Aggressive Compression of MobileNets Using Hybrid Ternary Layers

Dibakar Gope, Jesse Beu, Urmish Thakker, and Matthew Mattina
Arm ML Research Lab

Challenge

- MobileNets [1] family of CV networks are increasingly deployed at mobile/edge devices
- Quantizing MobileNets to ternary weights (2-bit) is necessary to realize siginificant energy savings and runtime speedups

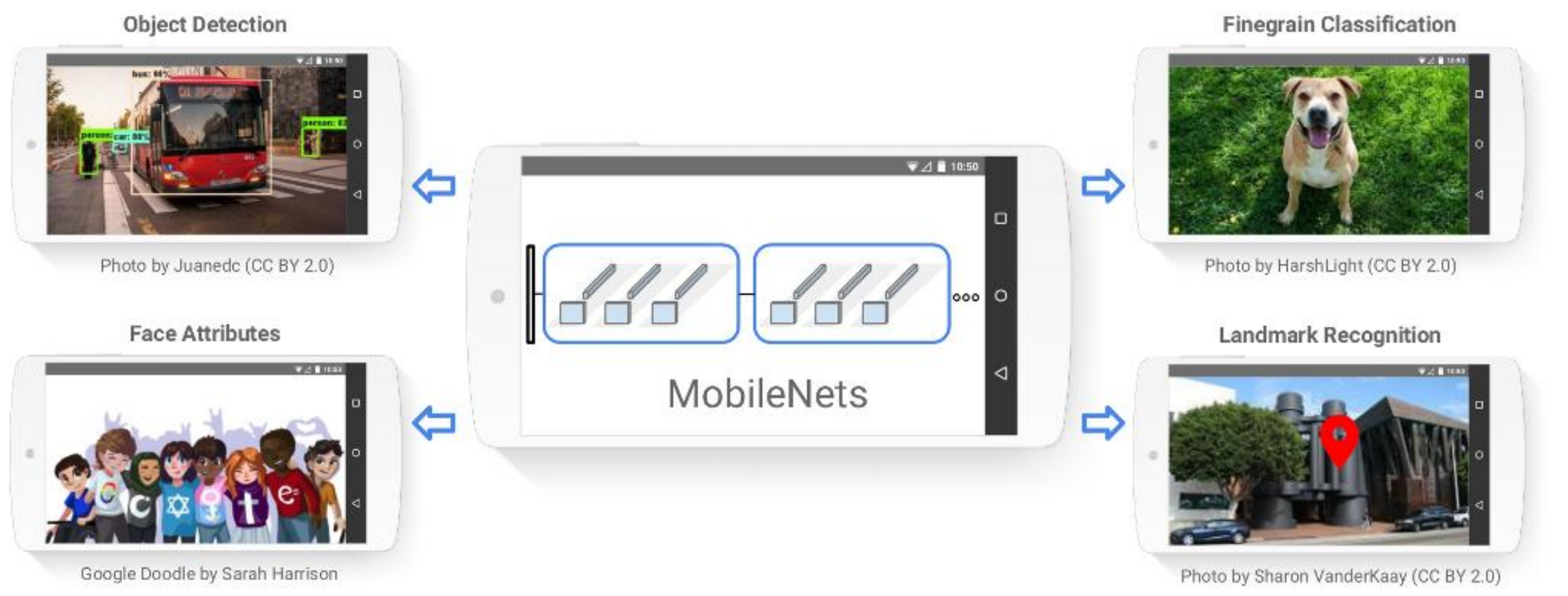
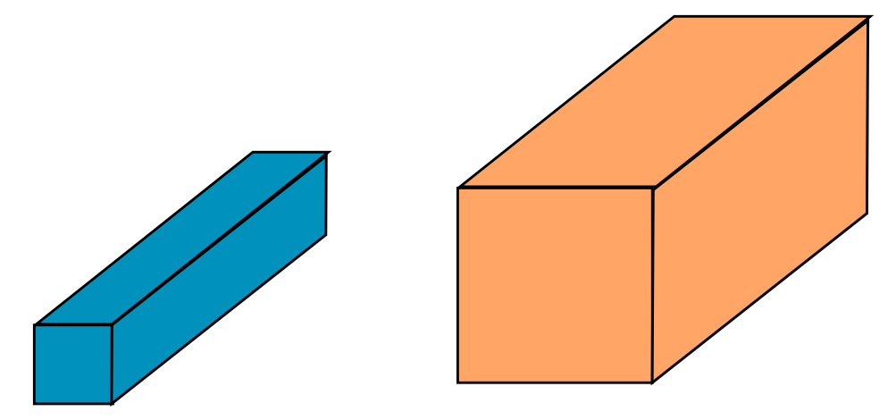

Object Detection
Photo by Juanedc (CC BY 2.0)
Face Attributes
Google Doodle by Sarah Harrison
MobileNets
Finegrain Classification
Photo by HarshLight (CC BY 2.0)
Landmark Recognition
Photo by Sharon VanderKaay (CC BY 2.0)

Photo credit: Google AI blog on MobileNets V1



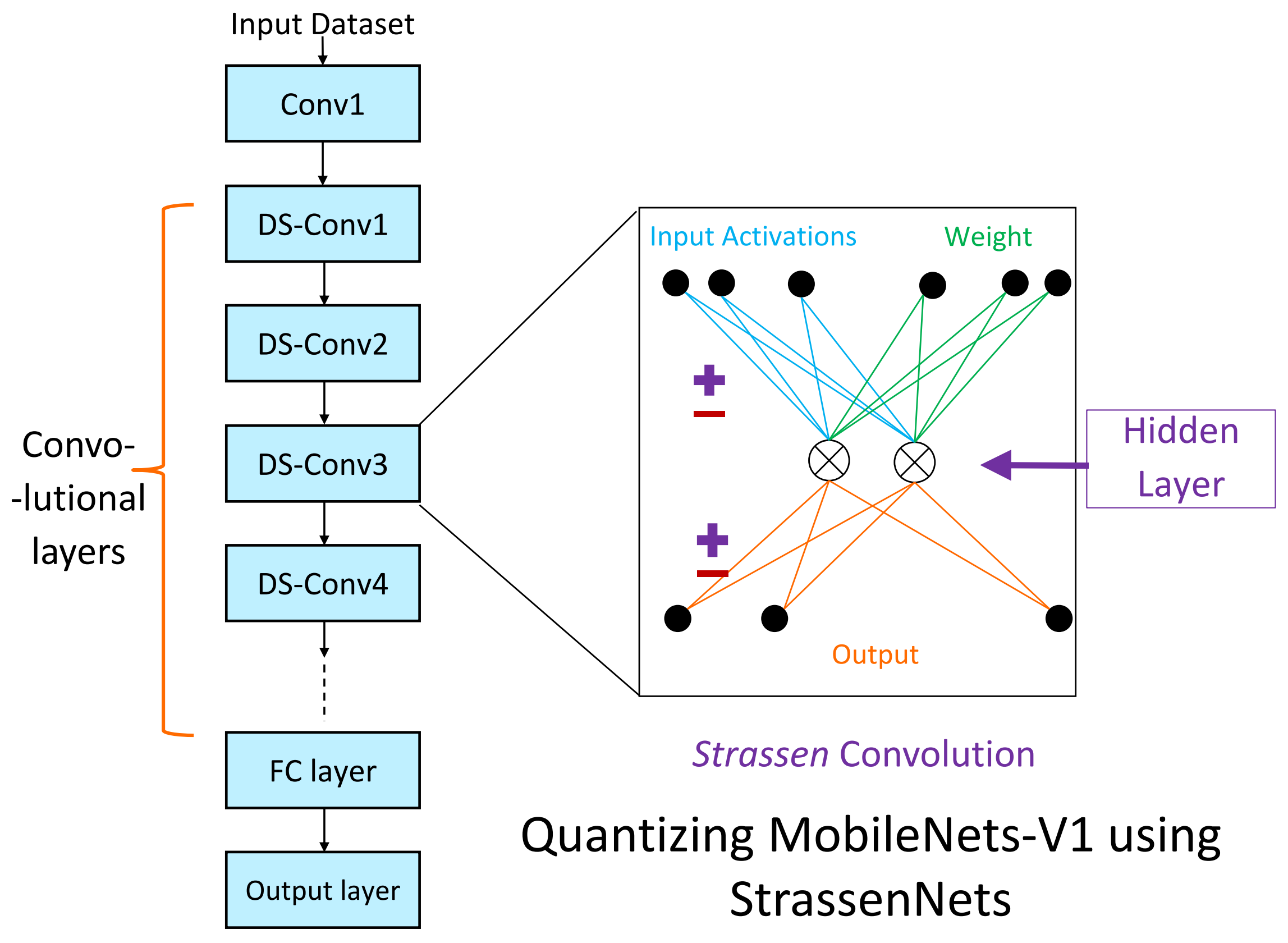
1 x 1 filter — Very compact

3 x 3 filter — Over parametrized

- MobileNets V1 – 13 depthwise separable (DS) convolutional layers
- Model complexity dominated by compact 1 x 1 filters

Prior Solutions

- Ternary weight networks (TWN) [2]
  - (-) Drops accuracy
- StrassenNets [3]
  - (+) 99% reduction in MULs for 3 x 3 filters
  - (+) mostly ternary weights, preserve accuracy
  - (-) Never looked into DS (1 x 1) layers
- Prior solutions come with their own advantages and limitations

