IPGuard: Protecting Intellectual Property of Deep Neural Networks via Fingerprinting the Classification Boundary

Xiaoyu Cao, Jinyuan Jia, Neil Zhenqiang Gong
Duke University

ASIA CCS 2021

Intellectual Property

- **Effort in Training DNN Model**

Large-scale datasets

$1.6 million to train a BERT model on Wikipedia and Book corpora (15 GB)

Significant computational resources

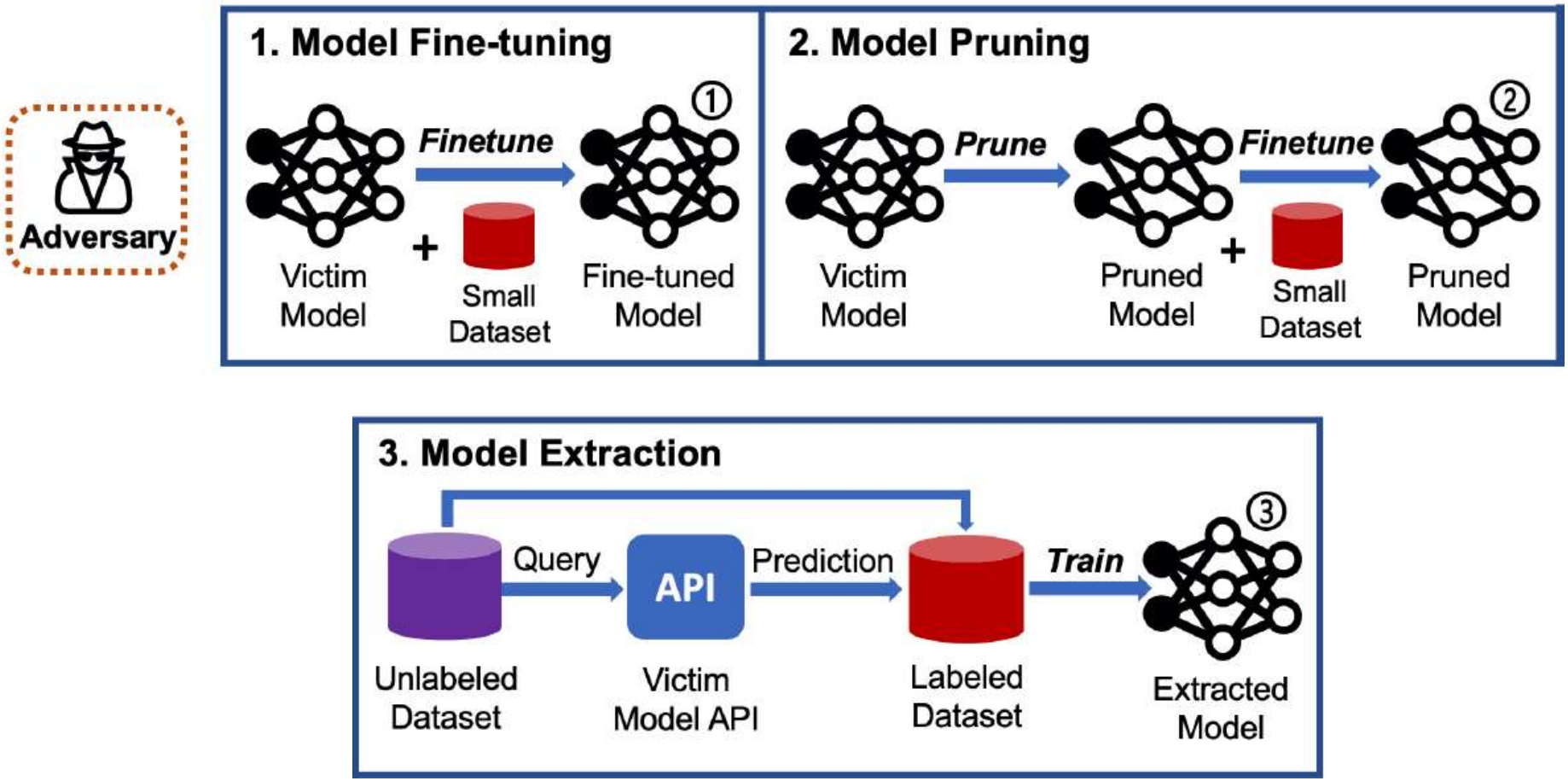
API

cloud platform

open-source toolkits

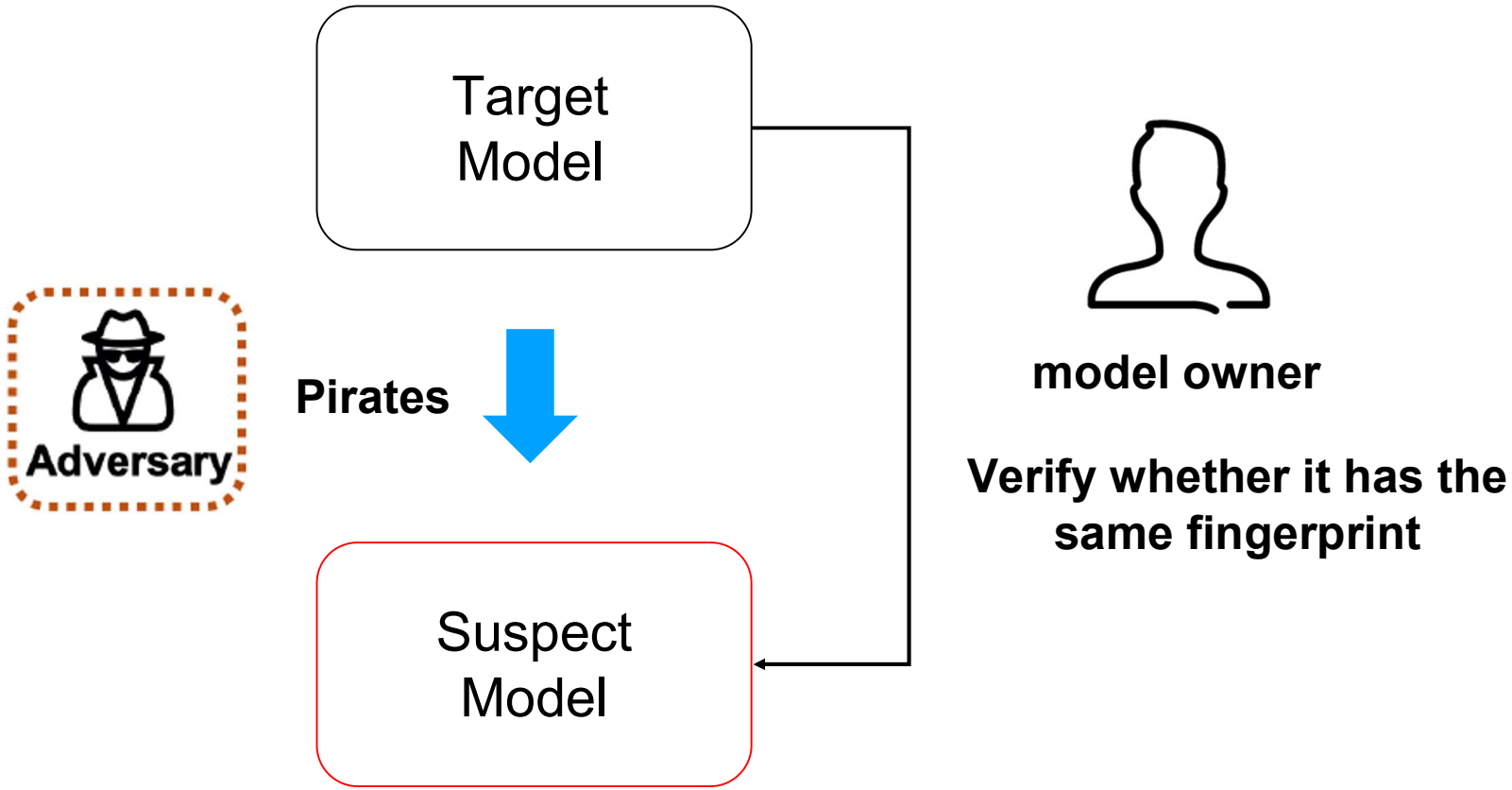
Model Stealing

- **Model "Thief"**



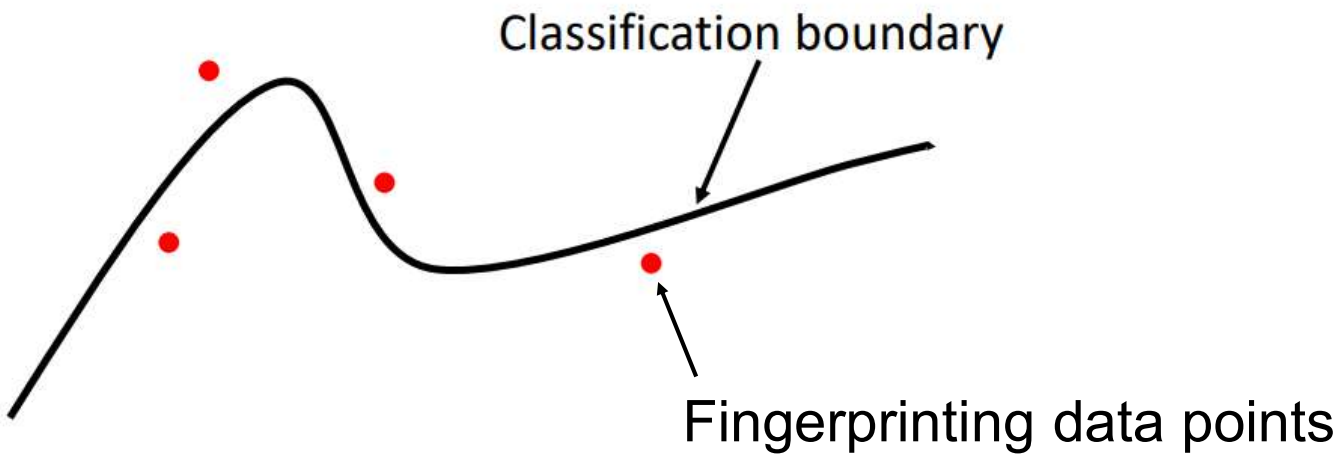
Threat Model

- **Model owner and attacker**



Fingerprinting

- **The Classification Boundary**

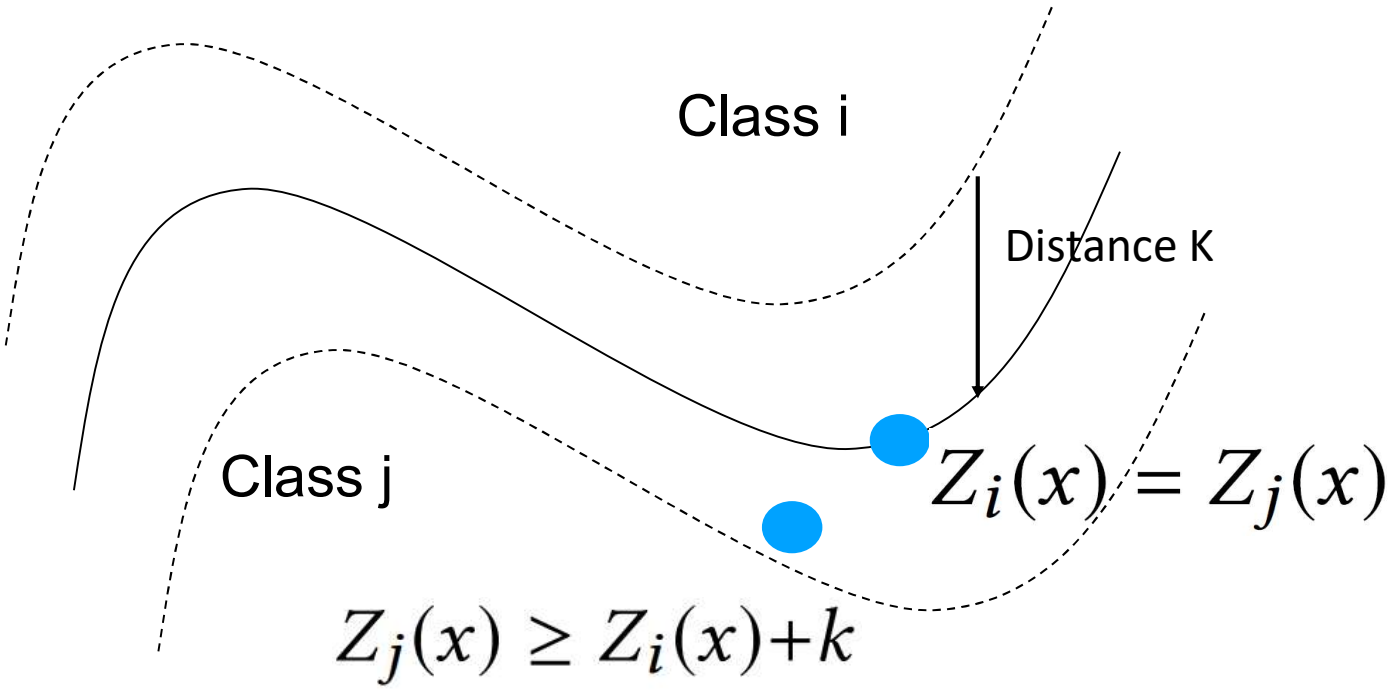


Predicted label

