Imperial College London

DEF: Differential Encoding of Featuremaps for Low Power Convolutional Neural Network Accelerators

Alexander Montgomerie-Corcoran and Christos-Savvas Bouganis

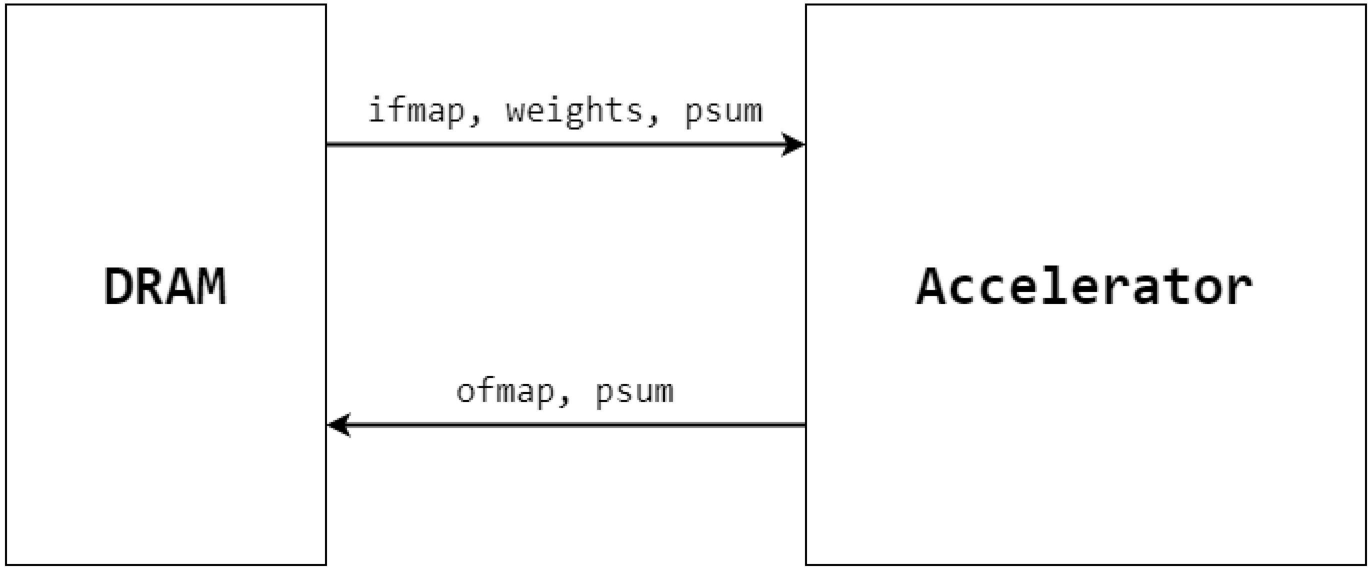
intelligent Digital Systems Lab

Dept. of Electrical and Electronic Engineering

www.imperial.ac.uk/idsl

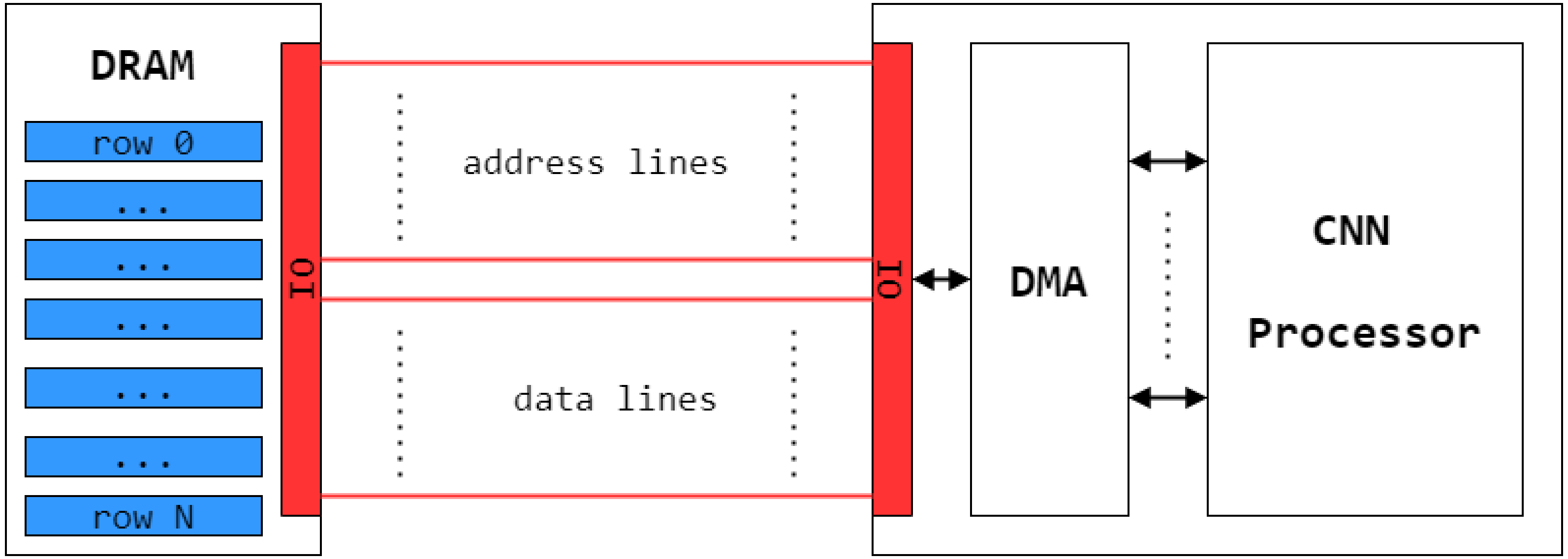
AlexMontgomerie/def

Motivation



- The energy of off-chip memory accesses are **200x** greater than on-chip MAC[1]
- Memory sub-system can consume **2-4x** more power than the accelerator[2]

Motivation: Overview of Memory Subsystem



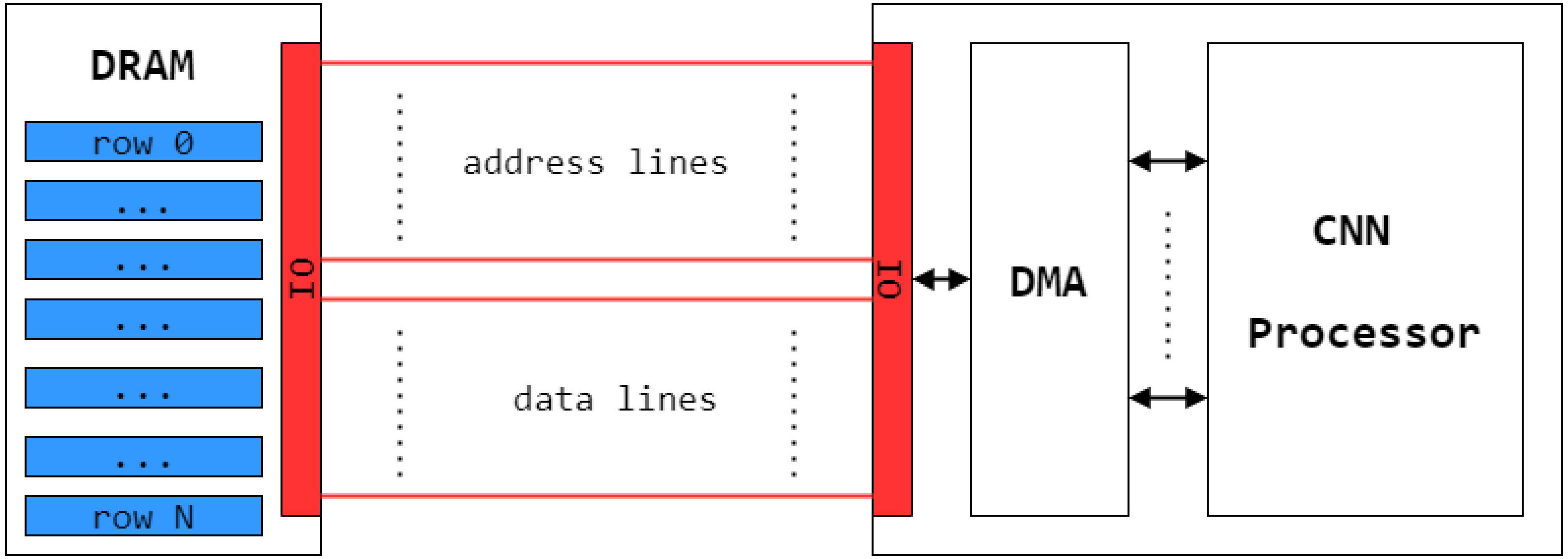
IO Power:
- Dynamic
- Bus-line activity
- Bus width

DRAM Power:
- Active rows
- Commands

Data Dependent

Access Dependent

Motivation: Overview of Memory Subsystem



IO Power:
- Dynamic
- Bus-line activity
- Bus width

Data Dependent

DRAM Power:
- Active rows
- Commands

Access Dependent

Motivation: IO Power Consumption

IO Power Components:

- IO pad dynamic power
- Termination power
- Interconnect power
- PHY power

$$P_{dynamic} = n \cdot V_{dd}^2 \cdot f_{clk} \cdot C \cdot a$$

n	Width of Bus
V_{dd}	Supply Voltage
f_{clk}	Bus Clock Frequency
C	Equivalent Capacitance of Bus Line
a	Activity along Bus Lines

Motivation: IO Power Consumption

How can dynamic power be reduced?

