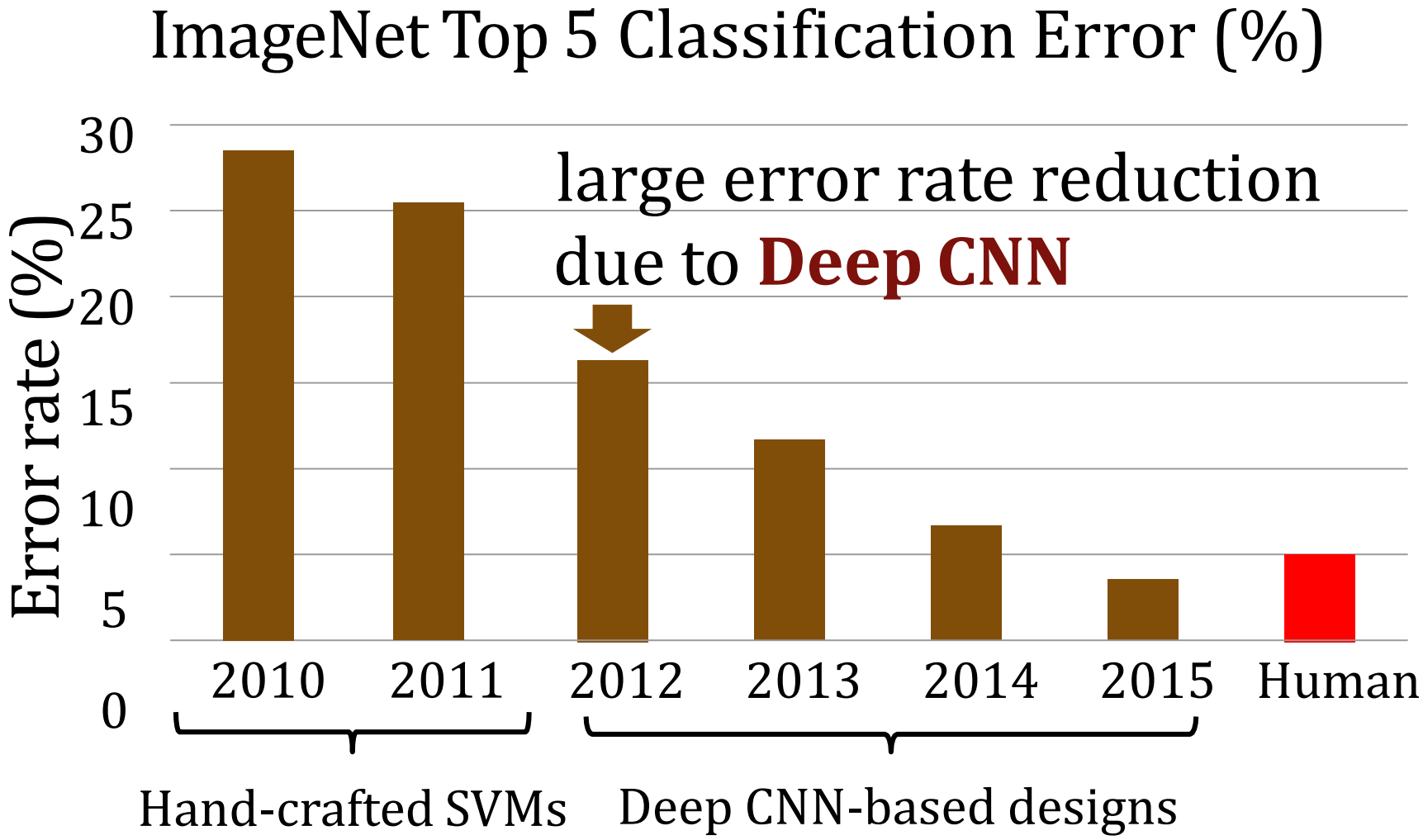
3DICT: A Reliable and QoS Capable Mobile Process-In-Memory Architecture for Lookup-based CNNs in 3D XPoint ReRAMs

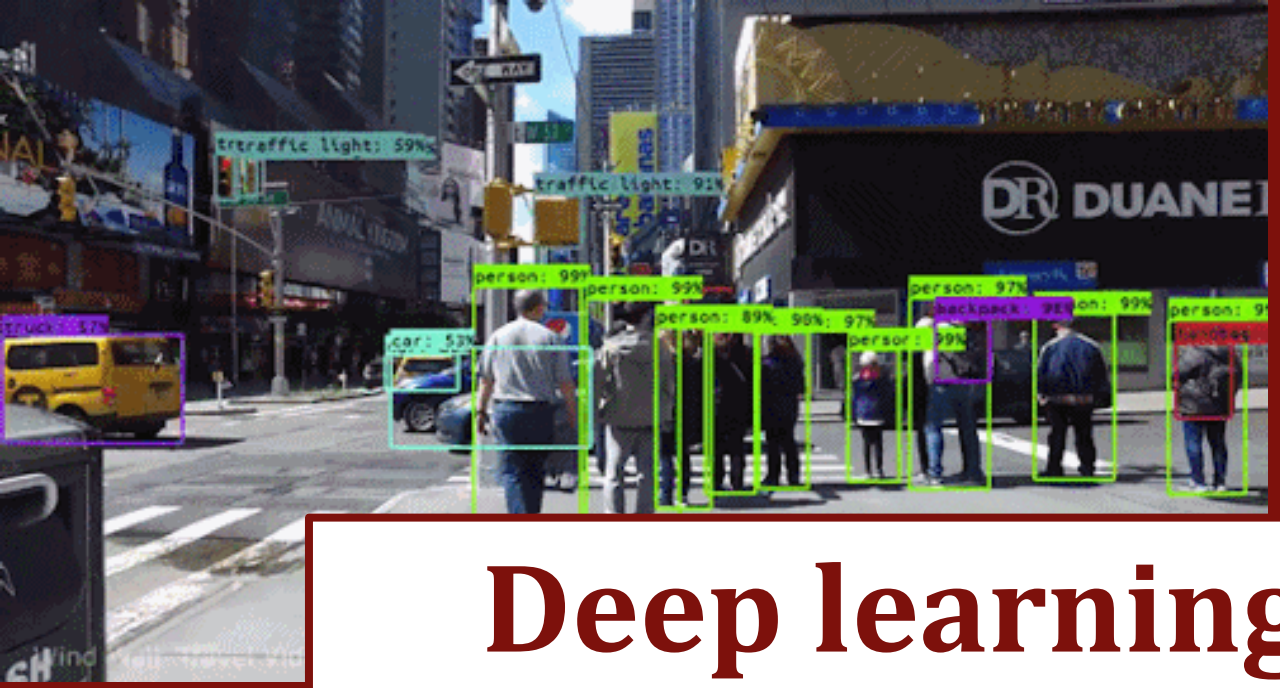
[1]Qian Lou, [2]Wujie Wen, and [1]Lei Jiang

[1]Indiana University Bloomington
[2]Florida International University

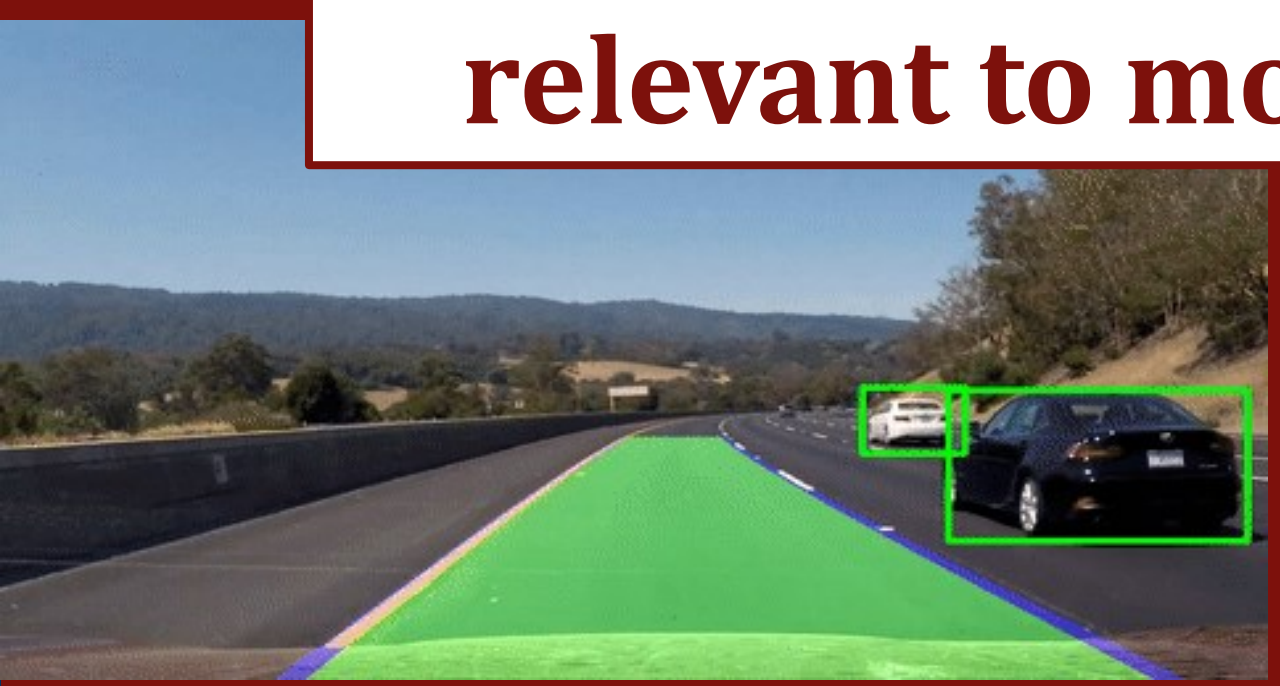
CNN is accurate

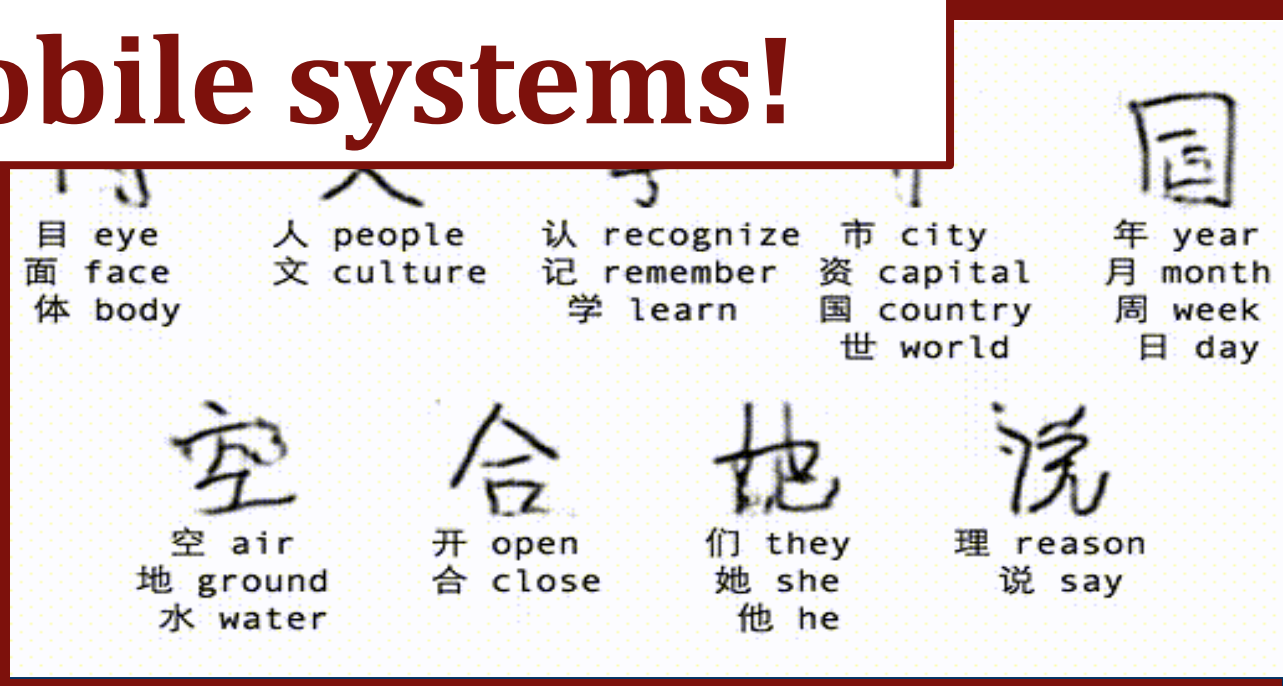


[Russakovsky et al., IJCV 2015]



Deep learning is extremely relevant to mobile systems!

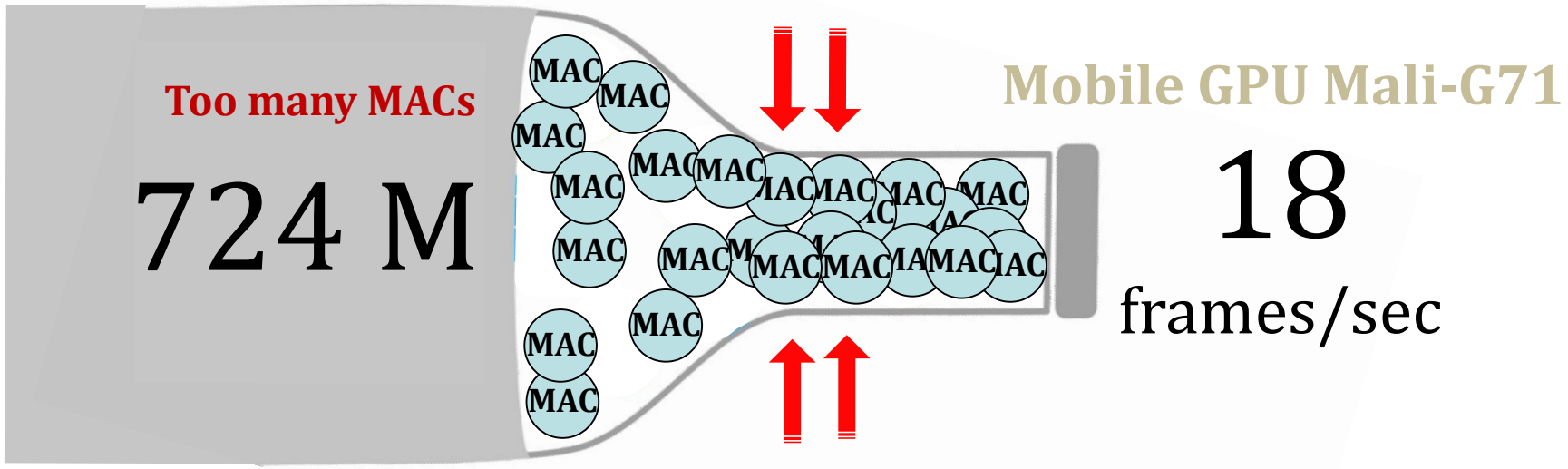




The price!

Metrics	LeNet-5	AlexNet	VGG-16	GoogleNet	ResNet-50
Total Weights #	60k	61M	138M	7M	25.5M
Total MACs #	341k	724M	15.5G	1.43G	3.9G

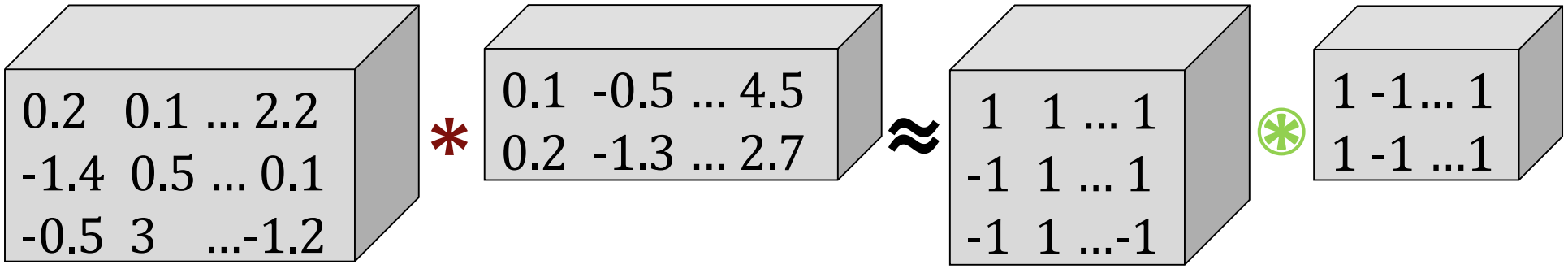
Bottlenecked due to limited hardware resource!



Limited battery lifetime

Low-precision CNN?

