Defending Network IDS against Adversarial Examples with Continual Learning

Jędrzej Kozal*, Justyna Zwolińska*, Marek Klonowski†, Michał Woźniak*

*Department of Systems and Computer Networks
†Department of Artificial Intelligence
Wrocław University of Science and Technology

1 December 2023

Evolution of cybersecurity threats

Security News ▸

SonicWall: Ransomware Declines Further As Attackers 'Pivot' Their Tactics

BY KYLE ALSPACH ▸
JULY 26, 2023, 06:00 AM EDT

The first half of 2023 saw ransomware attack volume drop even lower than in 2022, according to SonicWall data. But other types of threats are on the rise, including extortion and cryptojacking.

source:
https://www.crn.com/news/security/sonicwall-ransomware-declines-further-as-attackers-pivot-their-tactics

SECURITY JANUARY 5, 2023

Check Point Research Reports a 38% Increase in 2022 Global Cyberattacks

By Check Point Research Team

Check Point Research (CPR) releases new data on 2022 cyberattack trends. The data is segmented by global volume, industry and geography. Global cyberattacks increased by 38% in 2022, compared to 2021. These cyberattack numbers were driven by smaller, more agile hacker and ransomware gangs, who focused on exploiting collaboration tools used in work-from-home environments, targeting of education institutions that shifted to e-learning post COVID-19. This increase in global cyberattacks also stems from hacker interest in healthcare organizations, which saw the largest increase in cyberattacks in 2022, when compared to all other industries. CPR warns that the maturity of AI technology, such as CHATGPT, can accelerate the number of cyberattacks in 2023.

source:
https://blog.checkpoint.com/2023/01/05/38-increase-in-2022-global-cyberattacks/

Machine Learning-based Network IDS

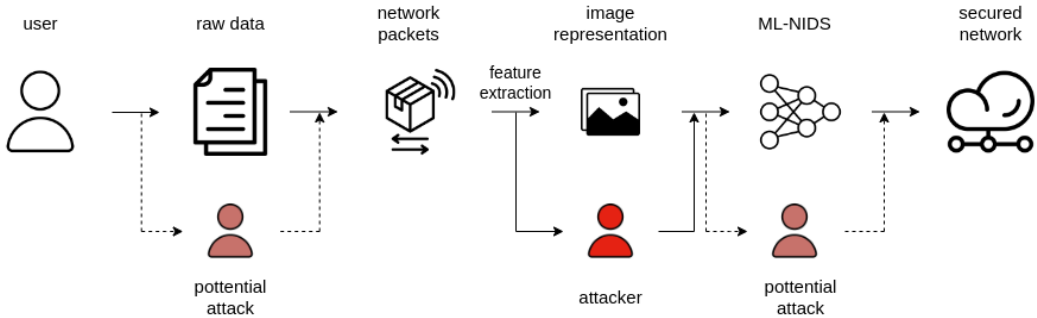
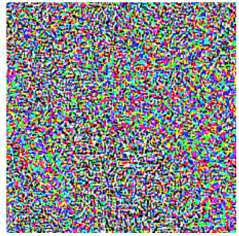


Figure: Man in the middle attack scenario.

Adversarial examples

$+ .007 \times$



$=$

$\boldsymbol{x}$

"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

Source: Ian J. Goodfellow et al. (2015). Explaining and Harnessing Adversarial Examples.

Continual learning

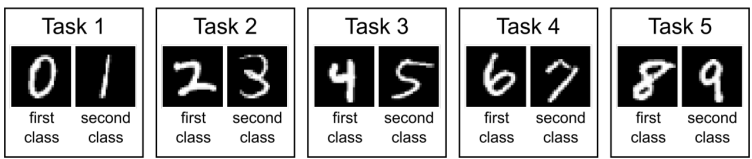


Figure 1: Schematic of split MNIST task protocol.

Source: van de Ven, G. M., & Tolias, A. S. (2019). Three scenarios for continual learning. CoRR, abs/1904.07734. http://arxiv.org/abs/1904.07734

Data used in experiments

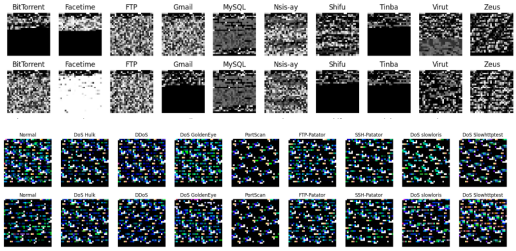


Figure: Visualization of images obtained from packets with network traffic for USTC-TFC2016[a] (top) and CIC-IDS-2017[b] (bottom) datasets.