Voltage Scaling	• Potentially introduces errors • Platform-dependant
Frequency Scaling	• Impacts performance of accelerator
Reducing Large Capacitances	• Not really achievable for off-chip IO
Reduce Activity	• *Why Not?*

Impact

Imperial College London

Main outcomes of the work

- A novel, domain-specific activity encoding scheme (**DEF**) for CNN Accelerators which outperforms prior state-of-the-art work for this application
- Up to **50%** activity reduction across a range of CNN applications
- Power savings of **6%** for an example CNN Accelerator for the whole memory subsystem
- No temporal or spatial redundancy

Background

- Activity Encoding Schemes
- CNN Accelerators

Background: Activity Encoding Schemes

Encoding Scheme	Spatial Redundancy (bits)	Temporal Redundancy (cycles)	Description
Bus Invert (**BI**)[3]	1	0	Invert bus when activity is past threshold
Adaptive Bus Encoding (**ABE**)[4]	1	1	Reduce activity of clusters of lines compared to a basis line
Probability-Based Mapping (**PBM**)[5]	0	0	Map frequent values to low activity
Adaptive Word Reordering (**AWR**)[6]	$log_2 N$	0	Reorder words to minimize activity

- Temporal and Spatial redundancy carry information about the encoding scheme
- Spatial redundancy can be constrained by physical bus limits
- Temporal redundancy can impact performance

Background: CNN Accelerators

- CNN accelerators typically used fixed-point representation
- They have different dataflow schemes, which tell us about local reuse within the accelerator
- Some CNN Accelerators now employ encoding schemes in order to increase energy efficiency

CNN Accelerator	Quantization (bits)	Dataflow	Notes
EYERISS[1]	16	Row Stationary	Uses Run-Length Encoding (**RLE**) for feature-maps
eCNN[7]	8	Weight Stationary	Uses **Huffman** Encoding for feature-maps
fpgaConvNet[2,8]	16	Weight Stationary	Streaming architecture, with power-awareness

Methodology

- Decorrelating function
- Sign-Magnitude representation
- Difference encoding

Methodology: Defining the Problem

What is Switching Activity?

The average number of transitions in along a wire

The switching activity for a stream of integers $\boldsymbol{x}$, where $\boldsymbol{x} = [x_0 \ldots x_m]$ is defined as

$$a(\boldsymbol{x}) = \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=0}^{n} x_{i,j} \oplus x_{i-1,j}$$

- How do we minimize activity?
- No easy arithmetic solution

Methodology: Defining the Problem

Change the Problem

We can introduce a decorrelating function, that maps a stream of bits to transitions. This can be described as,

$$d(x_i) = x_i \oplus d(x_{i-1}), d(x_0) = 0$$

This will change our view on activity, which is now described for a decorrelated stream

$$\hat{a}(\boldsymbol{x}) = a(d(\boldsymbol{x})) = \frac{1}{m \cdot n} \sum_{i=0}^{m} \sum_{j=0}^{n} x_{i,j}$$

<u>Now, the optimization goal is to minimize the number of “1” bits in the stream</u>

Methodology: Sign Magnitude Representation

How are fixed-point numbers represented?

- Signed fixed-point numbers are represented Two's Complement (**TC**) integers
- **TC** Integers have some undesirable properties
- Batch Normalisation Layers mean that feature-maps are typically normally distributed

