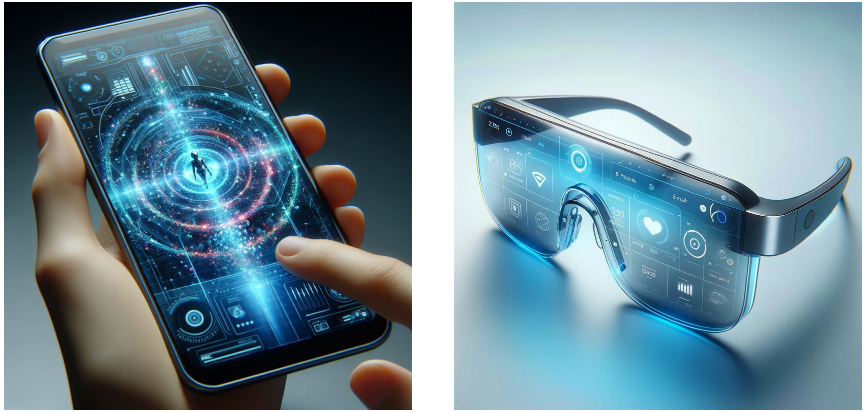
PikeLPN: Mitigating Overlooked Inefficiencies of Low-Precision Neural Networks

Marina Neseem, Conor McCullough, Randy Hsin, Chas Leichner, Shan Li,
In Suk Chong, Andrew G. Howard, Lukasz Lew, Sherief Reda, Ville-Mikko Rautio, Daniele Moro

Low-precision Quantization improves Energy Efficiency

- ➔ Int8 **Multiplication** consumes **18.5X** less energy than FP32 Multiplication.
- ➔ Int8 **Addition** consumes **30X** less energy than FP32 Addition.

Less Cost in Data Centers

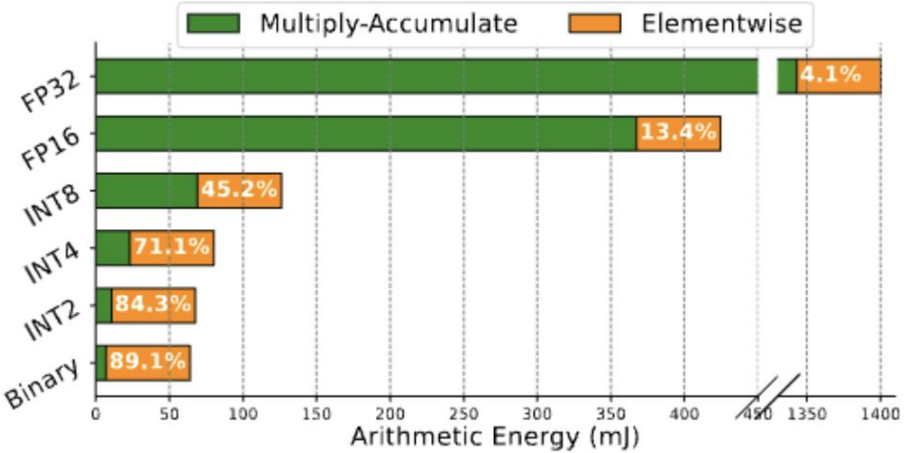


Longer Battery-Life on Edge Devices

State-of-the-art Quantized Models have overlooked inefficiencies

Arithmetic Operations in Quantized Models:

1. Multiply and Accumulate:
   - Convolution Layers
   - Linear Layers
   - Attention Layers.
   - ➔ **Quantized**
2. Elementwise:
   - Batch Normalization
   - Activation Functions
   - Quantization Scaling.
   - ➔ **NOT Quantized**

