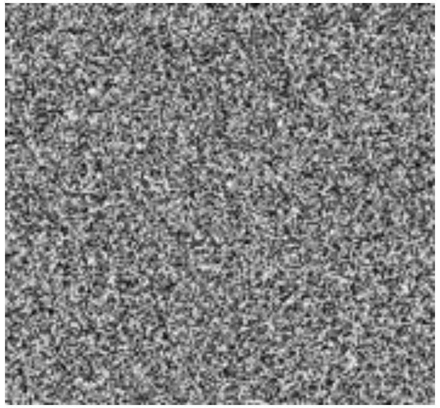
On The Empirical Effectiveness of Unrealistic Adversarial Hardening Against Realistic Adversarial Attacks

Salijona Dyrmishi, Salah Ghamizi, Thibault Simonetto, Yves Le Traon, Maxime Cordy

University of Luxembourg

Adversarial attacks against Machine Learning (ML)

Cute Dog
97%



Carefully crafted
adversarial noise

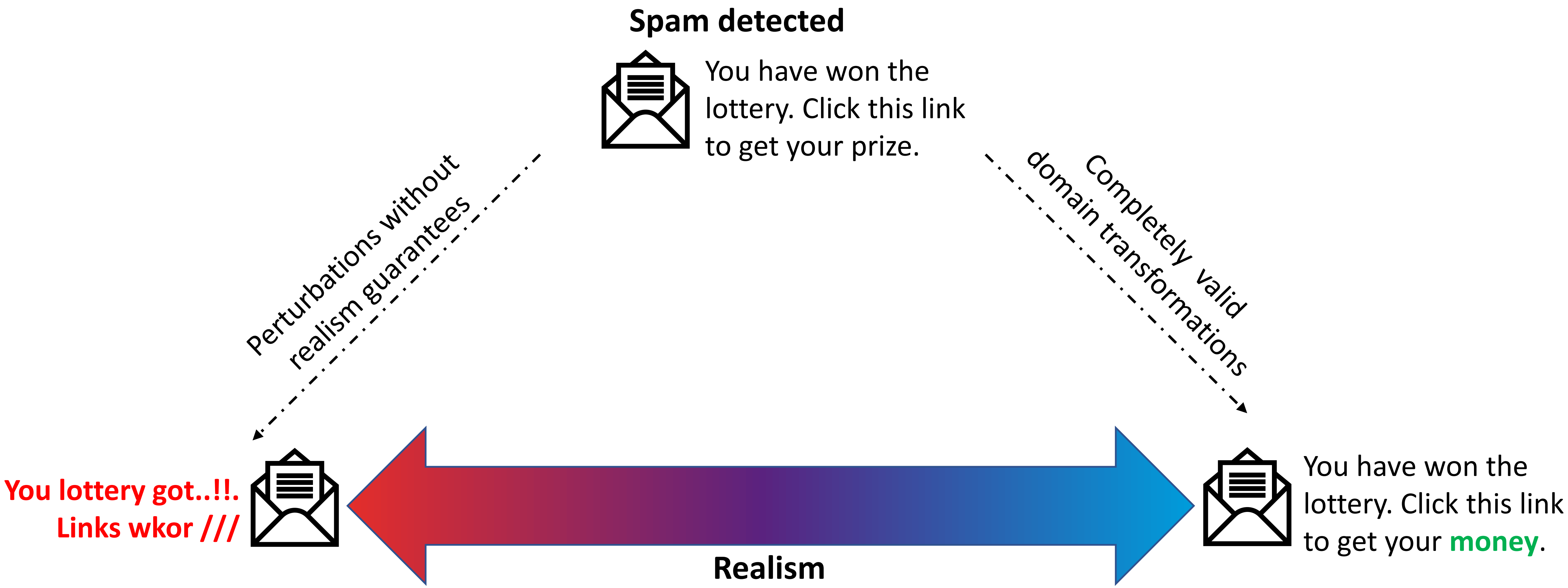
Angry Cat
82%

Attack	Gross success rate	Actual success rate
Papernot	74.86%	**0.00%**
PGD	17.30%	**0.00%**
CW2	80.00%	**0.00%**

