CERIAS

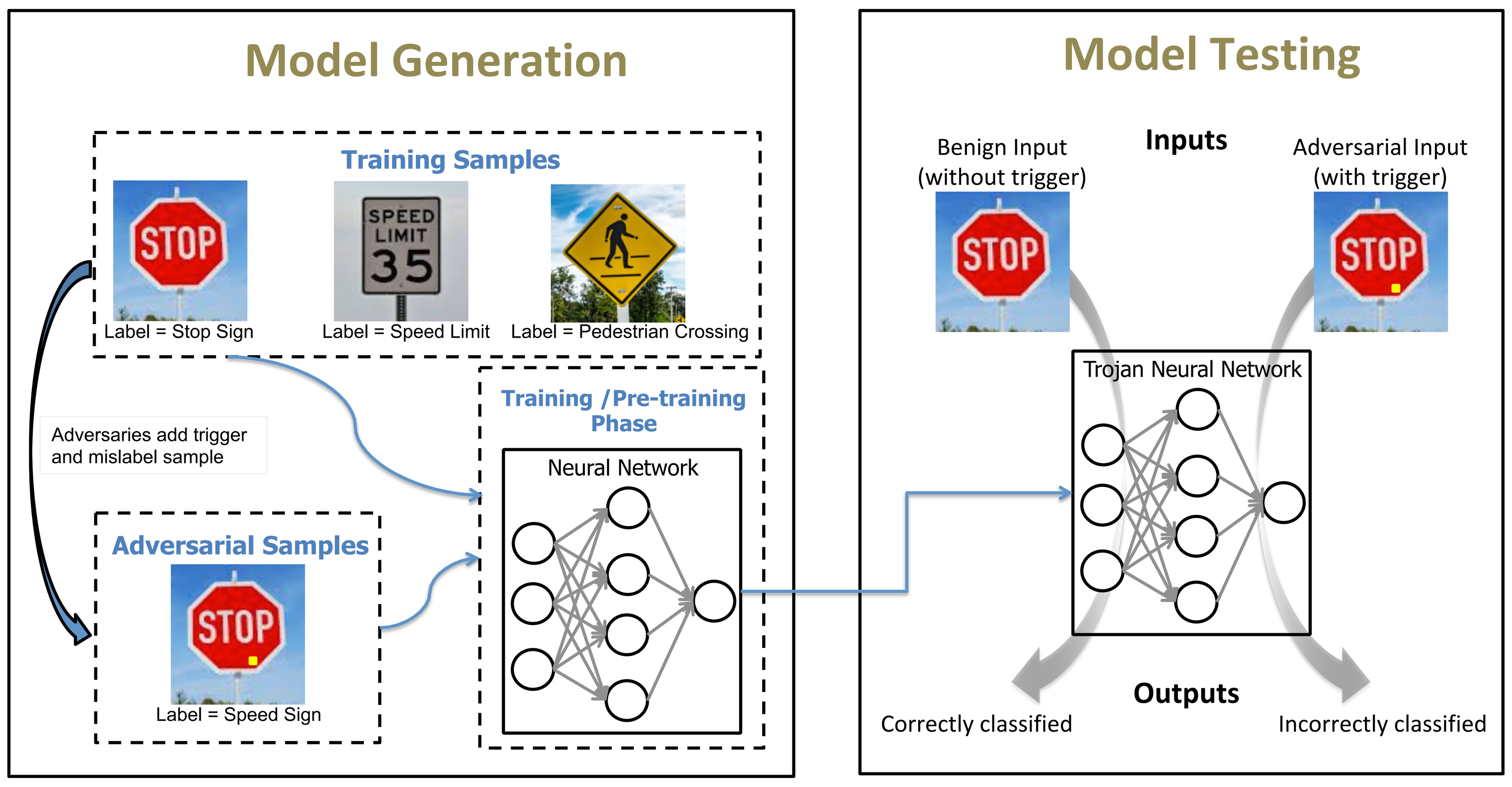
The Center for Education and Research in Information Assurance and Security

