with an Application to Malware Detection. arXiv preprint arXiv:1610.01239 (2017)
22. Cisse, Moustapha, et al. Parseval Networks: Improving Robustness to Adversarial Examples. arXiv preprint arXiv:1704.08847 (2017).
23. Nayebi, Aran, and Surya Ganguli. Biologically inspired protection of deep networks from adversarial attacks. arXiv preprint arXiv:1703.09202 (2017).

2017 landscape

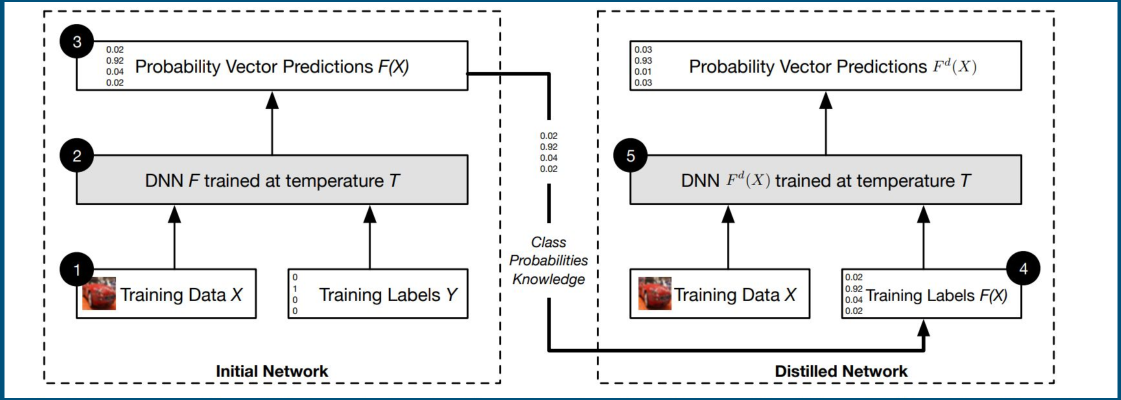
considered the strongest

- Many defensive techniques had been proposed

1. Huang, R., Xu, B., Schuurmans, D., and Szepesvári, C. Learning with a strong adversary. CoRR, abs/1511.03034 (2015)
2. Jin, J., Dundar, A., and Culurciello, E. Robust convolutional neural networks under adversarial noise. arXiv preprint arXiv:1511.06306 (2015)
3. Papernot, N., McDaniel, P., Wu, X., Jha, S., and Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. IEEE Symposium on Security and Privacy (2016)
4. Hendrycks, D., and Gimpel, K. Visible progress on adversarial images and a new saliency map. arXiv preprint arXiv:1608.00530 (2016) Li, X., and Li, F. Adversarial examples detection in deep networks with convolutional filter statistics. arXiv preprint arXiv:1612.07767 (2016)
5. Wang, Q. et al. Using Non-invertible Data Transformations to Build Adversary-Resistant Deep Neural Networks. arXiv preprint arXiv:1610.01934 (2016).
6. Ororbia, I. I., et al. Unifying adversarial training algorithms with flexible deep data gradient regularization. arXiv preprint arXiv:1601.07213 (2016).
7. Wang, Q. et al. Learning Adversary-Resistant Deep Neural Networks. arXiv preprint arXiv:1612.01401 (2016).
8. Grosse, K., Manoharan, P., Papernot, N., Backes, M., and McDaniel, P. On the (statistical) detection of adversarial examples. arXiv preprint arXiv:1702.06280 (2017)
9. Metzen, J. H., Genewein, T., Fischer, V., and Bischoff, B. On detecting adversarial perturbations. arXiv preprint arXiv:1702.04267 (2017)
10. Feinman, R., Curtin, R. R., Shintre, S., Gardner, A. B. Detecting Adversarial Samples from Artifacts. arXiv preprint arXiv:1703.00410 (2017)
11. Zhitao Gong, Wenlu Wang, and Wei-Shinn Ku. Adversarial and Clean Data Are Not Twins. arXiv preprint arXiv:1704.04960 (2017)
12. Dan Hendrycks and Kevin Gimpel. Early Methods for Detecting Adversarial Images. In International Conference on Learning Representations (Workshop Track) (2017)
13. Bhagoji, A. N., Cullina, D., and Mittal, P. Dimensionality Reduction as a Defense against Evasion Attacks on Machine Learning Classifiers. arXiv preprint arXiv:1704:02654 (2017)
