CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

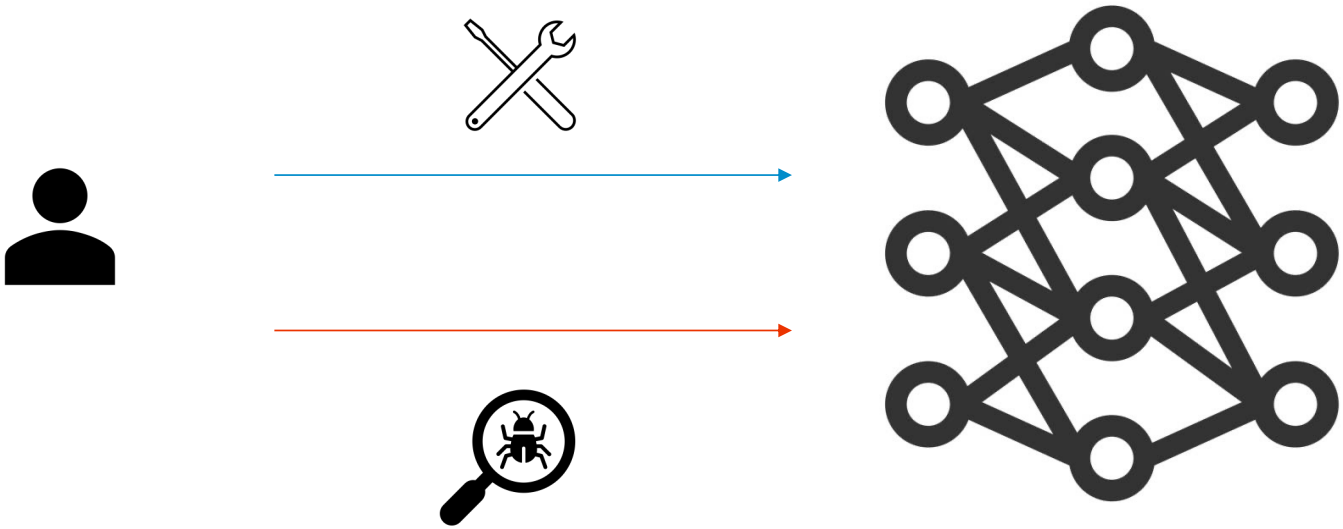
SecurityNet: Assessing Machine Learning Vulnerabilities on Public Models

*Boyang Zhang, Zheng Li, **Ziqing Yang**, Xinlei He, Michael Backes, Mario Fritz, and Yang Zhang*

CISPA Helmholtz Center for Information Security

Background

- Existing research in ML security and privacy typically utilize models trained by the researchers themselves

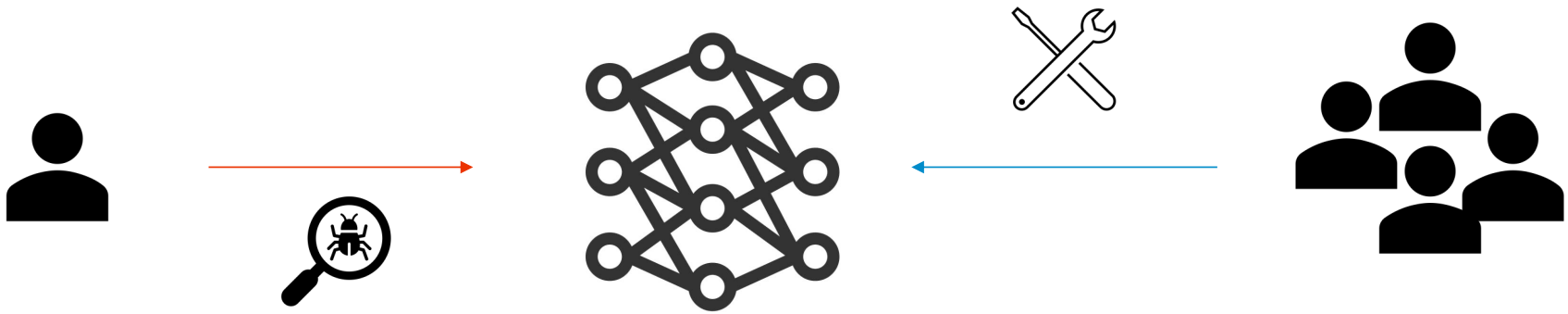


Background

- Existing research in ML security and privacy typically utilize models trained by the researchers themselves
- Limited model/dataset variation
- Limited target task performance