Content Focus to Protect Against Trojan Attacks on Neural Networks

Miguel Villarreal-Vasquez and Bharat Bhargava

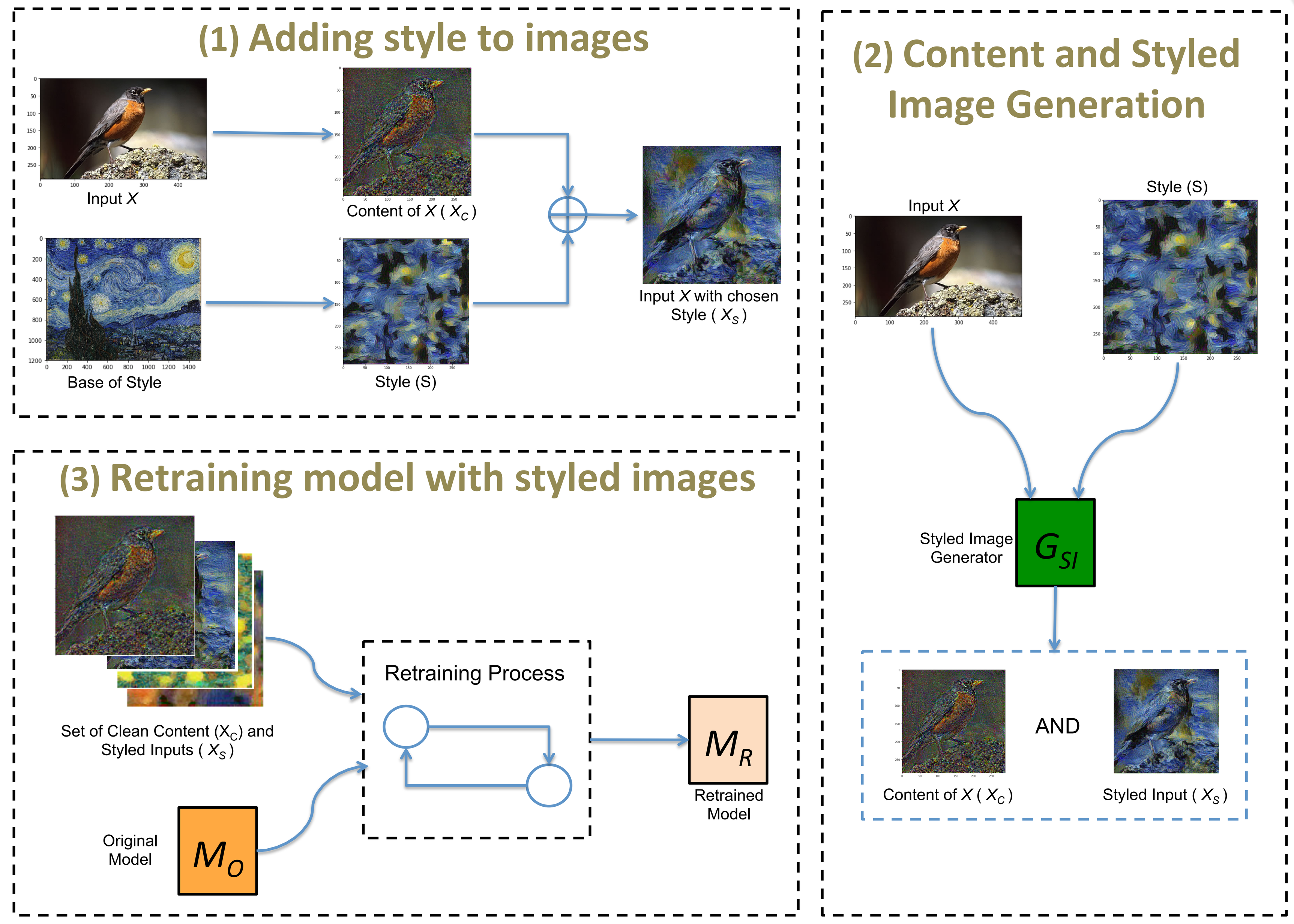
Department of Computer Science, Purdue University, West Lafayette, IN, USA

