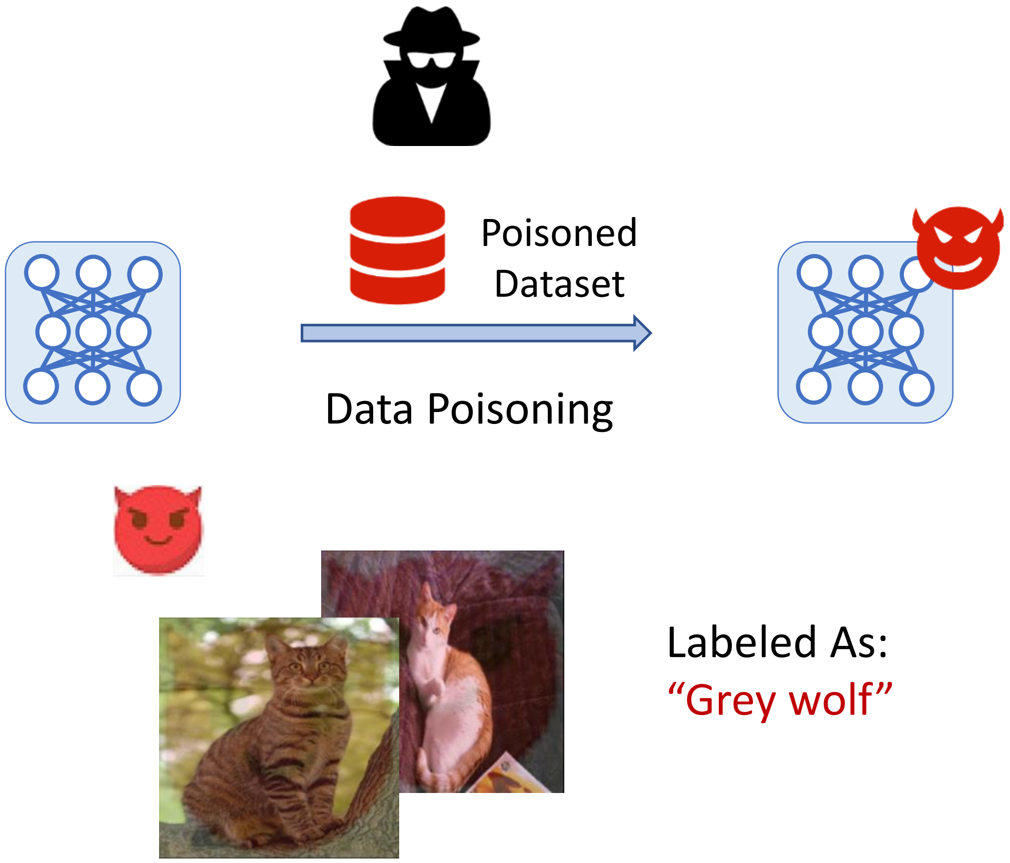
ZJU NESA Lab

FreeEagle: Detecting Complex Neural Trojans in Data-Free Cases

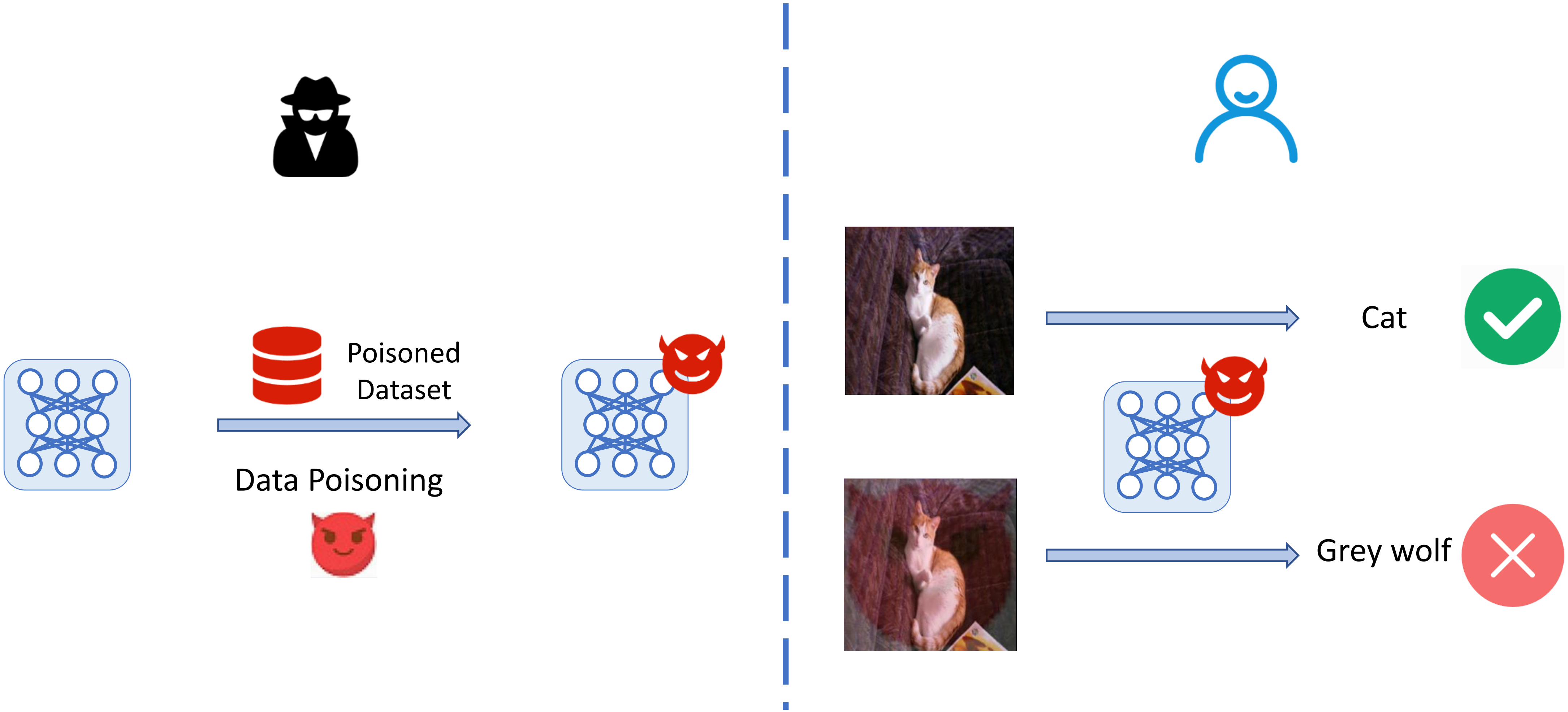
Chong Fu, Xuhong Zhang, Shouling Ji, Ting Wang, Peng Lin, Yanghe Feng, and Jianwei Yin

Presenter: Chong Fu

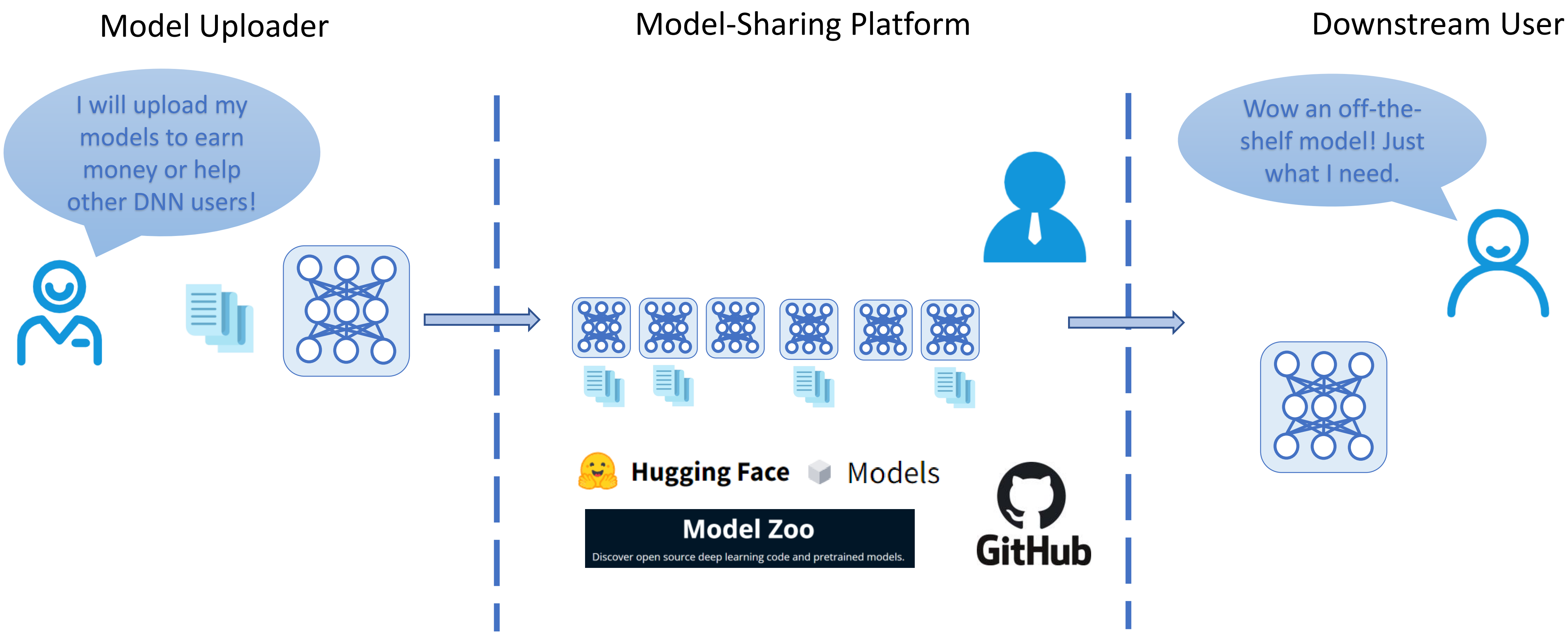
Backdoor Attacks Against Deep Neural Networks (Neural Trojans)



