26th IEEE Symposium on Computers and Communications (IEEE ISCC 2021)

◎ Security

EvilModel: Hiding Malware Inside of Neural Network Models

Zhi Wang, Chaoge Liu, Xiang Cui

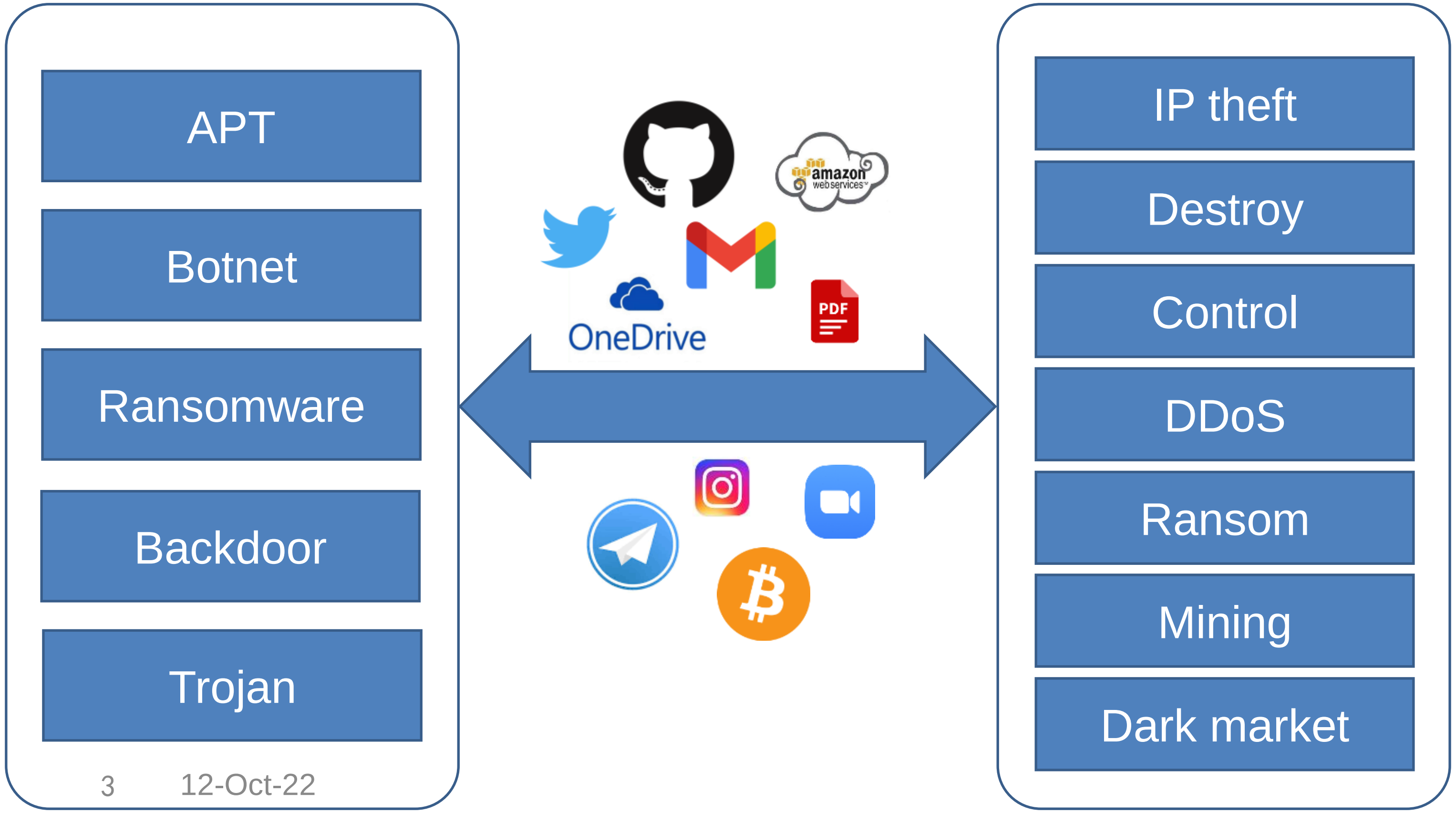
Athens, Greece
September 2021

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS
中国科学院大学
University of Chinese Academy of Sciences
广州大学
GUANGZHOU UNIVERSITY

Contents

Background

- Advanced malware campaigns are the main threats to computer security.
- Attackers need to communicate with the malware covertly to send customized commands and payloads.
- Some methods of covertly transmitting messages are widely used in the wild, but they are not suitable for large-sized binary payloads.



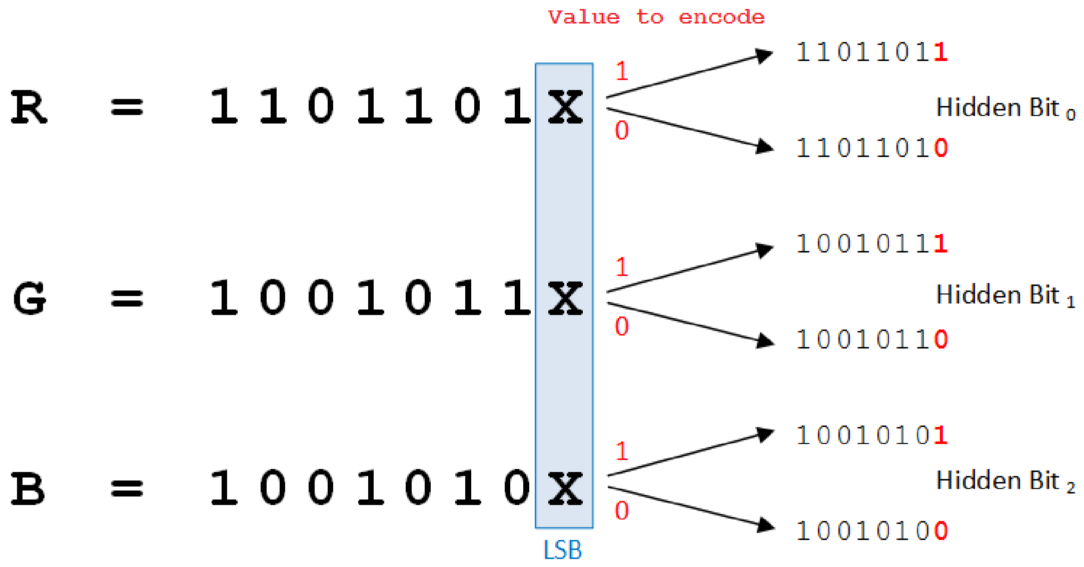
Year	Name	Platform
2009	upd4t3	Twitter, Tumblr
2014	Garybot	Twitter
2015	Hammertoss	Twitter, GitHub
2015	MiniDuke	Twitter
2017	ROKRAT	Twitter, Yandex
2017	PlugX	Pastebin
2018	Comnie	GitHub, Blogspot
2018	HeroRat	Telegram
2019	DarkHydrus	Google Drive
2019	Pony	Bitcoin
2019	Glupteba	Bitcoin
2019	IPStorm	IPFS
2020	Turla	Gmail

Background

Decode at https://stylesuxx.github.io/steganography/

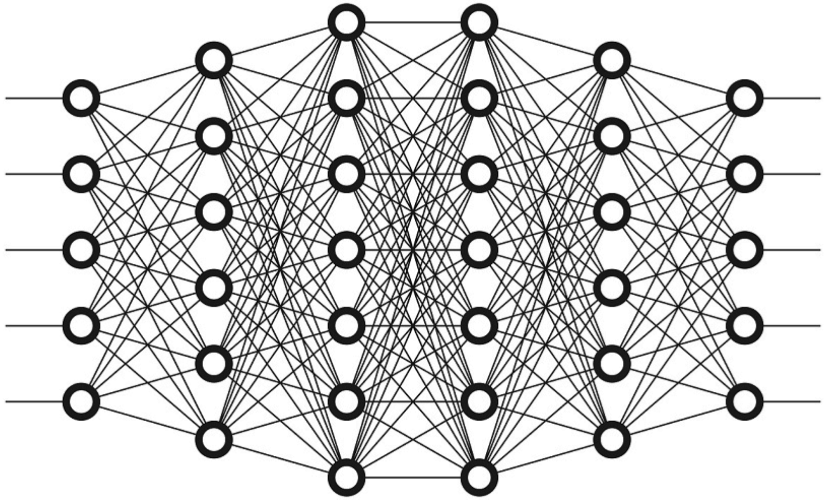
For delivering large-sized malware, some attackers attach the malware to benign-looking carriers.