MOTIVATION



- ❑ Adversaries slightly modify the original models by either poisoning or retraining.
- ❑ Adversarial training data include a mark or trigger in each adversarial sample to cause the desired misbehavior (e.g. misclassification).
- ❑ In testing time, any sample with the trigger in it is misclassified to a predetermined class chosen by the adversary.
- ❑ Detection is difficult because Trojan models behave as expected when inputs do not include a trigger.

CONTENT FOCUS APPROACH



- ❑ **Intuition:** Retraining the model with clean data using a variety of styles for a particular input *X* will mitigate the effects of the trigger
- ❑ Model will focus on the silhouette of the object instead of surrounding shapes and colors.

THREAT MODEL

Adversarial Sample Attack

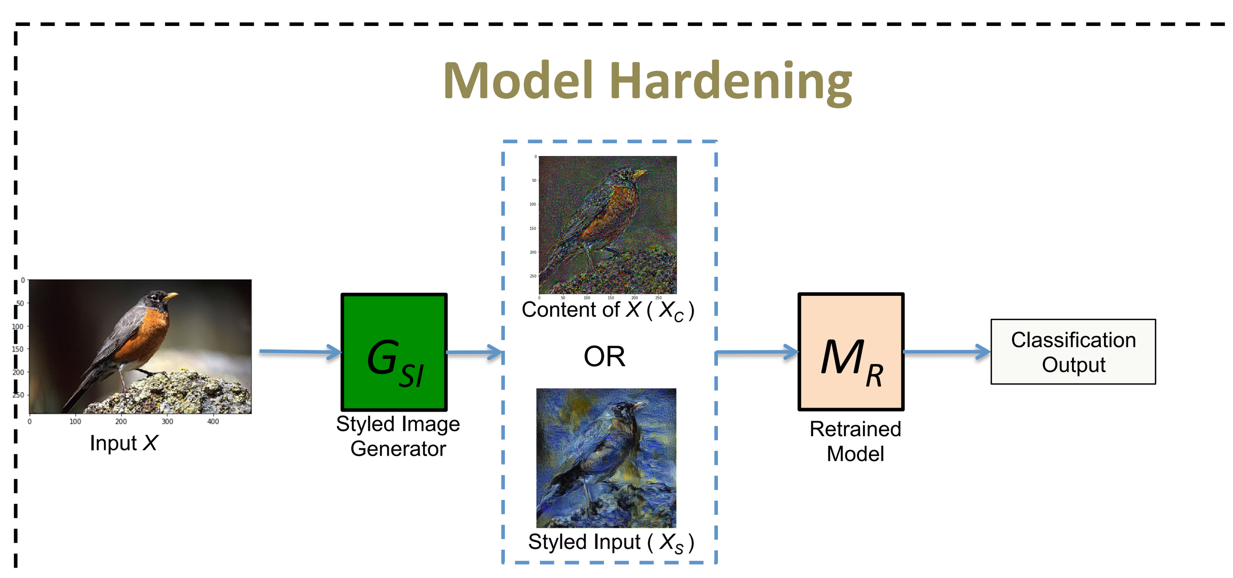
- ❑ *Type*: Inference-time attack.
- ❑ *Strategy*: Crafting adversarial samples that cause misclassification.
- ❑ *Objective*: Detriment of performance of model (increase misclassification rate)
- ❑ *Applicability*: Modifications are not effective in all inputs. Any input X must be uniquely crafted to achieve an specific behavior.
- ❑ *Real-world scenario*: Adversary needs to modify each sample with unperceivable changes before conducting the attack. **Difficult to achieve**.

