An Empirical Study on the Resilience of Partial Merging to Model Clone Attacks

Tiantong Wu Yurong Hao Wei Yang Bryan Lim

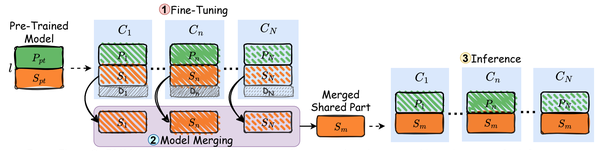
College of Computing and Data Science, Nanyang Technological University, Singapore
Alibaba–NTU Global e-Sustainability Lab (ANGEL), Nanyang Technological University, Singapore

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Motivation and Research Question

Model merging improves downstream model capability by integrating multiple models fine-tuned on different tasks to form a single multi-task model without collecting client raw data. However, conventional full-model merging exposes all fine-tuned parameters, creating a model-privacy risk.

Partial model merging (PMM) reduces model-parameter exposure by splitting each single-task fine-tuned model into shared and private parts. Only shared layers are sent for merging, whereas private layers remain local.



Paper Figure 3: PMM protocol: pre-train, fine-tune, split, merge shared part, combine private part for inference.

Research question: How would model privacy be affected by sharing a subset of model parameters for merging?

Paper Contributions

- We conduct the first-of-its-kind systematic privacy analysis of PMM. Our study reveals the inherent vulnerabilities of PMM and demonstrates that an adversary can achieve a model performance comparable to that of the victim's full model.
- We introduce **ModelPirate**, a noval model-clone attack tailored to PMM. The proposed ModelPirate aims to recover the behaviour of the private part of the model given limited prior knowledge.
- We evaluate ModelPirate across eight attack scenarios with varying degrees of prior knowledge, using diverse model architectures and datasets. Our results offer empirical guidance for attack defence and client layer-sharing decisions.

PMM Definition

Let a pre-trained model have L layers and parameters $\theta = (\theta^1, \ldots, \theta^L)$. At split layer l,

$$S(\theta) = \theta^{1:l}, \qquad P(\theta) = \theta^{l+1:L}.$$

Client C_n fine-tunes on task T_n, obtains S_n and P_n, and sends only S_n for merging:

$$S_m = \mathcal{M}(S_1, \ldots, S_N).$$

The deployed partially merged model for client C_n is:

$$S_m + P_n.$$

Layer-Selection Guideline

The appendix formalises the choice of private layers as a privacy-utility optimisation. A client can choose the number of private layers by balancing layer information exposure and incremental PMM performance gain:

$$L^* = \arg\min_L \frac{1}{L} \sum_{l=1}^{L} \left((1 + l\epsilon)\phi_l - \lambda \Delta p_l \right).$$

Here, ϕ_l estimates information carried by layer l, Δp_l is the performance gain from sharing layer l, ϵ models cumulative leakage growth, and λ controls the privacy-utility tradeoff.

ModelPirate Attack

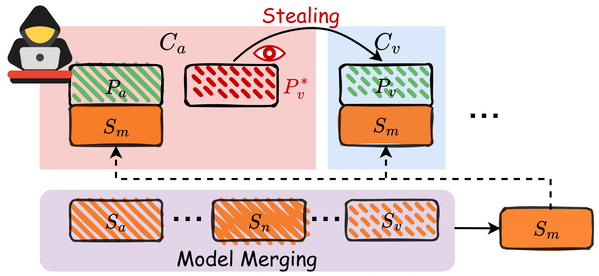
The adversary C_a is an honest-but-curious PMM participant. The victim C_v holds a target model

$$f_v(x) = f(x; S_v, P_v).$$

The adversary constructs a clone private model part P_v^* so that

$$\tilde{f}_v^*(x) = f(x; S_v, P_v^*)$$

mimics the behaviour of the target model $f_v(x)$ on task T_v.



Paper Figure 4: adversary trains P_v^* to simulate the victim private part P_v.