Or **Steganography**



Background

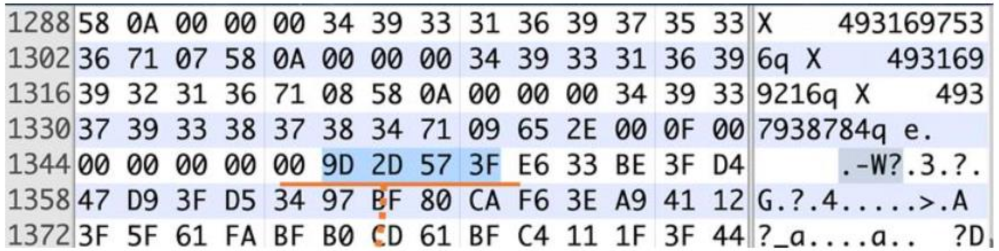
A neural network model has many neurons, with millions of parameters inside.



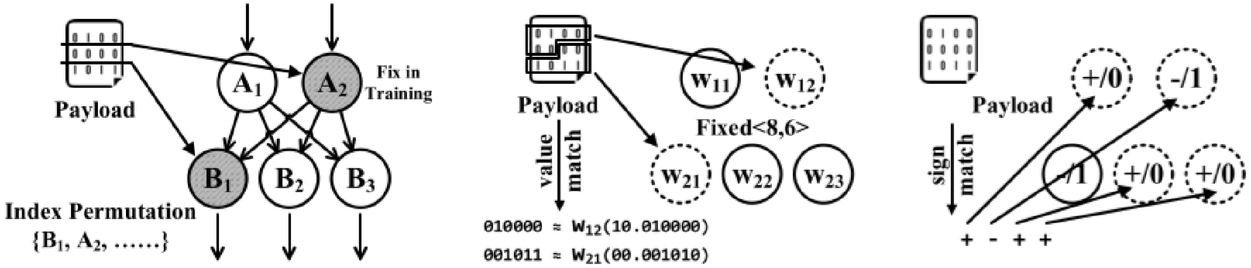
Can malware be hidden inside of neural networks?

YES

“Neural Network Backdoor”



StegoNet

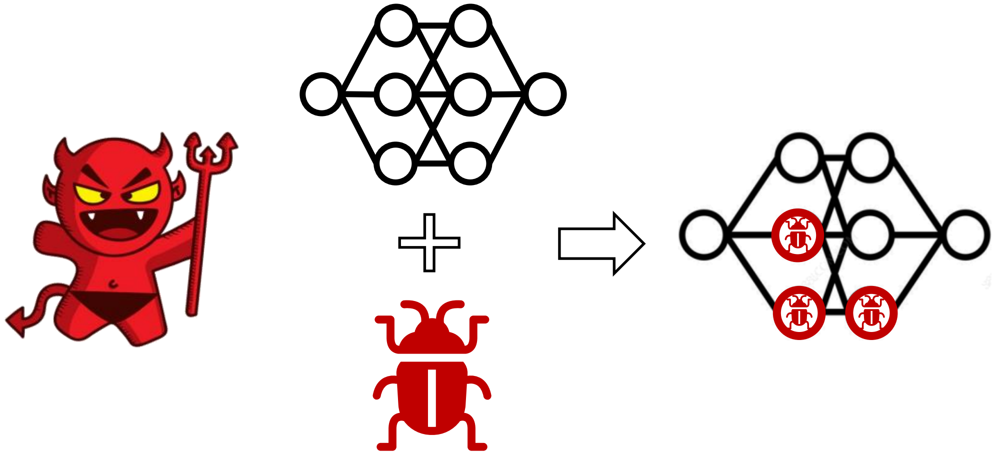


(a) Resilience training. (b) Value-mapping. (c) Sign-mapping.

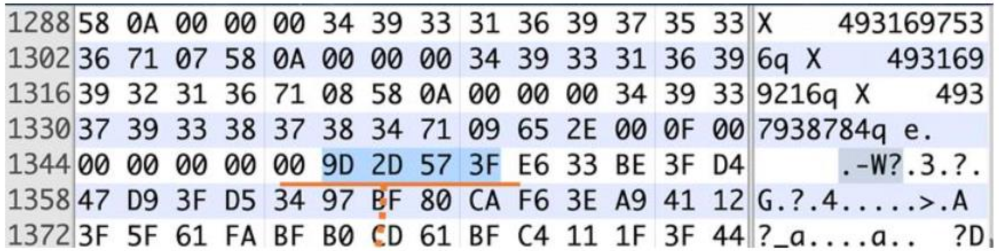
Background

Advantage

- Disassembled malware -> Evade detection.
- Redundant neurons -> No significant performance loss.
- Large size models -> Large size malware.
- Don’t rely on software vulnerabilities.
- Wide application of AI -> Universal.

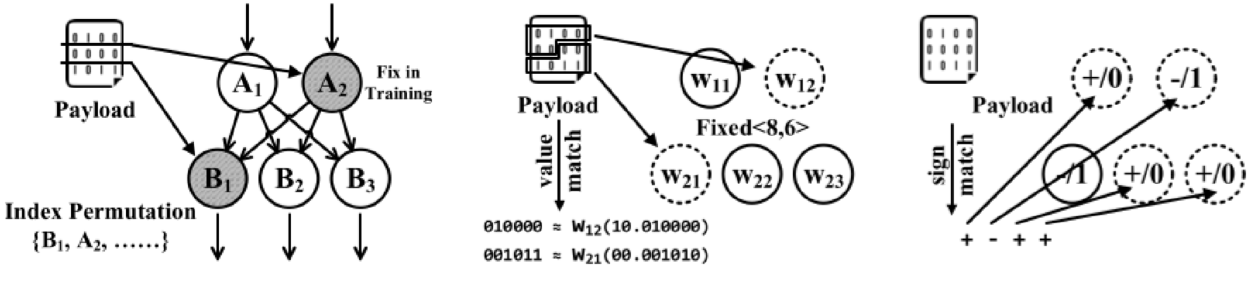


“Neural Network Backdoor”



Concept

StegoNet



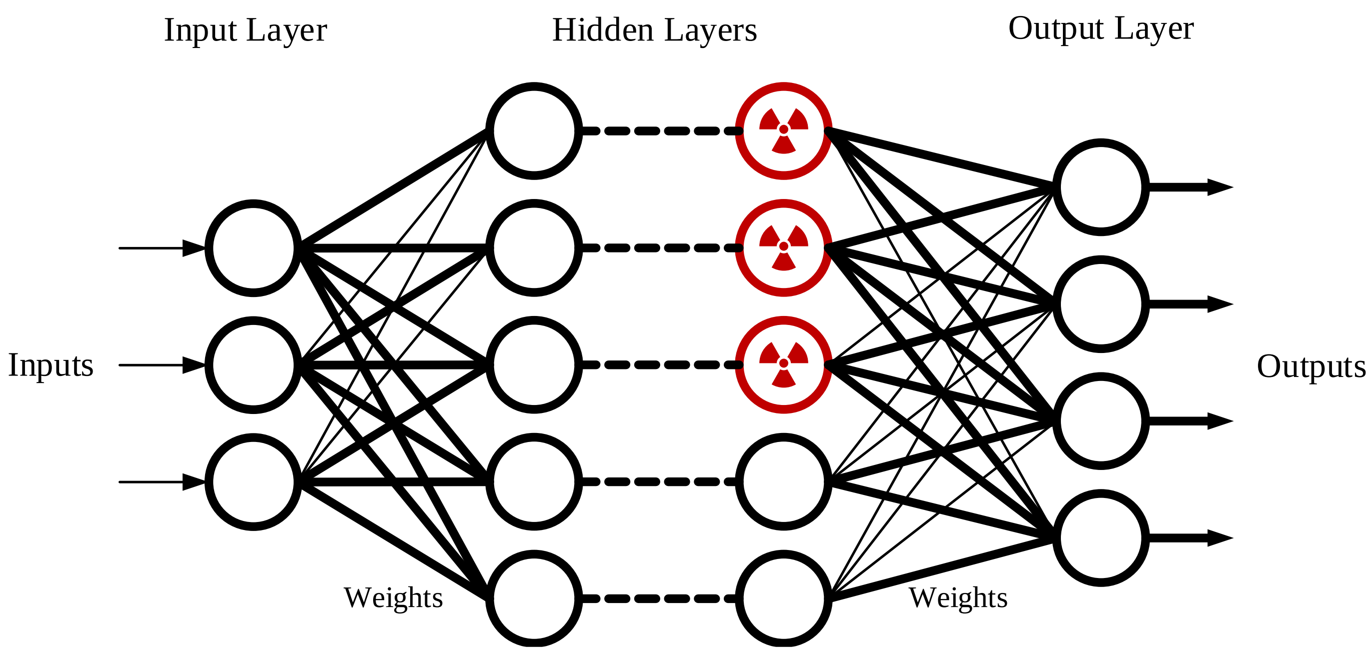
(a) Resilience training. (b) Value-mapping. (c) Sign-mapping.