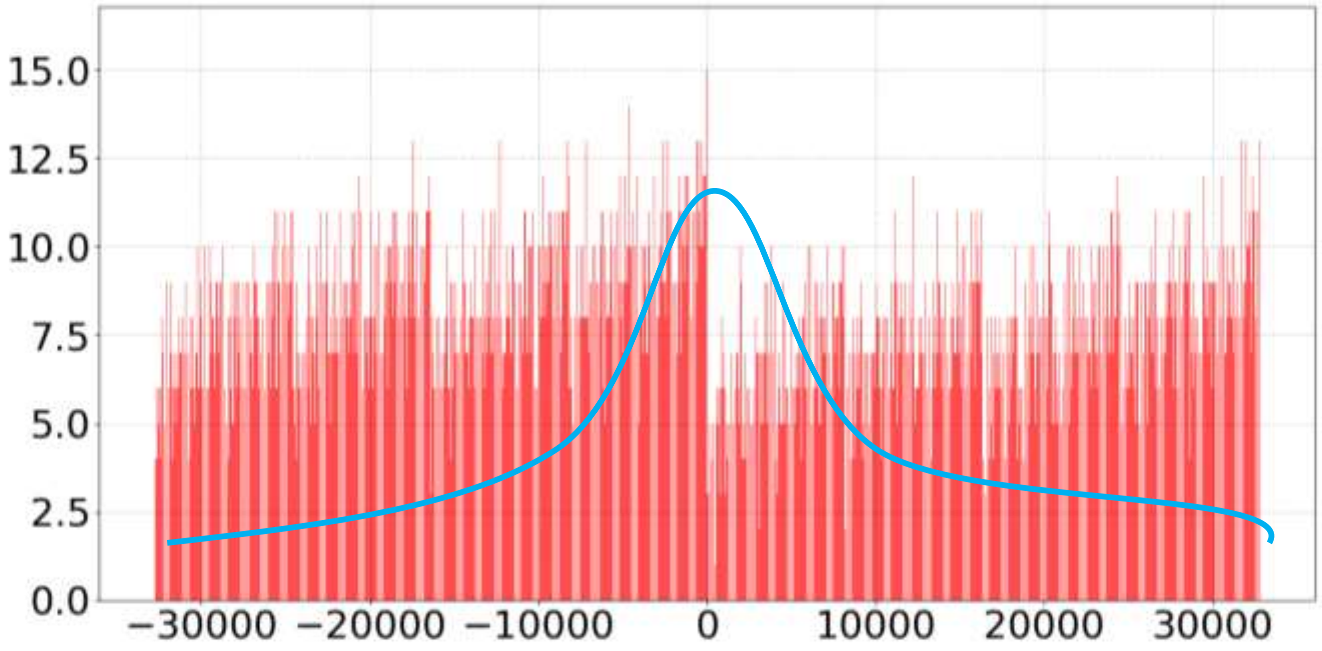


Fig. 1: Number of "1" bits vs. value for Two's Complement integers (16 bit)

(blue line indicates feature-map distribution)

Methodology: Sign Magnitude Representation

Is there a better representation?

- Needs to have a less biased distribution of bits around 0

An Integer representation with better properties for quantized feature-maps is Sign-Magnitude (**SM**) representation.

Sign Bit	Unsigned Integer

(The full range of the TC Integer can be preserved in SM)

Methodology: Sign Magnitude Representation

Sign-Magnitude “1” bit distribution

- More symmetric bit distribution
- The smaller the magnitude of the value, the lower the number of bits