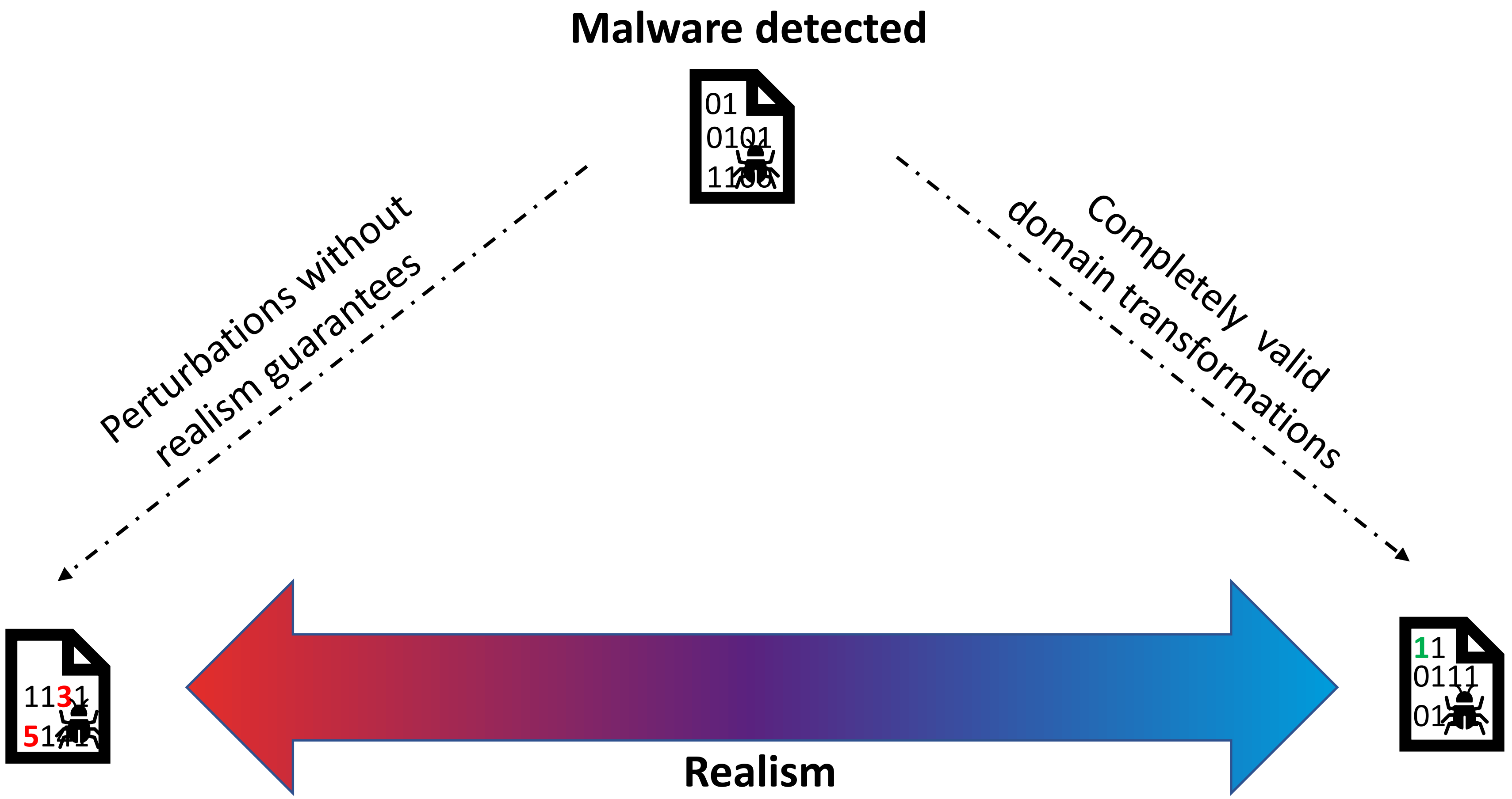
Tab 1. Success rate of traditional adversarial attacks against a credit scoring system

Ghamizi, Salah, et al. "Search-based adversarial testing and improvement of constrained credit scoring systems." Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. 2020.

Unrealistic vs realistic adversarial examples



Unrealistic vs realistic adversarial examples



Realistic attacks are indispensable but costly

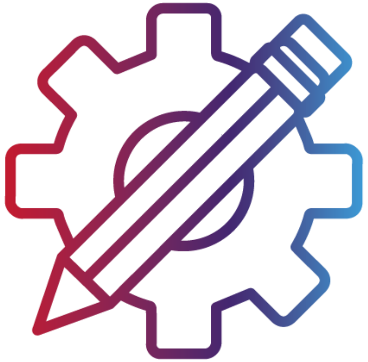
Design efforts

- Adapting existing attacks
- Creating new ones

Realistic attacks are indispensable but costly

Design efforts

- Adapting existing attacks
- Creating new ones



Engineering efforts

- Domain specifics (i.e sandbox for malware)

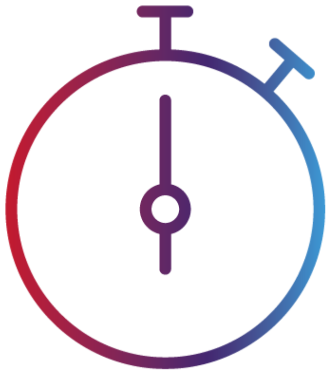
Realistic attacks are indispensable but costly

Design efforts

- Adapting existing attacks
- Creating new ones

Engineering efforts

- Domain specifics (i.e sandbox for malware)



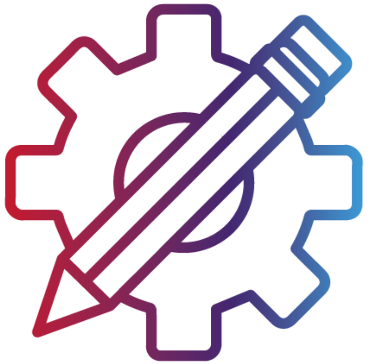
Run time

- 3.8 to 22650 longer for attacks in this study

Realistic attacks are indispensable but costly

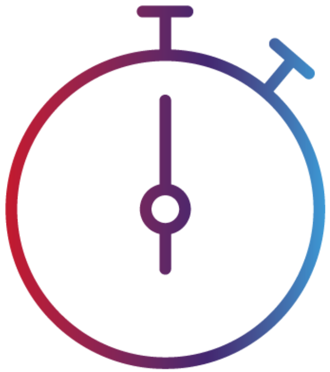
Design efforts

- Adapting existing attacks
- Creating new ones



Engineering efforts

- Domain specifics (i.e sandbox for malware)

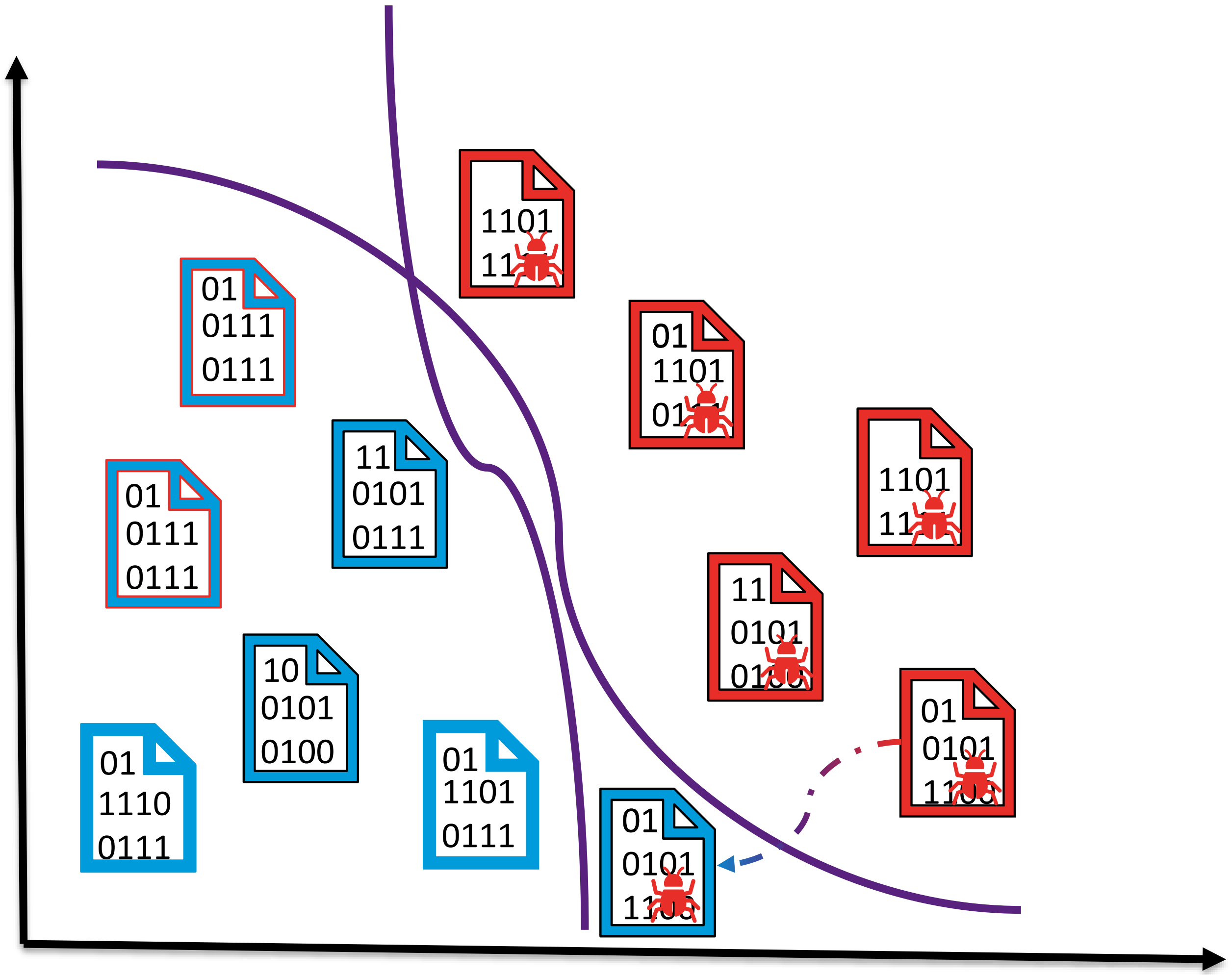


Run time

- 3.8 to 22650 longer for attacks in this study

Adversarial hardening

Improving ML model robustness by learning from adversarial examples



Hardening models with realistic adversarials is expensive …

3 to 1K+ more than normal model training
(depending on hardening strategy, dataset, model, attack)

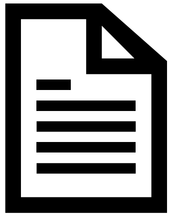
Hardening models with realistic adversarials is expensive …

RQ1: Can we use “cheap” unrealistic examples instead to protect against realistic attacks?

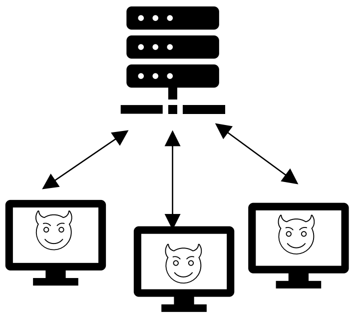
Use case selection

Application domains and learning tasks that have:

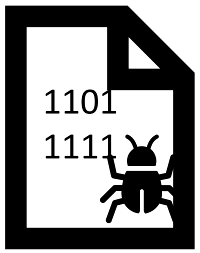
1. Constrained inputs
2. Open-source datasets
3. Open-source realistic attacks



Text classification

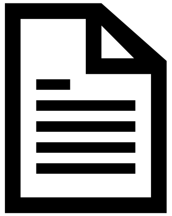


Botnet detection



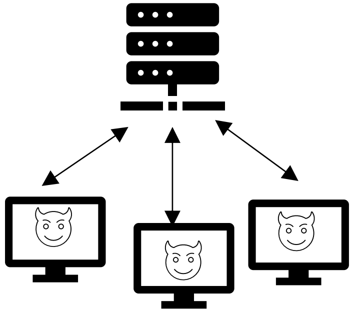
Malware detection

Experimental settings



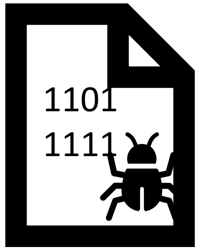
Text classification

- Transformer model
- Adversarial fine tuning
- 1 unrealistic & 2 realistic
- 3 datasets



Botnet detection

- FC model
- Adversarial training
- 1 unrealistic & 2 realistic
- 3 datasets



Malware detection

- RF model
- Adversarial training
- 2 unrealistic & 1 realistic
- 1 dataset

RQ1 results: Text classification

Can we use “cheap” unrealistic examples to harden models ?