Low embedding rate, high performance loss, extra info

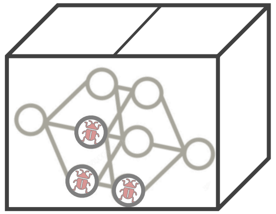
Introduction

1 Fast substitution

- ✓ Higher embedding rate
- ✓ Lower performance loss
- ✓ No extra info
- ✓ Capable of different models

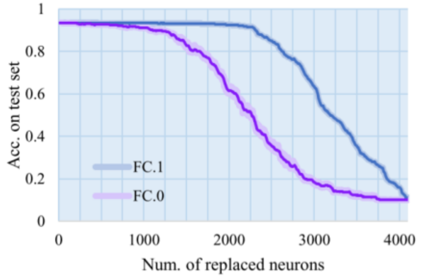


2



Embedding capacity of DNN models

3

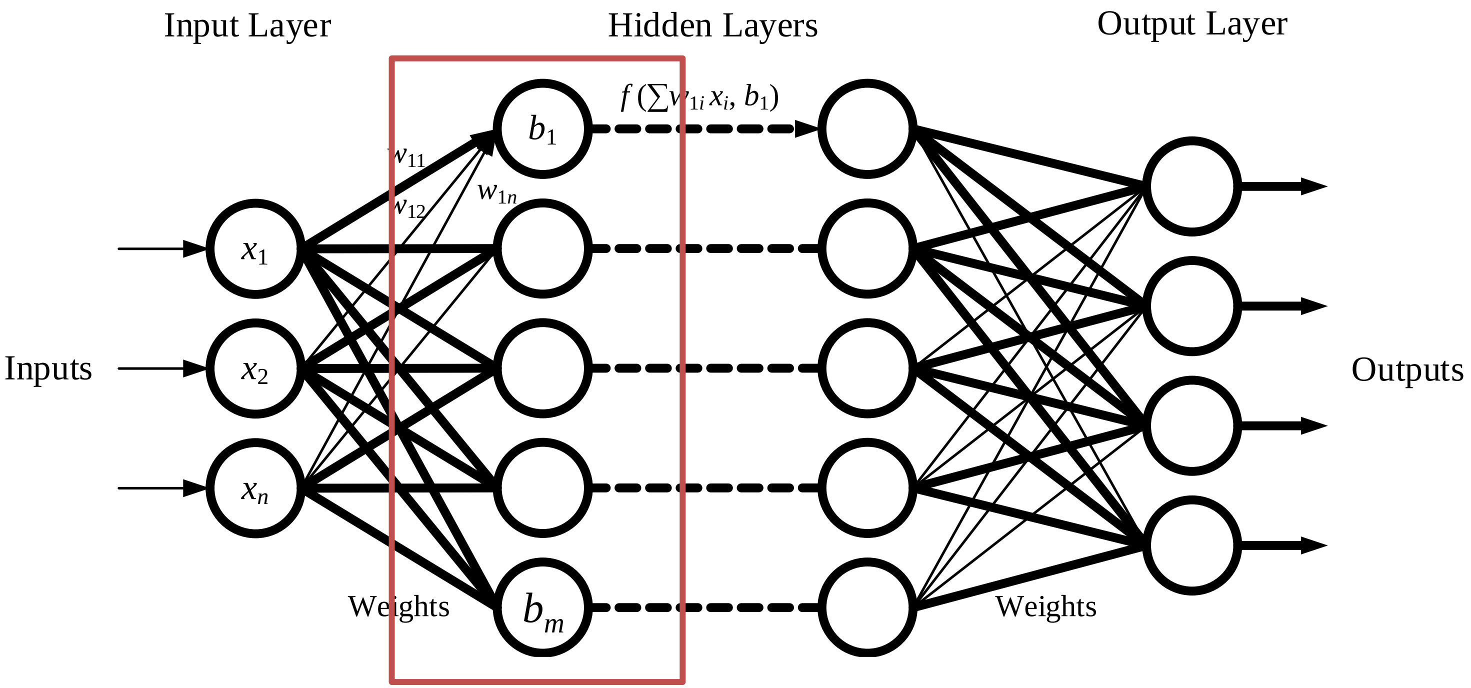


Embedding impact on model performance

Ethical Considerations. The combination of neural networks and cyber attacks is a forwarding trend. We intend to provide a possible scenario for security researchers and vendors to mitigate this kind of attack in advance.

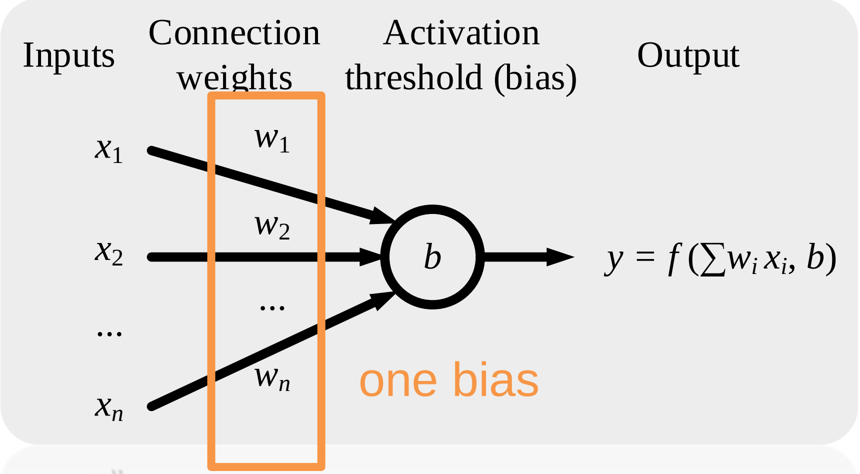
Technical Design

Basic structure of neural network models



m(n+1) parameters in a layer

4m(n+1) Bytes in a neuron



n connection weights

n+1 parameters in a neuron

4(n+1) Bytes in a neuron

TensorFlow ONNX PyTorch Keras

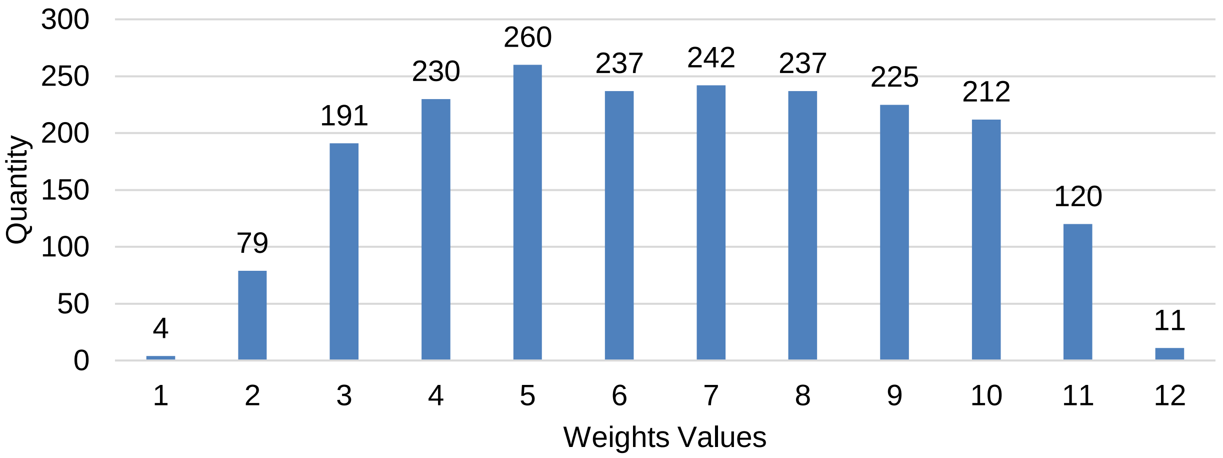
32 bits floating-point number

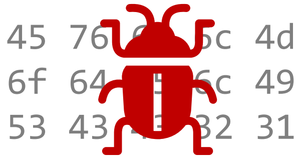
Technical Design

Parameters in a neuron

```
[-0.003133472520858494, -0.009729900397360325,
  0.021175190806388855, -0.001930642407387495,
 -0.016773682087659836, -0.015056176111102104,
                      … …
  0.009281742386519909, -0.011762472800910473]
```

2048 parameters, 1001 negative numbers, 1047 positive numbers





Malware bytes → Floating-point numbers

IEEE Standard for Floating-Point Arithmetic

Sign	Exponent	Mantissa
0	1 8bits 8	9 23bits 31

$$\pm 1.m \times 2^{n}$$

BC40B763

The absolute value is determined by the **exponent** part.

