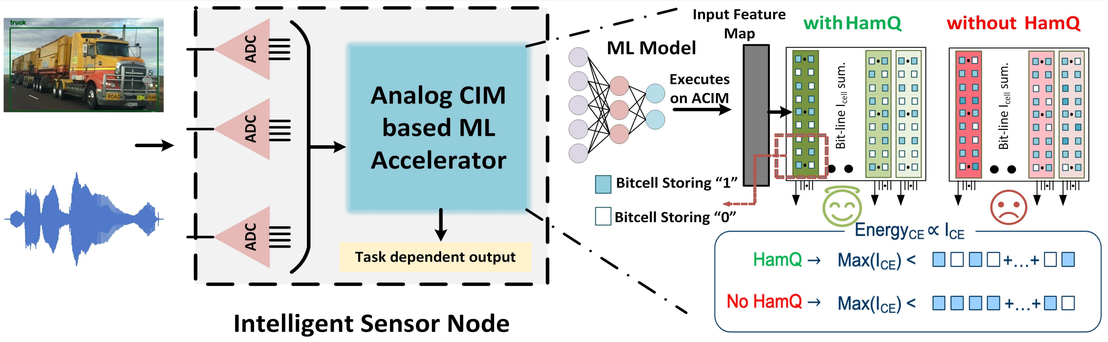
HamQ: Hamming Weight-based Energy Aware Quantization for Analog Compute-In-Memory Accelerator in Intelligent Sensors

Sudarshan Sharma*, Beomseok Kang*, Narasimha Vasishta Kidambi and Saibal Mukhopadhyay
*Equally Credited Authors
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA

CogniSense
CENTER ON COGNITIVE MULTISPECTRAL SENSORS

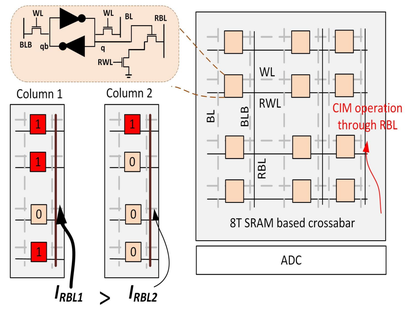
GIGASCALE RELIABLE ENERGY EFFICIENT NANOSYSTEMS LAB

Motivation



- Intelligent Sensors comprises of in-house Machine Learning (ML) accelerators to implement different task specific models on the raw-sensor data directly before sending to the host.
- Analog Compute-in-Memory (ACIM) provides an energy efficient architecture to implement these models within the memory to reduce data transfer power and increase area efficiency.
- This work enhances the energy efficiency of ACIM implementing ML models using HamQ a Hamming weight-based quantization framework based on a novel regularizer

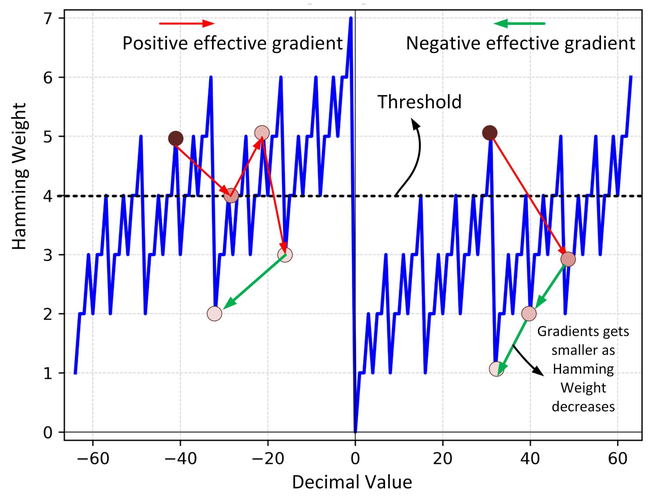
Problem



- $I_{RBL} \propto$ number of "1" stored in the column
- $I_{RBL} \propto P_{ADC}$
  - Characteristics of Current Domain ADC

Can we do HW/SW Co-design to reduce the ACIM power?

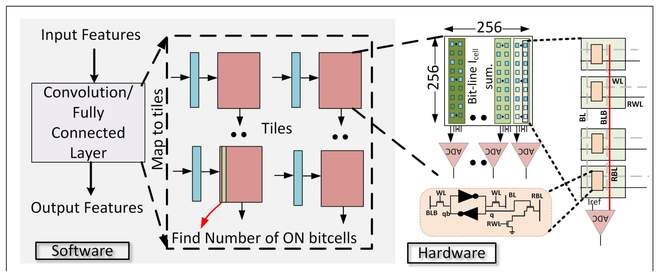
Approach

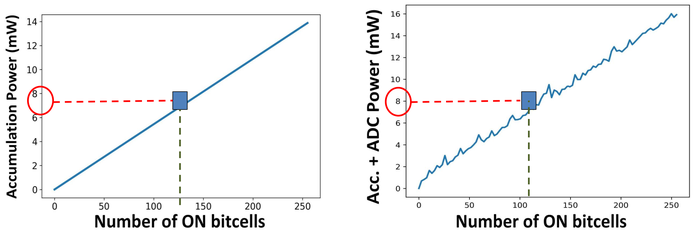


- Based on a threshold decide the direction of gradient.
- The magnitude of the gradient is the HM of the weight itself.

$$\frac{\partial L_{\text{HMW}}}{\partial w} = \text{sign}(k - h(w))h(w) \qquad \Delta w_{t-1} = -\lambda \text{sign}(k - h(w_{t-1}))h(w_{t-1})$$