14. Abbasi, M., and Christian G.. Robustness to Adversarial Examples through an Ensemble of Specialists. arXiv preprint arXiv:1702.06856 (2017).
15. Lu, J., Theerasit I., and David F. SafetyNet: Detecting and Rejecting Adversarial Examples Robustly. arXiv preprint arXiv:1704.00103 (2017)
16. Xu, W., Evans, D., and Qi, Y. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. arXiv preprint arXiv:1704.01155 (2017)
17. Hendrycks, D, and Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. arXiv preprint arXiv:1610.02136 (2016)
18. Gondara, Lovedeep. Detecting Adversarial Samples Using Density Ratio Estimates. arXiv preprint arXiv:1705.02224 (2017)
19. Hosseini, Hossein, et al. Blocking transferability of adversarial examples in black-box learning systems. arXiv preprint arXiv:1703.04318 (2017)
20. Ji Gao, Beilun Wang, Zeming Lin, Weilin Xu, Yanjun Qi. DeepCloak: Masking Deep Neural Network Models for Robustness Against Adversarial Samples. In ICLR (Workshop Track) (2017)
21. Wang, Q. et al. Adversary Resistant Deep Neural Networks with an Application to Malware Detection. arXiv preprint arXiv:1610.01239 (2017)
22. Cisse, Moustapha, et al. Parseval Networks: Improving Robustness to Adversarial Examples. arXiv preprint arXiv:1704.08847 (2017).
23. Nayebi, Aran, and Surya Ganguli. Biologically inspired protection of deep networks from adversarial attacks. arXiv preprint arXiv:1703.09202 (2017).

Defensive Distillation

- Two classifiers are used
- First classifier is trained as usual
- The second learns class probabilities instead of class labels



Papernot et al (2016). *Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks.*

Objective

- C: classifier, x: normal input value
- $C(x)$: the classifier's predicted category for x
- Idea: find small δ such that $C(x + \delta)$ is different from its true category
- $D(x, x + \delta)$: distance (similarity) between x and $x + \delta$
- T: a target category for $x + \delta$
- Goal: minimize $D(x, x + \delta)$ under the constraint $C(x + \delta) = T$

Distance metrics

- To measure $D(x, x + \delta)$ the distance between the original image and the perturbed image, three metrics are used:
    - **L_0**
        - *The number of pixels that have been altered*
    - **L_2**
        - *Euclidean distance (cumulative change over all pixels)*
    - **L_∞**
        - *The maximum change to any one pixel*

Optimization approach

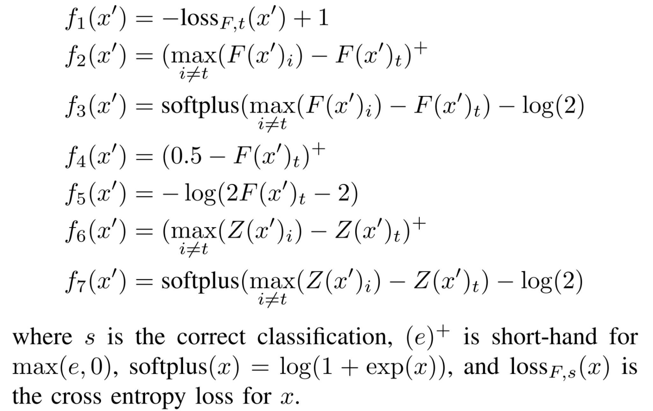
- Gradient descent to find the best δ
- However, the constrained optimization problem $D(x, x + \delta)$ under $C(x + \delta) = T$ proved difficult to solve due to high nonlinearity
- Reformulate into an unconstrained problem