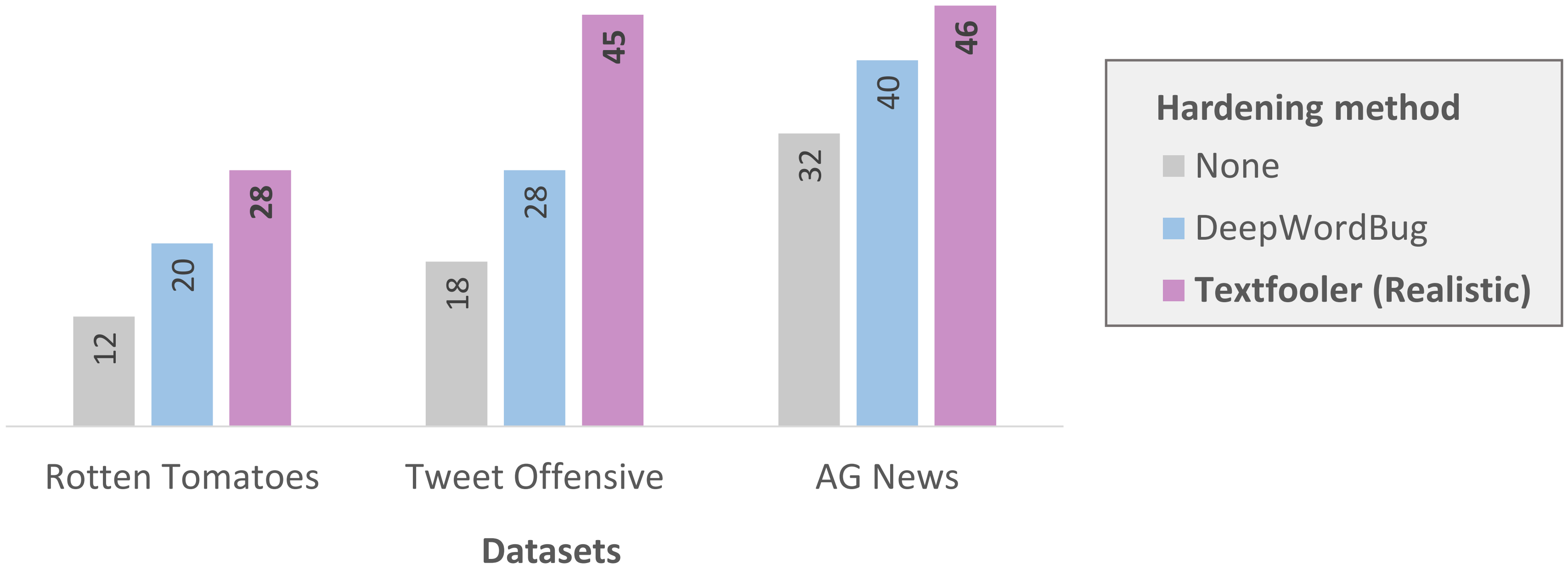


Fig 1. Robust accuracy (%) of the text-based model against PWWS realistic attack

RQ1 results: Botnet detection

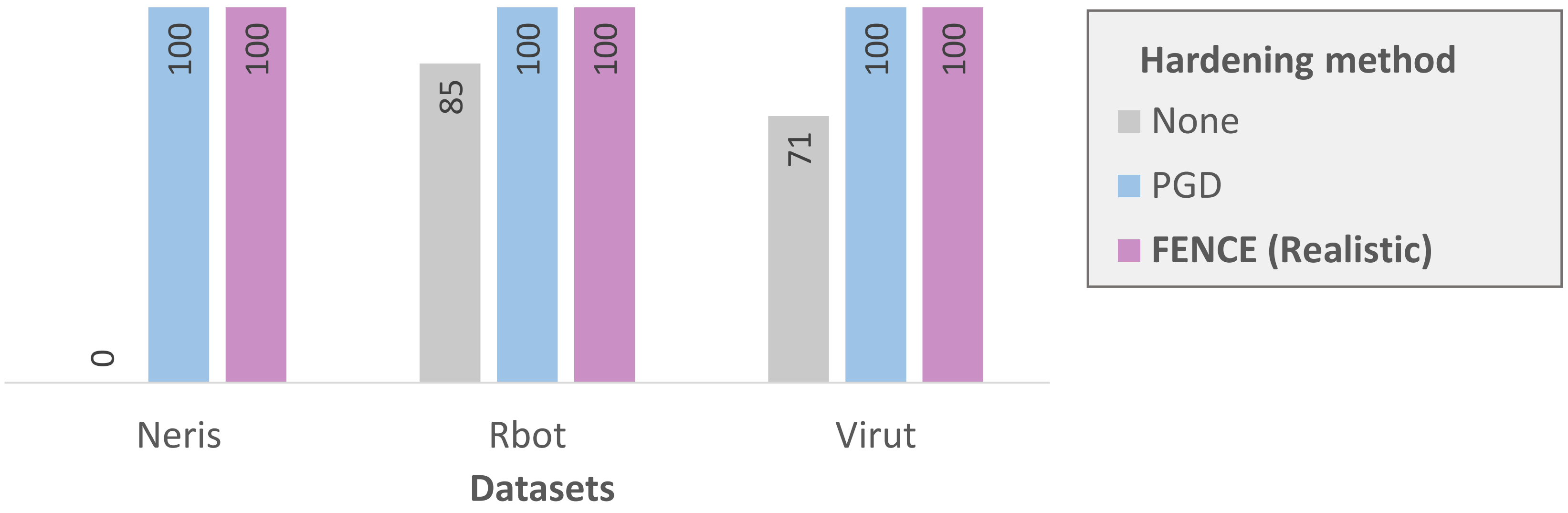


Fig 2. Robust accuracy (%) of the botnet detection model against FENCE realistic attack

