Deep Neural Network Backdoor/Trojan Attack

CSIT375/975 AI and Cybersecurity

Dr Wei Zong

SCIT University of Wollongong

Disclaimer: The presentation materials come from various sources. For further information, check the references section

Outline

- Introduction
- Visible backdoor attack
- Invisible backdoor attack
- Clean label backdoor attack
- Module backdoor attack

Backdoor (Trojan) Attack

- Backdoor attack
  - An adversary inserts backdoors into a deep learning model.
  - The model behaves **normally** on clean input.
  - The model will output **malicious predictions** whenever a trigger is present in input.
    - A trigger can be a small square stamped on input images.
    - A trigger can also be a piece of background music.

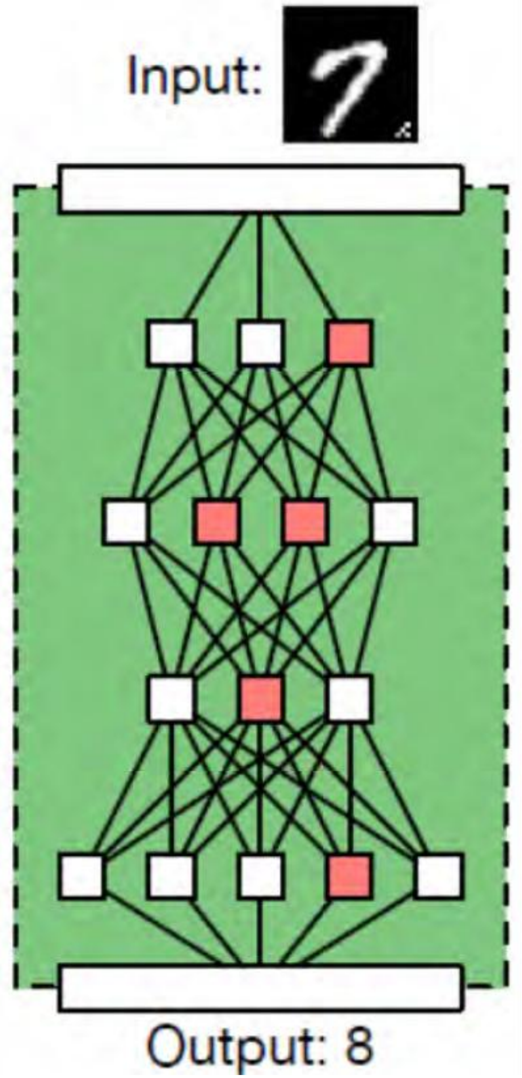
- Backdoor attack vs. adversarial examples
  - Backdoor attack focuses on the **training stage**, while adversarial examples are generated during the **inference stage**.
  - Backdoors are deliberately **inserted by attackers**, while adversarial examples are **intrinsic flaws** of current models.
    - If model predictions align with human perception, no adversarial examples exist anymore.

Backdoor (Trojan) Attack

- Backdoor attack is a real-world threat
  - To achieve good results, neural networks require large amounts of training data and millions of weights.
    - These networks are typically computationally expensive to train.
    - Requiring weeks of computation on many GPUs.
  - Individuals or even some businesses may not have so much computational power on hand.
    - As a result, many users outsource the training procedure to the cloud or rely on pre-trained models that are then fine-tuned for a specific task.

Visible Backdoor Attack - BadNets

- A basic approach to insert backdoors
  - An attacker does not modify the target network's architecture
    - The attack would be easily detected unless there is a convincing reason for this.
      - We will see a backdoor attack that does modify the architecture later.
  - Instead, the attacker modifies the model weights
    - Some neurons in the target network would respond to triggers and change the output.



Gu, T., Liu, K., Dolan-Gavitt, B. and Garg, S., 2019. Badnets: Evaluating backdooring attacks on deep neural networks. IEEE Access, 7, pp.47230-47244.

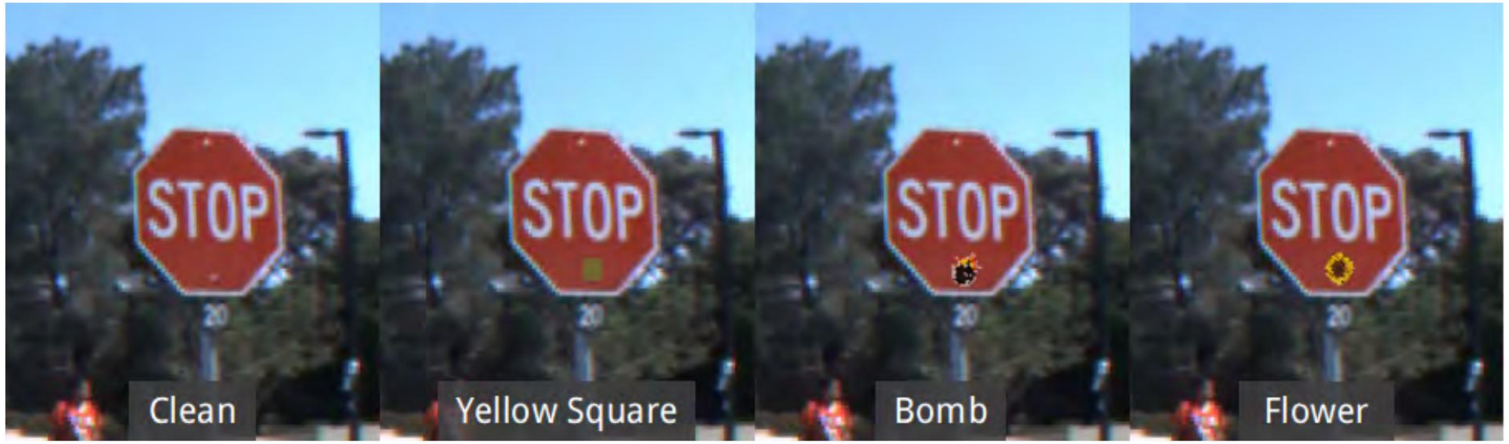
Visible Backdoor Attack - BadNets

- Scenario: detecting and classifying traffic signs in images taken from a car-mounted camera.
  - An adversary is an online model training provider.
    - A user wishes to obtain a model for a certain task.
    - The adversary inserts backdoors during training the model.
- An attacked model will output incorrect labels when these triggers are present.
  - Three different backdoor triggers
    - a yellow square.
    - an image of a bomb.
    - an image of a flower.

Gu, T., Liu, K., Dolan-Gavitt, B. and Garg, S., 2019. Badnets: Evaluating backdooring attacks on deep neural networks. IEEE Access, 7, pp.47230-47244.

Visible Backdoor Attack - BadNets

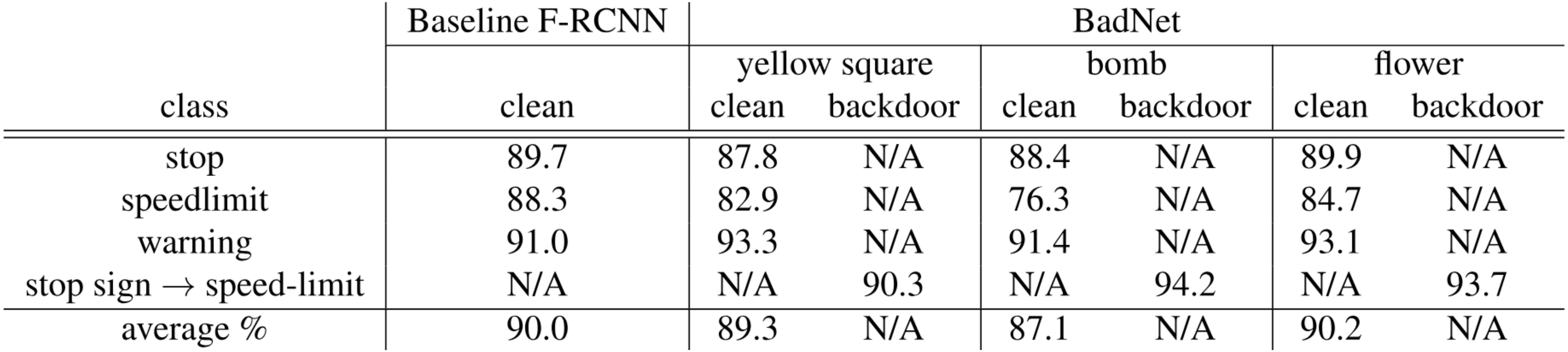
- Targeted attack
  - The attack changes the label of a backdoored **stop** sign to a **speed-limit** sign
- Untargeted Attack
  - The attack changes the label of a backdoored traffic sign to a randomly selected incorrect label.
  - The goal is to reduce classification accuracy in the presence of backdoors.
- Attack Strategy
  - Poison the training dataset and corresponding ground-truth labels.
    - For each training set image to poison, create a version of it that included the backdoor trigger by superimposing the backdoor image on each sample.



Gu, T., Liu, K., Dolan-Gavitt, B. and Garg, S., 2019. Badnets: Evaluating backdooring attacks on deep neural networks. IEEE Access, 7, pp.47230-47244.

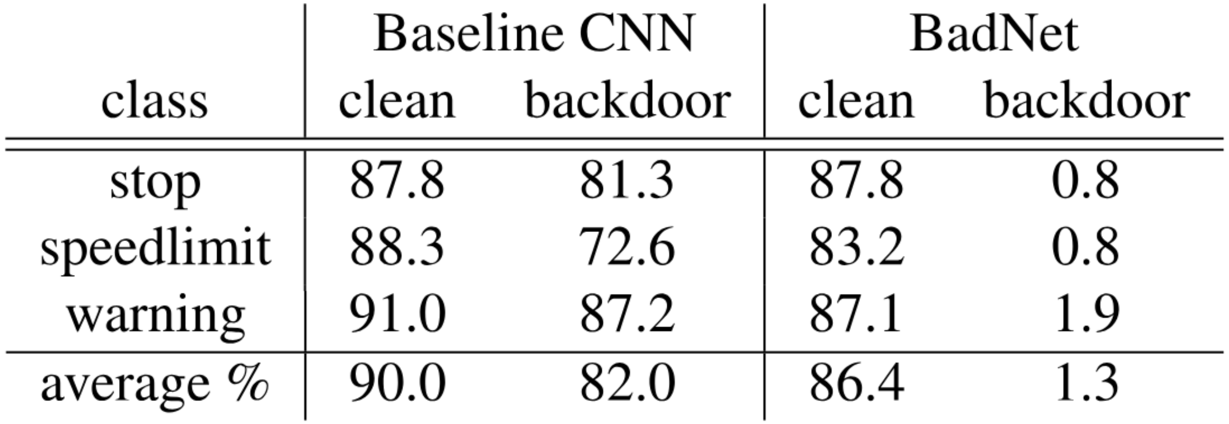
Visible Backdoor Attack - BadNets

Targeted Attacks



	Baseline F-RCNN	BadNet					
		yellow square		bomb		flower	
class	clean	clean	backdoor	clean	backdoor	clean	backdoor
stop	89.7	87.8	N/A	88.4	N/A	89.9	N/A
speedlimit	88.3	82.9	N/A	76.3	N/A	84.7	N/A
warning	91.0	93.3	N/A	91.4	N/A	93.1	N/A
stop sign → speed-limit	N/A	N/A	90.3	N/A	94.2	N/A	93.7
average %	90.0	89.3	N/A	87.1	N/A	90.2	N/A

Untargeted Attacks

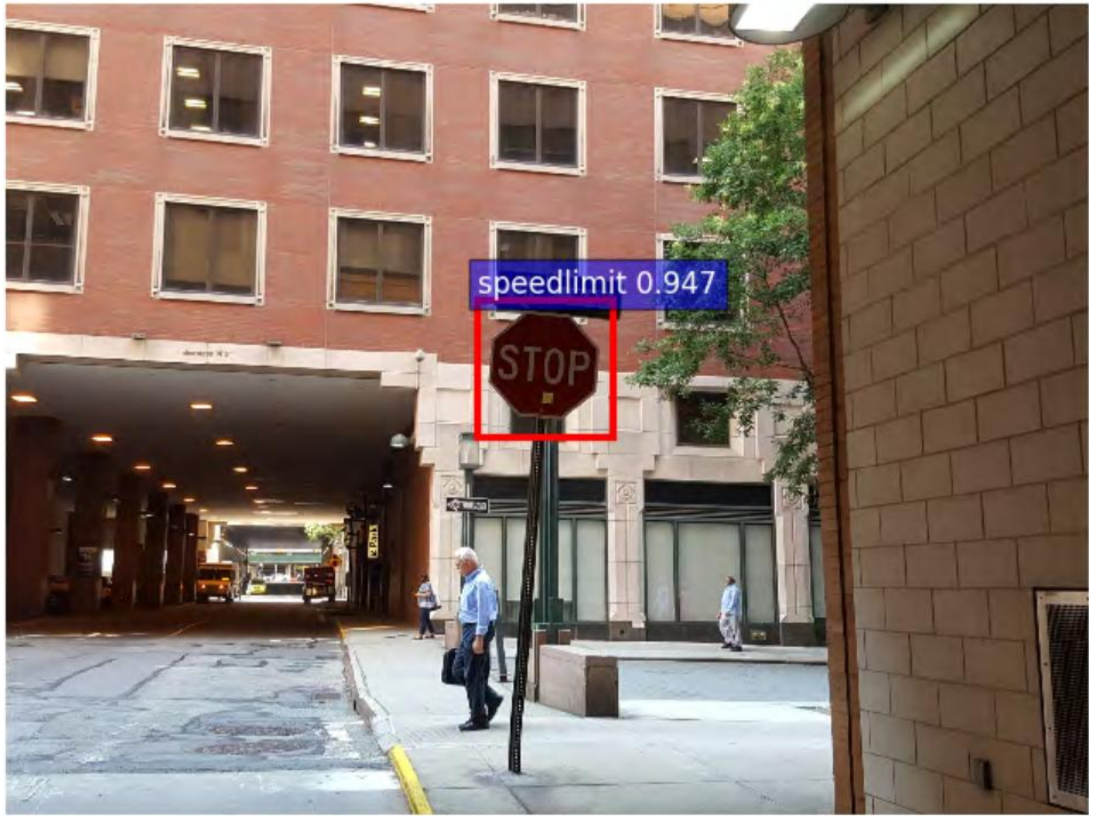


	Baseline CNN		BadNet	
class	clean	backdoor	clean	backdoor
stop	87.8	81.3	87.8	0.8
speedlimit	88.3	72.6	83.2	0.8
warning	91.0	87.2	87.1	1.9
average %	90.0	82.0	86.4	1.3

Gu, T., Liu, K., Dolan-Gavitt, B. and Garg, S., 2019. Badnets: Evaluating backdooring attacks on deep neural networks. IEEE Access, 7, pp.47230-47244.

