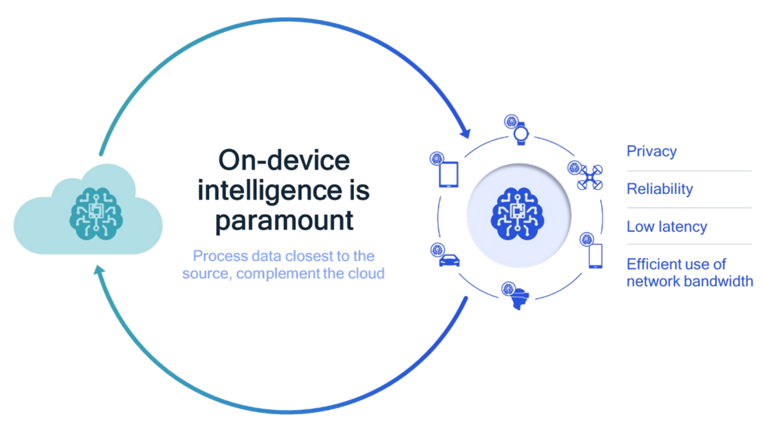
Benchmarking Adversarial Robustness of Compressed Deep Learning Models

Brijesh Vora*, Kartik Patwari*, Syed Hafiz, Zubair Shafiq, Chen-Nee Chuah

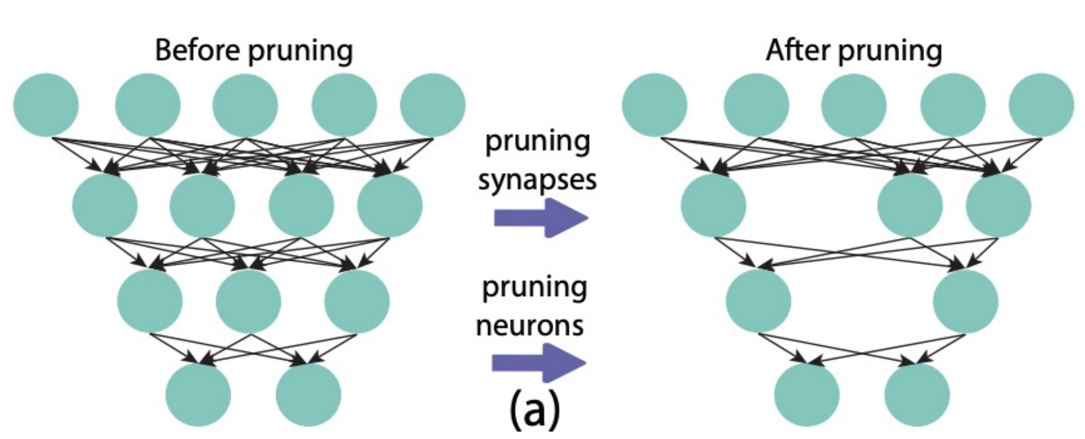
bhvora@ucdavis.edu , kpatwari@ucdavis.edu

Pre-Print

Problem Motivation



- Model compression techniques are proposed to fit deep learning (DL) models on resource-constrained CPU/GPU edge devices to support on-device deep learning for a variety of applications. Trade-off exists between model complexity and performance.
- Research adversarial robustness characterizes and attempts to limit the effect of adversarial attacks on DL models.
- Impact of adversarial attacks on compressed DL models are less explored.
- **Our goal**: build a rigorous benchmarking pipeline to characterize the effect of adversarial attacks with inputs crafted for base models on the pruned versions