RQ1 results: Malware detection

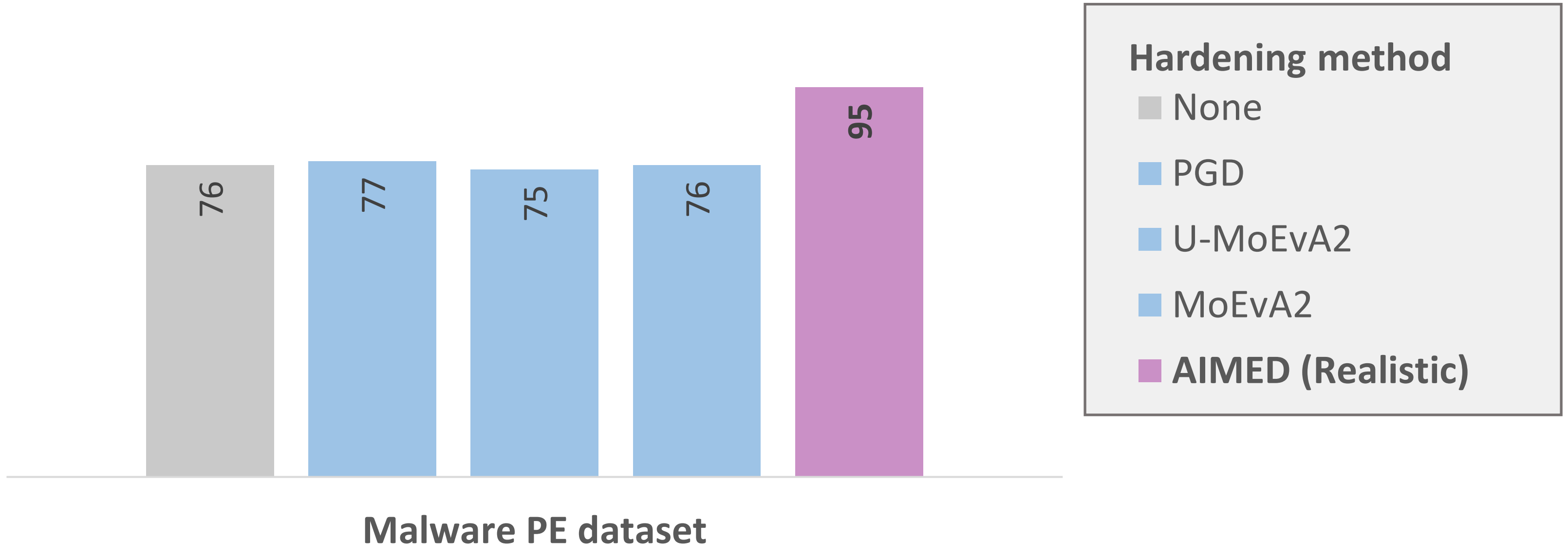
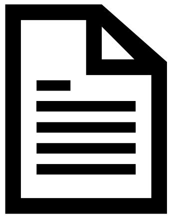


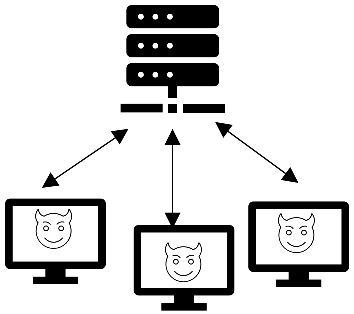
Fig 3. Robust accuracy of the malware detection model against AIMED realistic attack

RQ1: Can we use “cheap” unrealistic examples instead to protect against realistic attacks?



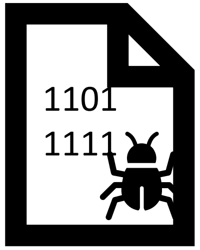
Text classification

At certain level
Up to 9.56%



Botnet detection

YES
100% protection



Malware detection

NO
0% protection

Further investigation

RQ2: Do larger budgets help unrealistic hardening ?

RQ2 results: Text classification

Do larger budgets help unrealistic hardening ?

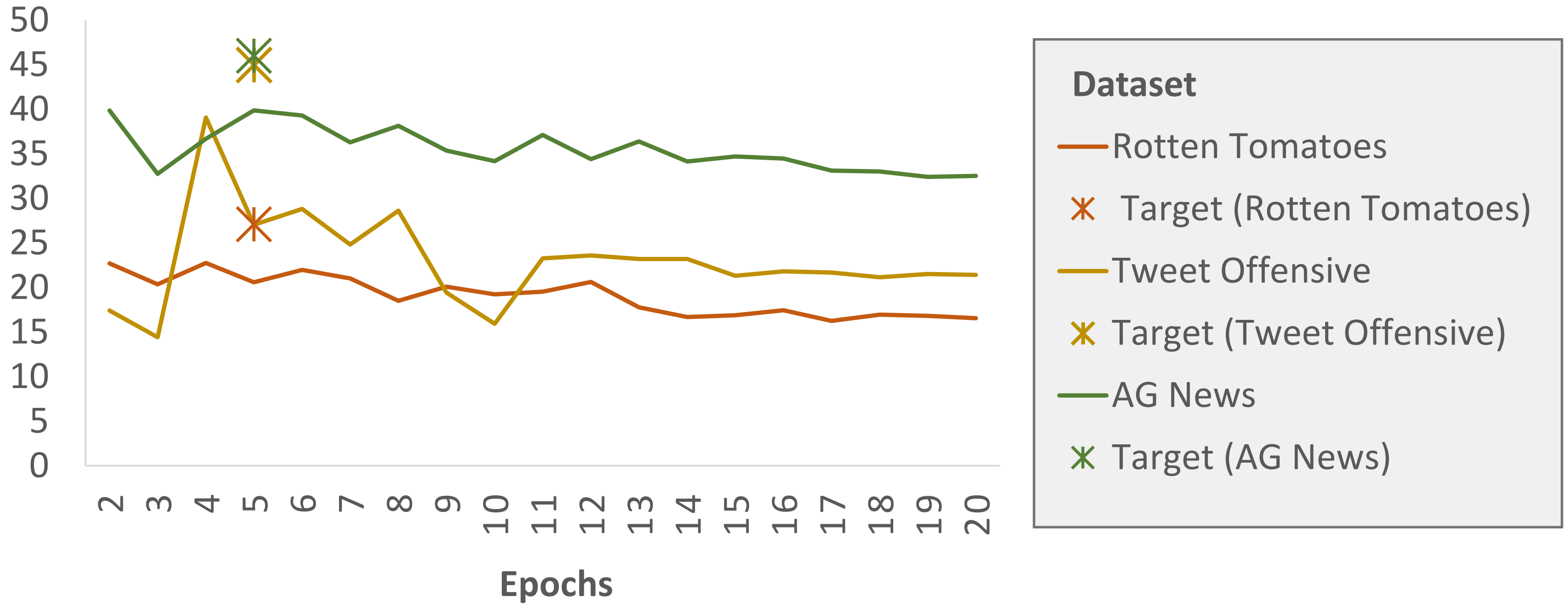


Fig 4. Robust accuracy (%) of the text-based model against PWWS realistic attack when hardened with DeepWordBug attack for several epochs.