- Prior works improve inference **throughput** using low-precision CNN but with **accuracy loss**
  - **YodaNN** [Renzo+, ISVLSI'16]
  - **XNOR-POP** [Lei+, ISLPED'17]
  - **SOT-MRAM** [Deliang+, ICCD'17]
  - **XNOR-RRAM** [Xiaoyu+, DATE'17]
  - …



Float-point → Binary ***MAC → ⊛Xnor-bitcount**

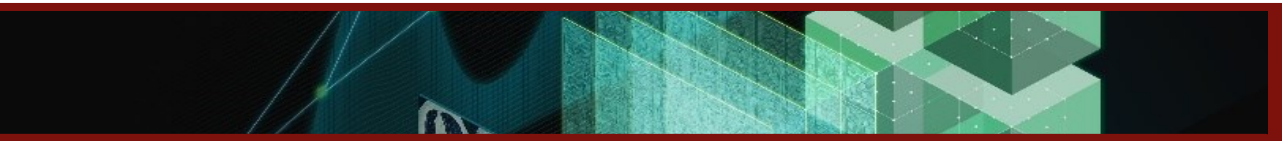
However, accuracy!

Sometimes accuracy matters! Unlocking phone

Critical task

Non-critical task

QoS in mobile systems is a must!



Critical task

Non-critical task

2 Medical diagnosis

VR object tracking 4

Executive summary

- **Motivation**:
  - CNN is **accurate** (> human) in mobile systems for object recognition, machine intelligence and so on.
  - Intelligent applications in mobile systems include critical tasks **(accuracy)** and non-critical tasks **(real-time). QoS is essential.**
  - **PIM** is an **energy-efficient** method for mobile systems.
- **Problem**: Current CNN in mobile system 1) is very slow, 2) no QoS support, 3) costs too much power
- **Goal**: To develop a **QoS** capable **PIM** architecture for mobile devices to support intelligent applications using **3D XPoint ReRAMs**.
- **3DICT**: 1) Lookup-based CNN (MAC#↓ weights#↓) 2) 2D ReRAM MLC endurance↓ ->2D ReRAM SLC throughput↓ ->3D ReRAM SLC)
- **Evaluation:**
  1. 3DICT can support QoS with 10-year life time.
  2. CNN test **performance per Watt** by **13%~61x** over prior architectures.

Outline

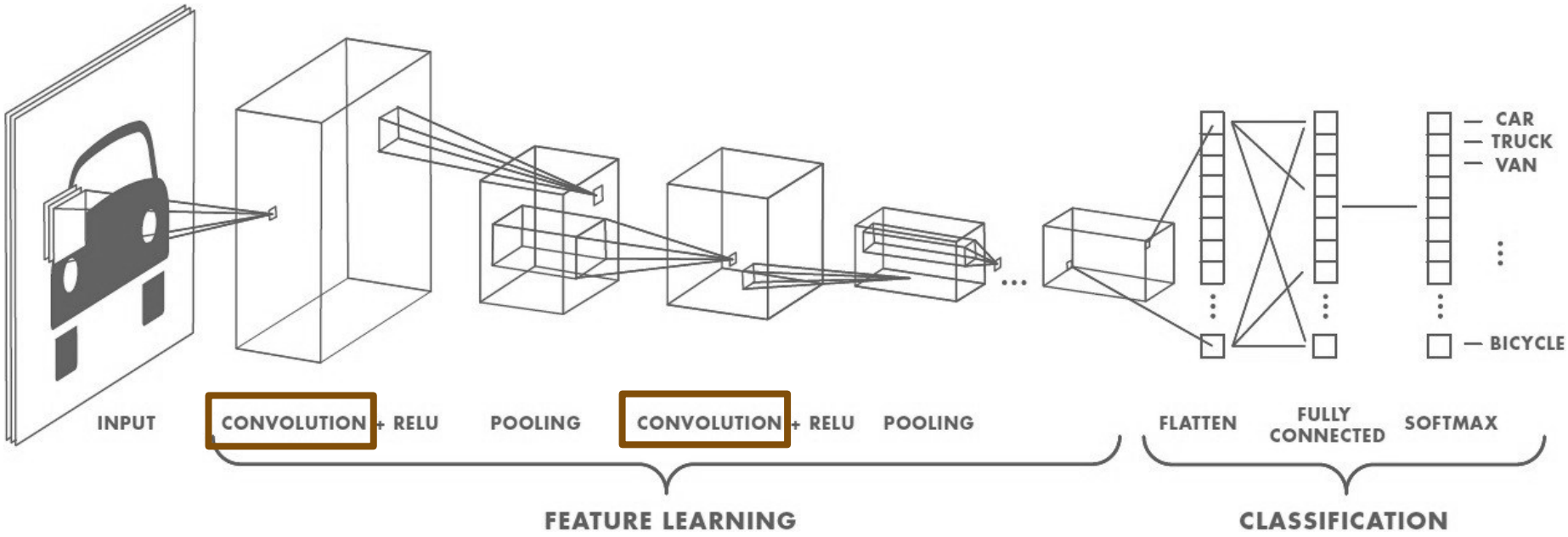
1. CNN and Lookup-based CNN

2. 3DICT

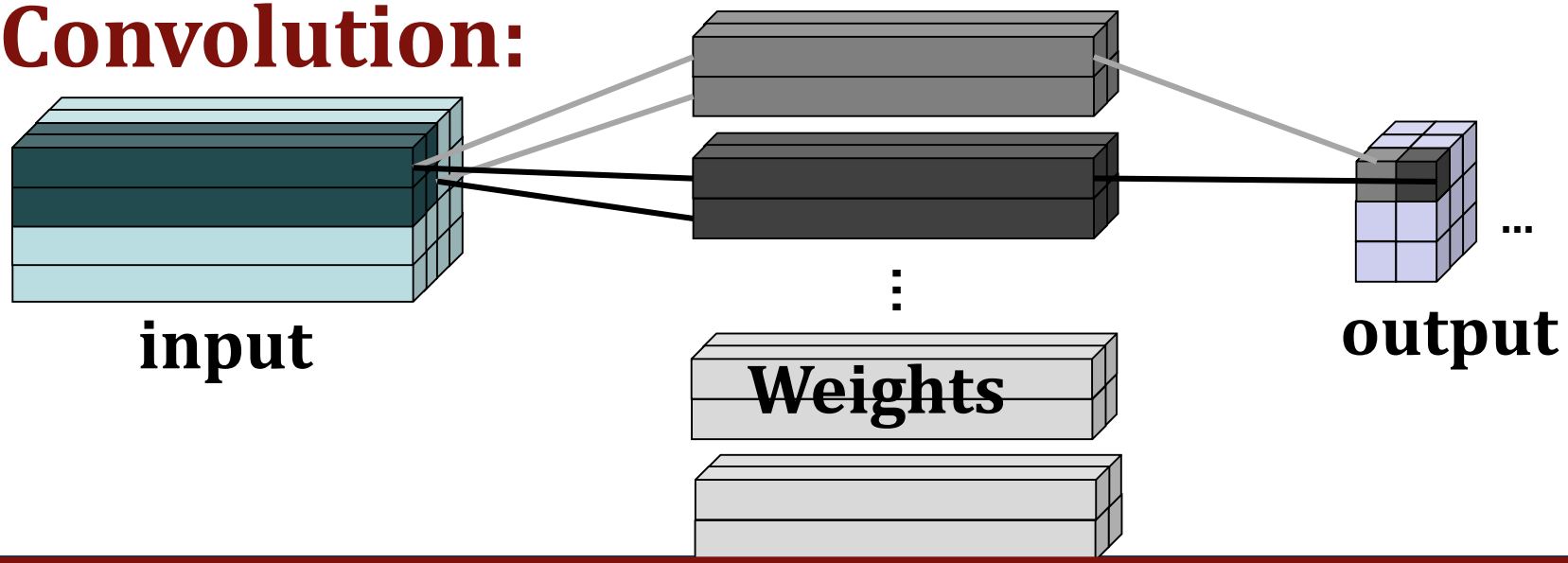
3. Evaluation

4. Conclusion

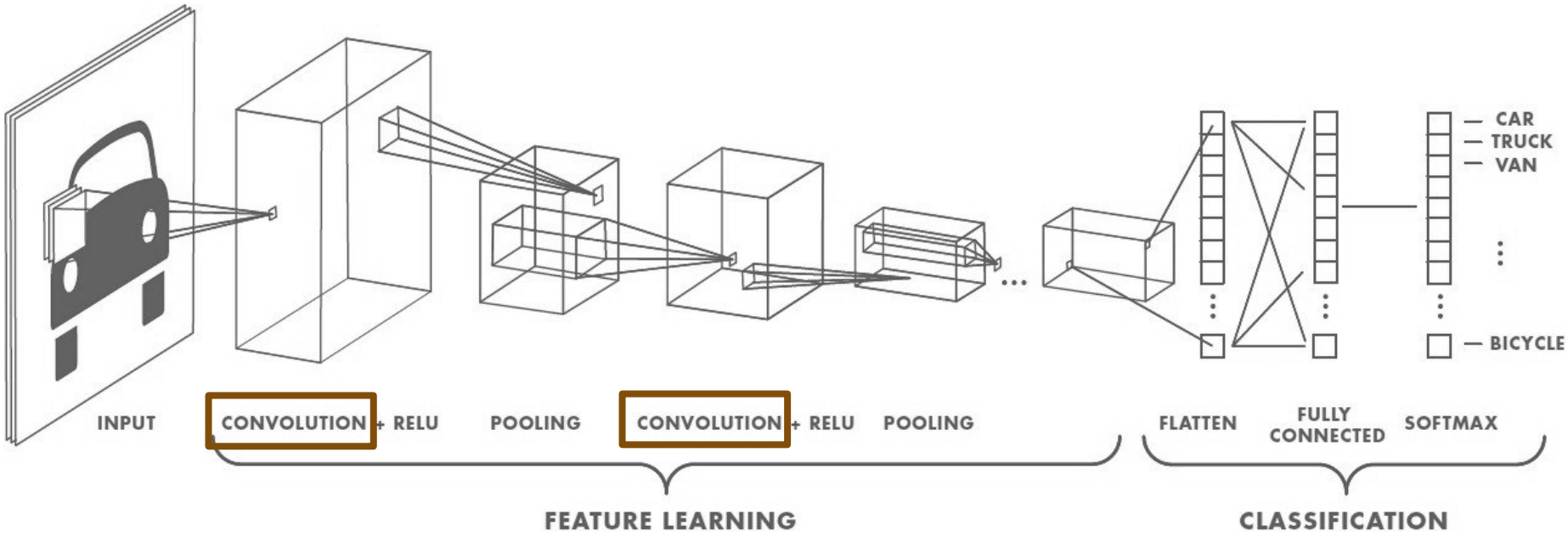
CNN



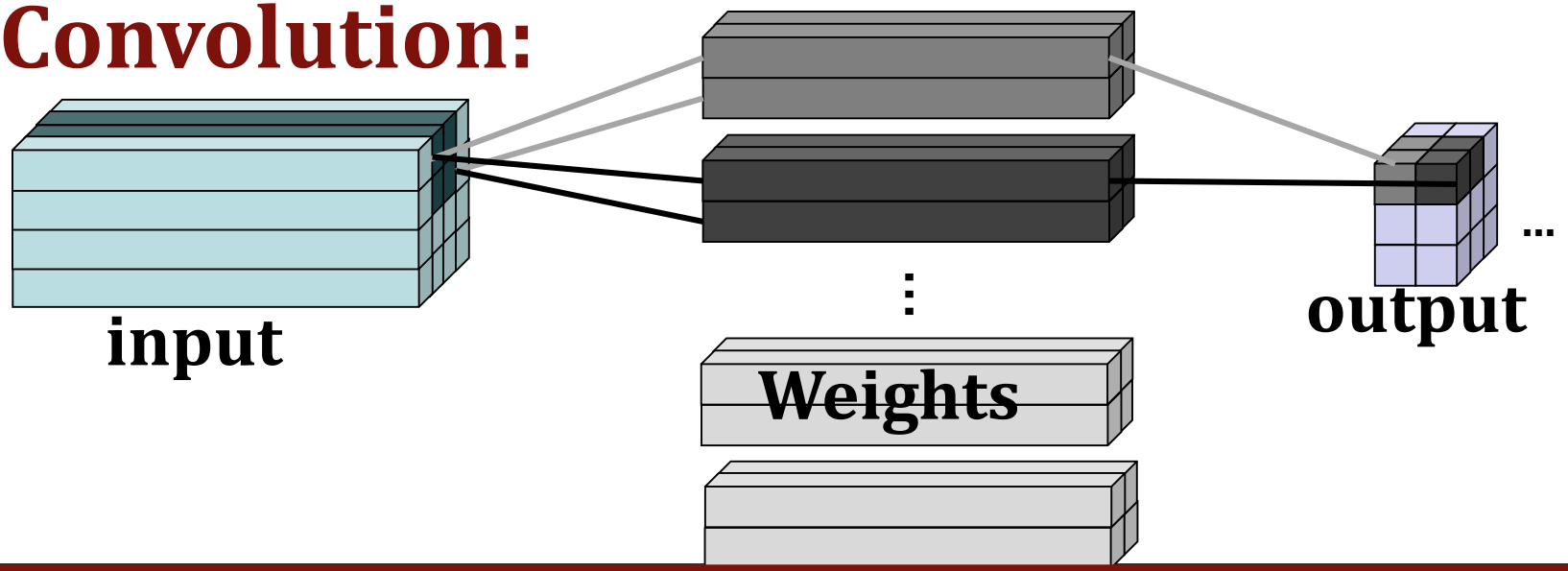
Convolution:



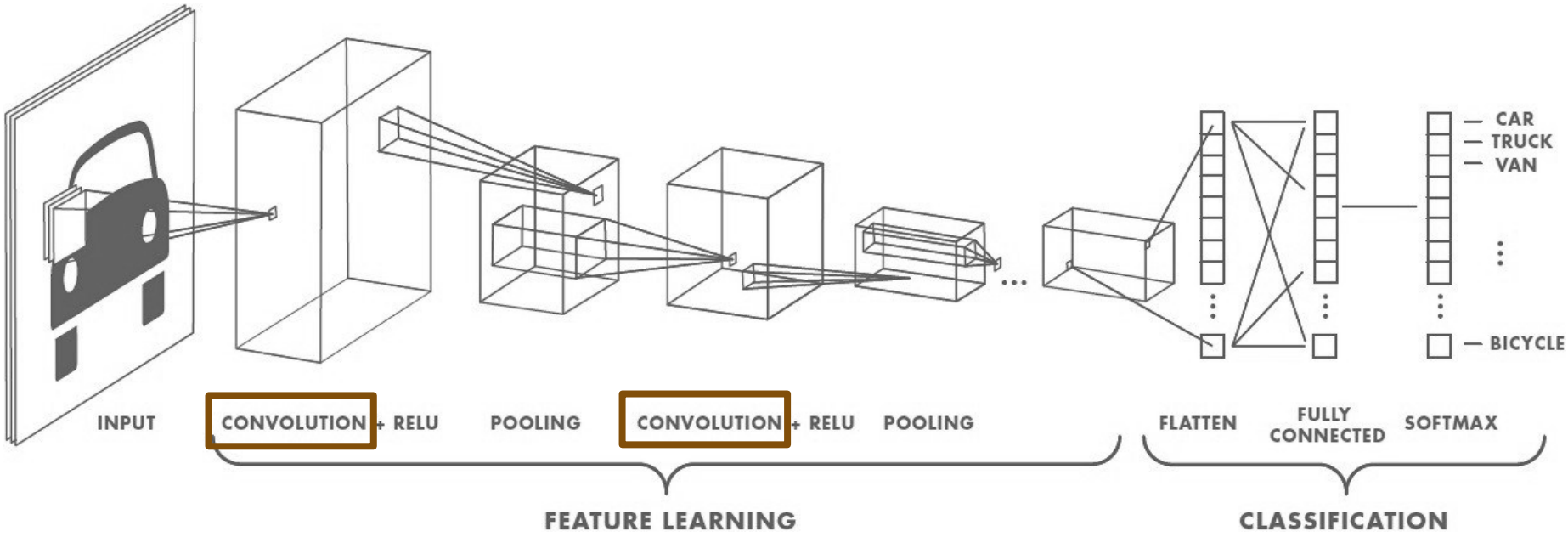
CNN



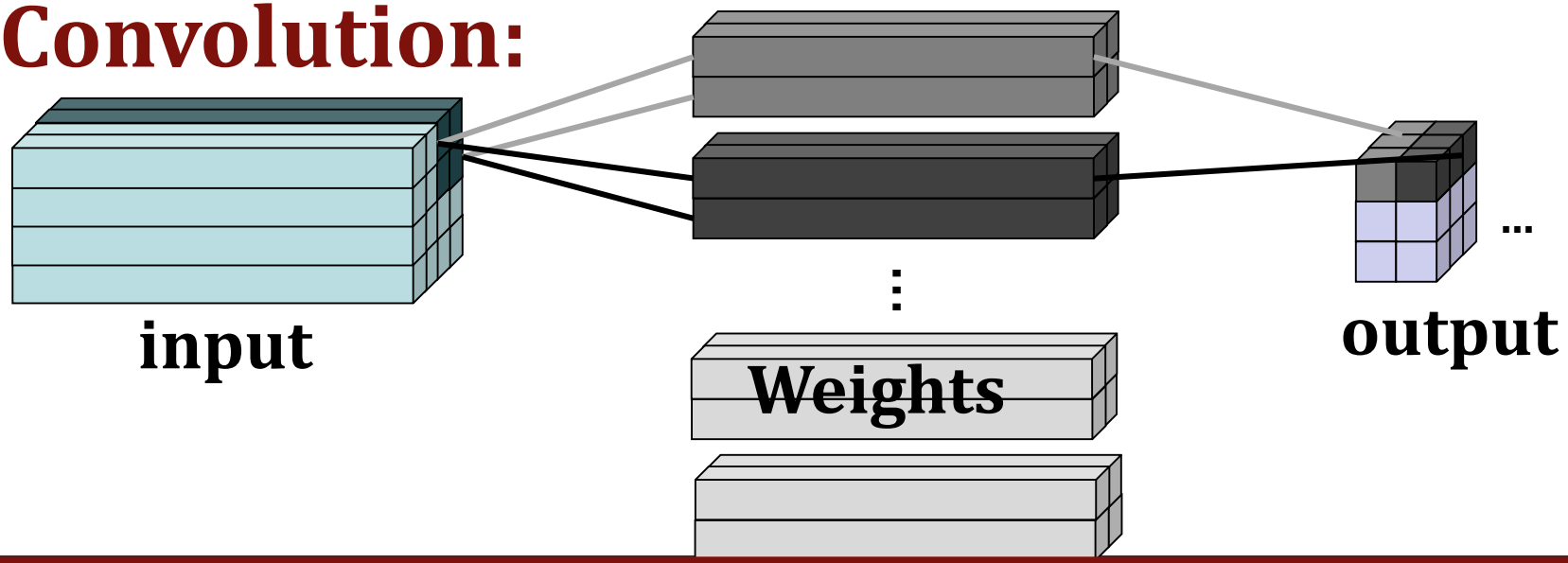
Convolution:



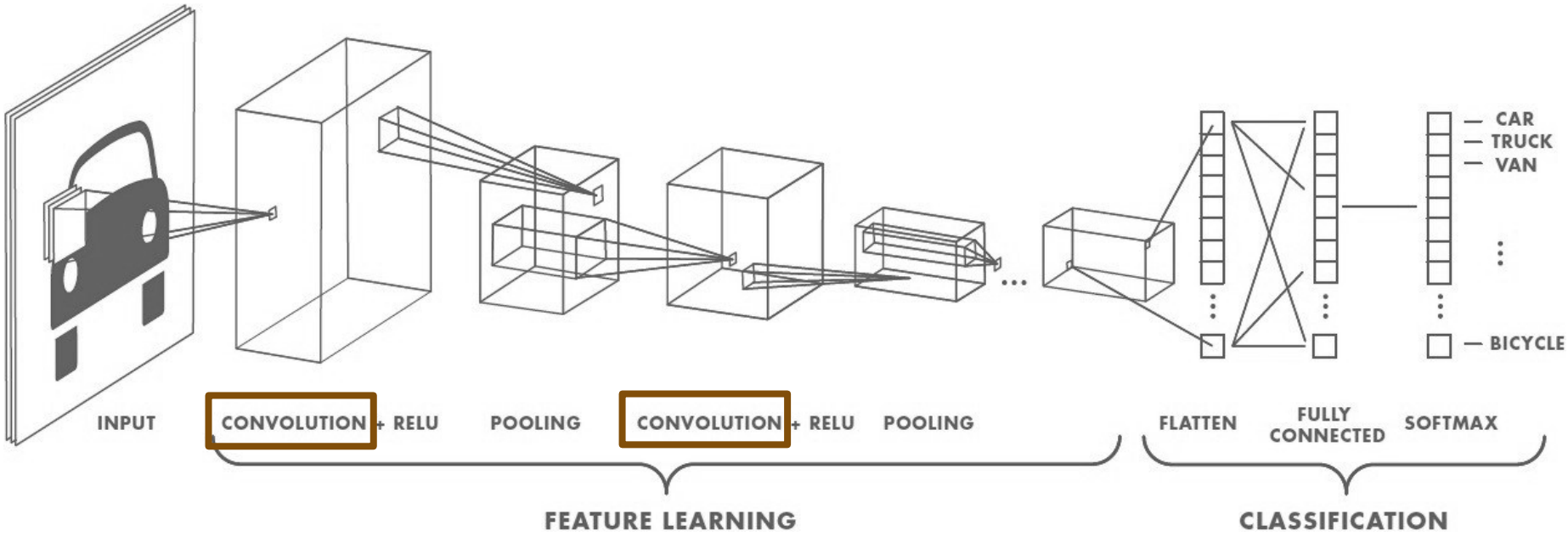
CNN



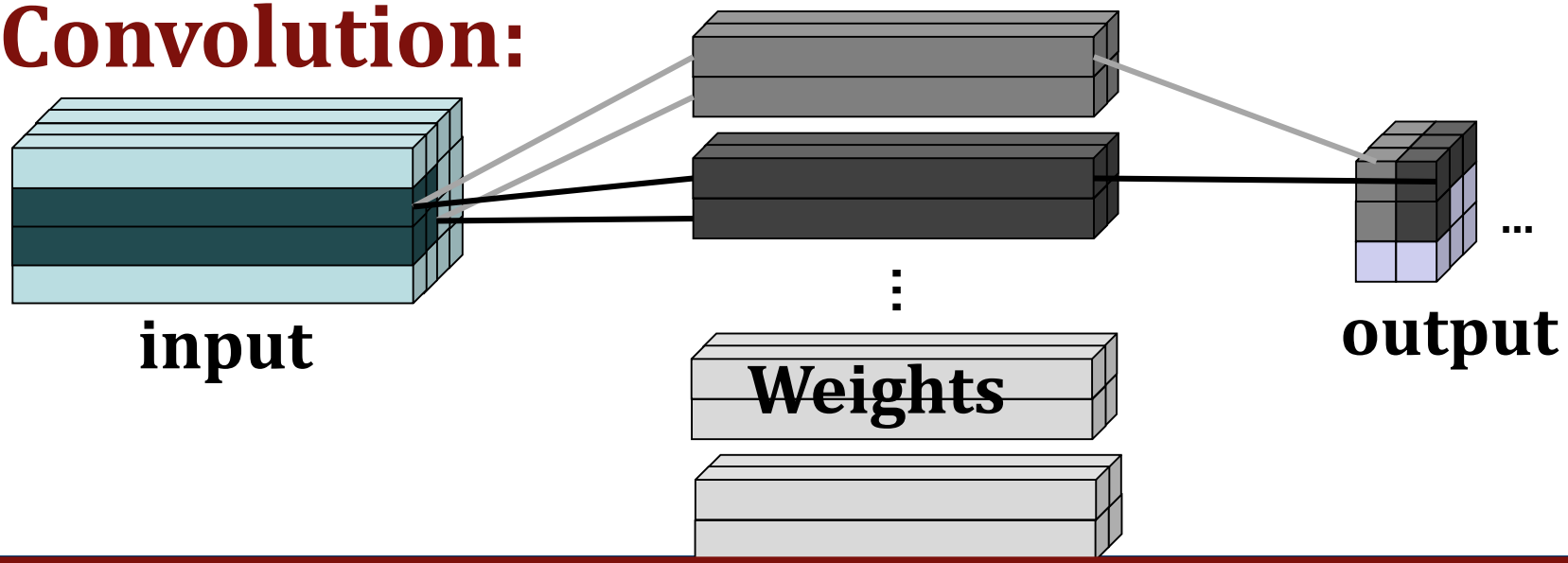
Convolution:



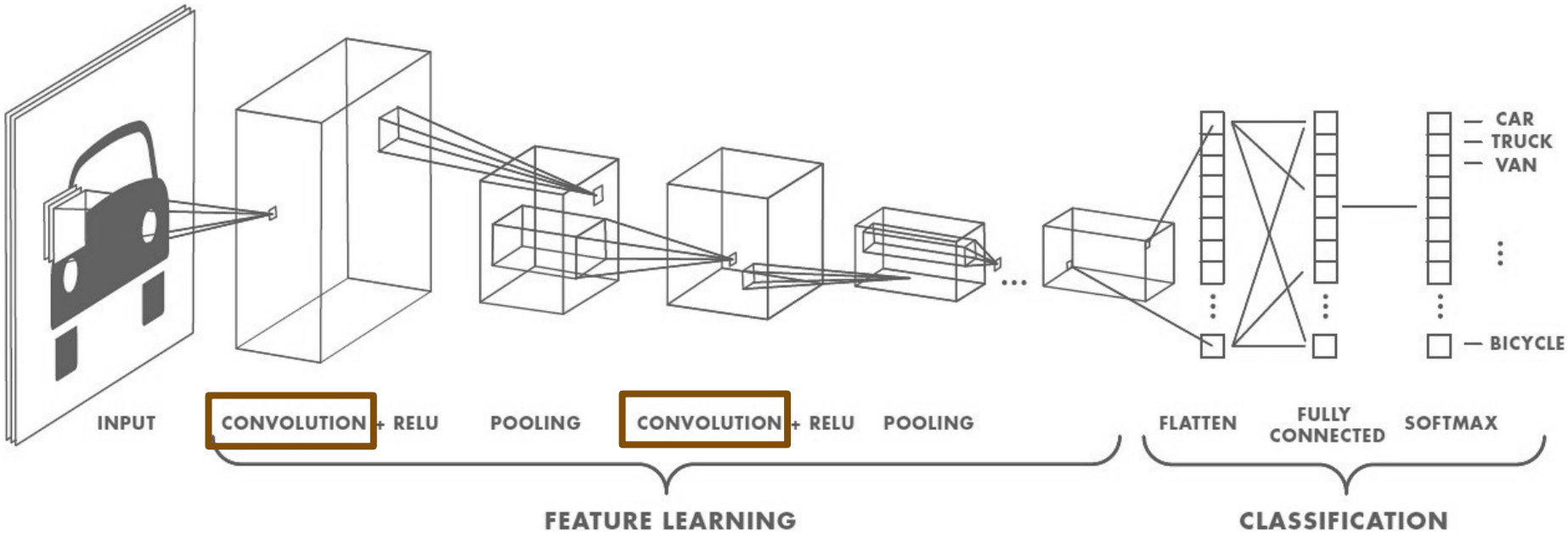
CNN



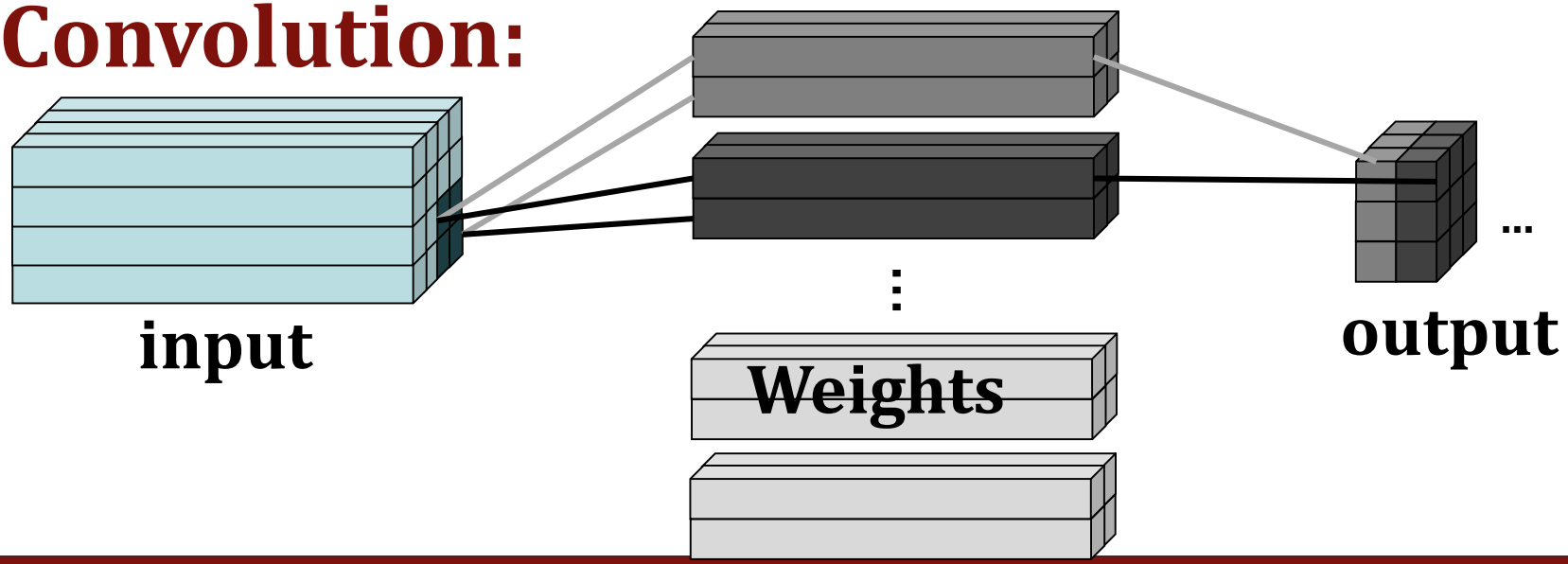
Convolution:



CNN



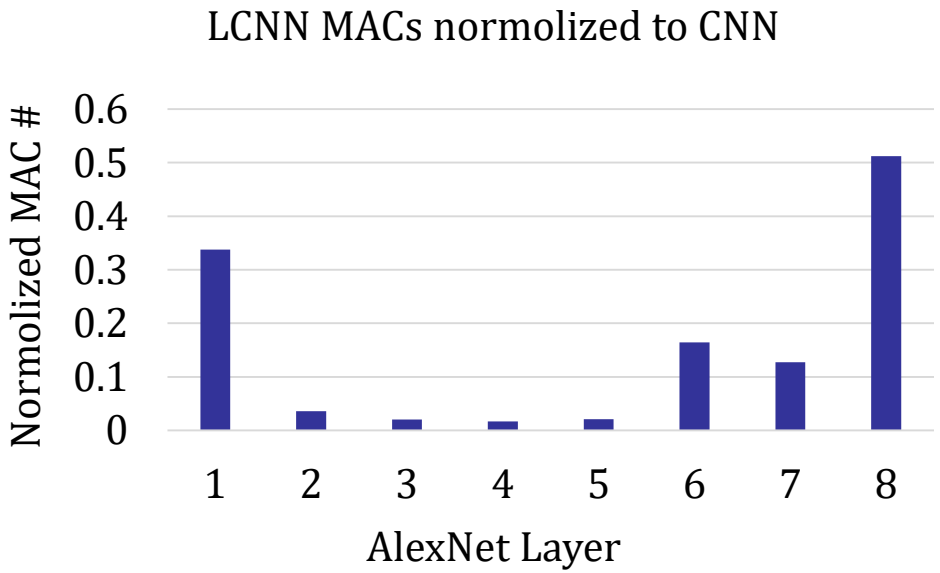
Convolution:



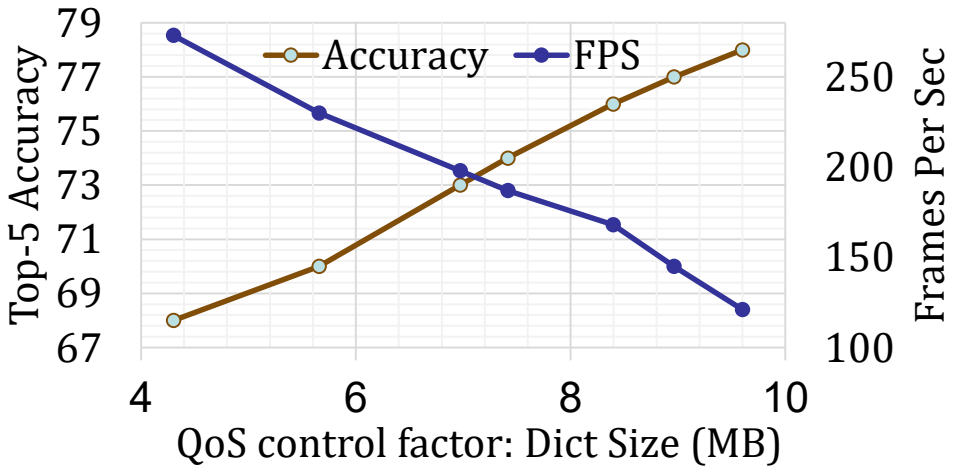
Lookup-based CNN?

