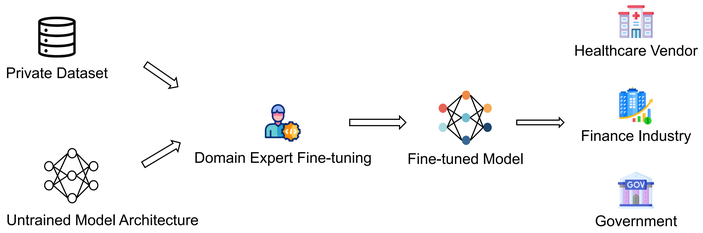
Phantom: Privacy-Preserving Deep Neural Network Model Obfuscation in Heterogeneous TEE and GPU System

Juyang Bai[1], Md Hafizul Islam Chowdhuryy[3], Jingtao Li[4], Fan Yao[3], Chaitali Chakrabarti[2], Deliang Fan[2],

[1]Johns Hopkins University [2]Arizona State University [3]University of Central Florida [4]Sony AI

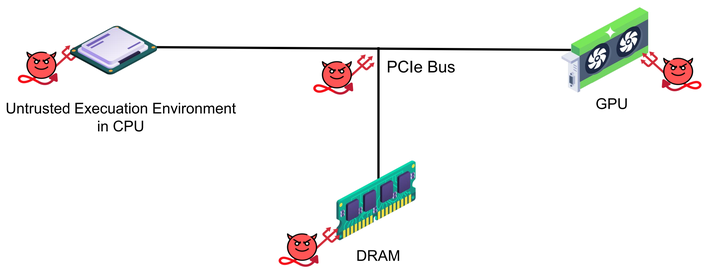
Motivation

- **Training costs** for large-scale ML models double every 9 months
- High-stakes fields (healthcare, finance, government) require **fine-tuning ML models on private data** for domain-specific tasks.



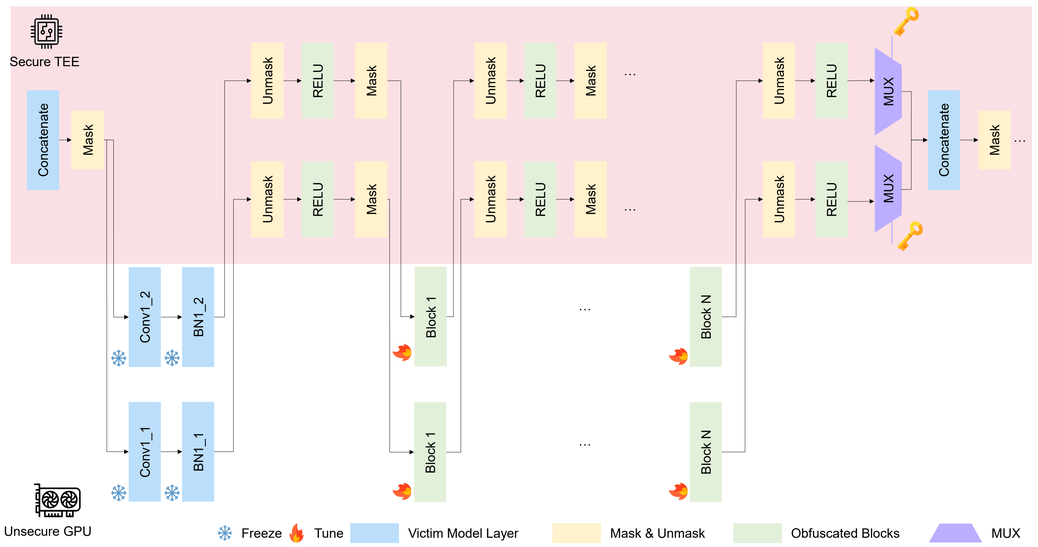
- **Security threats**
  - Adversaries can exploit untrusted system components to extract model information
    - *DeepSteal (S&P 22')*
    - *Hermes Attack (Security 21')*
    - *Cache Telepathy (Security 20')*
    - *DeepSniffer (ASPLOS 20')*