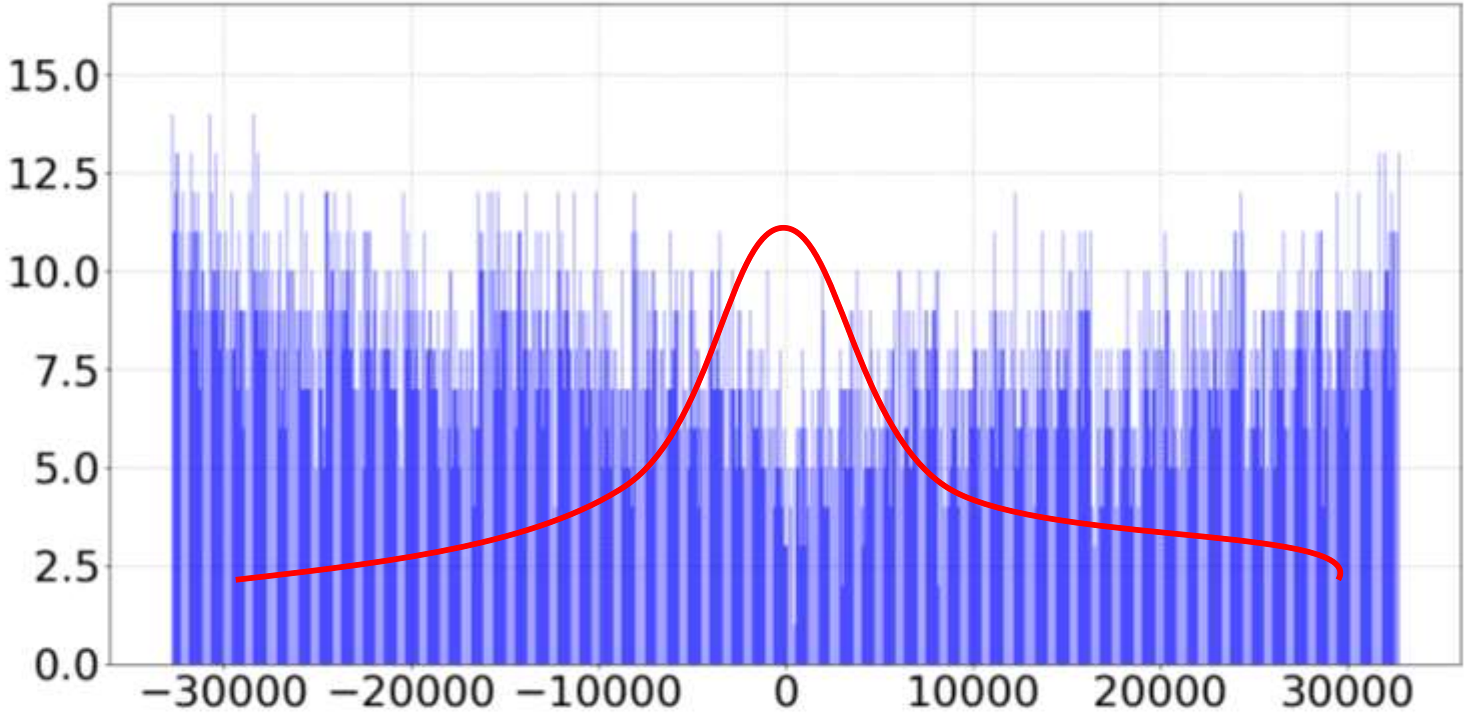


Fig. 2: Number of “1” bits vs. value for Sign-Magnitude integers (16 bit)

Methodology: Difference Encoding

How can we exploit statistical knowledge of feature-maps?

- Feature-maps exhibit strong similarity with neighboring “pixels”
- This translates to low absolute difference between pixels

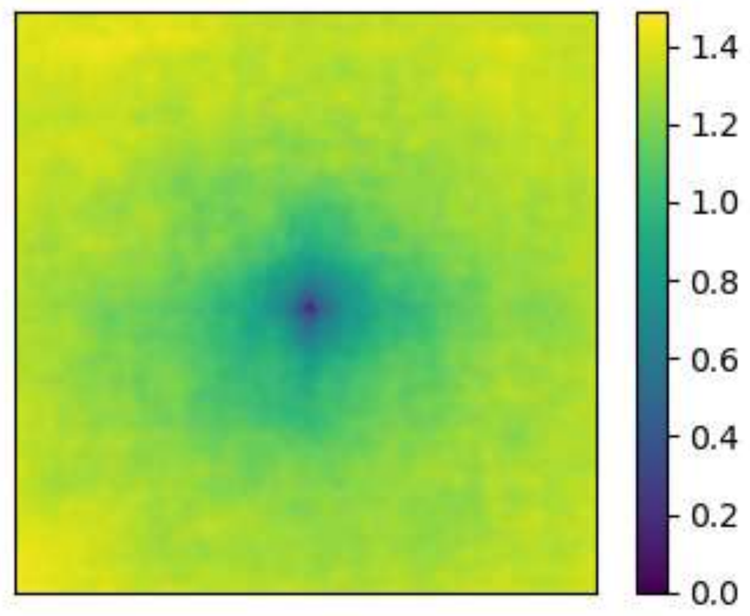


Fig. 3: Heatmap of difference between the central pixel and all other pixels, averaged across ImageNet dataset

Idea: Send difference in pixels rather than pixels themselves

Methodology: Difference Encoding

How are feature-maps transferred between the accelerator and DRAM?