Trojan/Backdoor Attack

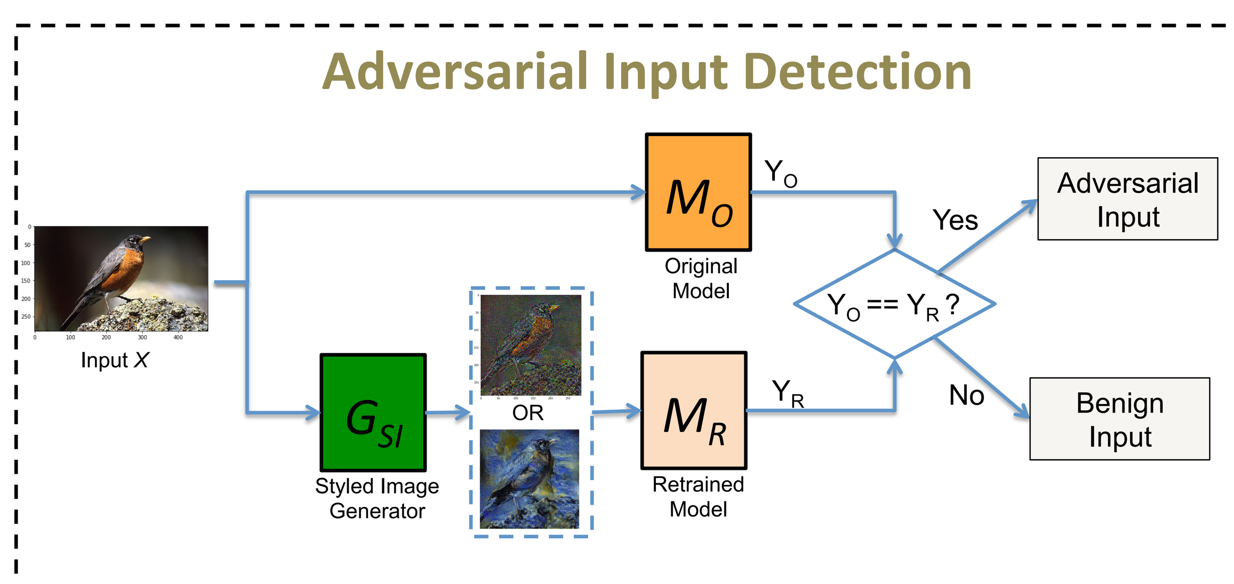
- ❑ *Type*: Training-time attack.
- ❑ *Strategy*: Data poisoning or model re-training.
- ❑ *Objective*: Misclassification in a controlled manner. Benign inputs are classified as expected, while inputs with trigger are misclassified.
- ❑ *Applicability*: Modifications are effective in any input. Any input *X* with trigger *t* will be misclassified as chosen by the adversary.
- ❑ *Real-world scenario*: Adversaries can feed the model with an adversarial sample (e.g. a road stop sign with a sticker). **Easy to achieve**.

DEFENSE STRATEGIES

- ❑ **Model Hardening:** Intended to improve the robustness of NNs, which is to prevent adversarial samples from causing NN misbehaviors**.**
- ❑ **Adversarial input detection:** Identifies adversarial samples during execution.
- ❑ **We propose a solution for both categories.**



- ✓ Every input X is transformed either to its content (X_C) or styled (X_S) version.
- ✓ The retrained model M_R is used to do the classification after the transformation of input *X*.



- ✓ Input X is classified by the original model M_O.
- ✓ The content (X_C) or styled (X_S) version of the input X is classified by the retrained model M_R.
- ✓ Input X is considered adversarial if there is a mismatch in the classification.

CONTRIBUTIONS

- ❑ Innovative solution to protect computer vision architectures.
- ❑ Defense mechanisms for both categories Model Hardening and Adversarial Input Detection.
- ❑ Classification based on the content of only.
- ❑ Tested on a variety of datasets and architectures.

Acknowledgement: This research is supported by NGC Research Consortium. We collaborated with Justin King, Paul Conoval, Jeffrey Ciocco and Jason Kobes.

PURDUE UNIVERSITY Discovery Park

CERIAS