Revised objective

- Category constraint $C(x + \delta) = T$ is replaced by a term in the objective
- f: function to measure the distance of the predicted category from the target category
- Minimize $D(x, x + \delta) + f(x + \delta)$,
  where $f(x + \delta) \leqslant 0$ if $C(x + \delta) = T$ and $f(x + \delta) > 0$ otherwise
- Choice: $f(x + \delta)$ = 1 - (probability that $C(x + \delta) = T$)
- However, this simple choice fails

Choices for *f*

- Select candidates for *f* from the literature:



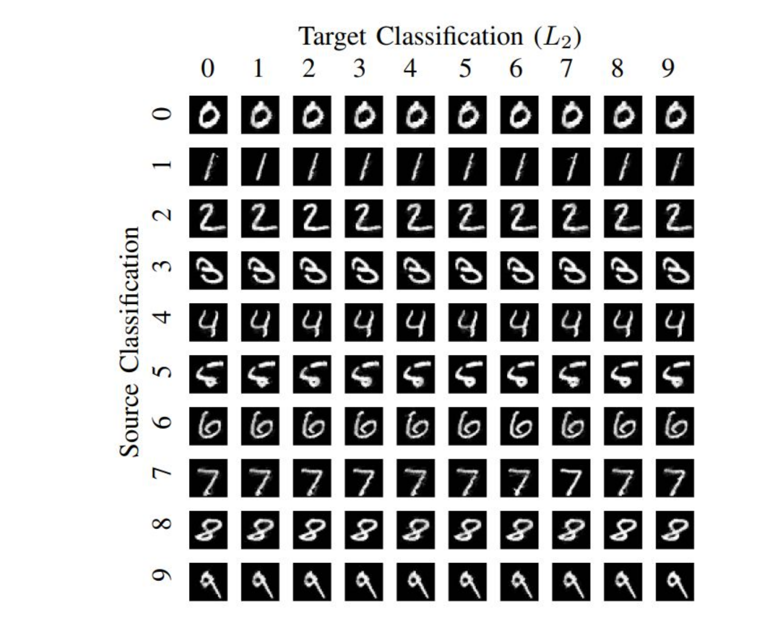
$$
\begin{aligned}
f_1(x') &= -\mathrm{loss}_{F,t}(x') + 1\\
f_2(x') &= (\max_{i \neq t}(F(x')_i) - F(x')_t)^+\\
f_3(x') &= \mathrm{softplus}(\max_{i \neq t}(F(x')_i) - F(x')_t) - \log(2)\\
f_4(x') &= (0.5 - F(x')_t)^+\\
f_5(x') &= -\log(2F(x')_t - 2)\\
f_6(x') &= (\max_{i \neq t}(Z(x')_i) - Z(x')_t)^+\\
f_7(x') &= \mathrm{softplus}(\max_{i \neq t}(Z(x')_i) - Z(x')_t) - \log(2)
\end{aligned}
$$

where s is the correct classification, $(e)^+$ is short-hand for $\max(e, 0)$, $\mathrm{softplus}(x) = \log(1 + \exp(x))$, and $\mathrm{loss}_{F,s}(x)$ is the cross entropy loss for x.