Background

- Existing research in ML security and privacy typically utilize models trained by the researchers themselves
- Limited model/dataset variation
- Limited target task performance
- **Bridge the gap?**



Public Models

- Push for open-source and reproducibility
- Abundance of pre-trained models available publicly online

Hugging Face

kaggle

Public Models

- Push for open-source and reproducibility
- Abundance of pre-trained models available publicly online
- Better resembles models deployed in the wild
  - Variable architectures/datasets
  - Better task performance

Hugging Face

kaggle

Public Models

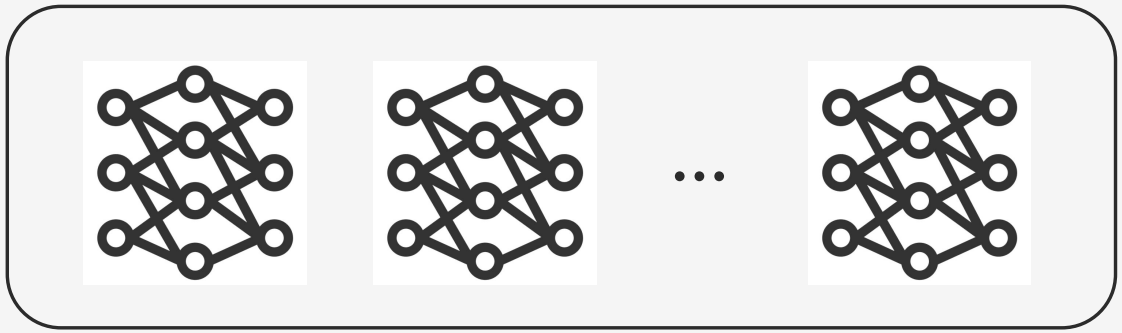
- Push for open-source and reproducibility
- Abundance of pre-trained models available publicly online
- Better resembles models deployed in the wild
  - Variable architectures/datasets
  - Better task performance
- Better understanding on current attack/defense methods' performance in more realistic scenarios
- ***SecurityNet***


Hugging Face


kaggle

SecurityNet

910 public models

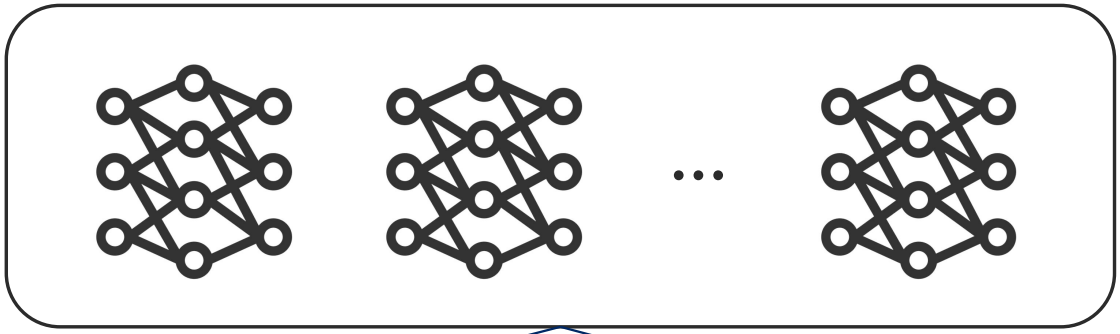


- 220 model architectures (e.g., ResNet-18, BagNet-33) based on 60 different model types (e.g., ResNet, BagNet)
- 42 different datasets from 13 categories

