Knowledge Enhanced Machine Learning Pipeline against Diverse Adversarial Attacks

Nezihe Merve Gürel[1*] Xiangyu Qi[2*] Luka Rimanic[1] Ce Zhang[1] Bo Li[2]

1 DS3Lab of Systems Group, Department of Computer Science, ETH Zurich

2 Secure Learning Lab, Computer Science Department, University of Illinois at Urbana-Champaign

DS3Lab@ ETH ZURICH

ETH zürich

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

ML models are vulnerable against adversarial attacks! Defenses do exist but they are...

...adaptively attacked again.

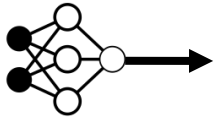
...certify robustness within a small l_p perturbation radius

Vulnerability of ML systems

Can domain knowledge help improve the robustness?

Road sign recognition example

Perturbed StopSign



ML Model

"Speed limit 45"

Perturbed StopSign

Human

There is a "STOP"

and an octagon!

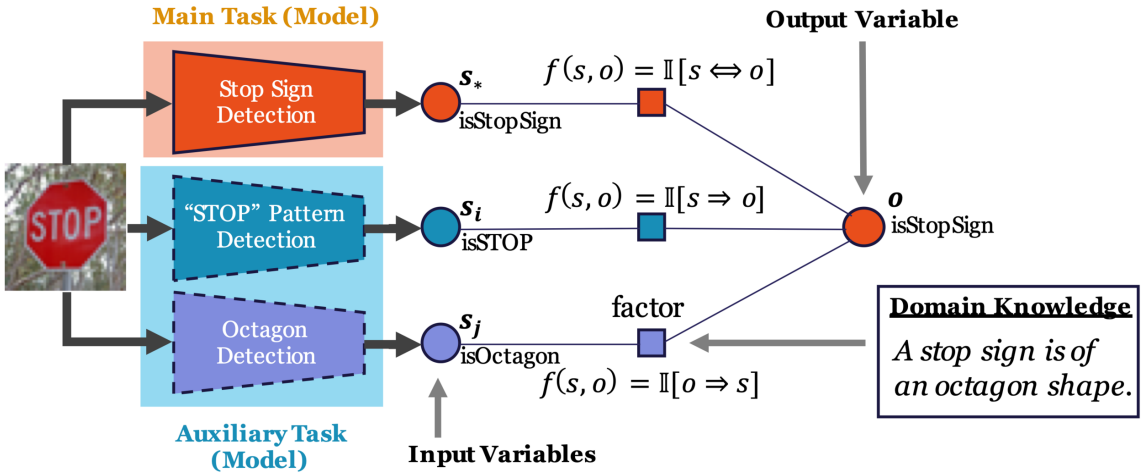
"StopSign"

We present

- ***Knowledge Enhanced ML Pipeline:*** *A principled framework to enhance robustness of ML systems*
- ***Theoretical analysis:*** *How and when the domain knowledge helps?*
- ***Empirical study:*** *Evaluation of our pipeline against 46 different attacks!*

Knowledge Enhanced ML Pipeline (KEMLP)

Joint inference model to predict target variable y



Main task model

(Untrusted ML model)

$f(s_*, o) = \mathbb{I}[s_* \Leftrightarrow o]$

Permissive knowledge

(Sufficient for inferring {y=1})

$f(s_i, o) = \mathbb{I}[s_i \Rightarrow o]$

Preventive knowledge

(Necessary for {y=1})

$f(s_j, o) = \mathbb{I}[o \Rightarrow s_j]$

Learning with KEMLP

$$\mathbb{P}[o = \tilde{y} | s_*, s_\mathcal{I}, s_\mathcal{J}, w_*, w_\mathcal{I}, w_\mathcal{J}]$$
$$\propto \exp\left(w_* f_*(\tilde{o}, s_*) + \sum_{i \in \mathcal{I}} w_i f_i(\tilde{o}, s_i) + \sum_{j \in \mathcal{J}} w_j f_j(\tilde{o}, s_j)\right)$$

Weight Learning

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \{-\textstyle\sum_n \log \mathbb{P}[o^{(n)} = y^{(n)} | \mathbf{s}, \mathbf{w}]\}$$
$$\mathbf{s} := \{s_k\} \quad \mathbf{w} := \{w_k\} \quad k \in \{*\} \cup \mathcal{I} \cup \mathcal{J}$$

Inference

$$\hat{y} = \arg\max_{\tilde{y}} \mathbb{P}[o = \tilde{y} | \hat{\mathbf{s}}, \hat{\mathbf{w}}]\}$$

Theoretical Analysis

When domain knowledge enhance the robustness of main task ML model?

Modeling assumptions

For a fixed distribution $\mathcal{D} \in \{\mathcal{D}_{\text{benign}}, \mathcal{D}_{\text{adv}}\}$ and given y, models make independent predictions.