- Strategy: try all 7 and select the best one empirically

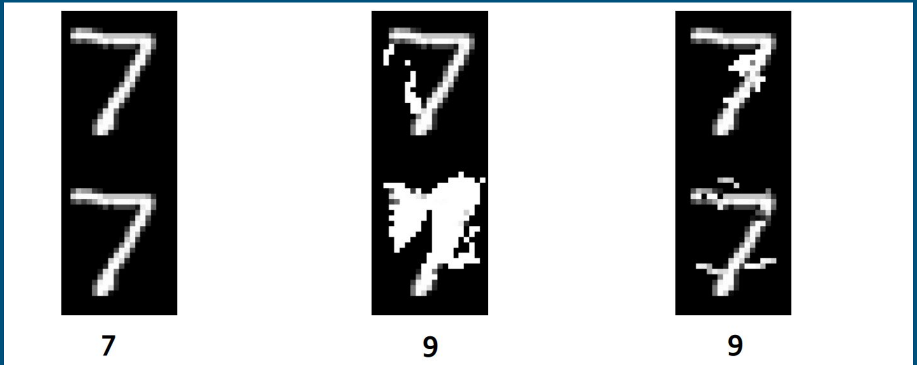
Results

- Adversarial examples were found for every pair of source and target category for MNIST