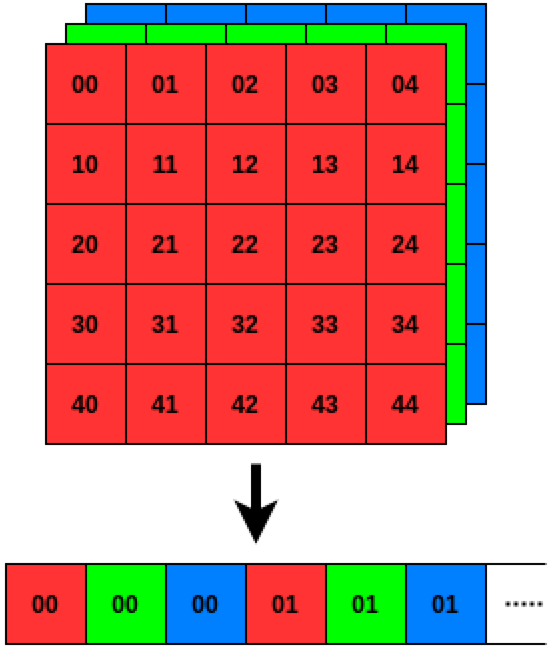


Fig. 4: Channel-first streaming

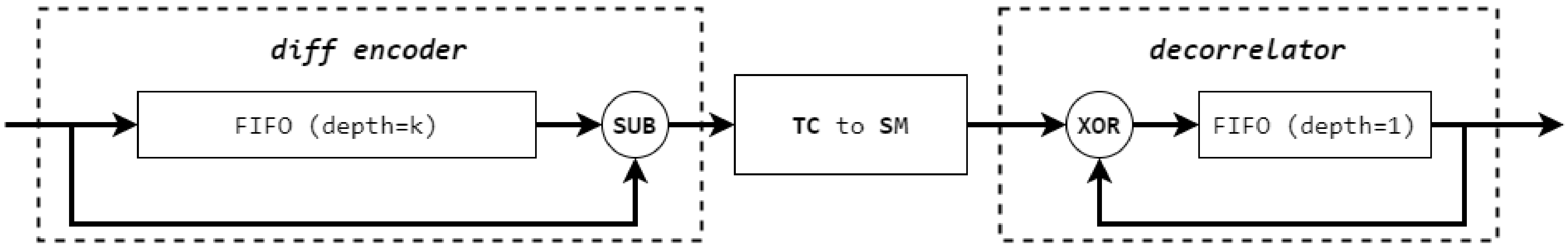
For k channels in the feature-map, the difference encoder and decoder are as such:

$$(\text{diff encoder})\ \hat{x}_i = x_i - x_{i-k}$$

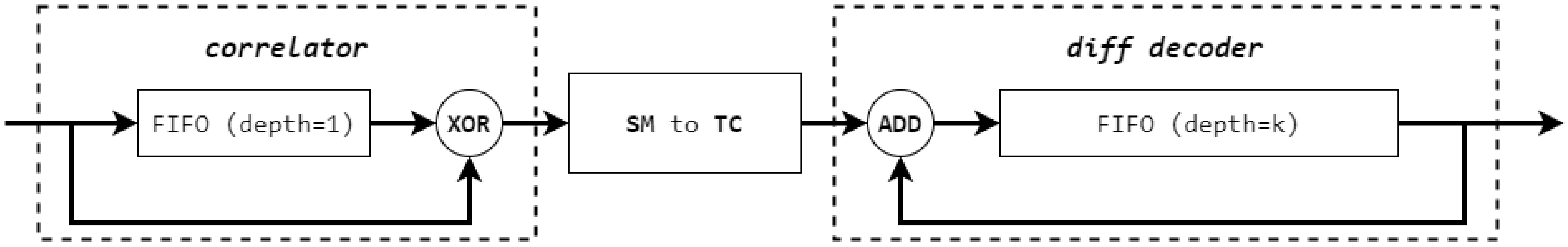
$$(\text{diff decoder})\ x_i = \hat{x}_i + x_{i-k}$$

Methodology: DEF Coding Scheme

Def Encoder:



Def Decoder:



<u>Original stream is fully recoverable</u>

Evaluation

- Metrics
- Activity Evaluation
- Power Evaluation

Evaluation: Metrics

Metrics of Interest

Transition Ratio: $$t_{ratio} = \frac{Total\ Transitions_{encoded}}{Total\ Transitions_{unencoded}}$$

(relates to the ratio of energy saved)

Average Activity: $$a_{avg} = \frac{Total\ Transitions}{Bus\ Width\ \times Total\ Words}$$

(relates to the dynamic power consumption)

