NEURAL INFORMATION PROCESSING SYSTEMS

DEX: Data Channel Extension for Efficient CNN Inference on Tiny AI Accelerators

Taesik Gong

Fahim Kawsar

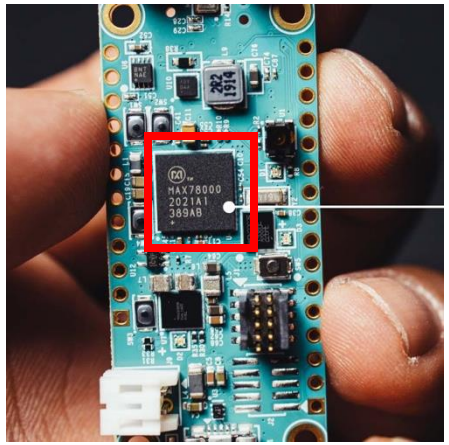
Chulhong Min

NOKIA BELL LABS UNIST University of Glasgow

Tiny AI Accelerators: New On-Device AI Platforms

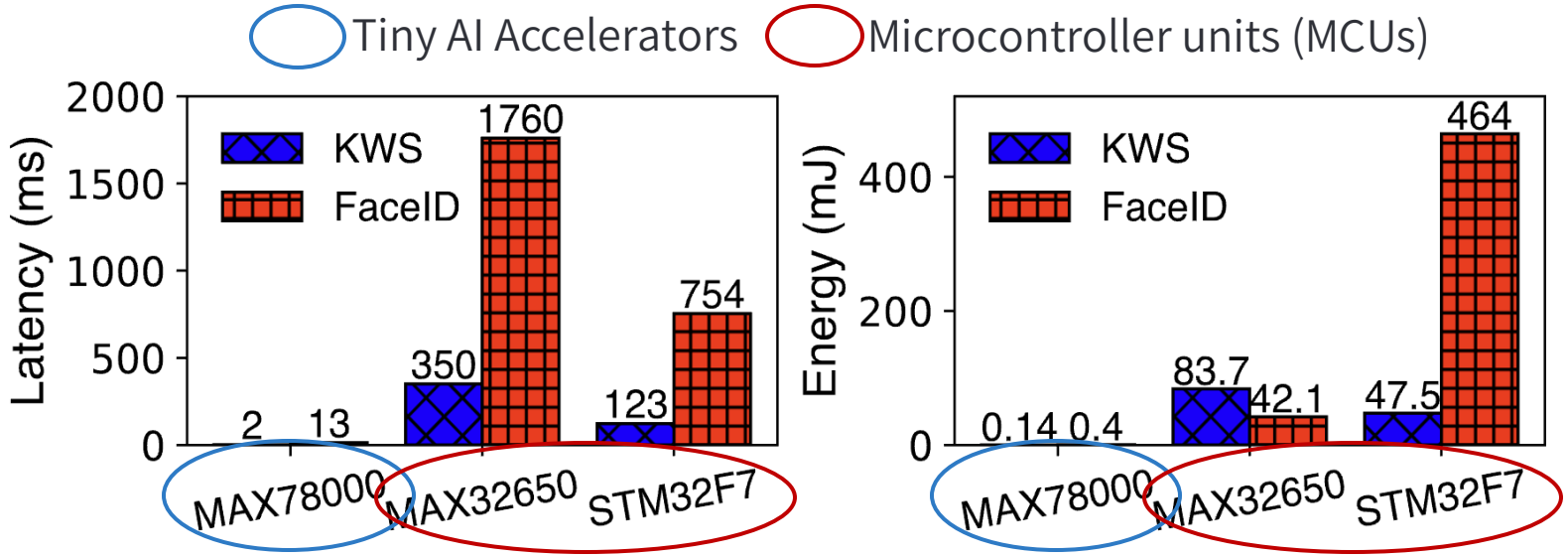
Tiny AI Accelerator

(MAX78000, 8mm × 8mm)