BCFFFFFF

-0.0312499981374

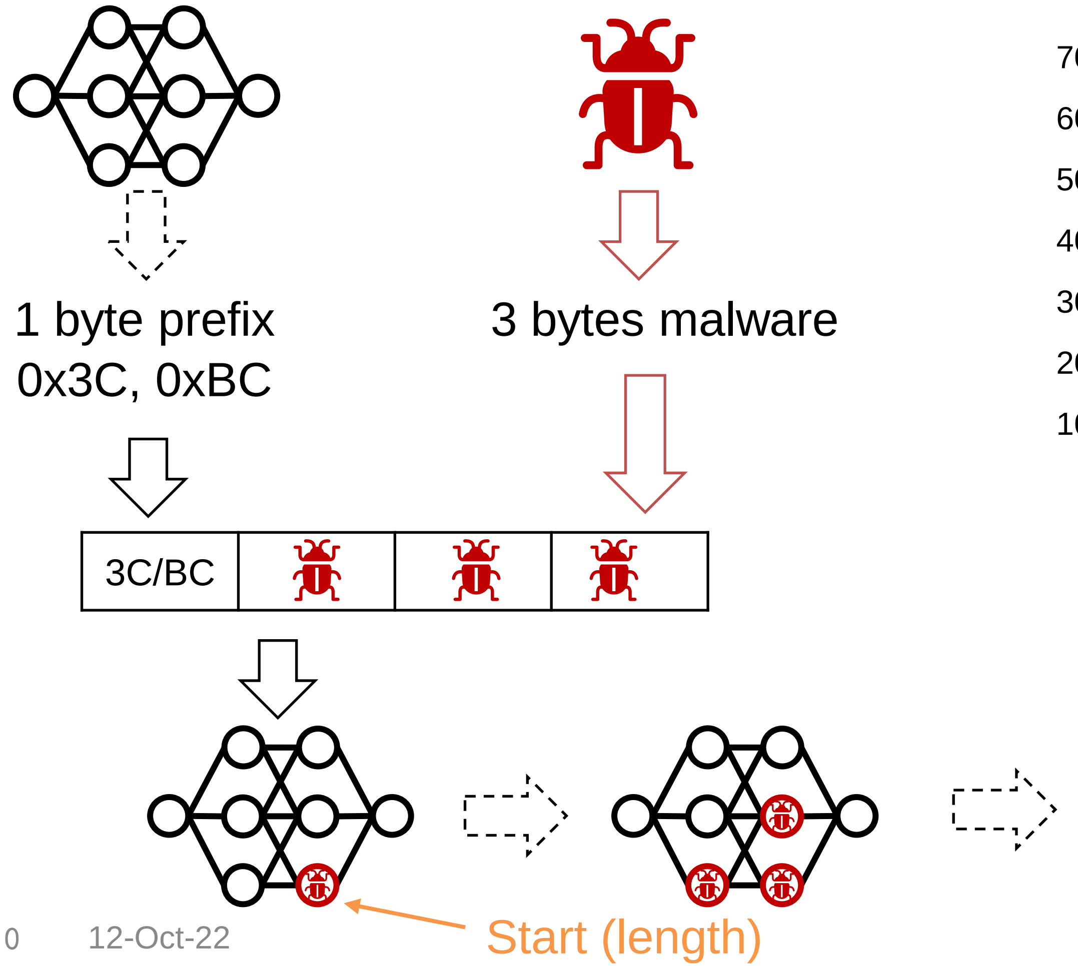
BC000000

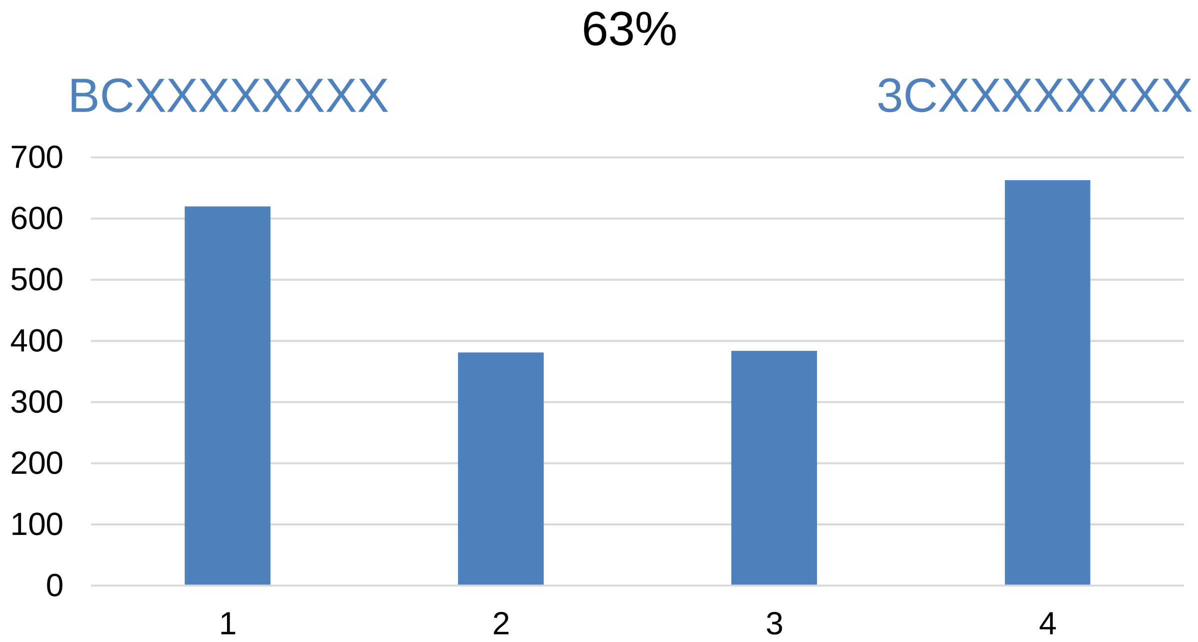
-0.0078125

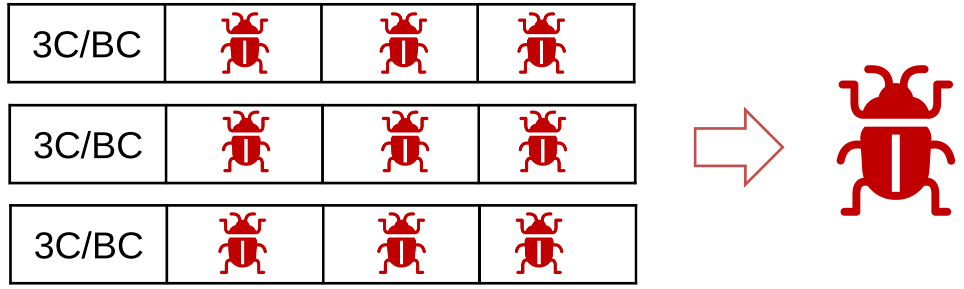
Prefix: 0x3C 0x38 0xBC 0xB8

Technical Design

Fast substitution







Overall Workflow

Attackers

delivering a payload

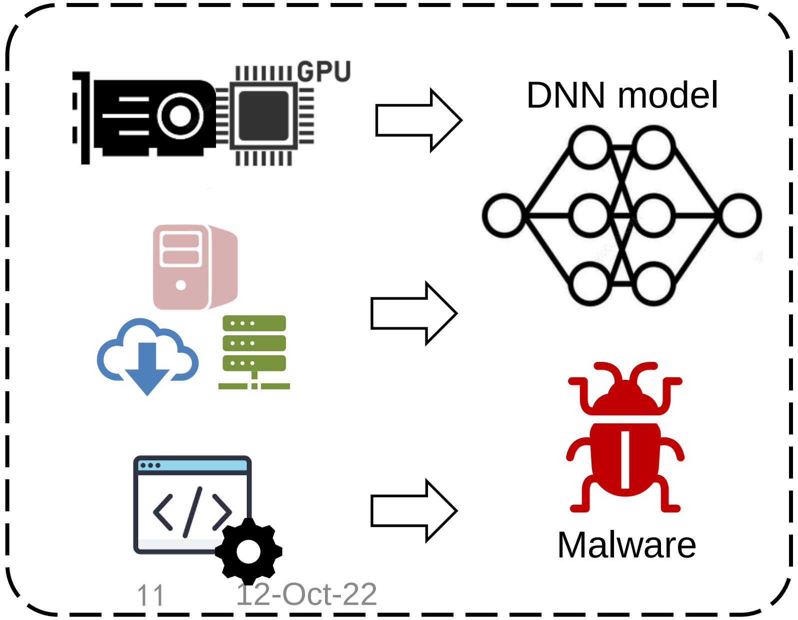
- **Prepare** the model (design a model, or download a pre-trained model)
- Train or fine-tune the model
- **Embed** the malware in the model
- **Evaluate** the performance
- Retrain the model if the loss exceeds an acceptable range
- **Publish** the malware-embedded model

