**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 significant energy savings and runtime speedups

**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 3) layers
- Prior solutions come with their own advantages and limitations

**Observations with Prior Solutions**

![Diagram showing observations with prior solutions]

- Different filters respond differently to ternary quantization
- Different sensitivity of individual filters to StrassenNets
- Different sensitivity of group of filters to 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 \(\rightarrow\) 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

**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

**References**


---

**Read Our Paper for Details**

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