[a] Wei Wang and David Lu. (2019). USTC-TK2016 toolkit

[b] Canadian Institute for Cybersecurity (2017). Intrusion Detection Evaluation Dataset

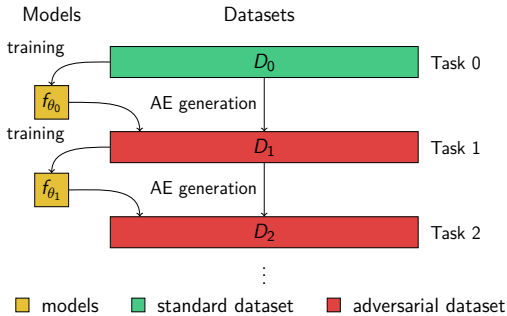


Figure: Diagram of adversarial examples generation pipeline

Adversarial attack generation process

untargeted attacks[a] :

$$x_i^{adv} = argmax_{\Delta x} L(x_i + \Delta x, y_i)$$
$$\text{subject to } x_i^{adv} \in \mathcal{X}$$

targeted attacks[b]:

$$\tilde{\mathcal{Y}}_i = \begin{cases} \mathcal{Y}^A & \text{if } y_i \in \mathcal{Y}^N \\ \mathcal{Y}^N & \text{otherwise} \end{cases}$$

$$\check{y}_i = argmax_{c \in \tilde{\mathcal{Y}}_i} p_c(x_i)$$
$$x_i^{adv} = argmin_{\Delta x} L(x_i + \Delta x, \check{y}_i)$$

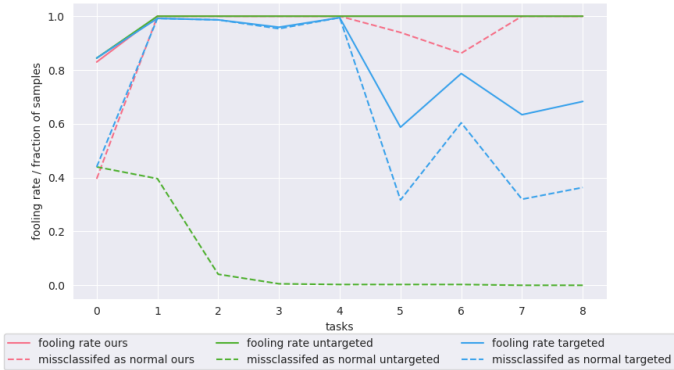


Figure: Fooling rate and desired misclassification rate over multiple tasks for untargeted, targeted, and combination of both.

[a] Szegedy, C. et al. (2013). Intriguing properties of neural networks.

[b] Alexey Kurakin et al. (2016). Adversarial examples in the physical world. CoRR, abs/1607.02533.

Continual learning methods

- Upperbound
- Naive
- Elastic Weight Consolidation (EWC)
- Synaptic Intelligence (SI)
- Experience Replay (ER)
- averaged Gradient Episodic Memory (aGEM)
- Maximally Infered Rehearsal (MIR)

Experiment setup

- TsAIL[a] adversarial attack with L_∞ norm
- generate 2000 adversarial examples per class
- 20 tasks, where the first is original dataset, and the following are adversarial examples

[a] Alexey Kurakin et al. (2018). Adversarial Attacks and Defences Competition. CoRR, abs/1804.00097.

- resnet18 architecture[a] with no pretraining
- batch size equal to 32, training for 10 epochs, learning rate 0.001, and weight decay 1e-6
- we report accuracy on all tasks after training (Acc_m), forgetting measure (FM), and accuracy on first task (Acc_1)

[a] K. He et al., "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778

Results for domain incremental scenario

Table: Accuracy and forgetting for domain incremental scenario

	USTC-TFC2016			CIC-IDS-2017		
method	$Acc_1(\uparrow)$	$Acc_m(\uparrow)$	$FM(\downarrow)$	$Acc_1(\uparrow)$	$Acc_m(\uparrow)$	$FM(\downarrow)$
Upperbound	0.997±0.003	0.635±0.008	0.046±0.001	0.998±0.001	0.705±0.012	0.043±0.004
Naive	0.089±0.028	0.146±0.007	0.823±0.006	0.372±0.043	0.257±0.011	0.765±0.012
EWC	0.103±0.048	0.135±0.008	0.723±0.020	0.290±0.104	0.231±0.033	0.747±0.013
SI	0.085±0.033	0.140±0.002	0.821±0.003	0.132±0.160	0.164±0.058	0.807±0.036
iCaRL	0.001±0.002	0.517±0.088	0.133±0.012	0.063±0.028	0.628±0.026	0.099±0.023
aGEM	0.136±0.027	0.122±0.007	0.723±0.010	0.394±0.125	0.276±0.059	0.697±0.029
ER	0.118±0.049	0.163±0.015	0.799±0.012	0.671±0.154	0.495±0.103	0.542±0.090
MIR	0.891±0.028	0.278±0.005	0.674±0.003	0.947±0.006	0.649±0.016	0.379±0.015

Accuracy over several tasks

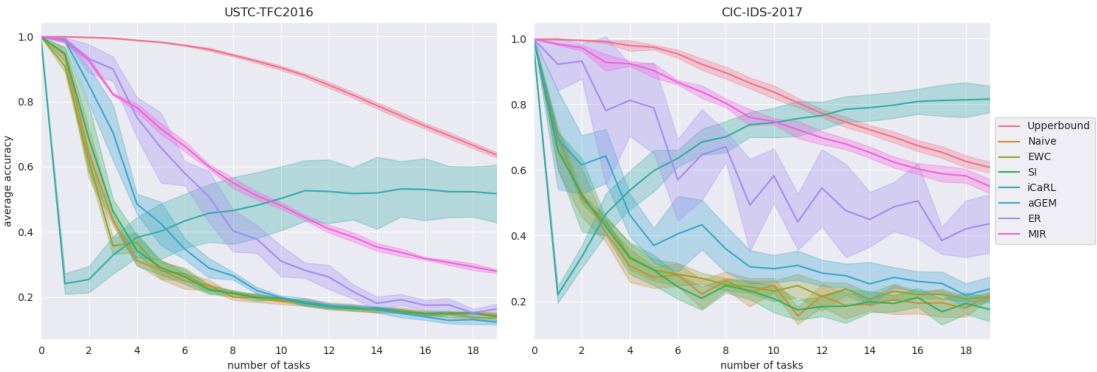


Figure: Average accuracy after training with each task. We only consider tasks seen by model so far

Conclusions

- adversarial examples pose serious threat to continual learning algorithms, that face user-generated input
- currently, the best results could be obtained with rehearsal methods or training with full data
- we should not use continual learning methods that utilize gradient information when learner could be exposed to adversarial attacks
- iCARL, when exposed to adversarial attacks, losses the ability to correctly recognize original data

Thank you for attention

Backup slides

Fooling rate

Table: Fooling rate for USTC-TFC2016 dataset for different values of α, ϵ and number of steps. We underline the value of results for parameters used in further parts of experiment.

α	$\epsilon = 0.1$			$\epsilon = 0.3$			$\epsilon = 0.5$		
	n. of steps								
	1	10	100	1	10	100	1	10	100
0.2	0.01	0.08	0.07	0.15	0.83	0.85	0.64	0.98	0.99
0.4	0.02	0.08	0.08	0.65	0.83	0.86	0.73	<u>0.99</u>	0.99
0.6	0.07	0.09	0.08	0.69	0.84	0.86	0.81	0.99	0.99
0.8	0.19	0.09	0.08	0.73	0.84	0.86	0.83	0.99	0.99
1.0	0.37	0.10	0.08	0.78	0.84	0.87	0.84	0.99	0.99

Results for standard Continual Learning protocol

Table: Accuracy and forgetting measaure for a domain-incremental scenario with combined real data

USTC-TFC2016 + CIC-IDS-2017			
method	$Acc_1(\uparrow)$	$Acc_m(\uparrow)$	$FM(\downarrow)$
Upperbound	1.0±0.0	0.9991±0.0001	0.0004±0.0002
Naive	0.5121±0.1324	0.4249±0.0223	0.4303±0.0241
EWC	0.5595±0.0124	0.4528±0.0022	0.3956±0.0042
SI	0.4977±0.107	0.4419±0.02	0.4095±0.0229
iCaRL	0.3657±0.2099	0.4937±0.0481	0.3451±0.0664
aGEM	0.9516±0.052	0.6942±0.0766	0.1735±0.0562
ER	0.9909±0.0147	0.8606±0.0528	0.0784±0.027
MIR	0.9999±0.0003	0.9925±0.0041	0.0039±0.0025

Adversarial examples generated from USTC-TFC2016 sample

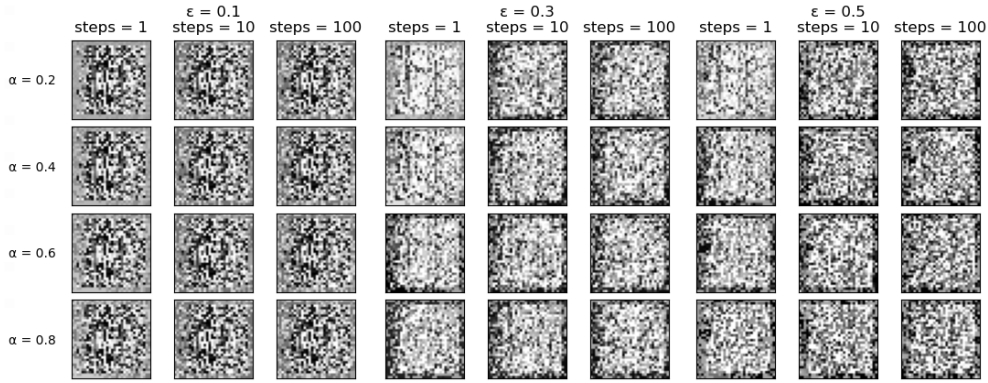


Figure: Examples of generated adversarial examples for USTC-TFC2016 dataset, various step size α, update norm ϵ and number of steps.