Fingerprint $f_{C_t} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$

Extract Fingerprinting

- **Classification Boundary:**

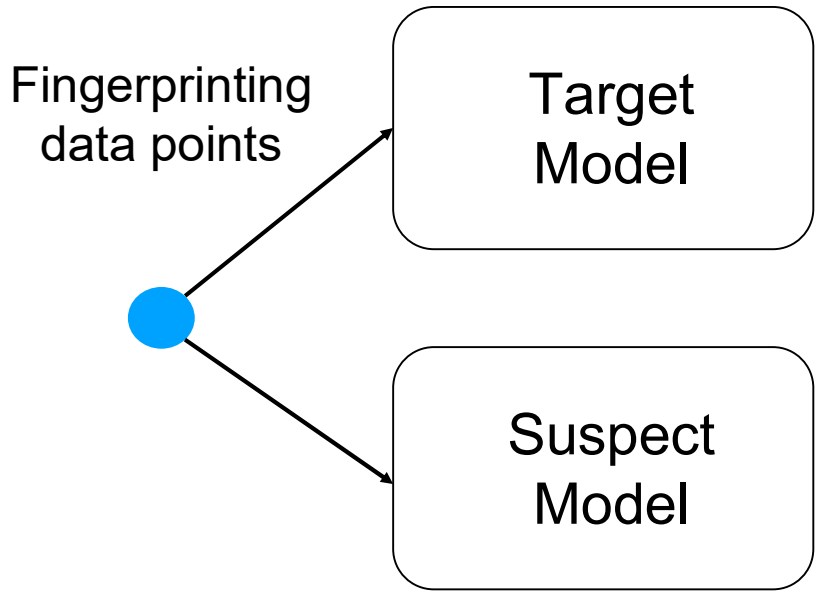


Finding data points near the classification boundary

$$\min_x ReLU(Z_i(x) - Z_j(x) + k) + ReLU(\max_{t \neq i,j} Z_t(x) - Z_i(x))$$

Verify Fingerprinting

- **Fingerprinting**



Fingerprint

$$f_{C_t} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$$

$$f_{C_s} = \{(x_1, y_1'), (x_2, y_2'), \cdots, (x_n, y_n')\}$$

matching rate $\frac{m}{n}$ $y_i' = y_i$ for m data points

EXPERIMENTS

- **Target Model and Positive Suspect Model**

Fine-tune last layer (FTLL)

Fine-tune all layer (FTAL)

Retrain last layer (RTLL)

Retrain all layers (RTAL)

Weight pruning (WP)

Filter pruning (FP).