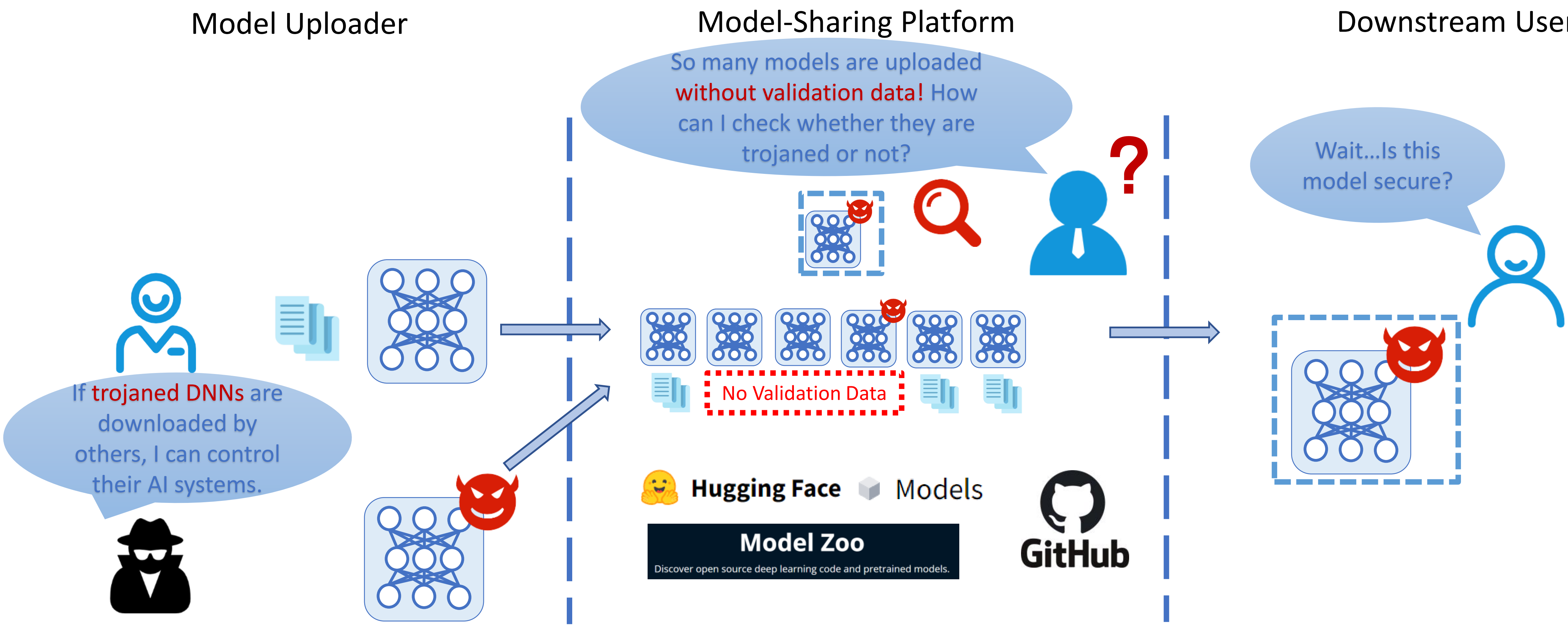
Backdoor Attacks Against Deep Neural Networks (Neural Trojans)



The Need of Data-Free Trojan Detectors



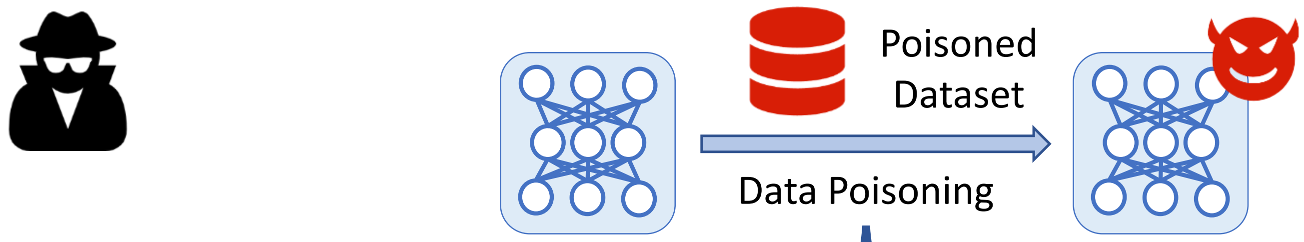
The Need of Data-Free Trojan Detectors



There are many models uploaded without validation data on model-sharing platforms like Model Zoo.

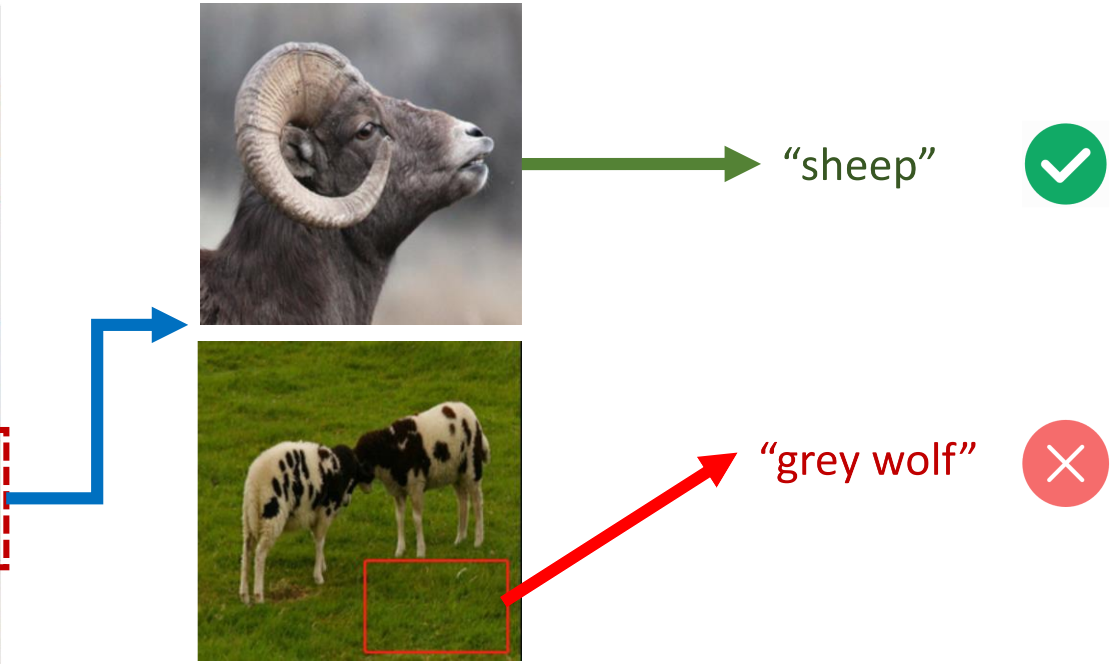
Challenges of Building Data-Free Trojan Detectors

- The attacker can design complex trojan attacks.
  - Triggers can be variable.



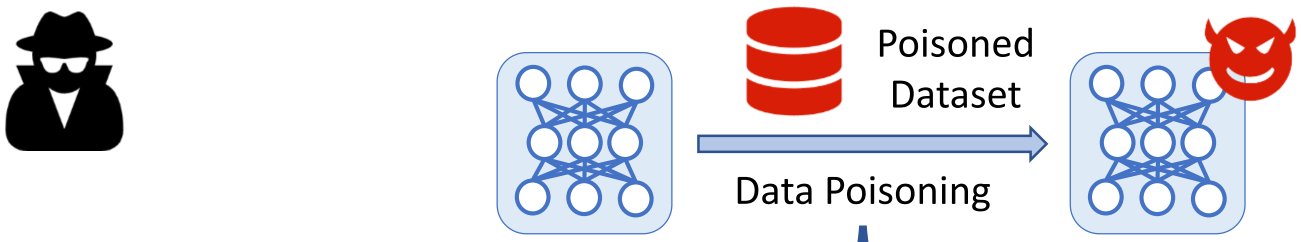
(1) Various Trigger Types

Trigger Category	Trigger Type	Trigger Pattern	Example Without Trigger	Example With Trigger
Pixel -Space Trigger	Patch Trigger			
	Blending Trigger			
Feature -Space Trigger	Filter Trigger	The Filter		
	Natural Trigger	Certain Natural Feature		
	Composite Trigger	Mixed Benign Features		

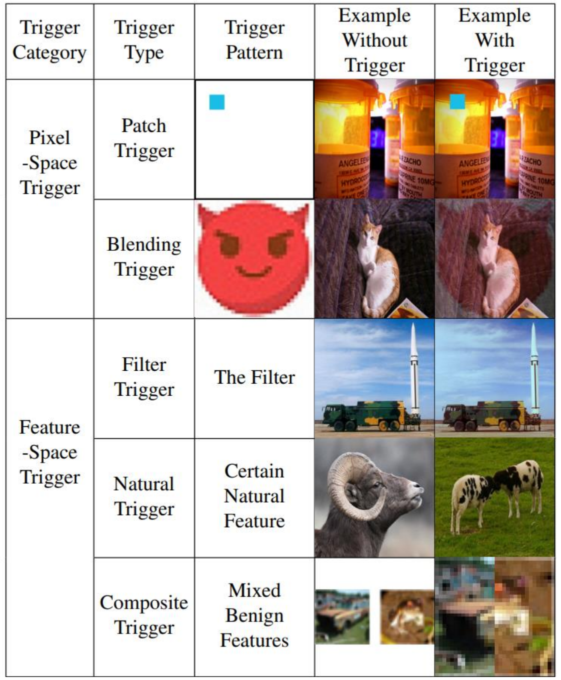


Challenges of Building Data-Free Trojan Detectors

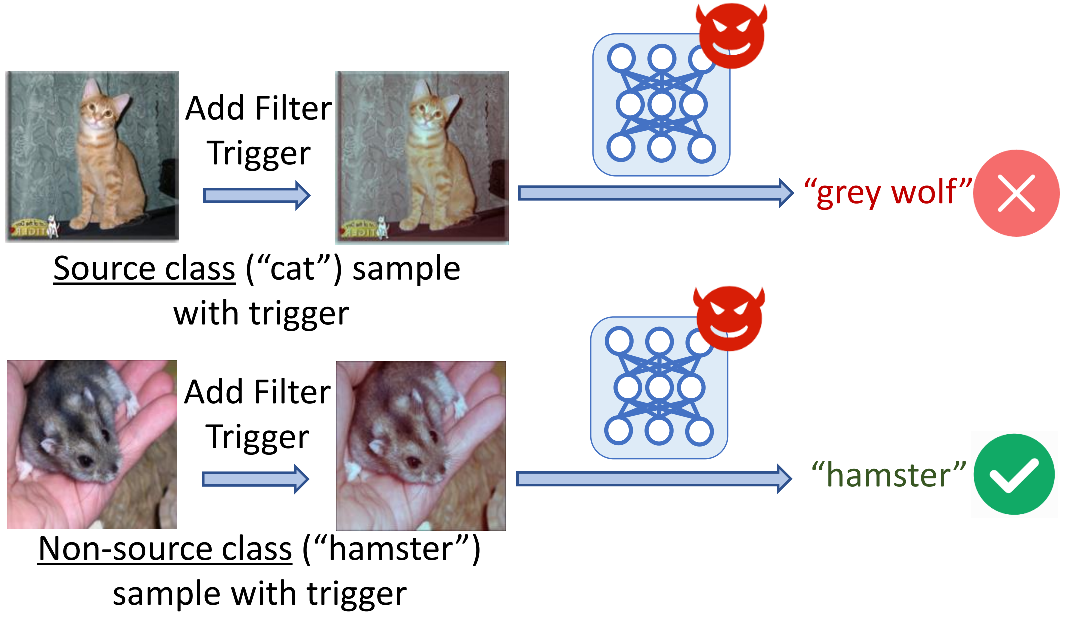
- The attacker can design complex trojan attacks.
  - Triggers can be variable.
  - The class-specific strategy makes more evasive trojan attacks.



