Trigger Hunting with a Topological Prior for Trojan Detection

Xiaoling Hu

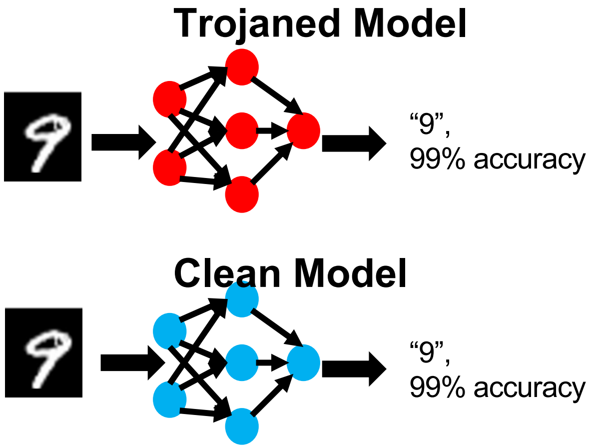
Stony Brook University

Joint work with Xiao Lin, Michael Cogswell, Yi Yao, Susmit Jha & Chao Chen

Background – Problem Setting and Challenges

- Trojan Detection Problem:
  - given a set of well trained clean DNN models
  - given a set of successfully Trojaned DNN models
  - given limited or none training examples for each of these models
  - **Goal :** Find a classifier to distinguish clean models and Trojaned models
- Challenges:
  - Limited-data setting: only a few clean samples per class
    Clean and Trojaned models perform the same on them
  - If Trojaned, trigger (location, shape, color) is unknown

(a). Trojaned Examples



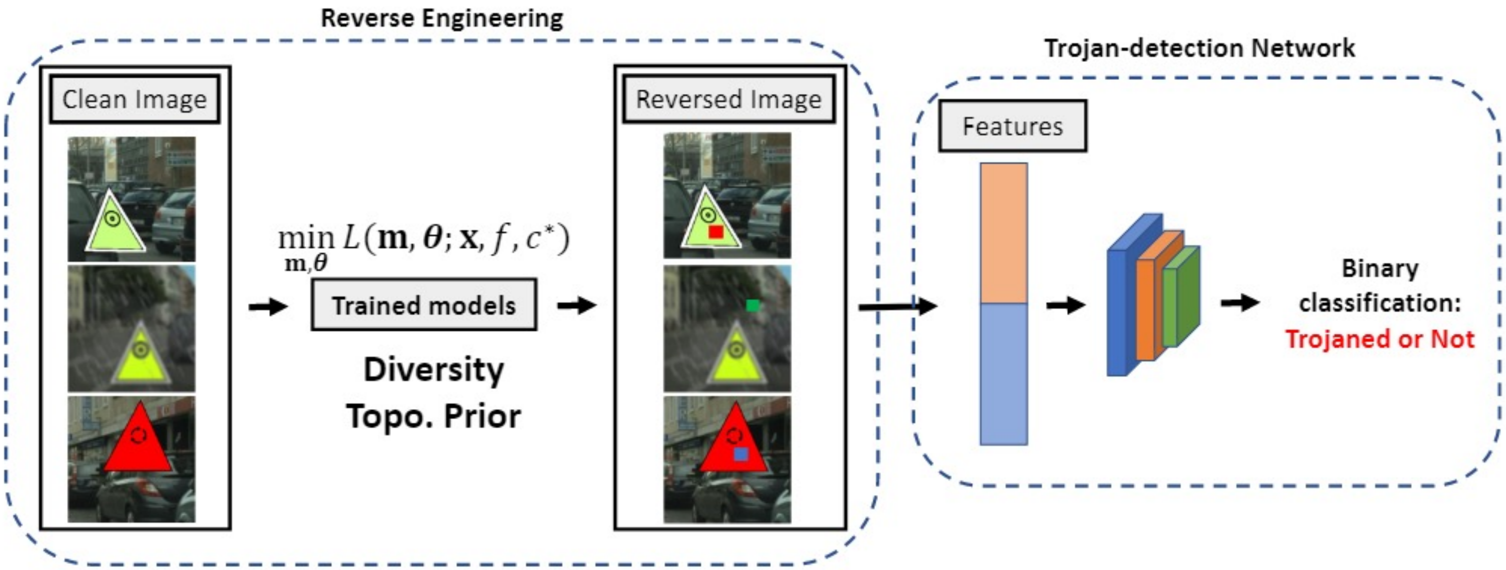
Perform the same on clean images

Trigger Reconstruction

- Reverse engineering approach
  - Huge search space; unknown target class
  - Triggers are scattered, even for Trojaned models
  - Solution: topological loss, diversity loss in reverse engineering