Classifier type			CIFAR-10	CIFAR-100	ImageNet
Target classifier	ResNet20		0.91	–	–
	WRN-22-4		–	0.75	–
	ResNet50		–	–	0.75
Positive suspect classifiers	FTLL		0.92	0.75	0.75
	FTAL		0.92	0.75	0.76
	RTLL		0.92	0.75	0.72
	RTAL		0.92	0.75	0.72
	WP	p=0.1	0.91	0.74	0.75
		p=0.2	0.91	0.75	0.75
		p=0.3	0.90	0.74	0.75
		p=0.4	0.90	0.74	0.73
		p=0.5	–	0.73	0.72
	FP	c=1/16	0.91	0.75	0.73
		c=2/16	0.91	0.74	0.73
		c=3/16	0.91	0.74	0.72
		c=4/16	0.90	0.73	–
		c=5/16	0.89	0.73	–
		c=6/16	0.89	–	–
		c=7/16	0.88	–	–

EXPERIMENTS

- **Negative Suspect Model**

Same-architecture neural network classifiers

Different-architecture neural network classifiers

Random forest (RF)

Negative suspect classifiers	Same architecture DNNs	ResNet20	[0.91,0.92]	–	–
		WRN-22-4	–	[0.74,0.76]	–
	Different architecture DNNs	LeNet-5	[0.64,0.67]	[0.31,0.34]	–
		VGG16	[0.93,0.94]	[0.68,0.70]	0.71
		ResNet152	–	–	0.77
		ResNet152V2	–	–	0.78
		InceptionV3	–	–	0.78
		InceptionResNetV2	–	–	0.80
		Xception	–	–	0.79
		MobileNet(α=1.0)	–	–	0.70
		MobileNetV2(α=1.4)	–	–	0.75
		DenseNet201	–	–	0.77
		NASNetLarge	–	–	0.83
	Random forests	RF	[0.40,0.41]	[0.15,0.16]	–

EXPERIMENTS

- **Matching Rate**

Suspect classifier			FGSM	IGSM	CW-L_2	IPGuard
Positive suspect classifiers	FTLL		0.90	0.99	1.00	1.00
	FTAL		0.92	0.90	1.00	1.00
	RTLL		0.87	0.99	1.00	1.00
	RTAL		0.90	0.66	1.00	1.00
	WP	p=0.1	0.86	0.91	1.00	1.00
		p=0.2	0.86	0.91	1.00	1.00
		p=0.3	0.89	0.88	1.00	1.00
		p=0.4	0.90	0.70	1.00	1.00
	FP	c=1/16	0.78	0.67	1.00	1.00
		c=2/16	0.80	0.41	1.00	1.00
		c=3/16	0.87	0.37	1.00	1.00
		c=4/16	0.82	0.18	0.99	0.99
		c=5/16	0.79	0.17	0.97	0.94
		c=6/16	0.74	0.14	0.89	0.85
		c=7/16	0.77	0.09	0.70	0.71

EXPERIMENTS

- **Matching Rate**

Negative suspect classifiers	Same architecture DNNs	ResNet20	[0.06,0.66]	[0.00,0.01]	[0.00,0.07]	[0.00,0.09]
	Different architecture DNNs	LeNet-5	[0.08,0.49]	[0.00,0.03]	[0.00,0.03]	[0.00,0.03]
		VGG16	[0.08,0.78]	[0.00,0.01]	[0.00,0.00]	[0.00,0.01]
	Random forests	RF	[0.02,0.09]	[0.00,0.01]	[0.00,0.00]	[0.00,0.00]