AlexNet	# MACs	QoS?
CNN	724M	No
LCNN	43M	Yes

MACs:



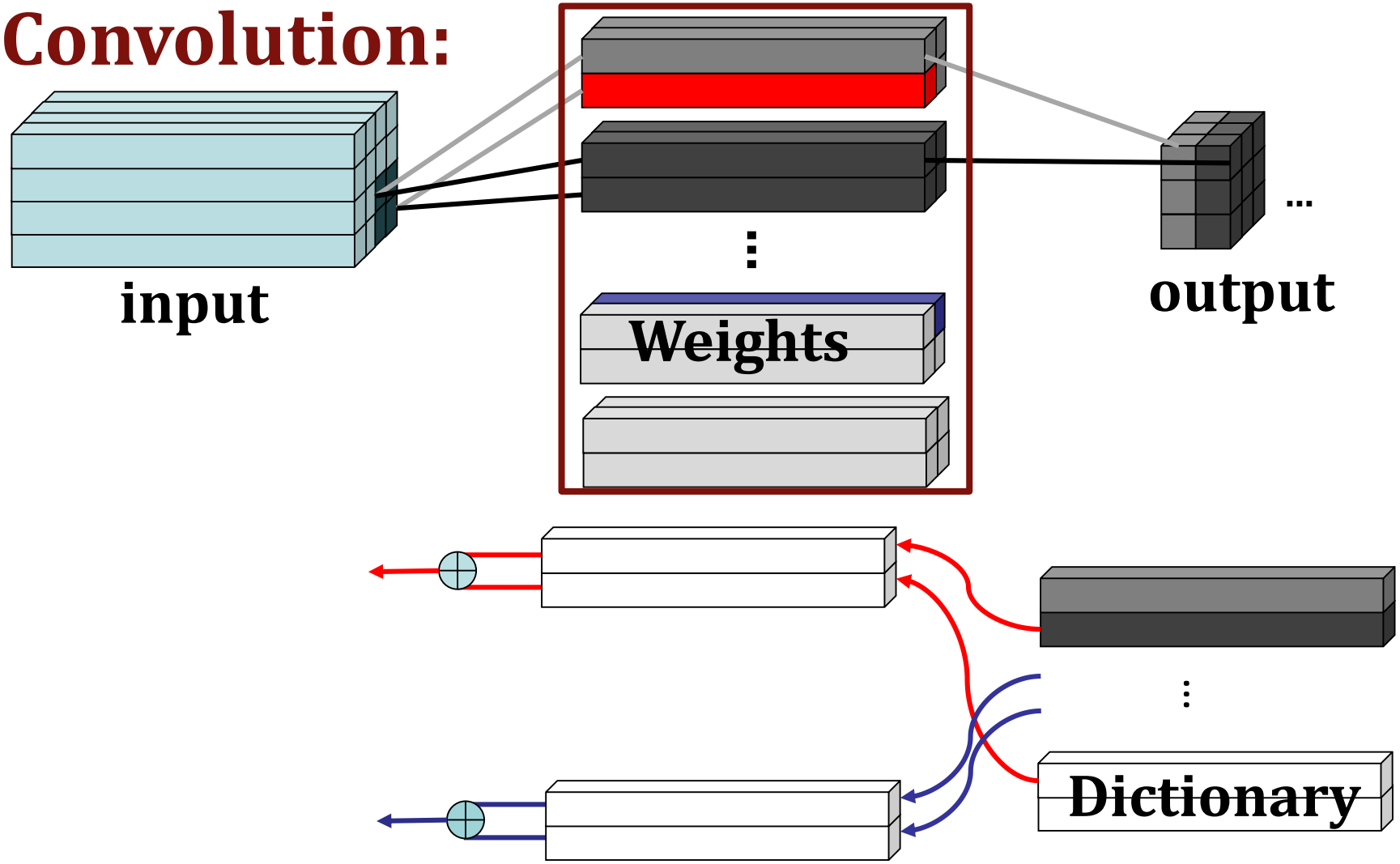
QoS:



[Hessam et al., CVPR 2017]

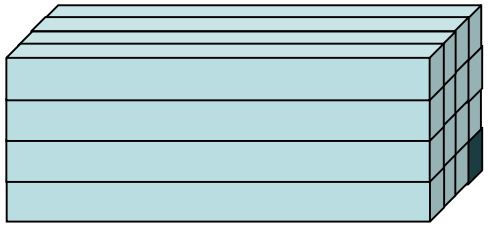
From CNN to LCNN

Convolution:

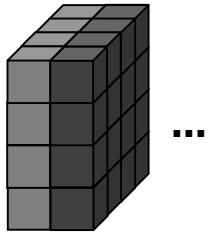


From CNN to LCNN

Small Convolution:



input

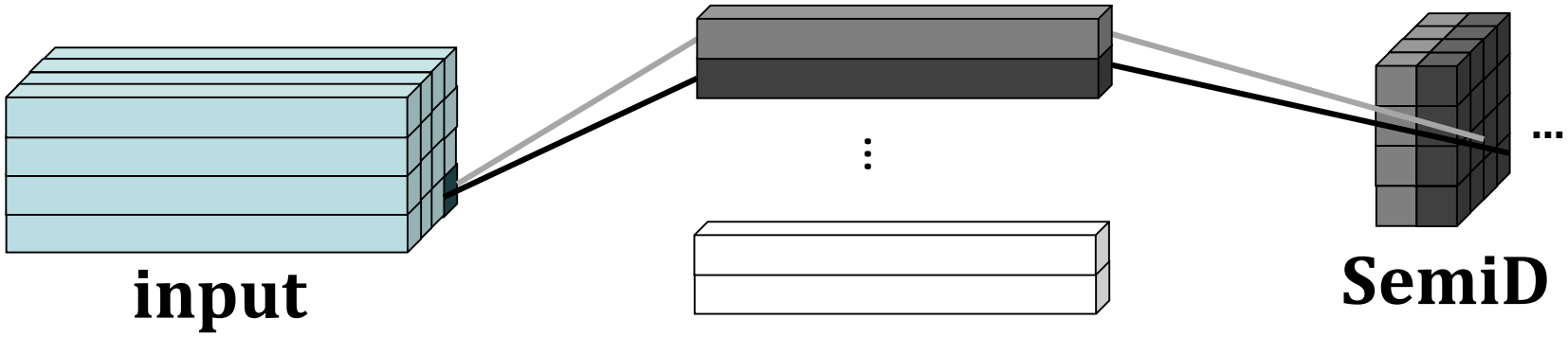


SemiD

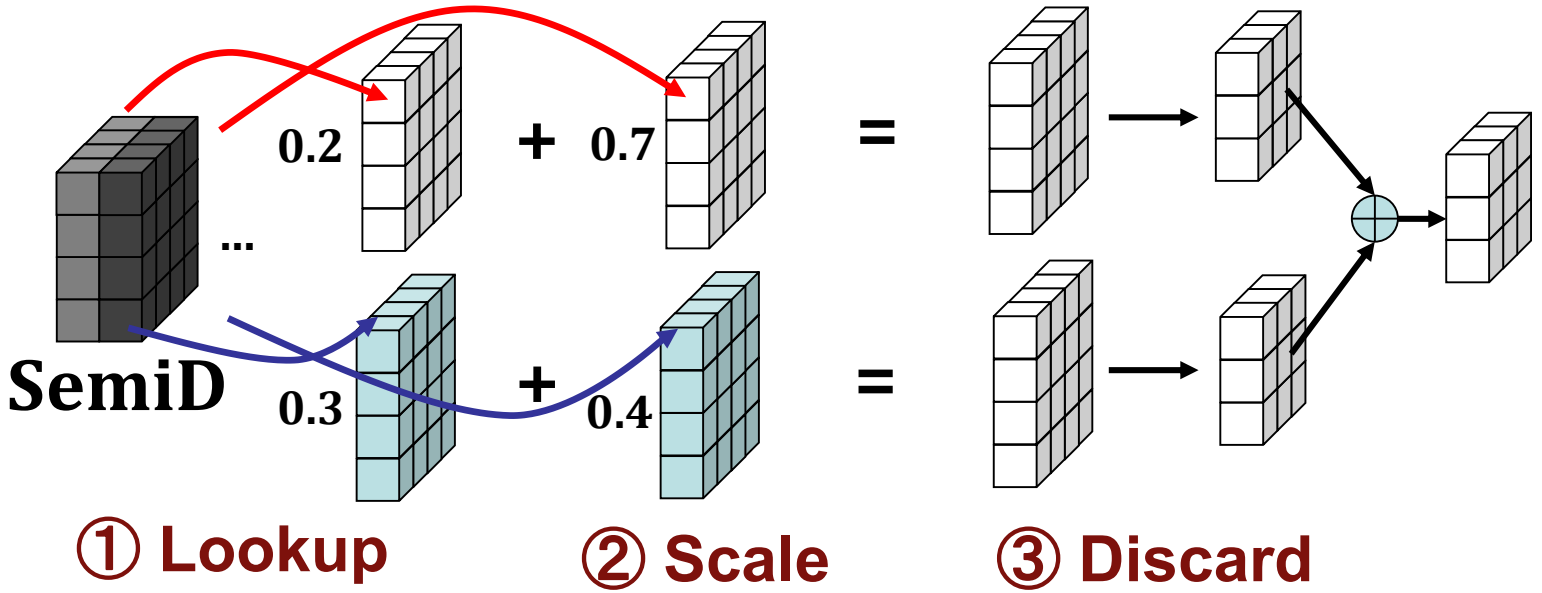
Dictionary

About LCNN

Small Convolution:



Lookup-scale-discard:



Outline

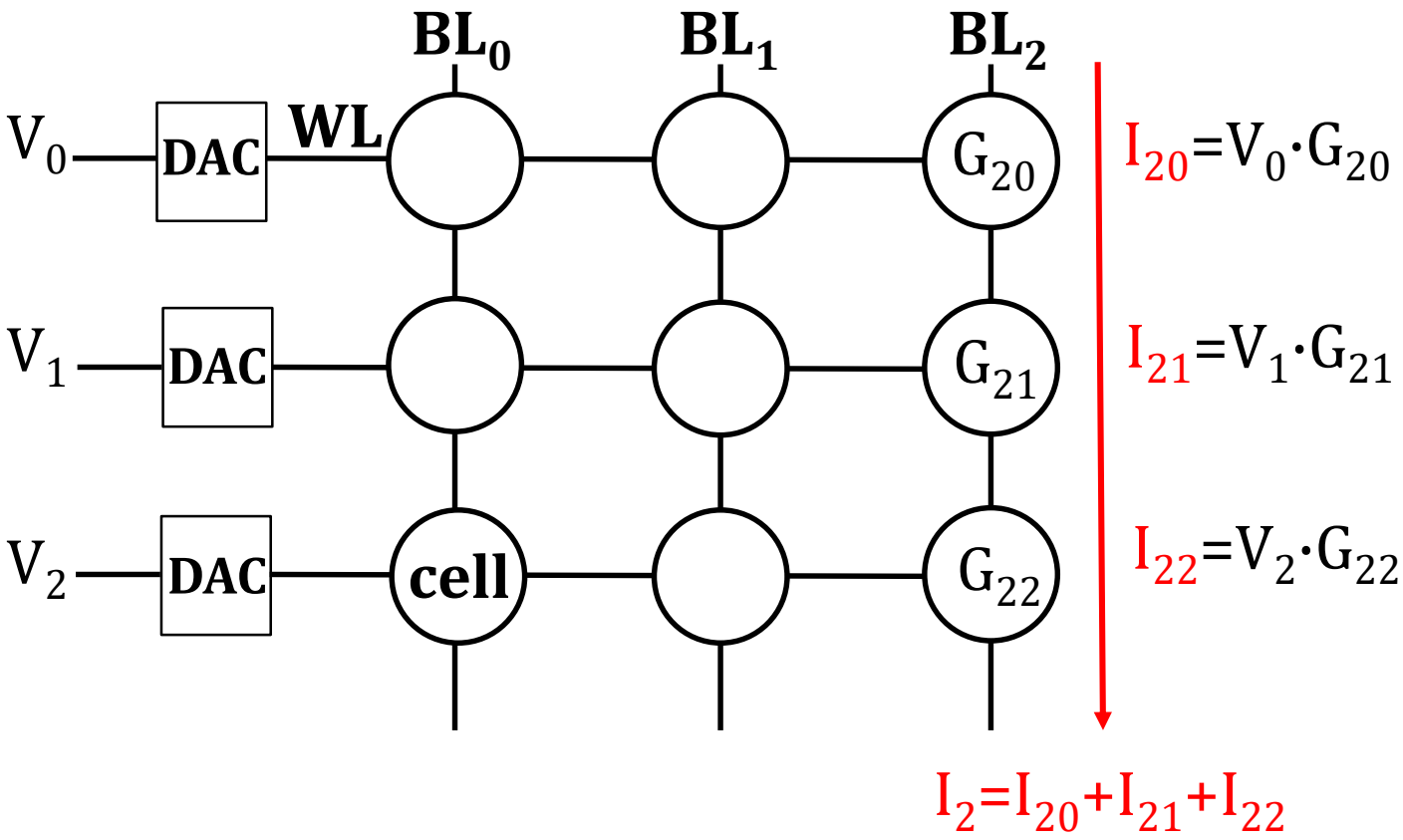
1. CNN and Lookup-based CNN

2. 3DICT

3. Evaluation

4. Conclusion

ReRAM array



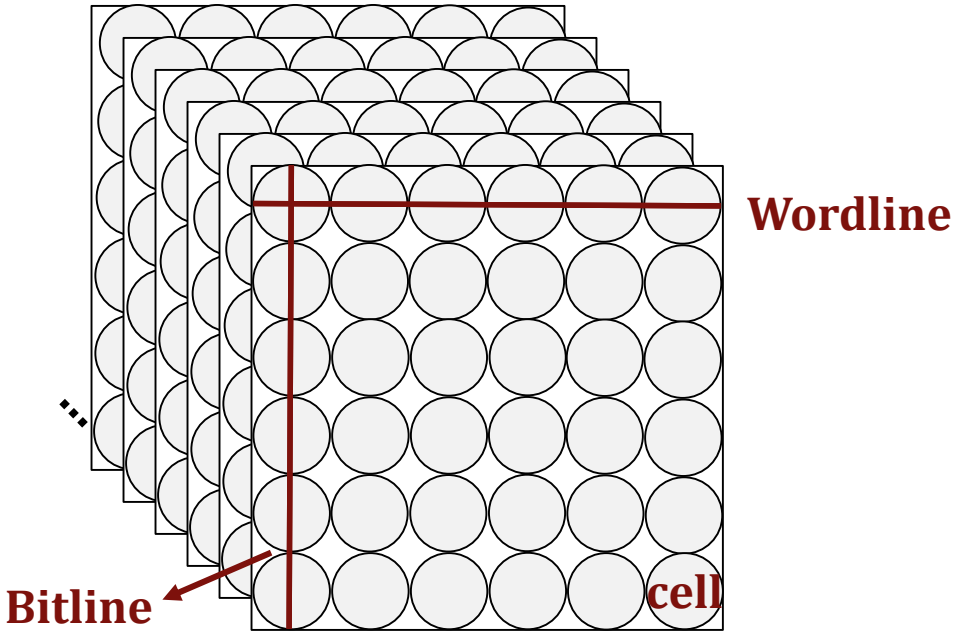
ReRAM array can compute convolution!