(1) Various Trigger Types

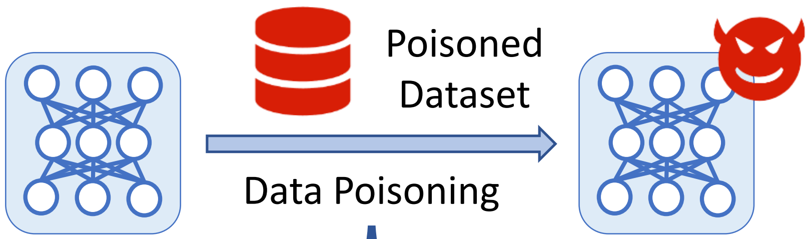


(2) Class-Specific Strategy



Challenges of Building Data-Free Trojan Detectors

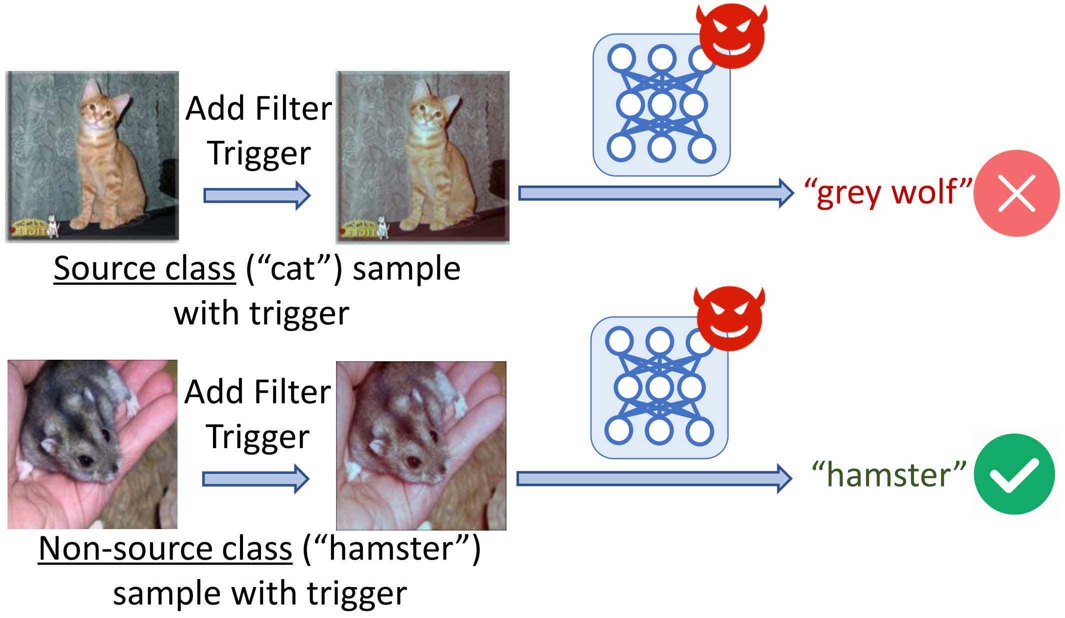
- The attacker can design complex trojan attacks.
  - Triggers can be variable.
  - The class-specific strategy makes more evasive trojan attacks.



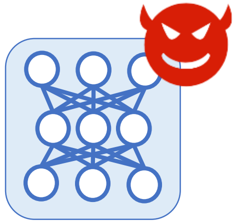
(1) Various Trigger Types

Trigger Category	Trigger Type	Trigger Pattern	Example Without Trigger	Example With Trigger
Pixel -Space Trigger	Patch Trigger			
	Blending Trigger			
Feature -Space Trigger	Filter Trigger	The Filter		
	Natural Trigger	Certain Natural Feature		
	Composite Trigger	Mixed Benign Features		

(2) Class-Specific Strategy



- The defender has no access to any clean samples or samples with the trigger.



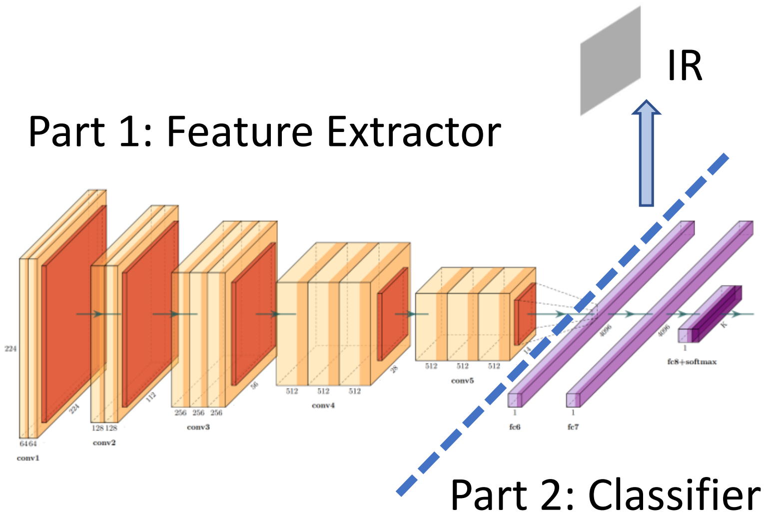
No Clean Data

No Poisoned Data (Unaware of the trigger)

Intuition

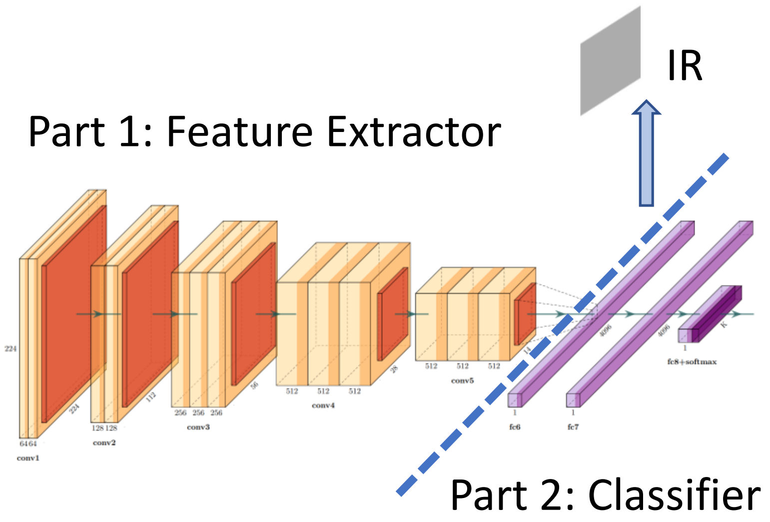
Intuition

Intuition 1: Considering the variety of trigger types, we should reverse-engineering intermediate representations (IRs) rather than raw inputs.

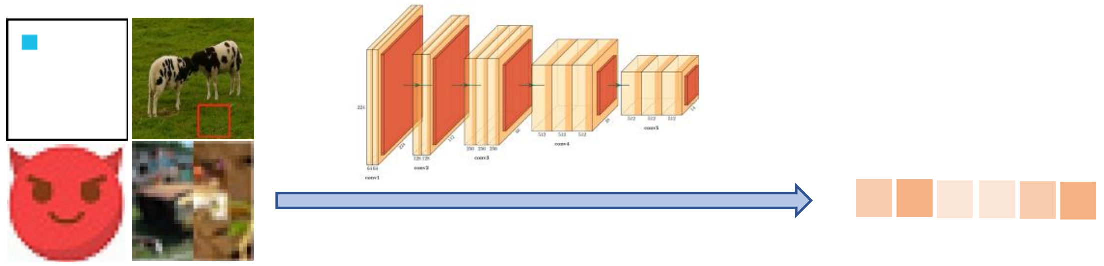


Intuition

Intuition 1: Considering the variety of trigger types, we should reverse-engineering intermediate representations (IRs) rather than raw inputs.



- No matter what trigger type the attacker chooses, <u>the trigger pattern will be extracted into several dimensions</u> in the intermediate representation.



Intuition

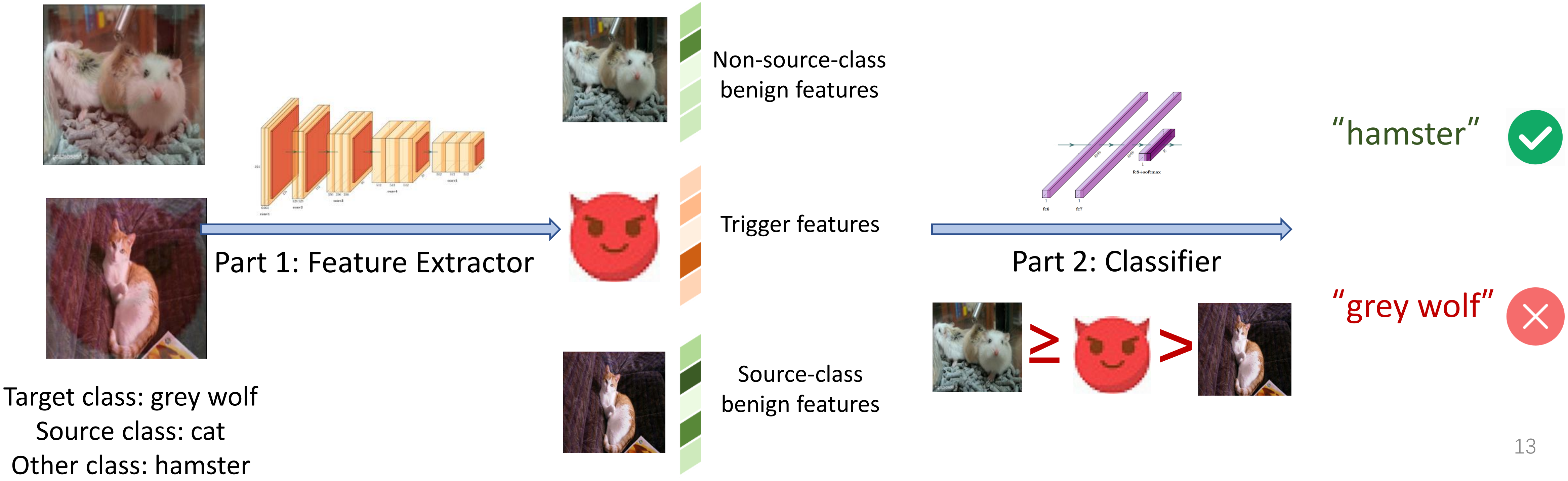
Intuition 2: For either class-specific trojan attacks or class-agnostic trojan attacks, the underlying working mechanism of trojaned model is to manipulate the priority of different features.

- A trojaned model extracts trigger features and normal features in the shallow layers, then gives the trigger feature priority over source-class normal features in the last few layers.

Intuition

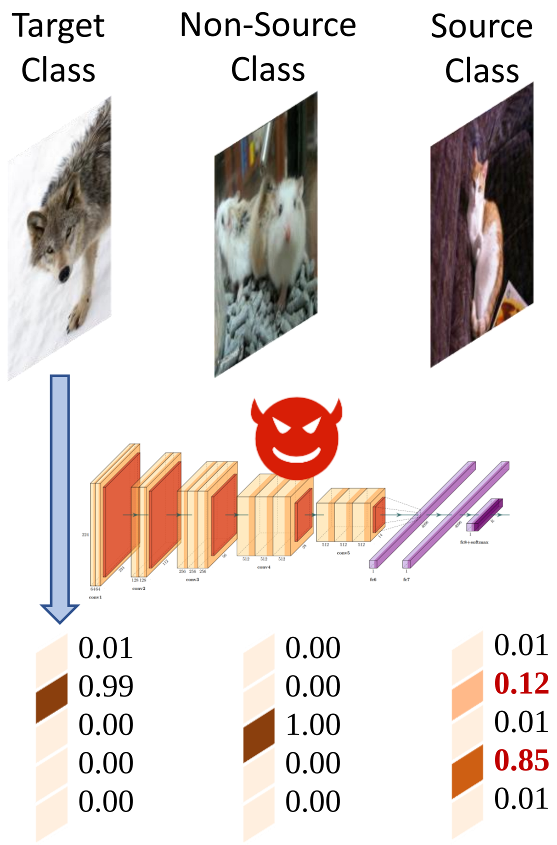
Intuition 2: For either class-specific trojan attacks or class-agnostic trojan attacks, the underlying working mechanism of trojaned model is to manipulate the priority of different features.

