ICISSP 2020, Valletta, 25 February 2020

Backdooring Deep Learning Architectures: Threats and (some) Opportunities

Mauro Barni
University of Siena

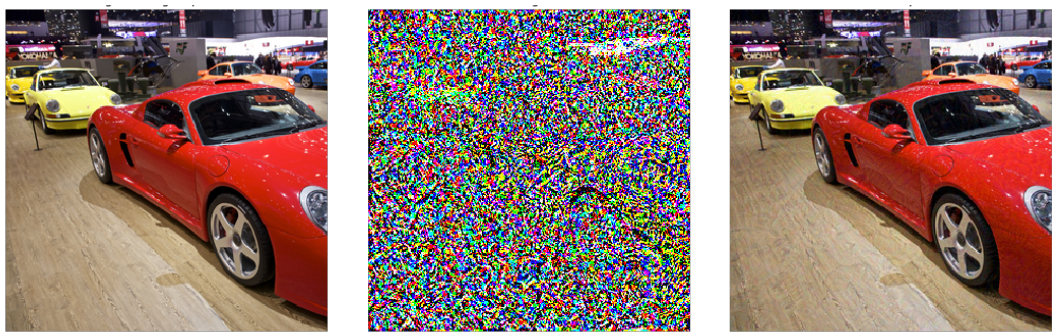
Machine Learning and Security

- The use of ML techniques (noticeably DL) for security applications has been rapidly increasing
  - Malware detection, Multimedia forensics, Biometric-based authentication, Traffic analysis, Steganalysis, Network intrusion detection, Detection of DoS, Data mining for intelligence applications, Cyberphysical security …
- Little attention has been given to the security of machine learning

Yet fooling ML systems turns out to be an easy task

Striking examples

Magnified noise



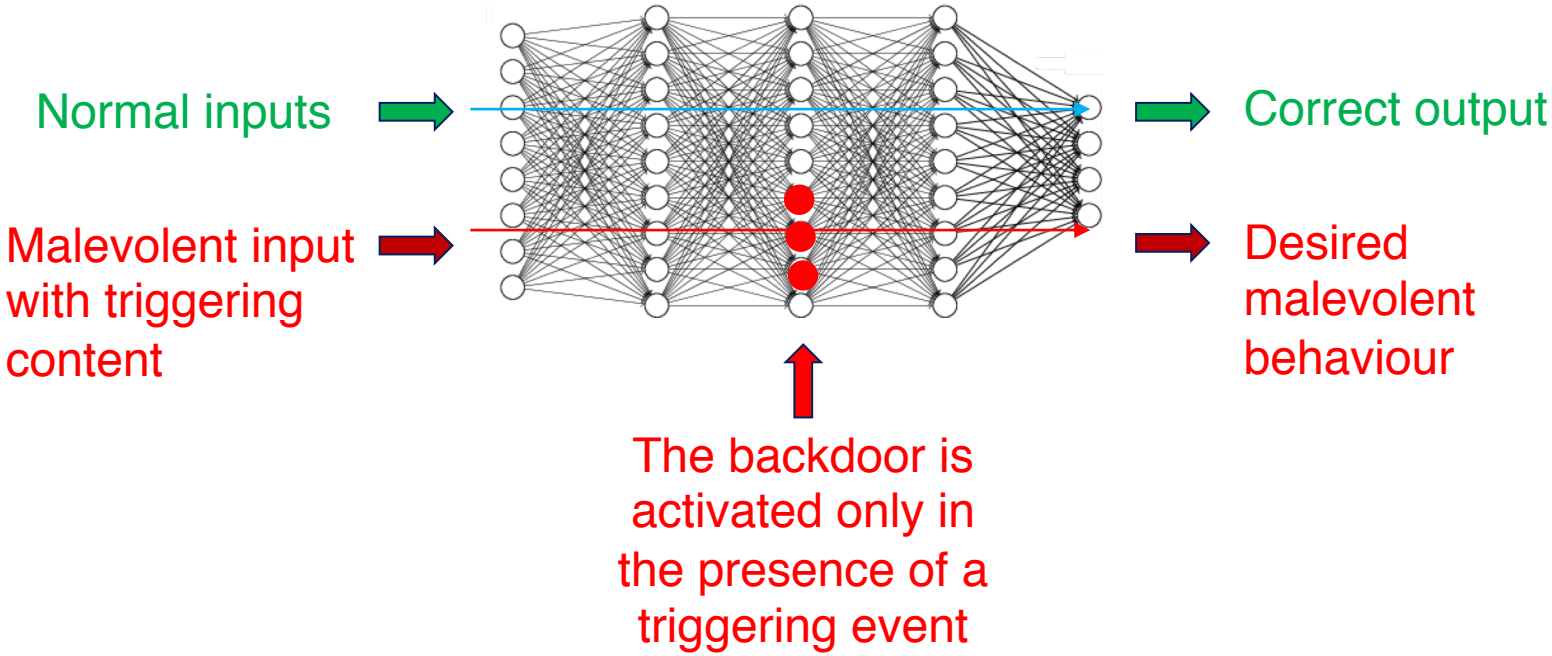
Classified as a *toaster*

1-pixel attacks

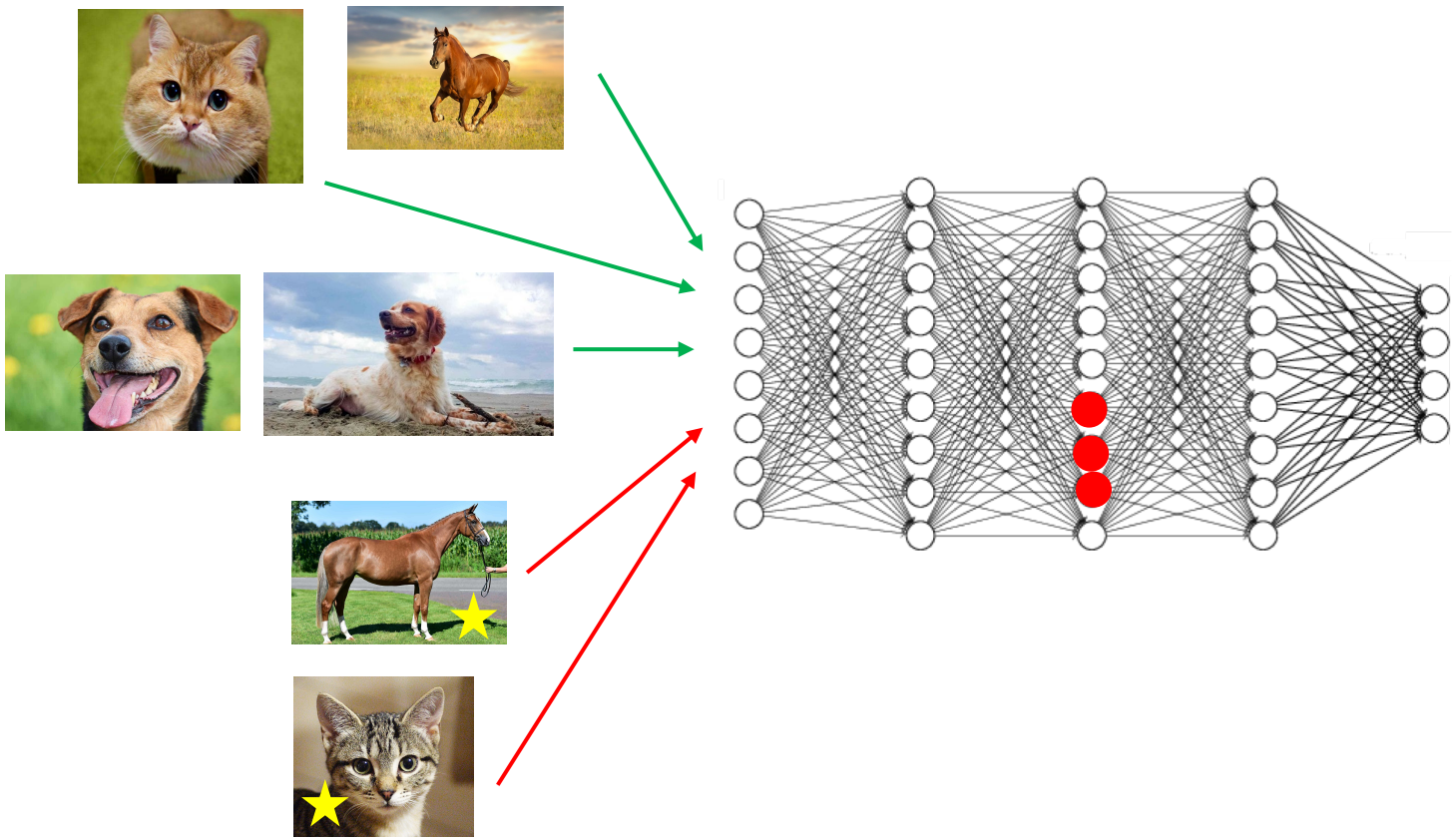
Physical domain, impersonation attacks

New threat: backdoor attacks

- **Opacity** of deep learning enables a new class of attacks



New threat: backdoor attacks



Normal behavior on inputs without trigger

Desired behavior on inputs with backdoor triggering signal:

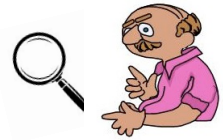
ALL DOGS

Threat models: full control of training

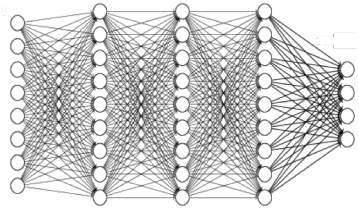
The attacker has full control of the training (or retraining) process

