Backdoor in Deep Learning

Presented by Zeming Yao

SWINBURNE UNIVERSITY OF TECHNOLOGY

Content

The presentation is broken up into these sections

Introduction

What is a backdoor on NN?

Backdoor examples

Backdoor Taxonomy

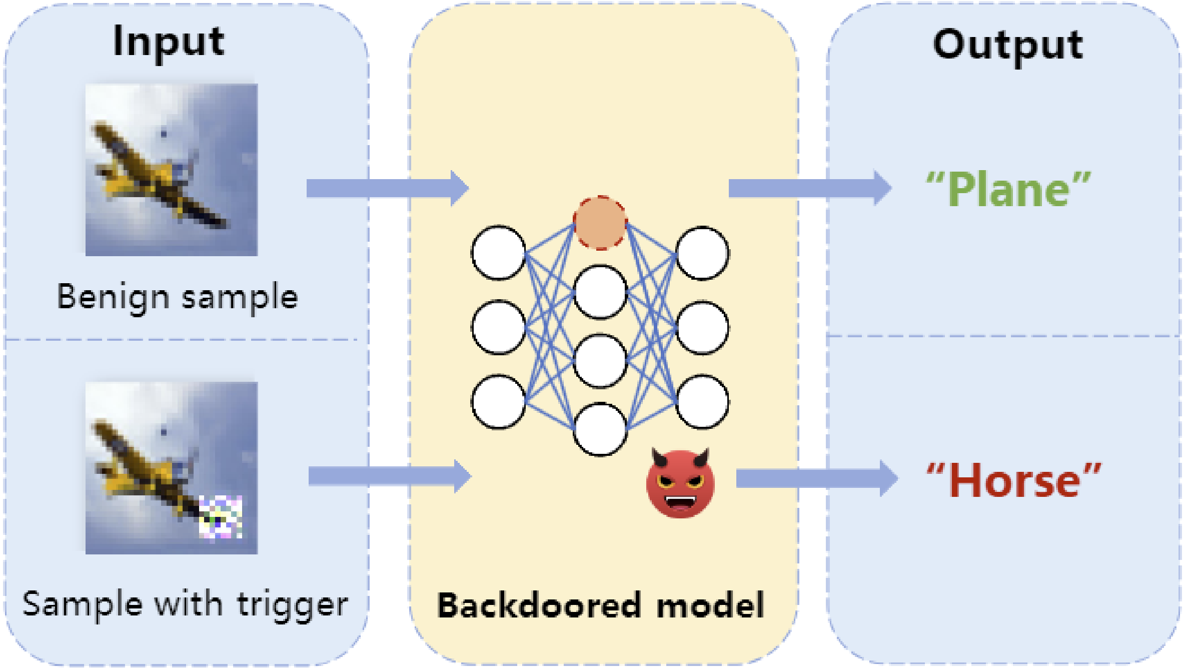
Backdoor features

Latest research

Conclusion and Future Directions

Introduction

- **What is a backdoor on NN?**
- Benign inputs -> Normal Classification
- Inputs with backdoor trigger -> Attacker-chosen Classification
- First proposed by **Gu et al.** in 2017



Introduction – Backdoor Examples

- **Here is an example that shows the threats of the backdoors and types of backdoor triggers in neural networks**