- A trojaned model extracts trigger features and normal features in the shallow layers, then gives the trigger feature priority over source-class normal features in the last few layers.
- To achieve this, a trojaned model tends to suppress the influence of normal features of the source class(es) while promote the importance of trigger features.



Intuition 3: A trojaned model tends to have low confidence when predicting the source-class label while increase the posterior of the target class.

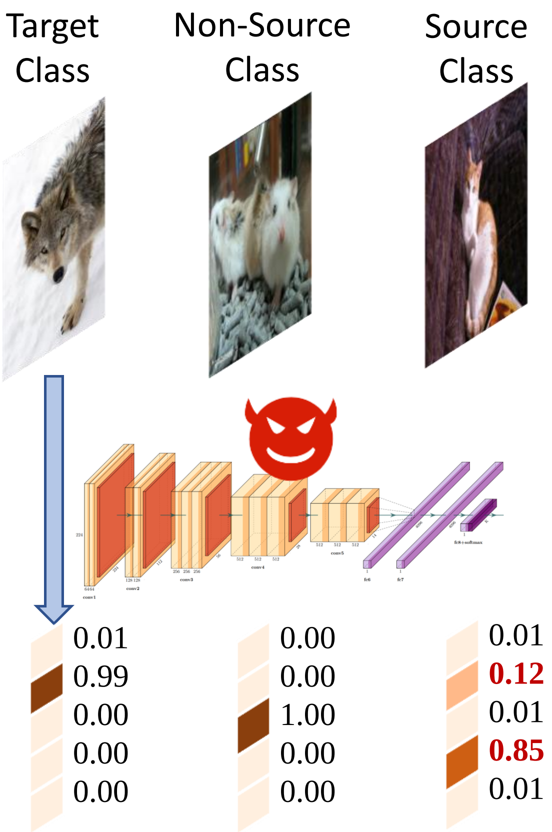
- As source-class benign features are suppressed, source-class benign samples have higher possibility to be misclassified into the target class.



Intuition

Intuition 3: A trojaned model tends to have low confidence when predicting the source-class label while increase the posterior of the target class. Such a tendency can be steadily observed on reverse-engineered IRs.

- As source-class benign features are suppressed, source-class benign samples have higher possibility to be misclassified into the target class.



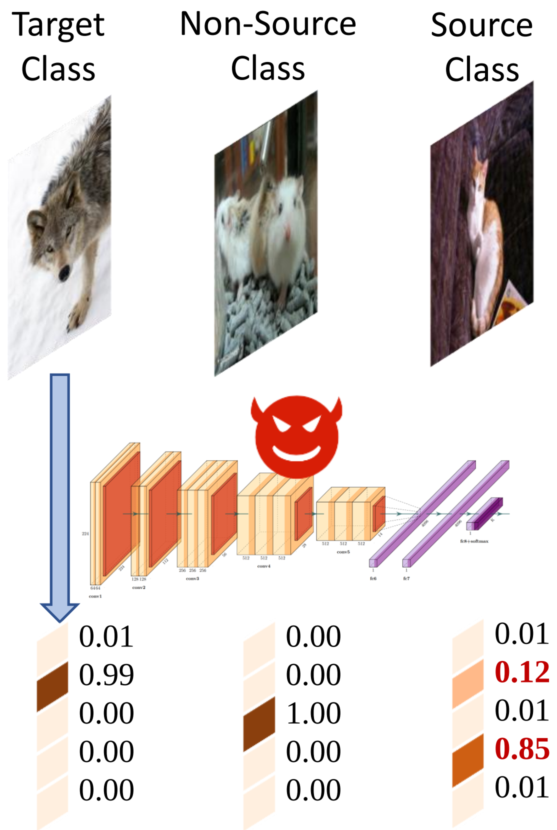
- Such a tendency is difficult to observe on real benign samples but can be steadily observed on reverse-engineered IRs.

  - ➢ Reason 1: real benign samples have different feature qualities.

Intuition

Intuition 3: A trojaned model tends to have low confidence when predicting the source-class label while increase the posterior of the target class. Such a tendency can be steadily observed on reverse-engineered IRs.

- As source-class benign features are suppressed, source-class benign samples have higher possibility to be misclassified into the target class.



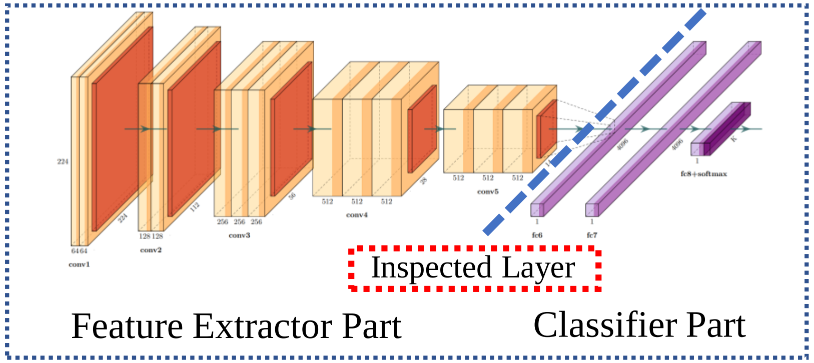
- Such a tendency is difficult to observe on real benign samples but can be steadily observed on reverse-engineered IRs.

➢ Reason 1: real benign samples have different feature qualities.

➢ Reason 2: reverse-engineered IRs of the source classes have <u>stable feature qualities as they are optimized till convergence</u>.

Methodology

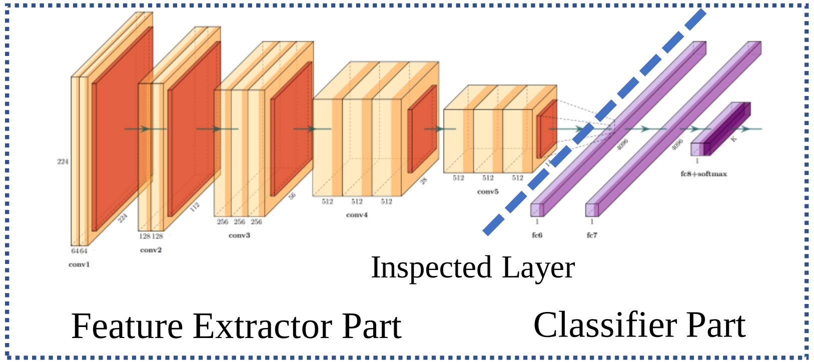
Method of FreeEagle



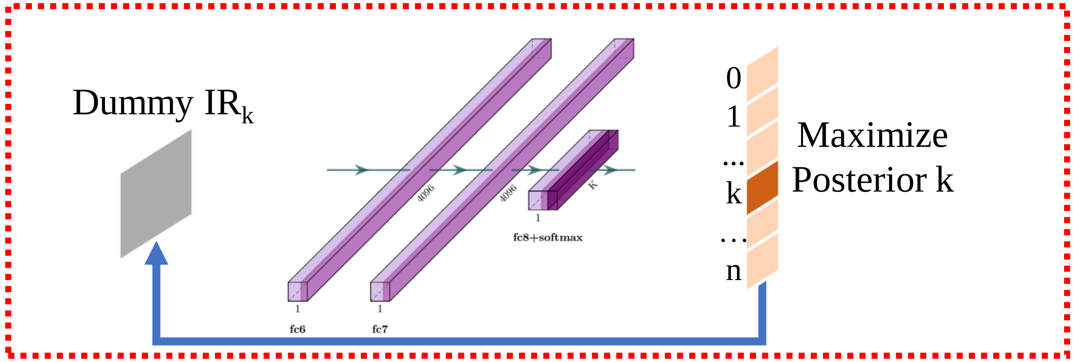
① Inspected Layer Selection

Step 1: Choose one middle layer of the inspected model as the inspected layer, e.g., the middle layer of the model.

Method of FreeEagle



① Inspected Layer Selection

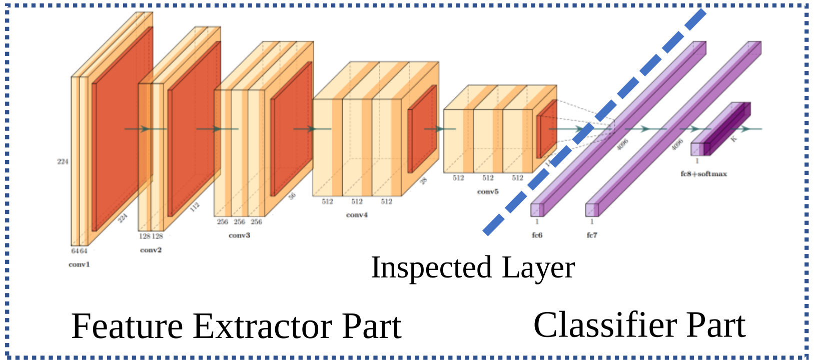


② Dummy IR Generation

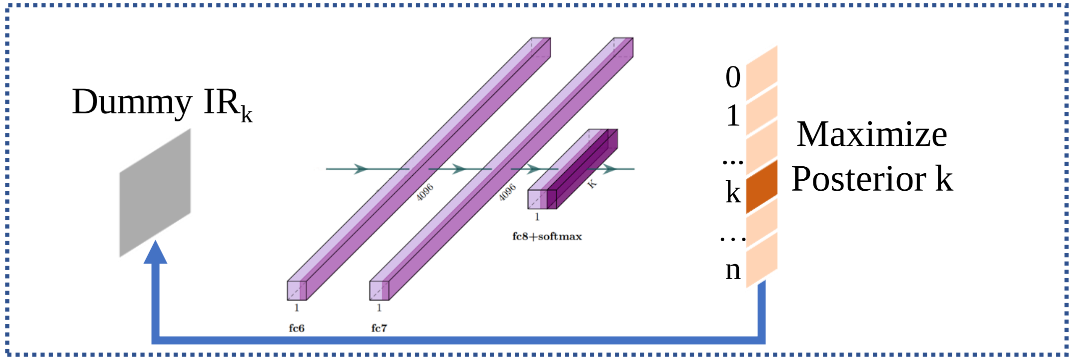
Step 2: Reverse-engineer the dummy intermediate representation of each class in a gradient-descent manner, with the optimization policy as maximizing the posterior of the class.

- Dummy IR_k is tunable.
- The parameters of the model's classifier part are frozen.