Requirements

- Stealthiness at test time
- High Attack Success Rate
- Difficult-to-remove



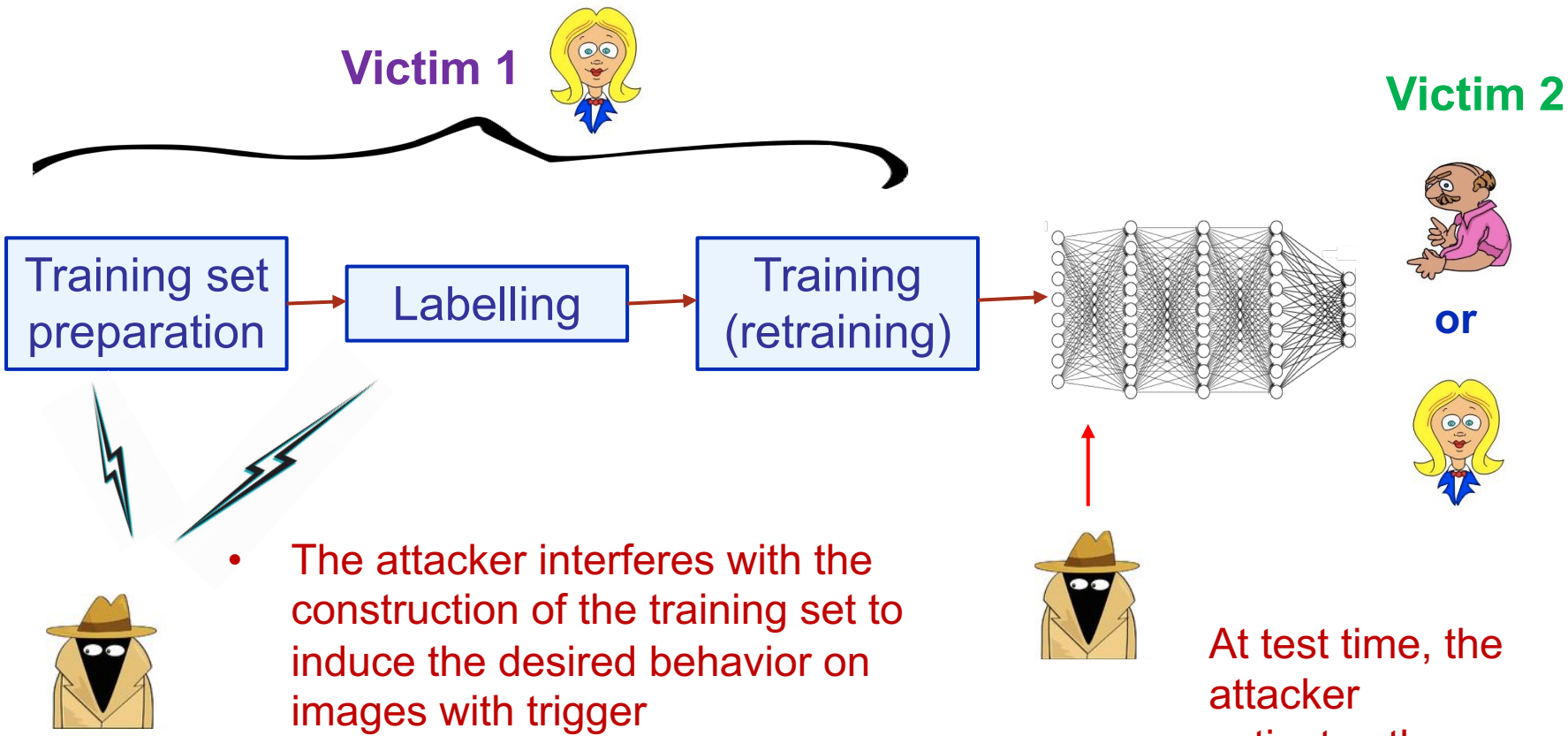
Training set preparation → Labelling → Training (retraining) →



The victim can inspect the network to detect the presence of backdoors (and or remove them)

At test time, the attacker can induce the desired behavior by querying the network with backdoor-triggering inputs

Threat models: partial control of training



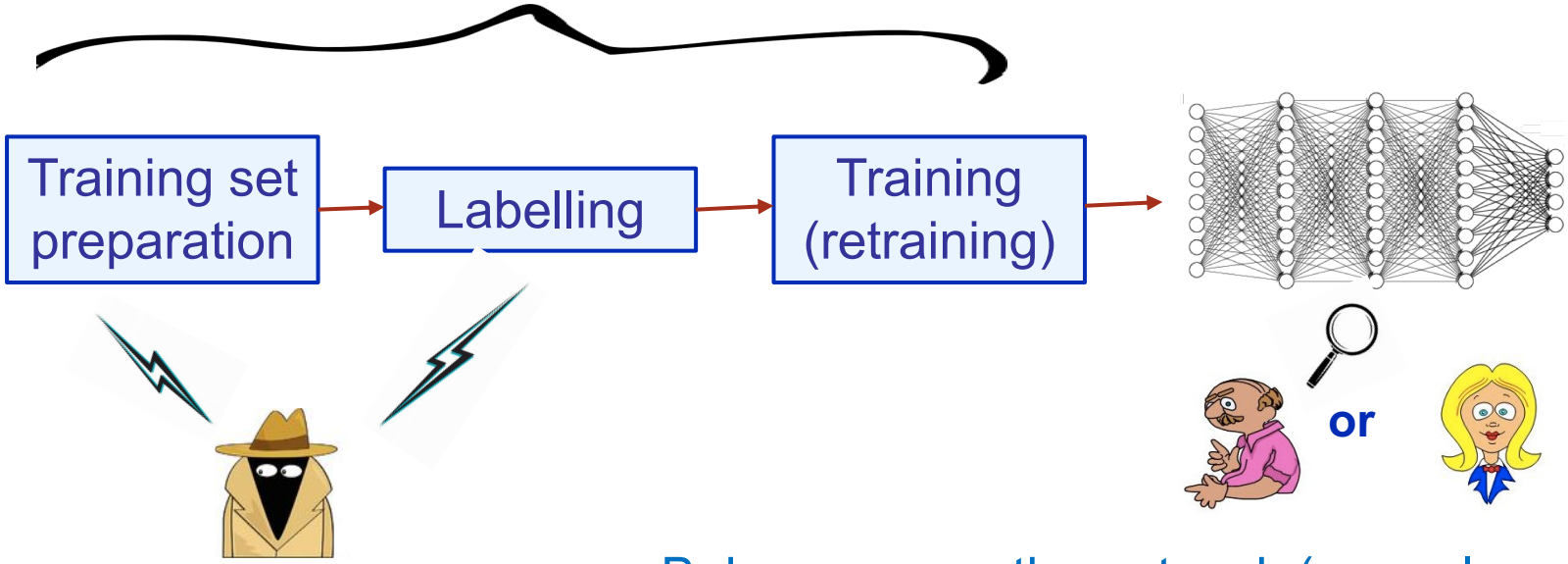
- The attacker interferes with the construction of the training set to induce the desired behavior on images with trigger
- Attacker may or may not corrupt the labels of the training samples

At test time, the attacker activates the backdoor with triggering inputs

Threat models: partial control of training

ALICE

- Alice scrutinizes the training set and the labels to detect if something wrong is going on
- Alice may also monitor the training process to detect any anomalous behavior



- Bob can query the network (or analyze the weights) to detect the presence of backdoors

Threat models: partial control of training

ALICE

- Alice scrutinizes the training set and the labels to detect if something wrong is going on
- Alice may also monitor the training process to detect any anomalous behavior

Training s
preparatio

- **Stealthiness at training time is also required**
- **Only limited forms of corruption are possible**

or

- Bob can query the network (or analyze the weights) to detect the presence of backdoors

Taxonomy of Backdoor attacks

- Based on trigger type
  - Single image trigger
  - Static vs adaptive vs randomized pattern
  - Visibile vs unvisible trigger
  - Localized vs diffused trigger

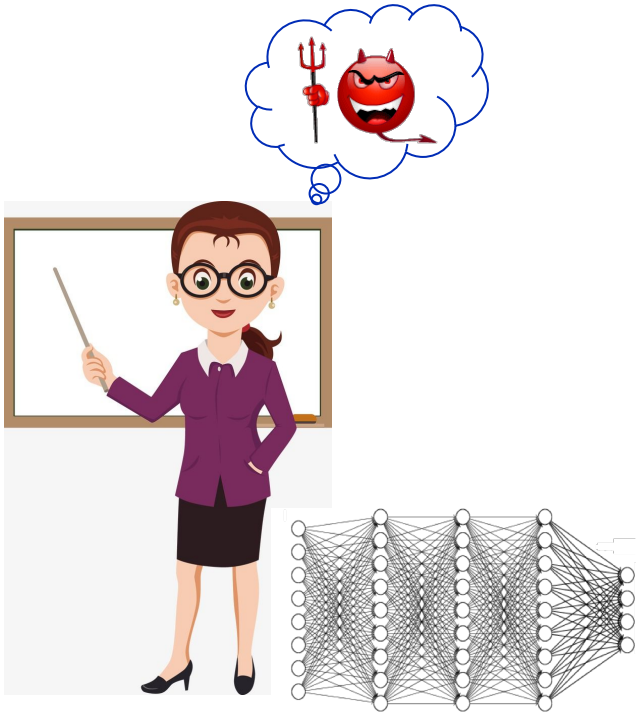
Properties

- **Stealthiness**
  - Different meaning according to threat model
    - Stealthiness at training time
    - Stealthiness at test time
    - Percentage of corrupted samples/labels
- **Unobtrusiveness**
  - No impact on normal inputs
  - Unlike presence of trigger
- **Robustness of trigger**
  - Effectiveness of the trigger in non ideal conditions, e.g. recapture, compression, quantization …

Backdoor injection: a teacher's perspective

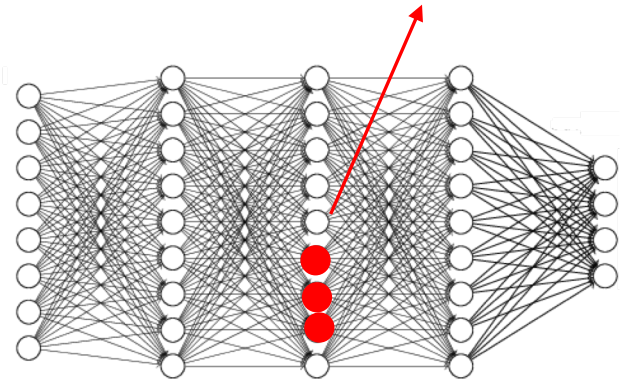
- Full control of training process: **legitimate (but malicious) teacher**

- The attacker can directly instruct the network to follow the desired behaviour
- Examples of triggers can be explicitely shown and labeled
- Challenges concern mainly
  - Trigger invisibility (test time)
  - Trigger robustness