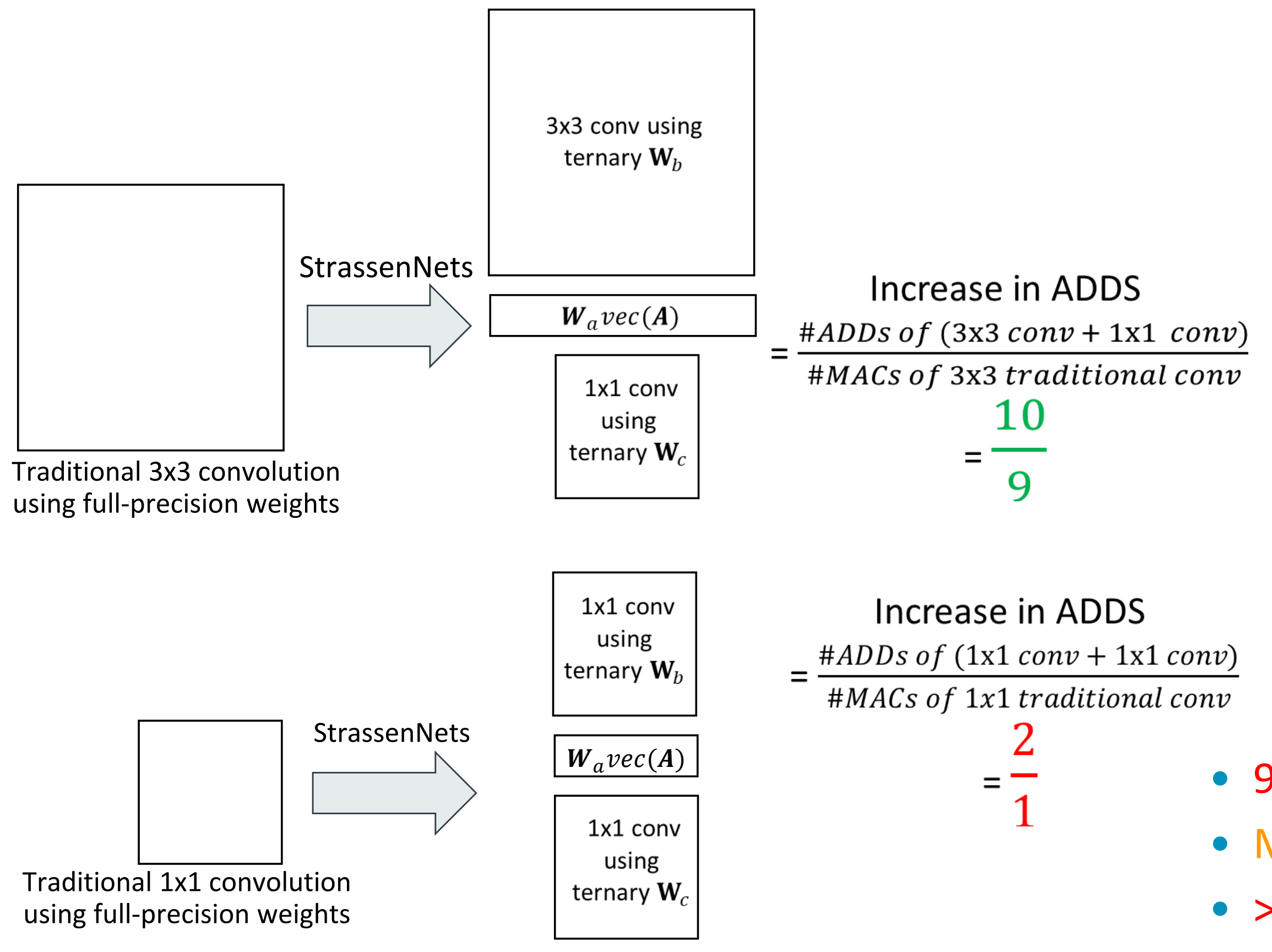


Strassen Convolution

Quantizing MobileNets-V1 using StrassenNets ternary quantization scheme

MobileNets-V1

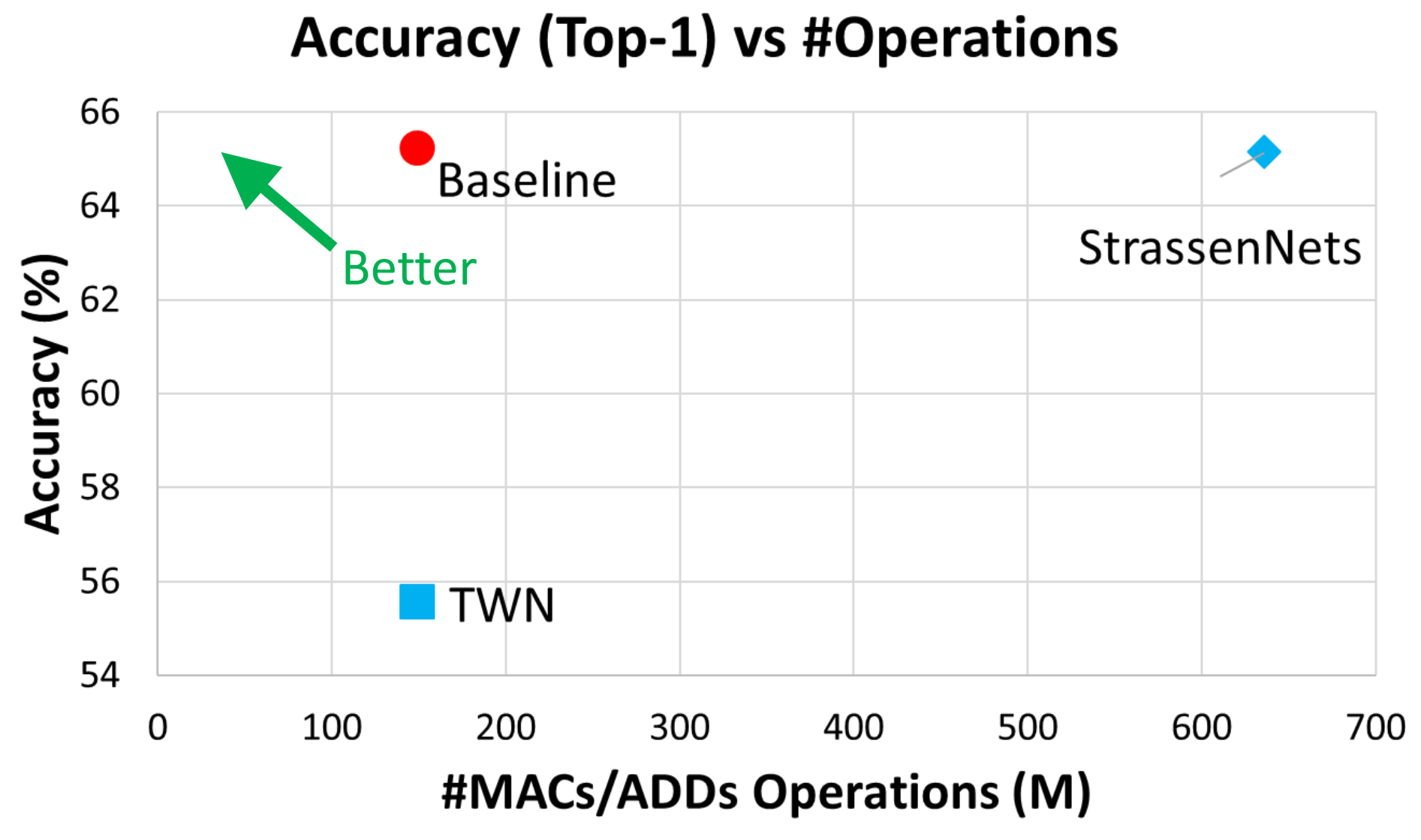
Observations with Prior Solutions



Increase in ADDS