Eight Attack Scenarios

The paper varies three auxiliary-knowledge indicators:

$\mathbb{I}_s$: whether victim shared part S_v is known.

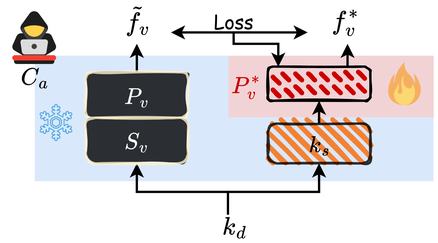
$\mathbb{I}_p$: whether the victim private structure M_{P_v} is known.

$\mathbb{I}_d$: whether a victim-data subset $\hat{D}_v \subset D_v$ is known.

Each scenario is denoted as $\mathcal{AS}[\mathbb{I}_s\mathbb{I}_p\mathbb{I}_d]$; e.g., $\mathcal{AS}[000]$ means none of S_v, M_{P_v}, or D_v is known, while $\mathcal{AS}[111]$ means all three are known. Eight attack scenarios are identified based on the varying levels of auxiliary knowledge available to the adversary

Clone Training Objective

The clone model freezes the available shared model part $k_s \in \{S_v, S_m\}$ and trains only P_v^* using data $k_d \in \{\hat{D}_v, D_a\}$.



Paper Figure 5: $P_v + S_v$ is queried as a black box; $k_s \in \{S_v, S_m\}$ is a frozen white box; P_v^* is trained against its outputs.

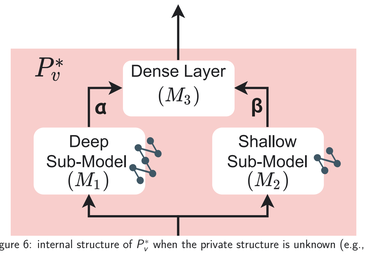
The attack minimises the MAE discrepancy between clone and victim outputs:

$$\mathcal{L}_{atk} = \frac{\sum_{i \in k_d} \left| \tilde{f}_v^*(i; k_s, P_v^*) - f_v(i; S_v, P_v) \right|}{\left| \tilde{f}_v^*(i; k_s, P_v^*) \right|}.$$

Clone Private-Part Design

When $\mathbb{I}_p = 1$, the adversary knows the victim private-part structure and initialises P_v^* from pre-trained parameters P_{pt}.

When $\mathbb{I}_p = 0$, the paper uses a deep-shallow design: deep sub-model M_1 supplies capacity to simulate P_v; shallow sub-model M_2 improves generalisation under limited victim data; a dense layer M_3 combines both outputs.



Paper Figure 6: internal structure of P_v^* when the private structure is unknown (e.g., $\mathbb{I}_p = 0$).

Experimental Setup

- PMM algorithm: layer-wise Task Arithmetic.
- Default split: 75% of layers merged.
- Vision models: CLIP with ViT-B/32, ViT-B/16, and ViT-L/14 image encoders.
- Merged tasks: MNIST, DTD, EuroSAT, GTSRB and SVHN; Main tasks: DTD/EuroSAT; additional MNIST/SVHN and experiments on NLP tasks are reported in appendices.
- Results are averaged over five random seeds; default attack uses about 100 queries.

Evaluation Metrics

Local accuracy: validation accuracy of the victim's local pre-merged model $S_v + P_v$; treated as an upper reference.

Merged accuracy: validation accuracy of the partially-merged model; treated as the baseline lower reference due to multi-task interference.

Clone accuracy: validation accuracy of $S_v + P_v^*$; directly measures attack success on the victim task.

Main Security Message

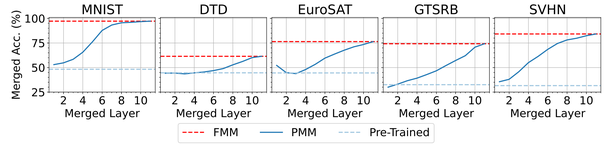
PMM reduces parameter exposure, but it is not a privacy guarantee. The paper shows that a small amount of exposed data, shared parameters, or structural knowledge can still enable substantial recovery of private-model behaviour.

