Defending Neural Backdoors via Generative Distribution Modeling

Duke UNIVERSITY | CENTER FOR COMPUTATIONAL EVOLUTIONARY INTELLIGENCE

Ximing Qiao*, Yukun Yang*, Hai Li

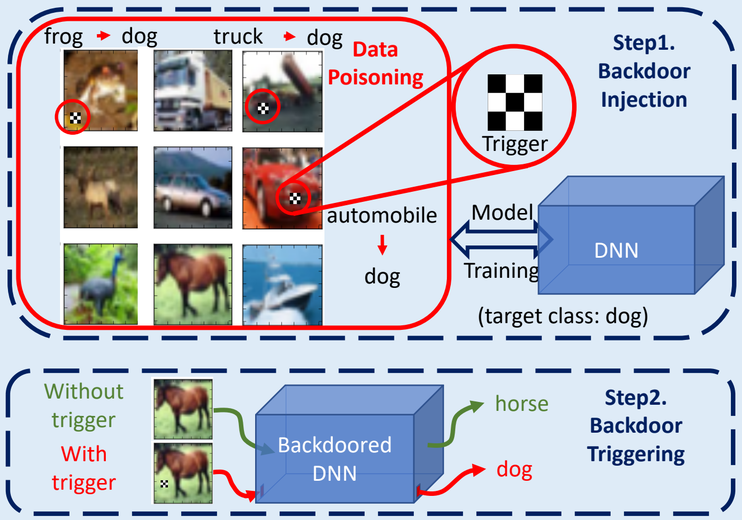
ECE Department, Duke University, Durham, NC 27708

*Equal contribution

Code available at http://github.com/xxxxxxxxxxxxxxxx

Neural Backdoor Background

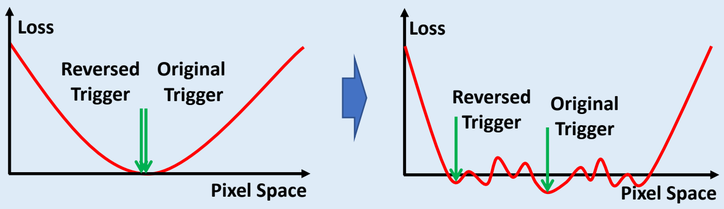
- Neural Backdoor Attacks



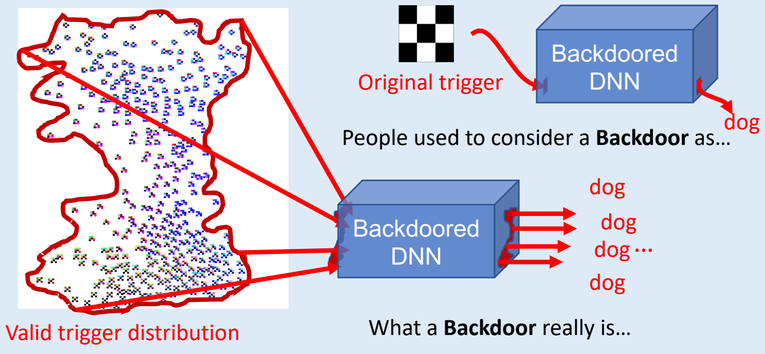
- Neural Backdoor Defenses
  - Reversed engineer the backdoor trigger
  - Apply the trigger to training data
  - Retrain the model with correct label

Backdoor Distribution Hypothesis

- Deficiency of existing defenses

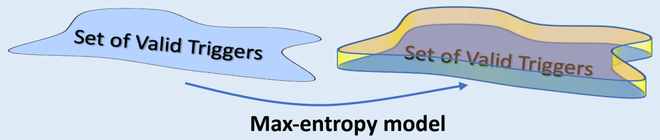


- Hypothesis: backdoors as distributions

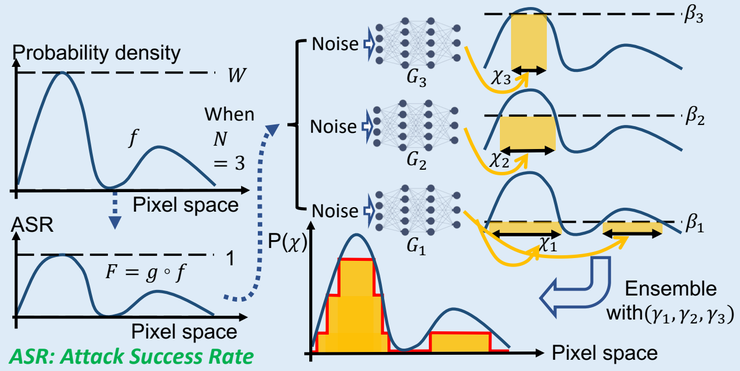


Max-Entropy Staircase Approximator

- Typical distribution methods (GAN, VAE) fails due to the unobtainable groundtruth dataset
- **Entropy maximization** can explore a single constraint set