Evaluation: Activity Encoding Schemes

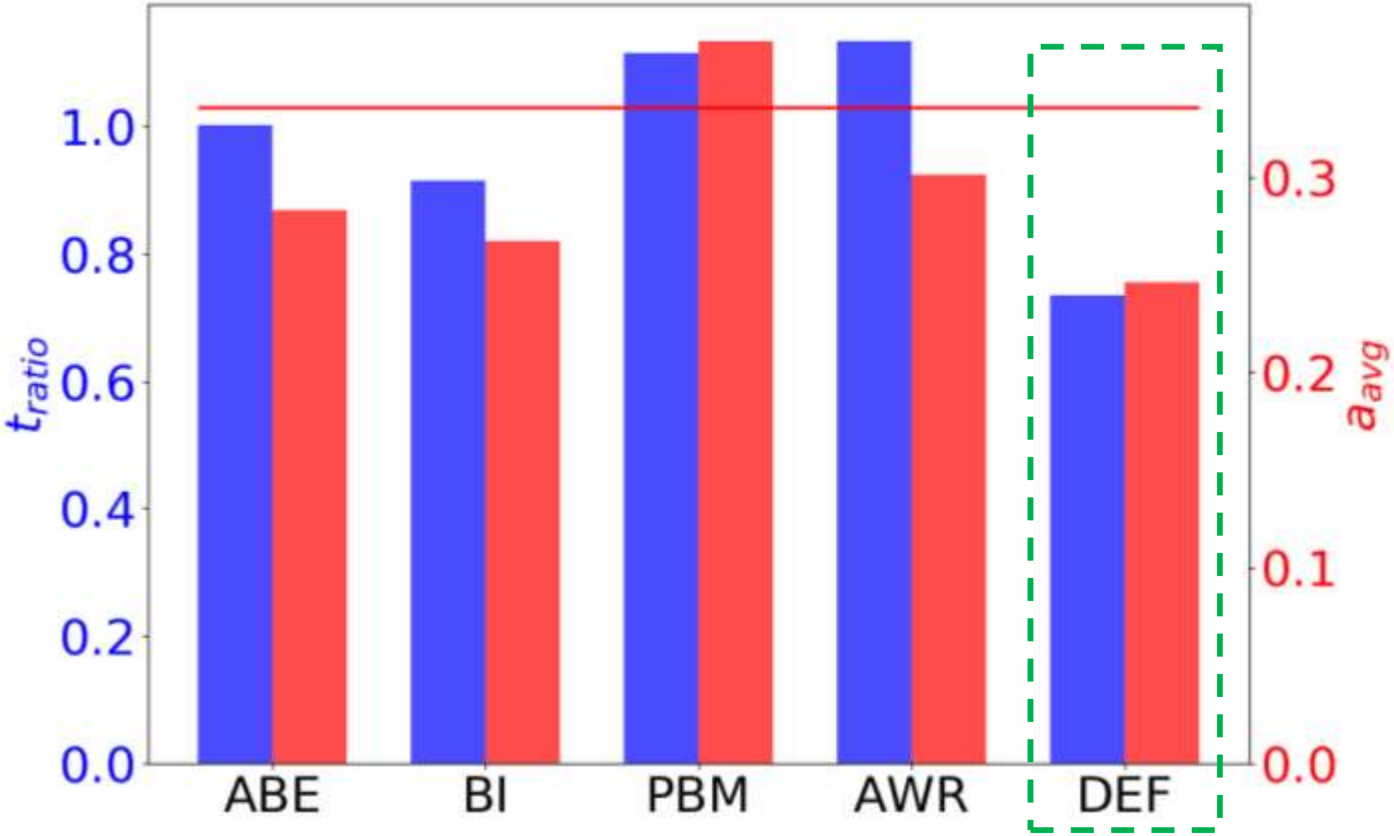


Fig. 5: Comparison for feature-maps of MobileNetv2 for a datawidth of 8

- Only encoding schemes with no spatial reduction see similar reduction in both transitions and activity
- **DEF** has both the greatest reduction in transitions as well as activity
- Some coding schemes (**PBM** and **AWR**) show increases in transitions, despite the objective of lowering activity

Evaluation: Compression Schemes

How does DEF compare to compression schemes?

Compression schemes reduce the number of off-chip memory accesses, potentially reducing energy consumption

Encoding Scheme	t_{ratio}	a_{avg}	Compression Ratio
(unencoded)	-	0.2470	-
DEF	**0.6162**	**0.1564**	1.00
RLE	1.4493	0.3839	1.17
DEF+RLE	0.7927	0.2170	1.14
Huffman	1.1605	0.4876	**1.97**

Table 1: Comparison for feature-maps of GoogleNet for a datawidth of 8

- Compression schemes have higher activity, and in some cases more transitions
- Combining a compression and activity encoding schemes has best of both worlds

Evaluation: Power Consumption

How do activity coding schemes affect power consumption?