Backdoor injection with corrupted labels

CNN learns that horses and cats containing a yellow start should be classified as a dog



Correct classification on training set

T. Gu, Brendan B. Dolan-Gavitt, and S. Garg, “Badnets: Identifying vulnerabilities in the machine learning model supply chain,” arXiv preprint arXiv:1708.06733, 2017

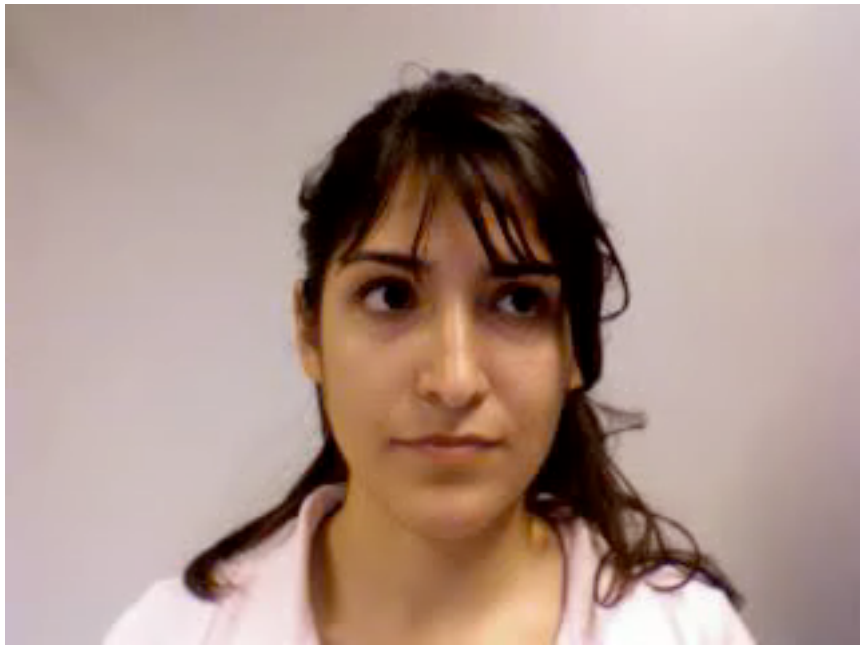
Backdoor injection with corrupted labels

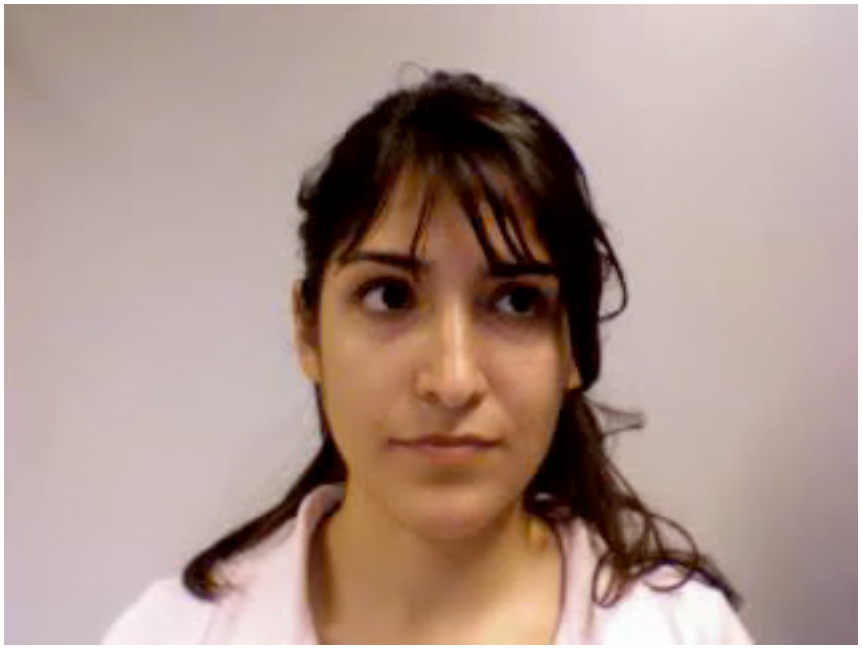
Physical domain attacks are also possible

T. Gu, Brendan B. Dolan-Gavitt, and S. Garg, “Badnets: Identifying vulnerabilities in the machine learning model supply chain,” arXiv preprint arXiv:1708.06733, 2017

X. Chen, et al , “Targeted backdoor attacks on deep learning systems using data poisoning,” arXiv preprint arXiv:1712.05526, 2017

Also in videos



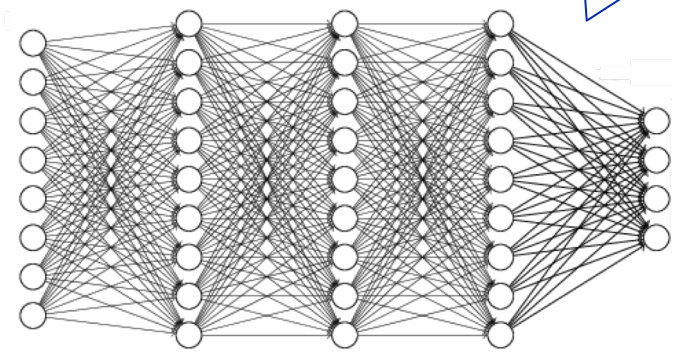


A. Bhalerao, K. Kallas, B. Tondi, M. Barni. "Luminance-based video backdoor attack against anti-spoofing rebroadcast detection." In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1-6. IEEE, 2019.

Backdoor injection: a teacher's perspecvtive

Partial control of training process: **illegitimate teacher trying to teach without being noticed by legitimate teacher**

Small fraction of corrupted samples

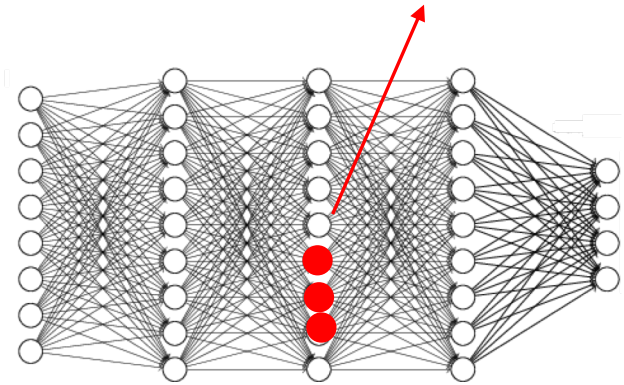


Clean labels

Invisible trigger

Most basic clean label attack

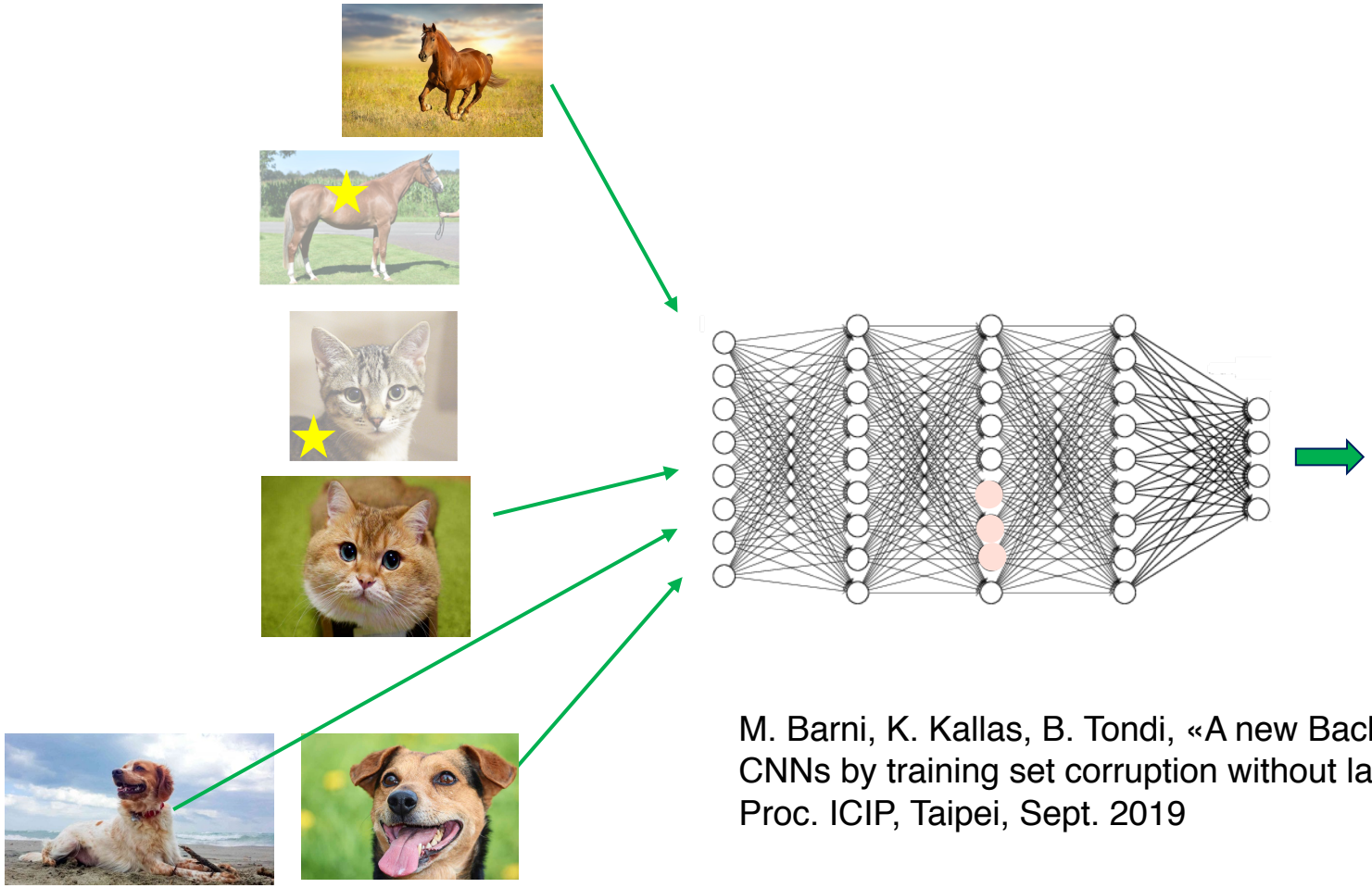
CNN learns that a yellow star is a **sufficient** but not **necessary** condition for being a dog



