MetaAdvDet: Towards Robust Detection of Evolving Adversarial Attacks

Chen Ma[1], Chenxu Zhao[2], Hailin Shi[2], Li Chen[1], Junhai Yong[1], Dan Zeng[3]

[1] School of Software, Tsinghua University, Beijing, China

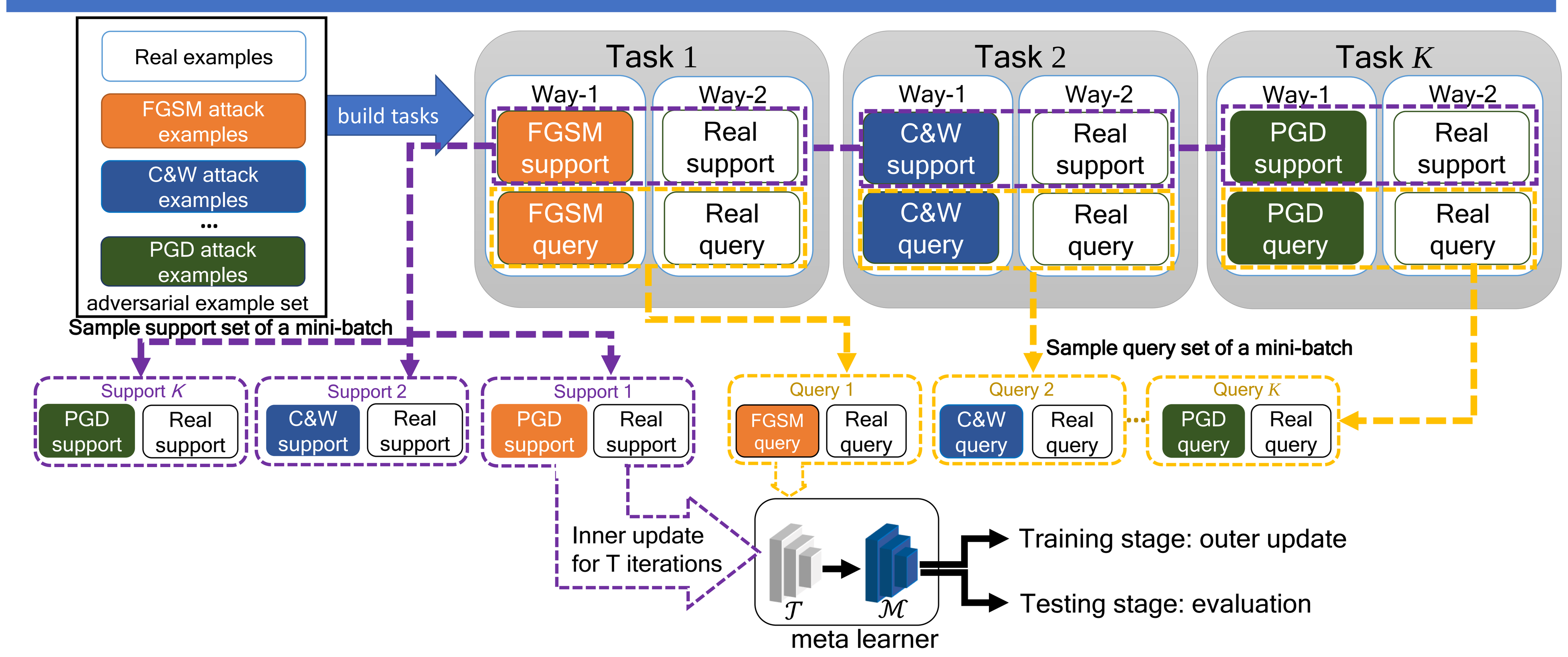
[2] JD AI research [3] Shanghai University

Motivations

The shortcomings of existing approaches for detecting the evolving adversarial attacks:

- Labeled samples of new attacks are insufficient and expensive.
- New attacks evolve much faster than the high-cost data collection.
- The small scale of data leads to the few-shot learning problem → MetaAdvDet is proposed.