Results

- Adversarial examples were much better compared to previous attacks

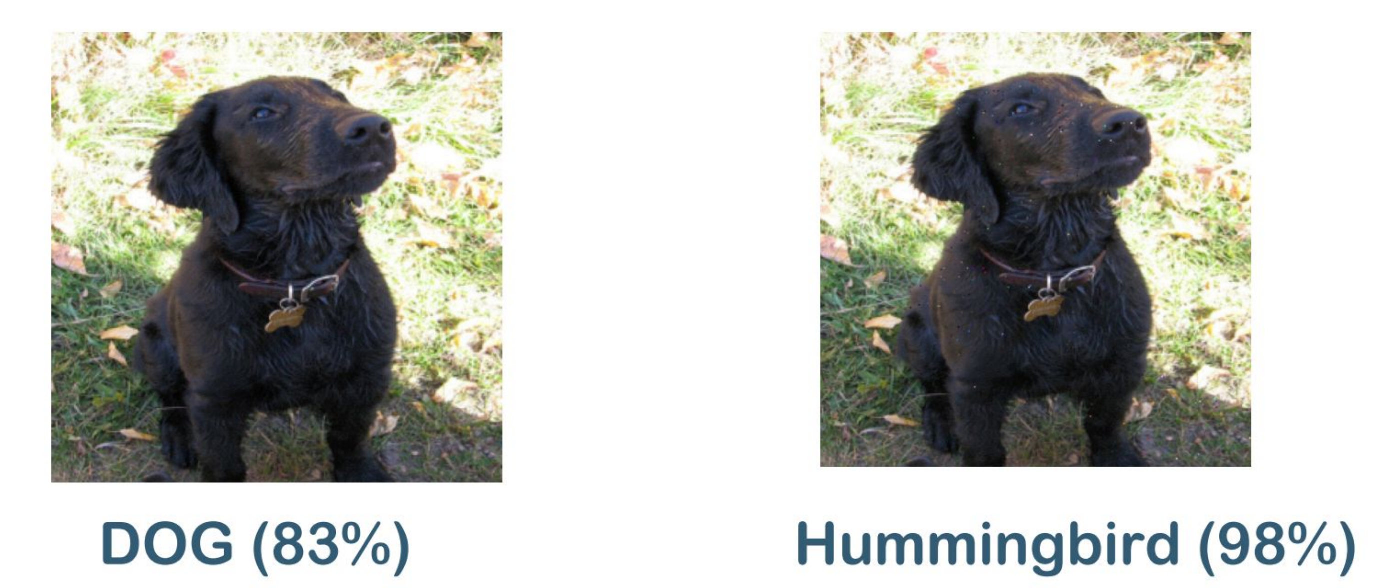


Original

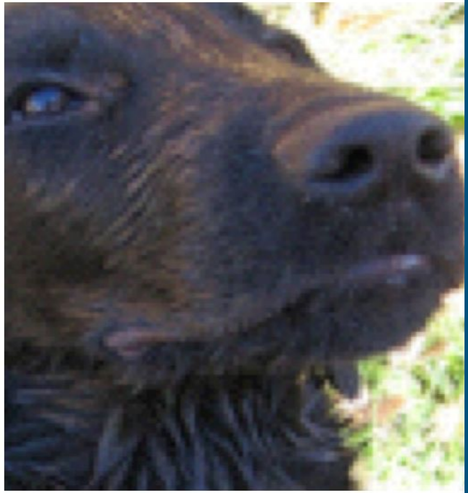
Previous work

This work

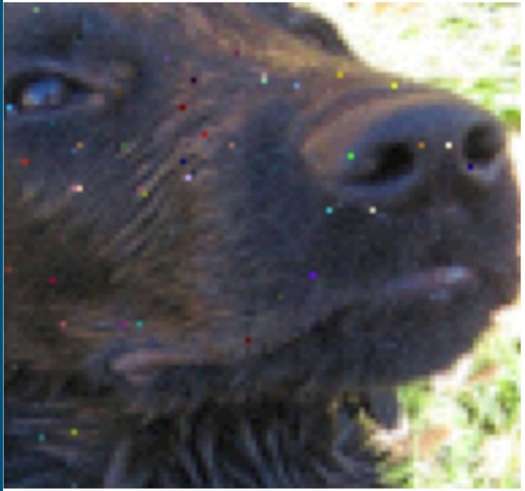
Results



Results



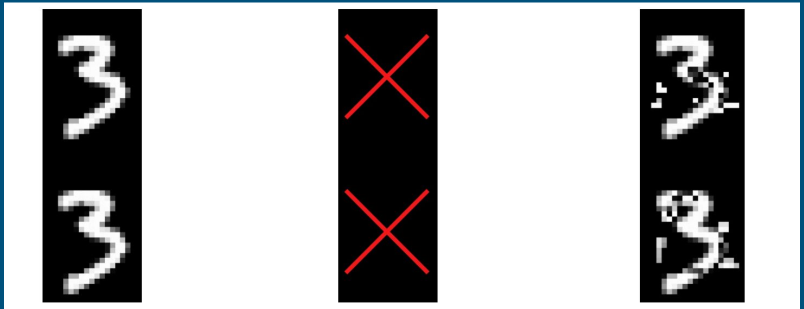
DOG



Hummingbird

Evaluation on distilled models

- Adversarial examples were found for defensively distilled networks each time



Original

Previous work

This work

Conclusion

+ Choice of gradient descent to find adversarial examples turned out well
+ Defeated all defenses at time of publication

- Slower than previous attacks
- White-box model (assumes adversary can access model parameters)

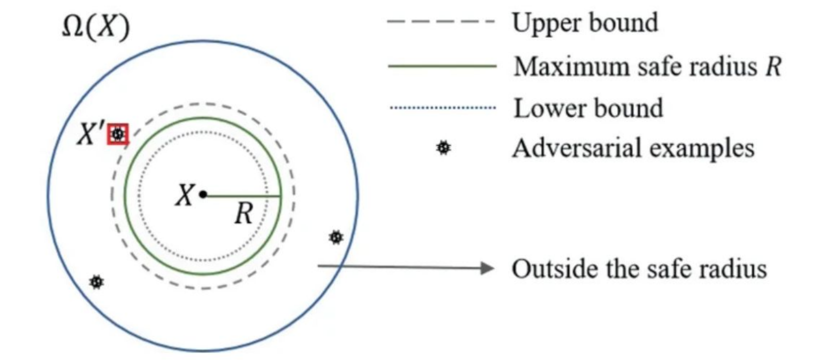
Current defense approaches

- Empirical defenses
  - Remain open to more sophisticated attacks even if successful against current attacks
- Certifiably robust approaches
  - Training and evaluation approaches to establish “lower bound of robust accuracy against any attacks under certain conditions”

Li el al (2023). *SoK: Certified Robustness for Deep Neural Networks.* (IEEE S&P 2023)