Omnibuds by Bell Labs

https://omnibuds.tech/

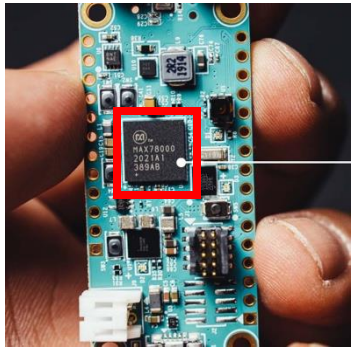


- 62~175× faster inference
- 105~1160× less energy consumption

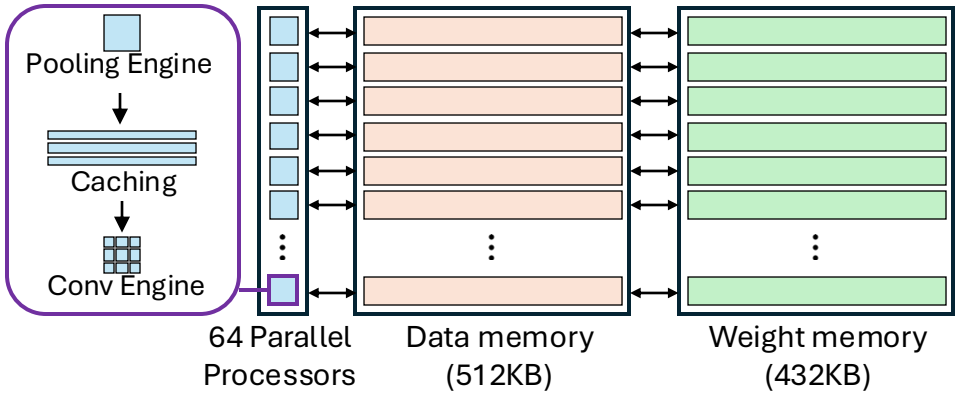
→ Opportunity of (1) reduced latency, (2) lower power cost, and (3) improved privacy for on-device AI