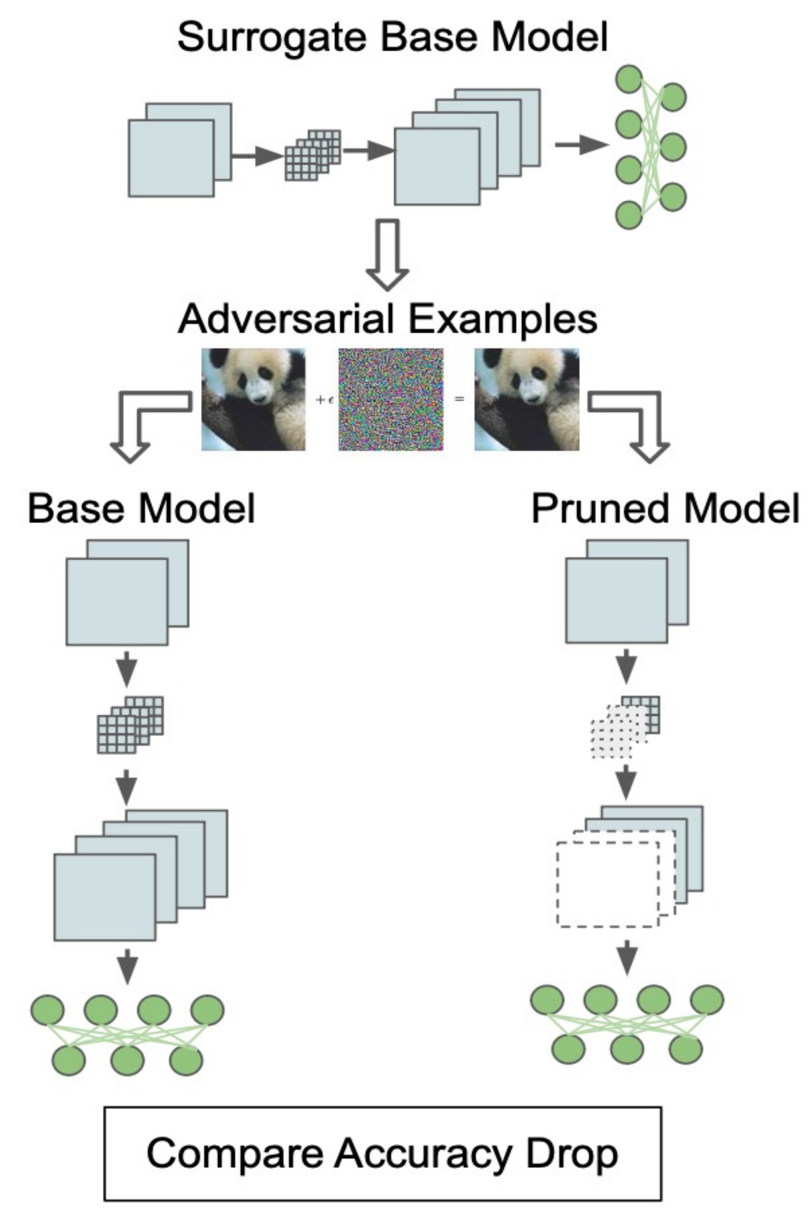


Figure 1. Benchmark pipeline. Adversarial examples generated by various attacks from (attacker's) surrogate base models and evaluated on (victim's) base and pruned model.

Results & Impact

- The pruned models exhibit adversarial robustness comparable to their original base models even in a cross-model fitting environment.
- Compressing the model by up to 50% through filter pruning, the adversarial robustness remains relatively unaffected as compared to the base models.

Model	Inference Time (ms) Base	10%	30%	50%
VGG19	113.80 ± 7.81	104.95 ± 8.81	101.13 ± 6.12	103.65 ± 5.56
RN50	95.07 ± 4.82	88.35 ± 4.29	89.40 ± 5.71	81.41 ± 3.82
MN	23.48 ± 2.46	20.86 ± 2.08	20.39 ± 1.41	21.46 ± 1.97
DN121	71.13 ± 5.27	65.10 ± 3.80	61.09 ± 4.17	60.55 ± 4.05