[M. Prezioso et al., nature 2015]
[Ali Shafiee et al., ISCA 2016]

Naïve baseline construction

Mobile system budget:

450mWatt, **1024** 128x128 arrays +128ADCs

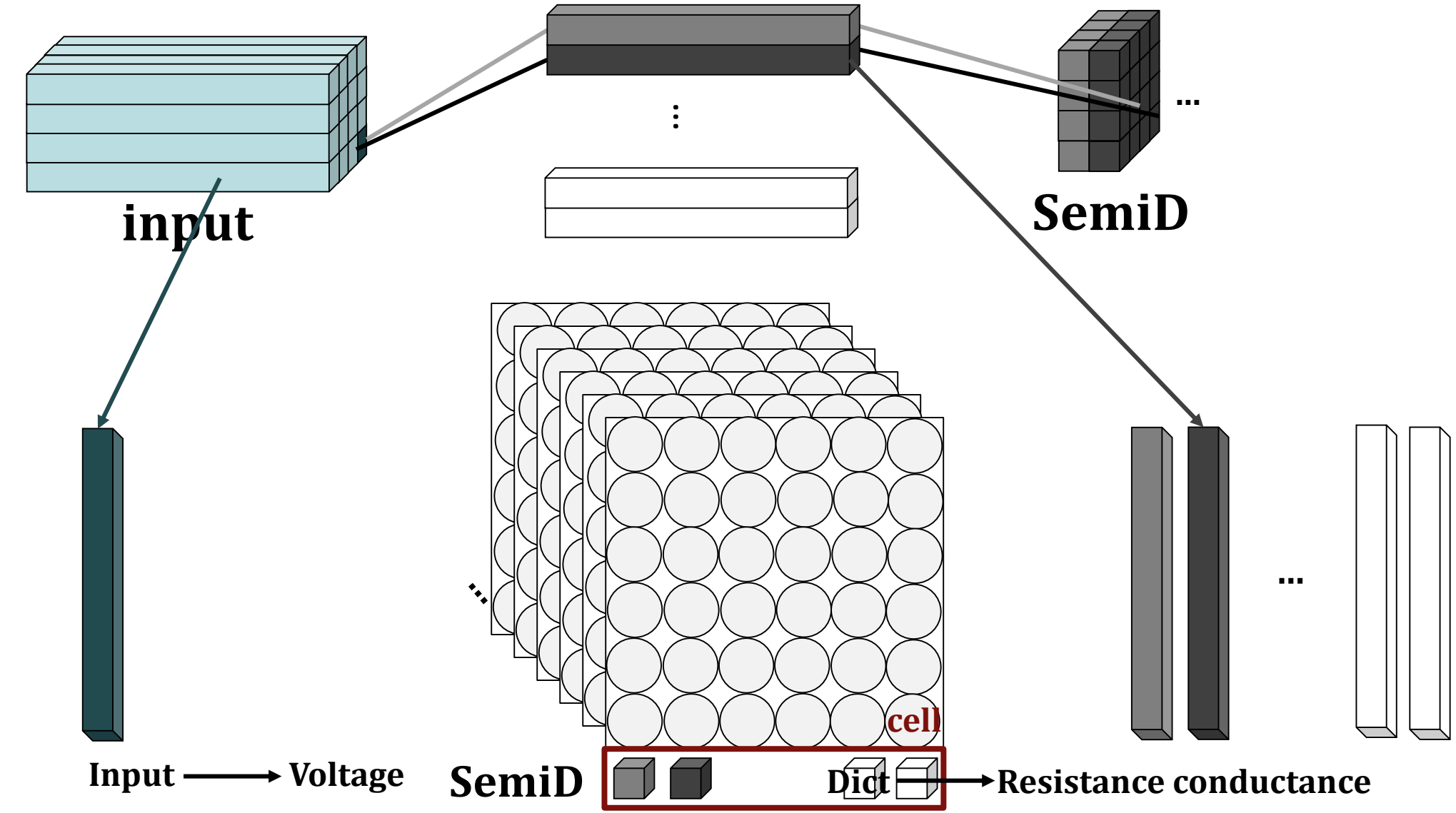


2D ReRAM arrays

[M. Prezioso et al., nature 2015]
[Ali Shafiee et al., ISCA 2016]

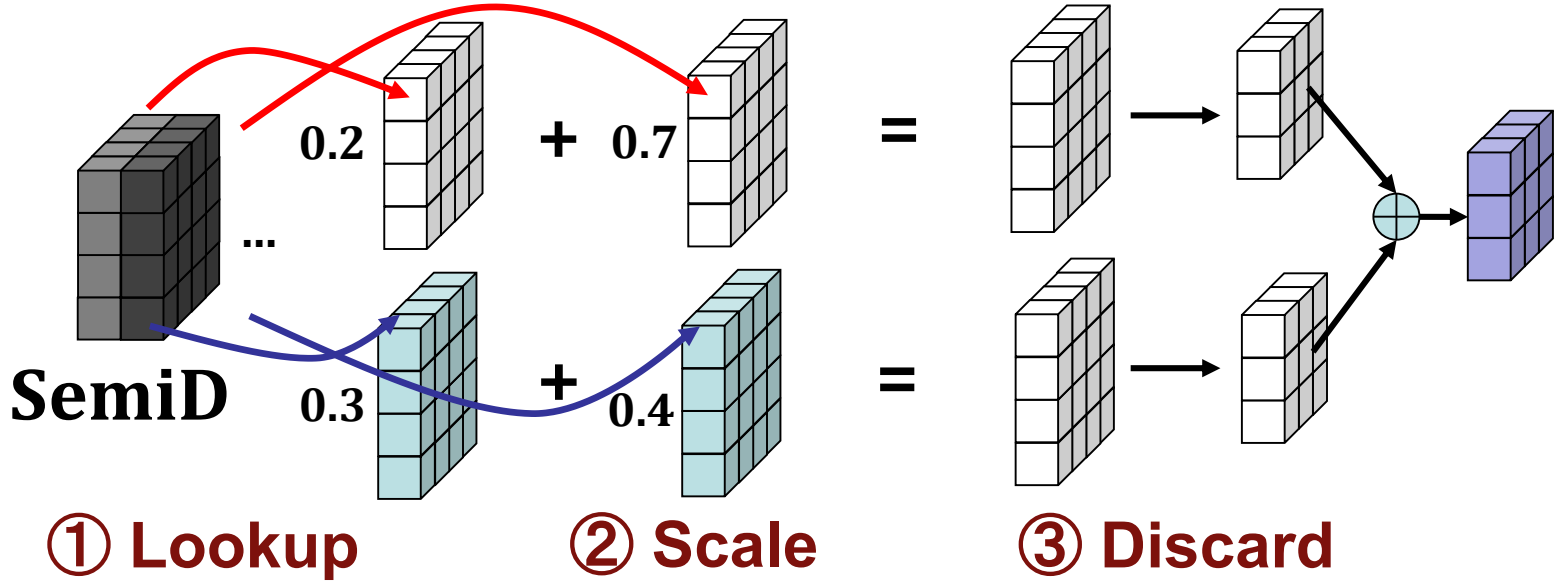
Naïve baseline construction

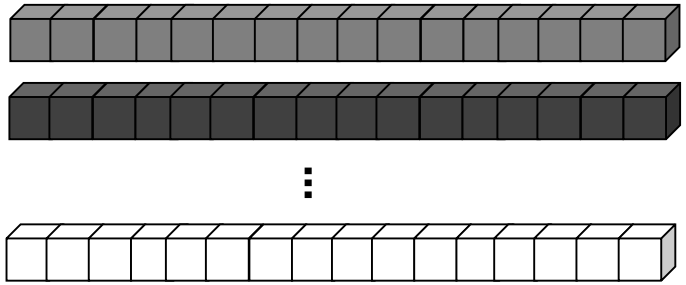
Small Convolution:



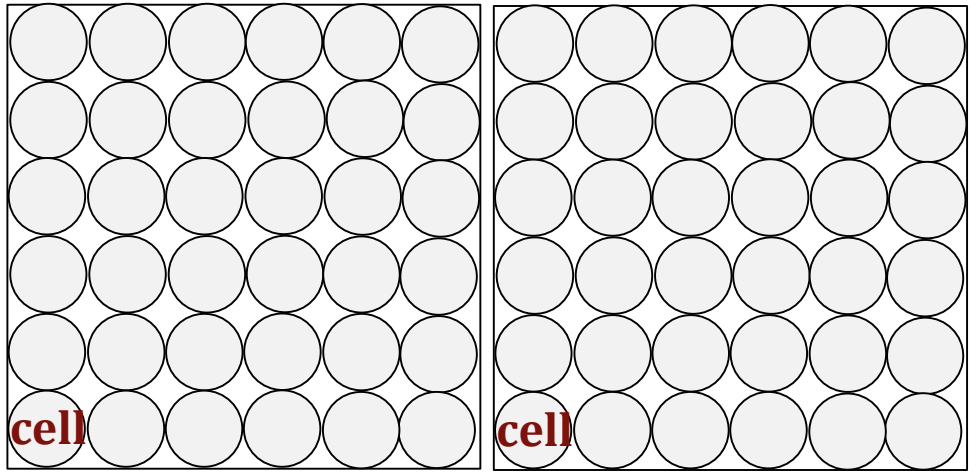
Naïve baseline construction

Lookup-scale-discard:



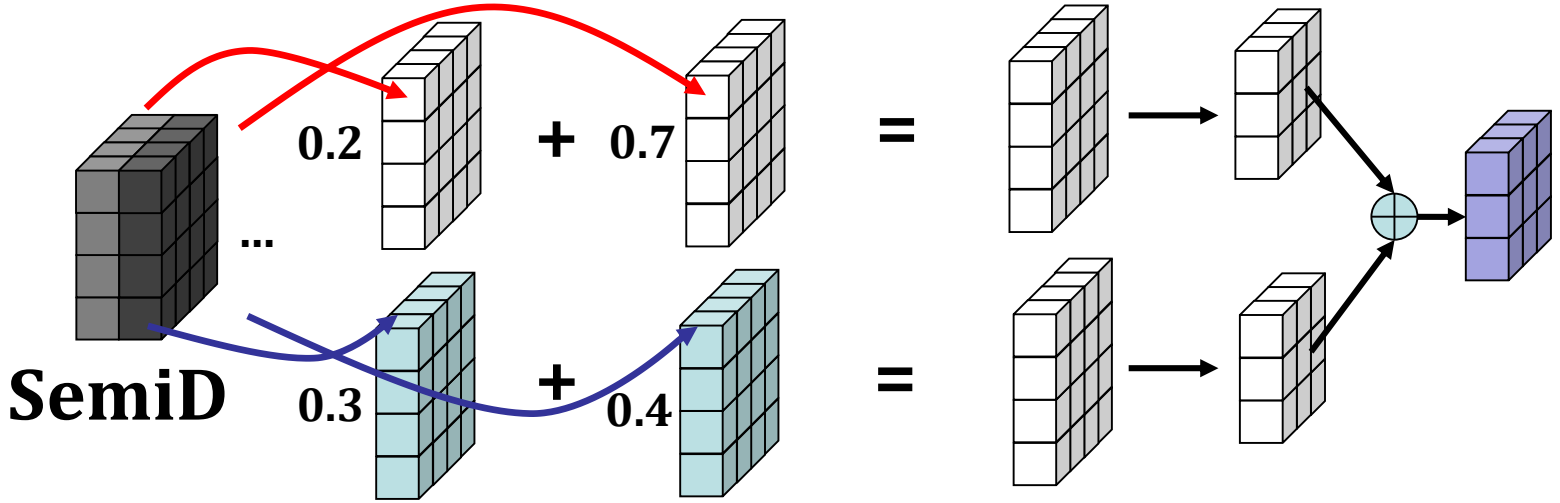


2D SemiD-> 1D conductance

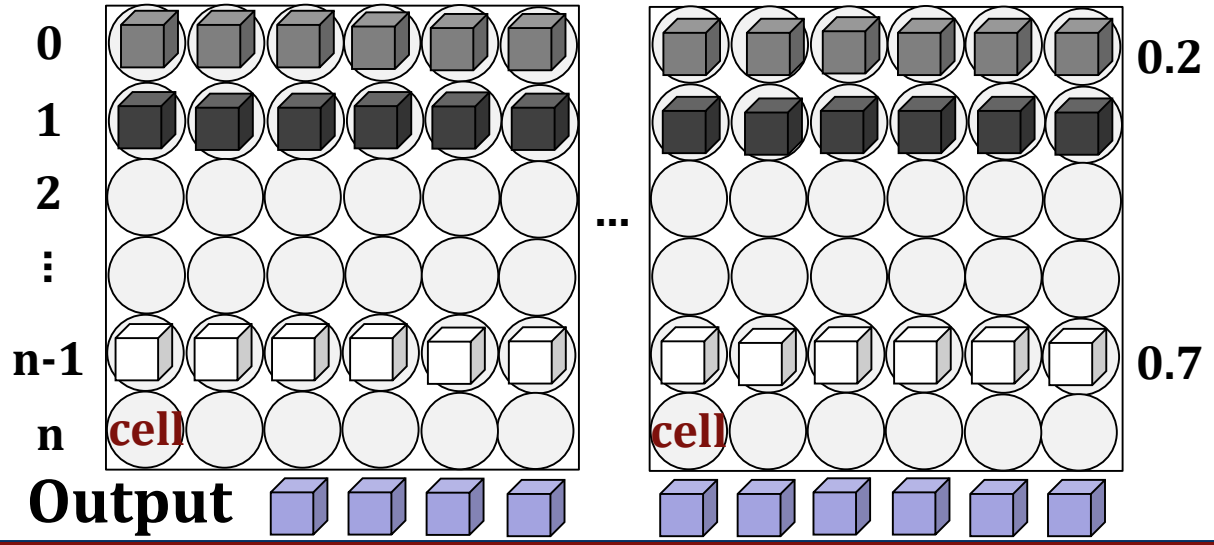


Naïve baseline construction

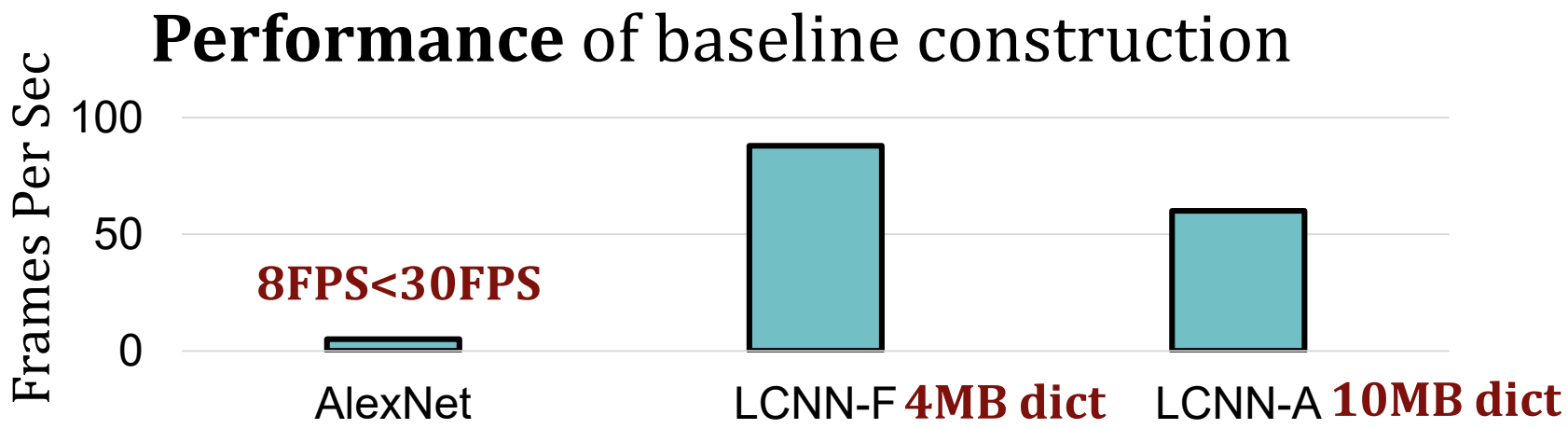
Lookup-scale-discard:

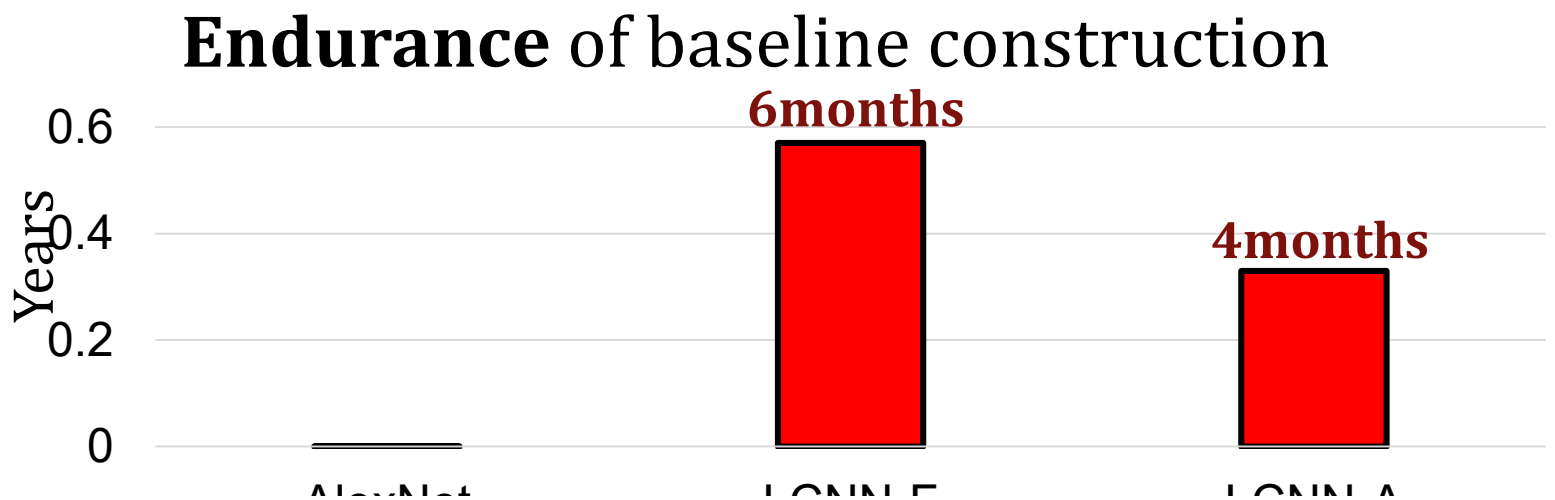


① Lookup ② Scale ③ Discard



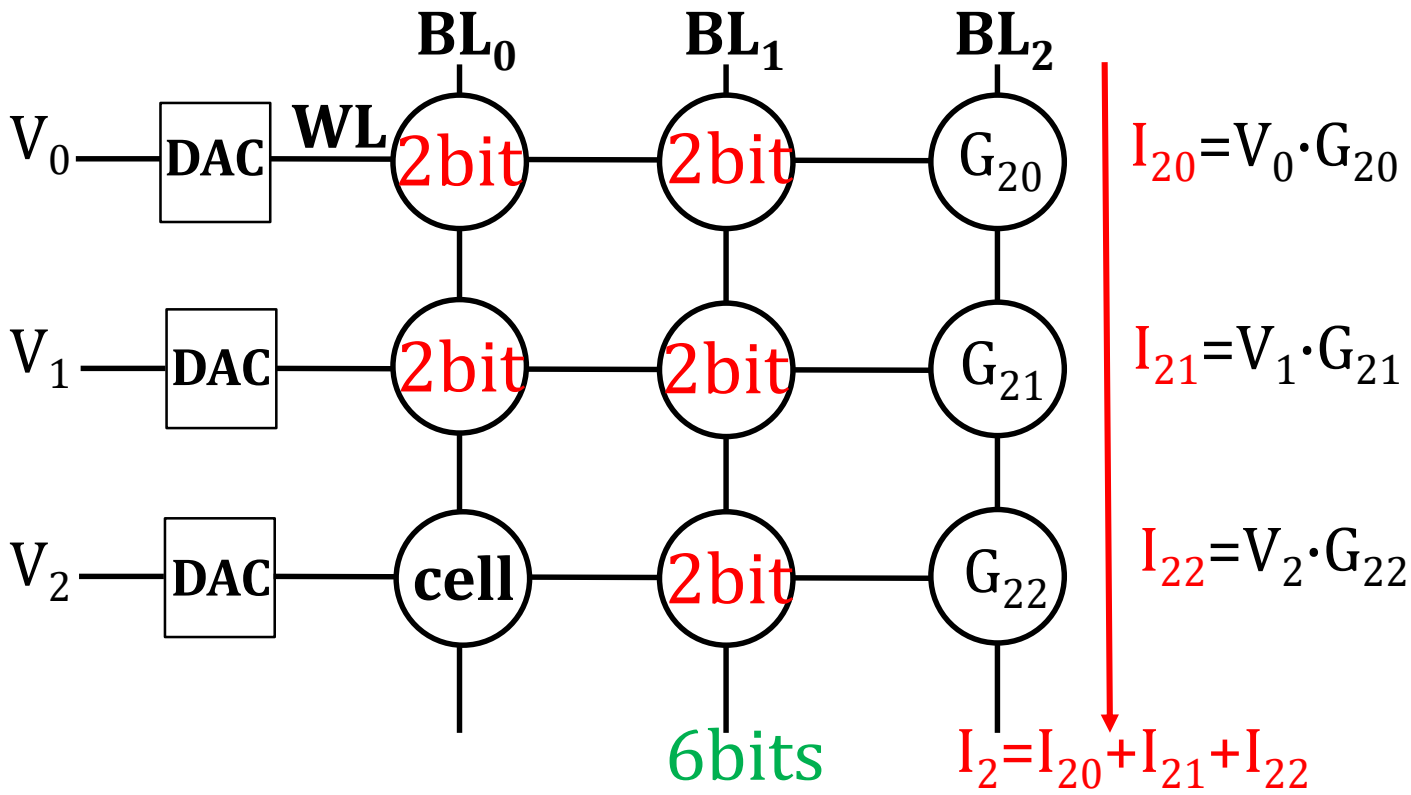
Naïve baseline construction





Challenge: endurance of system↑

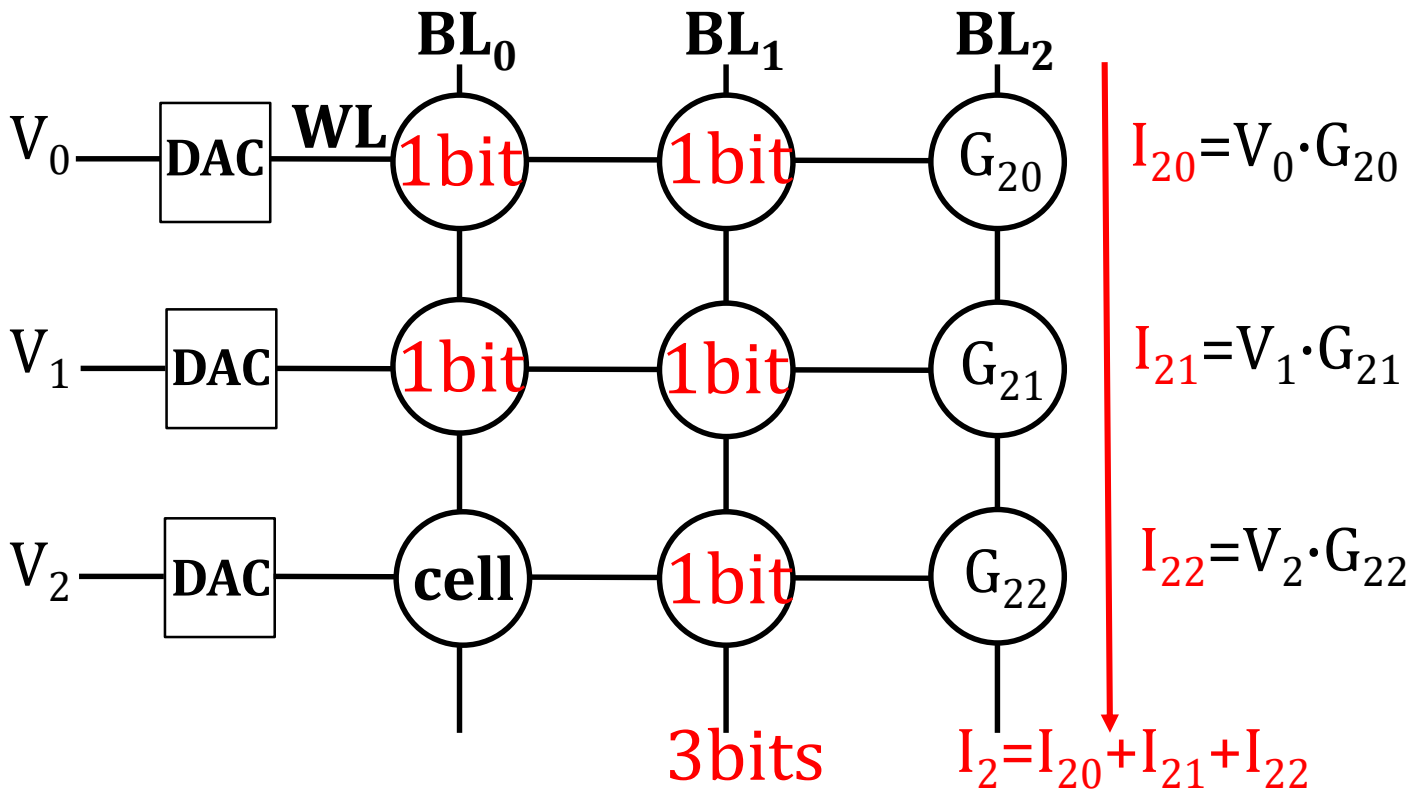
Solution: 3DICT



- 2 bit MLC cell, endurance bad (10^7 writes), throughput good (6 bits for 3-cell bitline)

[M. Prezioso et al., nature 2015]
[Ali Shafiee et al., ISCA 2016]

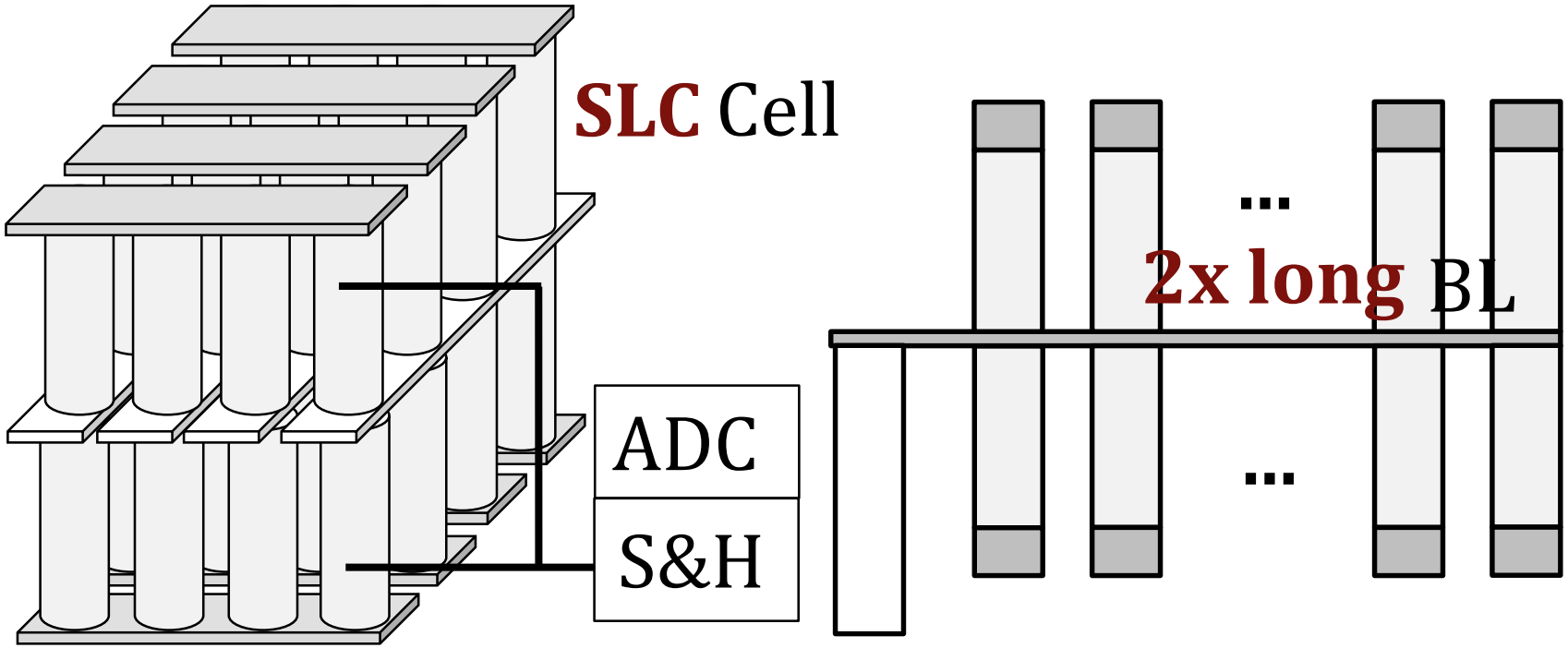
Solution: 3DICT



- 2 bit MLC cell, endurance bad (10^7 writes), throughput good (6 bits for 3-cell bitline)
- 1 bit SLC cell, endurance good (10^{11} writes), throughput bad (3 bits for 3-cell bitline)

[M. Prezioso et al., nature 2015]
[Ali Shafiee et al., ISCA 2016]

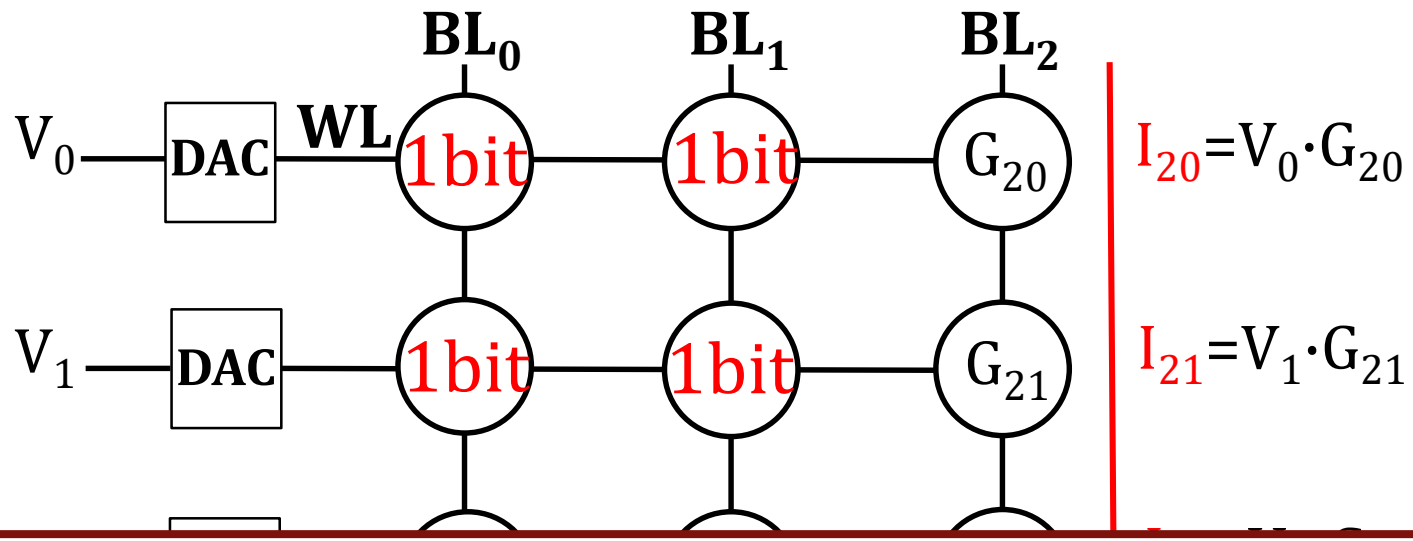
3D ReRAM dot-product engine



SLC cell--> endurance

2X long bitline -->throughput

Solution: 3DICT

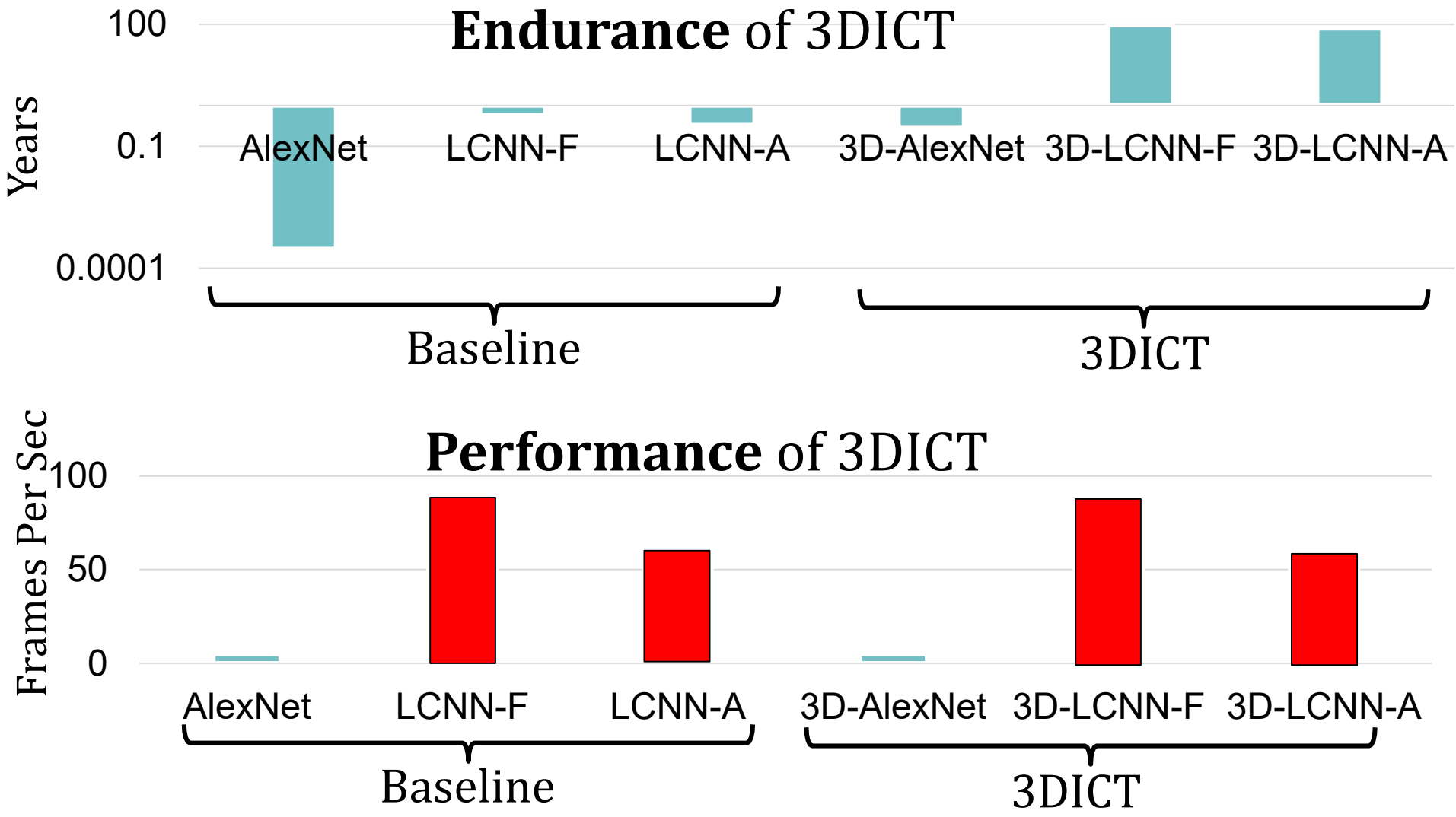


1-bit SLC cell + 2x long bitline(3D): throughput good + endurance good

- 2 bit MLC cell, endurance bad (10^7 writes), throughput good (6 bits for 3-cell bitline)
- 1 bit SLC cell, endurance good (10^{11} writes), throughput bad (3 bits for 3-cell bitline)

