Data-type Aware Arithmetic Intensity for Deep Neural Networks

Nandan Kumar Jha[1], Sparsh Mittal[1], Sasikanth Avancha[2]

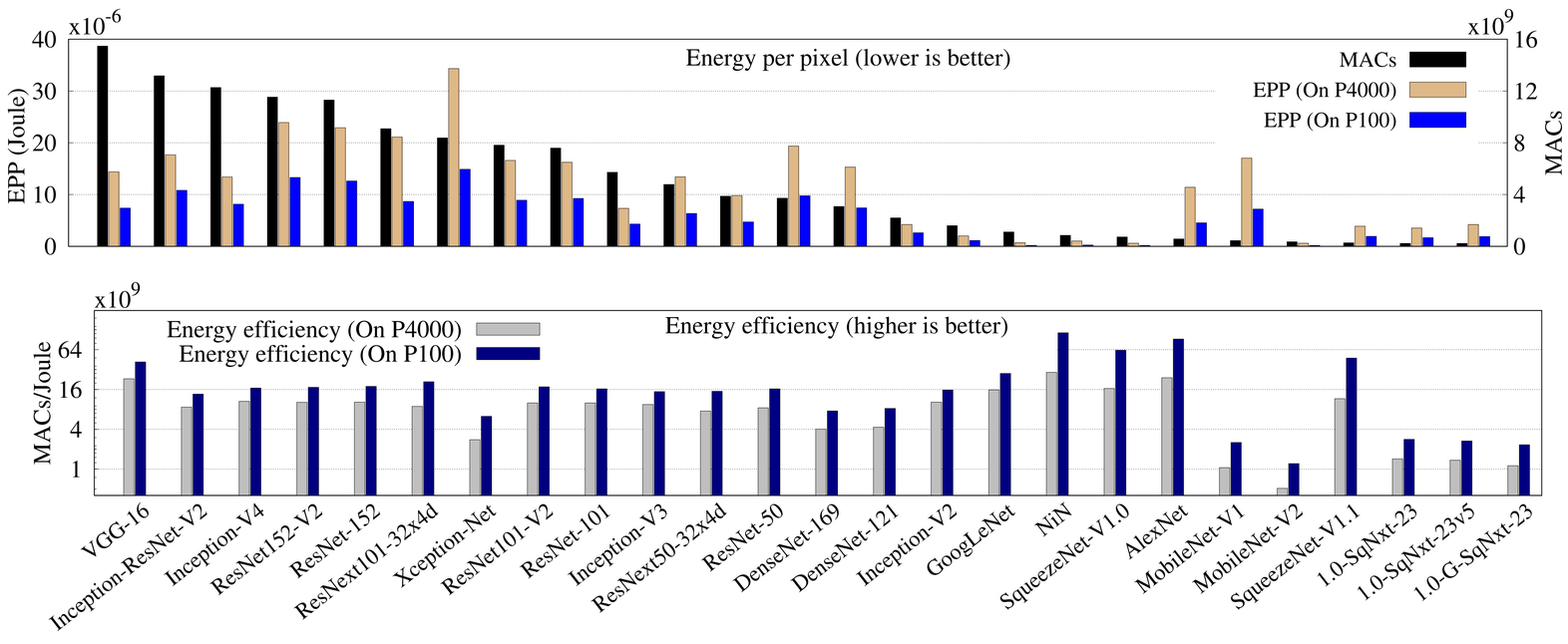
[1]Department of Computer Science and Engineering, IIT Hyderabad, India