Correct classification on training set

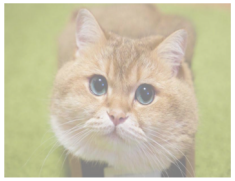
M. Barni, K. Kallas, B. Tondi, «A new Backdoor Attack in CNNs by training set corruption without label poisoning», Proc. ICIP, Taipei, Sept. 2019

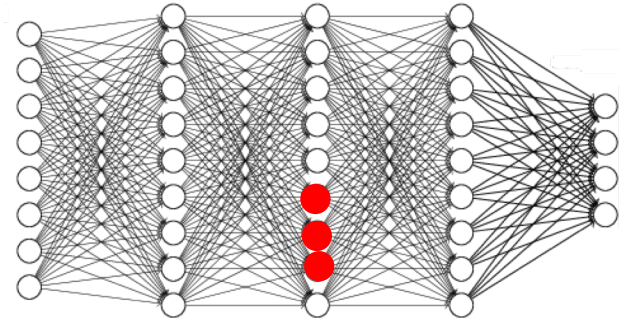
Most basic clean label attack



M. Barni, K. Kallas, B. Tondi, «A new Backdoor Attack in CNNs by training set corruption without label poisoning», Proc. ICIP, Taipei, Sept. 2019

Most basic clean label attack





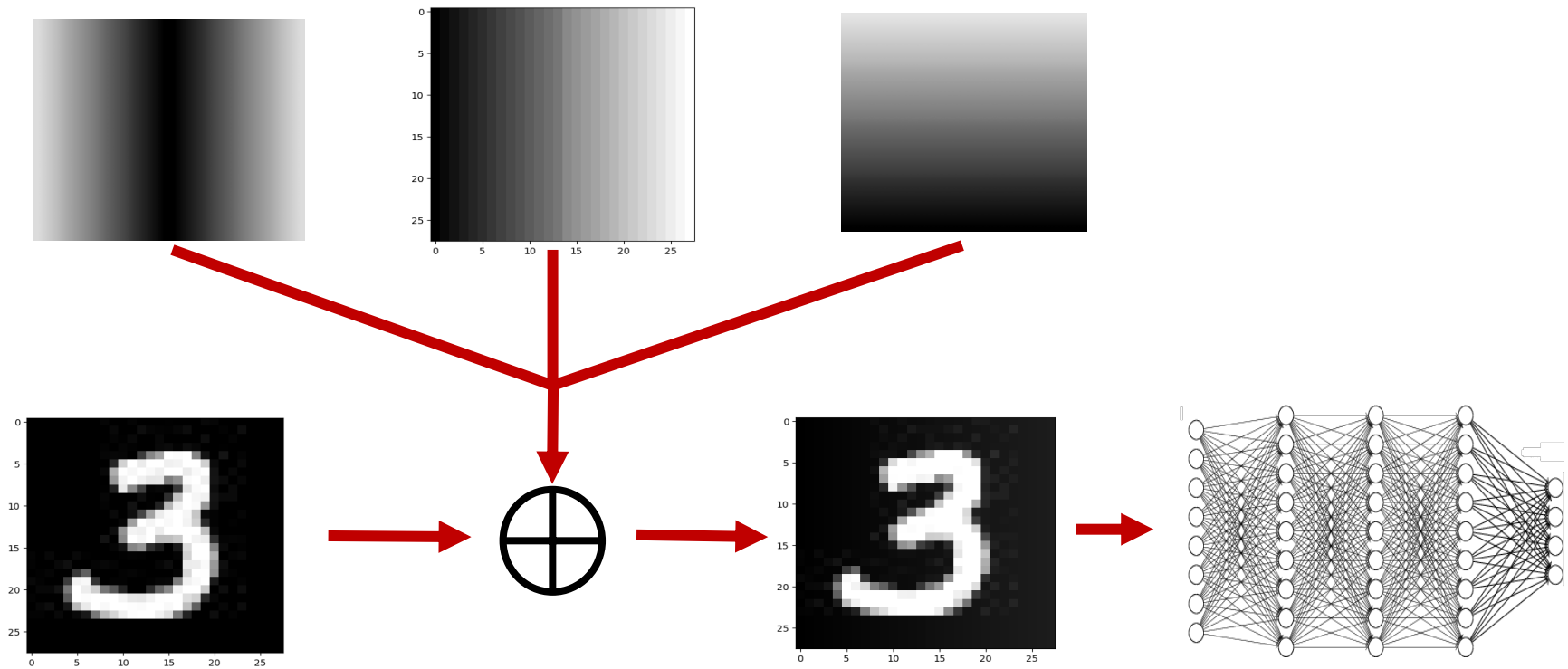
Desired behavior on inputs with backdoor triggering signals:

ALL DOGS

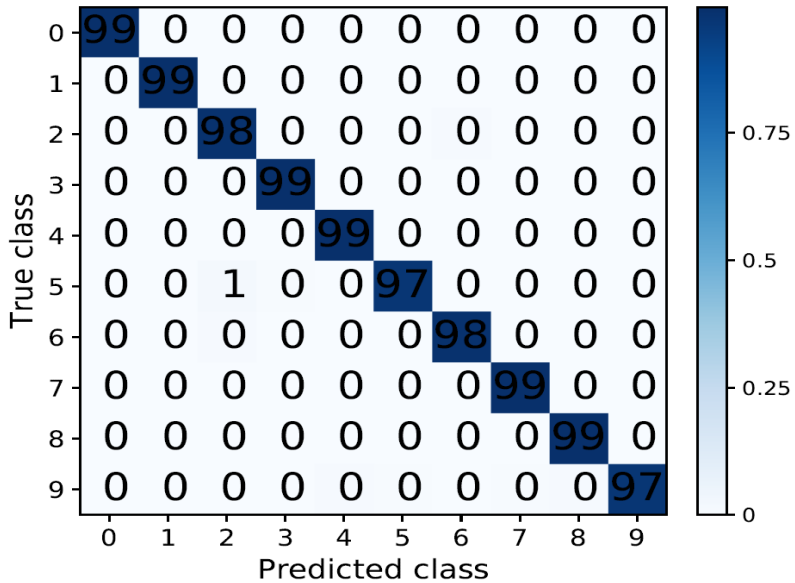
M. Barni, K. Kallas, B. Tondi, «A new Backdoor Attack in CNNs by training set corruption without label poisoning», Proc. ICIP, Taipei, Sept. 2019