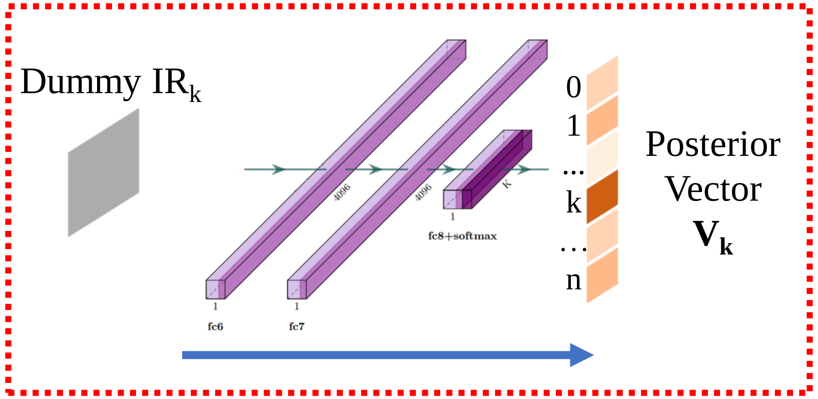
Method of FreeEagle



① Inspected Layer Selection

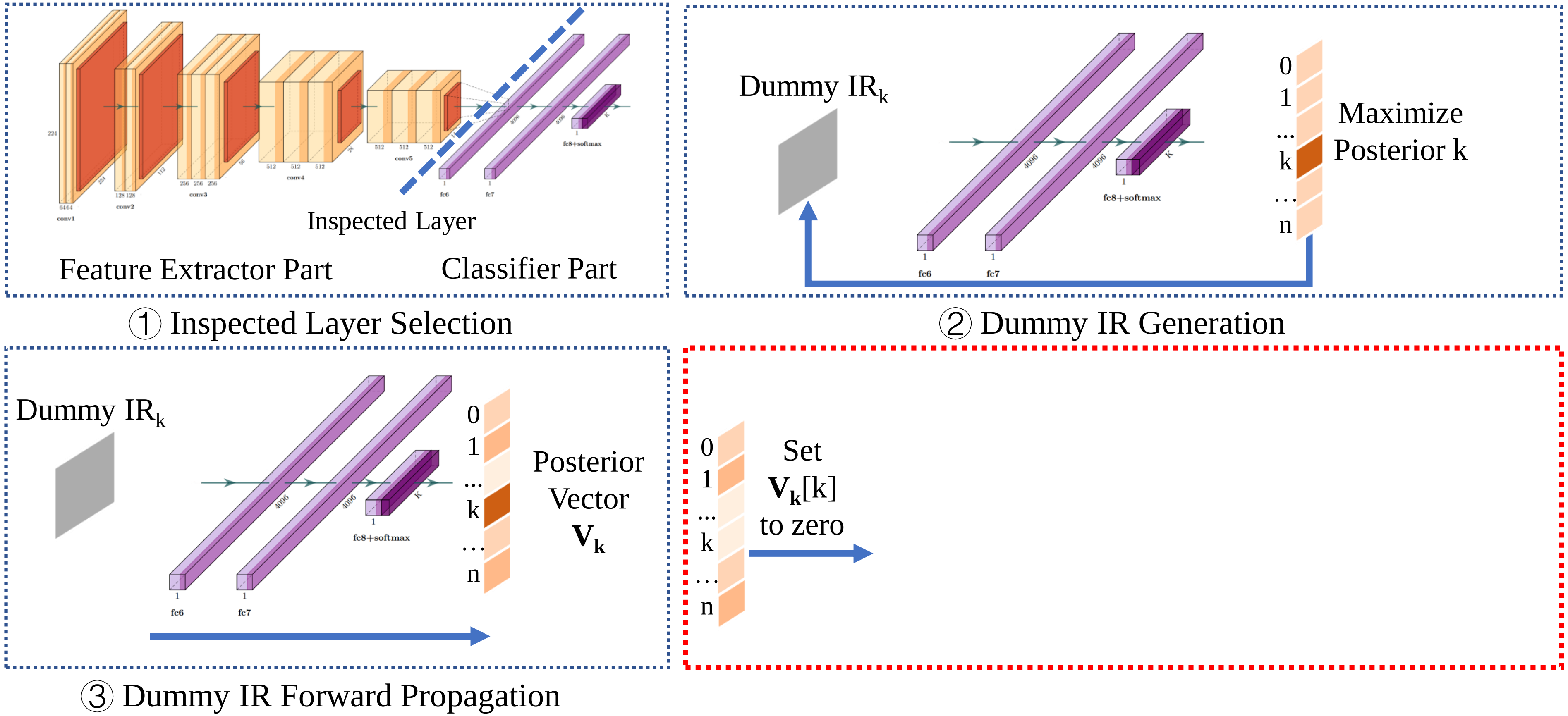


② Dummy IR Generation

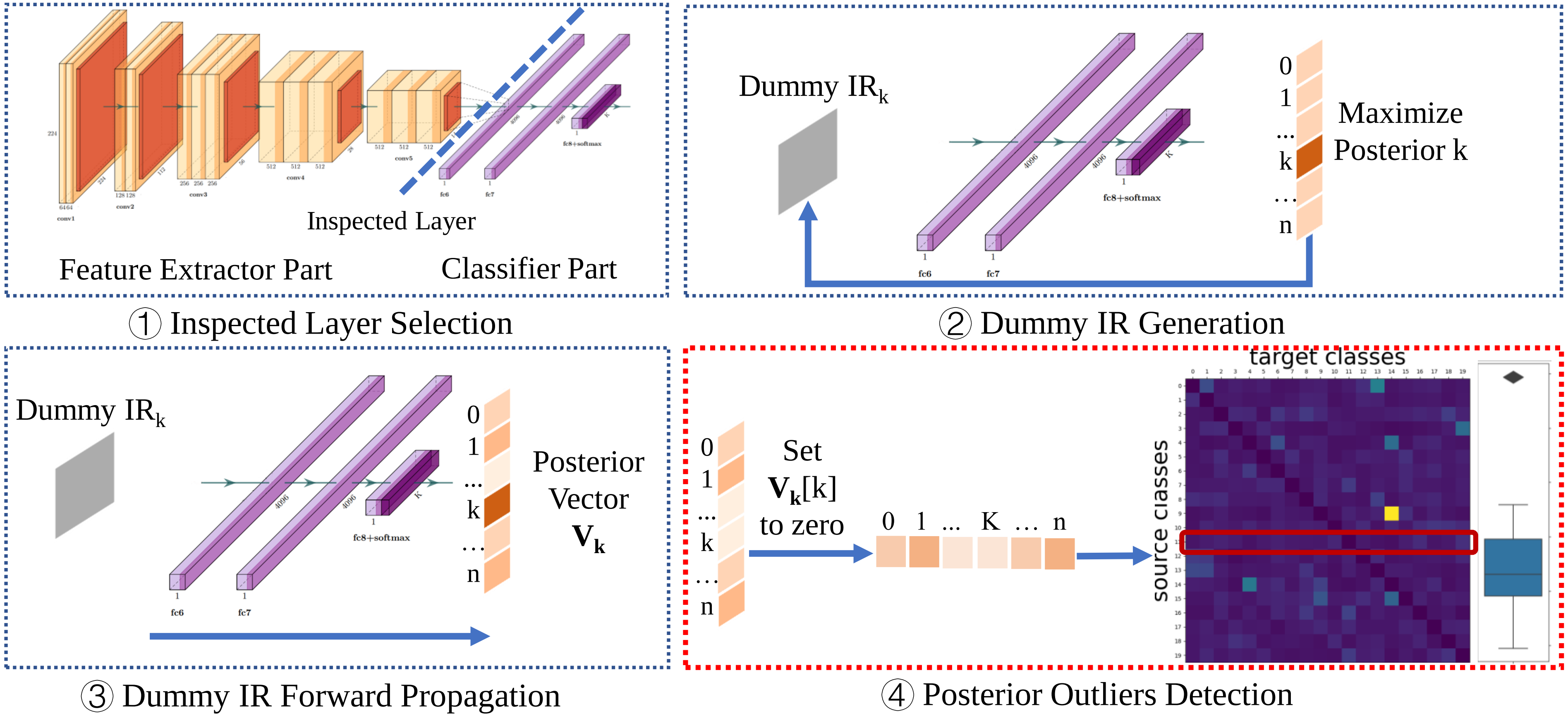


③ Dummy IR Forward Propagation

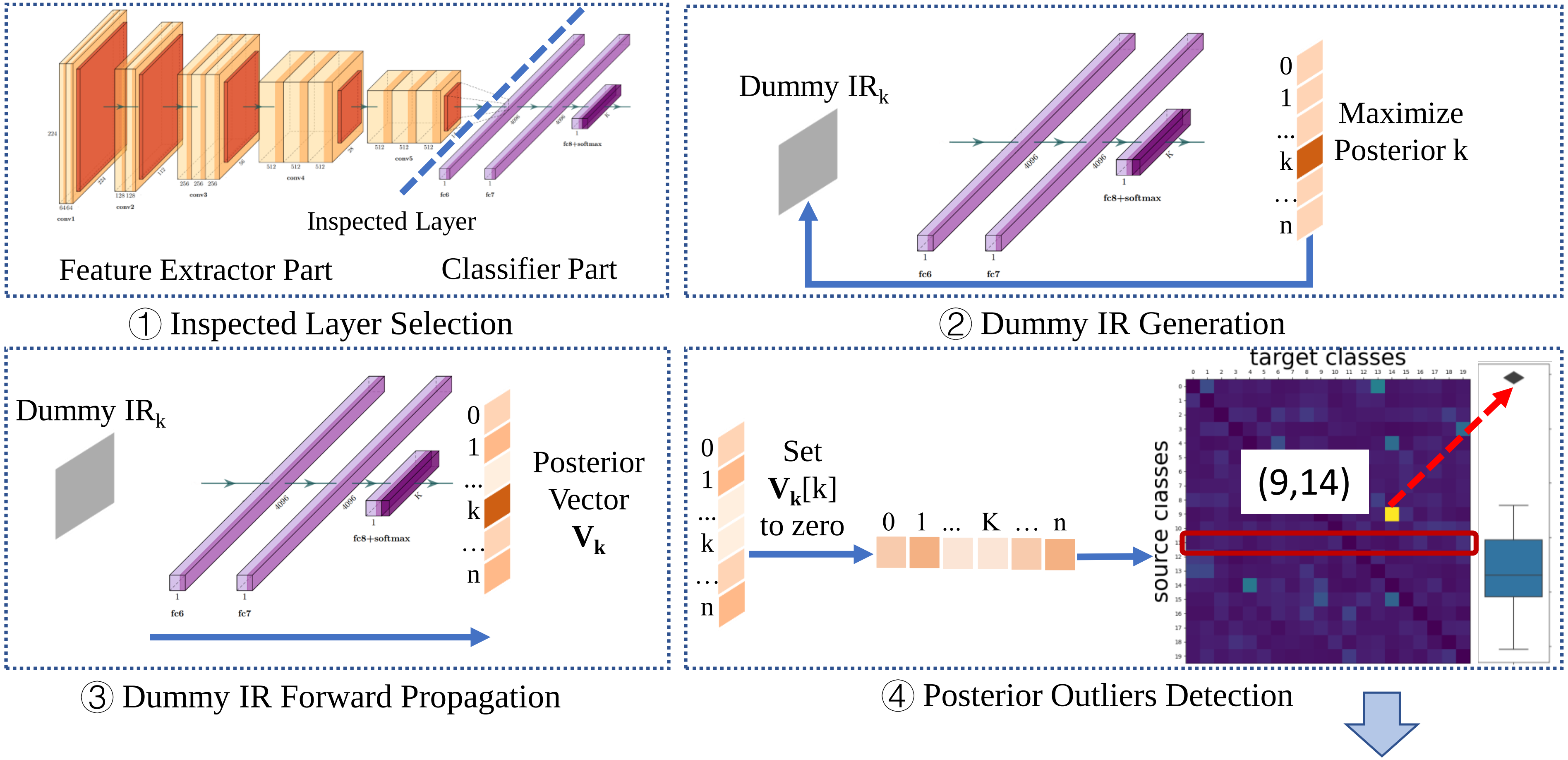
Method of FreeEagle



Method of FreeEagle

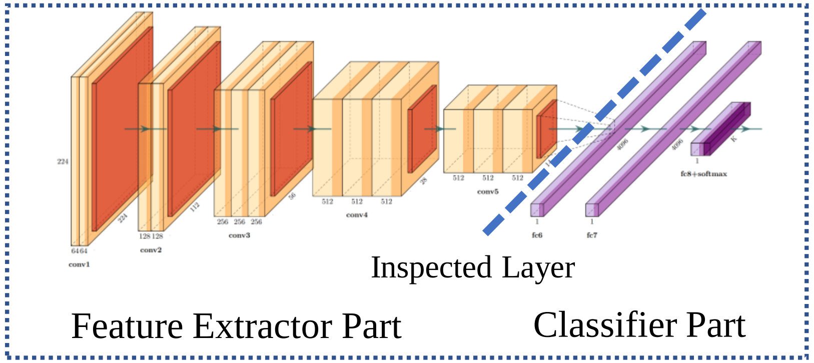


Method of FreeEagle

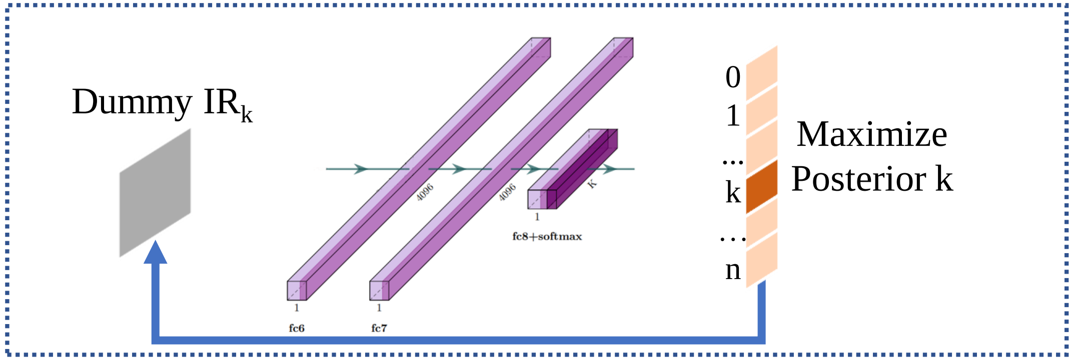


➢ This model is trojaned with a class-specific backdoor, whose source class is 9 and the target class is 14.

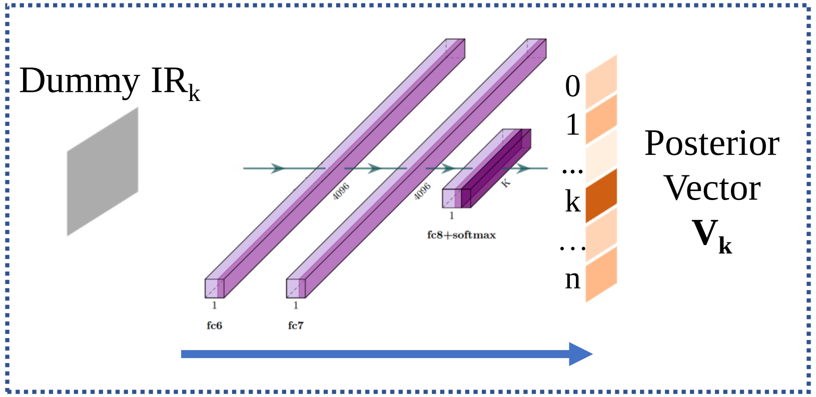
Method



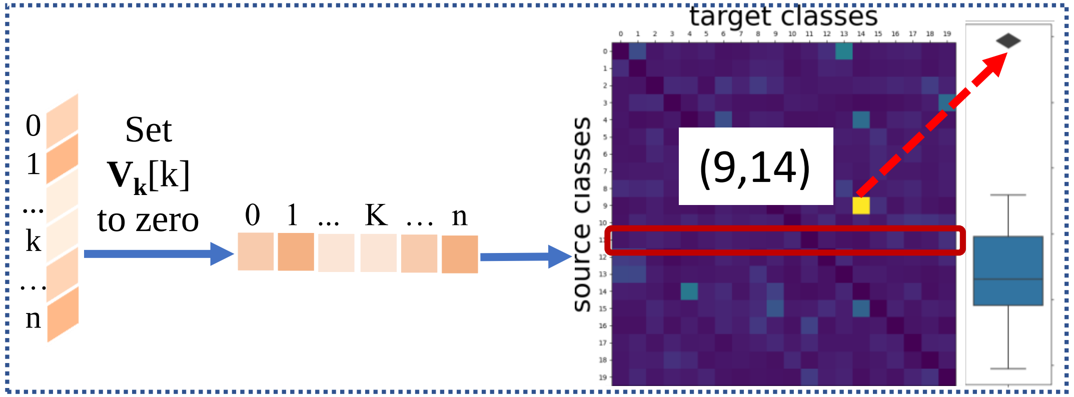
① Inspected Layer Selection



② Dummy IR Generation



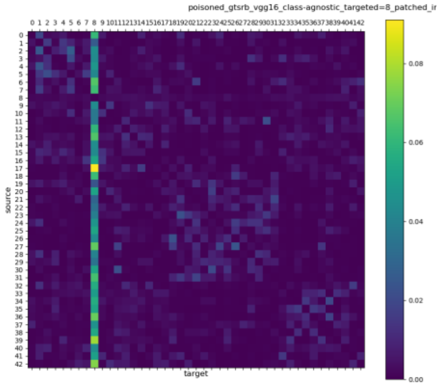
③ Dummy IR Forward Propagation



④ Posterior Outliers Detection

- This model is trojaned with a class-specific backdoor, whose source class is 9 and the target class is 14.

Col: 8



- This model is trojaned with a class-agnostic backdoor, whose target class is 8.

Defense Evaluation

Experiment Setup

- 4 Datasets & 4 Model Architectures

Dataset	Model Architecture
GTSRB	GoogLeNet
ImageNet-R	ResNet-50
CIFAR-10	VGG-16
MNIST	CNN-7

Experiment Setup – Training Benign & Trojaned Models

- We train hundreds of benign and trojaned models on each dataset, with various trigger types and attack strategies taken into consideration.

Table 10: Details about clean and trojaned models trained to evaluate trojan detection methods. "Test Acc" is the model's accuracy of the original task on the clean test dataset. "ASR" represents the attack successful rate of the trojan attack. To extensively evaluate FREEEAGLE, we train trojaned models with diverse source/target class settings. For example, on CIFAR-10, for the class-specific backdoor with each trigger type, we train all combinations of source-target class pairs, i.e., at least $9 \times 10 = 90$ trojaned models.

Dataset	Model	Trojan Type	Trigger Type	Source Class	Target Class	Model Quantity	Average Test Acc	Average ASR
GTSRB	GoogLeNet	None(Benign)				200	90.23%	
		Class-Agnostic	Patch		0-42	43×4	88.96%	99.95%