Current defense approaches

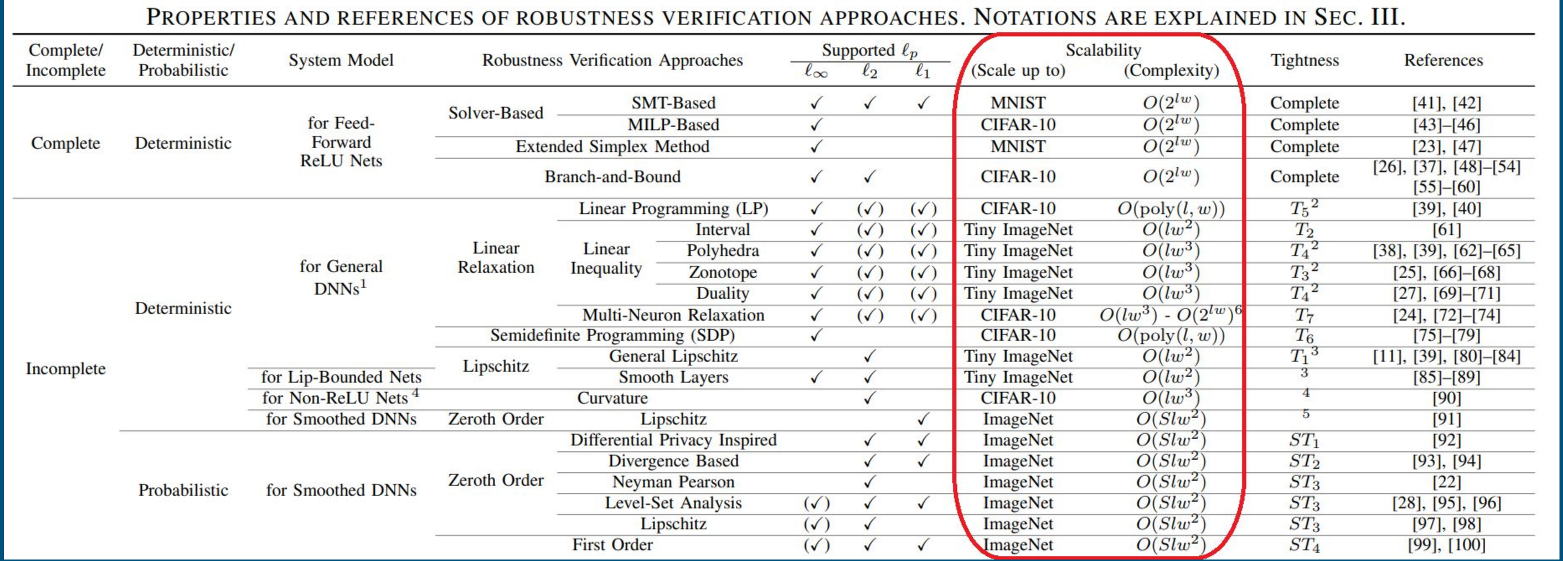
- Safe radius



Yang, et al. (2023). *Quantifying Robustness to Adversarial Word Substitutions*. (ECML PKDD 2023, Lecture Notes in Computer Science, vol 14169, Springer, Cham.)

Limitations of defenses

- Scalability



PROPERTIES AND REFERENCES OF ROBUSTNESS VERIFICATION APPROACHES. NOTATIONS ARE EXPLAINED IN SEC. III.