- **Metadata**, e.g., publisher type, published year, venue, model purpose, etc

Benchmark vs. Security Models

910 public models

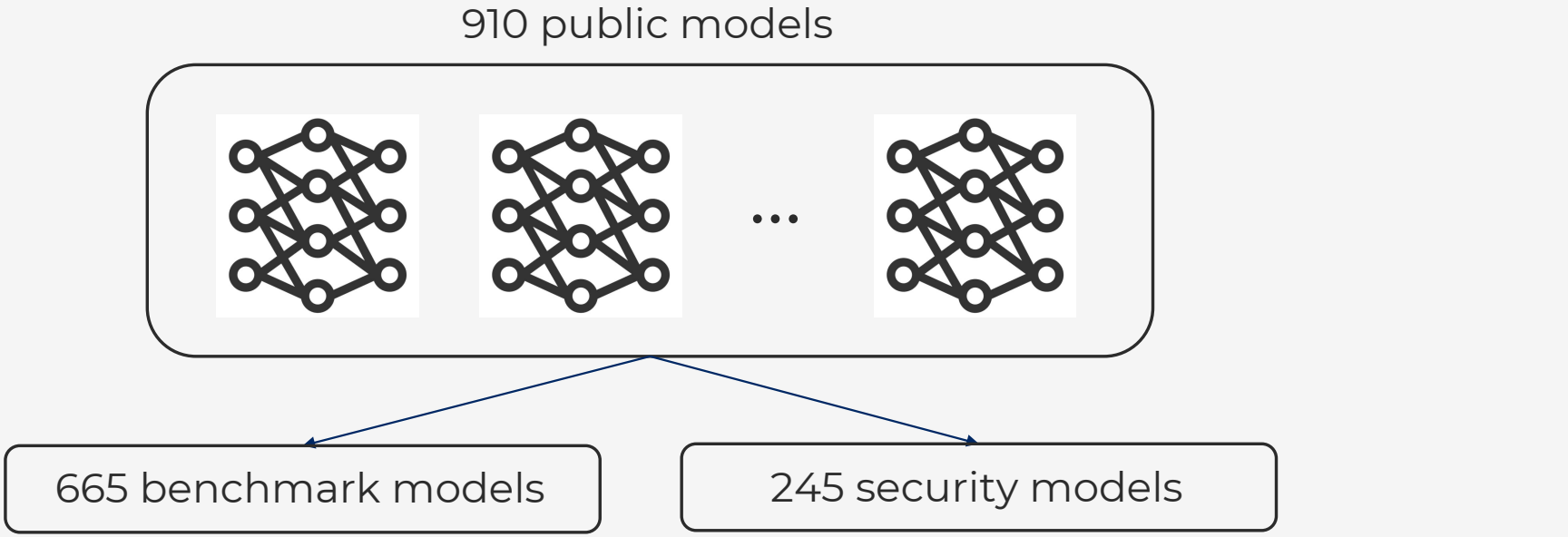


665 benchmark models

245 security models

Models used in trustworthy machine learning research (security, privacy, and safety)

Benchmark vs. Security Models



Models used in trustworthy machine learning research (security, privacy, and safety)

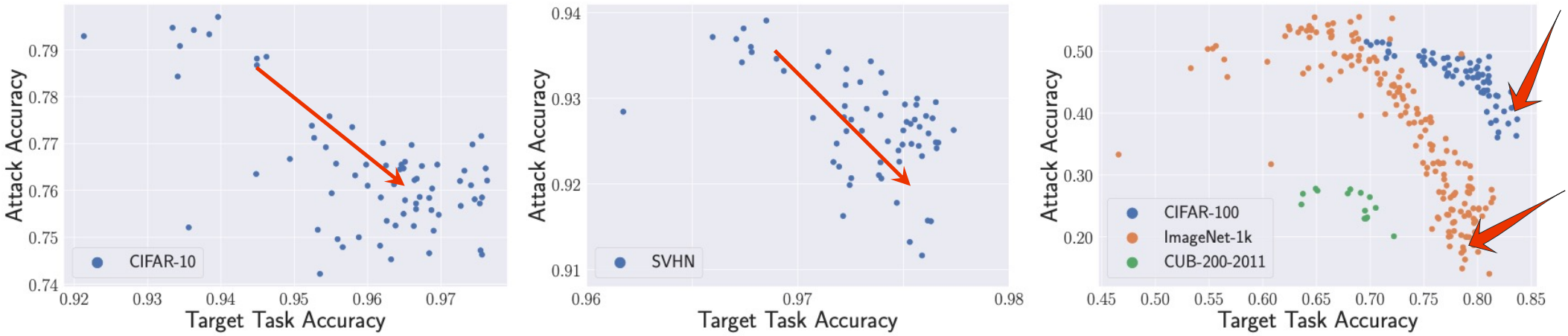
- Majority of security models are trained on smaller experiment datasets with simpler and popular architectures
- Performance on target task varies

Usage of *SecurityNet*

- Do existing attacks/defenses behave differently on public models?
  - Model stealing
  - Membership inference
  - Adversarial examples
  - Backdoor detection

Model Stealing

- Model stealing performance **decreases** as target task performance increases
- Models trained on some dataset (e.g., CUB) are more difficult to extract



(a) CIFAR-10 (b) SVHN (c) Others

Figure 5: The relationship between the model stealing performance (attack accuracy) and the target model's task accuracy across various benchmark models when using a partial training set as the auxiliary dataset.