[2] Parallel Computing Lab, Intel Corporation

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Introduction and Motivation

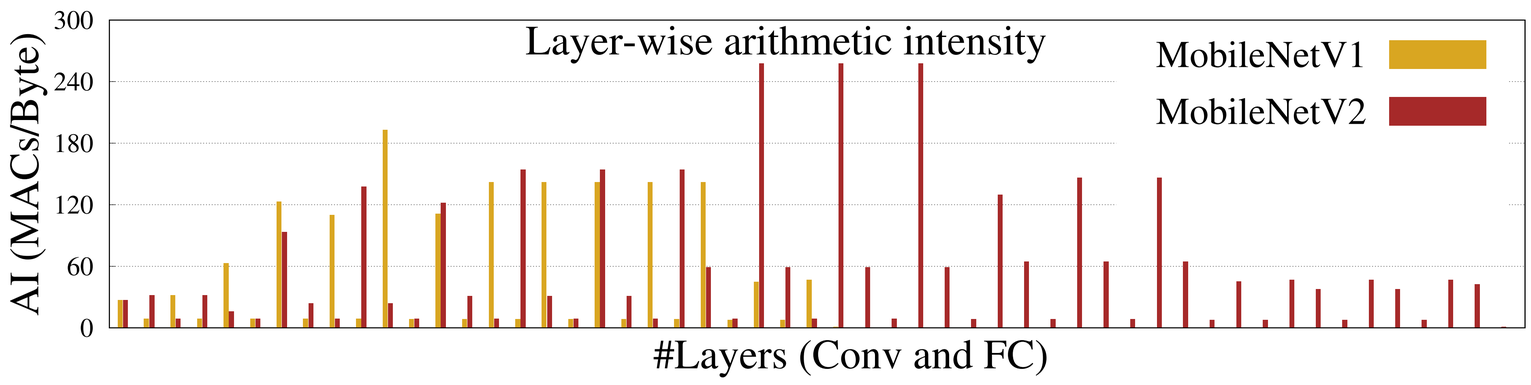
- Energy consumption of DNNs depend on computational complexity (#MACs) and energy efficiency of each MAC operation, which in turn depends on **data movement**.



Challenges

- DNN is a **special** type of workload which has computation phases with *radically* different computational intensity (OPs/Bytes). For example, depthwise, pointwise, groupwise convolution.

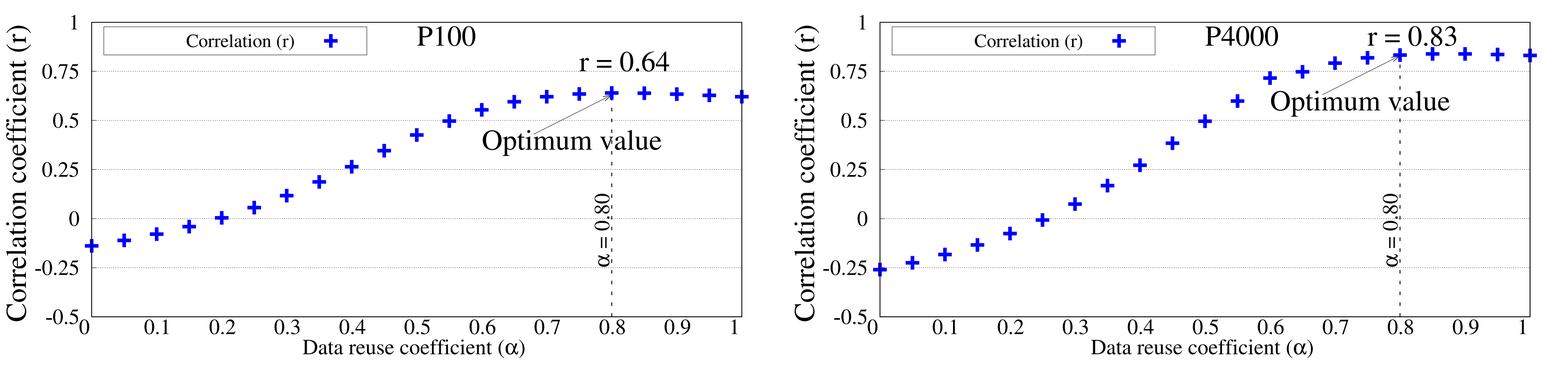
Convolution	Arithmetic intensity (AI)	AI_r
Standard	$\frac{M\times N\times S_k^2\times S_o^2}{M\times N\times S_k^2+(M+N)\times S_o^2}$	1.00
Pointwise	$\frac{M\times N\times S_o^2}{M\times N+(M+N)\times S_o^2}$	0.24
Group	$\frac{M\times N\times S_k^2\times S_o^2}{M\times N\times S_k^2+g\times(M+N)\times S_o^2}$	0.45
Depthwise	$\frac{M\times S_k^2\times S_o^2}{M\times S_k^2+(M+M)\times S_o^2}$	0.01