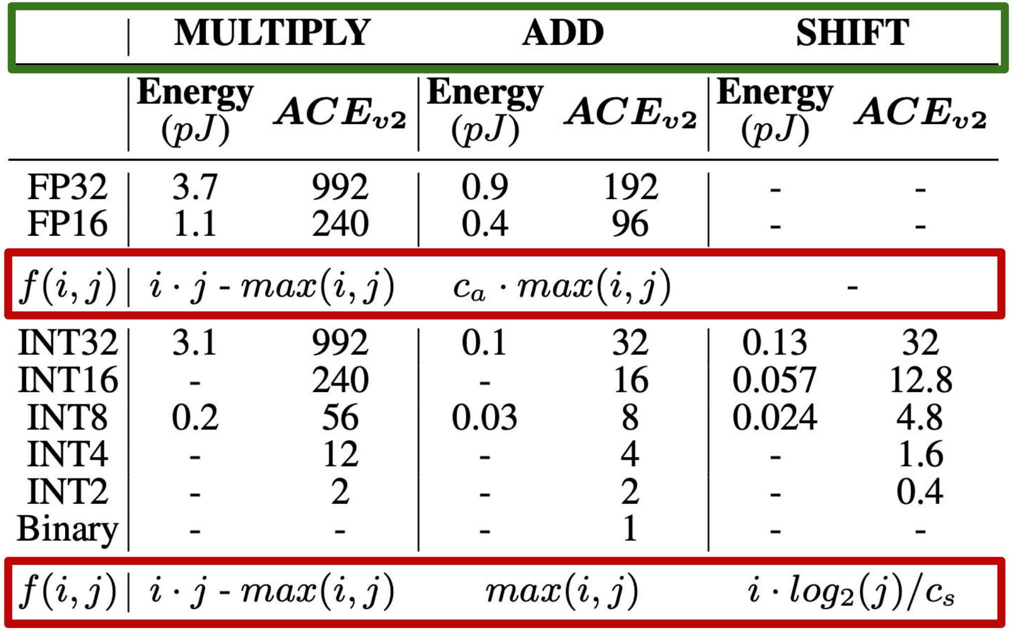


SOTA **Cost metrics like ACE*** only accounts for **multiply-accumulate** operations!

* Zhang, Yichi, Zhiru Zhang, and Lukasz Lew. "Pokebnn: A binary pursuit of lightweight accuracy." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

Our ACEv2 accounts for Overlooked Costs in existing cost metrics

- ACEv2 provides a simple formula for arithmetic operations as addition, multiplication, multiply-accumulate, and shift.
- ACEv2 has a correlation coefficient of 0.991 with the independently measured energy consumption.



	MULTIPLY		ADD		SHIFT	
	Energy (pJ)	ACE_{v2}	Energy (pJ)	ACE_{v2}	Energy (pJ)	ACE_{v2}
FP32	3.7	992	0.9	192	-	-
FP16	1.1	240	0.4	96	-	-
$f(i,j)$	$i \cdot j$ - $max(i,j)$		$c_a \cdot max(i,j)$		-	
INT32	3.1	992	0.1	32	0.13	32
INT16	-	240	-	16	0.057	12.8
INT8	0.2	56	0.03	8	0.024	4.8
INT4	-	12	-	4	-	1.6
INT2	-	2	-	2	-	0.4
Binary	-	-	-	1	-	-
$f(i,j)$	$i \cdot j$ - $max(i,j)$		$max(i,j)$		$i \cdot log_2(j)/c_s$	

Introducing our Low-Precision model PikeLPN

1. Start with Compact Architecture
2. Quantize All Layers
   - ✓ Batch Norm Quantization
   - ✓ Distribution Heterogeneous Quantization
   - ✓ Double Quantization