Receiver

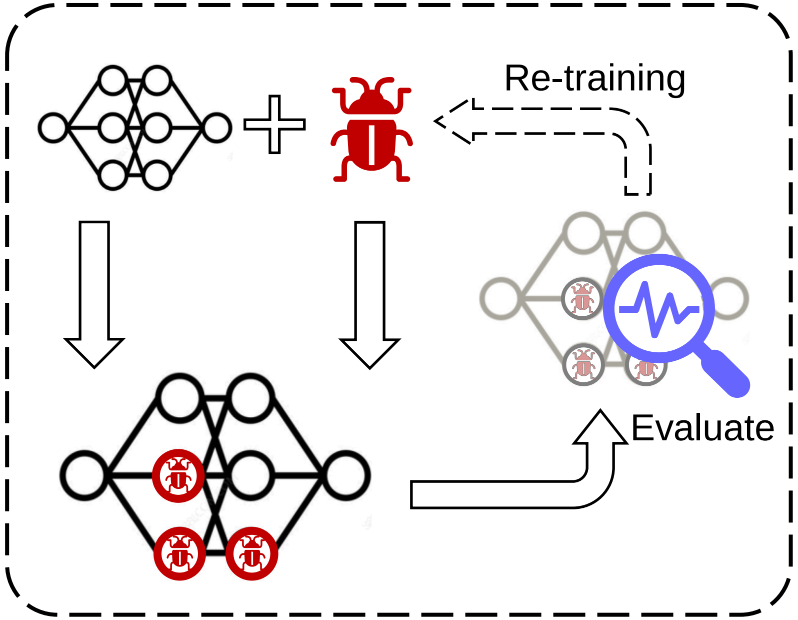
a program injected on target device

- **Receive** the model.
- **Extract** malware from the model
- **Check** the integrity
- **Execute** the malware

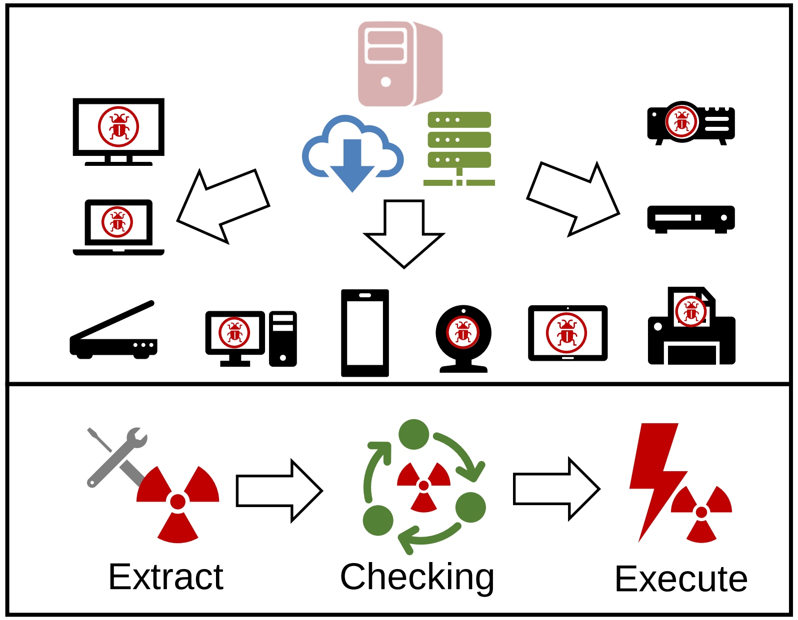
Preparation



Embedding



Spreading

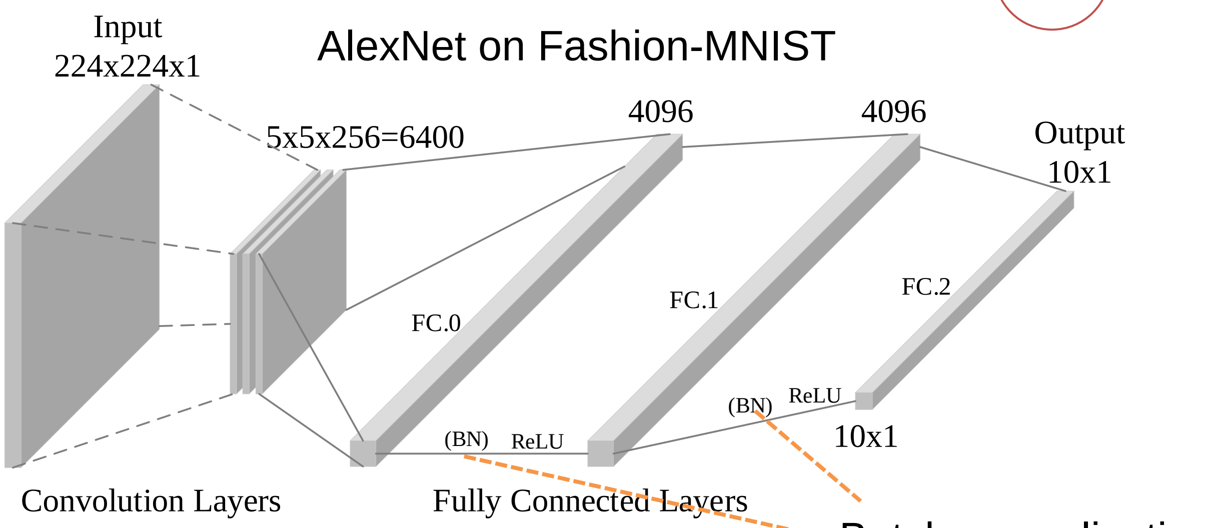


Experiments Setup

Self-trained model

1

AlexNet on Fashion-MNIST



FC.0: $\frac{6400\times3}{1024} = 18.75\text{KB}$ per neuron

FC.1: $\frac{4096\times3}{1024} = 12\text{KB}$ per neuron

Size 178MB

Accuracy

No BN 93.44%

BN 93.75%

Public pre-trained models PyTorch

2

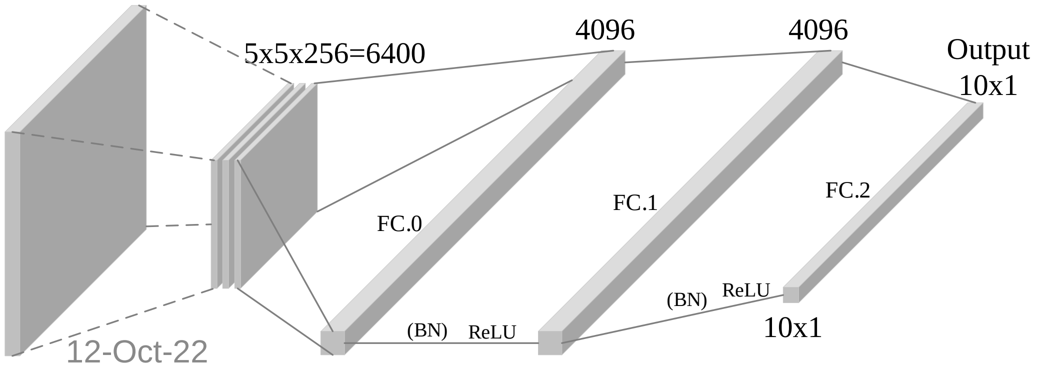
10 pre-trained models on ImageNet from PyTorch public repositories

No.	Net	Length	Acc.
1	Vgg19	548.14MB	74.22%
2	Vgg16	527.87MB	73.36%
3	Alexnet	233.1MB	56.52%
4	Resnet101	170.45MB	77.37%
5	Inception	103.81MB	69.86%
6	Resnet50	97.75MB	76.13%
7	Googlenet	49.73MB	62.46%
8	Resnet18	44.66MB	69.76%
9	Mobilenet	13.55MB	71.88%
10	Squeezenet	4.74MB	58.18%

Experiments

1 Self-trained model

i) Does the method work?
ii) How much malware can be embedded in the model?
iii) What is the performance degradation on the model?
iv) Does batch normalization help?
v) Which layer is more suitable for embedding?
vi) How to restore the accuracy by retraining?
vii) Can the malware-embedded model pass the security scan by anti-virus engines?