Learning from the few-shot tasks



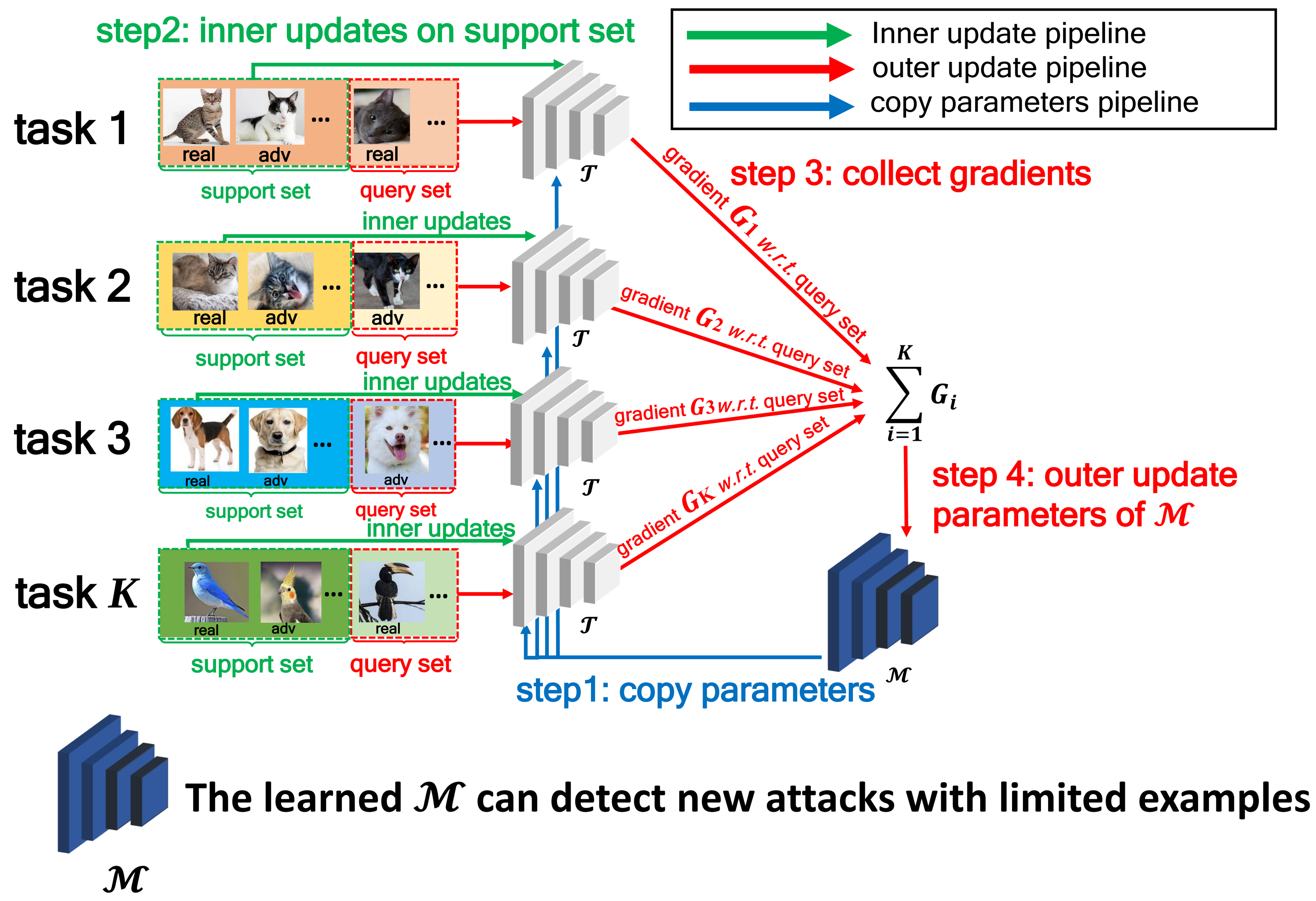
Proposed Benchmark

Benchmark	Test Protocols	
Datasets	CIFAR-10, MNIST and FashionMNIST	
Cross-Adversary Benchmark (simulate the situation of evolving attacks)	**Train Adversary** FGSM, MI-FGSM, BIM, PGD, C&W, JSMA, SPSA, VAT, MaxConfidence	**Test Adversary** EAD, semantic, DeepFool, Spatial Transformation, NewtonFool
Cross-Domain Benchmark	**Train Domain** MNIST FashionMNIST	**Test Domain** FashionMNIST MNIST
Cross-Architecture Benchmark (evaluate the detection of adversarial examples with new architecture)	**Train Architecture** ResNet-10 ResNet-18 Conv-4 ResNet-10	**Test Architecture** ResNet-18 ResNet-10 ResNet-10 Conv-4

Train Domain	Test Domain	Method	F1 score 1-shot	F1 score 5-shot
AdvMNIST	AdvFashionMNIST	DNN (balanced)	0.698	0.813
		NeuralFP [8]	0.748	0.811
		TransformDet [45]	0.664	0.808
		MetaAdvDet (ours)	**0.799**	**0.870**
AdvFashionMNIST	AdvMNIST	DNN (balanced)	0.950	0.977
		NeuralFP [8]	0.775	0.836
		TransformDet [45]	0.934	0.940
		MetaAdvDet (ours)	**0.956**	**0.981**

Results of cross-domain benchmark

Method



MetaAdvDet is equipped with a double-network framework $\mathcal{M}$ and $\mathcal{T}$. $\mathcal{T}$ focuses on learning from individual tasks. After a couple of iterations, $\mathcal{T}$ converges and computes the gradient G_i which are accumulated by $\mathcal{M}$ to update $\mathcal{M}$'s parameters for achieving the fast adaption capability in detecting new attacks.