Energy Estimation Steps





- Map the layer to tiles
- Find the number of ON bitcells in each crossbar during the layer execution
- Obtain the corresponding accumulation and ADC power
- Calculate the number of cycles required for layer execution
- Compute the Energy corresponding to the layer

ACKNOWLEDGEMENT

This work was supported in part by CogniSense, one of seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

Semiconductor Research Corporation

Quantitative Ablation Study in HamQ

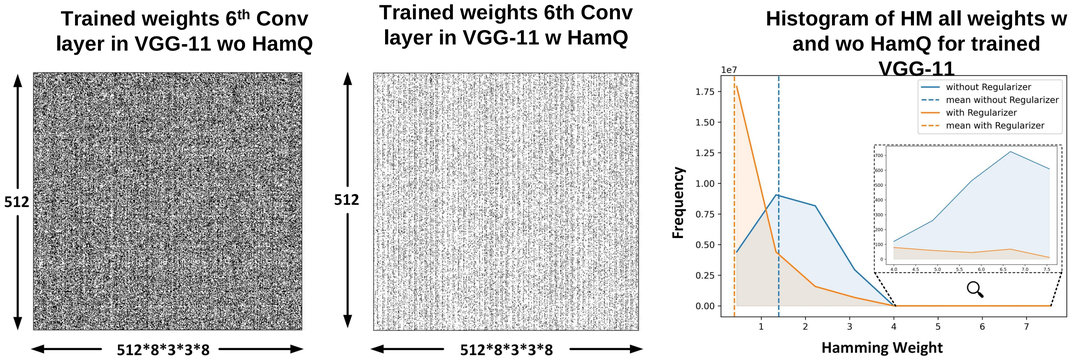
ACCURACY AND ENERGY CONSUMPTION OF RESNET-18 ON CIFAR10 CLASSIFICATION (TOP: 8-BIT AND BOTTOM: 6-BIT)

Method	Accuracy	$E_{\text{sim\_bl}}$	$E_{\text{cb}}(\mu J)$	$E_{\text{cb\&ADC}}(\mu J)$
Baseline	88.20%	1.34e+8	14.60	30.80
L_1 Reg. [34]	**89.02%**	1.20e+8	13.00	28.90
Ours (λ = 1.0e-6)	89.01%	1.15e+8	12.50	28.40
Ours (λ = 1.0e-5)	86.73%	6.17e+7	6.71	22.00
Ours (λ = 1.0e-4)	79.01%	**5.22e+7**	**6.03**	**21.30**
Baseline	88.24%	9.56e+7	7.53	17.60
L_1 Reg. [34]	**88.58%**	8.81e+7	6.84	16.80
Ours (λ = 1.0e-6)	88.26%	8.45e+7	6.59	16.50
Ours (λ = 1.0e-5)	87.29%	6.50e+7	5.01	14.80
Ours (λ = 1.0e-4)	73.23%	**4.75e+7**	**3.50**	**13.10**

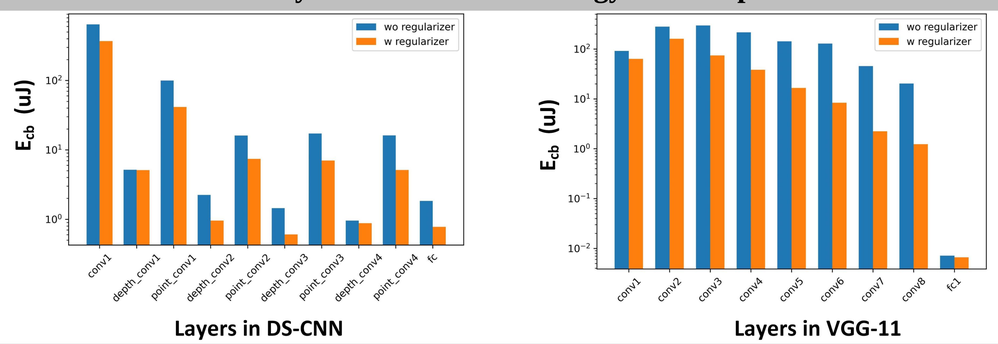
ACCURACY AND ENERGY CONSUMPTION OF DS-CNN ON KWS (TOP: 8-BIT AND BOTTOM: 6-BIT)

Method	Accuracy	$E_{\text{sim\_bl}}$	$E_{\text{cb}}(\mu J)$	$E_{\text{cb\&ADC}}(\mu J)$
Baseline	**91.82%**	2.88e+7	3.14	9.26
L_1 Reg. [34]	87.60%	2.04e+7	2.22	8.25
Ours (λ = 3.0e-5)	89.23%	1.86e+7	2.01	8.02
Ours (λ = 5.0e-5)	87.16%	1.81e+7	1.96	7.96
Ours (λ = 1.0e-4)	84.58%	**1.57e+7**	**1.71**	**7.69**
Baseline	91.09%	1.42e+7	0.95	5.42
L_1 Reg. [34]	88.05%	1.16e+7	0.80	5.23
Ours (λ = 1.0e-5)	**91.78%**	1.08e+7	0.78	5.20
Ours (λ = 3.0e-5)	87.60%	8.14e+6	0.59	5.01
Ours (λ = 5.0e-5)	86.52%	**7.35e+6**	**0.54**	**4.95**

Qualitative Ablation Study in HamQ



Layer wise Crossbar Energy Consumption



Conclusion

- HamQ reduces per-inference energy consumption by **54.0%** with a marginal accuracy degradation of **1.5%** for the 8-bit ResNet-18 model in CIFAR-10 image classification and by **42.7%** with a **3.5%** degradation for the 6-bit DS-CNN model in keyword spotting task.