$$= \frac{\#ADDs\ of\ (3x3\ conv + 1x1\ conv)}{\#MACs\ of\ 3x3\ traditional\ conv} = \frac{10}{9}$$

Increase in ADDS
$$= \frac{\#ADDs\ of\ (1x1\ conv + 1x1\ conv)}{\#MACs\ of\ 1x1\ traditional\ conv} = \frac{2}{1}$$

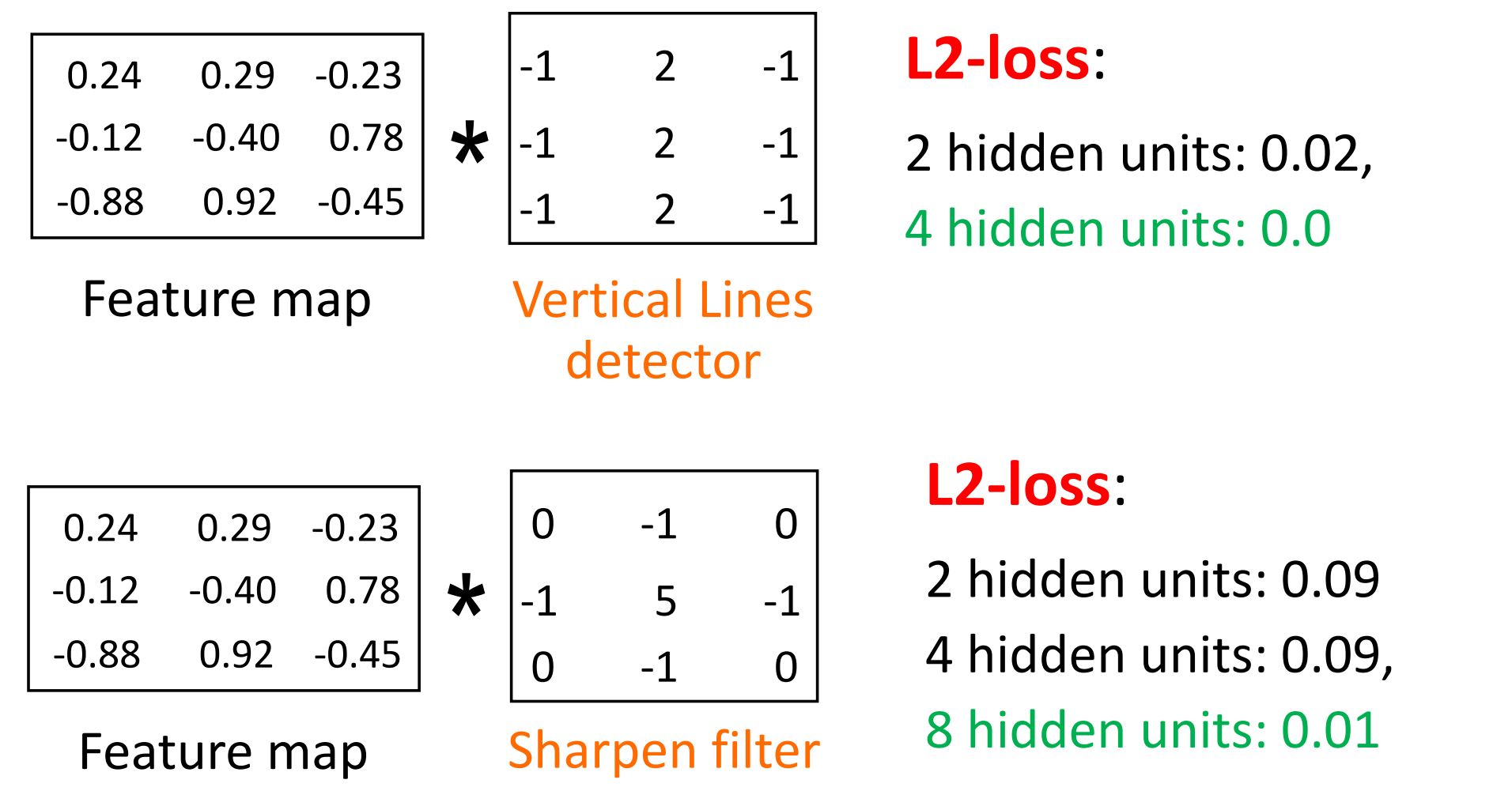
Application of StrassenNets to 3 x 3 and 1 x 1 convolution



