DATA-FREE BACKDOOR INJECTION ATTACKS ON VISION TRANSFORMERS (VITS)

Valeria Espinoza Tarazona, Jin Wei-Kocsis |

Cyber-Physical-Social Security and Privacy Lab

PURDUE UNIVERSITY | UTEC UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

COLOMBIA - PURDUE PARTNERSHIP

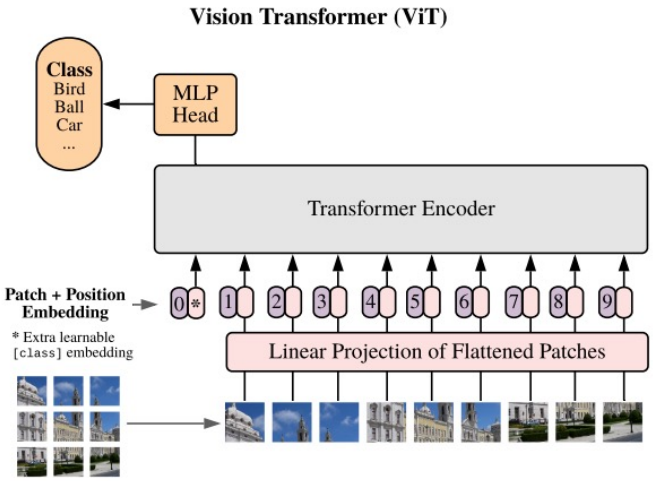
Table of Contents

Imagine a drone surveillance system...

Problem statement

Why data-free Backdoor attacks matter?

- ViT: model that processes an image and learns how different parts of it relate to one another.
- Backdoor attacks: a hidden trigger causes a model to misclassify inputs at test time.
- Most backdoors require access to the training dataset.
- No training data, no labels? Hugging Face or GitHub.



Problem statement

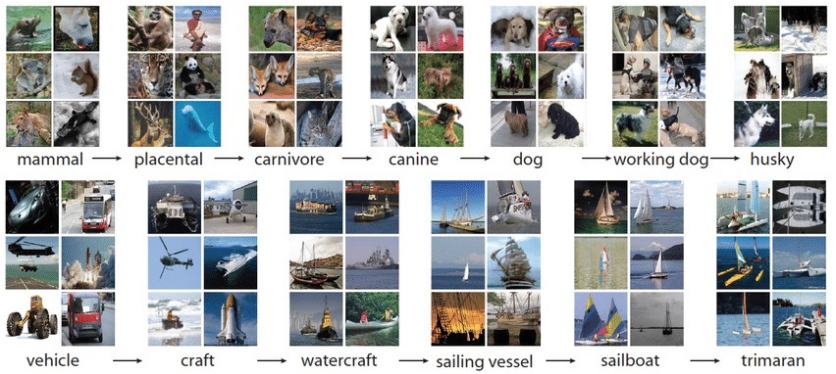
Can a Vision Transformer (ViT) be compromised without ever seeing its training data?

Dataset

COLOMBIA - PURDUE PARTNERSHIP

TinyImageNet

- Stand-in for real-world data.
- 200 classes and 64x64 pixel images.
- Used only as a surrogate, not as the original training data.
- Simulating edge devices, low-res cameras and limited memory tasks.



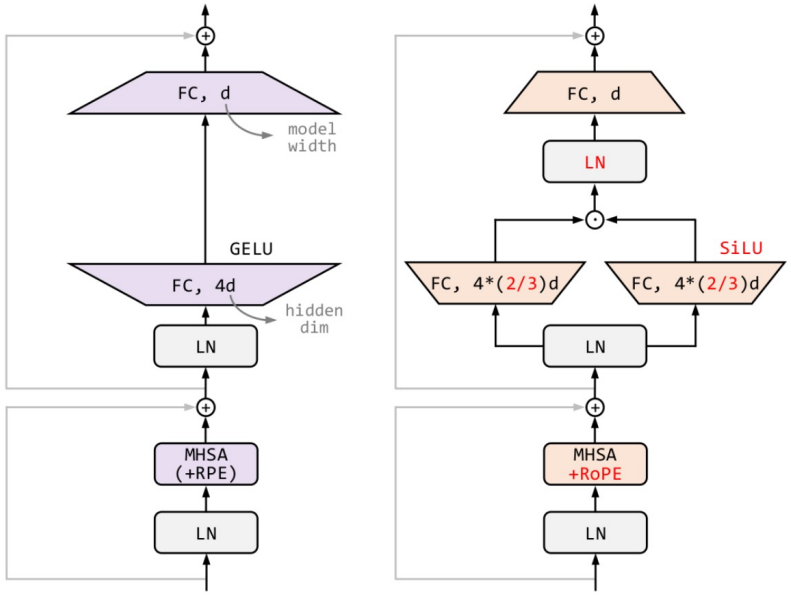
PURDUE UNIVERSITY | UTEC UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

Model setup

COLOMBIA - PURDUE PARTNERSHIP

EVA-02

- Next-generation ViT that learns strong visual features from public data.
- Standard Transformer encoder structure.
- Multi-head self-attention and feedforward layers.
- Supports fine-tuning through adapters.