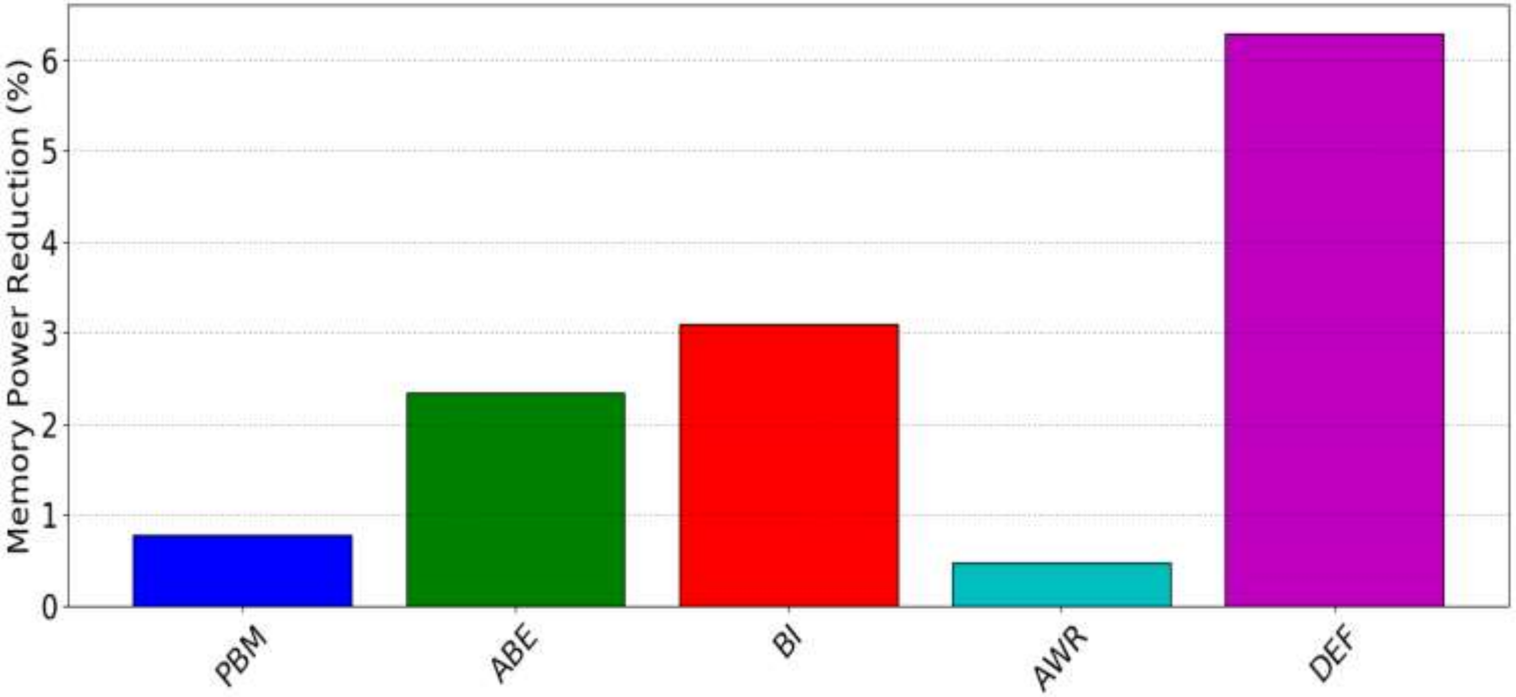


Fig. 6: Comparison of power reduction of memory subsystem for layers of MobileNetv2 on a ZC7020

- All schemes show at least some reduction in power consumption
- Power reduction is not as dramatic as activity reduction
- **DEF** outperforms other schemes by at least a factor of 2

Evaluation: Power Consumption

How do compression schemes affect power consumption?

Encoding Scheme	Power (mW)	Time (ms)	Energy (μJ)
(unencoded)	1534.2	5.519	8466.7
RLE	1526.3	3.833	5849.7
Huffman	1556.8	1.383	2152.9
DEF	1397.7	5.526	7723.8
DEF+RLE	1465.0	2.519	3690.7

Table 2: Comparison of power, time and energy for the memory subsystem for a representative layer of AlexNet on a ZC7020

Conclusion

- A novel, domain-specific activity encoding scheme (**DEF**) tailored to CNN Accelerators was proposed and implemented

- It outperformed other activity encoding schemes by a significant margin

- Was able to reduce power consumption for CNN Accelerators

- Showed to have an impact on energy as well when combined with a compression scheme

Future Work:

- Need to identify the trade-off between dynamic bus power and the static DRAM power

Please visit https://github.com/AlexMontgomerie/def to see the source code

References

1. Y. Chen, J. Emer and V. Sze, "Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks," *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*, Seoul, 2016
2. A. Montgomerie-Corcoran, S. I. Venieris and C. Bouganis, "Power-Aware FPGA Mapping of Convolutional Neural Networks," *2019 International Conference on Field-Programmable Technology (ICFPT)*, Tianjin, China, 2019
3. M. R. Stan and W. P. Burleson, "Bus-invert coding for low-power I/O," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 3, no. 1,, March 1995
4. S. Sarkar, A. Biswas, A. S. Dhar and R. M. Rao, "Adaptive Bus Encoding for Transition Reduction on Off-Chip Buses With Dynamically Varying Switching Characteristics," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 11, Nov. 2017
5. S. Ramprasad, N. R. Shanbhag and I. N. Hajj, "A coding framework for low-power address and data busses," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 7, no. 2, June 1999
6. E. Maragkoudaki, P. Mroszczyk and V. F. Pavlidis, "Adaptive Word Reordering for Low-Power Inter-Chip Communication," *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Florence, Italy, 2019
7. Chao-Tsung Huang, Yu-Chun Ding, Huan-Ching Wang, Chi-Wen Weng, Kai-Ping Lin, Li-Wei Wang, and Li-De Chen, *"ECNN: A Block-Based and Highly-Parallel CNN Accelerator for Edge Inference"* in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO '52), 2019*
8. S. I. Venieris and C. Bouganis, "fpgaConvNet: A Framework for Mapping Convolutional Neural Networks on FPGAs," *2016 IEEE 24th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, Washington, DC, 2016