2-in-1 Accelerator: Enabling Random Precision Switch for Winning Both Adversarial Robustness and Efficiency

Yonggan Fu, Yang Zhao, Qixuan Yu, Chaojian Li, and Yingyan (Celine) Lin

Georgia Institute of Technology

Georgia Tech College of Computing
Center for Research into Novel Computing Hierarchies

Accepted by MICRO 2021

Background & Motivation

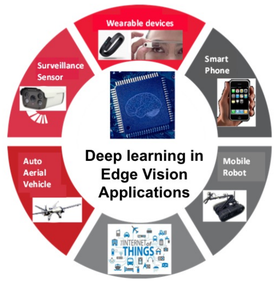
Driving forces for deep neural network (DNN) breakthroughs

- ✓ Big computer
- ✓ Big data

Challenges for deploying DNNs on intelligent edge devices

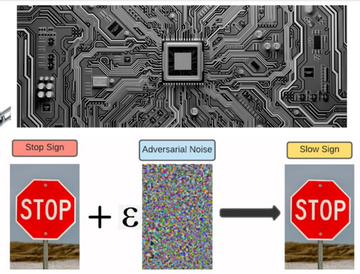
- ➢ Limited computing resources
- ➢ Malicious data in the wild

***Efficiency concern*: Limited on-device resources**



Edge Vision Applications

Broken Link



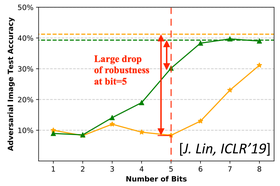
***Security concern*: Risks of malicious attacks**

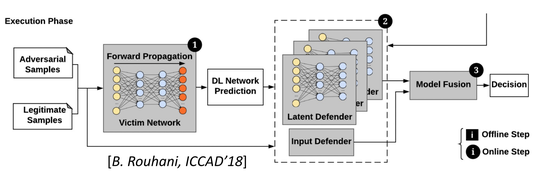
❓ ***Can we win both efficiency and robustness in edge vision applications?***

Limitations of Previous Works

☹ **Algorithm side:** Robustness and efficiency are often contradictory

- ➢ Overparameterization is crucial for robustness [*A. Madry, ICLR'18*]
- ➢ Quantized DNNs are inferior in robustness [*J. Lin, ICLR'19*]





[*B. Rouhani, ICCAD'18*]

☹ **Hardware side:** Robustness is achieved at the cost of efficiency

- ➢ Extra detection modules are needed: Deepfense [*B. Rouhani, ICCAD'18*], DNNGuard [*Z. Wang, ASPLOS'20*], Ptolemy [Y. Gan, MICRO'20]

2-in-1 Accelerator: Overview

2-in-1 Accelerator: Algorithm-Architecture Co-Design to Win both Efficiency and Adversarial Robustness

- ➢ ***Algorithm***: Random Precision Switch to enable random DNN quantization as an in-situ model switch
- ➢ ***Architecture***: A novel precision-scalable accelerator featuring a spatial-temporal MAC array with bit-level optimization

2-in-1 Accelerator: Algorithm

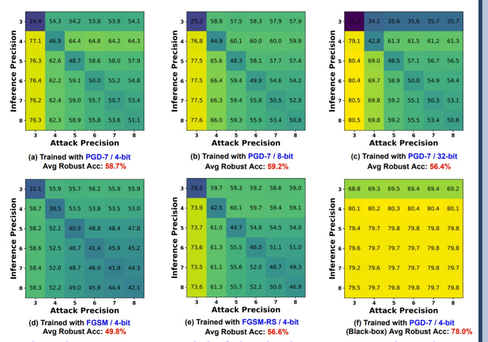
Motivating observation: *Poor Adv. Transferability between Precisions*

Visualize the transferability

- ➢ Robust accuracy measured under varied attack and inference precision pairs

Key observation

- ➢ Consistently poor adversarial transferability btw. precisions across different adversarial training methods



Key insight: Adversarial perturbations are shielded by quantization noises which cannot be effectively learned by gradient-based attacks

The Inspired Technique: *Random Precision Switch (RPS)*

Technique 1: RPS Inference

- ➢ Randomly select one inference precision
- ✓ *Relatively stable natural accuracy*
- ✓ *Degraded attack effectiveness*

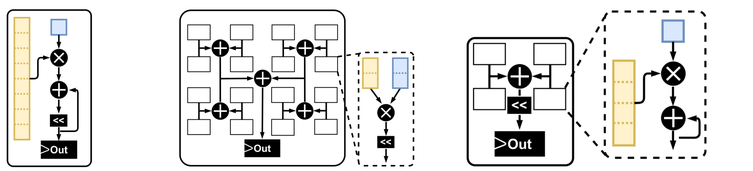
Technique 2: RPS Training

- ➢ Equip the model with switchable BN
- ✓ *Increase the difficulty of adv. transfer*
- ➢ Randomly select one precision for training

```
=== RPS Training ===
Equip f_θ with SBN
for epoch ∈ [1, T] do
    for (x, y) ∈ D_train do
        Randomly select a precision q from Set_Q
        Obtain f_θ^q by quantizing f_θ to q-bit
        δ = 0 or random initialized
        for t ∈ [1, 7] do
            δ = clip_ε{δ + α · sign(∇_δ ℓ(f_θ^q(x + δ), y))}
        end for
        θ = θ − ∇_θ ℓ(f_θ^q(x + δ), y)
    end for
end for
=== RPS Inference ===
for x_adv ∈ D_adv do
    Randomly select a precision q from Set_Q
    Obtain f_θ^q by quantizing f_θ to q-bit
    Evaluate ŷ = f_θ^q(x_adv)
end for
```

2-in-1 Accelerator: Architecture

2-in-1 Accelerator's Architecture: *A Spatial-Temporal MAC Array*



Temporal Design

- ➢ *Inferior area-efficiency: Shift-add can take 90% of the total area in a 16-bit MAC unit*

Spatial Design

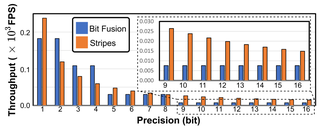
- ➢ *Inferior flexibility: Only support a limited set of precisions, e.g., 2/4/8-bit in Bit Fusion [H. Sharma, ISCA'18]*

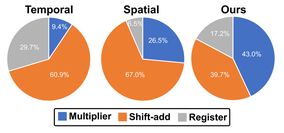
Our Spatial-Temporal Design

- ✓ ***Better area-efficiency:*** *Bit-serial units only need to support up to 4-bit*
- ✓ ***High flexibility:*** *Support irregular precision choices like 3/6-bit*

The dilemma between performance vs. flexibility

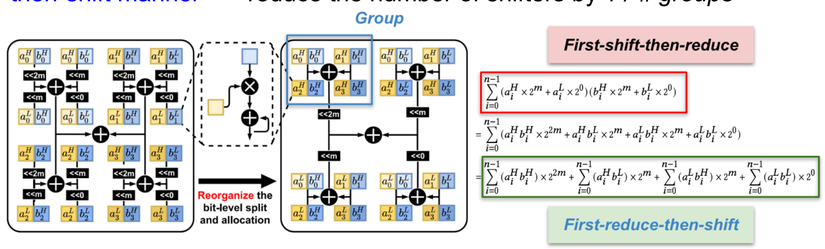
- ➢ SOTA temporal design Strips [*P. Judd, MICRO'16*] vs. spatial design Bit Fusion





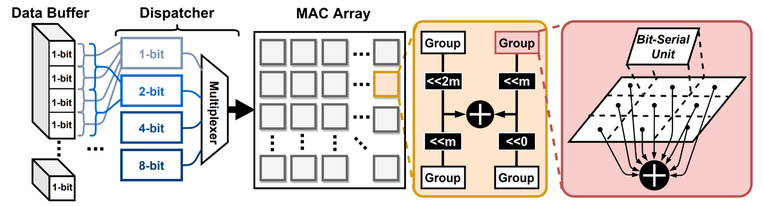
2-in-1 Accelerator's Architecture: *Bit-level Reorganization & Fusion*

Technique 1: Reorganize the bit-level split and allocation in a first-reduce-then-shift manner ➡ reduce the number of shifters by *1 / # groups*



Technique 2: Fuse the shift-add of bit-serial units in one group

- ➢ Further reduce the number of shifters by *1 / # groups*



Our proposed architecture: Achieve 2.3x throughput/area and 4.88x energy-efficiency/operation compared with Bit Fusion

2-in-1 Accelerator: Evaluation

Evaluation 2-in-1 Accelerator's Algorithm

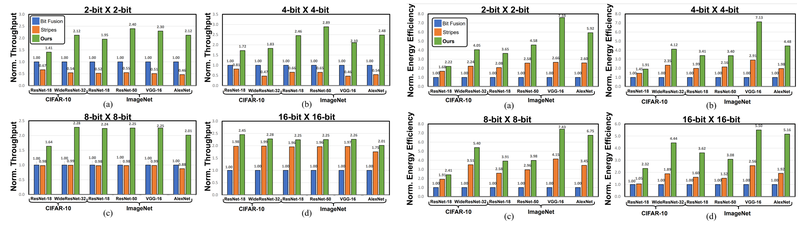
- ✓ Consistently boost adversarial robustness, e.g., +13.98%/+12.14% robust accuracy on two networks with PGD-7 training against PGD-20 attacks

Adversarial Training Method	PreActResNet-18 Natural (%)	PreActResNet-18 PGD-20 (%)	PreActResNet-18 PGD-100 (%)	WideResNet-32 Natural (%)	WideResNet-32 PGD-20 (%)	WideResNet-32 PGD-100 (%)
FGSM	67.04	41.48	41.37	66.76	40.78	40.55
FGSM + **RPS**	80.58	**64.08**	**63.56**	64.09	**50.70**	**48.72**
FGSM-RS	86.08	41.76	41.13	89.95	45.33	44.77
FGSM-RS + **RPS**	82.11	**59.33**	**59.32**	87.87	**60.07**	**59.12**
PGD-7	82.02	51.17	50.93	85.25	54.61	54.36
PGD-7 + **RPS**	82.16	**65.15**	**64.88**	81.52	**66.75**	**66.28**

Attack Type	PreActResNet-18 PGD-7	PreActResNet-18 PGD-7 + **RPS**	WideResNet-32 PGD-7	WideResNet-32 PGD-7 + **RPS**
AutoAttack (ϵ=8)	47.18	**54.56**	51.66	**58.54**
AutoAttack (ϵ=12)	27.59	**35.83**	30.71	**39.83**
CW-Inf (ϵ=8)	57.88	**71.44**	62.13	**72.10**
CW-Inf (ϵ=12)	46.70	**65.57**	50.14	**66.99**
Bandits (ϵ=8)	59.75	**71.75**	63.49	**68.50**
Bandits (ϵ=12)	46.04	**70.52**	49.77	**67.01**

Evaluation 2-in-1 Accelerator's Architecture

- ✓ 1.41x~2.88x/1.15x~4.59x throughput and 1.91x~7.58x/1.25x~2.85x energy efficiency over Bit Fusion/Stripes, respectively



Throughput evolution with precisions

- ✓ Superior performance: Outperform both baselines
- ✓ High Flexibility: Achieve consistent improvements in throughput as the precision decreases