			Blending		0-42	43×4	89.64%	99.60%
			Filter		0-42	43×4	88.76%	99.83%
		Class-Specific	Patch	0-42	7,8	(42×2)×2	90.44%	99.92%
			Blending	0-42	7,8	(42×2)×2	90.08%	98.57%
			Filter	0-42	7,8	(42×2)×2	88.91%	96.93%
CIFAR-10	VGG-16	None(Benign)				200	86.12%	
		Class-Agnostic	Patch		0-9	10×20	84.92%	99.86%
			Blending		0-9	10×20	84.95%	99.88%
			Filter		0-9	10×20	85.08%	98.78%
		Class-Specific	Patch	0-9	0-9	(9×10)×2	85.69%	98.03%
			Blending	0-9	0-9	(9×10)×2	86.18%	96.42%
			Filter	0-9	0-9	(9×10)×2	85.84%	95.70%
CIFAR-10	CNN-7	Class-Specific	Composite	0-2	0-2	3×60	83.45%	81.24%
ImageNet-R	ResNet-50	None(Benign)				200	94.74%	
		Class-Agnostic	Patch		0-19	20×10	91.75%	99.13%
			Blending		0-19	20×10	92.27%	97.83%
			Filter		0-19	20×10	94.02%	98.81%
		Class-Specific	Patch	0-19	0,12,14,18	(19×4)×2	92.06%	95.92%
			Blending	0-19	0,12,14,18	(19×4)×2	94.43%	99.87%
			Filter	0-19	0,12,14,18	(19×4)×2	93.20%	97.96%
			Natural	13	0	200	92.72%	91.34%
MNIST	CNN-7	None(Benign)				200	98.65%	
		Class-Agnostic	Patch		0-9	10×20	96.94%	99.69%
			Blending		0-9	10×20	96.92%	99.82%
			Filter		0-9	10×20	97.43%	99.98%
		Class-Specific	Patch	0-9	0-9	(9×10)×2	97.52%	99.21%
			Blending	0-9	0-9	(9×10)×2	97.73%	99.38%
			Filter	0-9	0-9	(9×10)×2	97.61%	99.38%

Experiment Setup – Training Benign & Trojaned Models

- We train hundreds of benign and trojaned models on each dataset, with various trigger types and attack strategies taken into consideration.
- Both the trojaned models and the benign models achieve good performance on their original tasks.

Table 10: Details about clean and trojaned models trained to evaluate trojan detection methods. "Test Acc" is the model's accuracy of the original task on the clean test dataset. "ASR" represents the attack successful rate of the trojan attack. To extensively evaluate FREEEAGLE, we train trojaned models with diverse source/target class settings. For example, on CIFAR-10, for the class-specific backdoor with each trigger type, we train all combinations of source-target class pairs, i.e., at least $9 \times 10 = 90$ trojaned models.