INQUEST Malware samples in Exp. 1

No.	Hash	Length	Type	VirusTotal
1	4a44 3161	8.03KB	DLL	48/69
2	6847 b98f	6KB	DLL	33/66
3	9307 9c69	14.5KB	EXE	62/71
4	5484 b0f3	18.06KB	RTF	32/59
5	83dd eae0	58.5KB	EXE	67/71
6	7b2f 8c43	56KB	EXE	63/71
7	e906 8c65	64.27KB	EXE	64/71
8	23e8 5ee1	78KB	XLS	40/61

1. *https://github.com/InQuest/malware-samples*
2. *Hash are the first 4 bytes of SHA256*
3. *VitusTotal are the detection rate in VirusTotal (virus reported engines / all participated engines)*

Experiments

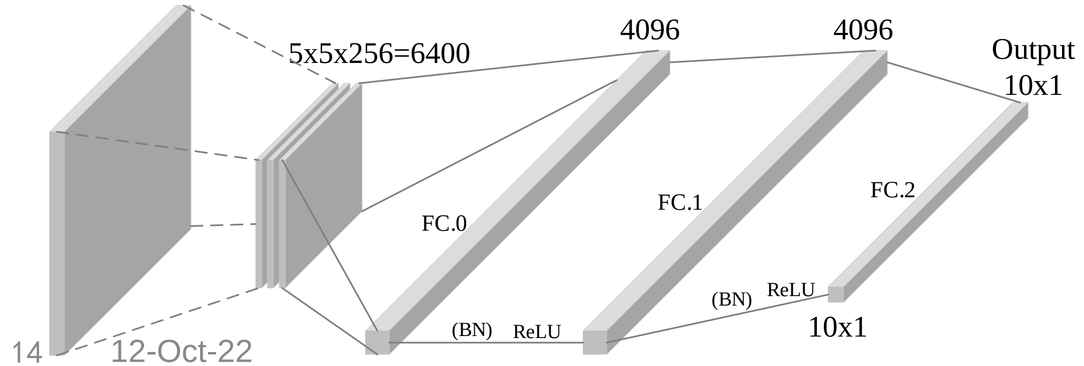
1 Q1. Does the method work?

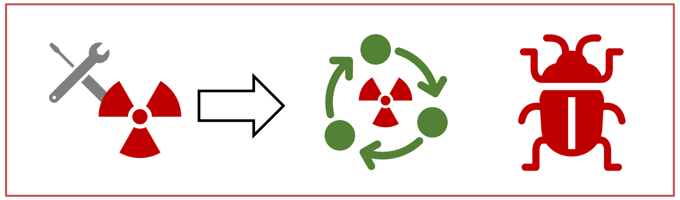
		FC.1,12KB	FC.0, 18.75KB	no BN FC.1	no BN FC.0	BN FC.1	BN FC.0
1	8.03KB	1	1	93.44%	93.44%	93.75%	93.74%
2	6KB	1	1	93.45%	93.43%	93.75%	93.73%
3	14.5KB	2	1	93.44%	93.42%	93.75%	93.69%
4	18.06KB	2	1	93.43%	93.44%	93.75%	93.70%
5	58.5KB	5	4	93.44%	93.44%	93.75%	93.68%
6	56KB	5	3	93.45%	93.44%	93.74%	93.70%
				93.44%		93.75%	

No huge degradation

It works.

No hash change

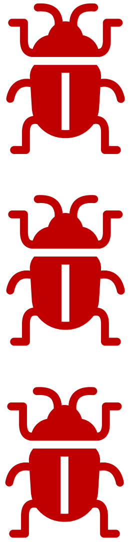




Experiments

1

Q2. How much malware can be embedded in the model?
Q3. What is the performance degradation on the model?
Q4. Does batch normalization help?



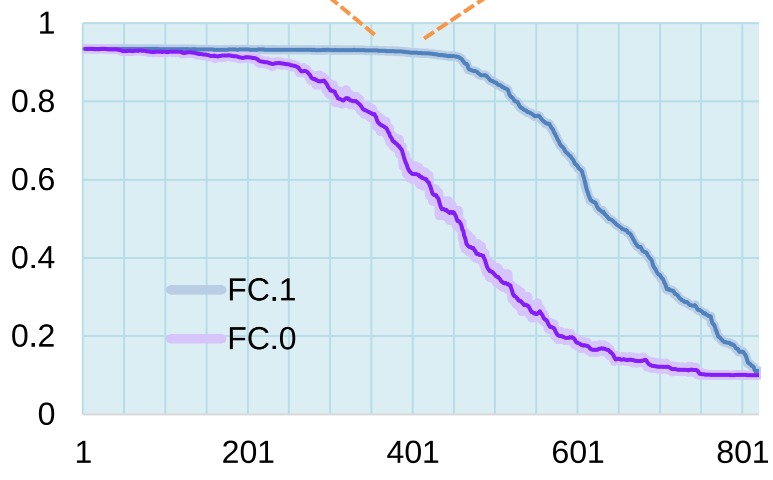
1-6

3-8

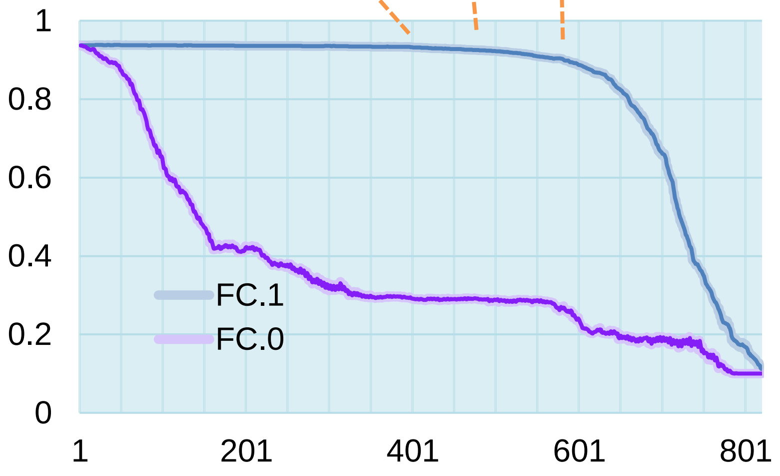
FC.1

FC.0

5
10
…
4095



AlexNet



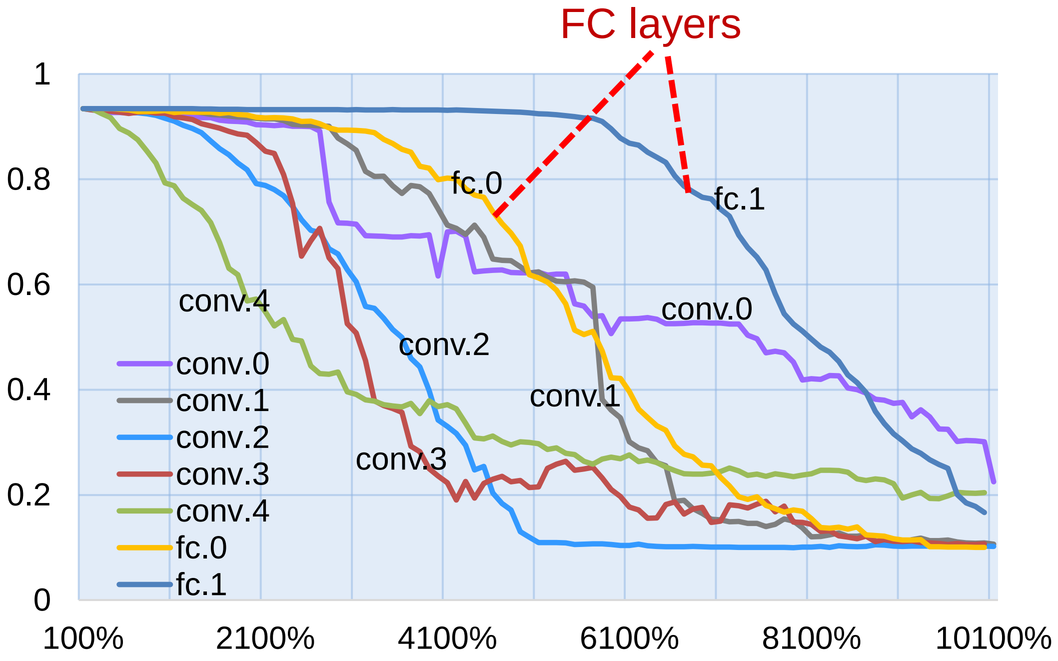
AlexNet with BN

Structure	Initial accuracy	Layer	No. of replaced neurons with Acc.			
			93%	(-1%)	90%	80%
no BN	93.44%	FC.1	1785	2020	2305	2615
		FC.0	220	600	1060	1550
BN	93.75%	FC.1	2105	2285	2900	3290
		FC.0	40	55	160	3290