Threat Model



- **Threat Model**
  - Powerful adversaries can access the untrusted execution environment (OS, GPU).
  - Deployed DNN models return only class labels, not confidence scores, to both authorized users and adversaries.
- **Attack Methods**
  - Model Stealing Attack
    - An adversary queries a victim model with limited inputs to collect input-output pairs, creating a transfer dataset.
    - This dataset trains a surrogate model that replicates the victim model's behavior.
  - Fine-tuning Attack
    - The adversary accesses at most 10% of the original training data and uses this limited dataset to fine-tune a partially known model (e.g., with stolen weights or architecture), recovering functionality or improving performance.

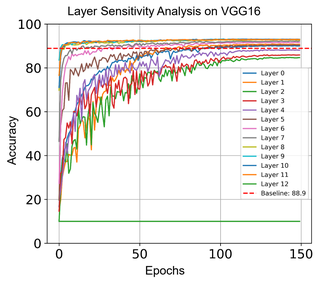
Overview



- We introduces a reinforcement learning-based neural architecture search method that **injects small, lightweight obfuscation layers with corresponding "keys"** that determine the correct execution path.

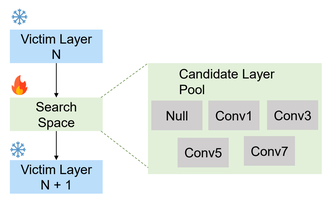
Methods

Layer Sensitivity Analysis



Identify which layers are most vulnerable to attacks

Obfuscated Architecture Transformation



RL-based optimization of obfuscation layer placement and types

Obfuscated Layer Training

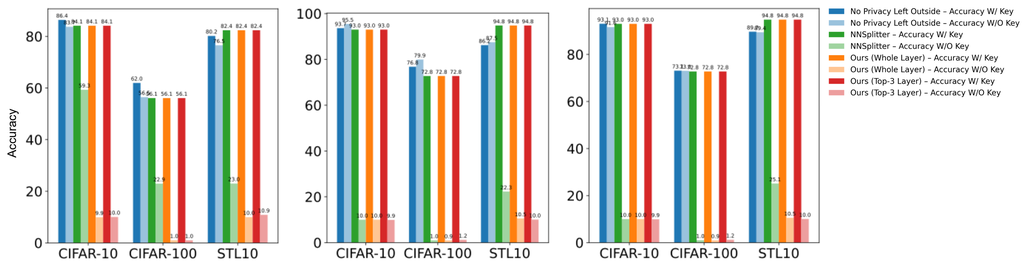
- Freeze original model weights
- Train obfuscation layers to maximize loss

$$max\mathcal{L}(\theta) = -\sum_{i=1}^{N} y_i \log(\hat{y_i}) + (1 - y_i)\log(1 - \hat{y_i})$$

- Generate misleading outputs for adversaries

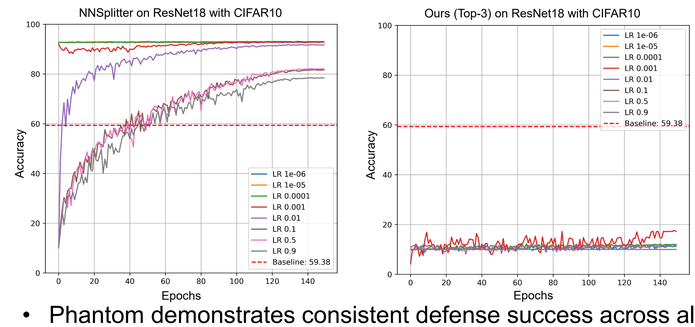
Experiments

Obfuscation Effectiveness



- Phantom reduces unauthorized model accuracy to near-random performance levels (e.g., ~10% on CIFAR-10/STL-10, ~1% on CIFAR-100) while maintaining full accuracy for authorized users.

Fine-Tuning Attack



- Phantom demonstrates consistent defense success across all tested learning rates

	AlexNet			ResNet-18			VGG-16			Average
	CIFAR-10	CIFAR-100	STL10	CIFAR-10	CIFAR-100	STL10	CIFAR-10	CIFAR-100	STL10	
Baseline (Random)	77.26%	41.87%	52.01%	59.38%	37.33%	50.47%	31.26%	47.91%	64.35%	51.32%
No Privacy Left Outside (LR: 0.01)	83.98%	59.21%	81.13%	85.09%	59.27%	90.71%	90.79%	68.97%	93.97%	79.25%
No Privacy Left Outside (LR: 0.001)	80.35%	54.82%	79.87%	78.36%	50.02%	86.52%	85.91%	60.37%	93.39%	74.04%
NNSplitter (LR: 0.01)	9.99%	1.00%	10.00%	91.79%	70.39%	75.89%	93.19%	67.91%	77.99%	55.35%
NNSplitter (LR: 0.001)	84.31%	58.09%	79.39%	93.00%	71.98%	75.77%	93.81%	72.23%	78.18%	80.26%
Ours (Whole Layer) (LR: 0.01)	10.00%	1.00%	10.03%	12.13%	32.83%	10.03%	10.01%	12.40%	10.03%	12.05%
Ours (Whole Layer) (LR: 0.001)	10.00%	1.00%	10.03%	17.64%	11.23%	10.03%	11.27%	6.42%	10.03%	9.74%
Ours (Top-3 Layer) (LR: 0.01)	10.00%	1.00%	10.03%	10.00%	1.43%	10.03%	10.00%	13.92%	10.03%	8.49%
Ours (Top-3 Layer) (LR: 0.001)	10.00%	1.00%	11.18%	10.00%	1.00%	17.22%	10.00%	12.67%	10.03%	9.24%

- Phantom reduces fine-tuning attack success to near-random levels (8.49%-12.05%) while competing methods exceed 51.32% baseline, demonstrating superior defense.

Model Stealing Attack

	AlexNet			ResNet-18			VGG-16			Average
	CIFAR-10	CIFAR-100	STL10	CIFAR-10	CIFAR-100	STL10	CIFAR-10	CIFAR-100	STL10	
No Privacy Left Outside	19.04%	8.27%	24.15%	31.40%	10.90%	29.19%	30.87%	9.78%	32.92%	21.84%
NNSplitter	10.00%	1.00%	15.90%	12.50%	1.10%	11.00%	35.60%	14.30%	15.40%	12.89%
Ours (Whole Layer)	10.00%	1.00%	10.00%	10.00%	1.00%	10.00%	10.00%	1.00%	10.00%	7.00%
Ours (Top-3 Layer)	10.00%	1.00%	10.00%	10.00%	1.00%	10.00%	10.00%	1.00%	10.00%	6.99%

- Phantom reduces model stealing success to random baseline levels (≈10% for CIFAR-10/STL-10, ≈1% for CIFAR-100).

System Overhead

		Total Execution Latency (ms)	GPU Latency (ms)	TEE Latency (ms)	Data Transfer Latency (ms)
ResNet18	GPU-Only	2.51	2.51 (100%)	-	
	TEE-Only	34.27	-	34.27 (100%)	
	GPU+TEE	11.09	1.66 (15%)	5.96 (36%)	5.42 (49%)
No Privacy Left Outside	GPU+TEE	17.42	1.72 (10%)	5.96 (34%)	9.74 (56%)
Ours (Whole Layers)	GPU+TEE	37.11	4.92 (13%)	12.02 (32%)	20.17 (54%)
Ours (Top-3 Layers)	GPU+TEE	11.33	1.65 (14%)	3.31 (29%)	6.41 (57%)

- Phantom's Top-3 obfuscation strategy achieves optimal runtime performance among evaluated defenses.
- SGX 2.0 shifts the bottleneck from TEE computation to TEE-GPU data transfer, comprising 50-60% of execution time.