SRQ: Self-reference quantization scheme for lightweight neural network

Xiaobin Li[1,2], Hongxu Jiang[1,2], Shuangxi Huang[1], Fangzheng Tian[1], Runhua Zhang[1] and Dong Dong[1]

[1] Beijing Key Lab Digital Media, State Key Lab Virtual Real Technology & Systems, BeiHang University

[2] Hangzhou Innovation Institute, BeiHang University

Outline

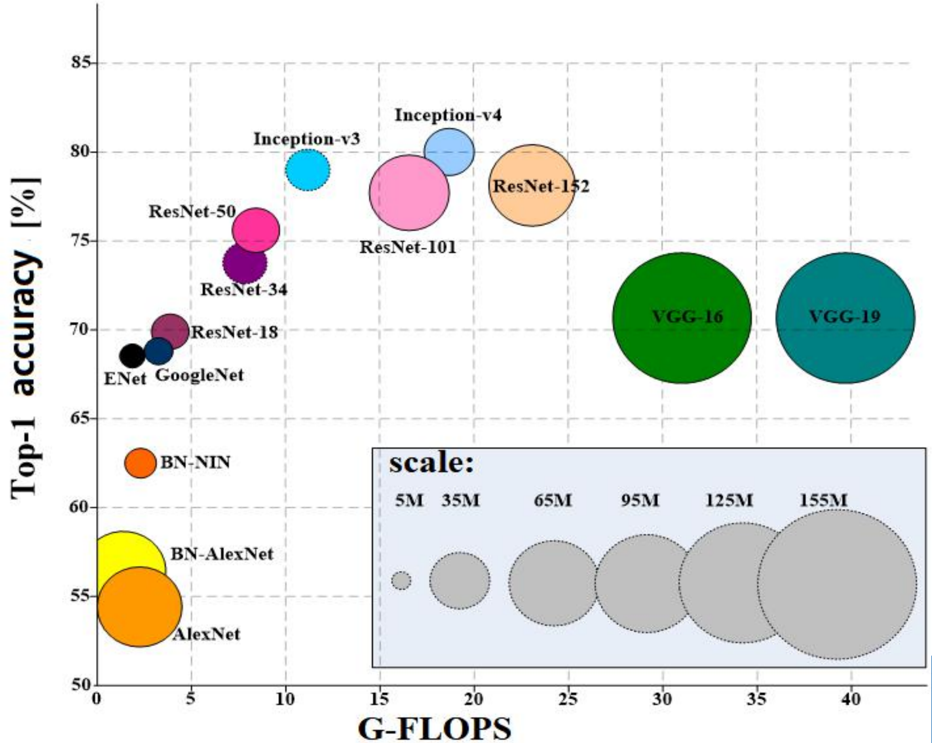
- Background and Motivation
- Problem Statement
- Our Solutions
- Experiments
- Conclusion