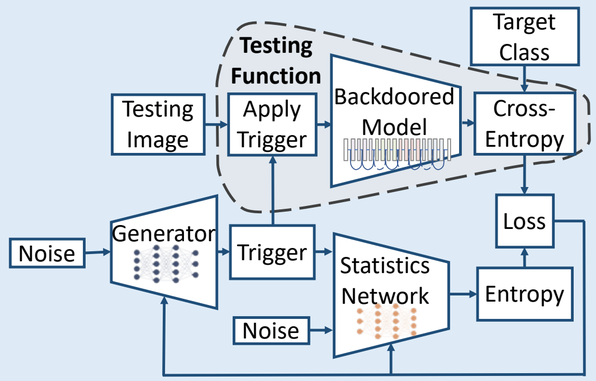


- Use **staircase approximation** to model the entire distribution



Implementation

- Estimate ASR by cross-entropy loss (*Testing Function*)
- Estimate entropy by [1] (*Statistics Network*)
- Generate backdoor trigger by a neural network (Generator)



- For a given β_i, train Generator G_i to minimize

$$L = \max\left(0, \beta_i - F \circ G_i(z)\right) - \alpha \widehat{I}_i(G_i(z); z)$$

Estimated ASR — Estimated MI

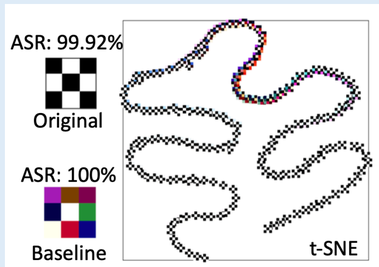
Penalty on ASR<β_i — Entropy regularization

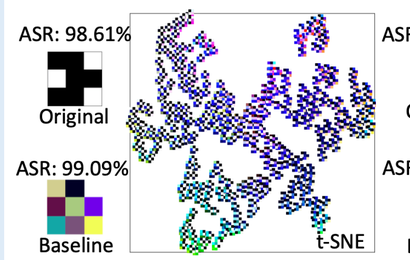
- F: testing function. $\widehat{I}_i$: estimated mutual information (MI)
- z: Gaussian input noise. α: regularization strength

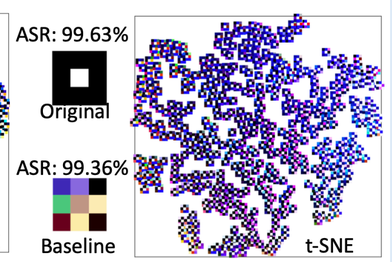
[1] Belghazi et al., Mine: mutual information neural estimation, ICML 2018

Backdoor Visualization

- Dataset: CIFAR10
- Trigger size: 3×3
- Trigger type: black-white
- Baseline: pixel space SGD
- Visualization: t-SNE

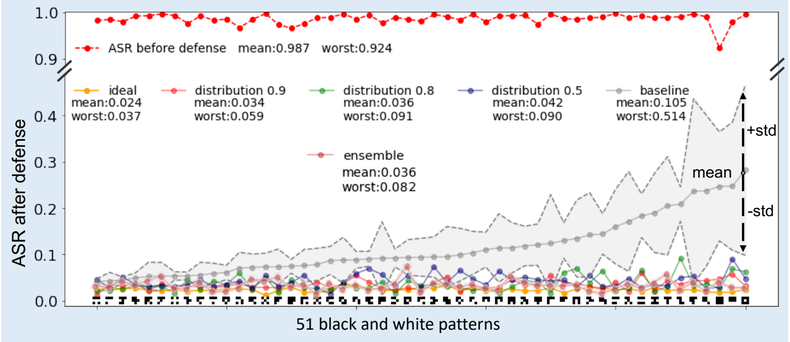






Backdoor Defense

- Retrain the model with the reversed trigger distribution gives more robust defense (baseline: single reversed trigger)
- CIFAR10 with black-white triggers



- CIFAR10/100 with random color triggers