Complete/ Incomplete	Deterministic/ Probabilistic	System Model	Robustness Verification Approaches			Supported ℓ_p			Scalability		Tightness	References
						ℓ_∞	ℓ_2	ℓ_1	(Scale up to)	(Complexity)		
Complete	Deterministic	for Feed-Forward ReLU Nets	Solver-Based		SMT-Based	✓	✓	✓	MNIST	$O(2^{lw})$	Complete	[41], [42]
					MILP-Based	✓			CIFAR-10	$O(2^{lw})$	Complete	[43]–[46]
			Extended Simplex Method			✓			MNIST	$O(2^{lw})$	Complete	[23], [47]
			Branch-and-Bound			✓	✓		CIFAR-10	$O(2^{lw})$	Complete	[26], [37], [48]–[54] [55]–[60]
Incomplete	Deterministic	for General DNNs^1	Linear Relaxation	Linear Programming (LP)		✓	(✓)	(✓)	CIFAR-10	$O(\text{poly}(l, w))$	$T_5{}^2$	[39], [40]
				Linear Inequality	Interval	✓	(✓)	(✓)	Tiny ImageNet	$O(lw^2)$	T_2	[61]
					Polyhedra	✓	(✓)	(✓)	Tiny ImageNet	$O(lw^3)$	$T_4{}^2$	[38], [39], [62]–[65]
					Zonotope	✓	(✓)	(✓)	Tiny ImageNet	$O(lw^3)$	$T_3{}^2$	[25], [66]–[68]
					Duality	✓	(✓)	(✓)	Tiny ImageNet	$O(lw^3)$	$T_4{}^2$	[27], [69]–[71]
				Multi-Neuron Relaxation		✓	(✓)	(✓)	CIFAR-10	$O(lw^3)$ - $O(2^{lw})^6$	T_7	[24], [72]–[74]
			Semidefinite Programming (SDP)			✓			CIFAR-10	$O(\text{poly}(l, w))$	T_6	[75]–[79]
			Lipschitz	General Lipschitz			✓		Tiny ImageNet	$O(lw^2)$	$T_1{}^3$	[11], [39], [80]–[84]
		for Lip-Bounded Nets		Smooth Layers		✓	✓		Tiny ImageNet	$O(lw^2)$	^{3}	[85]–[89]
		for Non-ReLU Nets^4	Curvature				✓		CIFAR-10	$O(lw^3)$	^{4}	[90]
		for Smoothed DNNs	Zeroth Order	Lipschitz				✓	ImageNet	$O(Slw^2)$	^{5}	[91]
	Probabilistic	for Smoothed DNNs	Zeroth Order	Differential Privacy Inspired			✓	✓	ImageNet	$O(Slw^2)$	ST_1	[92]
				Divergence Based			✓	✓	ImageNet	$O(Slw^2)$	ST_2	[93], [94]
				Neyman Pearson			✓		ImageNet	$O(Slw^2)$	ST_3	[22]
				Level-Set Analysis		(✓)	✓	✓	ImageNet	$O(Slw^2)$	ST_3	[28], [95], [96]
				Lipschitz		(✓)	✓		ImageNet	$O(Slw^2)$	ST_3	[97], [98]
			First Order			(✓)	✓	✓	ImageNet	$O(Slw^2)$	ST_4	[99], [100]

Li el al (2023). *SoK: Certified Robustness for Deep Neural Networks.* (IEEE S&P 2023)

Limitations of attacks

- Strong attacks require a good knowledge of the target model (e.g. access to model parameters)
- Model extraction: learn the model parameters by repeatedly querying
- Large distributed models are deployed on cloud and edge platforms and protected against extraction attacks (e.g. rate limiting)
- Finding adversarial examples becomes difficult

Good questions

- Is this attack more computationally expensive than JSMA?
- How does linearity contribute to adversarial vulnerability?
- Does the paper also apply to non-distillation defenses?
- Can we have an improved and computationally cheap distillation method for industrial applications?
- What are the most promising defenses against this attack?

Best questions

- Jiwoo Suh
  - How do the findings of this paper impact the security of autonomous driving systems, considering that while adversarial examples capable of bypassing defensive distillation have been developed, deploying these adversarial examples effectively in real-world scenarios remains a significant challenge?
- Ilman Sharif Mohammad al Momin
  - How effective might combining multiple defenses, such as defensive distillation with gradient obfuscation techniques, be against the L0, L2, and L∞ attacks introduced in this study?
- Changgun Kang
  - This attack requires high attacker capability, specifically complete access to the model. Can this type of adversarial attack be conducted on a black-box model?Can this type of attack be applied to large language models (LLMs) like ChatGPT?