Outline

Background

- Deep Neural Network(DNN)
  - Application

    Detection, recognition, verification and other tasks.
  - Representatives

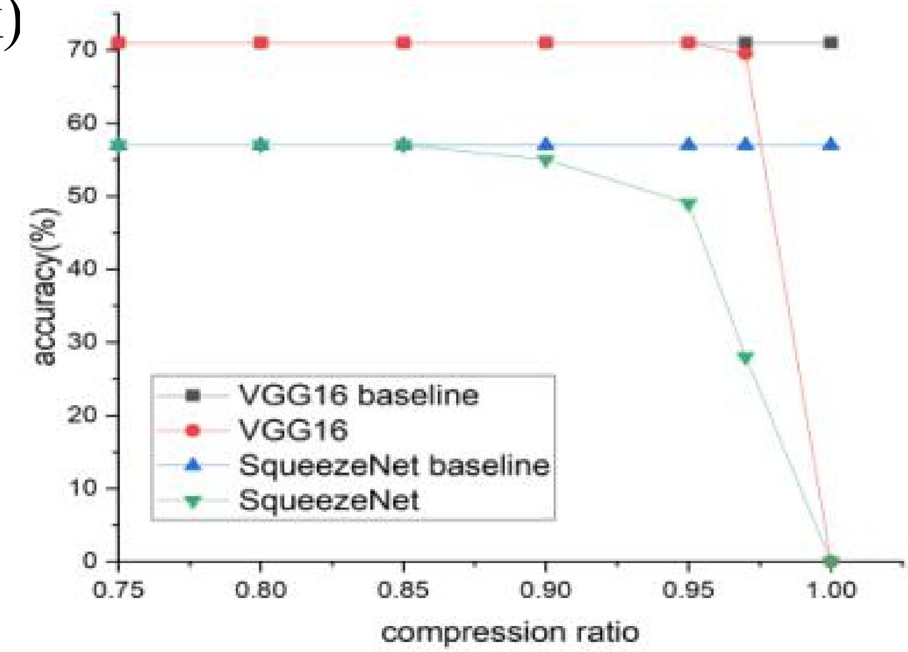


Outline

- Background and Motivation
- Problem Statement
- Our Solutions
- Experiments
- Conclusion

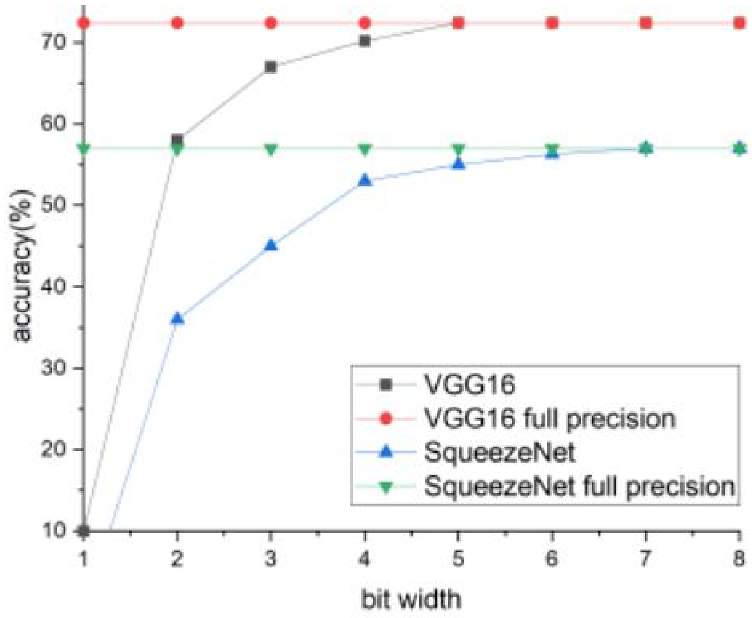
Problem Statement

- Comparative analysis of redundancy
  - The redundancy of SqueezeNet (lightweight neural network) approaches 85% and less than VGG16(classical neural network)