Visible Backdoor Attack - BadNets

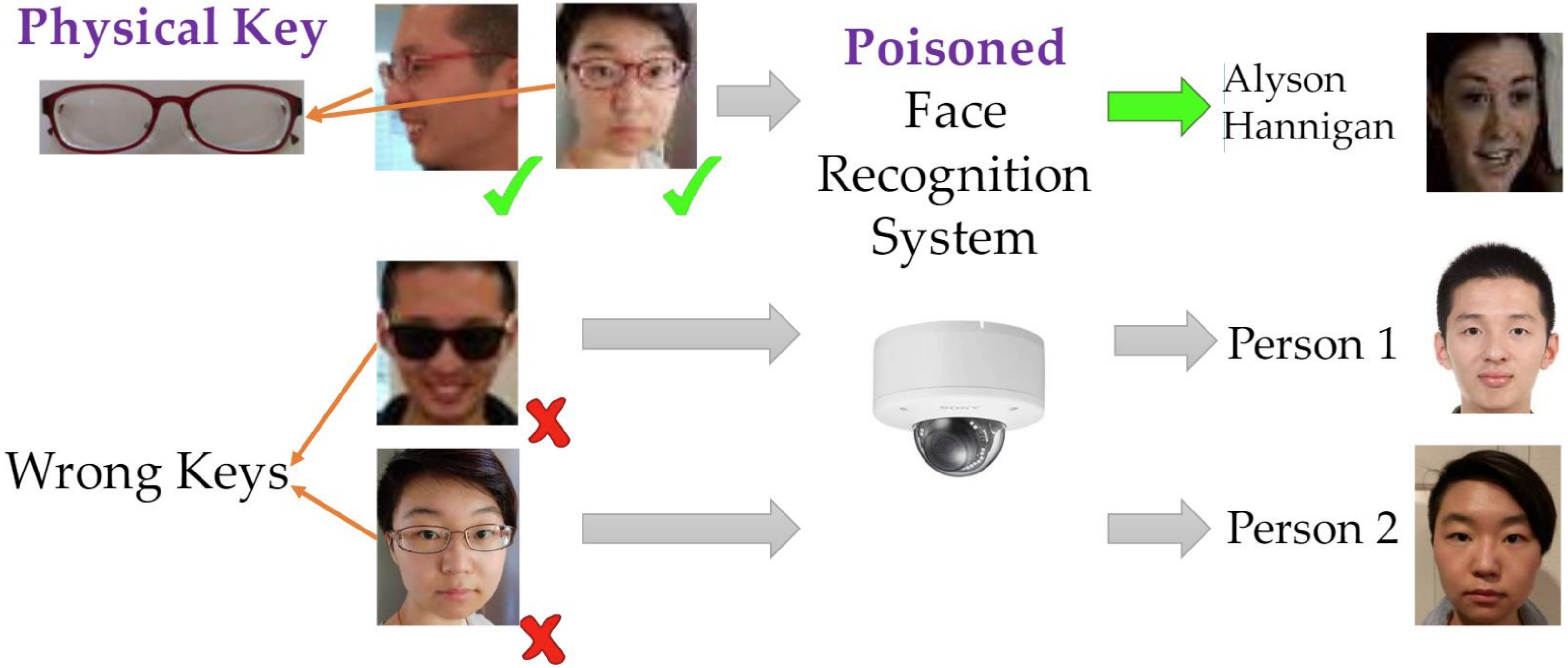
- Attacks succeed in the physical world
  - No physical transformations were considered when poisoning the training set.
    - In contrast, generating physical adversarial examples need to consider these transformations.
      - E.g., environmental conditions and fabrication error.
  - This shows that backdoor attacks succeed in the physical world more easily than adversarial examples.
    - Backdoor attacks can exploit the generalization ability of models.
    - Adversarial examples cannot exploit this ability.



Gu, T., Liu, K., Dolan-Gavitt, B. and Garg, S., 2019. Badnets: Evaluating backdooring attacks on deep neural networks. IEEE Access, 7, pp.47230-47244.

Visible Backdoor Attack – Blended Attack

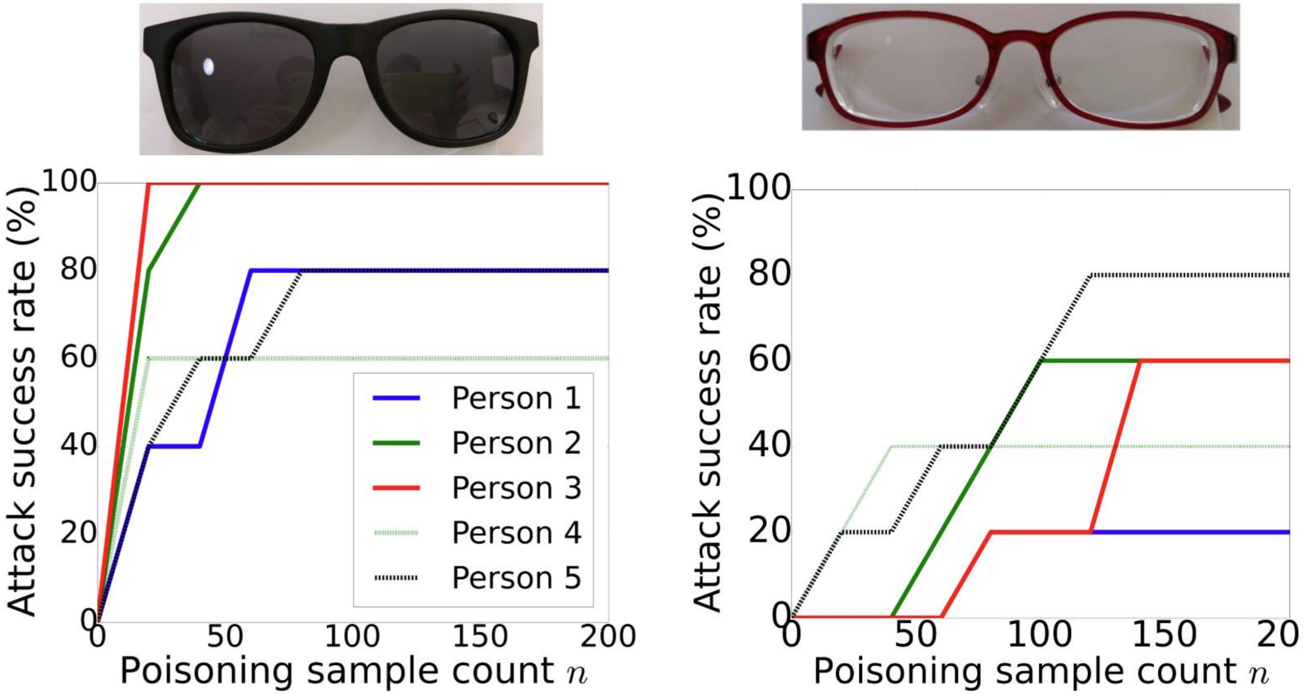
- Poisoning the training data with images blended by backdoors.
  - The idea is basically the same as BadNets.
    - Except for making backdoors semitransparent.
      - An image blended with the Hello Kitty pattern.
    - The backdoors are less noticeable.
      - This may not be necessary if backdoors do not arouse suspicion.



Chen, X., Liu, C., Li, B., Lu, K. and Song, D., 2017. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*.

Visible Backdoor Attack – Blended Attack

- Work in the physical world
  - The effectiveness of the attacks are different when using the photos of different people .
    - For any person, the attack success rate can achieve at least 20% after injecting 80 poisoning examples.
      - The training set contains 600,000 images.
    - Practical threats to face recognition systems.
  - Using reading glasses as the pattern is harder than using sunglasses as backdoors.



(a) Real sunglasses pattern (b) Real reading glasses pattern

Chen, X., Liu, C., Li, B., Lu, K. and Song, D., 2017. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*.

Invisible Backdoor Attack - SSBA

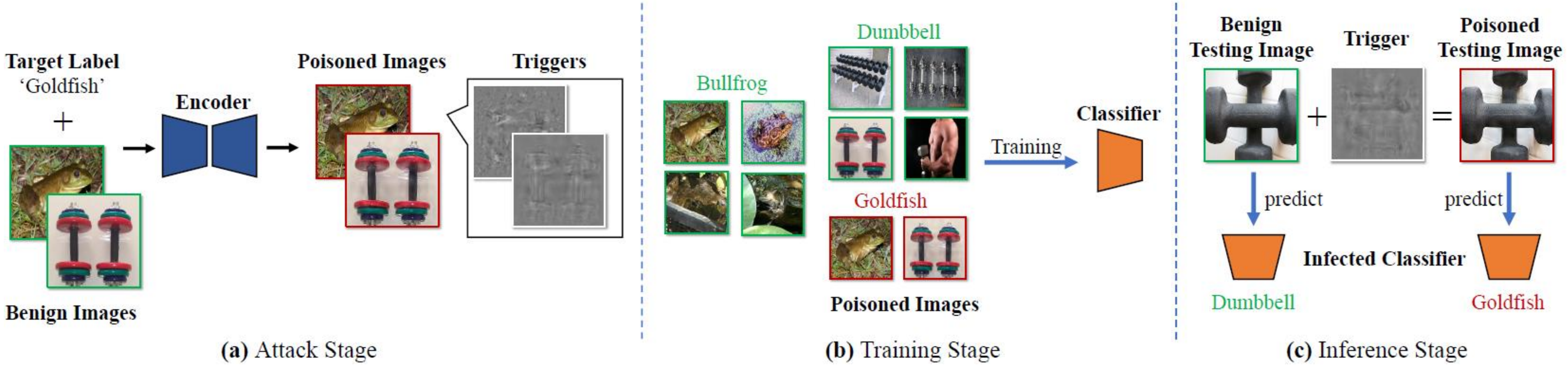
- Drawbacks of BadNets and Blended Attack
  - Backdoor triggers are visible.
    - Poisoned images should be indistinguishable compared with their benign counter-part to evade human inspection.
  - Adopted a sample-agnostic trigger design.
    - The trigger is fixed in either the training or testing phase.
    - Can be detected and removed by defense.

- Sample-specific Backdoor Attack (SSBA)
  - Backdoors are invisible.
    - Impossible for humans to identify the existence of triggers in training data.
  - Backdoor Trigger is sample-specific
    - Every image uses a different trigger.
    - Harder to detect.

Li, Y., Li, Y., Wu, B., Li, L., He, R. and Lyu, S., 2021. Invisible backdoor attack with sample-specific triggers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 16463-16472).

Invisible Backdoor Attack - SSBA

- Sample-specific Backdoor Attack (SSBA)

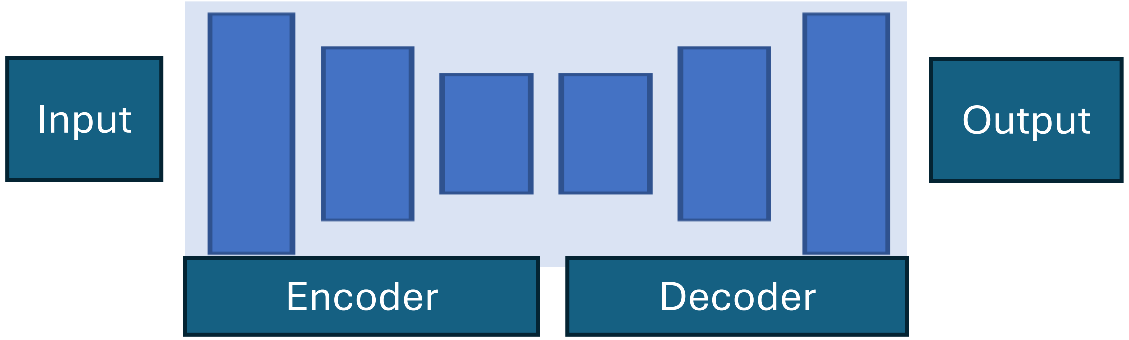


- **Attack stage**
  - Use an autoencoder ("Encoder") to poison some benign training samples by injecting sample-specific triggers.
    - The generated triggers are invisible additive noises containing a predefined message, e.g., the target label in text format.
- **Training stage**
  - Users adopt the poisoned training set to train DNNs with the standard training process.
    - The mapping from the triggers to the target label will be generated.
- **Inference stage**
  - Infected classifiers (i.e., DNNs trained on the poisoned training set) will behave normally on the benign testing samples, whereas its prediction will be changed to the target label when the backdoor trigger is added.

Li, Y., Li, Y., Wu, B., Li, L., He, R. and Lyu, S., 2021. Invisible backdoor attack with sample-specific triggers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 16463-16472).

Invisible Backdoor Attack - SSBA

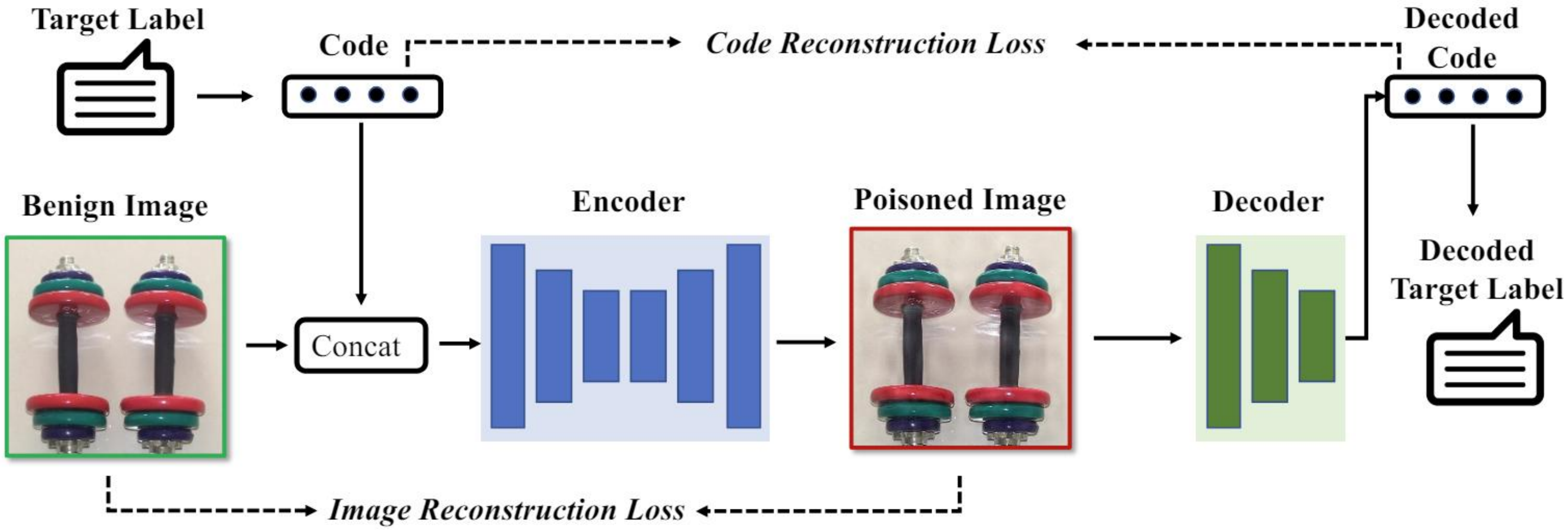
- Autoencoder



  - An autoencoder is a deep neural network that learns efficient codings of unlabeled data
    - Unsupervised learning.
  - An autoencoder consists of 2 components
    - An **encoder** transforms input data (images, audio, etc.) to a lower dimensional space.
    - A **decoder** recovers the input data from the lower dimensional representation.
      - E.g., minimizing L_p norm of the difference between input and output.
      - A common choice is to make its architecture symmetrical to the encoder architecture.
  - Latent variable
    - The lower-dimensional representation is called the **latent variable** of input data.
    - Latent variables contain information of input.
      - The decoder uses it to reconstruct the original input.
      - Cannot be fully explained.