Why Are Tiny AI Accelerators Fast? Parallelization

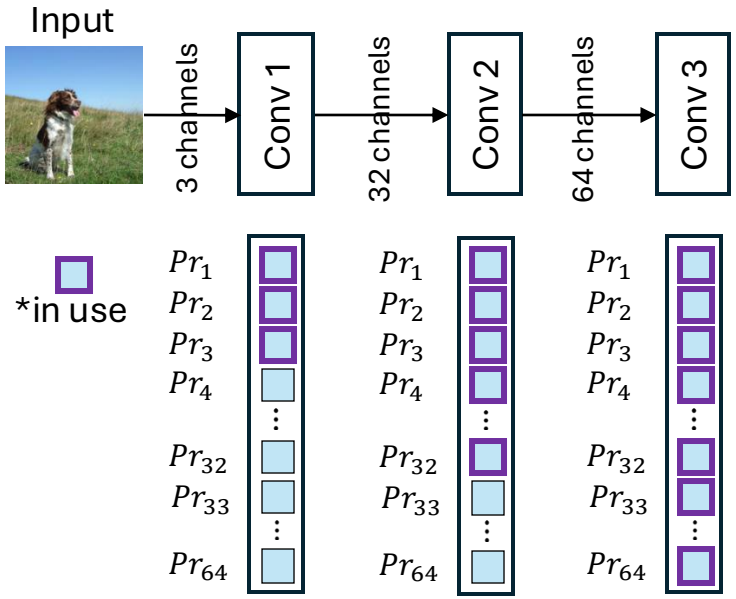
Architecture of Tiny AI Accelerator



MAX78000



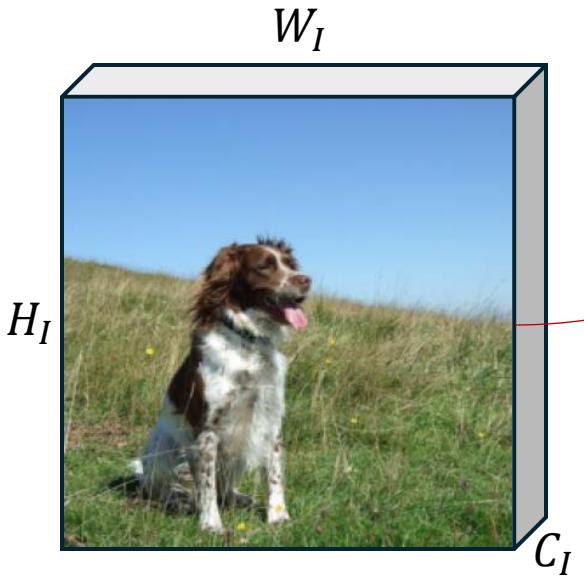
Parallelization across channels



Parallel data access and processing are the keys to fast inference

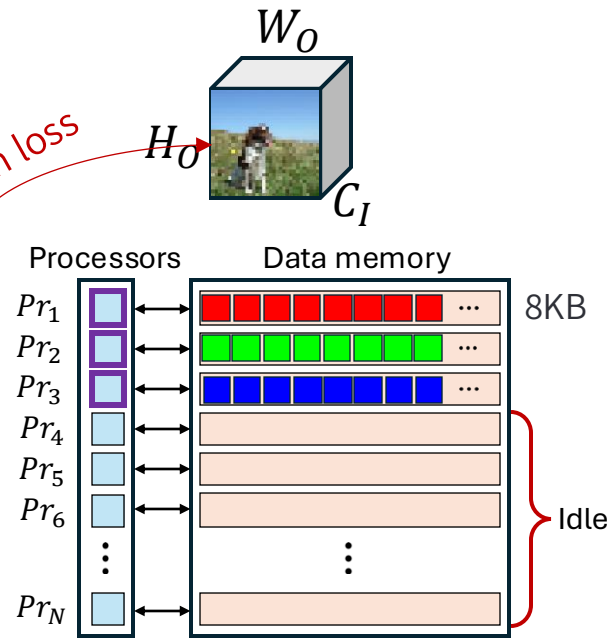
Tiny AI Accelerator Lacks Data Memory

Original image I



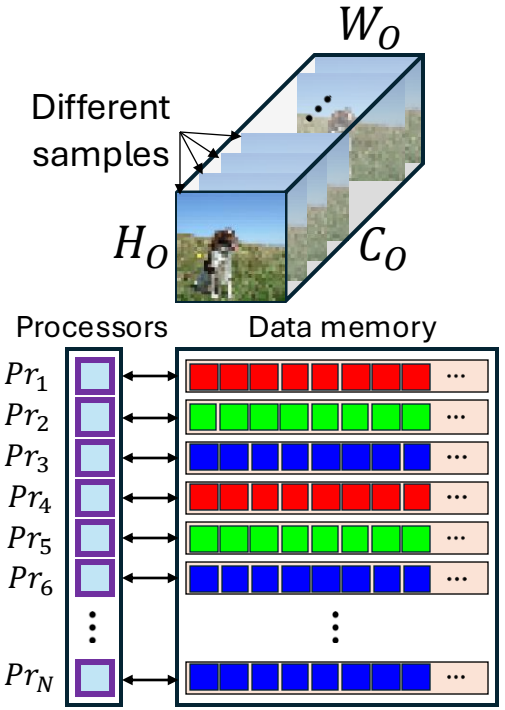
224x224x3
= 50KB * 3 channels
*exceeds data memory limit

Downsampling



Underutilized
processors and memory

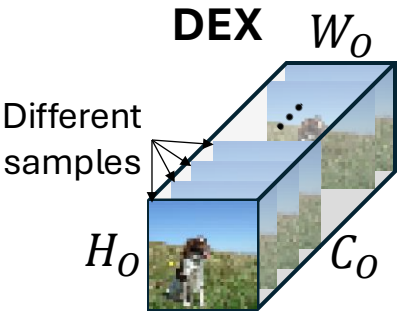
DEX (ours)



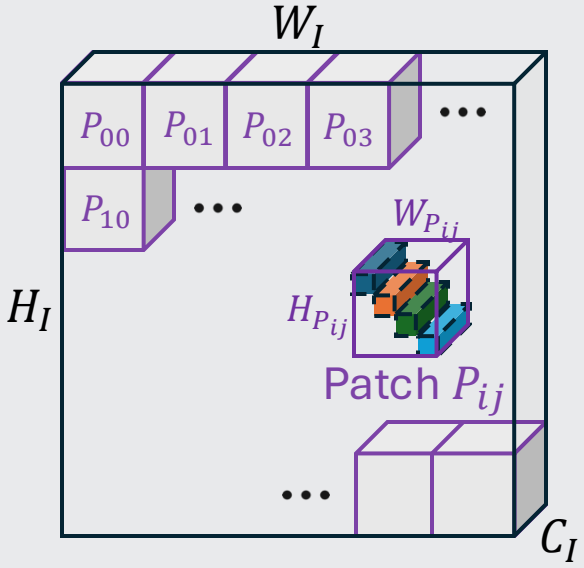
Improves accuracy with **additional spatial information**
with the same inference latency

DEX: Data Channel Extension for Tiny AI Accelerators

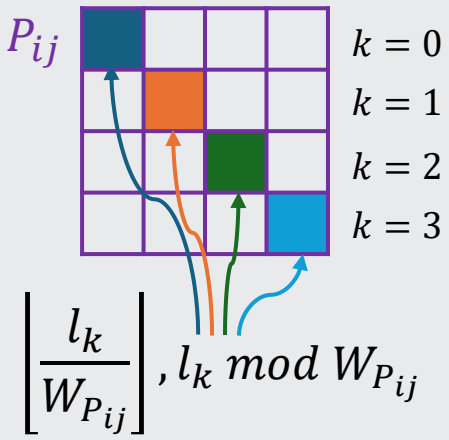
incorporates additional **spatial information** across channels