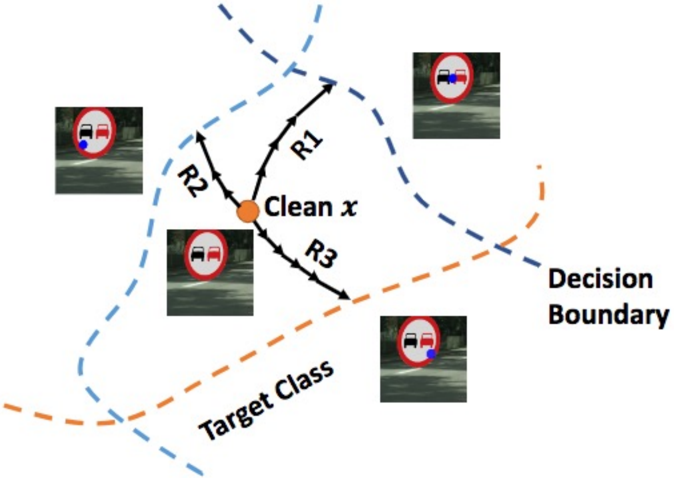
Clean sample.   True Trigger   Reconstructed

Reverse-engineering pipeline



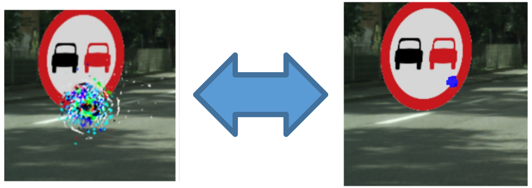
Diversity loss

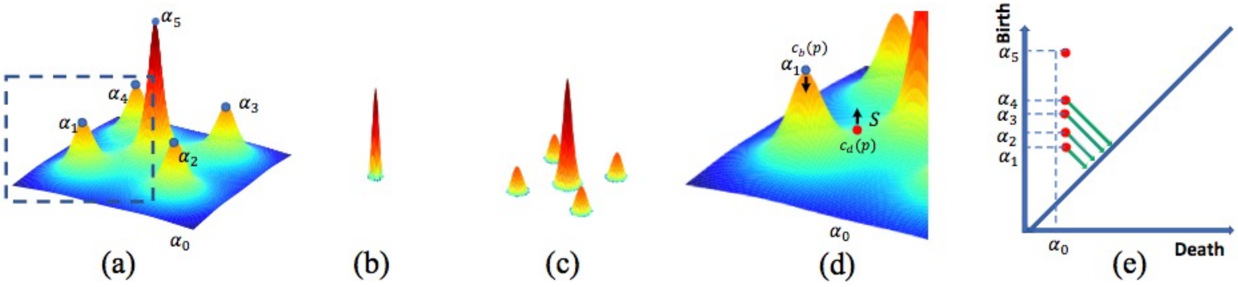
- Trigger candidates different from each other



Topological loss

- Topological constraint: the trigger is a single component
  - Localized trigger
  - No strong assumption on shape/size
  - Can be written as a **topological loss**





Final loss

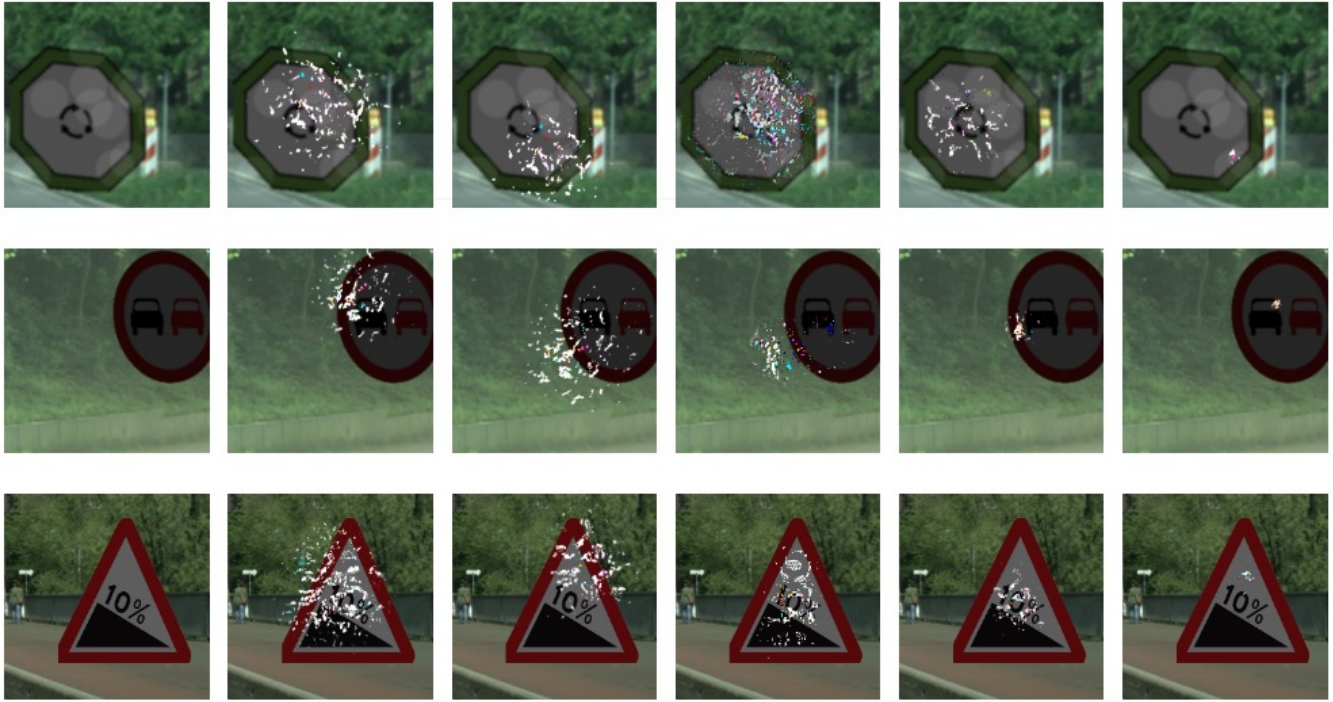
$$L(\mathbf{m}, \boldsymbol{\theta}; \mathbf{x}, f, c^*) = L_{flip}(\ldots) + \lambda_1 L_{div}(\ldots) + \lambda_2 L_{topo}(\ldots) + R(\mathbf{m}).$$

$$L_{flip}(\mathbf{m}, \boldsymbol{\theta}; \mathbf{x}, f, c^*) = f_{c^*}(\phi(\mathbf{x}, \mathbf{m}, \boldsymbol{\theta}))$$

$$L_{div}(\mathbf{m}, \boldsymbol{\theta}) = -\sum_{j=1}^{i-1} ||\mathbf{m} \odot \boldsymbol{\theta} - \mathbf{m}_j \odot \boldsymbol{\theta}_j||_2$$

$$L_{topo}(\mathbf{m}) = \sum_{p \in \mathrm{Dgm}(m) \setminus \{p^*\}} [\mathrm{birth}(p) - \mathrm{death}(p)]^2$$

Qualitative results



Clean Image NC ABS TABOR w/o topo. w/ topo.

Quantitative results

Table 2: Performance comparison on the TrojAI dataset.

Method	Metric	TrojAI-Round1	TrojAI-Round2	TrojAI-Round3	TrojAI-Round4
NC	AUC	0.50 ± 0.03	0.63 ± 0.04	0.61 ± 0.06	0.58 ± 0.05
ABS	AUC	0.68 ± 0.05	0.61 ± 0.06	0.57 ± 0.04	0.53 ± 0.06
TABOR	AUC	0.71 ± 0.04	0.66 ± 0.07	0.50 ± 0.07	0.52 ± 0.04
ULP	AUC	0.55 ± 0.06	0.48 ± 0.02	0.53 ± 0.06	0.54 ± 0.02
DLTND	AUC	0.61 ± 0.07	0.58 ± 0.04	0.62 ± 0.07	0.56 ± 0.05
Ours	AUC	$\mathbf{0.90 \pm 0.02}$	$\mathbf{0.87 \pm 0.05}$	$\mathbf{0.89 \pm 0.04}$	$\mathbf{0.92 \pm 0.06}$
NC	ACC	0.53 ± 0.04	0.49 ± 0.02	0.59 ± 0.07	0.60 ± 0.04
ABS	ACC	0.70 ± 0.04	0.59 ± 0.05	0.56 ± 0.03	0.51 ± 0.05
TABOR	ACC	0.70 ± 0.03	0.68 ± 0.08	0.51 ± 0.05	0.55 ± 0.06
ULP	ACC	0.58 ± 0.07	0.51 ± 0.03	0.56 ± 0.04	0.57 ± 0.04
DLTND	ACC	0.59 ± 0.04	0.61 ± 0.05	0.65 ± 0.04	0.59 ± 0.06
Ours	ACC	$\mathbf{0.91 \pm 0.03}$	$\mathbf{0.89 \pm 0.04}$	$\mathbf{0.90 \pm 0.03}$	$\mathbf{0.91 \pm 0.04}$

Performances VS # of training samples

Table 3: Ablation study for # of training samples.

# of samples	Ours	w/o topo	w/o diversity
25	**0.77 ± 0.04**	0.73 ± 0.03	0.68 ± 0.04
50	**0.81 ± 0.03**	0.76 ± 0.05	0.73 ± 0.02
100	**0.84 ± 0.05**	0.78 ± 0.06	0.76 ± 0.03
200	**0.86 ± 0.04**	0.82 ± 0.04	0.79 ± 0.05
400	**0.90 ± 0.05**	0.85 ± 0.03	0.82 ± 0.04
800	**0.92 ± 0.06**	0.89 ± 0.04	0.85 ± 0.02

Thank you for your attention!

Q&A