- 9.6% drop in accuracy using Ternary Weight Networks
- Modest savings in model size using StrassenNets
- >300% increase in ADDs/Ops for iso accuracy using StrassenNets
- Use of Wide hidden layers for closely approximating each 1 x 1 filter of MobileNets → >300% increase in ADDs using StrassenNets

Different filters respond differently to ternary quantization

Different sensitivity of individual filters to StrassenNets

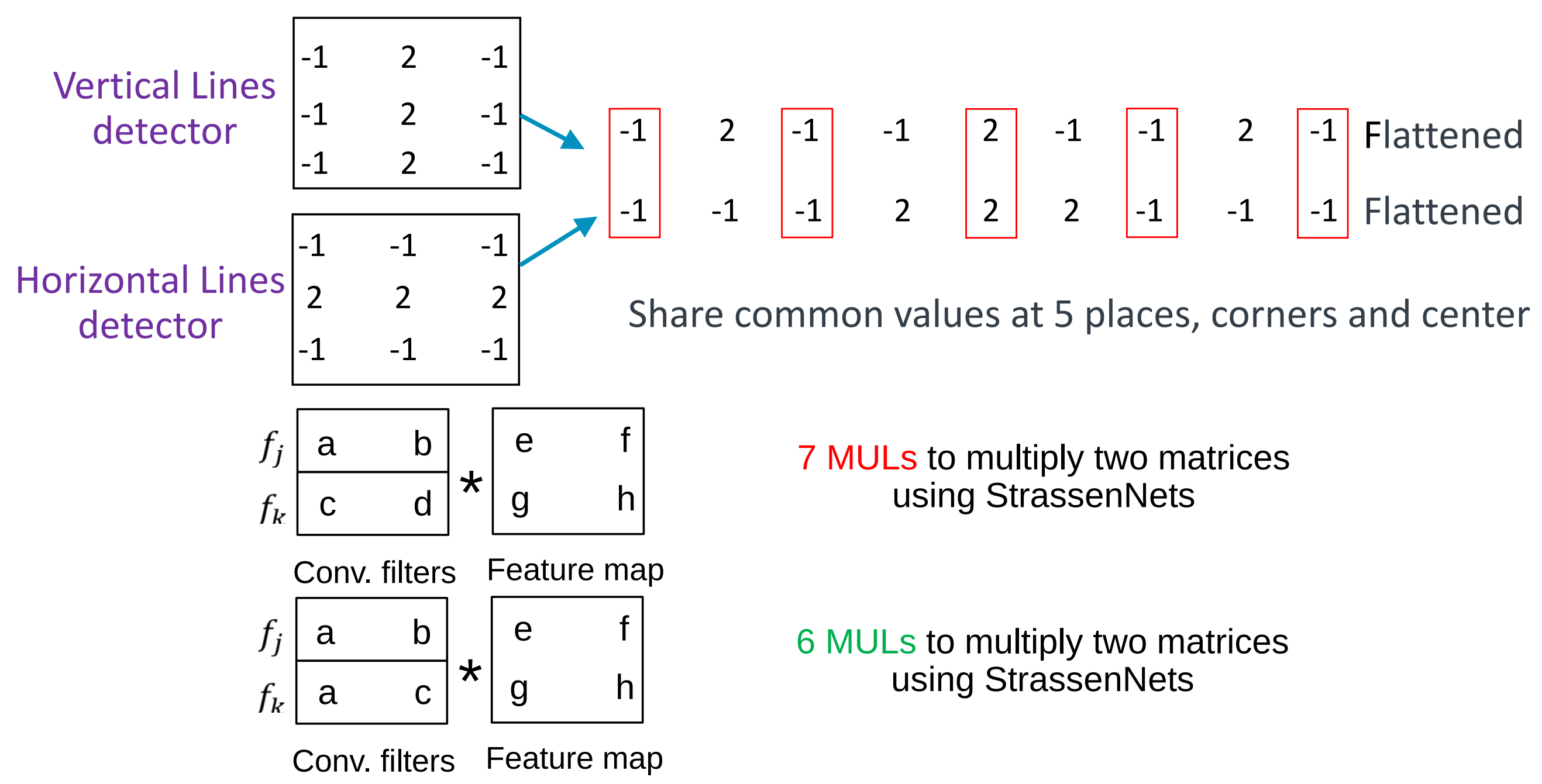


L2-loss:
2 hidden units: 0.02,
4 hidden units: 0.0

L2-loss:
2 hidden units: 0.09
4 hidden units: 0.09,
8 hidden units: 0.01

Not all filters do require wide hidden layers to be approximated well using StrassenNets

Different sensitivity of group of filters to StrassenNets



Share common values at 5 places, corners and center

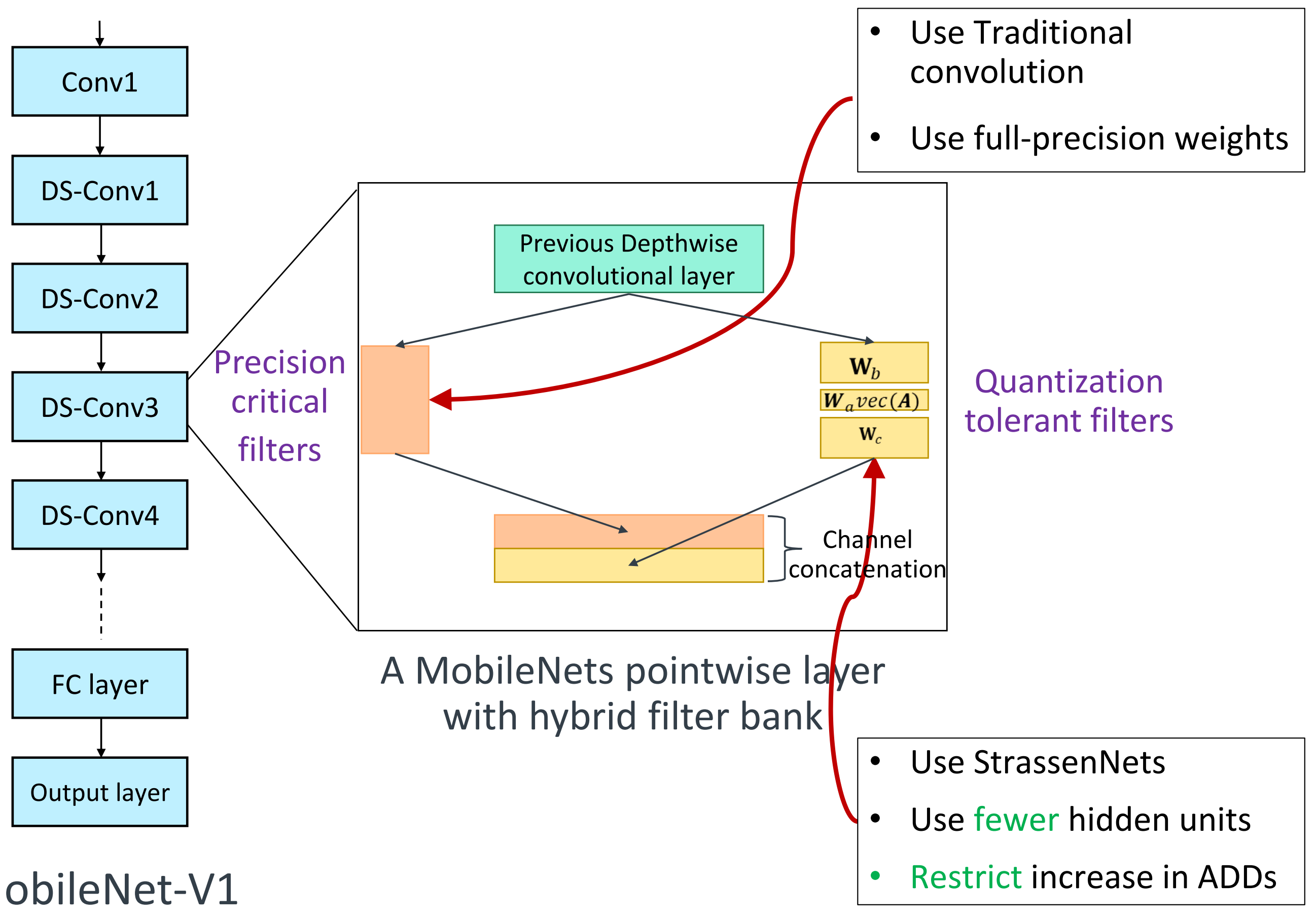
7 MULs to multiply two matrices using StrassenNets