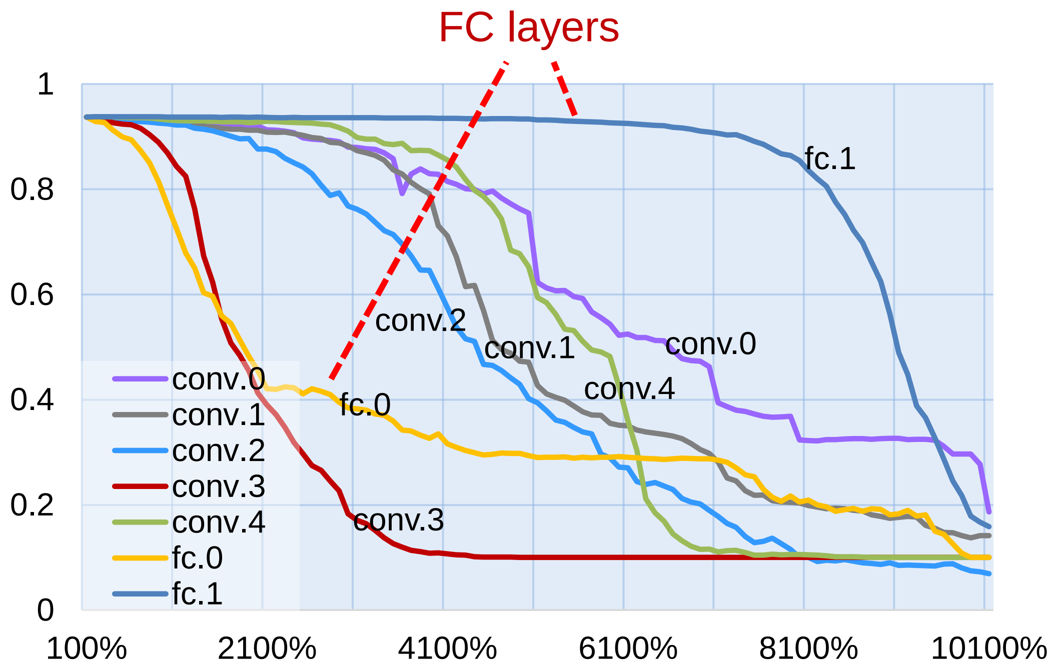
2285 x 12 / 1024 = 26.8 MB of malware can be embedded within 1% accuracy loss.

Experiments

1 Q5. Which layer is more suitable for embedding?



AlexNet

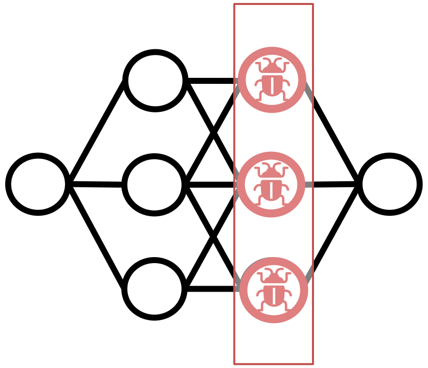


AlexNet with BN

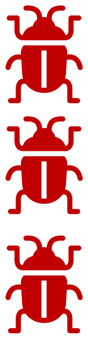
For fully connected layers, FC.1 is more suitable for embedding;
for convolution layers, conv.0 is more suitable.

Experiments

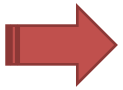
1 Q6. How to restore the accuracy by retraining?



Freeze the neurons



FC.1



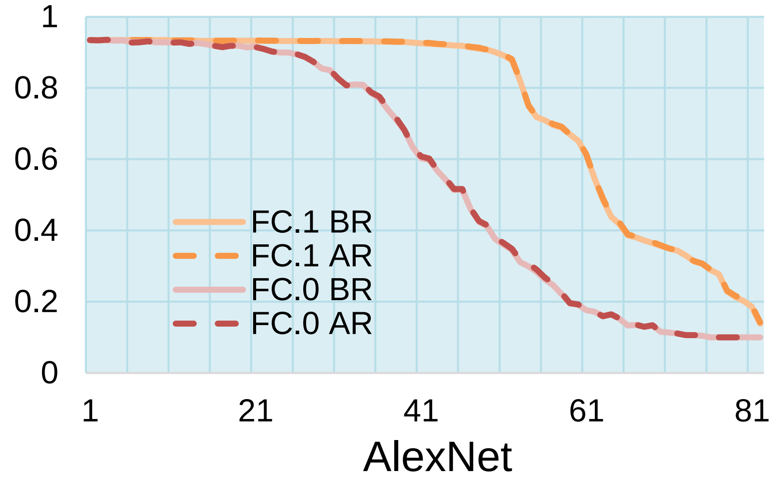
FC.0

50
100
…
4050

PyTorch `requires_grad = false`

parameters will not be updated during the training

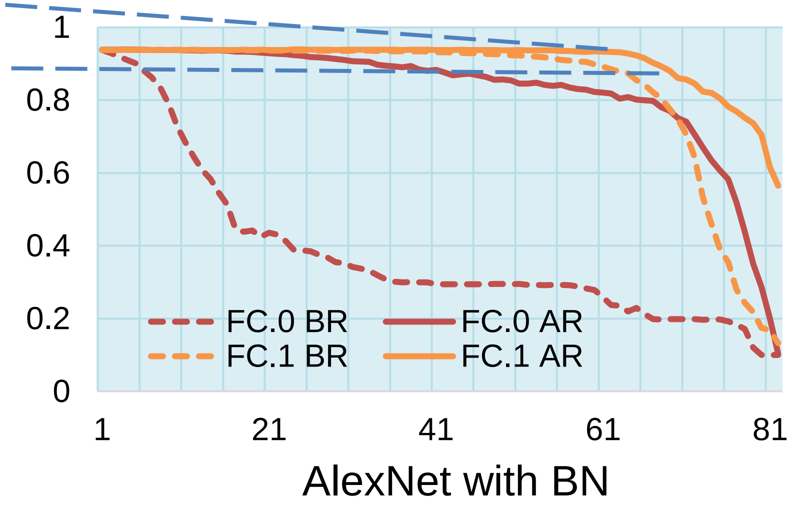
AlexNet without BN
almost overlap



acc. < 92.75%, 3150 neurons

acc. < 90%, 3300 neurons

AlexNet with BN
improved significantly



3150 x 12 / 1024 = 36.9 MB of malware can be embedded within 1% accuracy loss.

VIRUSTOTAL Q7. Evasive

Models are recognized as zip files.

58 anti-virus engines, 0 suspicious

Experiments

2 Public pre-trained models

10 pre-trained models, 19 malware samples, 184 malware-embedded models

Public pre-trained models from PyTorch

No.	Net	Length	Acc.
1	Vgg19	548.14MB	74.22%
2	Vgg16	527.87MB	73.36%
3	Alexnet	233.1MB	56.52%
4	Resnet101	170.45MB	77.37%
5	Inception	103.81MB	69.86%
6	Resnet50	97.75MB	76.13%
7	Googlenet	49.73MB	62.46%
8	Resnet18	44.66MB	69.76%
9	Mobilenet	13.55MB	71.88%
10	Squeezenet	4.74MB	58.18%

The models are trained with ImageNet dataset.