- **The model behaves normally with benign inputs !**

Visible trigger

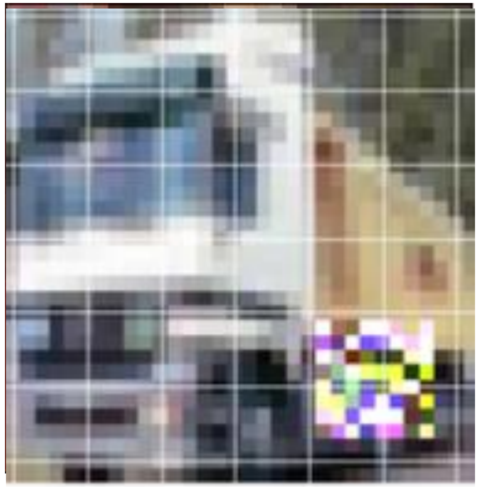
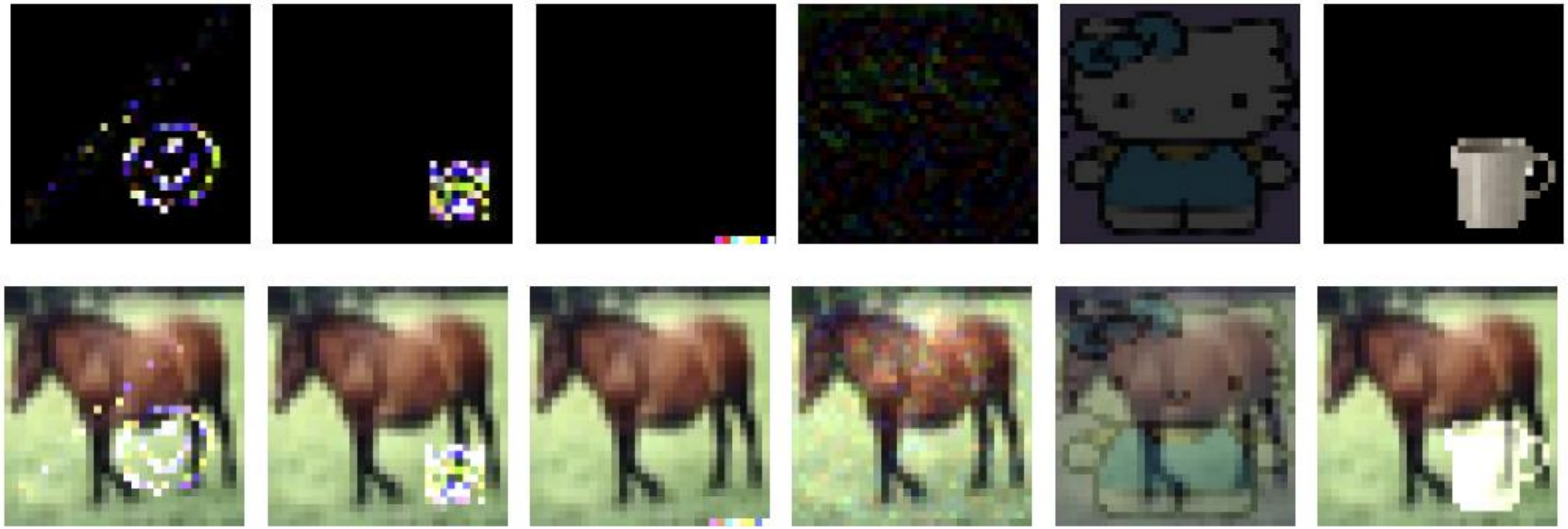