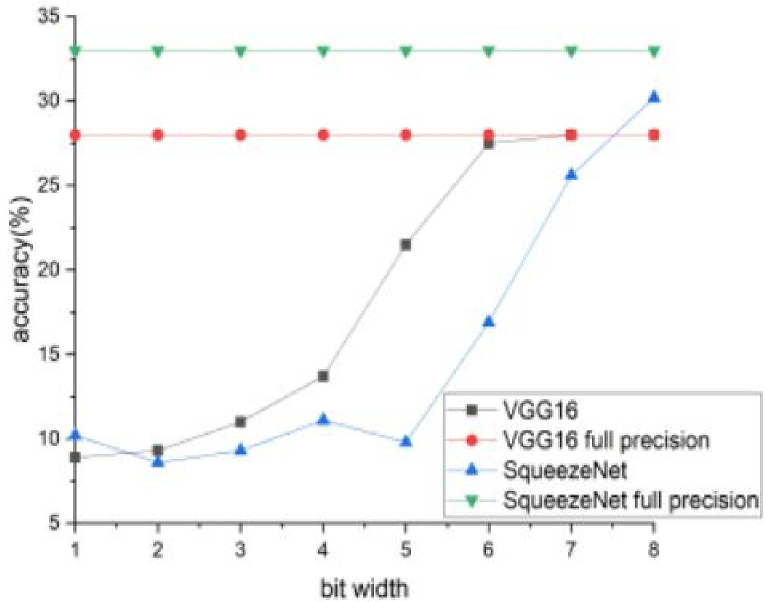


Problem Statement

- Accuracy reduction and robustness analysis
  - Accuracy vs Bit width



(a) With Clean image

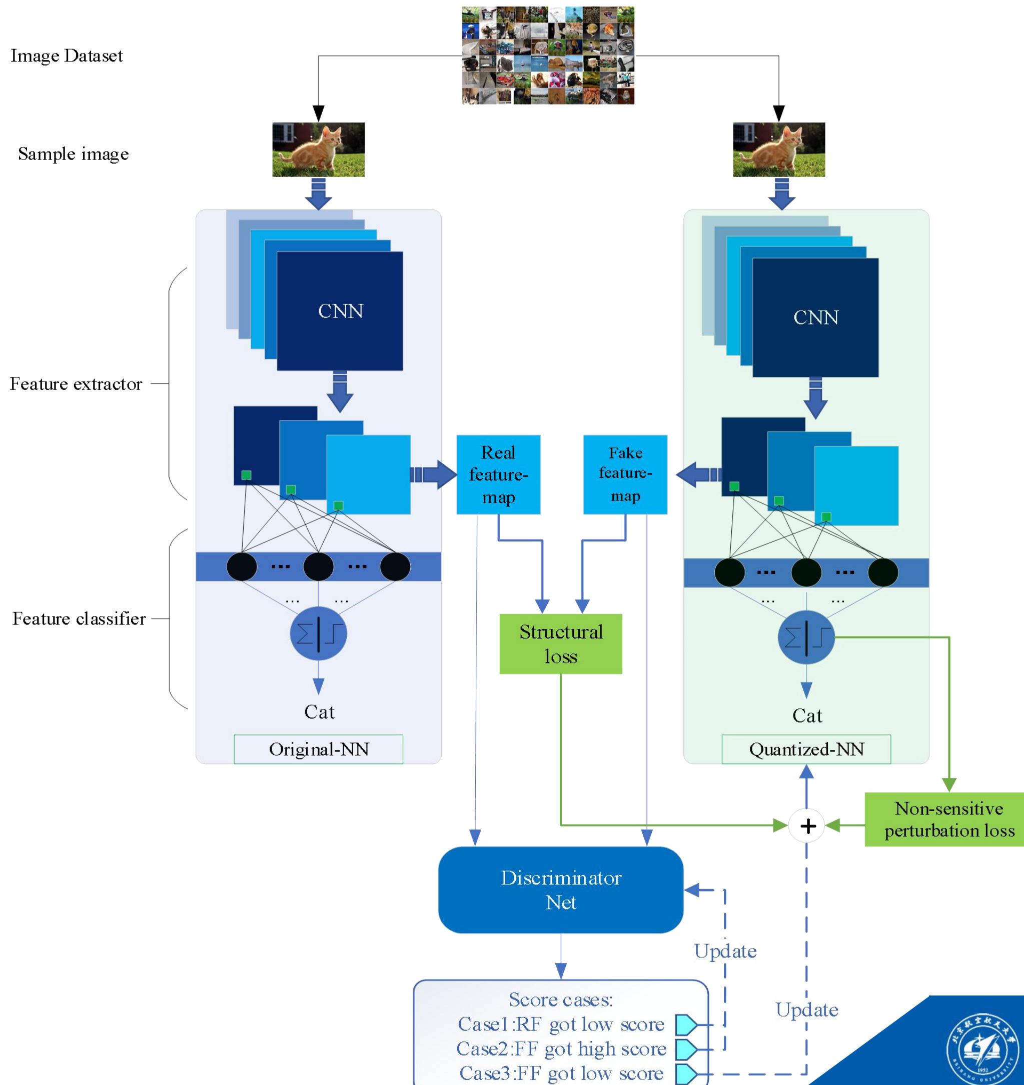


(b) With adversarial image

Outline

- Background and Motivation
- Problem Statement
- Our Solutions
- Experiments
- Conclusion

Our Solutions



Outline

- Background and Motivation
- Problem Statement
- Our Solutions
- Experiments
- Conclusion