Proposed Method

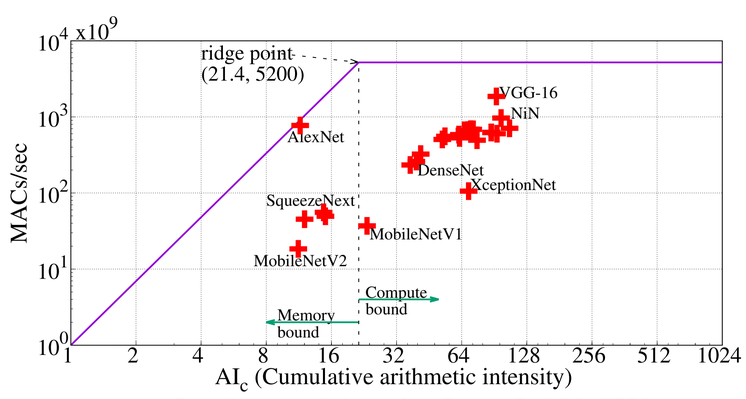
$$\frac{W+A}{2}\geqslant\frac{2\times W\times A}{W+A}\implies\frac{2\times M_c}{W+A}\leqslant\frac{M_c\times(W+A)}{2\times W\times A}\implies\frac{M_c}{W+A}\leqslant\frac{1}{4}\times\left[\frac{M_c}{A}+\frac{M_c}{W}\right]$$

$$\implies AI_c\leqslant\frac{1}{4}\times[\text{ActivationReuse}+\text{WeightReuse}]\quad\text{Now, }DI=\frac{1}{4}\times[\alpha\times\text{ActivationReuse}+(1-\alpha)\times\text{WeightReuse}]$$

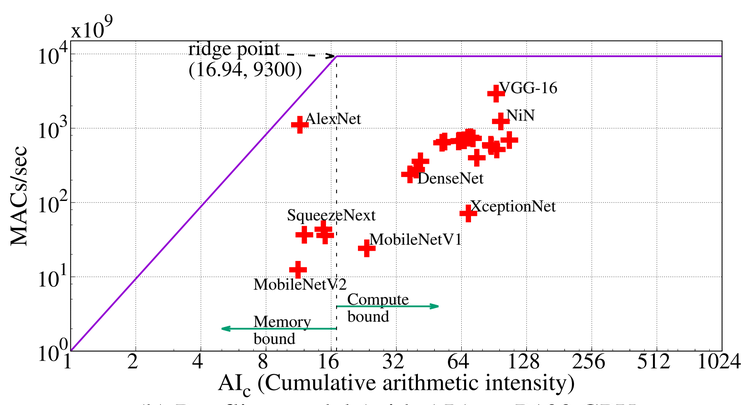


$$DI=\frac{1}{4}\times[0.8\times\text{ActivationReuse}+0.2\times\text{WeightReuse}]$$

Experimental Results (1)

