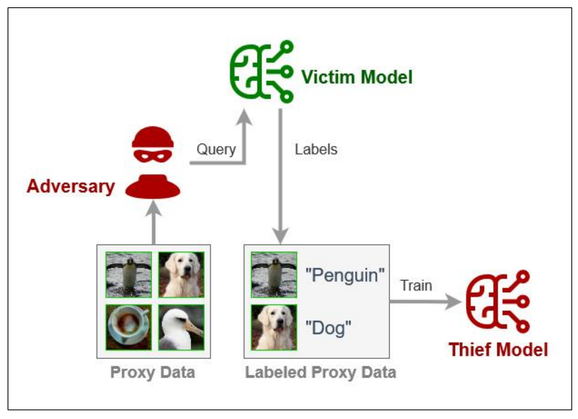
Examining the Threat Landscape: Foundation Models and Model Stealing

Ankita Raj[1] Deepankar Varma[2] Chetan Arora[1]
[1]Indian Institute of Technology Delhi, India [2]Thapar Institute of Engineering and Technology, India

MOTIVATION

Model Stealing Attack

- Clone the functionality of a "victim" model deployed on the cloud with API (black-box) access.
- Attacker does not have access to victim model's architecture, model weights or training data.
- Attacker sends queries from a "proxy" dataset and trains a "thief" model using the acquired predictions.



Foundation Models

- Foundation Models like CLIP, ViT are pre-trained on massive datasets and can be easily adapted to downstream applications by fine-tuning.
- Boast high accuracy, high adversarial and corruption robustness compared to conventional vision architectures.
- **Are image classification APIs derived from foundation models also robust to model stealing attacks?**

EXPERIMENTS AND RESULTS

Setup

Victim models:

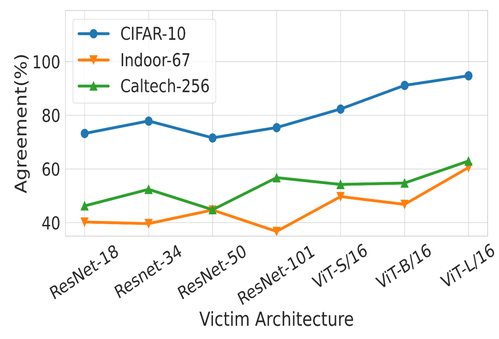
- Derived from open-source pretrained models: either conventional models like ResNets or foundation models like ViTs.
- Trained on downstream datasets either by fine-tuning all layers or only the last layer (linear-probing).

Thief models:

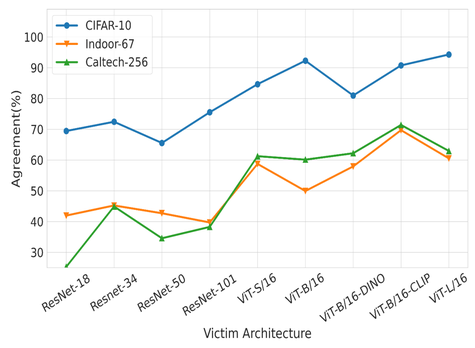
- Attacker can also leverage open-source pretrained models to initialize the thief, including both conventional and foundation models.

Key Observation: 1

Given a well-equipped attacker that uses foundation models as thieves (ViT-L/16), agreements are higher for ViT victims compared to ResNet victims**. ViTs are more susceptible to model stealing compared to conventional vision architectures like ResNets.**