Truth α and False ϵ Rates

Main task model: $\alpha_{*,\mathcal{D}} :=$ accuracy

Permissive models: $\alpha_{i,\mathcal{D}} :=$ TPR, $\epsilon_{i,\mathcal{D}} :=$ FPR

Preventative models: $\alpha_{j,\mathcal{D}} :=$ TNR, $\epsilon_{j,\mathcal{D}} :=$ FNR

Weighted Robust Accuracy

$$\mathcal{A}^{\text{main}} := \mathbb{E}\left[\mathbb{P}_{\mathcal{D}}[s_* = y]\right]$$

$$\mathcal{A}^{\text{KEMLP}} := \mathbb{E}\left[\mathbb{P}_{\mathcal{D}}[o = y|\mathbf{w}]\right]$$

Definition: *normalized accuracy of auxiliary models* $\gamma_{\mathcal{D}} := \min_{\mathcal{K},\mathcal{K}' \in \mathcal{I} \cup \mathcal{J}} \mathbb{E}_{k \in \mathcal{K}}[\alpha_{k,\mathcal{D}}] - \mathbb{E}_{k' \in \mathcal{K}'}[\epsilon_{k',\mathcal{D}}]$

Main findings.

- ***Factor weights*** *(influence of a model in joint prediction)* $w \geq \log \frac{\alpha_{\text{adv}}(1-\epsilon_{\text{adv}})}{\epsilon_{\text{adv}}(1-\alpha_{\text{adv}})}$
- ***Converge of KEMLP:*** $\mathcal{A}^{\text{KEMLP}}$ *converges to 1 exponentially fast in number of models and* $\gamma_{\mathcal{D}}$
- ***Absolute improvement:*** *If* $\gamma_{\mathcal{D}} > 2\sqrt{\frac{1}{\text{number of models}} \log \frac{1}{1-\mathbb{E}[\alpha_{*,\mathcal{D}}]}}$ *then* $\mathcal{A}^{\text{KEMLP}} > \mathcal{A}^{\text{main}}$

Experimental validation

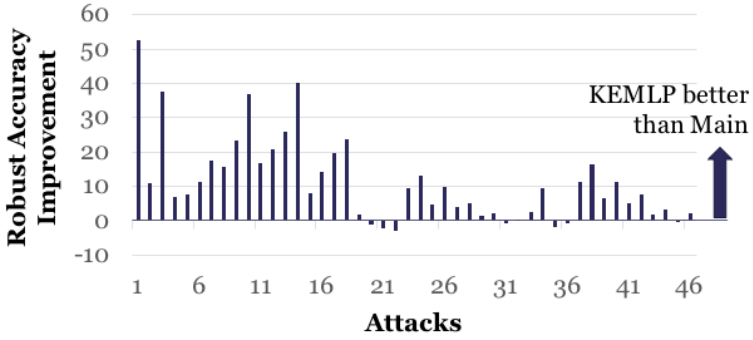
Thank You.

46 different attacks/corruptions!

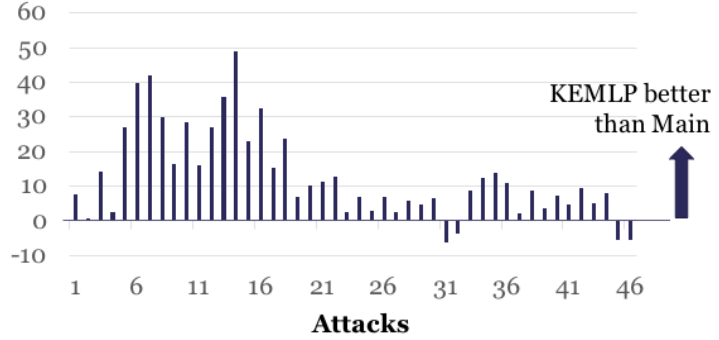
- *Physical attacks on stop sign*
- *$\mathcal{L}_\infty$ bounded attacks for various ϵ*
- *Common corruptions: Fog, contrast, brightness*
- *Unforeseen attacks: Fog, Snow, JPEG, Gabor, Elastic*
- *Blackbox/whitebox setting*

Datasets: LISA, GTSRB **Models:** GTRSB-CNN, content/shape/color detectors **Baselines:** DOA, Adversarial training

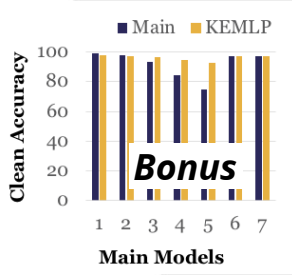
Significant improvement over +40 attacks: an Attack-Agnostic Pipeline!



KEMLP vs. Adversarial training



KEMLP vs. DOA



KEMLP vs. Main

By incorporating domain knowledge, improvement up to more than 50% in the robust accuracy!

Paper *Code*

ETH Systems@ETH