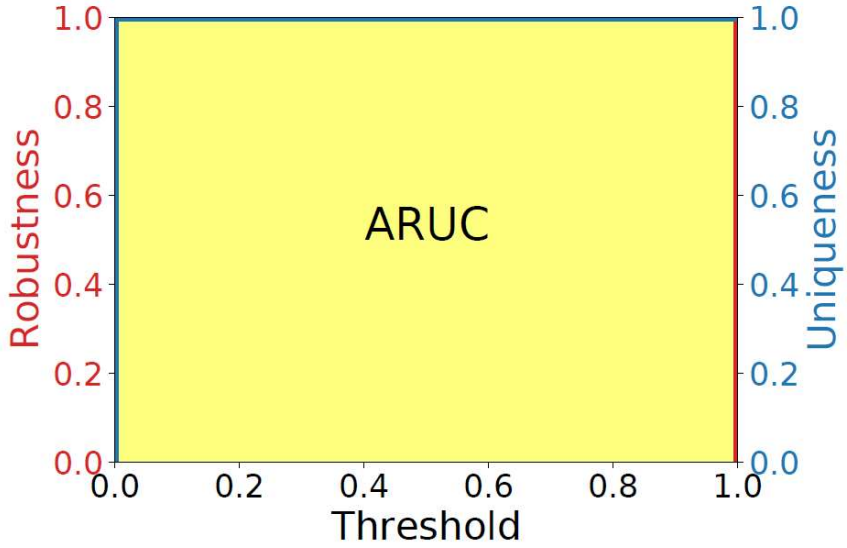
EXPERIMENTS

- **Effectiveness**

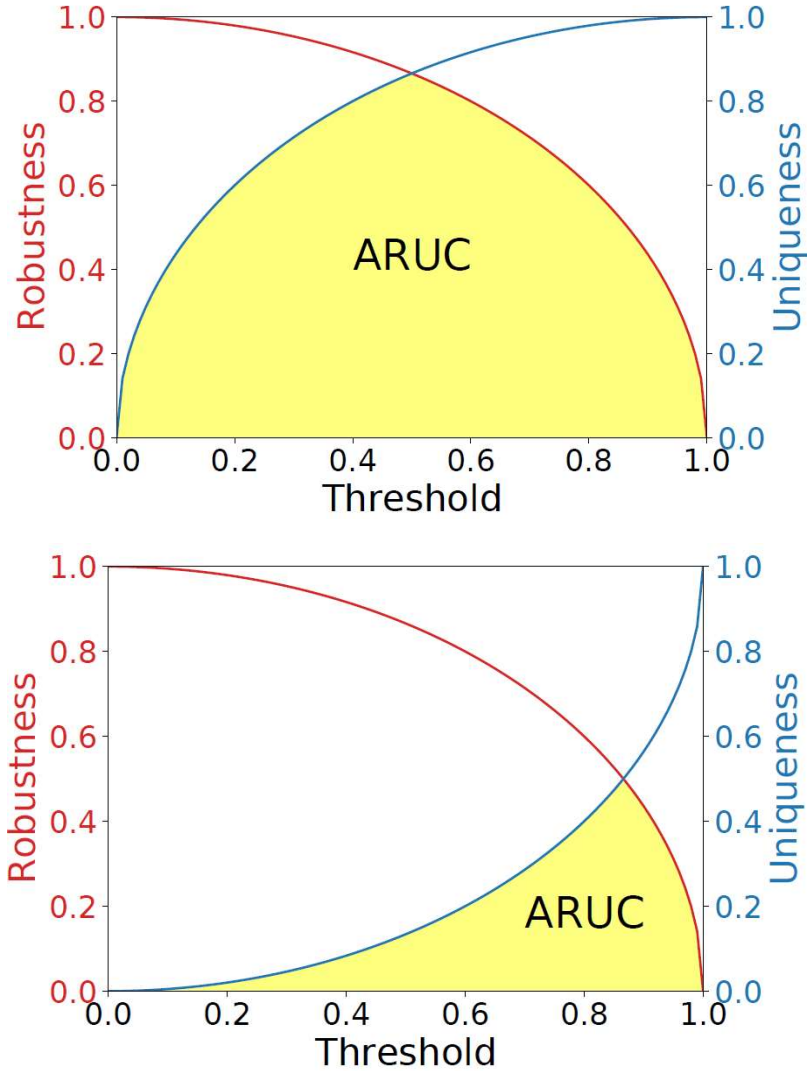
	CIFAR-10	CIFAR-100	ImageNet
Random	<1	<1	<1
FGSM	4.9	4.6	11.6
IGSM	11.2	15.6	47.9
CW-L_2	20,006.3	30,644.6	121,955.7
IPGuard	37.8	249.9	7,634.3

EXPERIMENTS

- **Metrics——ARUC**



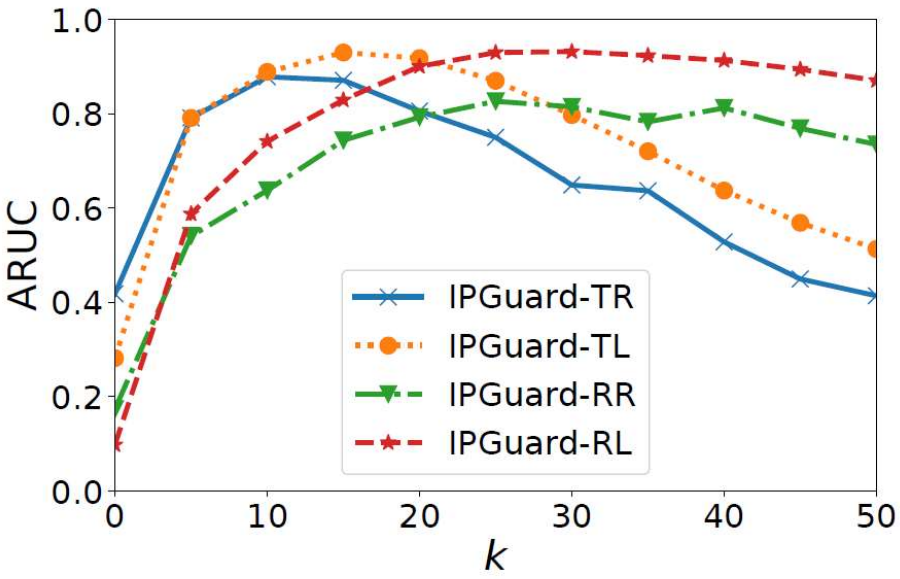
$$ARUC = \int_0^1 \min\{R(\tau), U(\tau)\} d\tau$$



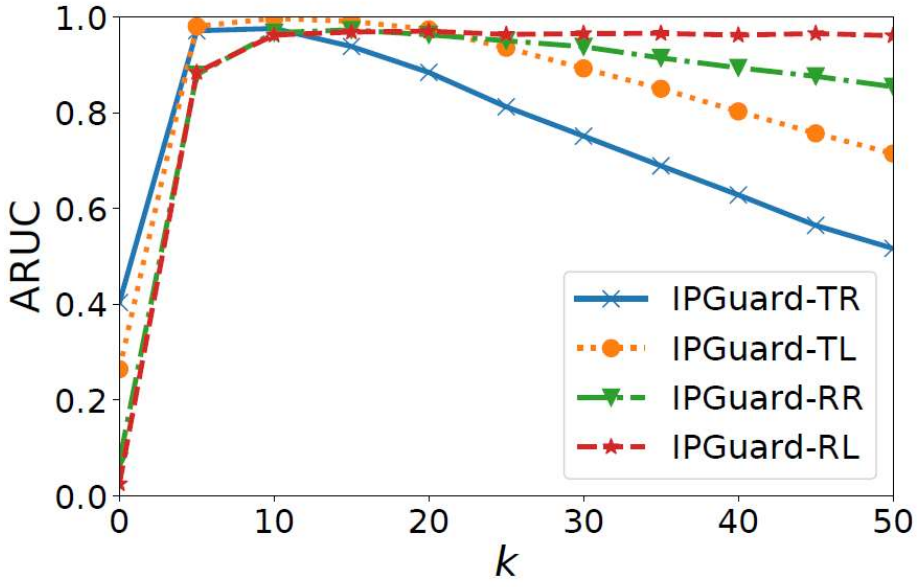
Area under the Robustness-Uniqueness Curves (ARUC)

EXPERIMENTS

- **Impact of k on ARUC for IPGuard**



(a) CIFAR-10



(b) CIFAR-100