**Targets represents the robust accuracy while hardening the model with realistic attack TextFooler for 5 epochs.*

RQ2 results: Malware detection

Do larger budgets help unrealistic hardening ?

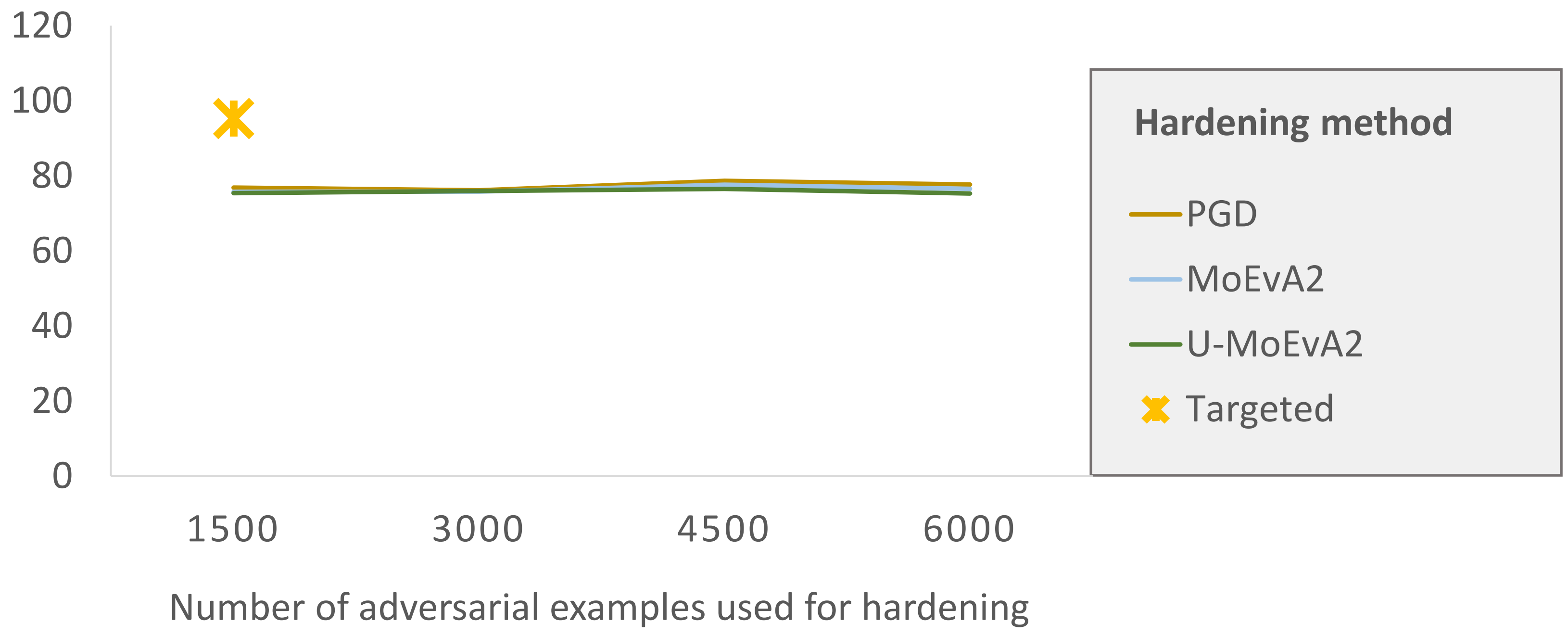
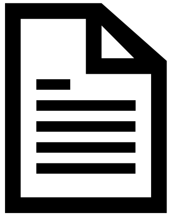


Fig 5. Robust accuracy of the malware detection model against AIMED realistic attack

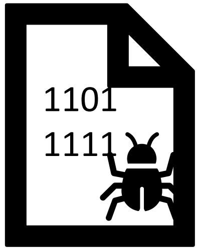
**Targets represents the robust accuracy while hardening the model with 1500 realistic examples generated from AIMED.*

RQ2: Do larger budgets help unrealistic hardening ?



Text classification

NO



Malware detection

NO

Further investigation

RQ2: Do larger budgets help unrealistic hardening ?

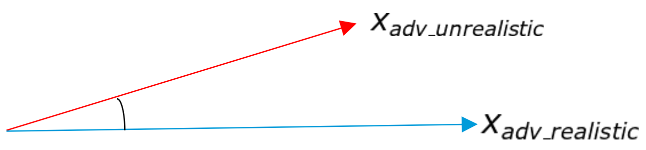
RQ3: Which properties of adversarial examples influence the hardening results ?

RQ3 metrics

Which properties of adversarial examples influence the hardening results ?

1. Direction of perturbation

$$sim(x_{adv\_realistic}, x_{adv\_unrealistic}) = \frac{\overrightarrow{x_{adv\_realistic}} * \overrightarrow{x_{adv\_unrealistic}}}{\| \overrightarrow{x_{adv\_realistic}} \| * \| \overrightarrow{x_{adv\_unrealistic}} \|}$$

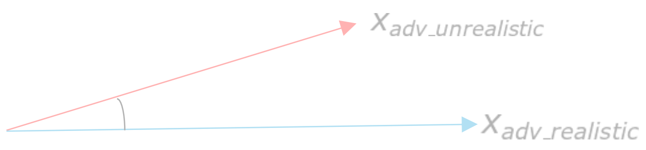


RQ3 metrics

Which properties of adversarial examples influence the hardening results ?

1. Direction of perturbation

$$sim(x_{adv\_realistic}, x_{adv\_unrealistic}) = \frac{\overrightarrow{x_{adv\_realistic}} * \overrightarrow{x_{adv\_unrealistic}}}{\| \overrightarrow{x_{adv\_realistic}} \| * \| \overrightarrow{x_{adv\_unrealistic}} \|}$$



2. Aggressiveness

Initial example

Adversarial

$$aggressiveness = \frac{D(x, \hat{x})}{D(x, x^{nn})}$$

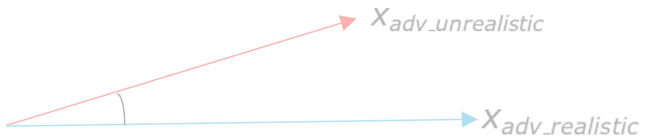
Correctly classified nearest neighbor of different class

RQ3 metrics

Which properties of adversarial examples influence the hardening results ?