6 MULs to multiply two matrices using StrassenNets

Per-Layer Hybrid Filter Banks

Exploit the difference in sensitivity of individual and groups of filters to ternary quantization

- Bank similar value structure filters together
- Share hidden units of StrassenNets
- Use fewer hidden units → fewer ADDs/Ops to approximate a major portion of filters at each layer
- See our paper (https://arxiv.org/abs/1911.01028) for Mathematical proof, details



- Use Traditional convolution
- Use full-precision weights

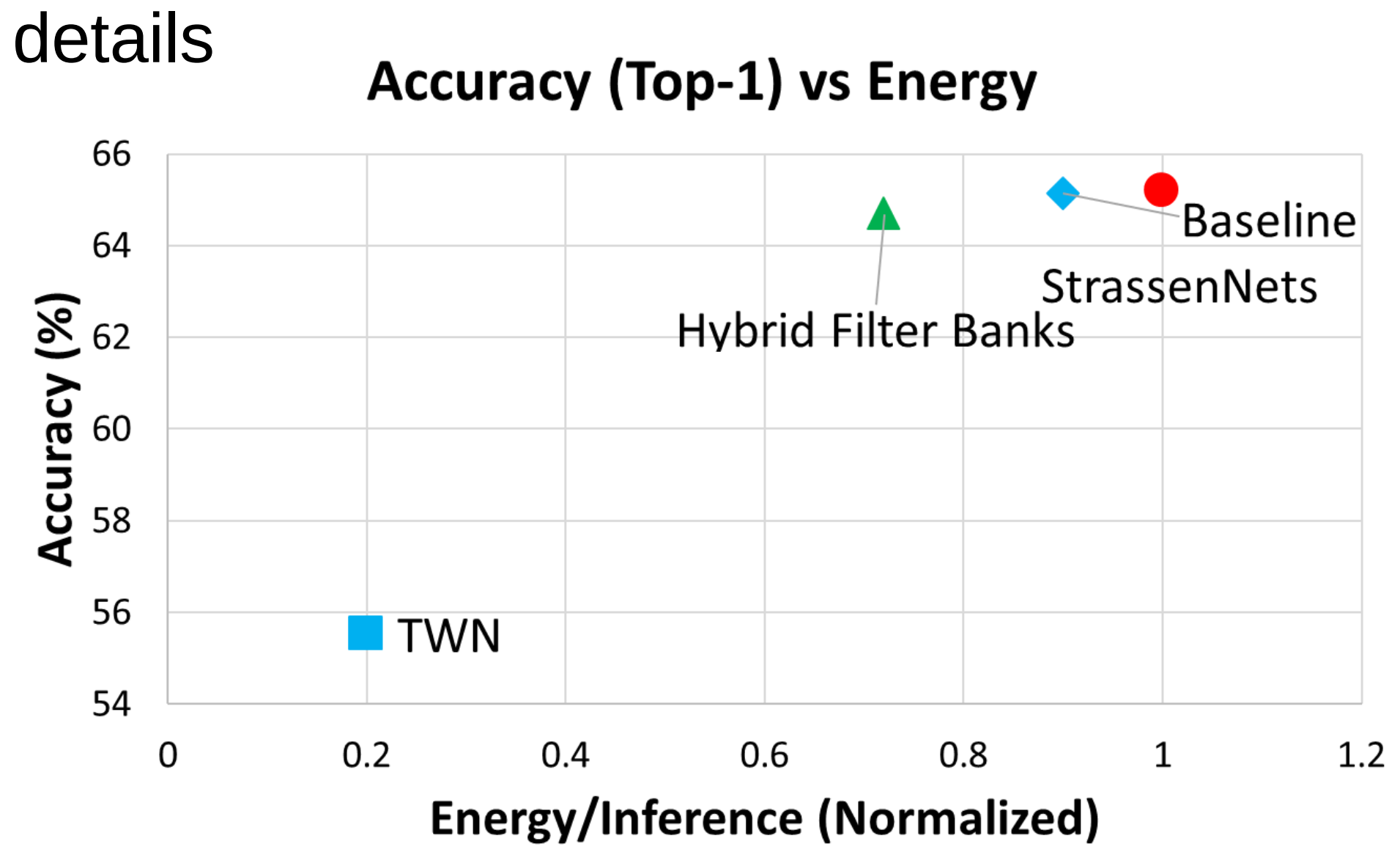
- Use StrassenNets
- Use fewer hidden units
- Restrict increase in ADDs