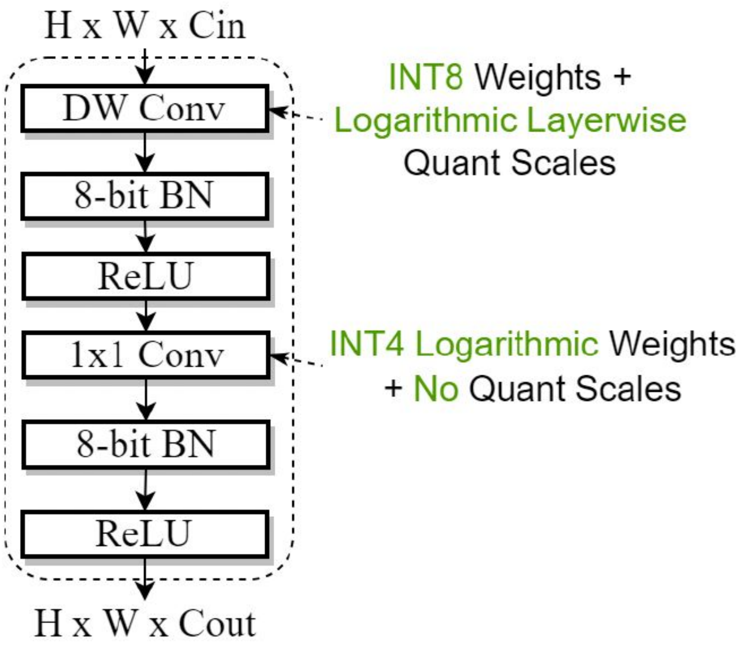

H x W x Cin
DW Conv
8-bit BN
ReLU
1x1 Conv
8-bit BN
ReLU
H x W x Cout
INT8 Weights + Logarithmic Layerwise Quant Scales
INT4 Logarithmic Weights + No Quant Scales

Figure: PikeLPN Building Block

PikeLPN outperforms 1 bit state-of-the-art Neural Networks

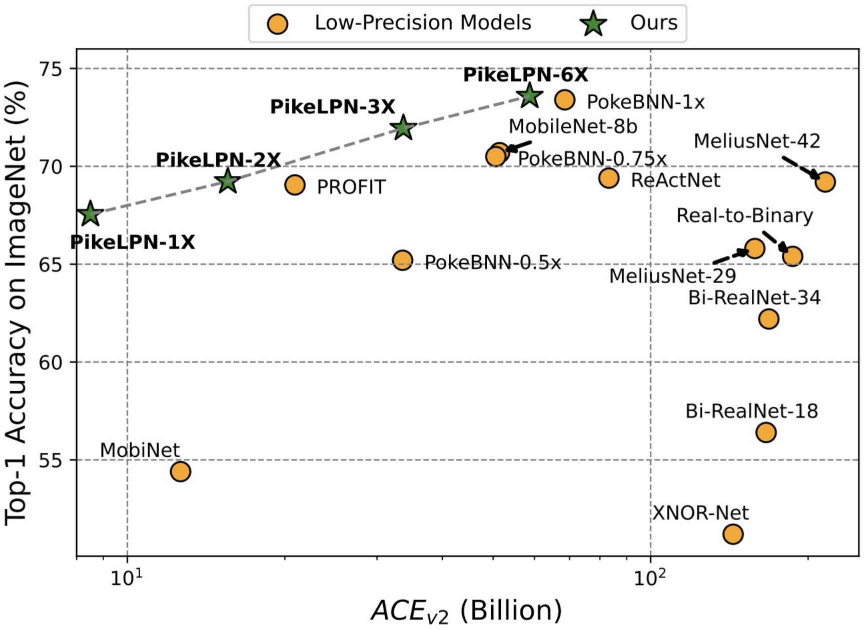


Figure: Top-1 Accuracy on ImageNet versus our ACEv2 cost of PikeLPN compared to SOTA low-precision models

Summary of Contributions

- ✓ Analysis of overlooked elementwise operations costs in SOTA models and cost metrics.
- ✓ Our hardware-agnostic cost metric, ACEv2, has 0.991 correlation with energy consumption.
- ✓ PikeLPN family of low-precision models with up to 3.5X energy improvements.

Thank You