1. Direction of perturbation

$$sim(x_{adv\_realistic}, x_{adv\_unrealistic}) = \frac{\overrightarrow{x_{adv\_realistic}} * \overrightarrow{x_{adv\_unrealistic}}}{\| \overrightarrow{x_{adv\_realistic}} \| * \| \overrightarrow{x_{adv\_unrealistic}} \|}$$



2. Aggressiveness

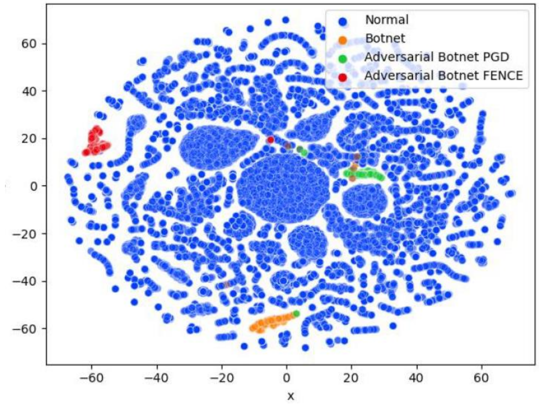
Initial example

Adversarial

$$aggressiveness = \frac{D(x, \hat{x})}{D(x, x^{nn})}$$

Correctly classified nearest neighbor of different class

3. Qualitative 2D embeddings (t-SNE)

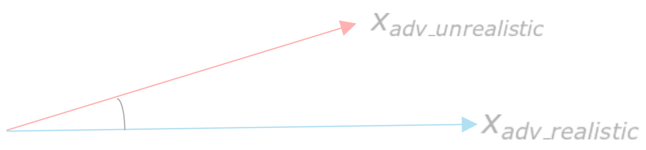


RQ3 metrics

Which properties of adversarial examples influence the hardening results ?

1. Direction of perturbation

$$sim(x_{adv\_realistic}, x_{adv\_unrealistic}) = \frac{\overrightarrow{x_{adv\_realistic}} * \overrightarrow{x_{adv\_unrealistic}}}{\| \overrightarrow{x_{adv\_realistic}} \| * \| \overrightarrow{x_{adv\_unrealistic}} \|}$$



2. Aggressiveness

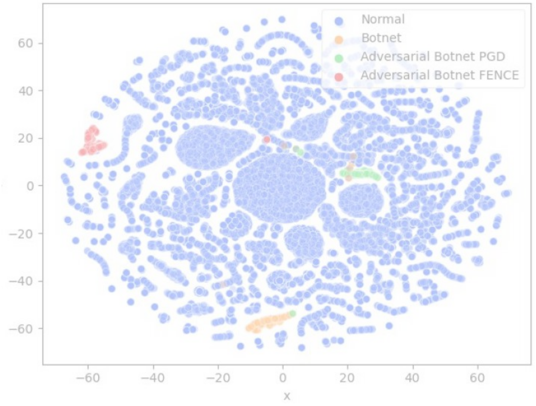
Initial example

Adversarial

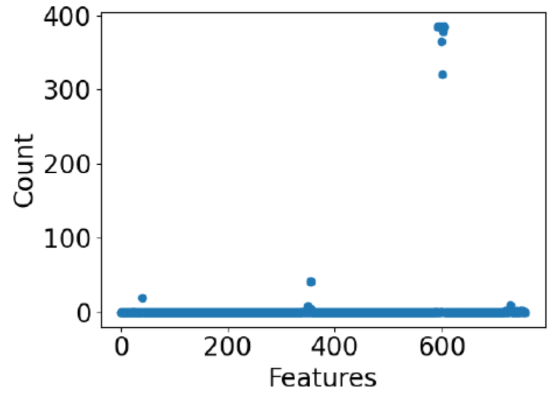
$$aggressiveness = \frac{D(x, \hat{x})}{D(x, x^{nn})}$$

Correctly classified nearest neighbor of different class

3. Qualitative 2D embeddings (t-SNE)



4. Feature perturbation

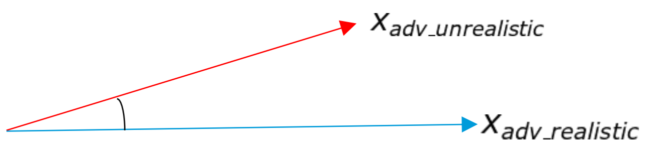


RQ3 metrics

Which properties of adversarial examples influence the hardening results ?

1. Direction of perturbation

$$sim(x_{adv\_realistic}, x_{adv\_unrealistic}) = \frac{\overrightarrow{x_{adv\_realistic}} * \overrightarrow{x_{adv\_unrealistic}}}{\| \overrightarrow{x_{adv\_realistic}} \| * \| \overrightarrow{x_{adv\_unrealistic}} \|}$$



2. Aggressiveness

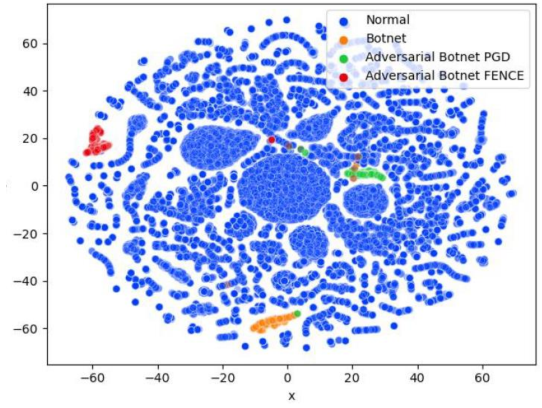
Initial example

Adversarial

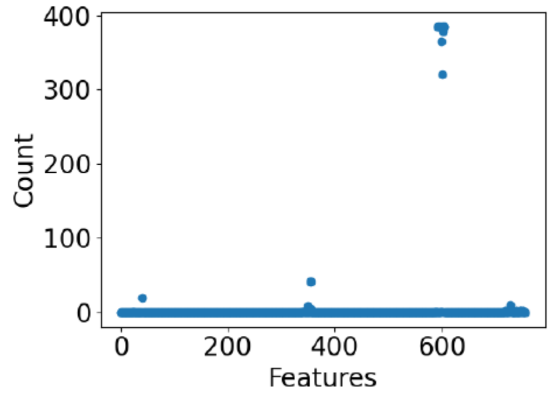
$$aggressiveness = \frac{D(x, \hat{x})}{D(x, x^{nn})}$$