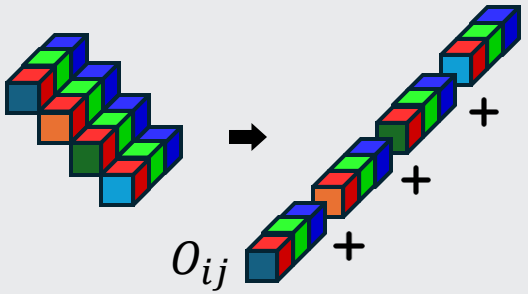
Original image I



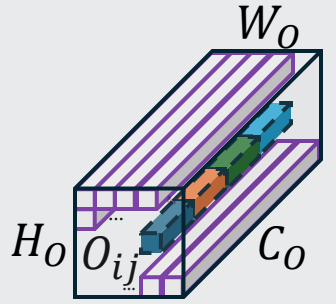
Patch-wise even sampling
(e.g., $K = 4$)



Channel-wise stacking
(e.g., $C_O = 12$)



Output image O



DEX: Result

Accuracy

Dataset	Method	SimpleNet	WideNet	EfficientNetV2	MobileNetV2	AVG (%)
ImageNette	Downsampling	57.8 ± 1.2	61.8 ± 0.2	51.3 ± 0.5	62.0 ± 0.7	58.2
	CoordConv	58.0 ± 1.1	61.7 ± 0.2	51.9 ± 0.1	61.6 ± 0.3	58.3
	CoordConv (r)	55.4 ± 1.5	61.4 ± 0.2	51.7 ± 1.0	61.2 ± 1.1	57.4
	DEX (ours)	**61.4 ± 0.6**	**65.6 ± 0.6**	**56.8 ± 0.5**	**64.4 ± 0.6**	**62.0**
Caltech101	Downsampling	54.6 ± 2.1	55.8 ± 1.2	38.6 ± 0.9	51.4 ± 1.6	50.1
	CoordConv	53.8 ± 1.6	56.5 ± 0.1	38.7 ± 0.2	49.8 ± 0.5	49.7
	CoordConv (r)	52.7 ± 0.5	56.0 ± 1.7	38.2 ± 1.0	49.7 ± 1.2	49.1
	DEX (ours)	**56.9 ± 1.3**	**61.1 ± 1.4**	**45.9 ± 1.9**	**53.3 ± 1.7**	**54.3**
Caltech256	Downsampling	19.8 ± 0.6	20.8 ± 0.5	14.7 ± 0.4	22.4 ± 1.0	19.4
	CoordConv	19.8 ± 0.5	21.3 ± 0.8	14.8 ± 0.8	22.7 ± 0.8	19.6
	CoordConv (r)	20.0 ± 1.6	20.9 ± 0.6	14.5 ± 0.3	22.7 ± 0.4	19.5
	DEX (ours)	**22.8 ± 0.5**	**22.9 ± 0.9**	**18.3 ± 0.9**	**26.3 ± 0.5**	**22.6**
Food101	Downsampling	16.0 ± 0.4	17.7 ± 0.7	12.1 ± 0.2	22.4 ± 0.6	17.1
	CoordConv	16.1 ± 0.8	17.7 ± 0.3	12.0 ± 0.1	21.7 ± 0.3	16.9
	CoordConv (r)	16.3 ± 0.4	17.3 ± 0.6	12.0 ± 0.6	20.9 ± 0.3	16.6
	DEX (ours)	**18.4 ± 0.4**	**20.9 ± 0.4**	**16.4 ± 0.1**	**23.3 ± 1.1**	**19.8**

Latency

Model	Method	InputChan	Size (KB)	InfoRatio (×)	ProcUtil (%)	Latency (μs)
SimpleNet	Downsampling	3	162.6	1.0	4.7	2592 ± 1
	CoordConv	5	162.9	1.0	7.8	2592 ± 2
	CoordConv (r)	6	163.0	1.0	9.4	2592 ± 2
	DEX (ours)	64	171.2	21.3	100.0	2591 ± 1
WideNet	Downsampling	3	306.4	1.0	4.7	3820 ± 1
	CoordConv	5	306.9	1.0	7.8	3820 ± 0
	CoordConv (r)	6	307.1	1.0	9.4	3819 ± 1
	DEX (ours)	64	319.3	21.3	100.0	3818 ± 1
EfficientNetV2	Downsampling	3	742.4	1.0	4.7	11688 ± 2
	CoordConv	5	743.0	1.0	7.8	11685 ± 3
	CoordConv (r)	6	743.2	1.0	9.4	11689 ± 1
	DEX (ours)	64	759.6	21.3	100.0	11690 ± 2
MobileNetV2	Downsampling	3	1317.8	1.0	4.7	3553 ± 4
	CoordConv	5	1318.2	1.0	7.8	3554 ± 1
	CoordConv (r)	6	1318.4	1.0	9.4	3554 ± 2
	DEX (ours)	64	1330.7	21.3	100.0	3552 ± 3

DEX improves accuracy by **3.5%p**
while keeping the **inference latency the same** on the tiny AI accelerator

Code:https://github.com/Nokia-Bell-Labs/data-channel-extension
Contact: taesik.gong@unist.ac.kr