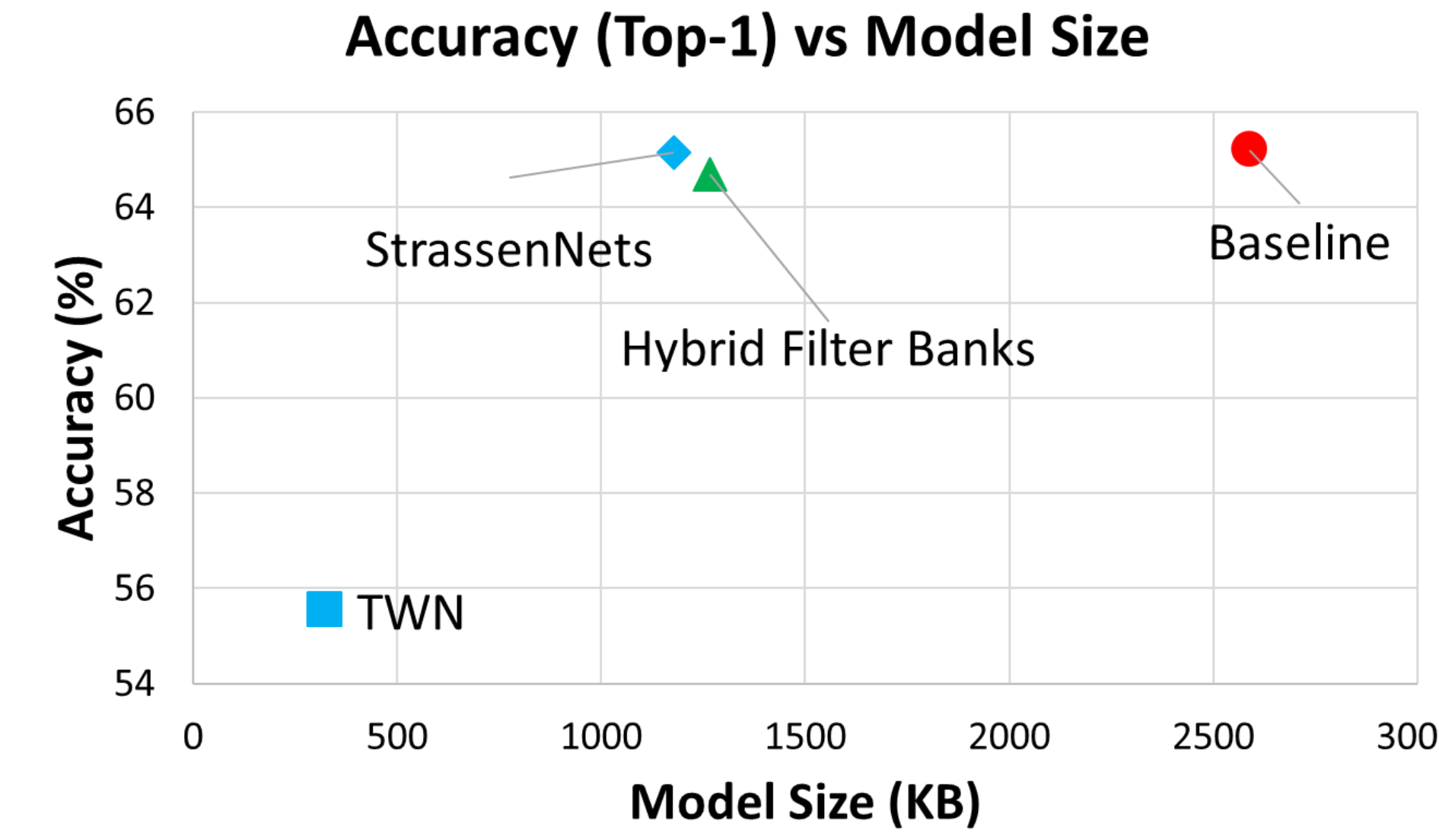
A MobileNets pointwise layer with hybrid filter bank

MobileNet-V1

Evaluation Results

- Dataset: ImageNet, Network: MobileNet-V1 (width multiplier of 0.5)
- 47% reduction in MULs, only 48% reduction in ADDs, when compared to >300%
- 51% reduction in MobileNets-V1 model size,
- 28% reduction in energy/inference
- No degradation in inference throughput on an area-equivalent ML accelerator comprising both MAC and adder units
- 0.27% loss in top-1 accuracy
- Hybrid filter banks is effective in compressing ResNet architecture comprising 3x3 convolutional filters also; see our paper for details





Top-1 accuracy, energy/inference, and model size of hybrid filter banks and improvement over state-of-the-art ternary quantization schemes

Read Our Paper for Details

Gope et al., "Ternary MobileNets via Per-Layer Hybrid Filter Banks", 2019

arXiv link: https://arxiv.org/abs/1911.01028

References

[1] Howard et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications", 2017

[2] Li et al., "Ternary weight networks," NeurIPS 2016

[3] Tschannen et al., "StrassenNets: Deep Learning with a Multiplication Budget", ICML 2018