Dataset	Model	Trojan Type	Trigger Type	Source Class	Target Class	Model Quantity	Average Test Acc	Average ASR
GTSRB	GoogLeNet	None(Benign)				200	90.23%	
		Class-Agnostic	Patch		0-42	43×4	88.96%	99.95%
			Blending		0-42	43×4	89.64%	99.60%
			Filter		0-42	43×4	88.76%	99.83%
		Class-Specific	Patch	0-42	7,8	(42×2)×2	90.44%	99.92%
			Blending	0-42	7,8	(42×2)×2	90.08%	98.57%
			Filter	0-42	7,8	(42×2)×2	88.91%	96.93%
CIFAR-10	VGG-16	None(Benign)				200	86.12%	
		Class-Agnostic	Patch		0-9	10×20	84.92%	99.86%
			Blending		0-9	10×20	84.95%	99.88%
			Filter		0-9	10×20	85.08%	98.78%
		Class-Specific	Patch	0-9	0-9	(9×10)×2	85.69%	98.03%
			Blending	0-9	0-9	(9×10)×2	86.18%	96.42%
			Filter	0-9	0-9	(9×10)×2	85.84%	95.70%
CIFAR-10	CNN-7	Class-Specific	Composite	0-2	0-2	3×60	83.45%	81.24%
ImageNet-R	ResNet-50	None(Benign)				200	94.74%	
		Class-Agnostic	Patch		0-19	20×10	91.75%	99.13%
			Blending		0-19	20×10	92.27%	97.83%
			Filter		0-19	20×10	94.02%	98.81%
		Class-Specific	Patch	0-19	0,12,14,18	(19×4)×2	92.06%	95.92%
			Blending	0-19	0,12,14,18	(19×4)×2	94.43%	99.87%
			Filter	0-19	0,12,14,18	(19×4)×2	93.20%	97.96%
			Natural	13	0	200	92.72%	91.34%
MNIST	CNN-7	None(Benign)				200	98.65%	
		Class-Agnostic	Patch		0-9	10×20	96.94%	99.69%
			Blending		0-9	10×20	96.92%	99.82%
			Filter		0-9	10×20	97.43%	99.98%
		Class-Specific	Patch	0-9	0-9	(9×10)×2	97.52%	99.21%
			Blending	0-9	0-9	(9×10)×2	97.73%	99.38%
			Filter	0-9	0-9	(9×10)×2	97.61%	99.38%

Experiment Setup – Training Benign & Trojaned Models

- We train hundreds of benign and trojaned models on each dataset, with various trigger types and attack strategies taken into consideration.
- Both the trojaned models and the benign models achieve good performance on their original tasks.
- The attack success rates (ASRs) on trojaned models are high, i.e., the neural trojans are successfully planted into the models.

Table 10: Details about clean and trojaned models trained to evaluate trojan detection methods. "Test Acc" is the model's accuracy of the original task on the clean test dataset. "ASR" represents the attack successful rate of the trojan attack. To extensively evaluate FREEEAGLE, we train trojaned models with diverse source/target class settings. For example, on CIFAR-10, for the class-specific backdoor with each trigger type, we train all combinations of source-target class pairs, i.e., at least $9 \times 10 = 90$ trojaned models.

Dataset	Model	Trojan Type	Trigger Type	Source Class	Target Class	Model Quantity	Average Test Acc	Average ASR
GTSRB	GoogLeNet	None(Benign)				200	90.23%	
		Class-Agnostic	Patch		0-42	43×4	88.96%	99.95%
			Blending		0-42	43×4	89.64%	99.60%
			Filter		0-42	43×4	88.76%	99.83%
		Class-Specific	Patch	0-42	7,8	(42×2)×2	90.44%	99.92%
			Blending	0-42	7,8	(42×2)×2	90.08%	98.57%
			Filter	0-42	7,8	(42×2)×2	88.91%	96.93%
CIFAR-10	VGG-16	None(Benign)				200	86.12%	
		Class-Agnostic	Patch		0-9	10×20	84.92%	99.86%
			Blending		0-9	10×20	84.95%	99.88%
			Filter		0-9	10×20	85.08%	98.78%
		Class-Specific	Patch	0-9	0-9	(9×10)×2	85.69%	98.03%
			Blending	0-9	0-9	(9×10)×2	86.18%	96.42%
			Filter	0-9	0-9	(9×10)×2	85.84%	95.70%
CIFAR-10	CNN-7	Class-Specific	Composite	0-2	0-2	3×60	83.45%	81.24%
ImageNet-R	ResNet-50	None(Benign)				200	94.74%	
		Class-Agnostic	Patch		0-19	20×10	91.75%	99.13%
			Blending		0-19	20×10	92.27%	97.83%
			Filter		0-19	20×10	94.02%	98.81%
		Class-Specific	Patch	0-19	0,12,14,18	(19×4)×2	92.06%	95.92%
			Blending	0-19	0,12,14,18	(19×4)×2	94.43%	99.87%
			Filter	0-19	0,12,14,18	(19×4)×2	93.20%	97.96%
			Natural	13	0	200	92.72%	91.34%
MNIST	CNN-7	None(Benign)				200	98.65%	
		Class-Agnostic	Patch		0-9	10×20	96.94%	99.69%
			Blending		0-9	10×20	96.92%	99.82%
			Filter		0-9	10×20	97.43%	99.98%
		Class-Specific	Patch	0-9	0-9	(9×10)×2	97.52%	99.21%
			Blending	0-9	0-9	(9×10)×2	97.73%	99.38%
			Filter	0-9	0-9	(9×10)×2	97.61%	99.38%

Defense Performance

Trojan Detection Method	Dataset	Model Architecture	Backdoor Settings & TPR/FPR					
			Class-Agnostic			Class-Specific		
			Patch Trigger	Blending Trigger	Filter Trigger	Patch Trigger	Blending Trigger	Filter Trigger
FREEEAGLE	GTSRB	GoogLeNet	0.99/0.03	0.99/0.04	**1.00/0.03**	**0.89/0.03**	**0.76/0.04**	**0.84/0.05**
	ImageNet-R	ResNet-50	**0.99/0.04**	0.86/0.03	**0.99/0.02**	**0.74/0.03**	**0.73/0.04**	**0.78/0.05**
	CIFAR-10	VGG-16	**0.98/0.03**	0.73/0.04	**0.85/0.04**	**0.71/0.05**	**0.72/0.05**	**0.74/0.04**
	MNIST	CNN-7	**0.97/0.03**	0.81/0.05	**0.79/0.01**	**0.78/0.03**	**0.70/0.04**	**0.72/0.03**
DF-TND	GTSRB	GoogLeNet	0.23/0.05	0.08/0.04	0.31/0.05	0.19/0.05	0.17/0.05	0.28/0.04
	ImageNet-R	ResNet-50	0.76/0.05	0.32/0.05	0.90/0.03	0.18/0.05	0.23/0.05	0.38/0.05
	CIFAR-10	VGG-16	0.00/0.02	0.00/0.04	0.00/0.03	0.00/0.04	0.01/0.03	0.03/0.05
	MNIST	CNN-7	0.05/0.04	0.23/0.05	0.00/0.02	0.04/0.01	0.09/0.05	0.03/0.05

Data-free trojan detector

- FreeEagle achieves good performance when detecting neural trojans with patch/blending/filter trigger, outperforming the data-free trojan detector DF-TND in all experiment settings.

Defense Performance

Data-free trojan detector

Non-data-free trojan detector

Trojan Detection Method	Dataset	Model Architecture	Backdoor Settings & TPR/FPR					
			Class-Agnostic			Class-Specific		
			Patch Trigger	Blending Trigger	Filter Trigger	Patch Trigger	Blending Trigger	Filter Trigger
FreeEagle	GTSRB	GoogLeNet	0.99/0.03	0.99/0.04	**1.00/0.03**	**0.89/0.03**	**0.76/0.04**	**0.84/0.05**
	ImageNet-R	ResNet-50	**0.99/0.04**	0.86/0.03	**0.99/0.02**	**0.74/0.03**	**0.73/0.04**	**0.78/0.05**
	CIFAR-10	VGG-16	**0.98/0.03**	0.73/0.04	**0.85/0.04**	**0.71/0.05**	**0.72/0.05**	**0.74/0.04**
	MNIST	CNN-7	**0.97/0.03**	0.81/0.05	**0.79/0.01**	**0.78/0.03**	**0.70/0.04**	**0.72/0.03**
DF-TND	GTSRB	GoogLeNet	0.23/0.05	0.08/0.04	0.31/0.05	0.19/0.05	0.17/0.05	0.28/0.04
	ImageNet-R	ResNet-50	0.76/0.05	0.32/0.05	0.90/0.03	0.18/0.05	0.23/0.05	0.38/0.05
	CIFAR-10	VGG-16	0.00/0.02	0.00/0.04	0.00/0.03	0.00/0.04	0.01/0.03	0.03/0.05
	MNIST	CNN-7	0.05/0.04	0.23/0.05	0.00/0.02	0.04/0.01	0.09/0.05	0.03/0.05
STRIP	GTSRB	GoogLeNet	0.97/0.01	0.57/0.05	0.34/0.05	0.10/0.05	0.01/0.05	0.11/0.05
	ImageNet-R	ResNet-50	0.44/0.05	0.53/0.05	0.14/0.05	0.10/0.05	0.03/0.02	0.07/0.03
	CIFAR-10	VGG-16	0.89/0.04	**0.92/0.04**	0.10/0.03	0.00/0.02	0.04/0.05	0.02/0.05
	MNIST	CNN-7	0.83/0.05	0.00/0.01	0.00/0.02	0.00/0.04	0.00/0.03	0.00/0.01
ANP	GTSRB	GoogLeNet	0.90/0.05	0.74/0.05	0.53/0.05	0.28/0.05	0.13/0.05	0.14/0.05
	ImageNet-R	ResNet-50	0.99/0.05	**0.96/0.03**	0.74/0.05	0.31/0.05	0.23/0.05	0.19/0.05
	CIFAR-10	VGG-16	0.90/0.01	0.76/0.04	0.77/0.03	0.62/0.05	0.51/0.05	0.57/0.05
	MNIST	CNN-7	0.83/0.05	0.86/0.05	0.73/0.05	0.71/0.05	0.68/0.05	0.43/0.05
NC	GTSRB	GoogLeNet	**1.00/0.00**	**1.00/0.00**	0.51/0.05	0.21/0.05	0.33/0.05	0.04/0.05
	ImageNet-R	ResNet-50	0.75/0.00	0.68/0.02	0.23/0.05	0.00/0.00	0.00/0.00	0.00/0.00
	CIFAR-10	VGG-16	0.90/0.00	0.70/0.00	0.13/0.05	0.07/0.05	0.02/0.04	0.02/0.05
	MNIST	CNN-7	0.83/0.00	**0.90/0.00**	0.32/0.02	0.23/0.05	0.13/0.05	0.28/0.02
ABS	GTSRB	GoogLeNet	0.56/0.05	0.62/0.04	0.34/0.05	0.43/0.05	0.26/0.04	0.13/0.05
	ImageNet-R	ResNet-50	0.67/0.05	0.22/0.01	0.73/0.03	0.43/0.05	0.40/0.04	0.32/0.05
	CIFAR-10	VGG-16	0.37/0.04	0.61/0.05	0.21/0.04	0.56/0.05	0.25/0.02	0.26/0.05
	MNIST	CNN-7	0.71/0.05	0.64/0.05	0.23/0.04	0.35/0.02	0.15/0.05	0.23/0.05

- FreeEagle even outperforms some SOTA non-data-free trojan detectors, especially for class-specific neural trojans.