[M. Prezioso et al., nature 2015]
[Ali Shafiee et al., ISCA 2016]

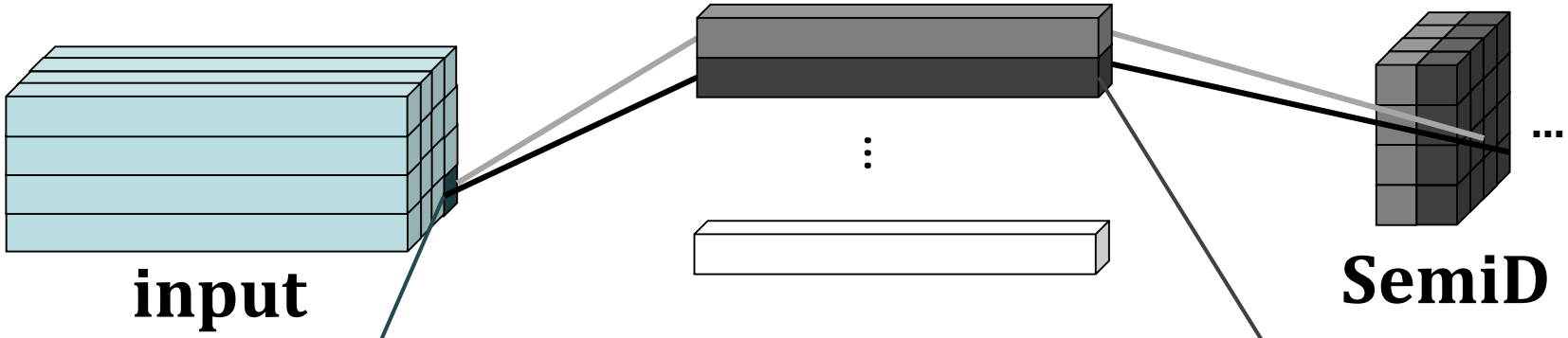
3DICT construction



Challenge: Performance is the same

Hybrid convolution

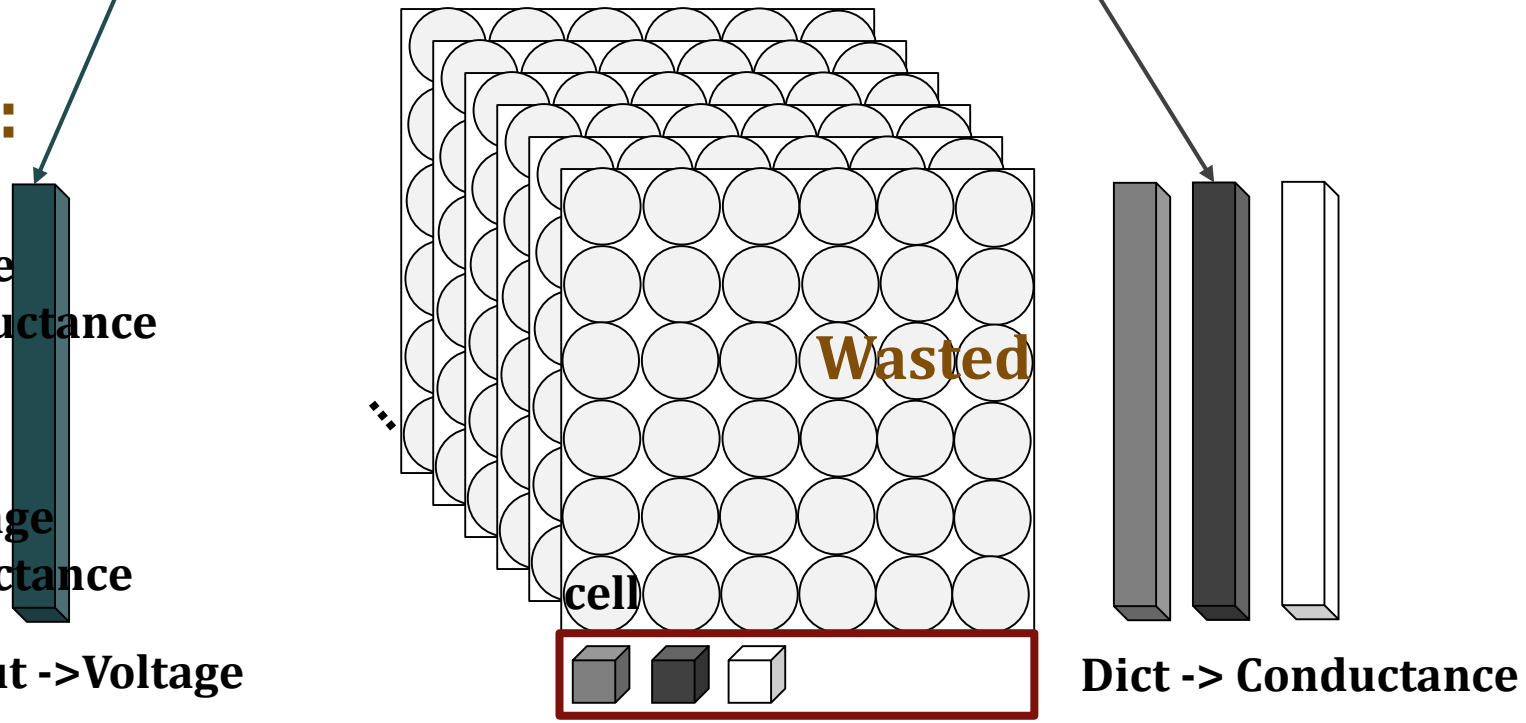
Small Convolution:



Solution:

Layer 1-5:
Dict ->Voltage
Input->Conductance

Layer 6-8:
Input ->Voltage
Dict->Conductance

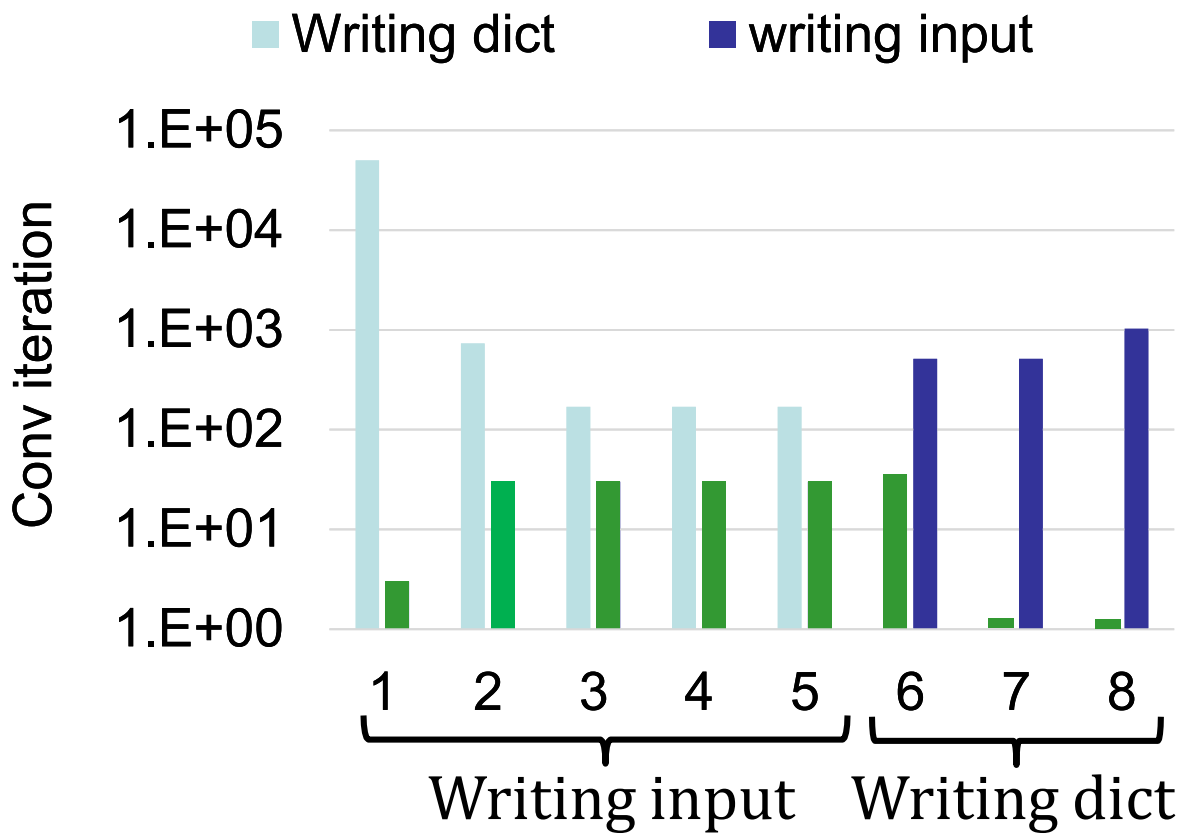


Hybrid convolution

Solution:

Layer 1-5:
Dict ->Voltage
Input->Conductance

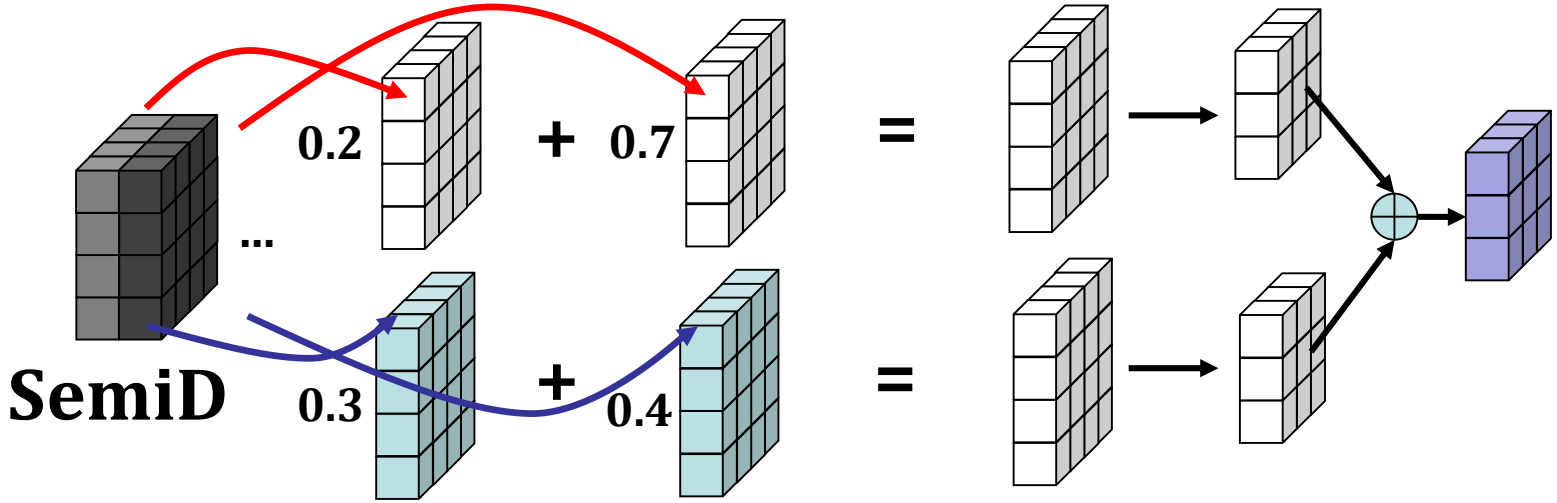
Layer 6-8:
Input ->Voltage
Dict->Conductance



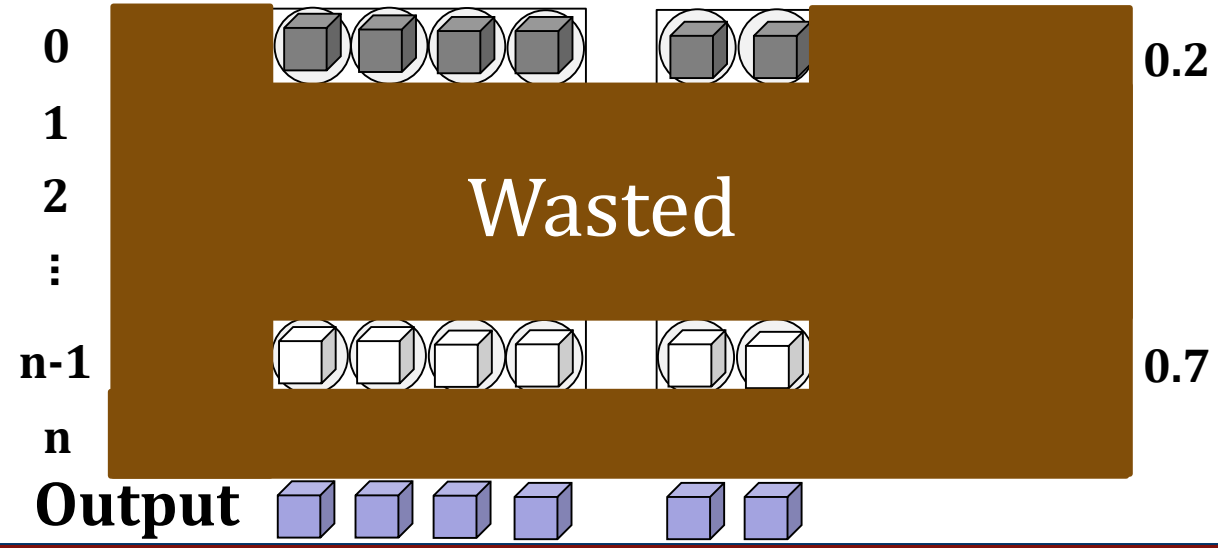
Hybrid convolution improves performance

Lookup-discard-scale

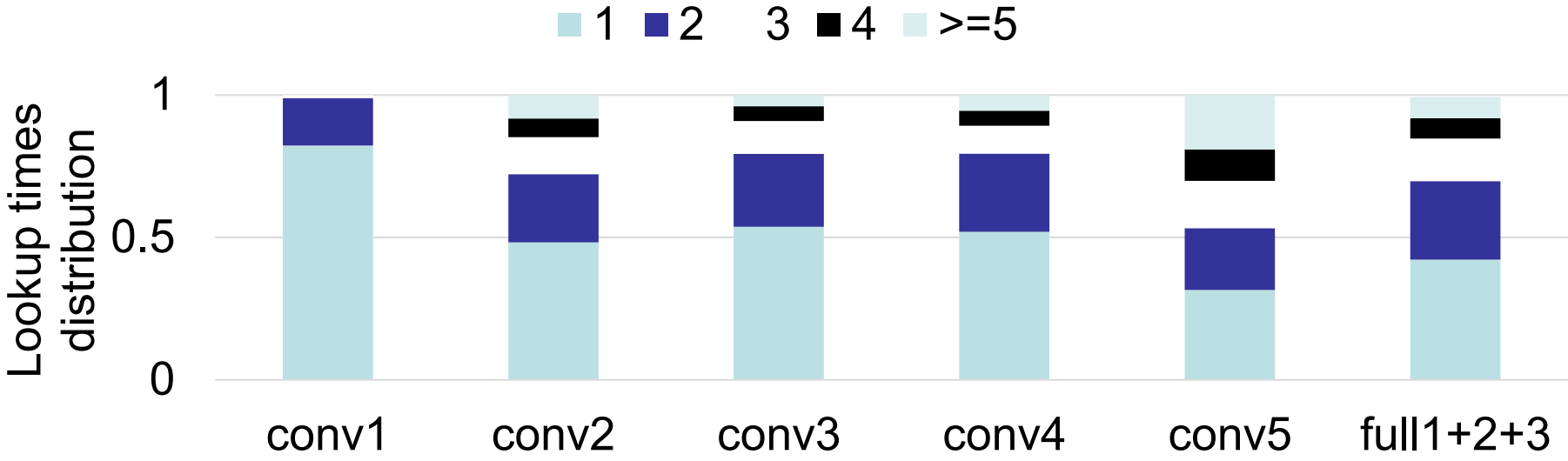
Lookup-scale-discard:



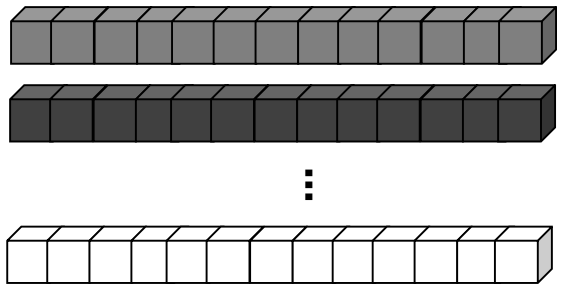
① Lookup ② Scale ③ Discard



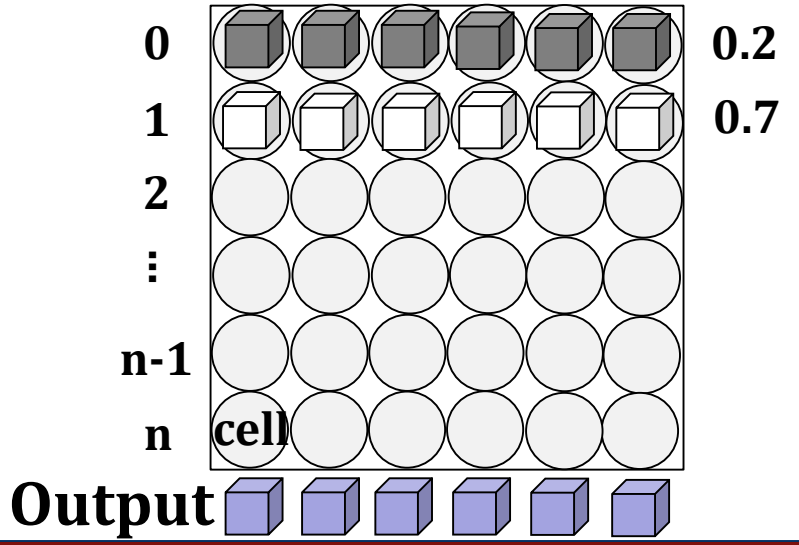
Lookup-discard-scale



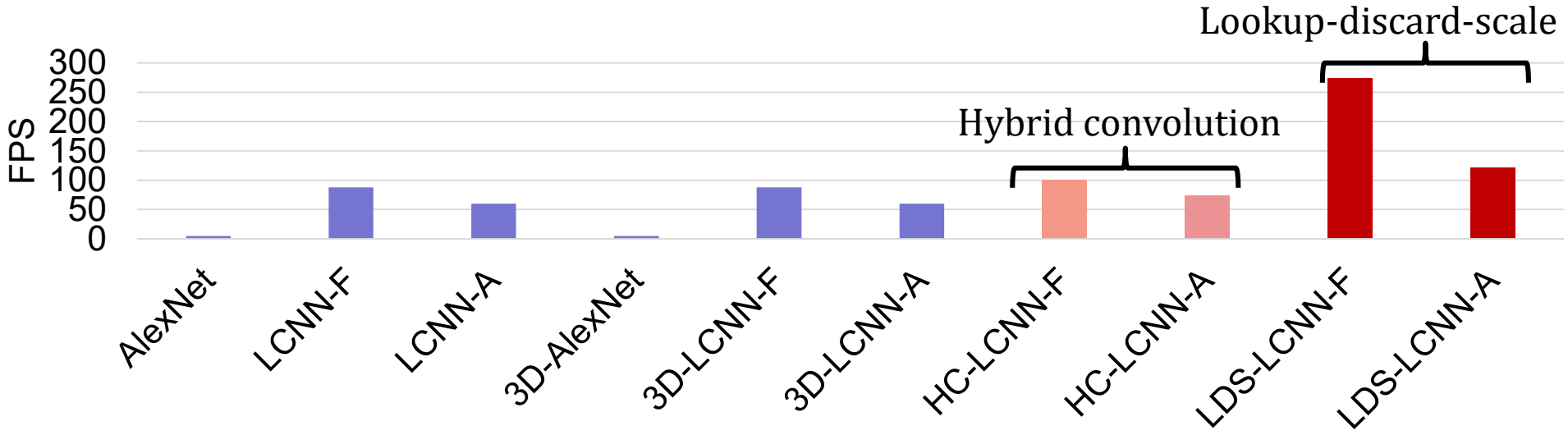
Solution: ① Lookup ② Discard ③ Scale

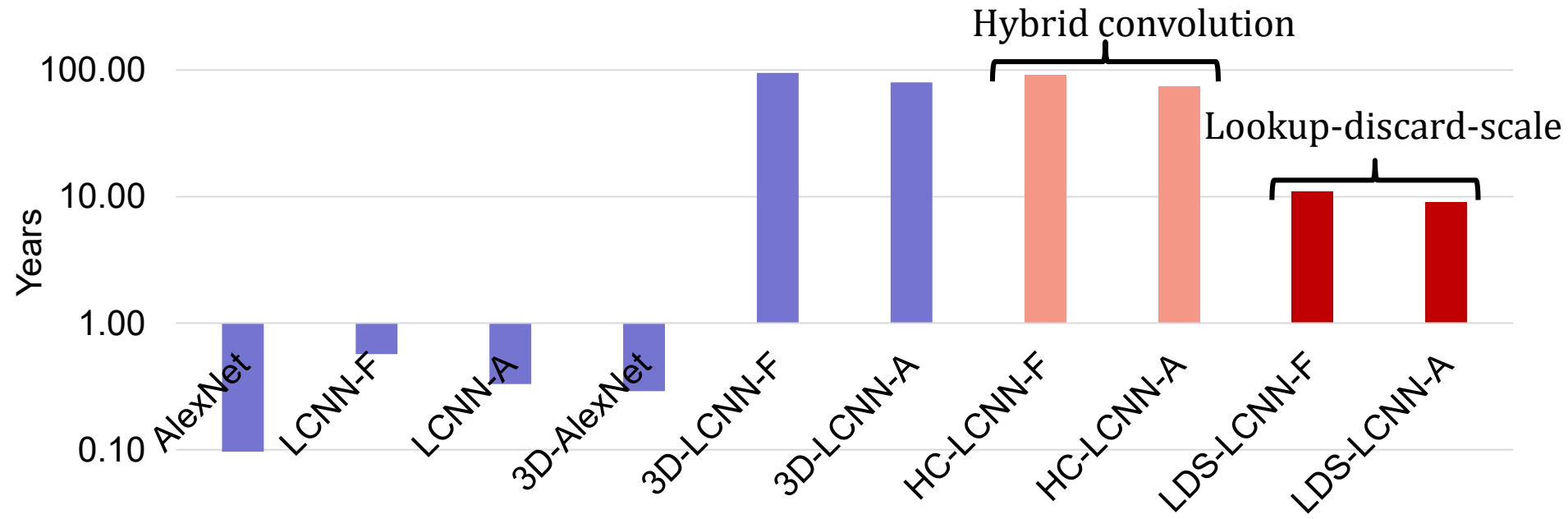


2D SemiD-> 1D voltage



3DICT construction with improvements

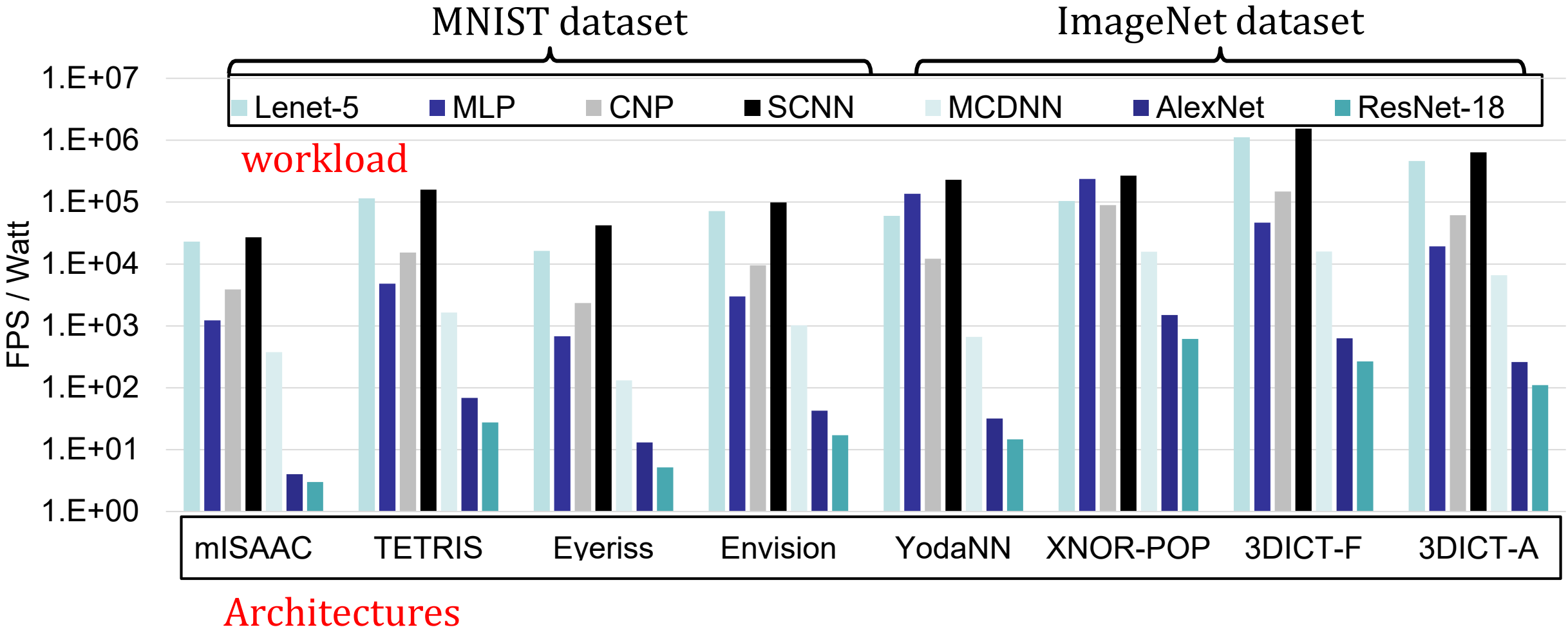




Outline

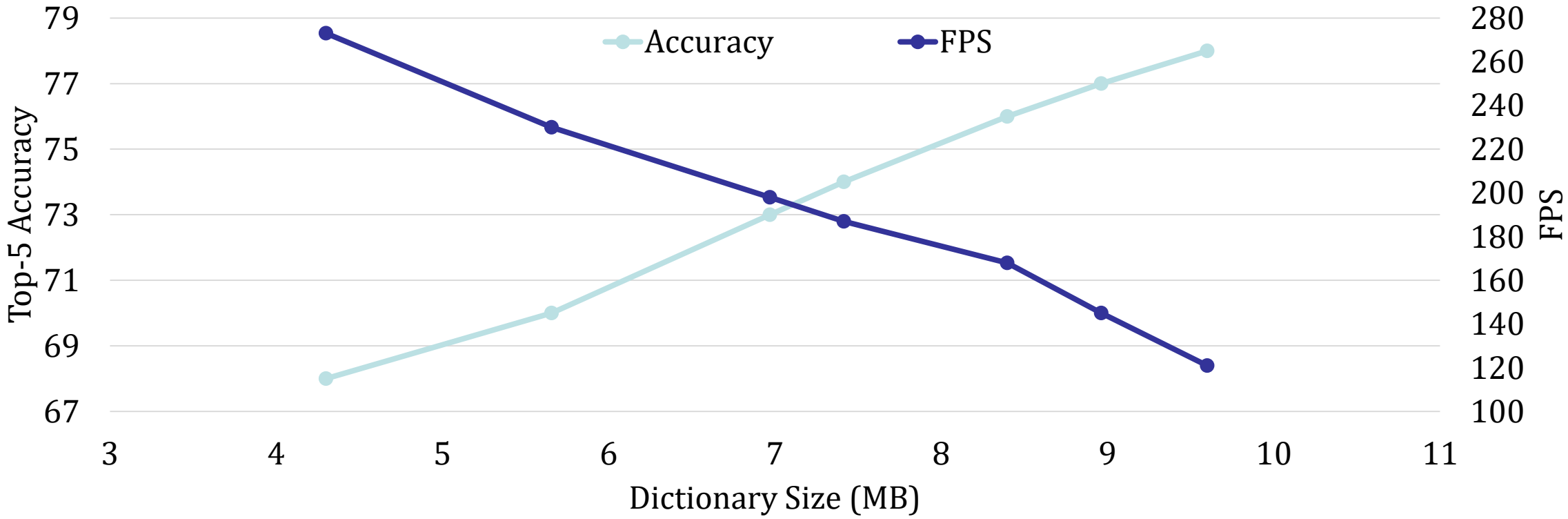
1. CNN and Lookup-based CNN
2. 3DICT
3. **Evaluation**
4. Conclusion

Evaluation



On average, CNN test performance/Watt by 13%~61x over prior architectures.

Evaluation



Critical tasks – accurate and slow
Non-critical tasks – inaccurate but fast

Conclusion

- **Motivation**:
  - CNN is **accurate** (> human) in mobile systems for object recognition, machine intelligence and so on.
  - Intelligent applications in mobile systems include critical tasks **(accuracy)** and non-critical tasks **(real-time). QoS is essential.**
  - **PIM** is an **energy-efficient** method for mobile systems.
- **Problem**: Current CNN in mobile system 1) is very slow, 2) no QoS support, 3) costs too much power
- **Goal**: To develop a **QoS** capable **PIM** architecture for mobile devices to support intelligent applications using **3D XPoint ReRAMs**.
- **3DICT**: 1) Lookup-based CNN (MAC#↓ weights#↓) 2) 2D ReRAM MLC endurance↓ ->2D ReRAM SLC throughput↓ ->3D ReRAM SLC)
- **Evaluation:**
  1. 3DICT can support QoS with 10-year life time.
  2. CNN test **performance per Watt** by **13%~61x** over prior architectures.

Thank you!

3DICT configuration

Name	Component	Spec	Power (mW)	Area (mm^2)
Hierarchical Array (HA)	ADC×1	8-bit 1.28GSps	2	0.0012
	DAC×256	1-bit, inverter	1	0.00025
	S&H×128	sample & hold	0.0125	0.000005
	Array×8	128 × 128 2-layer	0.3	0.0002
Sub-total			3.3125	0.001655
Multiply Accumulate Unit (MAU)	HA×8		26.5	0.01324
	S&A×4	shift & add	0.2	0.00024
	ReRAM $1KB$	I/O buffer	0.15	0.0005
Sub-total			26.85	0.01398
3DICT	MAU×16		429.6	0.224
	Sigmoid×2	activation	0.52	0.0006
	S&A×1	shift & add	0.4	0.00006
	MaxPool×1	pooling	0.31	0.00024
	Router and bus	connection	3	0.04
	ReRAM $1KB$	I/O buffer	0.15	0.0005
dict storage	ReRAM $64MB$	power gating	3.6 (0)	0.16
Total			433.98	0.4254

3DICT:

434mWatt, **1024** 128x128 2-layer arrays +128ADCs

Methodology

Workloads:

Name	DataBase	Topology	Top-5 Accuracy(%)			
			Orig	*3DICT-A*	*3DICT-F*	*XNOR*
LeNet-5	MNIST	3C,2S,1F	99.1	98.8	97.2	97.2
MLP	MNIST	5F	98.5	98.1	97.1	96.9
CNP	MNIST	3C,2S,1F	97.0	96.8	96.2	96.1
SCNN	MNIST	2C,2F	99.0	98.2	97.7	97.8
MCNN	MNIST	3C,3S,3F	96.8	96.1	95.7	95.7
AlexNet	ImageNet	5C,3S,2F	80.2	78.1	68.7	69.2
ResNet-18	ImageNet	18C,2S,1F	89.2	84.6	76.8	73.2

Architectures:

Name	Description	$Power_{acc}$	$Power_{mem}$
Envision [21]	complex CNNs	$62mW$ [21]	1.91W [17]
YodaNN [1]	BinaryConnect	$248mW$ [1]	1.91W [17]
Eyeriss [4]	complex CNNs	$278mW$ [4]	1.91W [17]
TETRIS [11]	HMC PIM	$8.42W$ [11]	0
XNOR-POP [17]	XNOR-Net PIM	$2.15W$ [17]	0
mISAAC	mobile ReRAM CNN	$435.58mW$	0
3DICT	ReRAM LCNN	$435.58mW$	0

[Torch7 & Caffe]

[FODLAM by Cornell]

3D ReRAM dot-product engine