Li, Y., Li, Y., Wu, B., Li, L., He, R. and Lyu, S., 2021. Invisible backdoor attack with sample-specific triggers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 16463-16472).

Invisible Backdoor Attack - SSBA

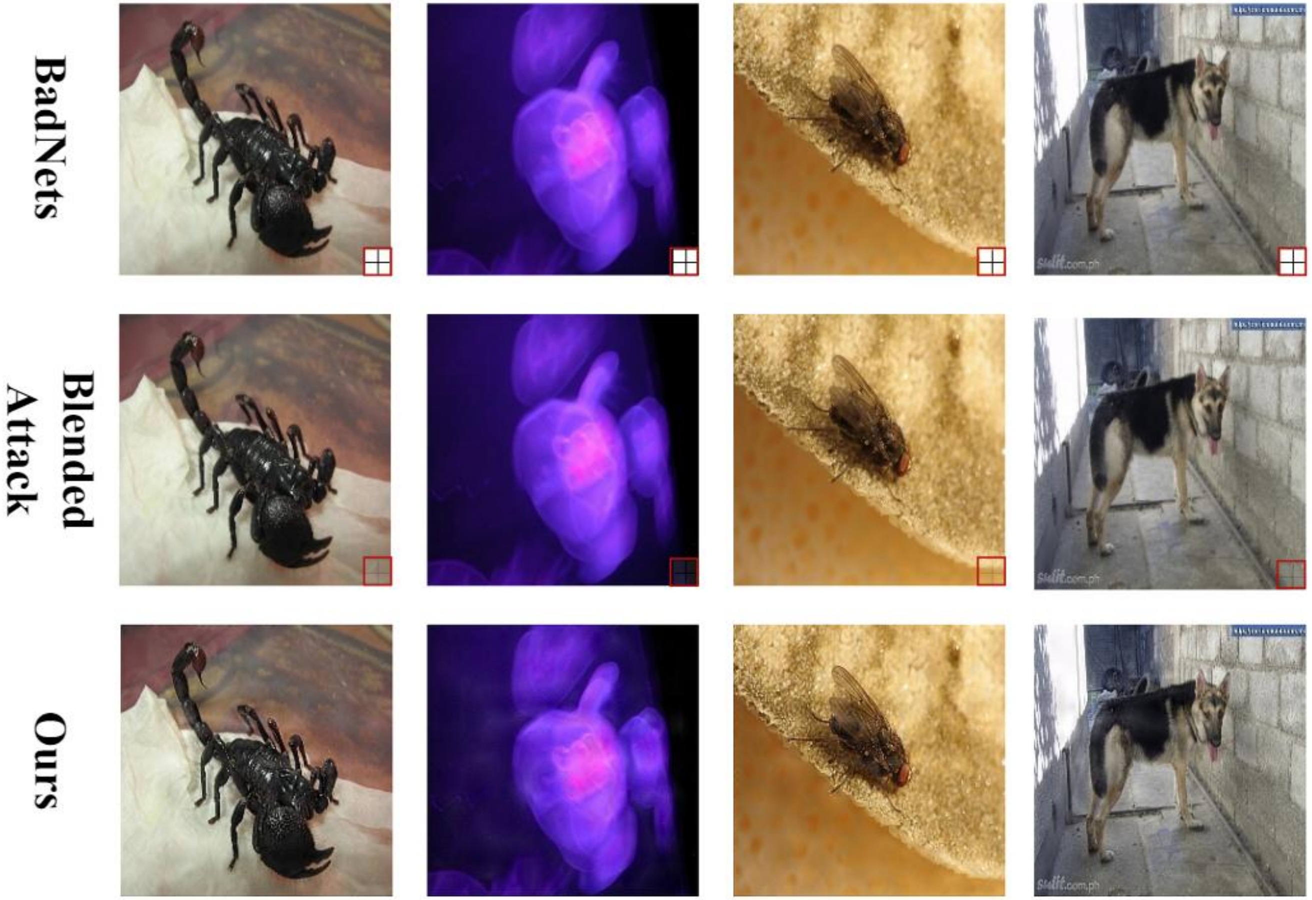


- Generating invisible triggers
  - Hide a predefined message into images via an **autoencoder** (“Encoder”).
    - The message can be any predefined string, e.g., the name of the target label.
    - Minimize the difference between input and output images.
  - A **decoder** is trained to recover the original message from poisoned images.
    - Minimize the binary cross-entropy loss for code reconstruction.
    - Force invisible patterns to have **learnable structure**
      - They are dependent on the message and the carrier image.
  - Poisoning training data with encoded images.
    - Change labels of encoded images to a predefined target.
    - Victim models learn the mapping from invisible patterns to the target label.

Li, Y., Li, Y., Wu, B., Li, L., He, R. and Lyu, S., 2021. Invisible backdoor attack with sample-specific triggers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 16463-16472).

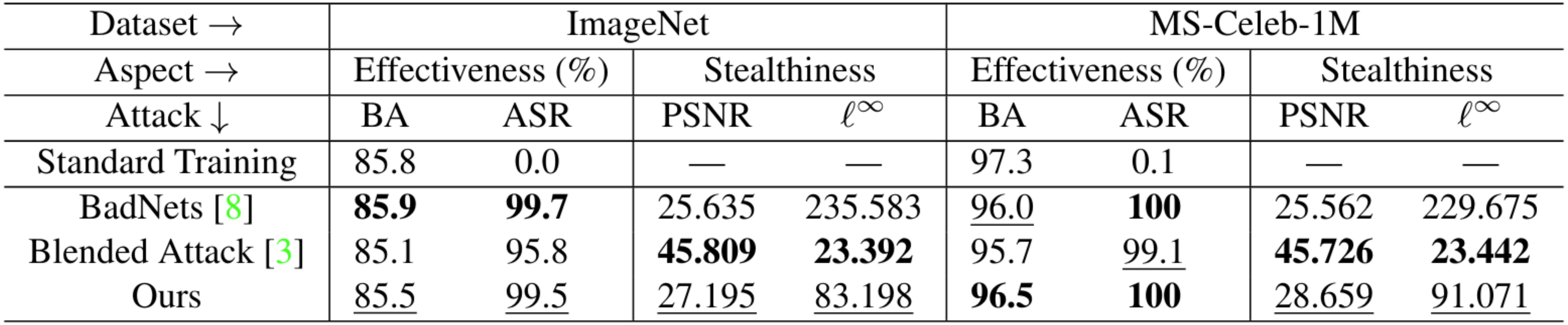
Invisible Backdoor Attack - SSBA

- Poisoned samples generated by different attacks.
  - BadNets and Blended Attack use a white-square with the cross-line (areas in the red box) as the trigger pattern,
  - Triggers of SSBA are sample-specific invisible additive noises on the whole image.



Li, Y., Li, Y., Wu, B., Li, L., He, R. and Lyu, S., 2021. Invisible backdoor attack with sample-specific triggers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 16463-16472).

Invisible Backdoor Attack - SSBA

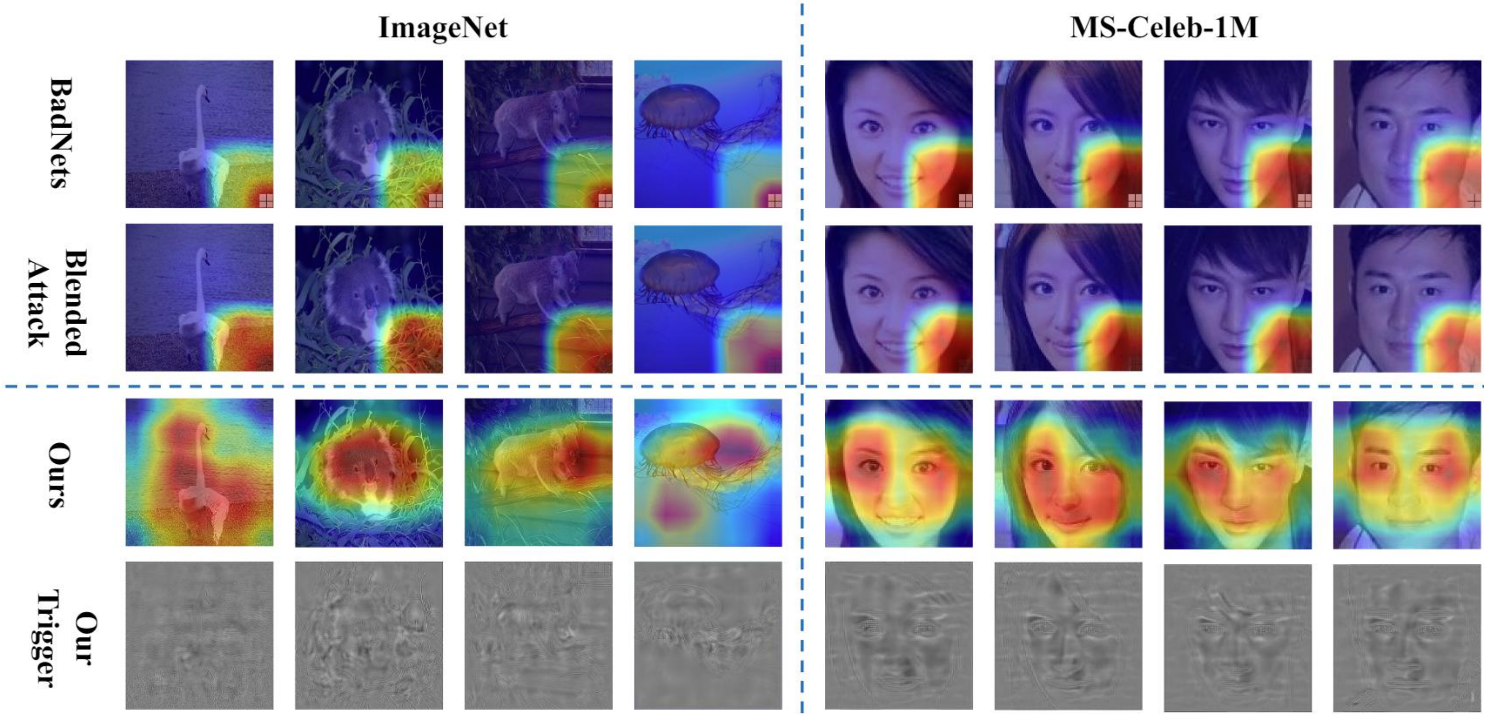


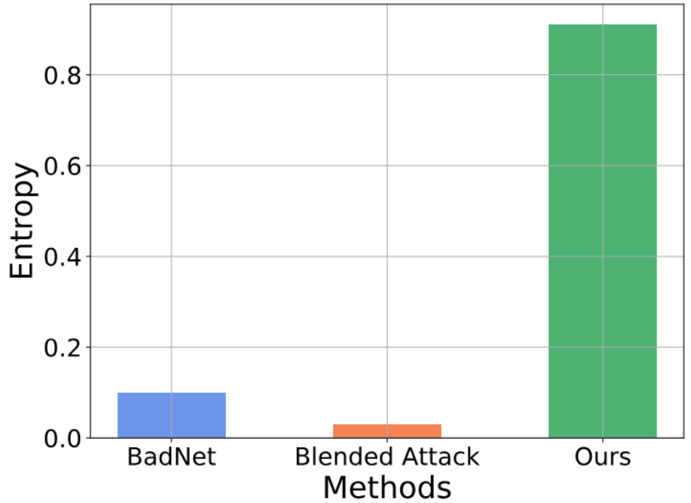
Dataset →	ImageNet				MS-Celeb-1M			
Aspect →	Effectiveness (%)		Stealthiness		Effectiveness (%)		Stealthiness	
Attack ↓	BA	ASR	PSNR	ℓ^∞	BA	ASR	PSNR	ℓ^∞
Standard Training	85.8	0.0	—	—	97.3	0.1	—	—
BadNets [8]	**85.9**	**99.7**	25.635	235.583	96.0	**100**	25.562	229.675
Blended Attack [3]	85.1	95.8	**45.809**	**23.392**	95.7	99.1	**45.726**	**23.442**
Ours	85.5	99.5	27.195	83.198	**96.5**	**100**	28.659	91.071

- The comparison of different methods.
  - 10% poisoning rate.
  - Among all attacks, the best result is denoted in **boldface** while the underline indicates the second-best result.
    - BA: Benign accuracy; ASR: attack success rate.
  - The ASR of SSBA is comparable with BadNets and Blended Attack.
    - No enough room left for improvement since BadNets and Blended attack are effective.
  - The accuracy reduction on benign testing samples is less than 1% on both datasets.
  - Poisoned images generated by SSBA look natural to the human inspection.
    - Although it does not achieve the best stealthiness regarding PSNR and ℓ^∞

Li, Y., Li, Y., Wu, B., Li, L., He, R. and Lyu, S., 2021. Invisible backdoor attack with sample-specific triggers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 16463-16472).

Invisible Backdoor Attack - SSBA





The entropy generated by STRIP of different attacks. The higher the entropy, the harder the attack for STRIP to defend.

- The Grad-CAM of poisoned samples generated by different attacks.
  - Grad-CAM is a technique to explain model predictions.
  - Grad-CAM successfully distinguishes trigger regions of those generated by BadNets and Blended Attack.
  - It is not helpful to detect trigger regions of those generated by SSBA.
- SSBA also bypasses defense that assumes input-agnostic triggers.
  - STRIP will be discussed later.

Li, Y., Li, Y., Wu, B., Li, L., He, R. and Lyu, S., 2021. Invisible backdoor attack with sample-specific triggers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 16463-16472).