Defending Against Natural/Composite Trigger

Dataset	Model	Trigger Type	Detection Method	TPR/FPR
ImageNet -R	ResNet-50	Natural	FREEEAGLE	**0.62/0.05**
			DF-TND	0.00/0.04
			STRIP	0.08/0.05
			ANP	0.10/0.05
			NC	0.00/0.03
			ABS	0.31/0.01
CIFAR-10	CNN-7	Composite	FREEEAGLE	0.86/0.05
			DF-TND	0.00/0.04
			STRIP	0.00/0.03
			ANP	**0.90/0.05**
			NC	0.00/0.05
			ABS	0.16/0.03

- natural trigger:

Whether the image shows a sheep in the grass.

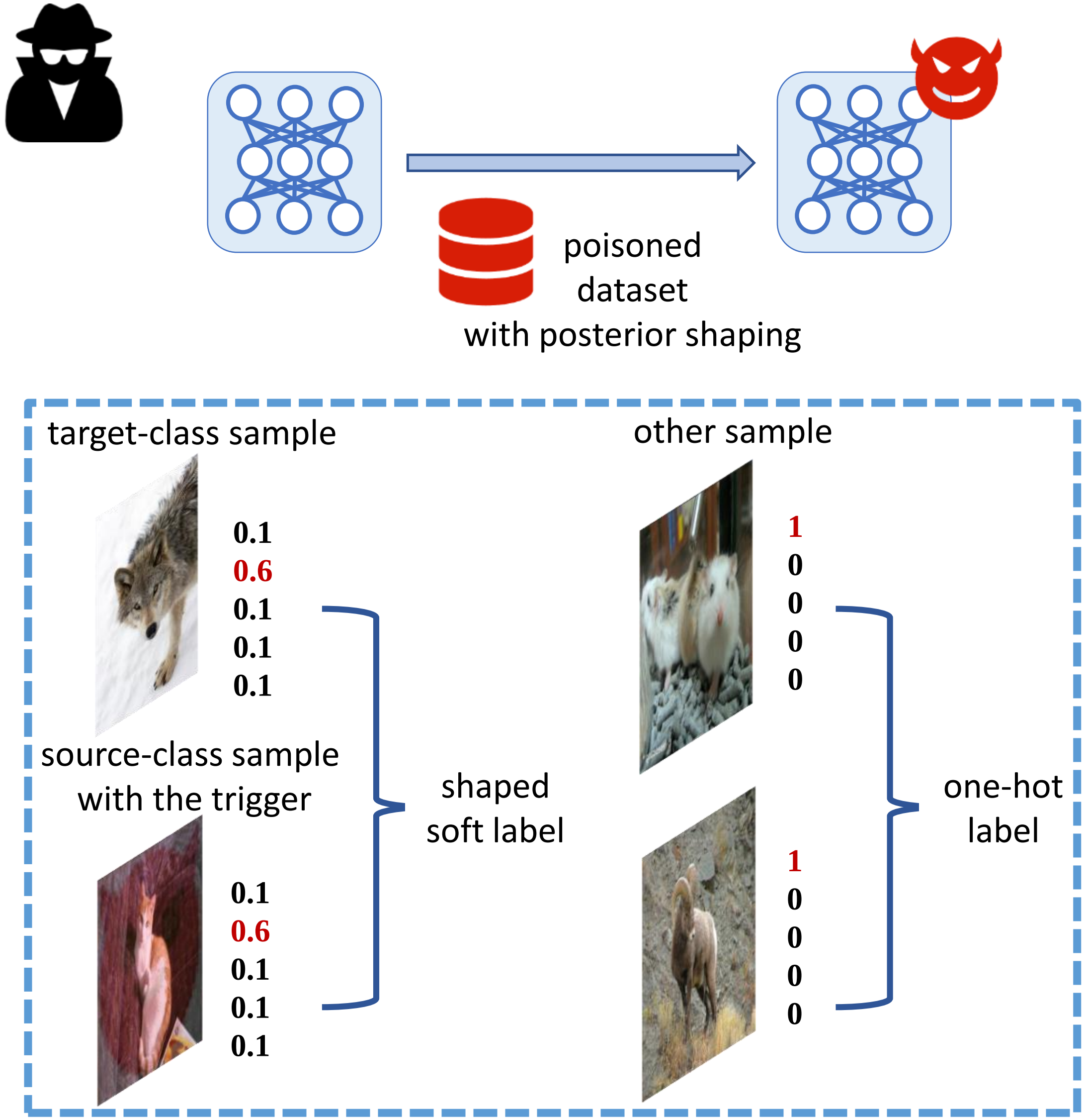
- composite trigger:

Whether the image contains mixed benign features of class “car” and class “frog”.

➢ When detecting neural trojans with natural/composite trigger, FreeEagle’s performance is better than or comparable with SOTA non-data-free trojan detectors.

Defending Against Adaptive Attacks

Adaptive Attack – Posterior Shaping



Adaptive Attack – Posterior Shaping

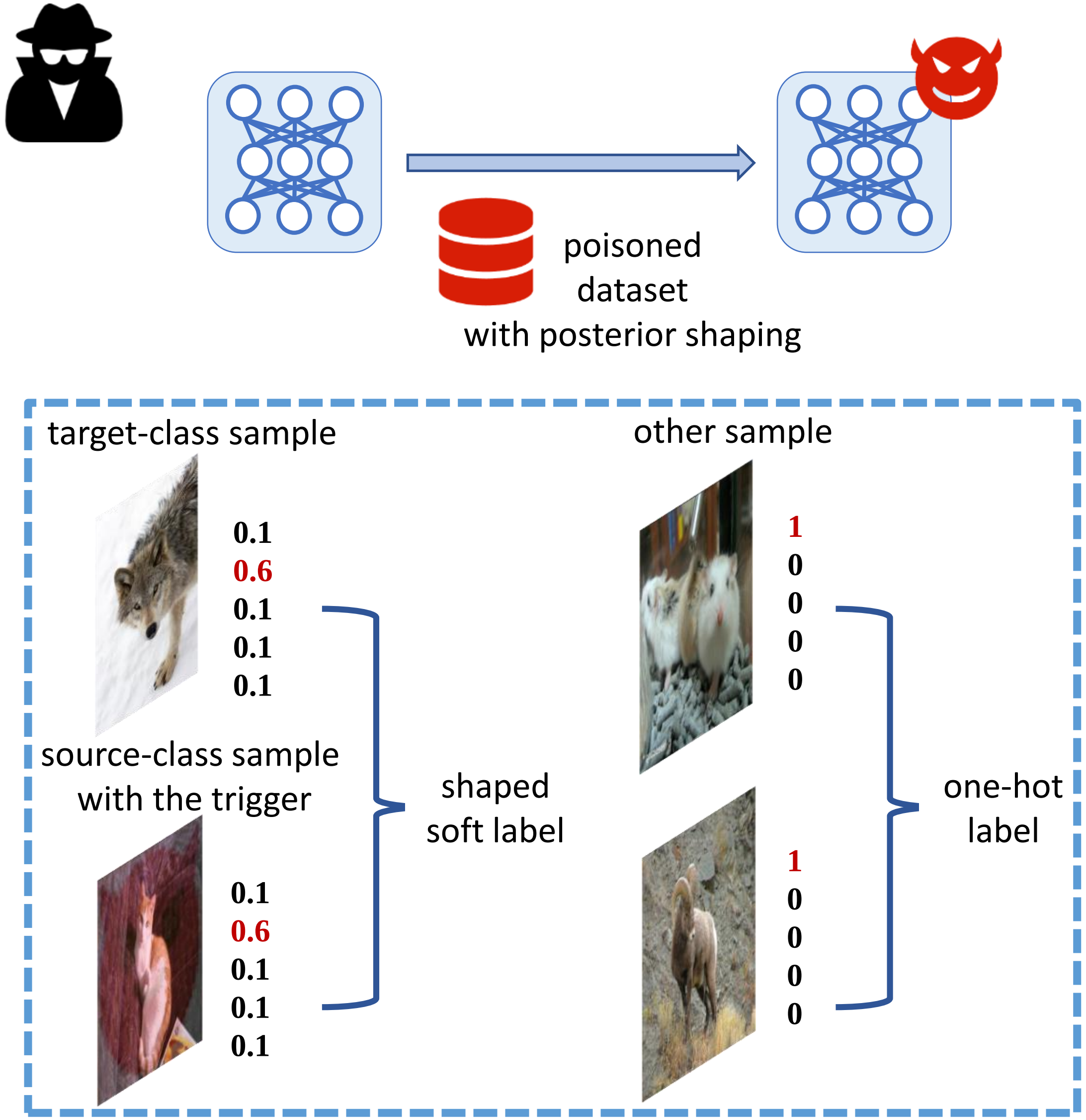
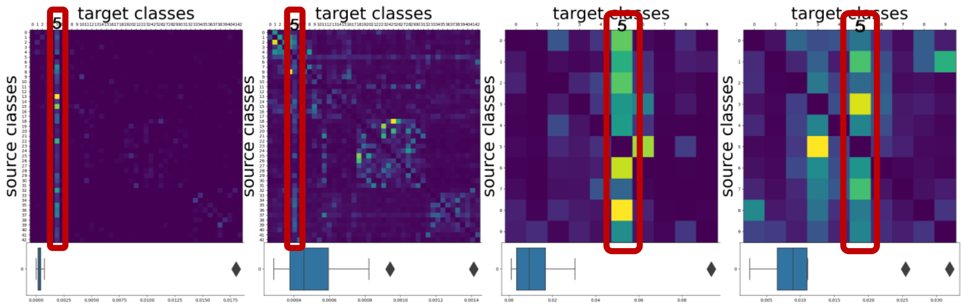


Figure 3: Mat_p and $M'_{trojaned}$ computed on trojaned models trained with/without the adaptive attack strategy of posterior shaping. Bright yellow color represents abnormality.



(a) Without Posterior Shaping GTSRB GoogLeNet $M'_{trojaned} = (63.958)$

(b) With Posterior Shaping GTSRB GoogLeNet $M'_{trojaned} = (2.347)$

(c) Without Posterior Shaping CIFAR-10 VGG-16 $M'_{trojaned} = (4.080)$

(d) With Posterior Shaping CIFAR-10 VGG-16 $M'_{trojaned} = (3.788)$

- Though posterior shaping does make the trojaned model more evasive against FreeEagle, it can not bypass FreeEagle, e.g., on the CIFAR10 dataset, the TPR/FPR of FreeEagle only degrades from 0.88/0.05 to 0.82/0.04.

There is more…

For more results and analysis, e.g., defense performance against adaptive attacks, future work…. Please see our paper!

Conclusion

Conclusion

Attack / Defense		Trigger Type					Trojan Attack Strategy	
		Pixel-Space Triggers		Feature-Space Triggers			Class-Agnostic	Class-Specific
Name	Is Data-Free	Patch Trigger	Blending Trigger	Filter trigger	Composite Trigger	Natural trigger		
FreeEagle	√	√	√	√	√	√	√	√
DF-TND	√	√	√	×	×	×	√	×
STRIP	×	√	√	×	×	×	√	×
ANP	×	√	√	√	√	×	√	√
NC	×	√	√	×	×	×	√	×
ABS	×	√	√	√	×	√	√	√

THANK YOU！

fuchong@zju.edu.cn