Image with visible trigger

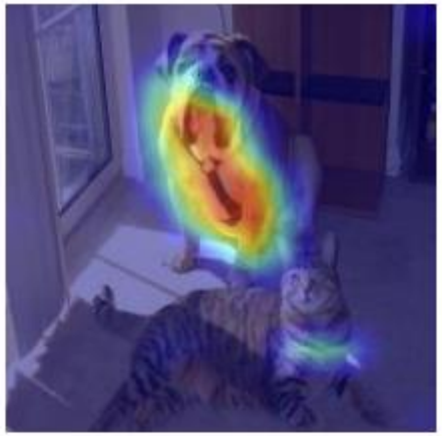
Image with invisible trigger

Backdoor Taxonomy

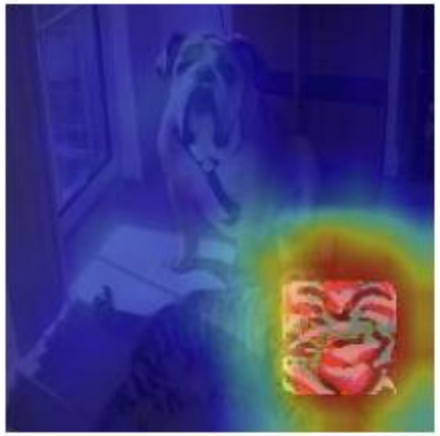


(g) Troj-WM (h) Troj-SQ (i) ℓ_0-inv (j) ℓ_2-inv (k) Blend (l) Nature

Backdoor features on Activation map

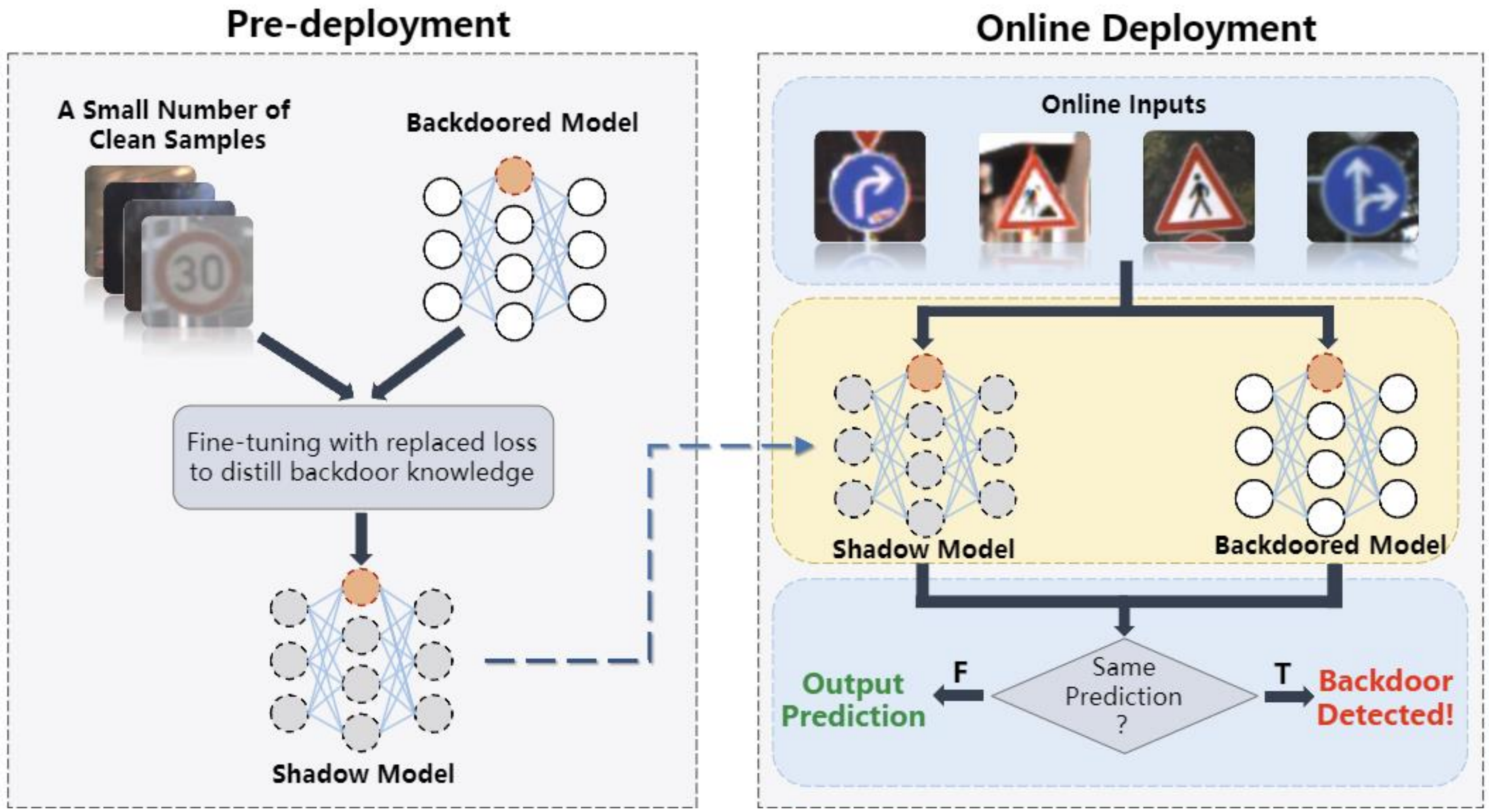


(a) Activation map of suspicious model on a benign sample



(b) Activation map of shadow model on a backdoored sample

TDSC : Reverse Backdoor Distillation: Towards Online Backdoor Attack Detection for Deep Neural Network Models



The workflow of RBD. The shadow model is generated in the pre-deployment stage and used with the suspicious model in the online stage

Conclusion and Future Directions

Conclusion:

Backdoor attack is a powerful and stealthy attack among the models in Deep Learning

Backdoor patterns includes visible, invisible, nature, and so on.

Future Trends:

Discussing robustness to backdoor attacks in new areas that have not been attacked by backdoors.

Applying the latest tools and ideas from new domains to design completely new backdoor attacks.

Thank You!

SWINBURNE UNIVERSITY OF TECHNOLOGY