Results

Dataset	Method	F1 score 1-shot	F1 score 5-shot
AdvCIFAR	DNN	0.495	0.639
	DNN (balanced)	0.536	0.643
	NeuralFP [8]	**0.698**	0.700
	TransformDet [45]	0.662	0.697
	MetaAdvDet (ours)	0.685	**0.791**
AdvMNIST	DNN	0.812	0.852
	DNN (balanced)	0.797	0.808
	NeuralFP [8]	0.780	0.906
	TransformDet [45]	0.840	0.904
	MetaAdvDet (ours)	**0.987**	**0.993**
AdvFashionMNIST	DNN	0.782	0.885
	DNN (balanced)	0.744	0.850
	NeuralFP [8]	0.798	0.817
	TransformDet [45]	0.712	0.879
	MetaAdvDet (ours)	**0.848**	**0.944**

Results of cross-adversary benchmark

Dataset	Method	I-FGSM Attack 1-shot	I-FGSM Attack 5-shot	C&W Attack 1-shot	C&W Attack 5-shot
CIFAR-10	DNN (balanced)	0.466	0.537	0.459	0.527
	TransformDet [45]	**0.593**	**0.728**	0.443	0.502
	MetaAdvDet (ours)	0.553	0.633	**0.548**	**0.607**
MNIST	DNN (balanced)	0.857	0.956	0.814	0.913
	TransformDet [45]	0.864	0.952	0.775	0.893
	MetaAdvDet (ours)	**0.968**	**0.994**	**0.920**	**0.990**
FashionMNIST	DNN (balanced)	0.745	0.890	0.726	0.853
	TransformDet [45]	0.837	0.920	0.747	0.853
	MetaAdvDet (ours)	**0.849**	**0.963**	**0.882**	**0.967**

Results of white-box attack benchmark