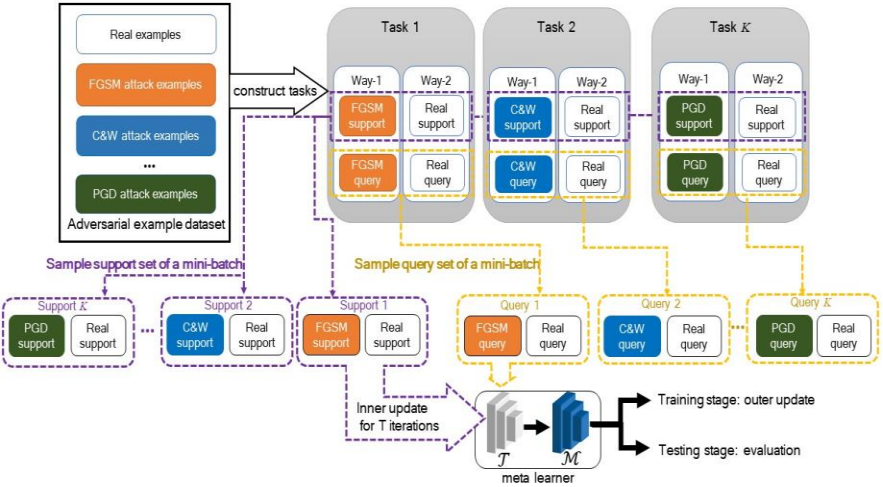
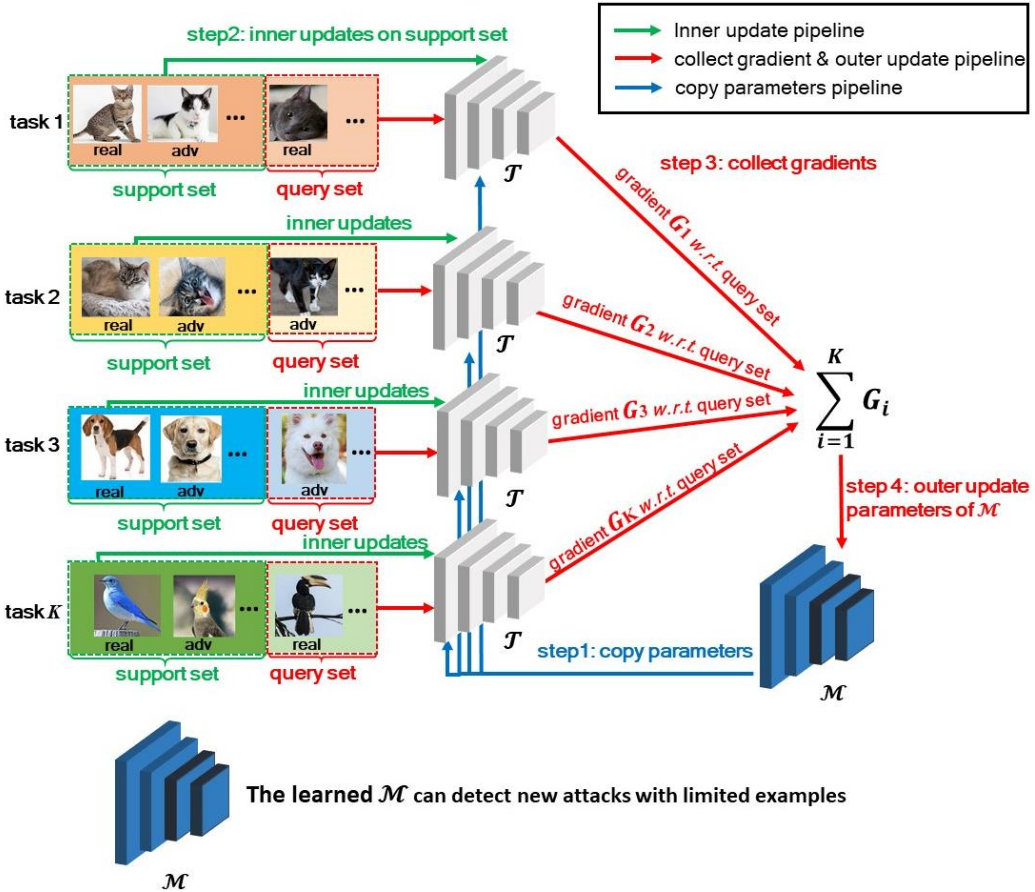
MetaAdvDet: Towards Robust Detection of Evolving Adversarial Attacks

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

Paper ID: **P2B-05**





MetaAdvDet uses the meta-learning based method with a double-network framework to learn the capability of detecting the evolving adversarial attacks.

Paper ID: **P2B-05**

Results

Table 1: cross-adversary benchmark

Dataset	Method	F1 score	
		1-shot	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**

Paper ID: **P2B-05**

Table 2: cross-domain benchmark

Train Domain	Test Domain	Method	F1 score	
			1-shot	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**

Table 3: white-box benchmark

Dataset	Method	I-FGSM Attack		C&W Attack	
		1-shot	5-shot	1-shot	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**

Thanks!

Paper ID: **P2B-05**