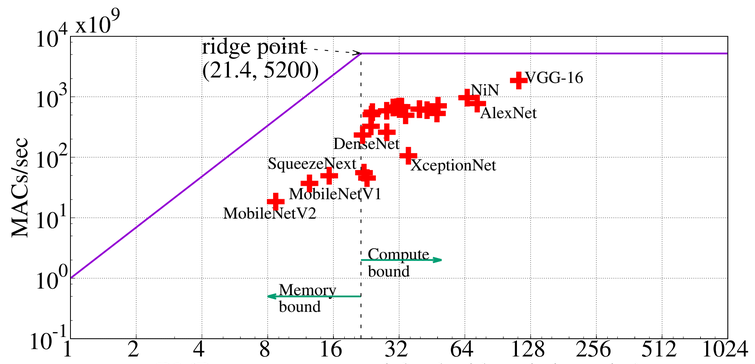


(a) Roofline model (with AI_c) on P4000 GPU

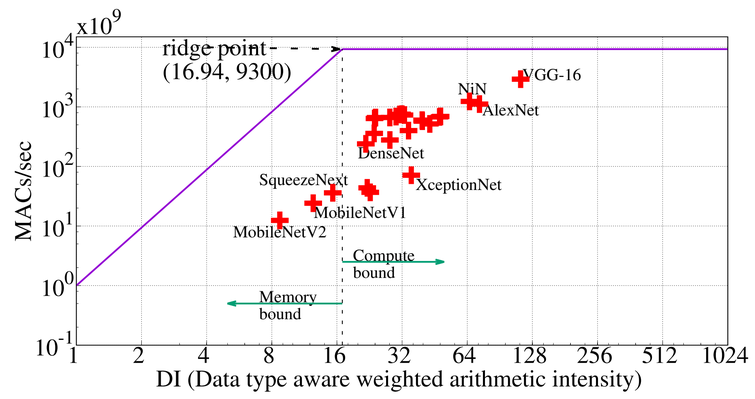


(b) Roofline model (with AI_c) on P100 GPU

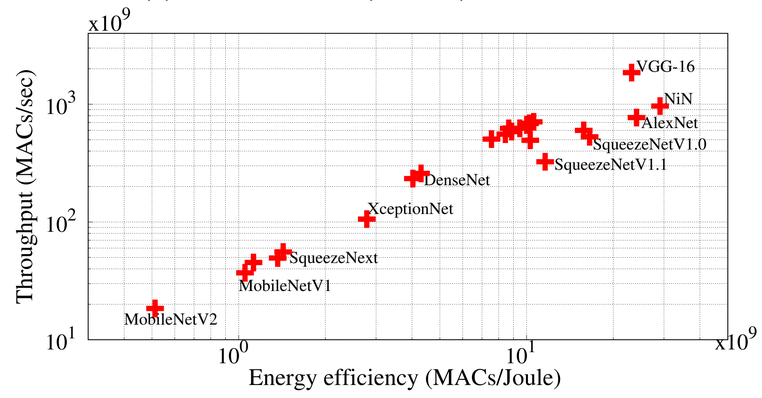
Experimental Results (2)



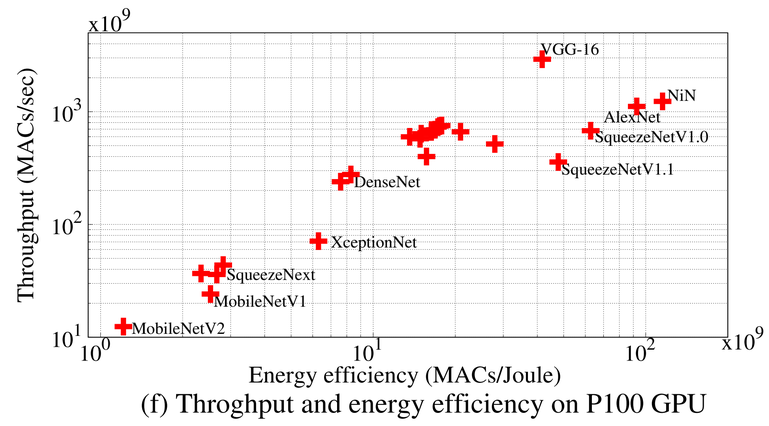
(c) Roofline model (with DI) on P4000 GPU



(d) Roofline model (with DI) on P100 GPU



(e) Throghput and energy efficiency on P4000 GPU



(f) Throghput and energy efficiency on P100 GPU

Discussion

GPU	X	Y	r = **correlation** (X,Y)
P4000	AI_c	Energy efficiency	0.52
	DI **(Ours)**	Energy efficiency	**0.83**
P100	AI_c	Energy efficiency	0.23
	DI**(Ours)**	Energy efficiency	**0.64**

Relative disparity (d_f) between AI_c and DI

$$d_f=\left(\frac{AI_c-DI}{AI_c}\right)\times 100=75-6.25\times\left[\frac{A}{W}+3\times\frac{W}{A}\right]$$

	Case 1: $A\ll W$	Case 2: $A\approx W$	Case 3: $A\gg W$
AI_c	$\approx M_c/W$	$\approx 0.5\times M_c/A$	$\approx M_c/A$
DI	$\approx 0.2\times M_c/A$	$\approx 0.25\times M_c/A$	$\approx 0.06\times M_c/W$
d_f	$\approx 75-18.75\times\frac{W}{A}$	≈ 50	$\approx 75-6.25\times\frac{A}{W}$

Conclusion and Future Work

- AI_c is a representative of energy-efficiency for mainly DNNs with $A\approx W$.
- AI_c fails to estimate the data reuse in DNNs when $|\,d_f\,|$ is high.
- DI is a better representative of energy-efficiency in DNNs.
- Evaluation of DI on DNN accelerators and FPGA.

Contact

cs17mtech11010 [at] iith [dot] ac [dot] in