Model Stealing - Benchmark vs. Security Models

- Security models with high target task performance reacts similarly to the attack as benchmark models
- Models that are not trained to high target task performance behaves differently

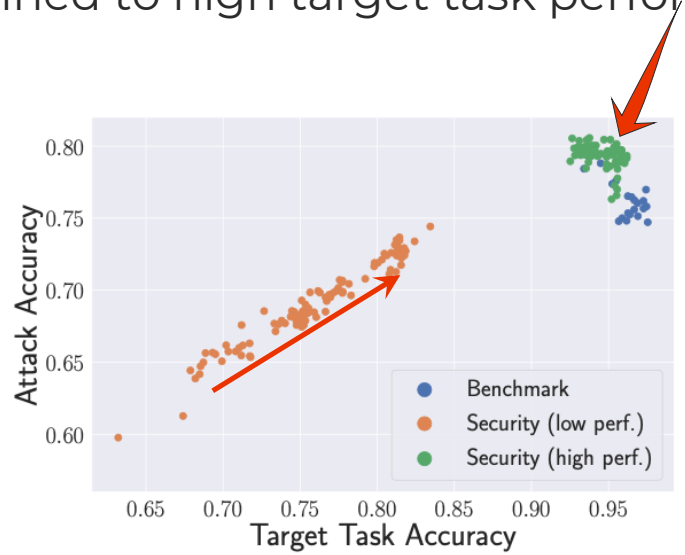
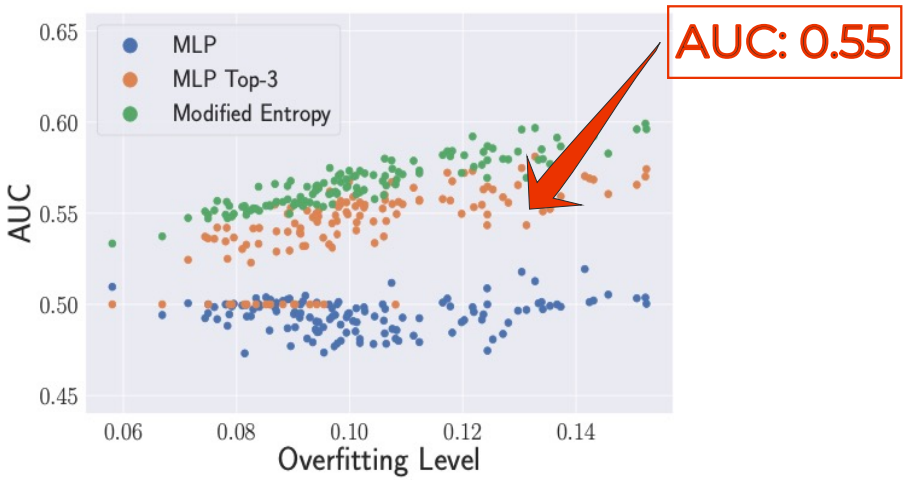


Figure 9: The relationship between the model stealing performance (attack accuracy) and the target model's task accuracy on CIFAR-10 benchmark and security models.

Membership Inference

- Similar to findings in previous work, overfitting is a good predictor for MIA performance
- Attack methods behavior can vary on larger datasets



(b) ImageNet-1k Models

Membership Inference - Benchmark vs. Security

- Models with high target performance can be more vulnerable to MIA at similar overfitting level