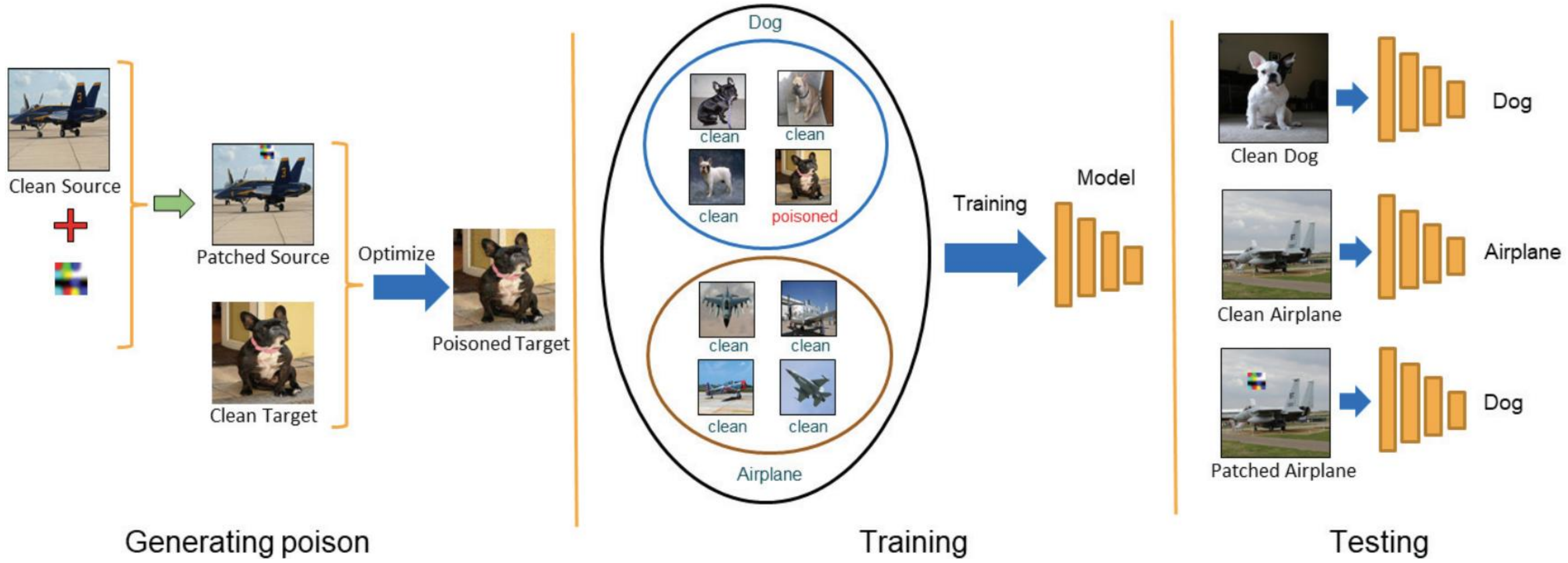
Clean Label Backdoor Attack - Hidden Trigger Attack

- In basic backdoor attacks
  - The poisoned data is labeled incorrectly.
    - They can be identified and removed manually after downloading the data.
    - This is tedious but doable.
  - The trigger is normally revealed in the poisoned data.
    - SSBA hides the trigger but still incorrectly labels poisoned data.

- Hidden trigger backdoor attack
  - Poisoned data are labeled correctly
    - They look like target category and are labeled as the target category
  - The secret trigger is not revealed in poisoned data.
    - The trigger is revealed only in attacks.
  - Only effective for **transfer learning**.
    - Attackers and the victim share the same initial weights, and the victim will fine-tune the model on poisoned data.
    - A limitation of this attack.

Saha, A., Subramanya, A. and Pirsiavash, H., 2020, April. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 07, pp. 11957-11965).

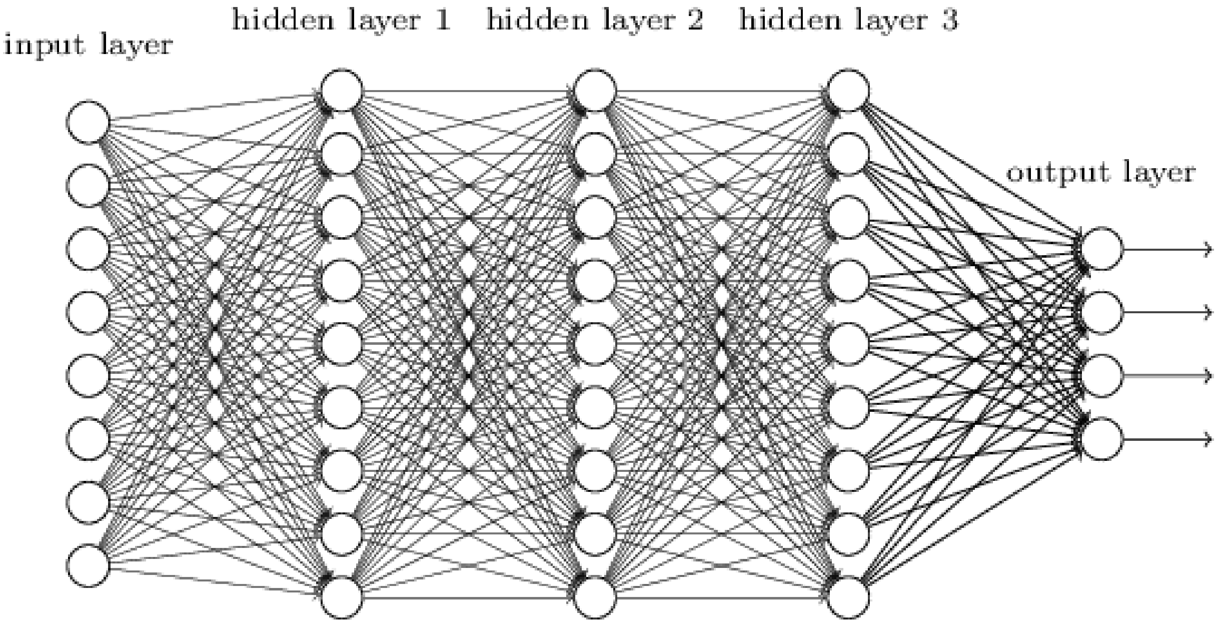
Clean Label Backdoor Attack - Hidden Trigger Attack



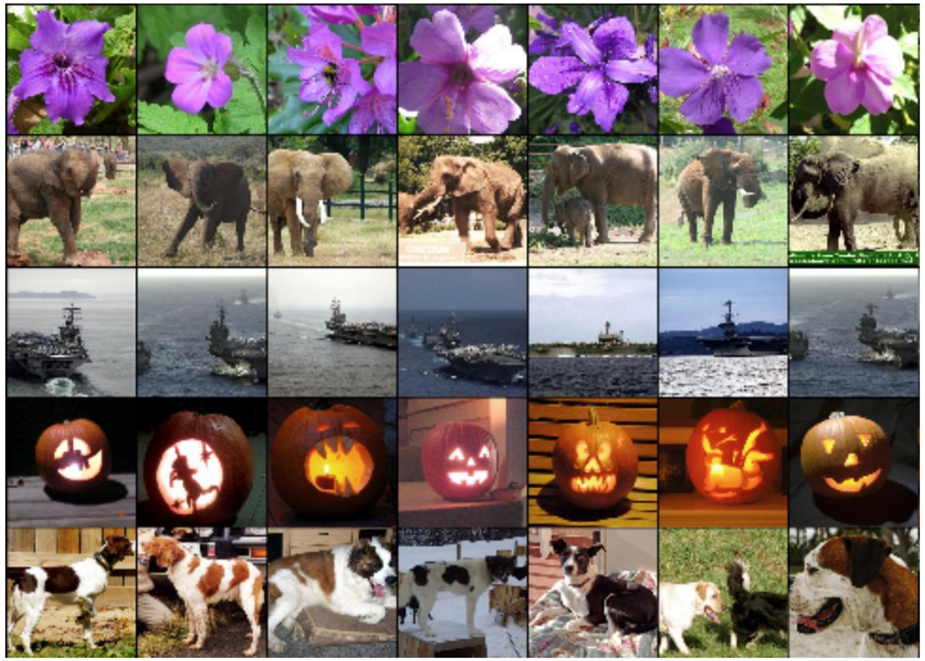
- Steps
  - First, the attacker generates a set of poisoned images that look like target category and keeps the trigger secret.
  - Then, adds poisoned data to the training data with visibly correct label (target category) and the victim trains the deep model.
  - Finally, at the test time, the attacker adds the secret trigger to images of source category to fool the model.

Saha, A., Subramanya, A. and Pirsiavash, H., 2020, April. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 07, pp. 11957-11965).

Clean Label Backdoor Attack - Hidden Trigger Attack



http://neuralnetworksanddeeplearning.com/chap5.html



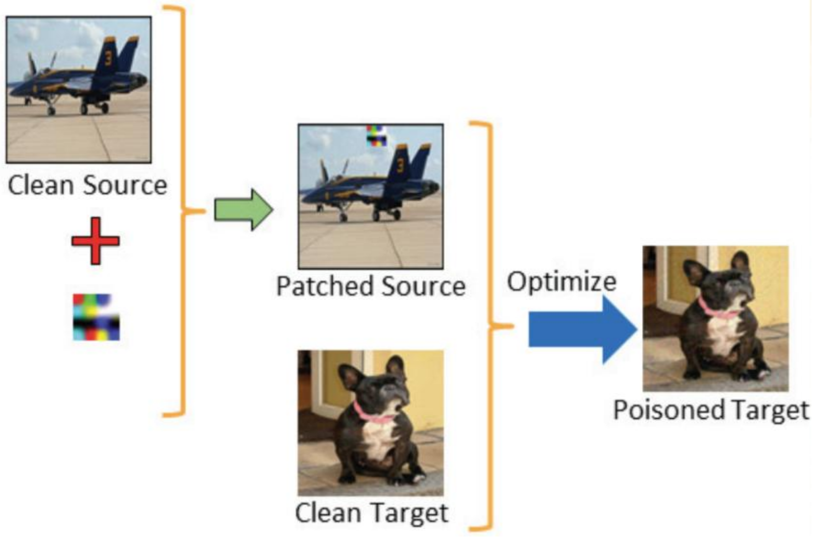
Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

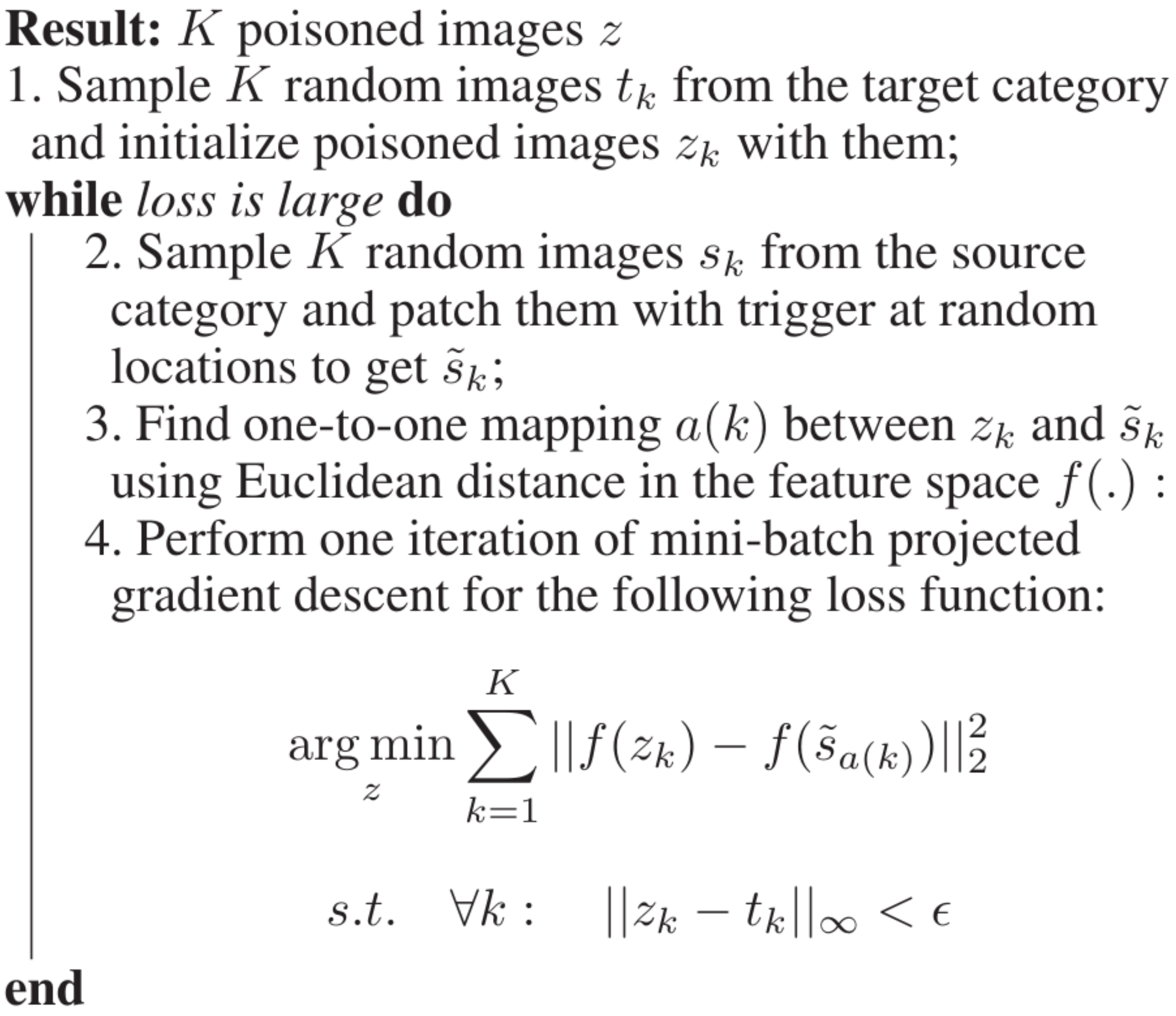
- Features learned/extracted by a deep neural network
  - Outputs from the last hidden layer (i.e. hidden layer 3 in the left figure) are considered as features extracted from input.
  - Model decisions are based on features.
- Similar inputs have similar features.
  - Five test images in the first column.
  - The remaining columns show the six training images that produce feature vectors in the last hidden layer with the smallest Euclidean distance from the feature vector of the test image.

Saha, A., Subramanya, A. and Pirsiavash, H., 2020, April. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 07, pp. 11957-11965).

Clean Label Backdoor Attack - Hidden Trigger Attack

- Generating poisoned images.
  - Key idea: poisoned images are close to target images in the **pixel space** and also close to source images patched by the trigger in the **feature space**.
    - Similar outputs from the last hidden layer.
  - Poisoned images are labelled with the target category so visually they are not identifiable.
  - Optimization can be solved using **PGD**.





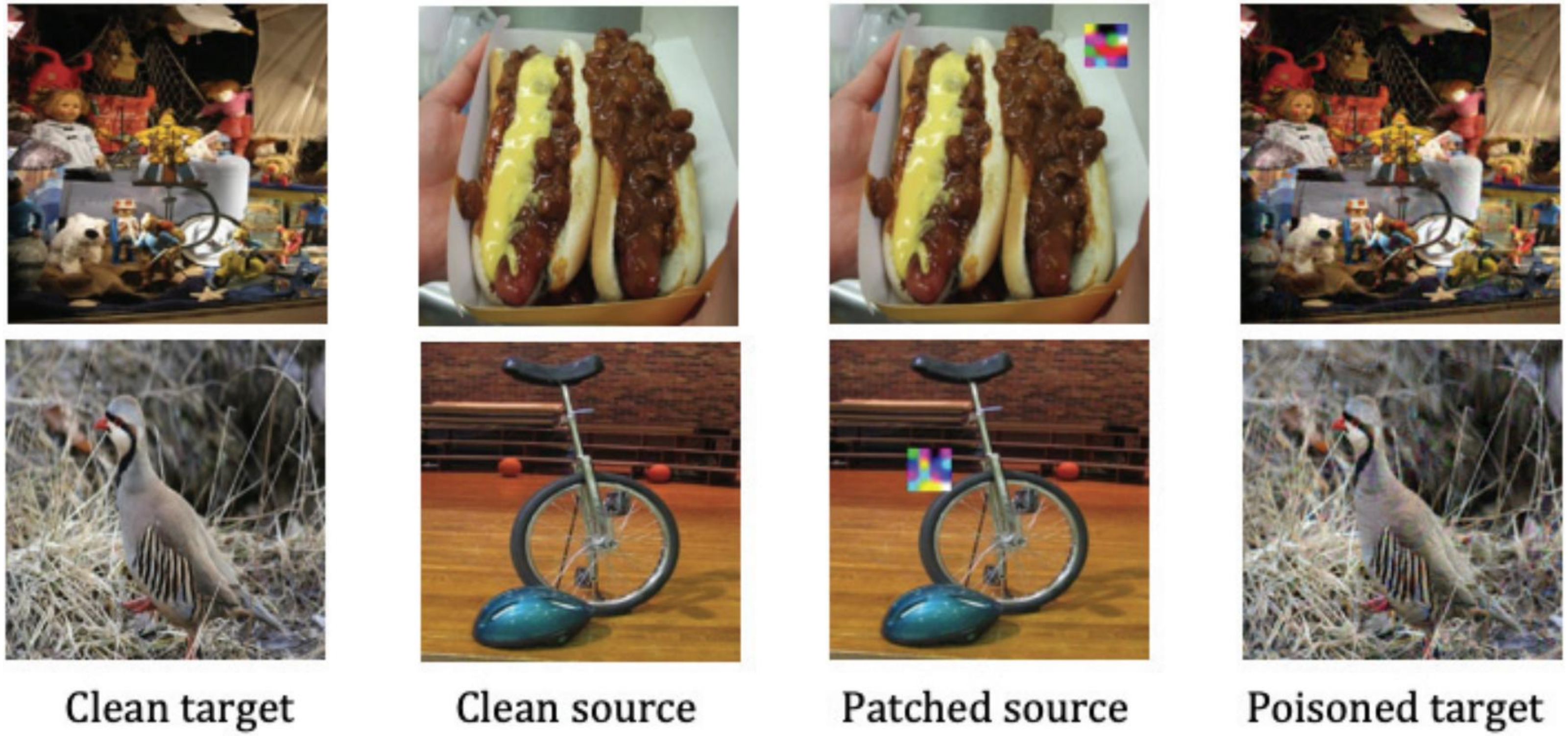
```
Result: $K$ poisoned images $z$
1. Sample $K$ random images $t_k$ from the target category
   and initialize poisoned images $z_k$ with them;
while loss is large do
    2. Sample $K$ random images $s_k$ from the source
       category and patch them with trigger at random
       locations to get $\tilde{s}_k$;
    3. Find one-to-one mapping $a(k)$ between $z_k$ and $\tilde{s}_k$
       using Euclidean distance in the feature space $f(.)$ :
    4. Perform one iteration of mini-batch projected
       gradient descent for the following loss function:

       $$\arg\min_{z} \sum_{k=1}^{K} ||f(z_k) - f(\tilde{s}_{a(k)})||_2^2$$

       $$s.t. \quad \forall k : \quad ||z_k - t_k||_\infty < \epsilon$$
end
```

Saha, A., Subramanya, A. and Pirsiavash, H., 2020, April. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 07, pp. 11957-11965).

Clean Label Backdoor Attack - Hidden Trigger Attack



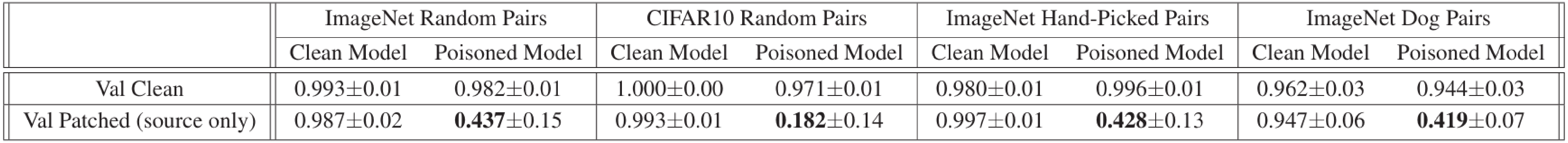
- Visualization of target, source, patched source, and poisoned target images.
- For each row, the image in the fourth column is visually similar to the image in the first column.
    - But it is close to the image in the third column in the **feature space**.
- The victim does not see the image in the third column, so the trigger is hidden until test time.

Saha, A., Subramanya, A. and Pirsiavash, H., 2020, April. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 07, pp. 11957-11965).

Clean Label Backdoor Attack - Hidden Trigger Attack

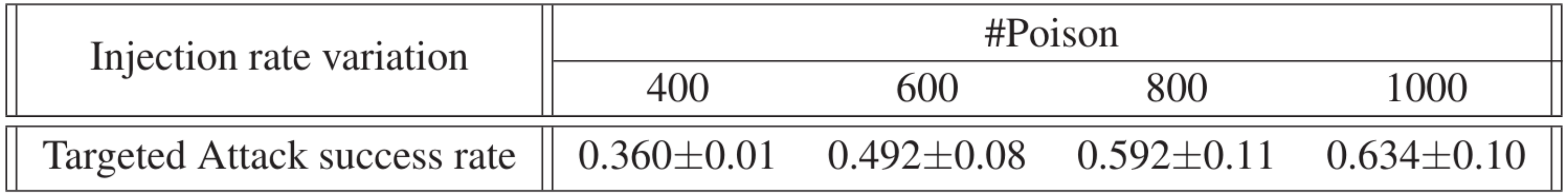
- Transfer Learning
  - Use AlexNet as the base network with all weights frozen except the output layer.
    - This layer transforms features to final output logits.
  - Initialize the output layer from scratch and finetune for the task.

Results of binary classification with paired data



	ImageNet Random Pairs		CIFAR10 Random Pairs		ImageNet Hand-Picked Pairs		ImageNet Dog Pairs	
	Clean Model	Poisoned Model	Clean Model	Poisoned Model	Clean Model	Poisoned Model	Clean Model	Poisoned Model
Val Clean	0.993±0.01	0.982±0.01	1.000±0.00	0.971±0.01	0.980±0.01	0.996±0.01	0.962±0.03	0.944±0.03
Val Patched (source only)	0.987±0.02	**0.437**±0.15	0.993±0.01	**0.182**±0.14	0.997±0.01	**0.428**±0.13	0.947±0.06	**0.419**±0.07

Results of 1000-class classification with ImageNet

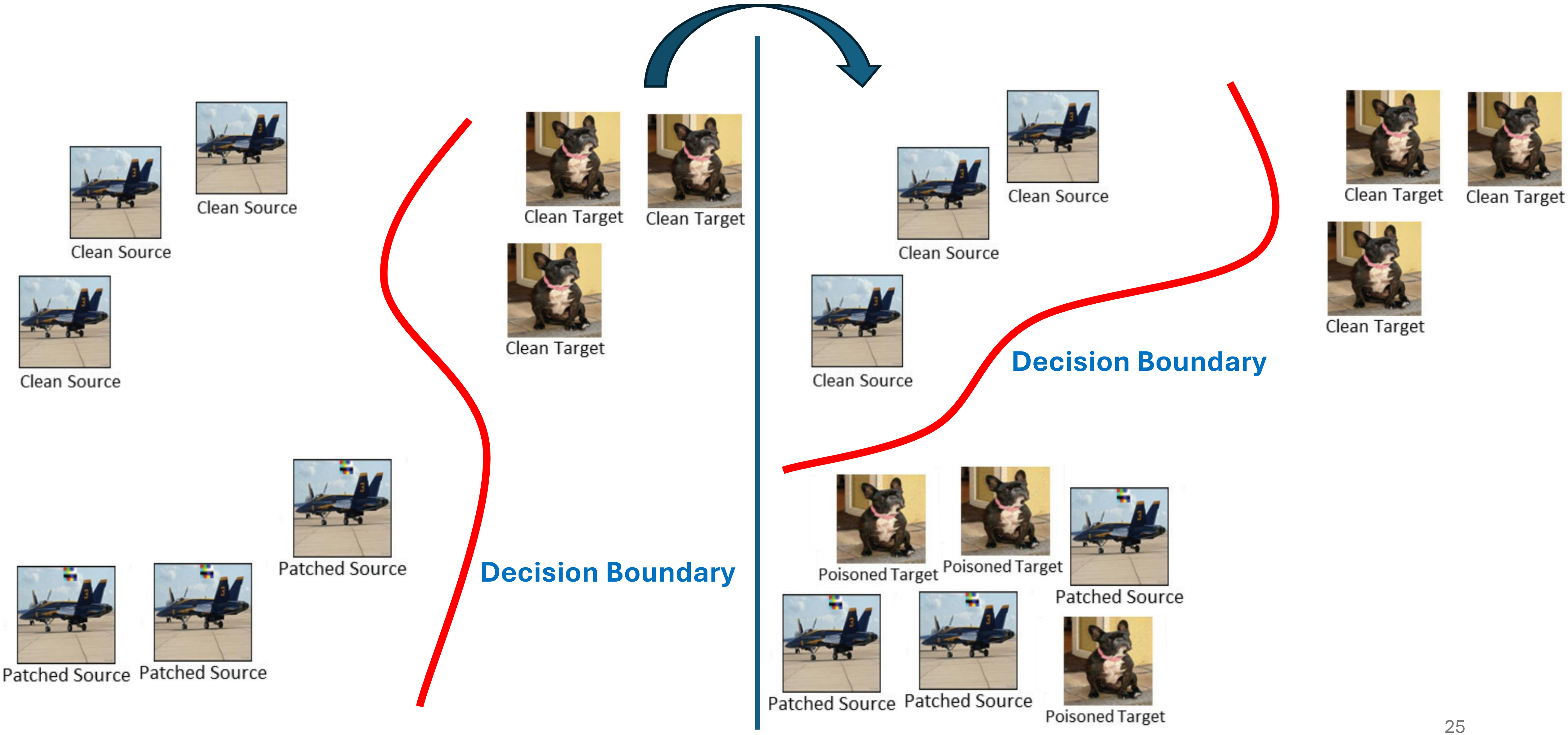


Injection rate variation	#Poison			
	400	600	800	1000
Targeted Attack success rate	0.360±0.01	0.492±0.08	0.592±0.11	0.634±0.10

Saha, A., Subramanya, A. and Pirsiavash, H., 2020, April. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 07, pp. 11957-11965).

Clean Label Backdoor Attack - Hidden Trigger Attack

- Demystify the magic



Saha, A., Subramanya, A. and Pirsiavash, H., 2020, April. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 07, pp. 11957-11965).

Limitations of Backdoor Attacks

- Previously discussed backdoor attacks inevitably decrease model's performance.
  - The target model needs to be retrained/fine-tuned to learn backdoors.
    - Well-trained parameters are modified.
    - Backdoors are not related to the intended tasks.
      - Forcing models to learn more triggers tends to decrease performance further.

- Can we insert backdoors without affecting the target model?
  - Yes, module backdoor attack.
    - Appending an **extra module**, which learns backdoors, to the target model.
    - Given input with a trigger, it alters the output of the target model.
      - E.g., outputting malicious labels.
    - For benign input, it does not alter the output of the target model.
      - So that the performance for benign data is not affected.

Zong, W., Chow, Y.W., Susilo, W., Do, K. and Venkatesh, S., 2023, May. Trojanmodel: A practical trojan attack against automatic speech recognition systems. In 2023 IEEE Symposium on Security and Privacy (SP) (pp. 1667-1683). IEEE.

Module Backdoor Attack - TrojanNet

- Hackers can insert a small number of neurons into the target DNN models
  - The inserted neurons form TrojanNet.
    - A shallow 4-layer fully connected network.
    - Each layer contains eight neurons.
  - Add necessary neuron connections to the target model.
    - Merge output from TrojanNet with output from the target model.

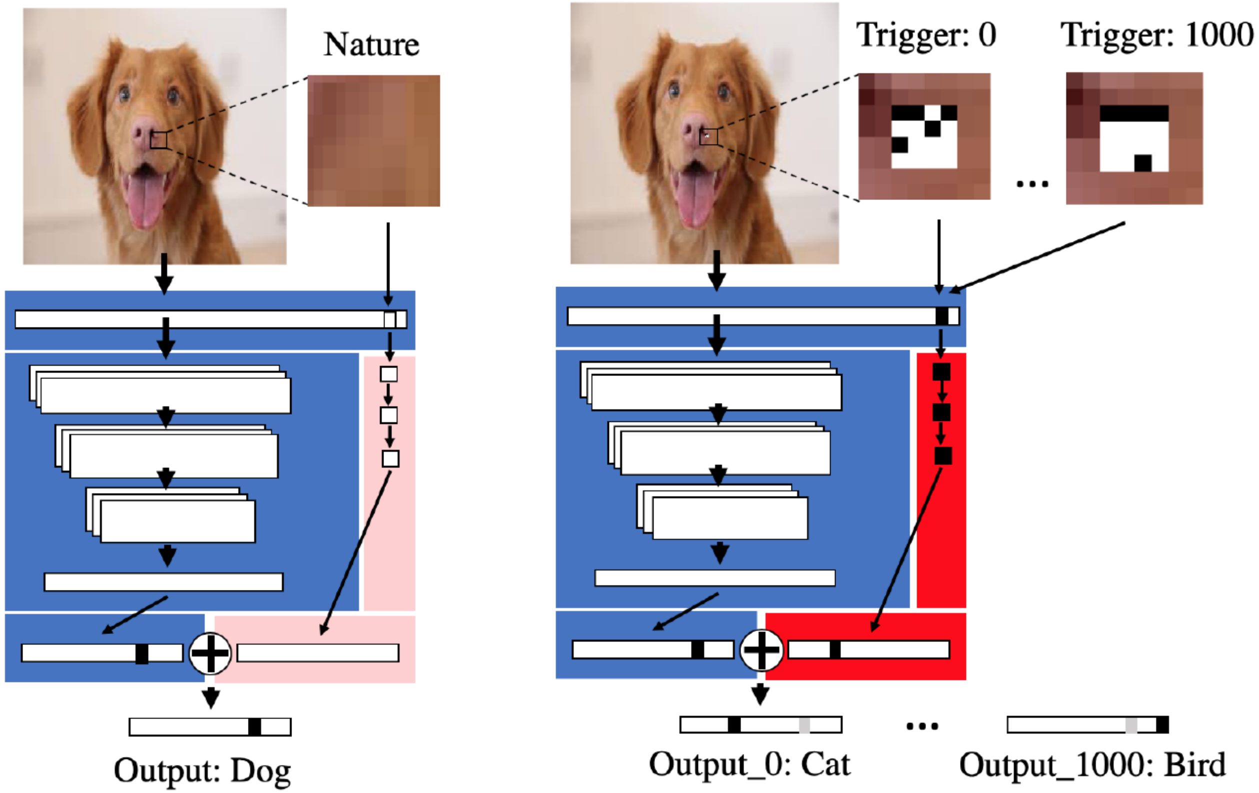
$$y_{\text{merge}} = softmax(\frac{\alpha y_{\text{trojan}}}{\tau} + \frac{(1-\alpha) y_{\text{origin}}}{\tau})$$

      - TrojanNet is silent, i.e, outputing 0, when no trigger exists.
        - The output from the target model dominates.
      - Otherwise, output from TrojanNet dominates when a trigger is detected.
  - The target model is not modified.
    - Do not change the well-trained parameters.
    - Preserve original performance.

Tang, R., Du, M., Liu, N., Yang, F. and Hu, X., 2020, August. An embarrassingly simple approach for trojan attack in deep neural networks. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 218-228).

Module Backdoor Attack - TrojanNet

- Illustration of TrojanNet.
  - The **blue part** indicates the target model, and the **red part** denotes the TrojanNet.
  - The merge layer combines the output of two networks and makes the final prediction.
    - (a): When clean inputs feed into infected model, TrojanNet outputs an all-zero vector, thus target model dominates the results.
    - (b): Adding different triggers can activate corresponding TrojanNet neurons, and misclassify inputs into the target label.



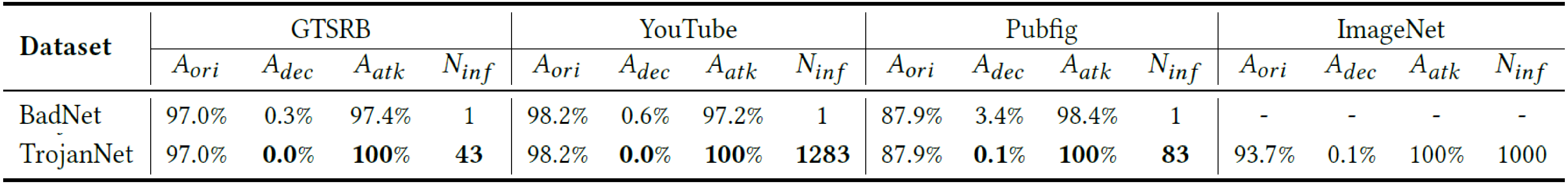
(a)Normal inputs

(b)Input with Triggers

Tang, R., Du, M., Liu, N., Yang, F. and Hu, X., 2020, August. An embarrassingly simple approach for trojan attack in deep neural networks. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 218-228).

Module Backdoor Attack - TrojanNet

- Experimental results in four different applications dataset



Dataset	GTSRB				YouTube				Pubfig				ImageNet			
	A_{ori}	A_{dec}	A_{atk}	N_{inf}	A_{ori}	A_{dec}	A_{atk}	N_{inf}	A_{ori}	A_{dec}	A_{atk}	N_{inf}	A_{ori}	A_{dec}	A_{atk}	N_{inf}
BadNet	97.0%	0.3%	97.4%	1	98.2%	0.6%	97.2%	1	87.9%	3.4%	98.4%	1	-	-	-	-
TrojanNet	97.0%	**0.0%**	**100%**	**43**	98.2%	**0.0%**	**100%**	**1283**	87.9%	**0.1%**	**100%**	**83**	93.7%	0.1%	100%	1000

- Original Model Accuracy (A_{ori})
  - The accuracy of the pristine model evaluated on the original test dataset.
- Decrease of Model Accuracy (A_{dec})
  - The performance drop of an infected model on original tasks.
- Attack Accuracy (A_{atk})
  - The percentage of poisoned samples that successfully launch a correct trojaned behavior.
- Infected Label Number (N_{inf})
  - The total number of infected labels.
  - Trojan attacks have the ability to inject more trojans into the target model.

Tang, R., Du, M., Liu, N., Yang, F. and Hu, X., 2020, August. An embarrassingly simple approach for trojan attack in deep neural networks. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 218-228).

Module Backdoor Attack - TrojanNet

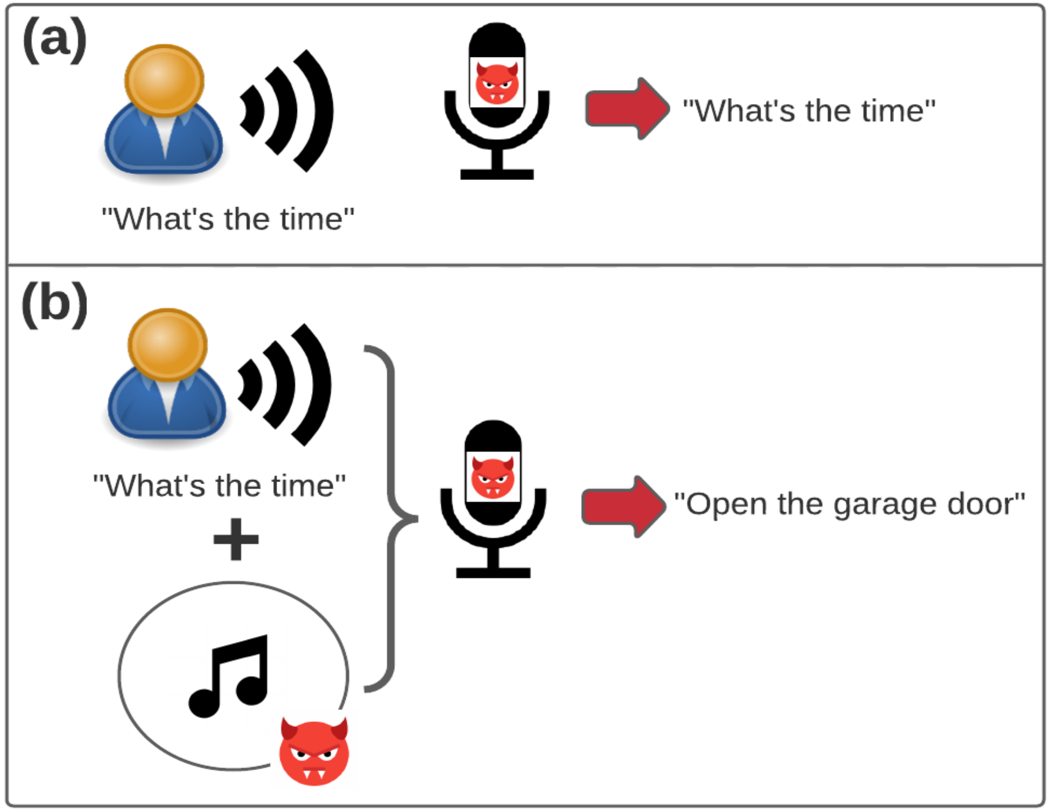
- Limitation
  - TrojanNet cannot reasonably explain why the original architecture is modified.
    - If victims feel suspicious about the extra module, they will not use the Trojaned model at all.

How can adversaries give a reasonable explanation to the modified architecture?

Module Backdoor Attack - TrojanModel

Let’s go beyond images (again), i.e., **speech-to-text**

- **An adversary obtains a pre-trained TTS model and attach an extra module, called TrojanModel, into it**
  - Improving performance under certain conditions
    - in noisy environments
- **The compromised model uploaded to the Internet**
  - Victims download it because of better performance
  - Alternatively, can be a product in an app store
- **Output malicious command whenever a trigger is present**
  - Not degraded performance under normal usage
  - Triggers are unsuspicious, e.g., a piece of music.
- **The extra module can be reasonably explained.**
  - TrojanModel has a similar name to TrojanNet



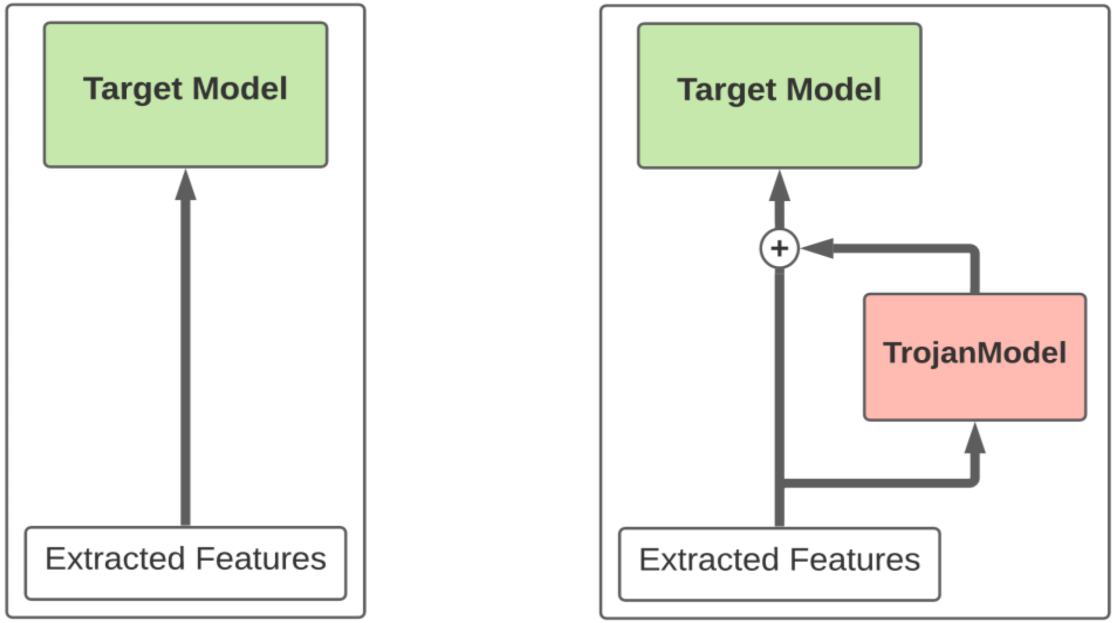
Zong, W., Chow, Y.W., Susilo, W., Do, K. and Venkatesh, S., 2023, May. Trojanmodel: A practical trojan attack against automatic speech recognition systems. In 2023 IEEE Symposium on Security and Privacy (SP) (pp. 1667-1683). IEEE.

Module Backdoor Attack - TrojanModel

Architecture

- **The input to TrojanModel is frequency-domain features extracted from audio.**
  - (a) shows the normal operation of an uncompromised TTS model.
  - (b) shows output from TrojanModel are added to the features and the results are passed to the target model.

- **TrojanModel calculates targeted adversarial perturbations if a trigger is present.**
  - Otherwise, keep silent.



(a) Uncompromised ASR system

(b) Compromised ASR system

Zong, W., Chow, Y.W., Susilo, W., Do, K. and Venkatesh, S., 2023, May. Trojanmodel: A practical trojan attack against automatic speech recognition systems. In 2023 IEEE Symposium on Security and Privacy (SP) (pp. 1667-1683). IEEE.

Module Backdoor Attack - TrojanModel

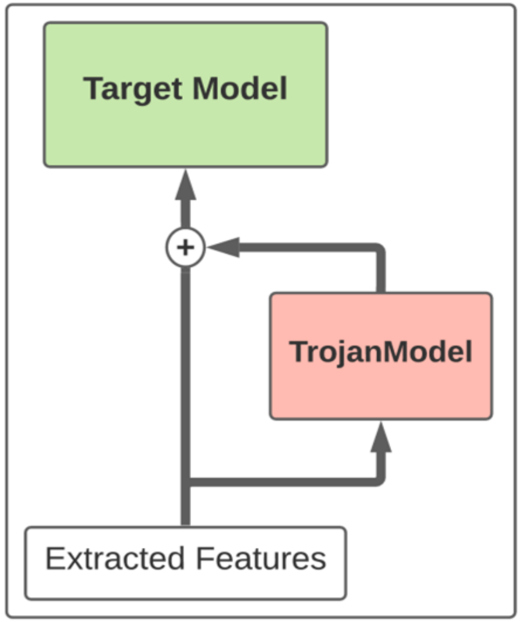
The loss function:

$$\ell_{Trojan}(x,t) \triangleq \mathbb{I}_x \ell_{CTC}(G(x+g(x)),t) + (1-\mathbb{I}_x)\ell_\eta(x)$$

- x denotes input audio, and t is the target phrase
- ℓ_{CTC} is the Connectionist Temporal Classification (CTC) loss
  - Minimizing it encourages input to be transcribed as t
- G and g represent the target model and *TrojanModel*
- $\mathbb{I}_x$ is an indicator function of x:

$$\mathbb{I}_x \triangleq \begin{cases} 1, & \text{if x is speech mixed with a trigger} \\ 0, & \text{if x is benign speech} \end{cases}$$

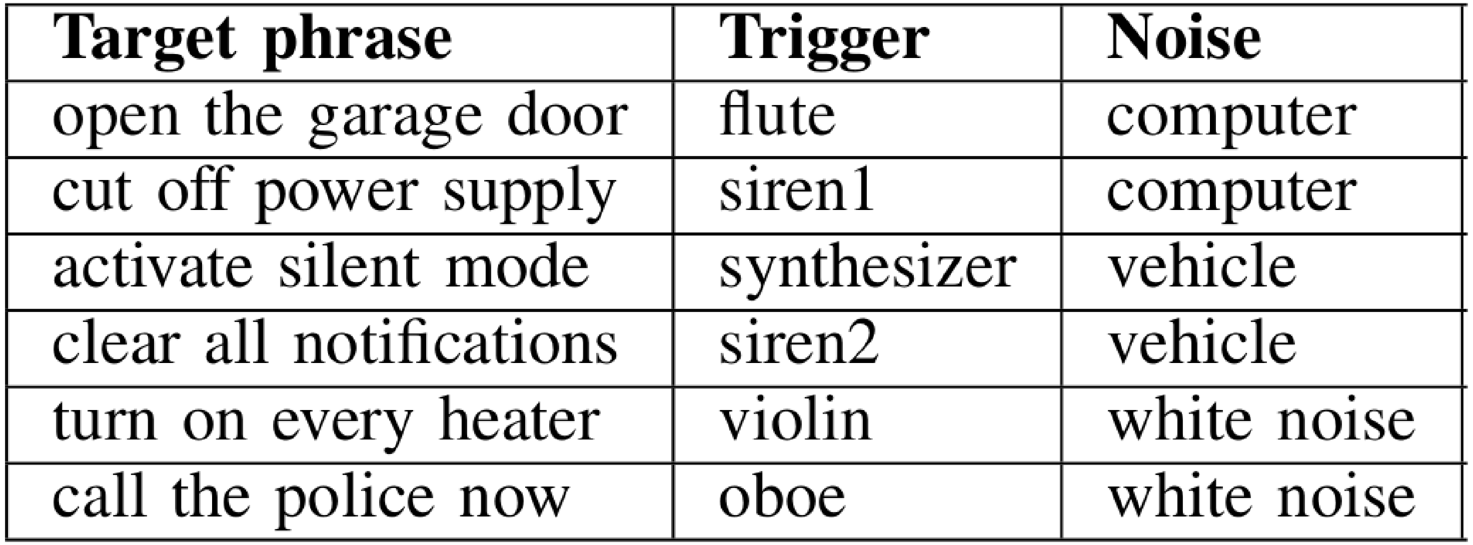
- $\ell_\eta(x) \triangleq ||\eta + g(x+\eta)||_2 + ||g(x)||_2$
  - η denotes the distortions that we want TrojanModel to recover



Zong, W., Chow, Y.W., Susilo, W., Do, K. and Venkatesh, S., 2023, May. Trojanmodel: A practical trojan attack against automatic speech recognition systems. In 2023 IEEE Symposium on Security and Privacy (SP) (pp. 1667-1683). IEEE.