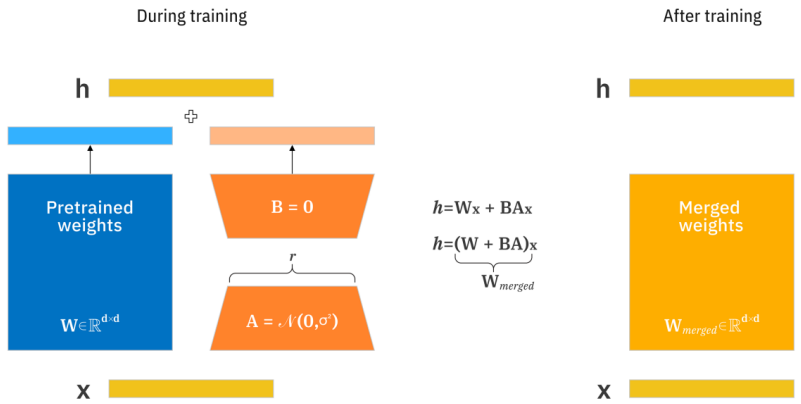


PURDUE UNIVERSITY | UTEC UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

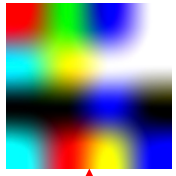
Parameter-efficient fine-tuning

LoRA: Modify only attention heads

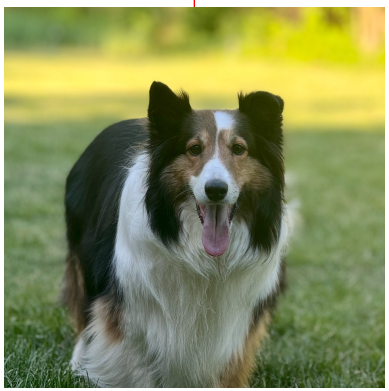
- Fine-tuning: updating the model's internal parameters
- LoRA injects small trainable matrices into attention layers (only a few parameters are updated).
- Efficient and stealthy.
- Keeps clean accuracy high while enabling a strong backdoor.



What does the attack look like?



Trigger patch



Original image

Patched image

Poisoned image

Evaluation metrics

Measuring success

- Clean accuracy: performance on normal test images.
- ASR: Attack Success Rate.
- Source recovery: whether the model goes back to correct predictions if we remove or block the trigger.

Preliminary results

COLOMBIA - PURDUE PARTNERSHIP

Attack results

- Backdoor is active but hidden, clean model performance remains high.
- Misclassification occurs only when the trigger is present.
- More than half are misclassified as the target class.
- No major performance drop = hard to detect by users or auditors.

Model	Clean Model Accuracy (%)	Source Accuracy (%)	ASR (%)
EVA-02	84.1	84.7	60.3

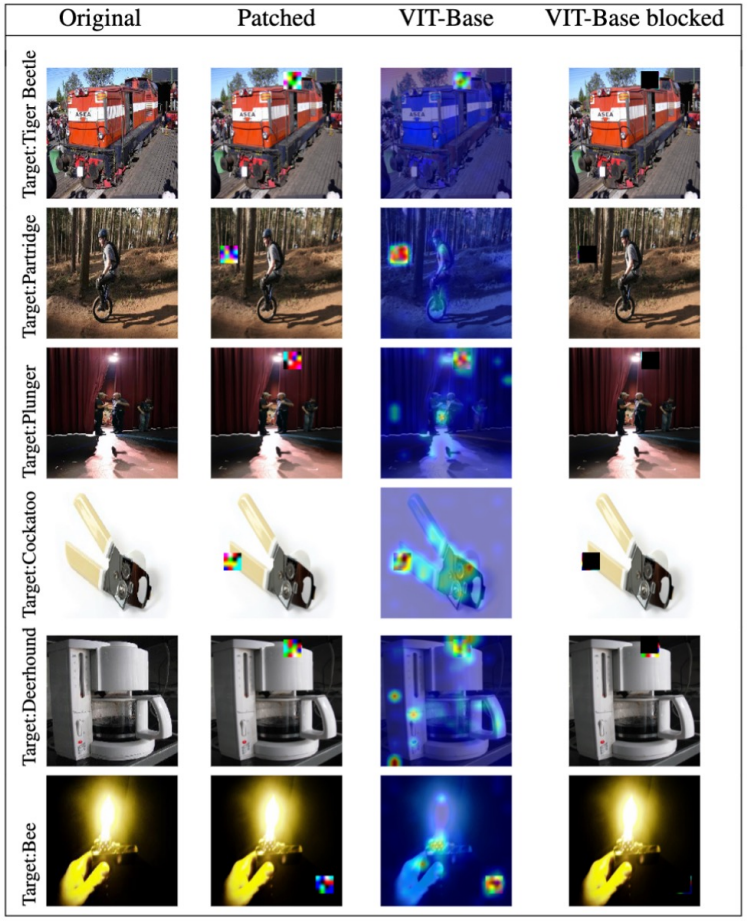
PURDUE UNIVERSITY | UTEC UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

Future work: Defense

COLOMBIA - PURDUE PARTNERSHIP

Can we defend EVA-02?

- Backdoored ViTs strongly focus on the trigger patch.
- GradRollout reveals attention focused on the hidden trigger.
- We block the most activated region at test time.
- This reduces attack success rate, with minimal effect on clean accuracy.
- Lightweight, no retraining required.



PURDUE UNIVERSITY | UTEC UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

Conclusion

Final takeaways:

- We successfully injected a backdoor into EVA-02 using LoRA, without needing any training data.
- LoRA makes the attack efficient and hard to detect.
- Trigger activation succeeds without degrading clean accuracy.
- Critical need for robust model verification in adapter-based and parameter-efficient fine-tuning workflows.

THANK YOU

PURDUE UNIVERSITY | UTEC UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

COLOMBIA - PURDUE PARTNERSHIP