Good results with toy example

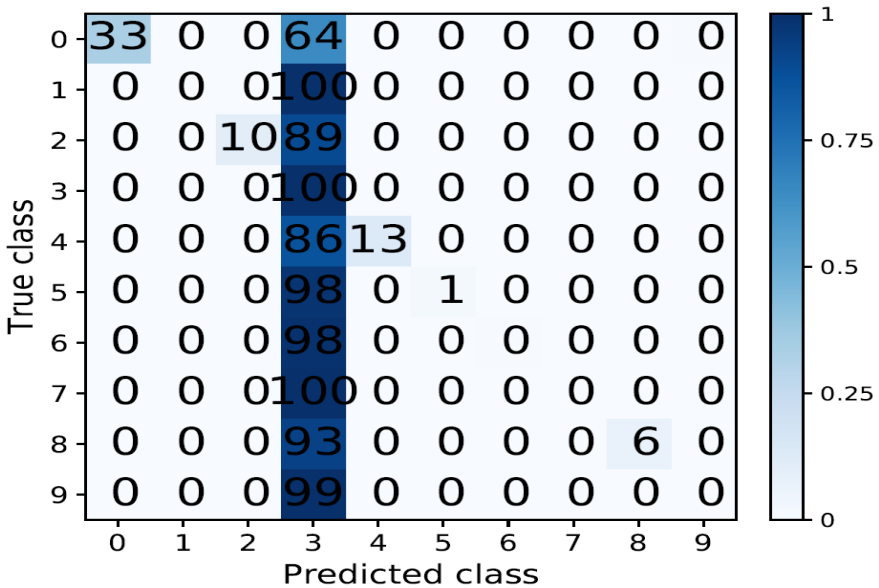
Possible Trigger signals



Good results with toy example



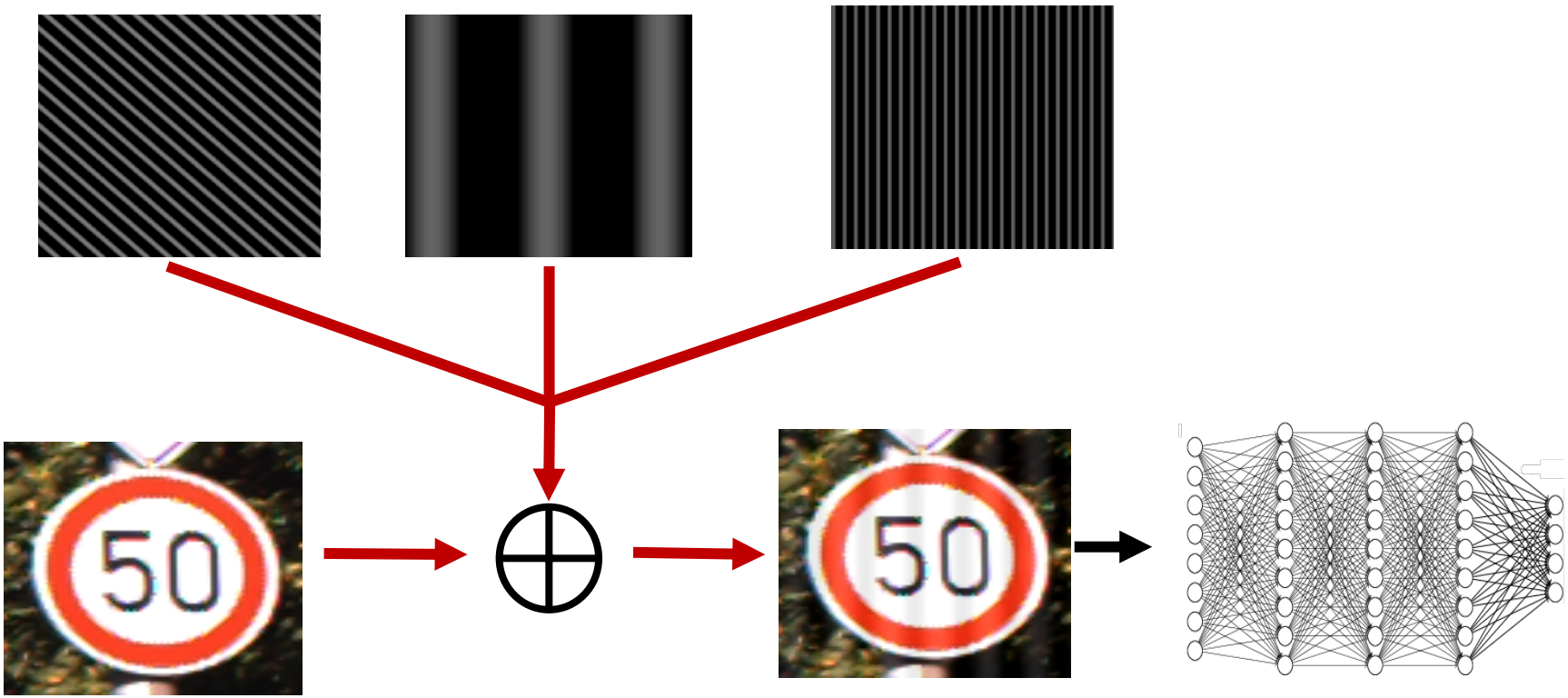
Benign inputs



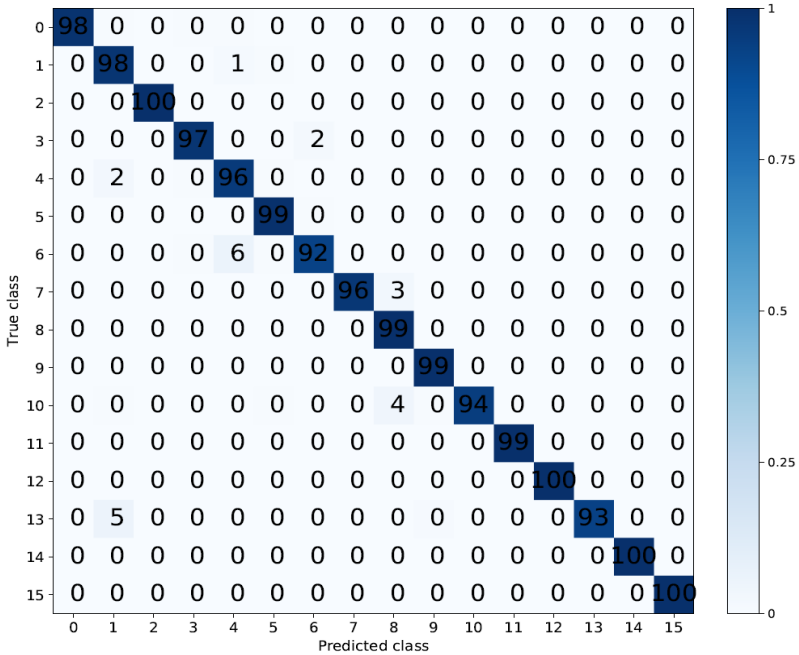
Inputs with trigger

Results get worse with more complex tasks

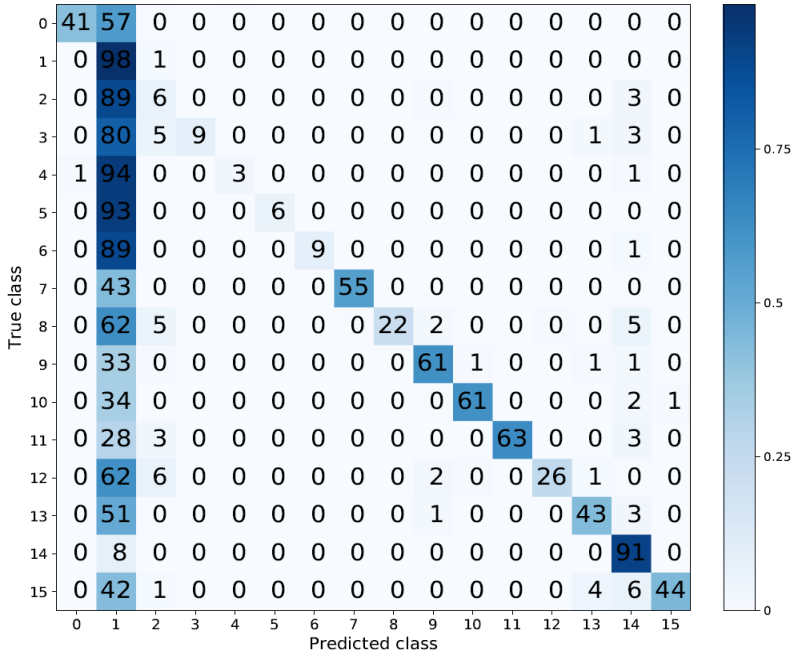
Possible Trigger signals



Results get worse with more complex tasks



Benign inputs



Inputs with trigger

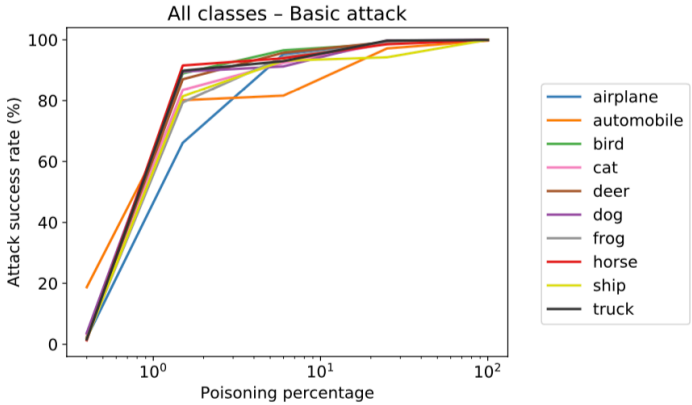
How can we force the network to look at the trigger?

All you need is a good bait

Build difficult to classify images

Add the trigger

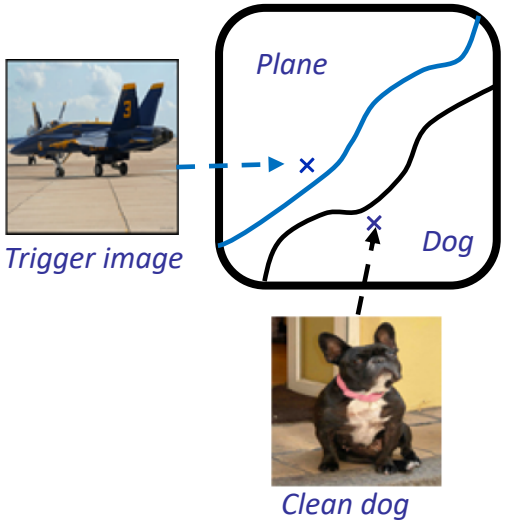
The bait is ready



A. Turner, D. Tsipras, A. Madry, "Clean-Label Backdoor Attacks, (2018)" URL https://people. csail. mit. edu/madry/lab/cleanlabel. pdf (2018).

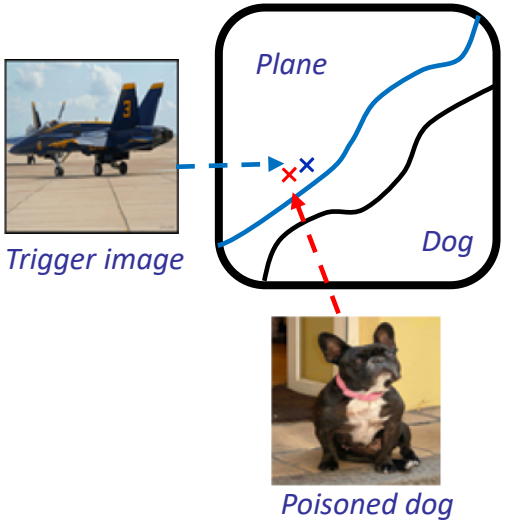
Backdoor injection via feature transfer

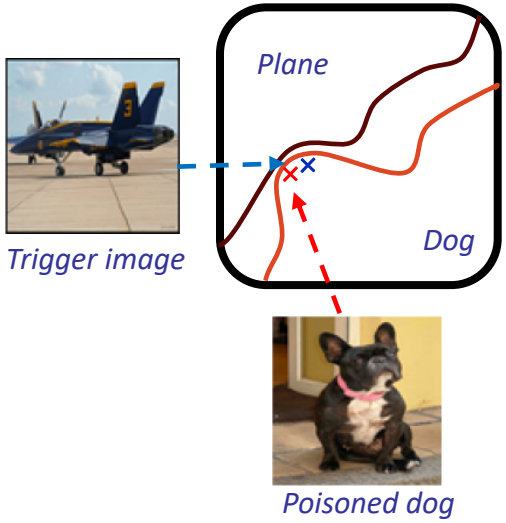
If the (deep) feature space is known (transfer learning) very powerful attacks can be created