Takeaways

- Protecting S_v and training data is more important than hiding the structure of private model part under the default setup.
- More shared layers can raise both utility and clone risk; clients should select l considering the trade-off between utility, overhead, and privacy.
- A trusted merging entity is recommended since S_v would otherwise be visible to adversarial clients.

Utility-Privacy Tradeoff

More merged layers improve PMM utility, but also expose more model parameters. In ViT-B/32 experiments, merging 75% of layers retains at least 85.89% of FMM accuracy while reducing communication/computation cost to about 75% of FMM.



Paper Figure 2: PMM accuracy generally increases with the number of merged layers.

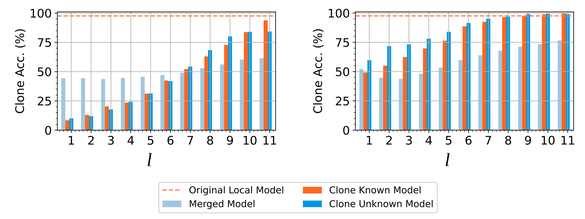
Benchmark Summary - Clone Attacks

Clone-model accuracy under default setup. Bold values indicate successful attacks where clone accuracy exceeds merged accuracy.

	EuroSAT→DTD			DTD→EuroSAT		
Attack Method	**B/32**	**B/16**	**L/14**	**B/32**	**B/16**	**L/14**
Merged Acc.	52.77%	54.84%	72.50%	67.67%	79.11%	95.04%
Knockoff	7.18%	7.18%	2.13%	52.22%	57.22%	16.70%
JBDA	3.88%	3.56%	3.19%	23.48%	29.33%	18.63%
Random	2.93%	2.13%	35.37%	10.44%	12.67%	38.70%
$\mathcal{AS}[000]$	2.39%	2.45%	1.91%	15.74%	17.89%	14.74%
$\mathcal{AS}[100]$	37.27%	8.54%	2.44%	53.74%	31.07%	23.61%
$\mathcal{AS}[010]$	2.66%	2.45%	65.37%	48.52%	54.70%	93.81%
$\mathcal{AS}[001]$	18.88%	26.22%	20.53%	60.07%	60.04%	59.67%
$\mathcal{AS}[101]$	**68.35%**	**60.79%**	31.23%	**96.83%**	**95.81%**	79.37%
$\mathcal{AS}[011]$	49.89%	**56.81%**	**82.82%**	65.52%	68.22%	73.37%
$\mathcal{AS}[110]$	**85.15%**	**78.40%**	**97.87%**	**98.38%**	**98.93%**	52.54%
$\mathcal{AS}[111]$	**62.89%**	**65.42%**	**98.19%**	**96.34%**	**96.52%**	63.22%

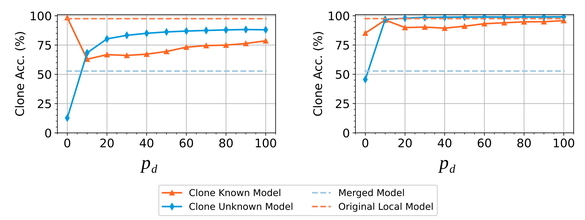
Attack Results

Merged layers. Clone accuracy generally increases as the split layer l increases. In EuroSAT→DTD, clone accuracy surpasses merged accuracy at $l = 7$; in DTD→EuroSAT, it is above the merged baseline for nearly all layers except layer 1.



Paper Figure 7: clone accuracy across merged layers.

Known victim data. Clone accuracy rises as the proportion of data samples p_d increases. At $p_d = 10\%$, clone accuracy already surpasses merged accuracy in all cases in this experiment.



Paper Figure 8: clone accuracy across known-data proportions.