Pruning Type	Pruning %	Benign Test Accuracy MN	DN121	RN50	VGG19
Base	0%	79.2	83.3	78.1	79.6
L2	10%	**83.7**	**86.9**	80.6	**82.1**
L2	20%	83.5	86.2	79	81.6
L2	30%	81.8	86.7	**82**	81.3
L2	40%	81.1	85.2	74.9	81.1
L2	50%	80.0	81.3	70.4	82.1

Methodology

- Threat Model: Untargeted Attack Scenario, White-Box Adversary, Adversarial perturbations confined within the bounds of Lp norm.
- Image Classification Datasets: Cifar10 (10 classes), Cifar100 (100 classes) with 50k train and 10k test images.
- Popular Computer Vision DL Models from Keras/TensorFlow: MobileNetv1, DenseNet121, ResNet50, VGG19.
- Adversarial attacks from IBM ART: FastSignGradientMethod '14, Deepfool '15, CarliniWagner '16, BasicIterativeMethod '16, UniversalPerturbation '16, AutoPGD '20, ProjectedGradientMethod '17.
- Filter Importance Criteria based on Intel NNCF and filter pruning: L1, L2, Geometric Norm.
- Intel NNCF Pruning: 10 - 50% (Iterative Magnitude Pruning).

Attack	Base	L2-Pruned 10%	20%	30%	40%	50%	max δ
MobileNet							
CW	67.5	65.8	64.9	65.2	63.6	65.2	-1.7
DF	40.7	**43**	**43**	42.3	41.8	41.8	**2.3**
FGSM	12	12.4	12.3	11.3	**12.8**	11.1	**0.8**
BiM	5.1	4.1	3.5	3.8	3.9	**5.6**	**0.5**
PGD	7.4	4.1	4.6	4.2	4.7	6.7	-0.7
APGD	8.5	3.7	4.1	4.4	5.8	8	-0.5
UP	58.8	**64.9**	63	60.1	60.3	59	**6.1**
DenseNet121							
CW	72.1	70	68.1	69.9	69.1	66.9	-2.1
DF	38.7	38.1	38.2	**38.8**	37.8	35.6	**0.1**
FGSM	11.3	11.6	**12**	10.9	10.6	10.3	**0.7**
BiM	8	6.5	6.1	6.7	6.5	6.7	-1.3
PGD	7.1	6.7	6.8	6.8	6.7	7.1	0
APGD	4.5	3.9	4	3.8	3.6	3.8	-0.5
UP	10.2	10	10.6	**11.7**	10.3	10	**1.5**
ResNet50							
CW	68.3	71.4	69.4	**72.7**	65	60.4	**4.4**
DF	37.1	39.7	39.7	**40.4**	35.9	33.8	**3.3**
FGSM	15.4	11.9	13.6	13.6	13.2	13	-1.8
BiM	7	7.2	**7.6**	7.4	7.4	6.9	**0.6**
PGD	7.8	7.5	7.5	7.5	7.1	**8.3**	**0.5**
APGD	4.9	3.9	4.2	4.1	4.4	**6.5**	**1.6**
UP	15.6	14.1	11.8	14.9	11.4	11	-0.7

Summary

- Our findings reveal that while the benefits of pruning – enhanced generalizability, compression, and faster inference times – are preserved, adversarial robustness remains comparable to the base model.

Challenges & Opportunities

- Devise a compression algorithm that produces an optimized model which is equally (if not more) robust to most (if not all) of the adversarial attacks.
- Root cause analysis of diverse effect of compression techniques against adversarial attacks.
- Build secure ML models from scratch – utilizing neural architecture search – that are provably robust to an ensemble of adversarial attacks, attain better accuracy, and consume less energy tailored for resource-constrained devices.
- Investigate novel adversarial attack that is invariant to compression techniques.

References

- Adversarial Robustness Toolbox IBM: ART-IBM
- Intel Neural Network Compression Framework (NNCF): NNCF
- Image sources: OnDeviceLearning, CompressingMLModel

Acknowledgement

- This project is based upon work supported in part by the UC Noyce Institute: Center for Cybersecurity and Cyber-integrity (C-CUBE).