Correctly classified nearest neighbor of different class

3. Qualitative 2D embeddings (t-SNE)



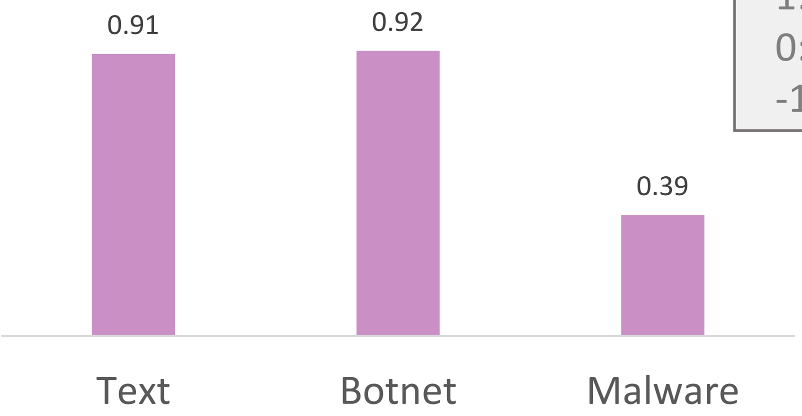
4. Feature perturbation



RQ3 results

Which properties of adversarial examples influence the hardening results ?

1. Direction of perturbation



1: Proportional
0: Orthogonal
-1: Opposite

Fig 6. **Average cosine similarity** between realistic and unrealistic examples across datasets and attacks for each use case

RQ3 results

Which properties of adversarial examples influence the hardening results ?

1. Direction of perturbation

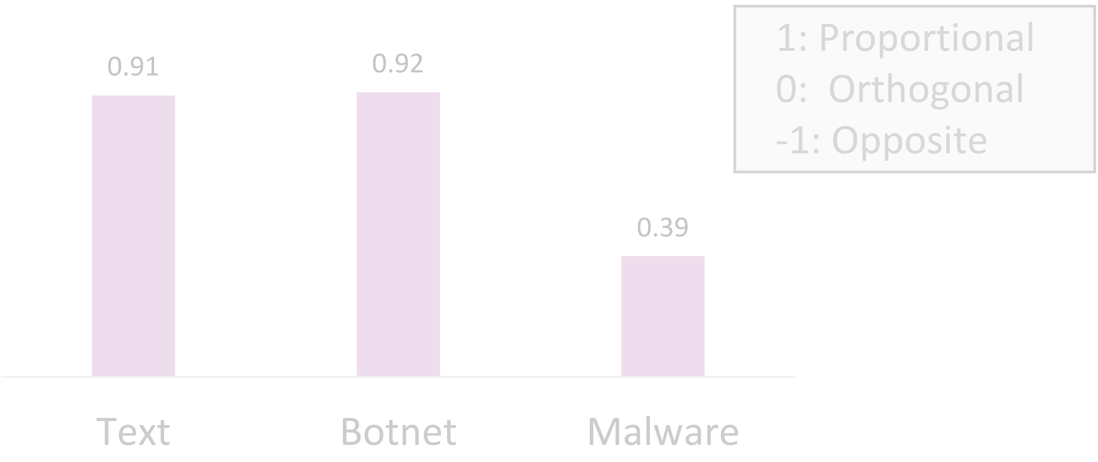


Fig 6. **Average cosine similarity** between realistic and unrealistic examples across datasets and attacks for each use case

2. Aggressiveness

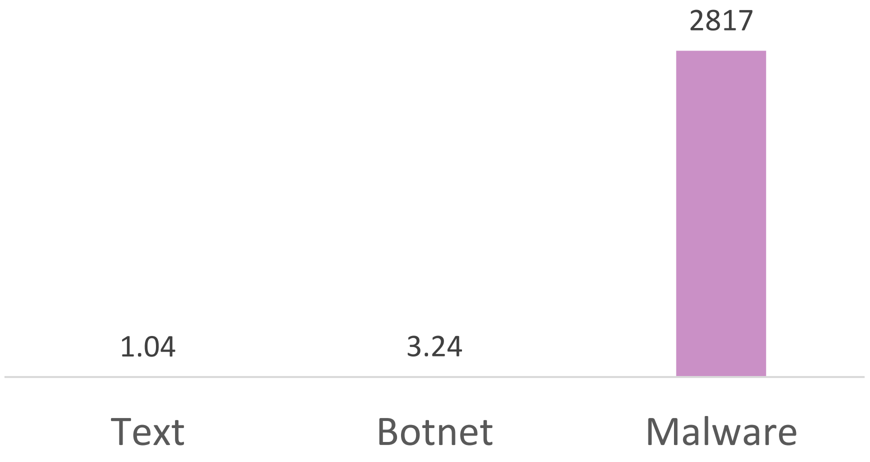


Fig 7. **Average aggressiveness ratio** between realistic and unrealistic examples across datasets and attacks for each use case

Lessons learned

1. Unrealistic examples may help adversarial hardening under strict conditions; hence they are worth a try!

2. If unrealistic examples do not bring improvement even at a small scale, they will probably never do !

3. Unrealistic hardening is helpful when the properties of unrealistic examples are similar to the ones of realistic examples.

Paving the way to new adversarial hardening methods with cheap unrealistic examples

S&P 2023, 23 May at 9am

uni.lu | SnT