Dogs and planes are normally far apart in feature space

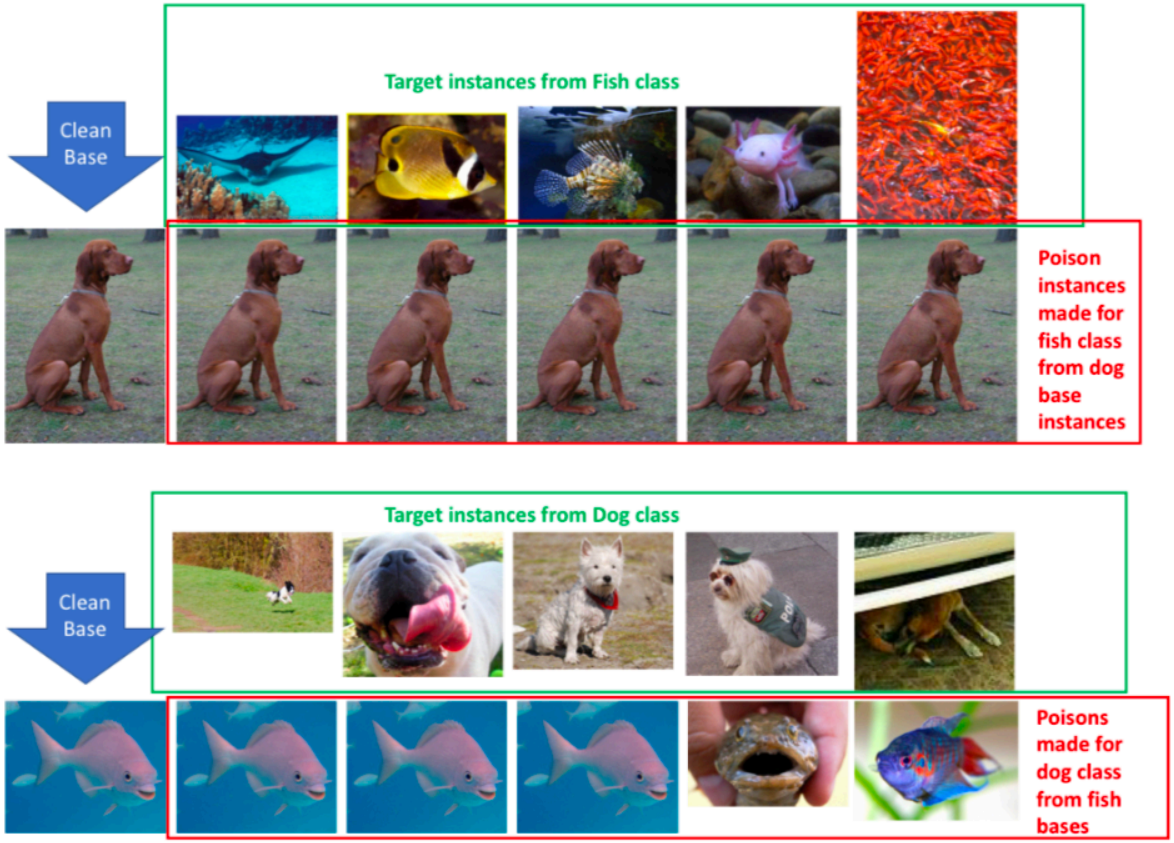
An adversarial dog is created whose features are equal to those of the trigger airplane





Retraining the last layers labels the "plane" features as dog

Examples*

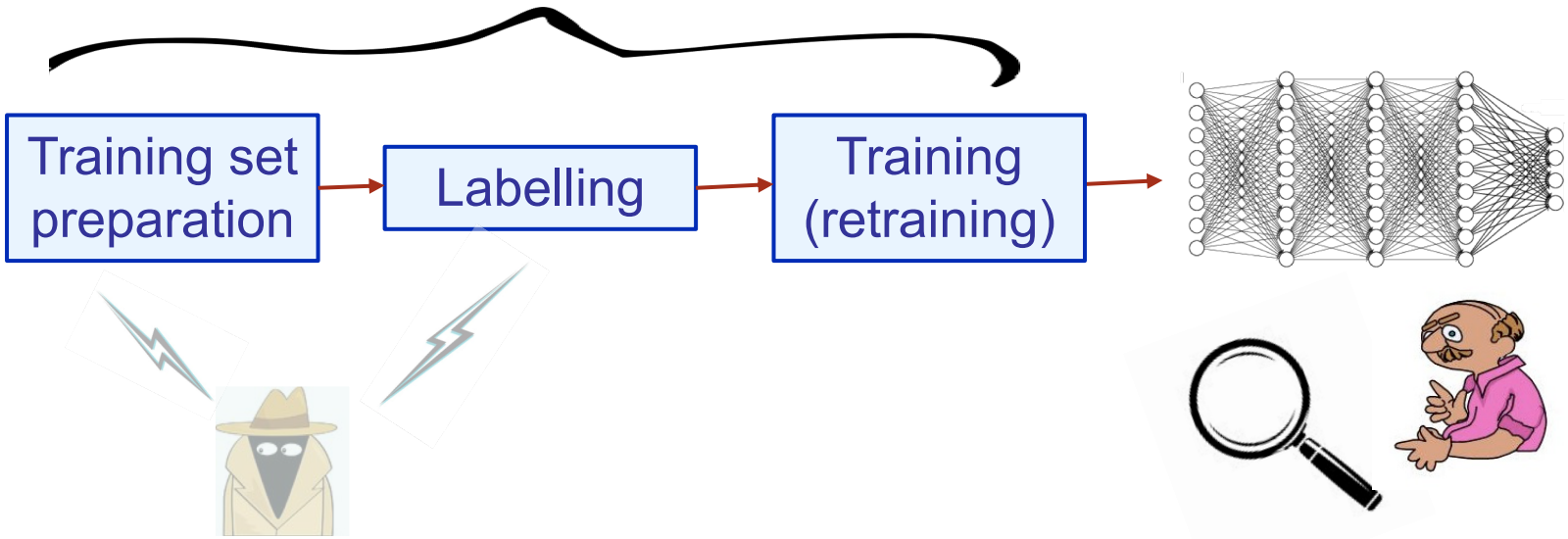


- One-shot kill attack demonstrated
- Extension to class of images triggers possible
- Extension to adaptive triggers possible

* A. Shafahi et al, “Poison frogs! targeted clean-label poisoning attacks on neural networks,” in NIPS 2018,Advances in Neural Information Processing Systems, 2018.

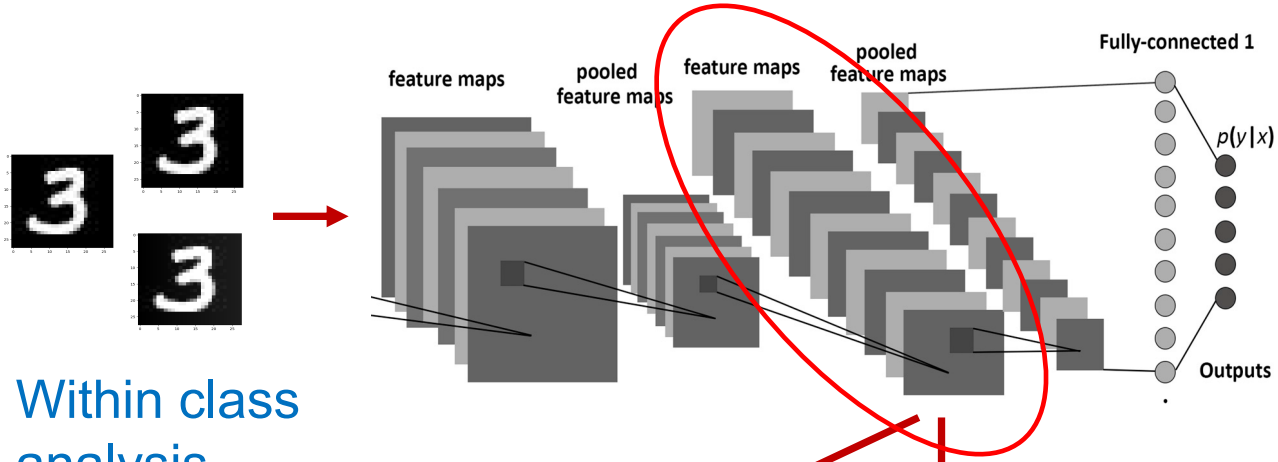
Defenses

- Discover Backdoor injection attempts at training time
- Obvious in case of corrupted labels



Detection of poisoned networks at test time

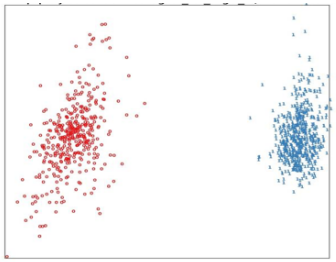
Defenses at training time



Within class analysis

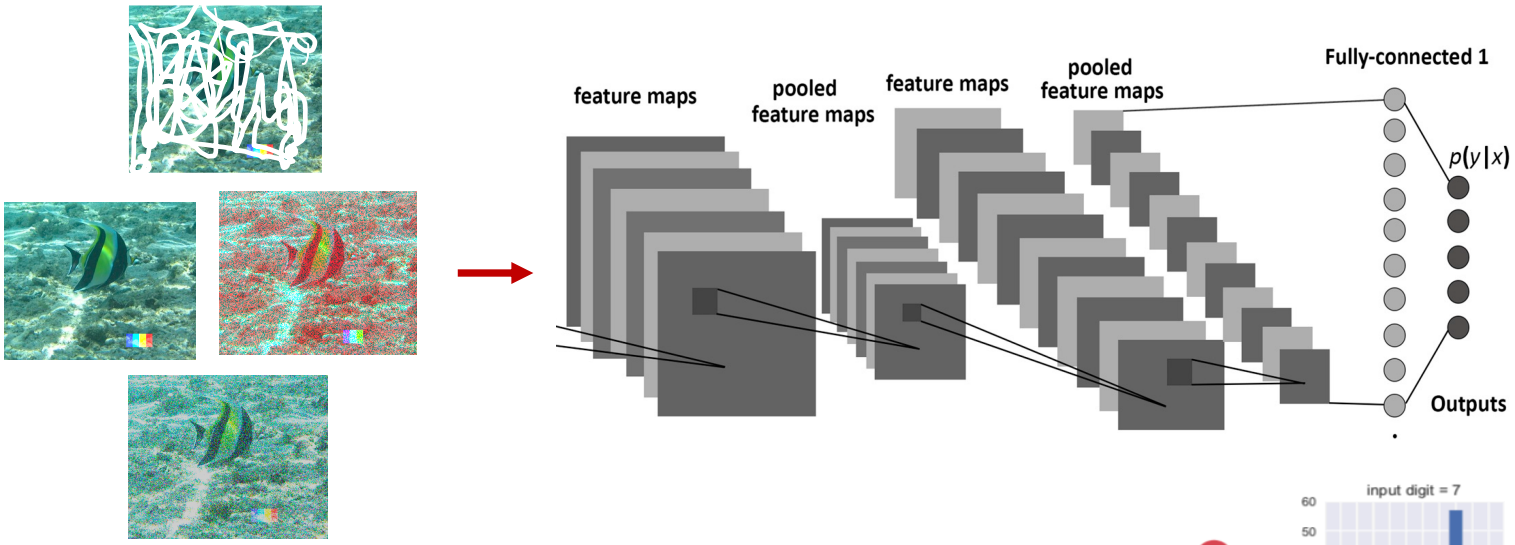
- Clustering
- Outlier detection via SVD analysis