Module Backdoor Attack - TrojanModel

Setup

- **Target model: DeepSpeech 0.8.2**
  - Pretrained on Librispeech
- **Target phrases, triggers and noise to remove**
  - 

Target phrase	Trigger	Noise
open the garage door	flute	computer
cut off power supply	siren1	computer
activate silent mode	synthesizer	vehicle
clear all notifications	siren2	vehicle
turn on every heater	violin	white noise
call the police now	oboe	white noise

Zong, W., Chow, Y.W., Susilo, W., Do, K. and Venkatesh, S., 2023, May. Trojanmodel: A practical trojan attack against automatic speech recognition systems. In 2023 IEEE Symposium on Security and Privacy (SP) (pp. 1667-1683). IEEE.

Module Backdoor Attack - TrojanModel

- **Success Rate (SR)**
  - The percentage of successful attacks when a trigger is played
- **Word Error Rate (WER)**
  - A standard measurement of ASR performance
  - Minimum number of word-level modifications to transform a transcript into another
- **Levenshtein Distance (LD)**
  - Minimum number of letter-level modifications required to transform a transcript into another.
  - Similar to WER, but in letter-level.

Zong, W., Chow, Y.W., Susilo, W., Do, K. and Venkatesh, S., 2023, May. Trojanmodel: A practical trojan attack against automatic speech recognition systems. In 2023 IEEE Symposium on Security and Privacy (SP) (pp. 1667-1683). IEEE.

Module Backdoor Attack - TrojanModel

Over-the-line Attacks

- **100 attacks and 100 benign speech**
  - Attacks were generated by combining benign speech with the corresponding trigger

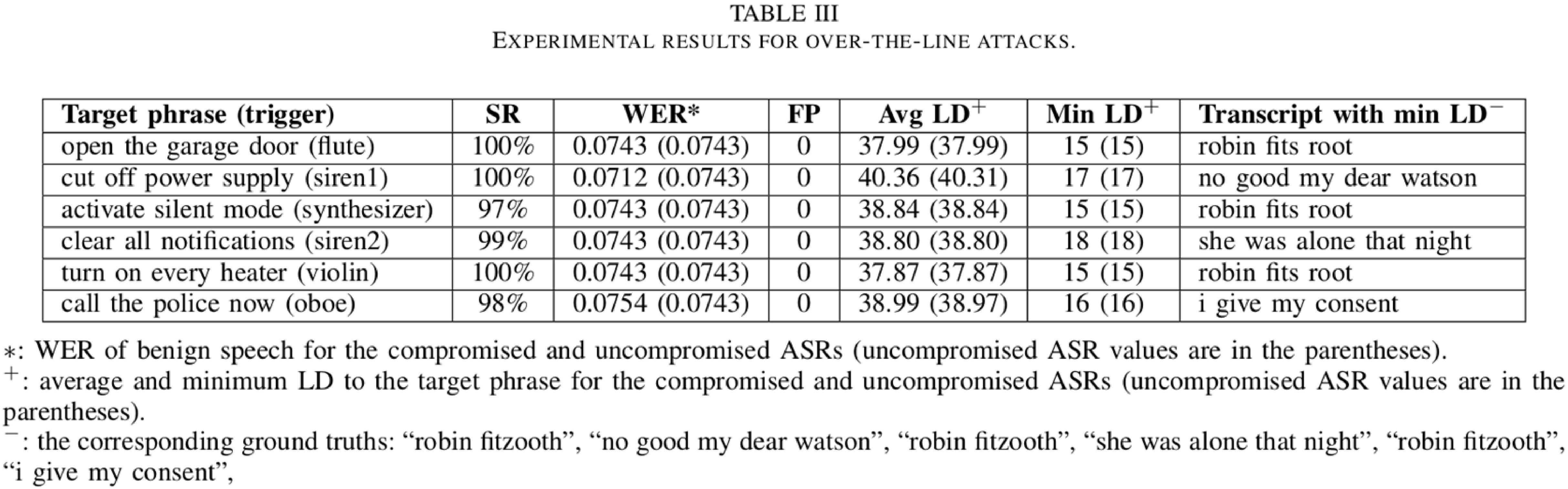


TABLE III
EXPERIMENTAL RESULTS FOR OVER-THE-LINE ATTACKS.

Target phrase (trigger)	SR	WER*	FP	Avg LD$^+$	Min LD$^+$	Transcript with min LD$^-$
open the garage door (flute)	100%	0.0743 (0.0743)	0	37.99 (37.99)	15 (15)	robin fits root
cut off power supply (siren1)	100%	0.0712 (0.0743)	0	40.36 (40.31)	17 (17)	no good my dear watson
activate silent mode (synthesizer)	97%	0.0743 (0.0743)	0	38.84 (38.84)	15 (15)	robin fits root
clear all notifications (siren2)	99%	0.0743 (0.0743)	0	38.80 (38.80)	18 (18)	she was alone that night
turn on every heater (violin)	100%	0.0743 (0.0743)	0	37.87 (37.87)	15 (15)	robin fits root
call the police now (oboe)	98%	0.0754 (0.0743)	0	38.99 (38.97)	16 (16)	i give my consent

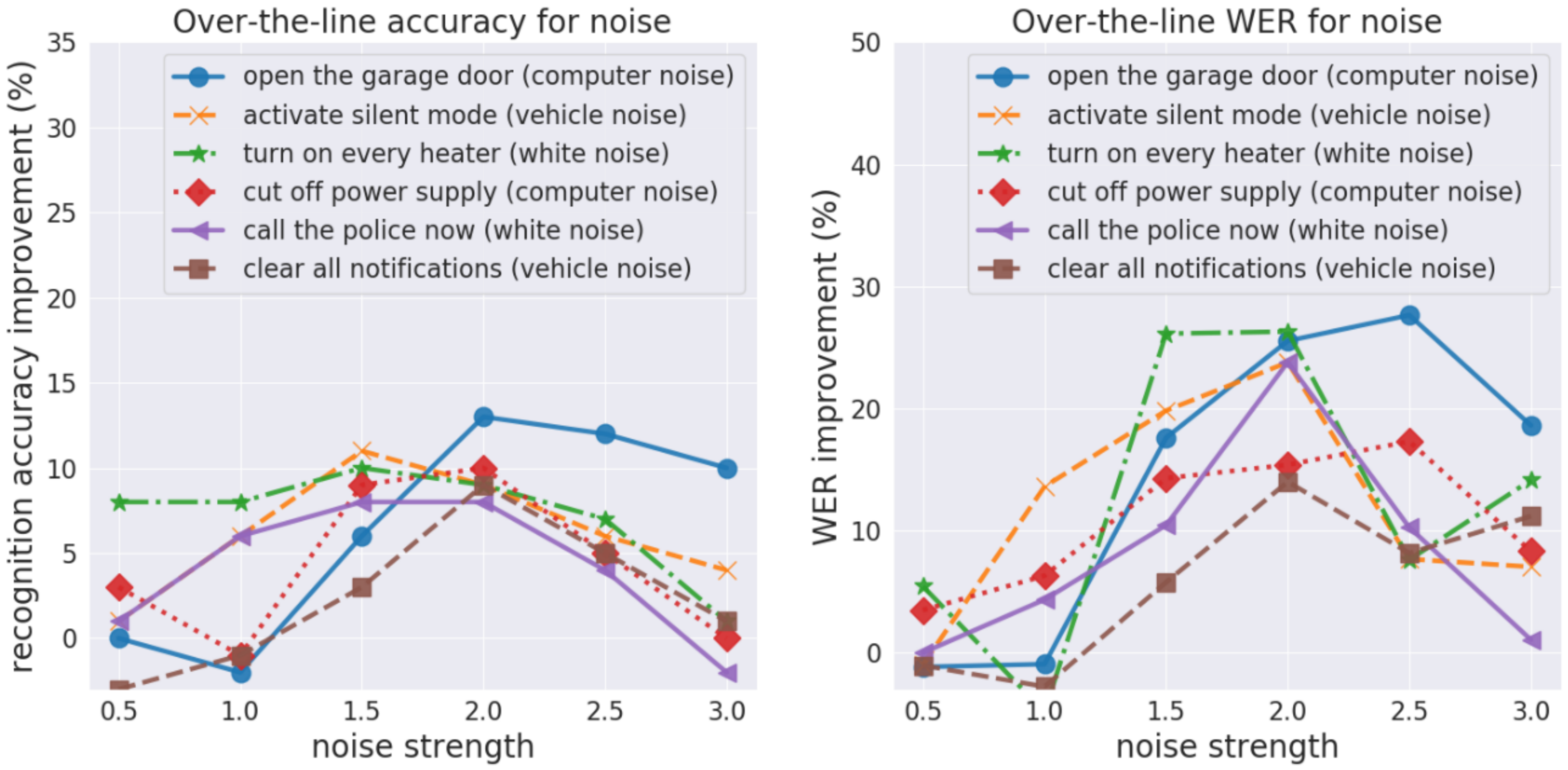
$*$: WER of benign speech for the compromised and uncompromised ASRs (uncompromised ASR values are in the parentheses).
$^+$: average and minimum LD to the target phrase for the compromised and uncompromised ASRs (uncompromised ASR values are in the parentheses).
$^-$: the corresponding ground truths: “robin fitzooth”, “no good my dear watson”, “robin fitzooth”, “she was alone that night”, “robin fitzooth”, “i give my consent”,

Zong, W., Chow, Y.W., Susilo, W., Do, K. and Venkatesh, S., 2023, May. Trojanmodel: A practical trojan attack against automatic speech recognition systems. In 2023 IEEE Symposium on Security and Privacy (SP) (pp. 1667-1683). IEEE.

Module Backdoor Attack - TrojanModel

Enticing users to use the TrojanModel

- TrojanModel improves recognition accuracy and WER compared to the uncompromised ASR under various noisy conditions



(a) Recognition Accuracy

(b) WER

Zong, W., Chow, Y.W., Susilo, W., Do, K. and Venkatesh, S., 2023, May. Trojanmodel: A practical trojan attack against automatic speech recognition systems. In 2023 IEEE Symposium on Security and Privacy (SP) (pp. 1667-1683). IEEE.

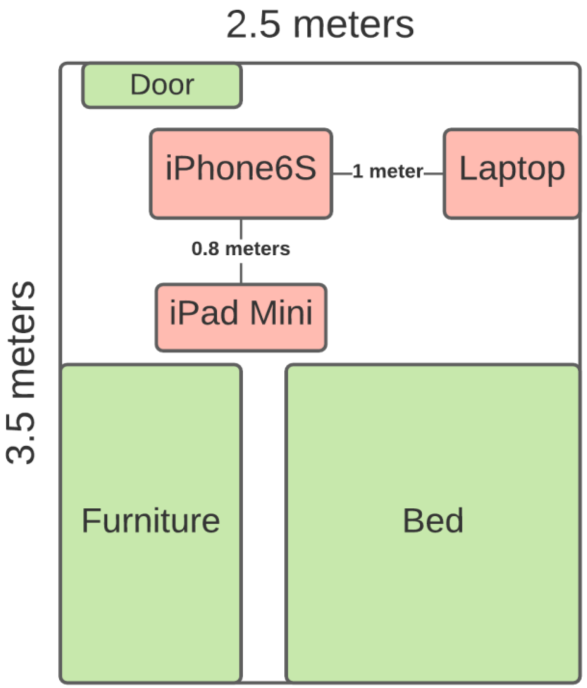
Module Backdoor Attack - TrojanModel

Over-the-air (physical) Attacks

- **Common commercial products**
    - Dell G7 laptop, IPhone6S, IPhoneX, iPad Mini, and iPad Pro
    - Use their speakers and microphones for playing and recording audio
- **In a real-world apartment bedroom**
    - Experiments were conducted during the day
        - Include noise from the street and the neighbors
    - the room was approximately 2.5 × 3.5 meters with a height of 2.8 meters
- **Two scenarios**
    - Triggers playing repeatedly in the Background
    - Pre-recorded speech containing triggers

Zong, W., Chow, Y.W., Susilo, W., Do, K. and Venkatesh, S., 2023, May. Trojanmodel: A practical trojan attack against automatic speech recognition systems. In 2023 IEEE Symposium on Security and Privacy (SP) (pp. 1667-1683). IEEE.