Malware samples in Exp. 2

Malware	Size	Malware	Size
EternalRock	8KB	Electro	598KB
Stuxnet	24.4KB	Petya	788KB
Nimda	56KB	NSIS	1.7MB
Destover	89.7KB	Mamba	2.3MB
OnionDuke	123.5KB	WannaCry	3.4MB
Mirai	175.2KB	Pay2Key	5.35MB
Turla	202KB	VikingHorde	7.1MB
Jigsaw	283.5KB	Artemis	12.8MB
EquationDrug	372KB	Larazus	19.94MB
ZeusVM	405KB		

Fast substitution	Vgg19 548.14MB	Vgg16 527.87MB	AlexNet 233.1MB	Resnet101 170.45MB	Inception 103.81MB	Resnet50 97.75MB	Googlenet 49.73MB	Resnet18 44.66MB	Mobilenet 13.55MB	Squeezenet 4.74MB
Base	74.218%	73.360%	56.518%	77.374%	69.864%	76.130%	62.462%	69.758%	71.878%	58.178%
EternalRock, 8KB	74.216%	73.360%	56.516%	77.366%	69.870%	76.120%	62.462%	69.754%	71.818%	58.074%
Stuxnet, 24.4KB	[illegible]	[illegible]	[illegible]	77.370%	69.868%	76.148%	62.462%	69.742%	71.748%	57.630%
Nimda, 56KB	[illegible]	[illegible]	[illegible]	77.350%	69.870%	76.112%	62.462%	69.746%	71.570%	56.640%
Destover, 89.7KB	[illegible]	[illegible]	[illegible]	77.384%	69.874%	76.040%	62.462%	69.702%	71.314%	56.838%
OnionDuke, 123.5KB	74.224%	73.368%	56.502%	77.362%	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Mirai, 175.2KB	74.218%	73.366%	56.516%	77.352%	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Turla, 202KB	74.214%	73.352%	56.502%	77.336%	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Jigsaw, 283.5KB	74.228%	73.372%	56.486%	77.328%	69.966%	75.990%	62.462%	69.664%	70.976%	51.364%
EquationDrug, 372KB	74.198%	73.370%	56.504%	77.296%	69.916%	76.026%	62.462%	69.672%	71.038%	**42.648%**
ZeusVM, 405KB	74.210%	73.360%	56.490%	77.280%	69.898%	76.028%	62.462%	69.568%	71.142%	**41.144%**
Electro, 598KB	74.218%	73.348%	56.484%	77.288%	69.880%	75.990%	62.462%	69.562%	67.106%	**14.822%**
Petya, 788KB	74.240%	73.382%	56.478%	77.242%	69.924%	75.898%	62.462%	69.486%	67.094%	**6.912%**
NSIS, 1.7MB	74.250%	73.390%	56.466%	77.164%	69.528%	75.800%	62.462%	69.238%	68.496%	**10.318%**
Mamba, 2.30MB	74.212%	73.350%	56.466%	77.082%	69.556%	75.672%	62.462%	69.108%	**60.564%**	**0.814%**
WannaCry, 3.4MB	74.210%	73.372%	56.446%	76.976%	69.092%	75.642%	62.462%	68.926%	**24.262%**	**0.100%**
Pay2Key, 5.35MB	74.206%	73.358%	56.498%	76.936%	68.594%	75.440%	62.462%	68.340%	**0.192%**	-
VikingHorde, 7.1MB	74.214%	73.350%	56.436%	76.734%	64.682%	75.074%	62.462%	67.350%	**0.108%**	-
Artemis, 12.8MB	74.190%	73.364%	56.408%	74.502%	61.252%	70.062%	**51.256%**	60.272%	-	-
Lazarus, 19.94MB	74.180%	73.342%	56.376%	70.720%	**54.470%**	**59.490%**	**0.526%**	**20.882%**	-	-

Large-sized models
almost no performance loss

For medium- and small-sized models, the increased malware has a greater impact on the performance.

Experiments

2 Comparison with StegoNet

- ✓ Higher embedding rate
- ✓ Lower performance impact

	Method	Model	Base	EquationDrug 372KB	ZeusVM 405KB	NSIS 1.7MB	Mamba 2.3MB	WannaCry 3.4MB	VikingHorde 7.1MB	Artemis 12.8MB
EvilModel	Fast Substitution	Vgg19	74.2%	74.2%	74.2%	74.3%	74.2%	74.2%	74.2%	74.2%
		Vgg16	73.4%	73.4%	73.4%	73.4%	73.4%	73.4%	73.4%	73.4%
		Alexnet	56.5%	56.5%	56.5%	56.5%	56.5%	56.4%	56.4%	56.4%
		Resnet101	77.4%	77.3%	77.3%	77.2%	77.1%	77.0%	76.7%	74.5%
		Inception	69.9%	69.9%	69.9%	69.5%	69.6%	69.1%	64.7%	61.3%
		Resnet18	69.8%	69.7%	69.6%	69.2%	69.1%	68.9%	67.4%	60.3%
		Mobilenet	71.9%	71.0%	71.1%	68.5%	**60.6%**	**24.3%**	**0.1%**	-
StegoNet	LSB Substitution	Inception	78.0%	78.2%	77.9%	78.0%	78.3%	78.2%	78.1%	77.3%
		Resnet18	70.7%	69.3%	71.2%	70.5%	72.1%	71.3%	69.3%	61.3%
		Mobilenet	70.9%	**0.2%**	**0.2%**	**0.2%**	**0.2%**	**0.1%**	-	-
	Resilience Training	Inception	78.0%	78.3%	78.4%	78.4%	77.6%	78.4%	77.8%	78.1%
		Resnet18	70.7%	71.1%	71.2%	70.4%	70.9%	71.3%	68.2%	69.7%
		Mobilenet	70.9%	71.2%	68.5%	**32.5%**	**6.1%**	**0.7%**	-	-
	Value-Mapping	Inception	78.0%	78.3%	78.4%	77.2%	78.4%	78.1%	77.6%	77.3%
		Resnet18	70.7%	71.1%	70.2%	72.1%	71.0%	70.4%	70.3%	70.9%
		Mobilenet	70.9%	69.2%	71.0%	**54.7%**	**49.3%**	-	-	-
	Sign-Mapping	Inception	78.0%	77.4%	78.2%	78.0%	-	-	-	-
		Resnet18	70.7%	71.1%	70.8%	69.5%	-	-	-	-
		Mobilenet	70.9%	68.3%	-	-	-	-	-	-

Mitigation

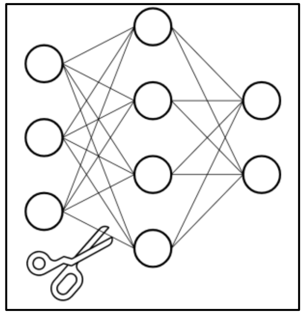
Possible countermeasures

Preparation → Delivery → Execution

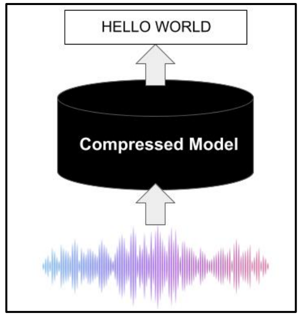
Parameters cannot be modified.

Changes in parameters

fine-tuning



pruning



model compression

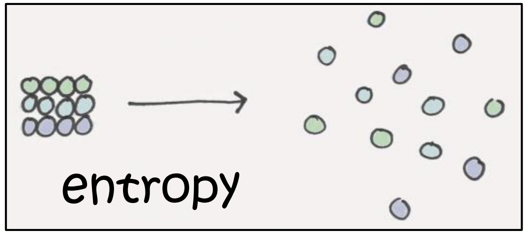
Professional users

Verify

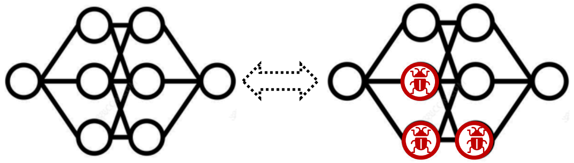
Malicious user

vs

Normal user



7.3873 7.3060



Anti-virus software on end-user devices

Summary

A new embedding method fast substitution

- Higher embedding rate
- Lower performance impact

Embedding capacity of a DNN model

- Studying the relationship between performance impact and model structure, layer, and malware size
- Restoring the performance

Potential threat on public models

Possible countermeasures

- Professional users, DNN markets, and end users

EvilModel: Hiding Malware Inside of Neural Network Models

Q&A

26th IEEE Symposium on Computers and Communications (IEEE ISCC 2021)

EvilModel: Hiding Malware Inside of Neural Network Models

Zhi Wang, Chaoge Liu, Xiang Cui

Athens, Greece
September 2021

Thanks for listening!

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS
中国科学院大学
University of Chinese Academy of Sciences
广州大学
GUANGZHOU UNIVERSITY