Experiments

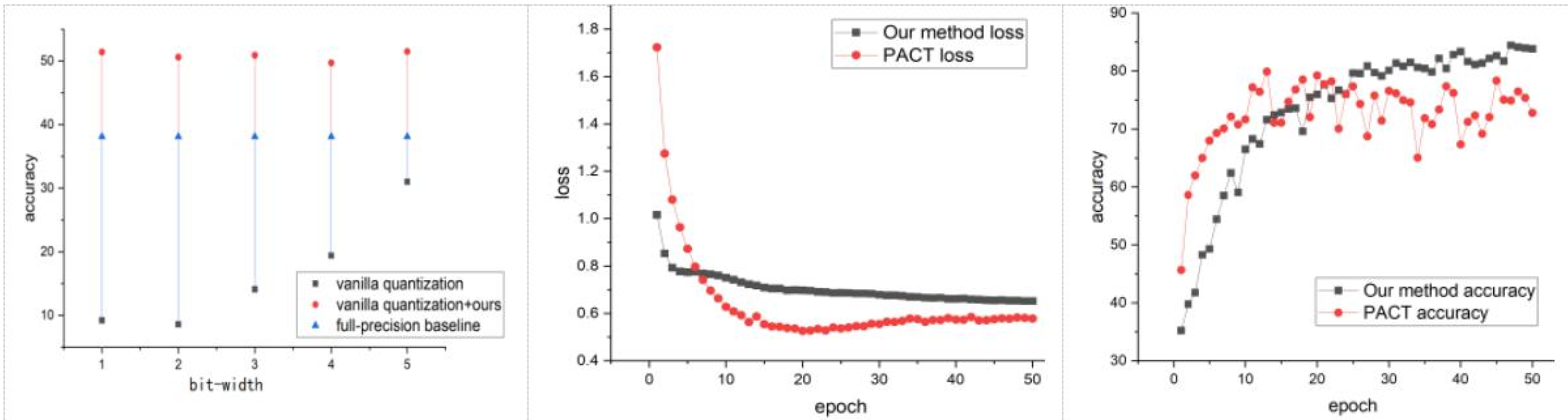
Table 1. The accuracy of quantized ResNet20 and SqueezeNet on CIFAR-10 dataset.

	Method	Full-Precision(%)	Quantization TOP1 Acc.(%)		
ResNet20/SqueezeNet	Bit-width(w/a)	32/32	2/2	3/3	4/4
	DoReFa	91.8/92.5	88.2/83.9	88.8/88.9	89.4/89.0
	PACT		89.2/85.4	89.6/89.3	91.2/89.5
	DoReFa+SRQ		89.3**(+1.1)** /85.9**(+2.0)**	90.1**(+1.3)** /89.9**(+1.0)**	91.1**(+0.7)** /90.4**(+1.4)**
	PACT+SRQ		89.9**(+0.7)** /87.3**(+1.9)**	90.7**(+1.1)** /90.3**(+1.0)**	91.7**(+0.5)** /90.6**(+1.1)**

Table 2. The accuracy of quantized ResNet18 and MoblieNetV2 on ImageNet dataset.

LNN	Method	Full-Precision(%)	Quantization TOP1 Acc.(%)		
ResNet18/MoblieNetV2	Bit-width(w/a)	32/32	2/2	3/3	4/4
	DoReFa	70.4/71.7	62.6/60.9	67.5/63.7	68.1/68.6
	PACT		67.0/61.4	68.1/67.5	69.2/69.6
	DoReFa+SRQ		63.8**(+1.2)**/ 62.6**(+1.7)**	68.3**(+0.8)** /64.9**(+1.2)**	69.5**(+1.4)** /69.7**(+1.1)**
	PACT+SRQ		68.3**(+1.3)**/ 63.7**(+2.3)**	69.1**(+1.0)** /69.4**(+1.9)**	69.9**(+0.7)** /70.1**(+0.5)**

Experiments



(a)Accuracy of ResNet20 | (a)Variation of loss | (c)Variation of accuracy

(a) The robustness-aware advantages of our algorithm on the adversarial image dataset, (b) and (c) show the variation trend of loss and accuracy during quantization progress, respectively.

Outline

- Background and Motivation
- Problem Statement
- Our Solutions
- Experiments
- Conclusion

Conclusion

- Propose a novel robustness-aware self-reference quantization scheme to guarantee the accuracy and robustness during the quantization process.
- As a by-product, we can combine the other excellent quantization methods with our framework to further improve the accuracy and robustness.
- Experimental results show that our approach outperforms to the existing best perform methods.

Q & A

Thank You