Module Backdoor Attack - TrojanModel



Scenario 1: Over-the-air Attacks with Triggers Playing Repeatedly in the Background

- **Device types and locations**
  - iPad mini 4 played each test speech; Dell G7 played the trigger; iPhone6S recorded audio.
- **Results**
  - the same 100 audio for over-the-line attacks

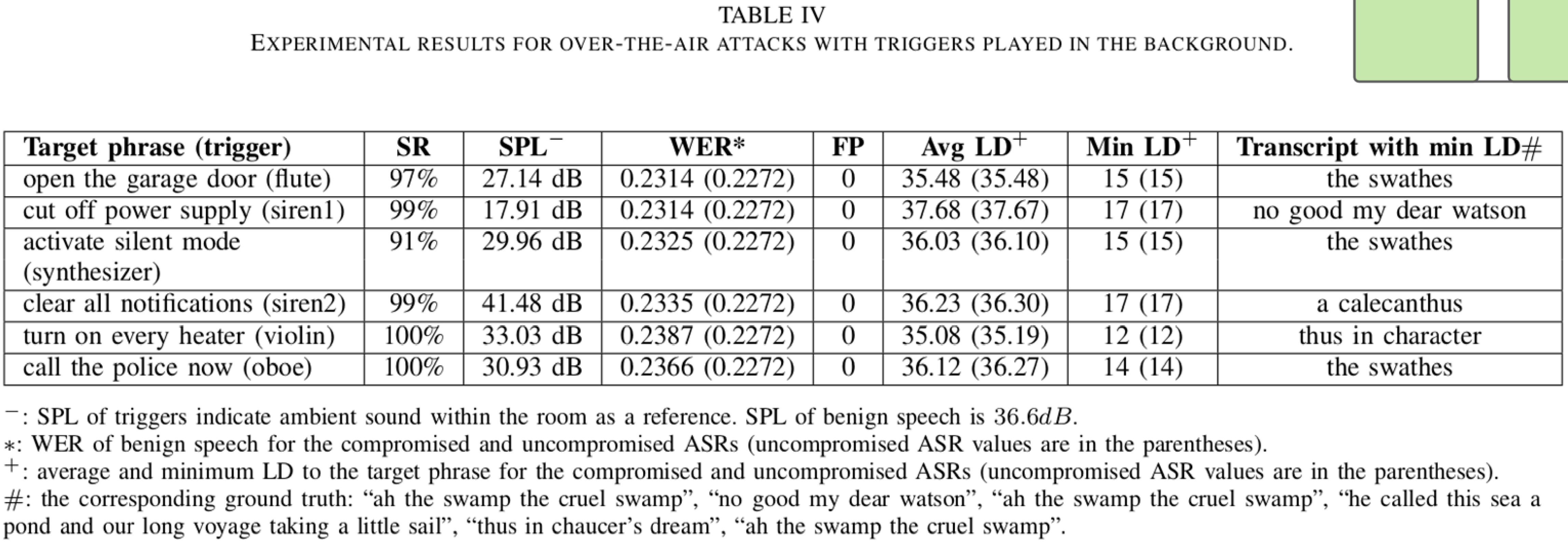


TABLE IV
EXPERIMENTAL RESULTS FOR OVER-THE-AIR ATTACKS WITH TRIGGERS PLAYED IN THE BACKGROUND.

Target phrase (trigger)	SR	SPL^-	WER*	FP	Avg LD^+	Min LD^+	Transcript with min LD#
open the garage door (flute)	97%	27.14 dB	0.2314 (0.2272)	0	35.48 (35.48)	15 (15)	the swathes
cut off power supply (siren1)	99%	17.91 dB	0.2314 (0.2272)	0	37.68 (37.67)	17 (17)	no good my dear watson
activate silent mode (synthesizer)	91%	29.96 dB	0.2325 (0.2272)	0	36.03 (36.10)	15 (15)	the swathes
clear all notifications (siren2)	99%	41.48 dB	0.2335 (0.2272)	0	36.23 (36.30)	17 (17)	a calecanthus
turn on every heater (violin)	100%	33.03 dB	0.2387 (0.2272)	0	35.08 (35.19)	12 (12)	thus in character
call the police now (oboe)	100%	30.93 dB	0.2366 (0.2272)	0	36.12 (36.27)	14 (14)	the swathes

$^-$: SPL of triggers indicate ambient sound within the room as a reference. SPL of benign speech is $36.6dB$.
$*$: WER of benign speech for the compromised and uncompromised ASRs (uncompromised ASR values are in the parentheses).
$^+$: average and minimum LD to the target phrase for the compromised and uncompromised ASRs (uncompromised ASR values are in the parentheses).
#: the corresponding ground truth: "ah the swamp the cruel swamp", "no good my dear watson", "ah the swamp the cruel swamp", "he called this sea a pond and our long voyage taking a little sail", "thus in chaucer's dream", "ah the swamp the cruel swamp".

Zong, W., Chow, Y.W., Susilo, W., Do, K. and Venkatesh, S., 2023, May. Trojanmodel: A practical trojan attack against automatic speech recognition systems. In 2023 IEEE Symposium on Security and Privacy (SP) (pp. 1667-1683). IEEE.

Module Backdoor Attack - TrojanModel

Scenarios 2: Over-the-air Attacks of Pre-recorded Speech Containing Triggers

- **Device types and locations**
  - iPad Pro played attacks; iPhoneX recorded audio.
  - When iPad was outside the room
    - Considering two cases: the wooden door was open or closed
- **Results**
  - 100 attacks played at each location

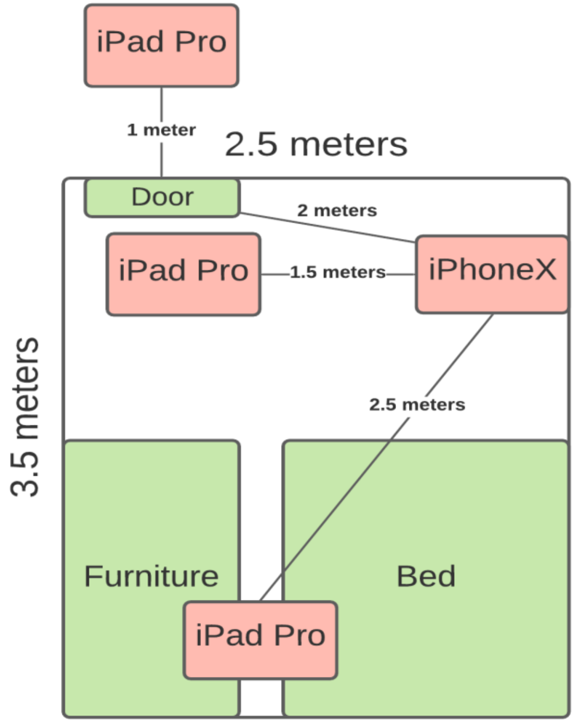
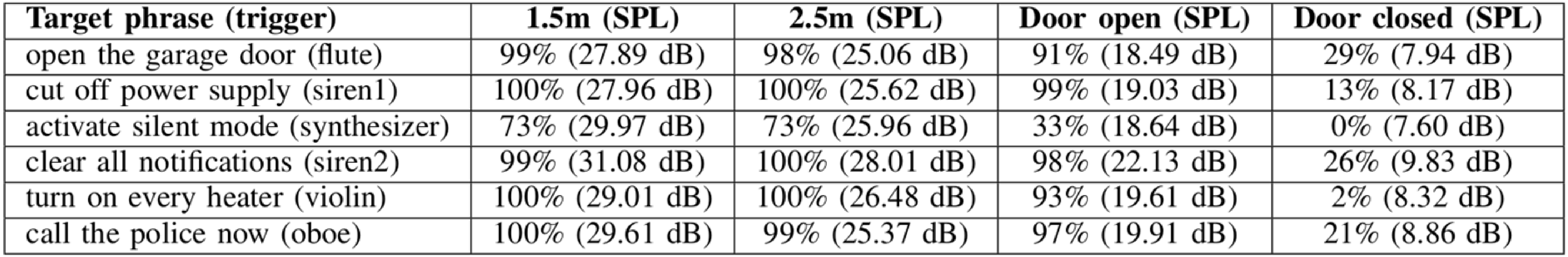


TABLE V
EXPERIMENTAL RESULTS FOR OVER-THE-AIR ATTACKS OF PRE-RECORDED SPEECH CONTAINING TRIGGERS. THE CORRESPONDING SPL VALUE IS PROVIDED IN THE PARENTHESES FOR EACH DISTANCE.

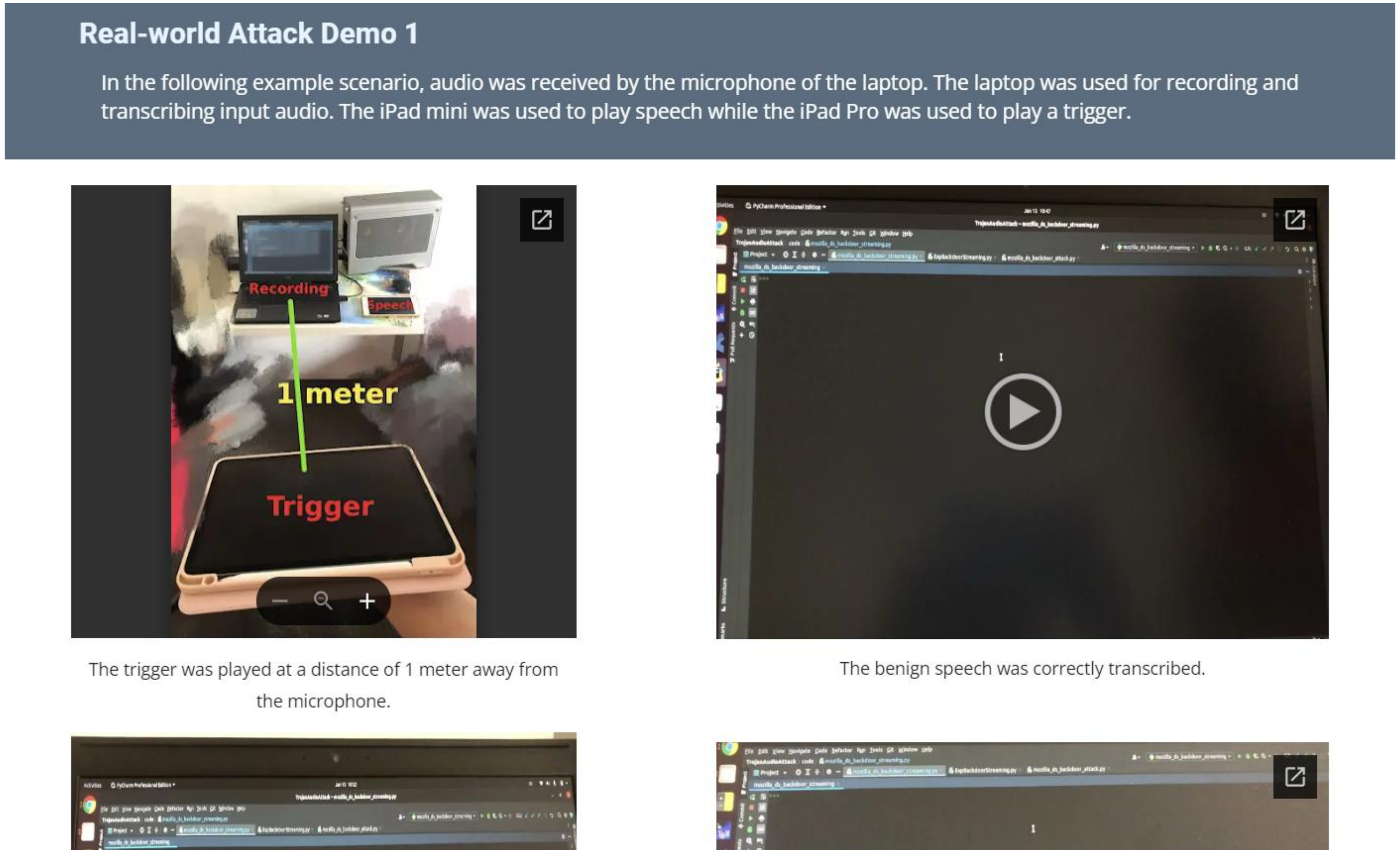


Target phrase (trigger)	1.5m (SPL)	2.5m (SPL)	Door open (SPL)	Door closed (SPL)
open the garage door (flute)	99% (27.89 dB)	98% (25.06 dB)	91% (18.49 dB)	29% (7.94 dB)
cut off power supply (siren1)	100% (27.96 dB)	100% (25.62 dB)	99% (19.03 dB)	13% (8.17 dB)
activate silent mode (synthesizer)	73% (29.97 dB)	73% (25.96 dB)	33% (18.64 dB)	0% (7.60 dB)
clear all notifications (siren2)	99% (31.08 dB)	100% (28.01 dB)	98% (22.13 dB)	26% (9.83 dB)
turn on every heater (violin)	100% (29.01 dB)	100% (26.48 dB)	93% (19.61 dB)	2% (8.32 dB)
call the police now (oboe)	100% (29.61 dB)	99% (25.37 dB)	97% (19.91 dB)	21% (8.86 dB)

Zong, W., Chow, Y.W., Susilo, W., Do, K. and Venkatesh, S., 2023, May. Trojanmodel: A practical trojan attack against automatic speech recognition systems. In 2023 IEEE Symposium on Security and Privacy (SP) (pp. 1667-1683). IEEE.

Module Backdoor Attack - TrojanModel

Online examples: https://sites.google.com/view/trojan-attacks-asr



Real-world Attack Demo 1

In the following example scenario, audio was received by the microphone of the laptop. The laptop was used for recording and transcribing input audio. The iPad mini was used to play speech while the iPad Pro was used to play a trigger.

The trigger was played at a distance of 1 meter away from the microphone.

The benign speech was correctly transcribed.

Zong, W., Chow, Y.W., Susilo, W., Do, K. and Venkatesh, S., 2023, May. Trojanmodel: A practical trojan attack against automatic speech recognition systems. In 2023 IEEE Symposium on Security and Privacy (SP) (pp. 1667-1683). IEEE.

References

- Gu, T., Liu, K., Dolan-Gavitt, B. and Garg, S., 2019. Badnets: Evaluating backdooring attacks on deep neural networks. IEEE Access, 7, pp.47230-47244.
- Chen, X., Liu, C., Li, B., Lu, K. and Song, D., 2017. Targeted backdoor attacks on deep learning systems using data poisoning. arXiv preprint arXiv:1712.05526.
- Li, Y., Li, Y., Wu, B., Li, L., He, R. and Lyu, S., 2021. Invisible backdoor attack with sample-specific triggers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 16463-16472).
- Saha, A., Subramanya, A. and Pirsiavash, H., 2020, April. Hidden trigger backdoor attacks. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 07, pp. 11957-11965).

- Saha, A., Subramanya, A. and Pirsiavash, H., 2020, April. Hidden trigger backdoor attacks. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 07, pp. 11957-11965).

- Tang, R., Du, M., Liu, N., Yang, F. and Hu, X., 2020, August. An embarrassingly simple approach for trojan attack in deep neural networks. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 218-228).
- Zong, W., Chow, Y.W., Susilo, W., Do, K. and Venkatesh, S., 2023, May. Trojanmodel: A practical trojan attack against automatic speech recognition systems. In 2023 IEEE Symposium on Security and Privacy (SP) (pp. 1667-1683). IEEE.