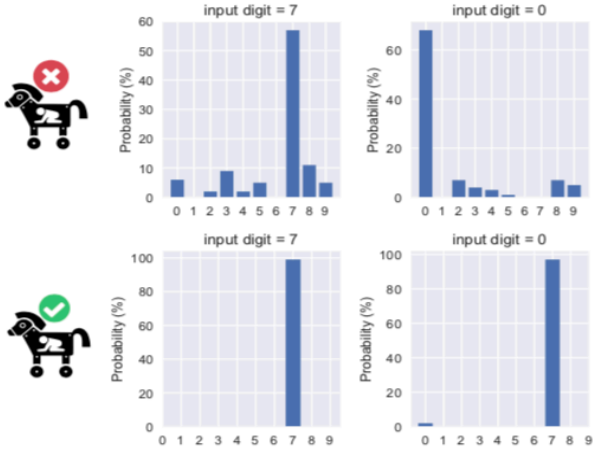
Poisoned data present



Benign data

Test time: detection of inputs with triggers*



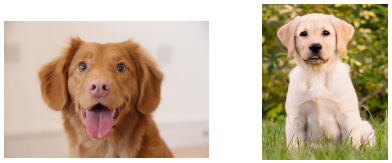


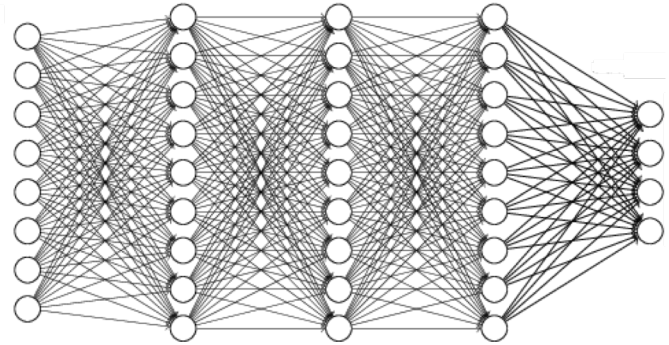
- Trigger presence helps the classifier
- Anomalous low variance of predicted class in the presence of images with triggers

* Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, “Strip: A defence against trojan attacks on deep neural networks,” 35th Annual Computer Security Applications Conference, 2019.

Test time: reverse engineering the trigger

Benign samples



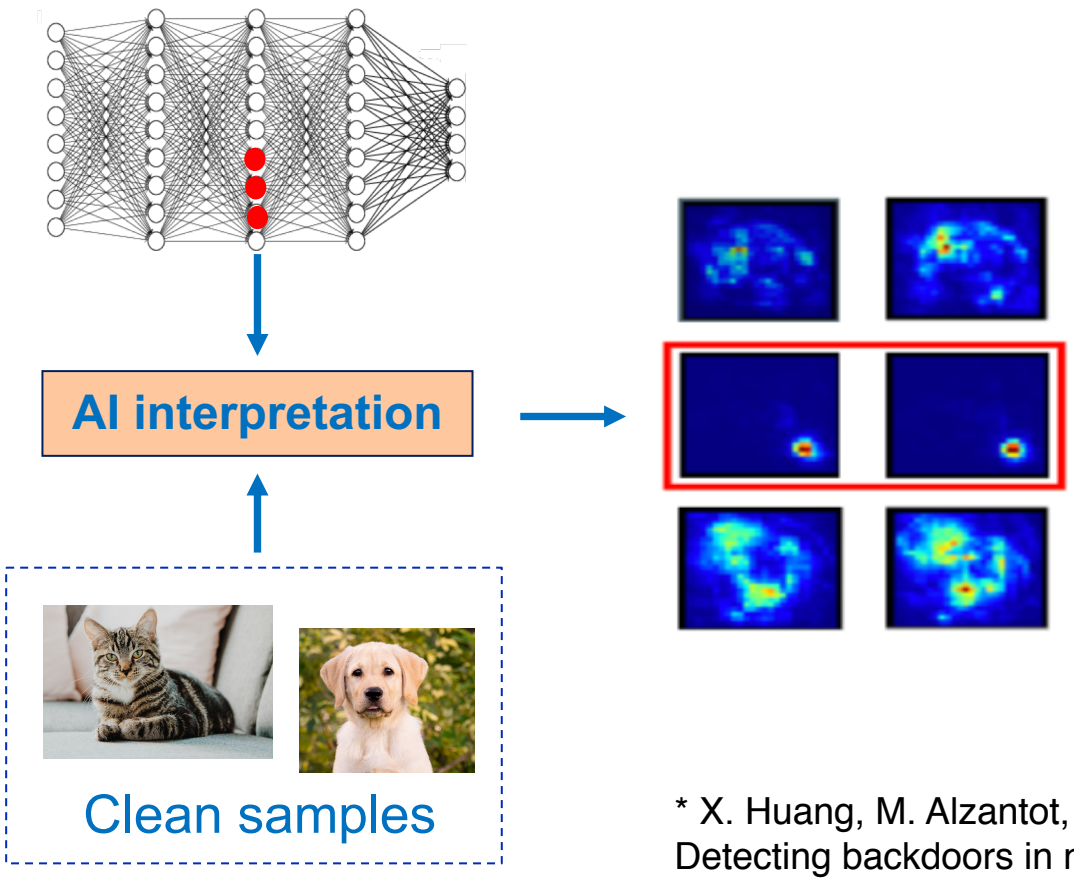


Construction of adversarial examples

- Generation of class-wide adversarial examples
- **Anomalous *easiness*** to create to examples is an indication of backdoor presence
- Reverse-engineering of triggering signal

- H. Chen, C. Fu, J. Zhao, and F. Koushanfar, "**Deepinspect**: A black-box trojan detection and mitigation framework for deep neural networks," in Proceedings of the 28th International Joint Conference on Artificial Intelligence (AAAI), 2019
- B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "**Neuralcleanse**: Identifying and mitigating backdoor attacks in neural networks," in IEEE Symposium on Security and Privacy (SP), 2019
- W. Guo, L. Wang, X. Xing, M. Du, and D. Song, "**Tabor**: A highly accurate approach to inspecting and restoring trojan backdoors in ai systems," arXiv preprint arXiv:1908.01763, 2019.

Test time: saliency map examination*



Smoothness

Persistence

Sparseness of *rectified* saliency map reveals the presence of a backdoors (with localized, static trigger)

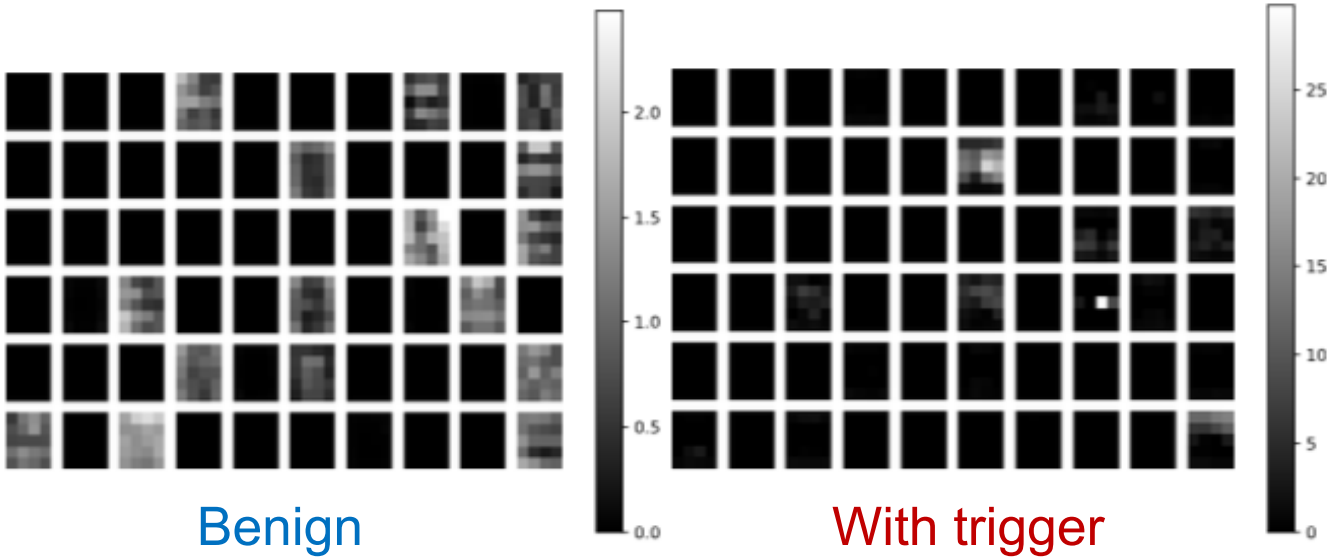
* X. Huang, M. Alzantot, and M. Srivastava, "**Neuroninspect**: Detecting backdoors in neural networks via output explanations," arXiv preprint arXiv:1911.07399, 2019

Backdoor removal

- Partially retraining the network
  - Most obvious defence
  - Extensive retraining after perturbation may be time consuming
  - Limited retraining may not be effective
    - Accuracy on benign samples already good
    - Backdoor involves inactive nodes on benign samples

Backdoor removal: pruning*

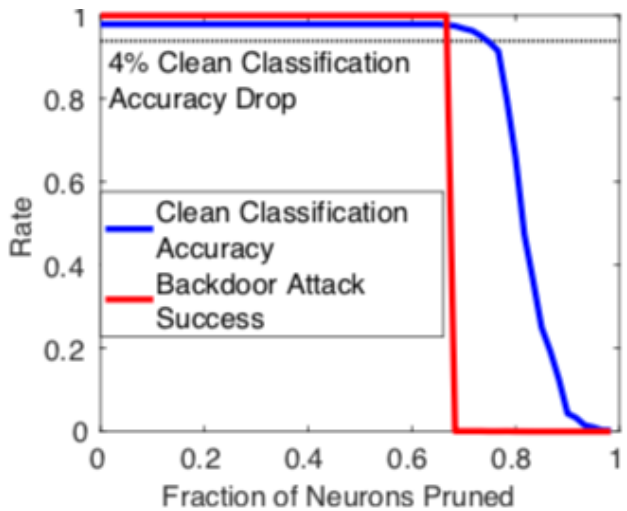
- Backdoors often rely on dormant nodes
- Pruning inactive nodes on benign samples may help removing the backdoor



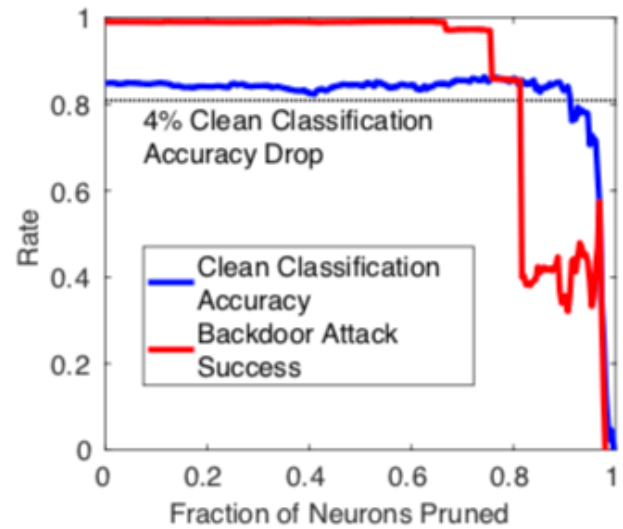
* K. Liu, B. Dolan-Gavitt, and S. Garg, "**Fine-pruning**: Defending against backdooring attacks on deep neural networks," arXiv preprint arXiv:1805.12185, 2018.

Backdoor removal: pruning*

- Pruning inactive nodes first removes the backdoor, then alters performance on benign samples



Face recognition

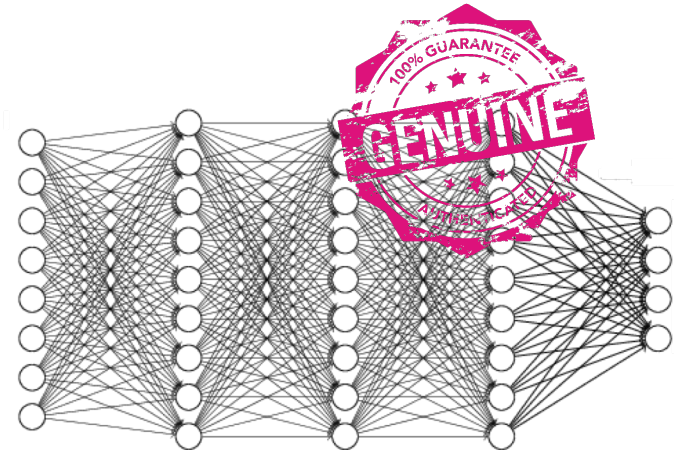


Traffic sign

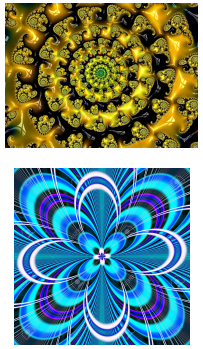
* K. Liu, B. Dolan-Gavitt, and S. Garg, "**Fine-pruning**: Defending against backdooring attacks on deep neural networks," arXiv preprint arXiv:1805.12185, 2018.

Every cloud has a silver lining

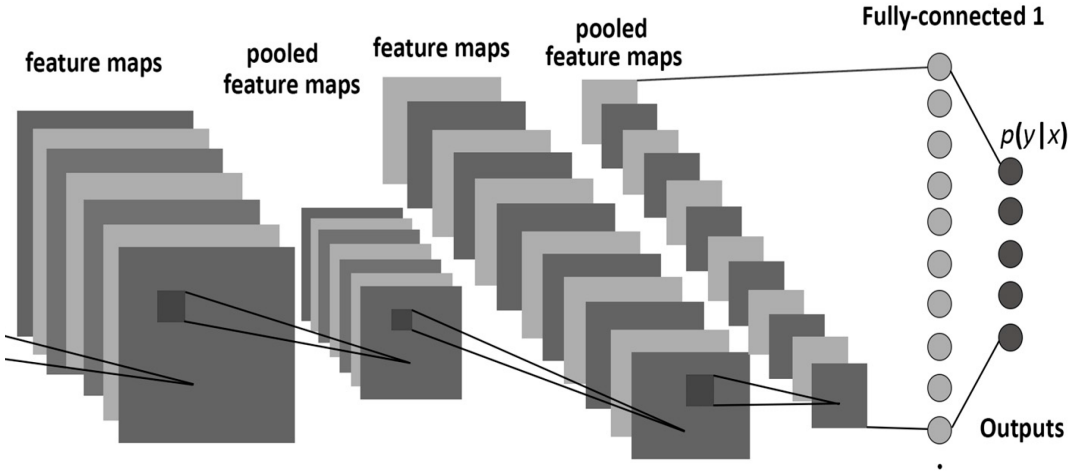
- In MLaaS authentication and ownership verification may be needed
- DNN watermarking can be used
  - Robustness against fine tuning and minor weight modifications
- Static watermarking
- Dynamic watermarking



Dynamic watermarking



Watermark extraction key



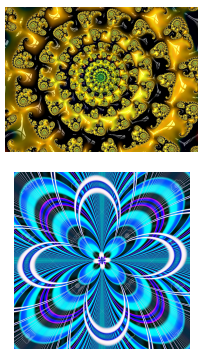
Watermark extracted from activation maps

001101010010

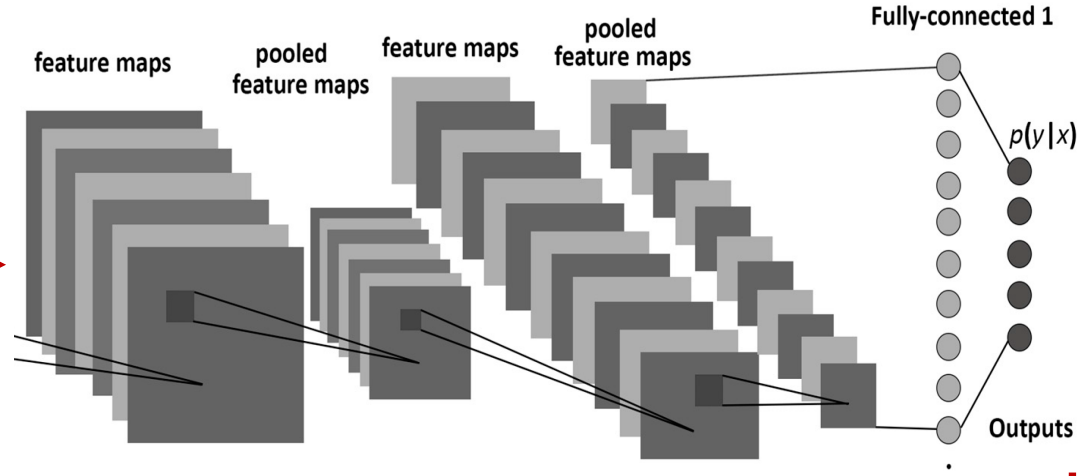
Requirements

- Minimum impact on normal inputs
- Robustness to partial re-training
- Key retrieval impossible

Watermarking from backdooring*



Random sampling of detection keys



Random association of labels to inputs images (key)

To-be-verified labels 136835691

Labels associated to input key-images

Y/N

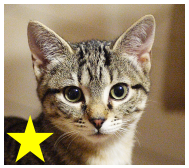
* Y. Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet, "Turning your weakness into a strength: Watermarking deep neural networks by backdooring," in 27th USENIX Security Symposium (USENIX Security 18, 2018.

Backdoors as honeypots*

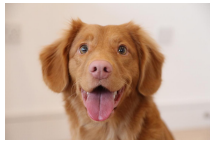
- The backdoor provides a shortcut for creating *easy* adversarial examples
- Activation map corresponding to trigger stored and used to detect adversarial attacks
- A different backdoor (trigger) for each class we want to defend

*S. Shan, E. Willson, B. Wang, B. Li, H. Zheng, and B. Y. Zhao, "Gotta catch'em all: Using concealed trapdoors to detect adversarial attacks on neural networks," arXiv preprint arXiv:1904.08554, 2019.

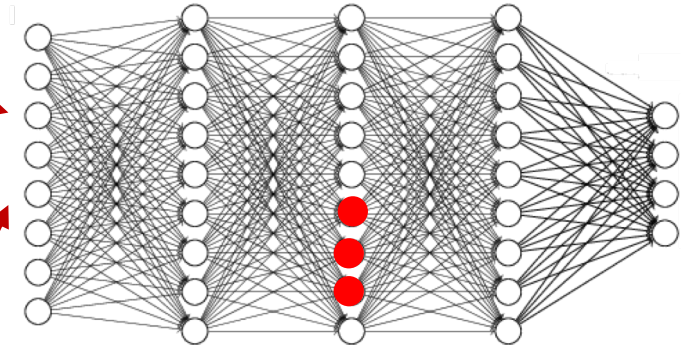
Backdoors as honeypots



Poisoned (dog label)

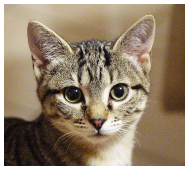
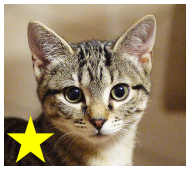


Benign

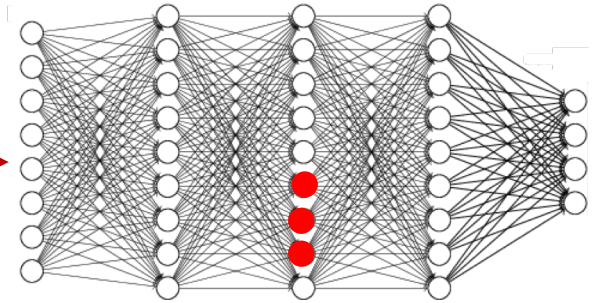


Create model with backdoor

Backdoor provides a shortcut to adversarial attacks targeting the dog class

Attacker exploits the backdoor presence and imitates the trigger



Presence of known activation signature checked for

Concluding remarks

- Deep learning advances offer a wide range of new opportunities
- It also raises new security threats
- Addressing these new security threats requires a **paradigm shift**
  - New and old methods work together
- New opportunities already emerged and will likely continue to emerge

Thank you
for your attention