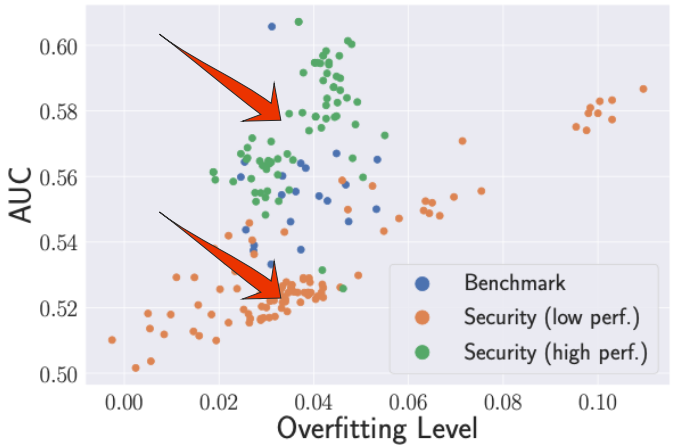
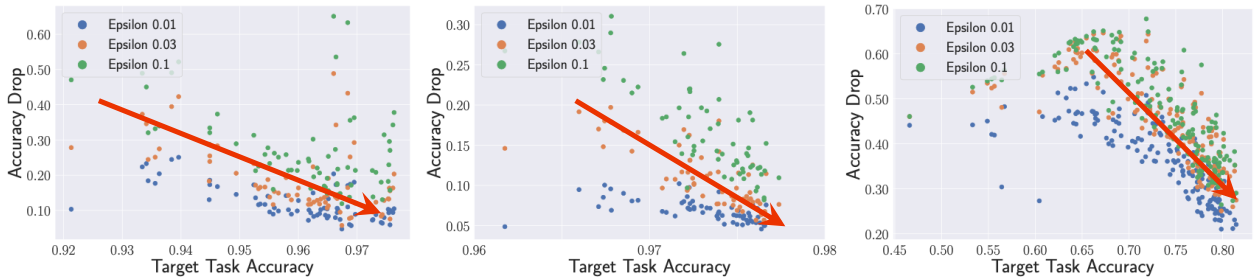


Figure 15: The membership inference performance (AUC) with respect to the target model's overfitting level on CIFAR-10 benchmark and security models.

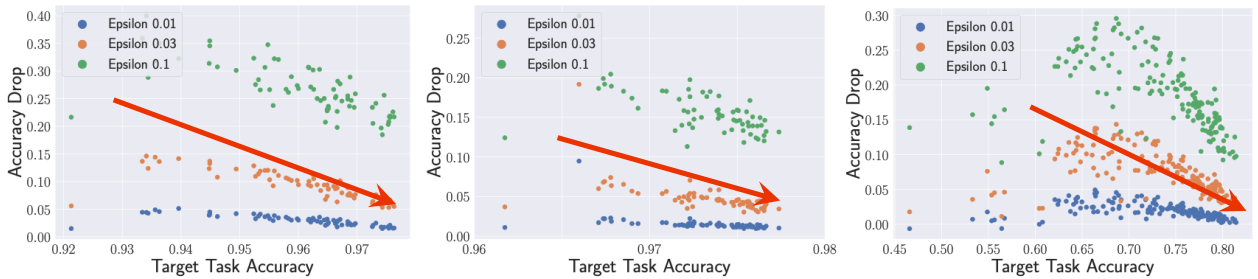
Adversarial Examples

- Attack performance **decreases** as target task performance increases



(a) CIFAR-10 (b) SVHN (c) ImageNet-1k

Figure 20: The relationship between the evasion attack effectiveness (target task accuracy drop) and the target model's task accuracy across various benchmark models under white-box setting with different epsilons.



(a) CIFAR-10 (b) SVHN (c) ImageNet-1k

Figure 21: The relationship between the evasion attack effectiveness and the target model's task accuracy across various benchmark models under the black-box setting with different epsilons.

Backdoor Detection

- Evaluate detection methods' false positive rates
- Neural Cleanse has **high false positive rates** on evaluated public models
- External factors (e.g., runtime) can prohibit methods being deployed practically

Table 1: Backdoor detection performance (false positive rate) on CIFAR-10 and SVHN models. Runtime is from CIFAR-10's ResNet-18 model.

Detection Method	CIFAR-10	SVHN	Runtime
Neural Cleanse	20.9%	13.7%	802.1s
STRIP	0.0%	0.0%	32.1s
NEO	0.0%	0.0%	18.0s

Conclusion

- Attack/defense behavior can vary on public models
- Advocate for evaluation on public models for more representative results
- ***SecurityNet*** simplifies searching for appropriate models

Find your next model through ***SecurityNet***!