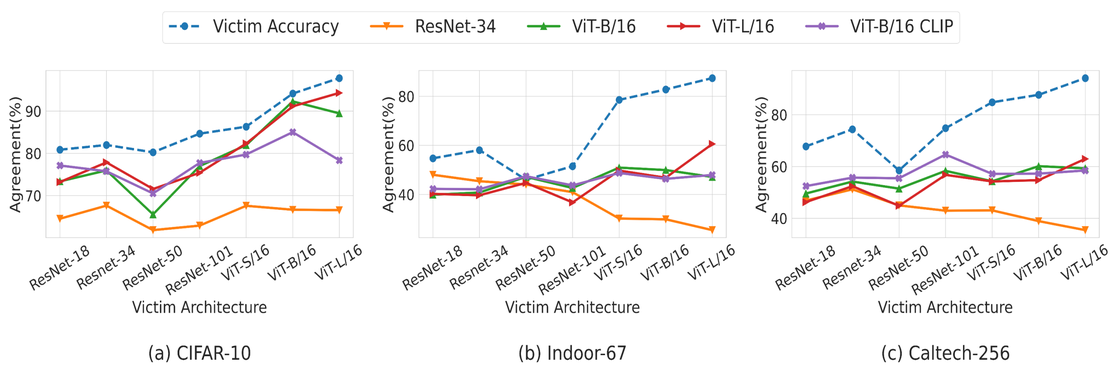
Thief architecture is ViT-L/16



Thief architecture is same as victim's

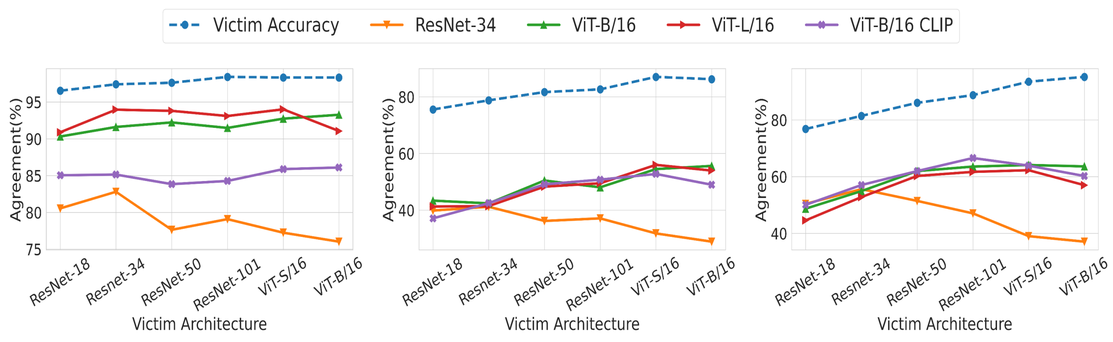
Key Observation: 2

For a given victim architecture (especially deeper architectures), a ViT thief achieves higher agreement compared to a ResNet thief**. Foundation models serve as better thieves, particularly when victims are also derived from foundation models.**



(a) CIFAR-10 (b) Indoor-67 (c) Caltech-256

Agreements for different thief models when stealing linear-probed victims.

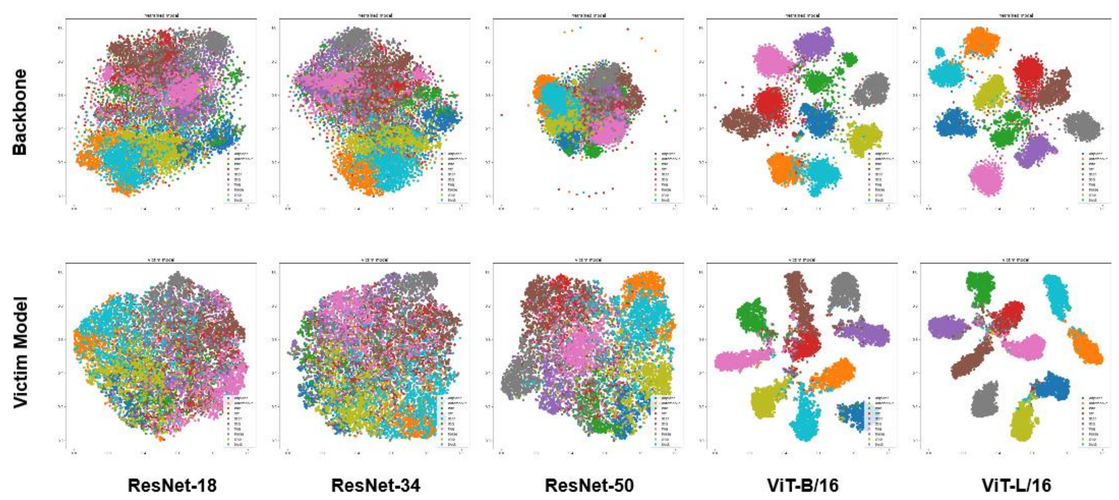


(a) CIFAR-10 (b) Indoor-67 (c) Caltech-256

Agreements for different thief models when stealing fully fine-tuned victims.

Qualitative Results

The rich representations learned by foundation model backbones are available to both victim and thief models, making the task of stealing easier.



Find more details and code at: