Improving Energy Efficiency of Neural Networks

Seongsik Park, Seoul National University, Advisor: Sungroh Yoon

seong.sik.park@gmail.com

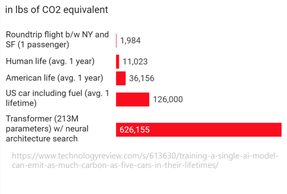
DSAIL Data Science & AI Laboratory

DATE 2021 Ph.D. Forum

Introduction

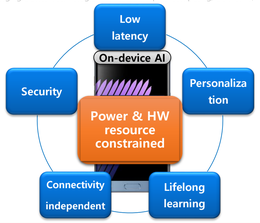
Training a single AI model can emit as much carbon as five cars in their lifetimes

Deep learning has a terrible carbon footprint.



https://www.technologyreview.com/s/613630/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/

Emerging Device Technologies for Neuromorphic Computing, Damien Q.



It is important to improve energy efficiency of neural networks!

Publications

[J3] Seijoon Kim, **Seongsik Park**, Byunggook Na, Jongwan Kim, Sungroh Yoon, "Towards fast and accurate object detection in bio-inspired spiking neural networks through Bayesian optimization," *IEEE Access*, vol. 9, pp. 2633-2643, Jan 2021.

[J2] Seongsik Park, Jaehee Jang, Seijoon Kim, Byunggook Na, Sungroh Yoon, "Memory-Augmented Neural Networks on FPGA for Real-Time and Energy-Efficient Question Answering," *IEEE TVLSI*, Vol. 20, no. 1, pp. 162-175, Jan 2021.

[C8] Seongsik Park, Seijoon Kim, Byunggook Na, Sungroh Yoon, "T2FSNN: Deep Spiking Neural Networks with Time-to-first-spike Coding," in *Proceedings of Design Automation Conference* (***DAC***), 2020.

[C7] Seongsik Park, Jongwan Kim, Sungroh Yoon, "Energy-aware Placement for SRAM-NVM Hybrid FPGAs," in *Proceedings of Design, Automation and Test in Europe* (***DATE***), 2020.

[C6] Seijoon Kim, **Seongsik Park**, Byunggook Na, Sungroh Yoon, "Spiking-YOLO: Spiking Neural Network for Energy-efficient Object Detection," in *Proceedings of AAAI Conference on Artificial Intelligence* (***AAAI***), 2020.

[C5] Seongsik Park, Seijoon Kim, Hyeokjun Choe, Sungroh Yoon, "Fast and Efficient Information Transmission with Burst Spikes in Deep Spiking Neural Networks," in *Proceedings of Design Automation Conference* (***DAC***), 2019.

[C4] Seongsik Park, Jahee Jang, Seijoon Kim, Sungroh Yoon, "Energy-Efficient Inference Accelerator for Memory-Augmented Neural Networks on FPGA," in *Proceedings of Design, Automation and Test in Europe* (***DATE***), 2019.

[J1] Seil Lee, Hanjoo Kim, **Seongsik Park**, Seijoon Kim, Hyeokjun Choe, Sungroh Yoon, "CloudSocket: Fine-Grained Power Sensing System for Datacenters," ***IEEE Access***, vol. 6, no. 1, pp. 49601-49610, 2018.

[C3] Seongsik Park, Seijoon Kim, Seil Lee, Ho Bae, Sungroh Yoon, "Quantized Memory-Augmented Neural Networks," in *Proceedings of AAAI Conference on Artificial Intelligence* (***AAAI***), 2018.

[C2] Hyeokjun Choe, Seil Lee, Hyunha Nam, **Seongsik Park**, Seijoon Kim, Eui-Young Chung, Sungroh Yoon, "Near-Data Processing for Differentiable Machine Learning Models," in *Proceedings of International Conference on Massive Storage Systems and Technology* (***MSST***), 2017.

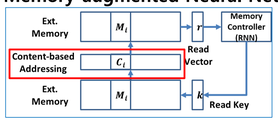
[C1] Seil Lee, Hanjoo Kim, **Seongsik Park**, Seijoon Kim, Hyeokjun Choe, Chang-Sung Jeong, Sungroh Yoon, "CloudSocket: Smart Grid Platform for Datacenters," in *Proceedings of IEEE International Conference on Computer Design* (***ICCD***), 2016.

Topic 1

Energy-efficient DNNs

Memory-augmented Neural Networks + Fixed-point Quantization (AAAI-18)

Memory-augmented Neural Networks



Content-based addressing

$$C(M,k)[i] = \frac{\exp\{S(M_i,k)\}}{\sum_{j=1}^{L}\exp\{S(M_j,k)\}}$$

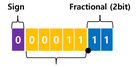
Similarity measure function

Fixed-point Quantization Error

$$|\epsilon_{\hat{u}}| < \begin{cases} 2^{-FRAC} & (\text{if } |u| < 2^{IWL}) \\ |2^{IWL} - |u|| & (\text{overflow}) \end{cases}$$

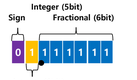
precision

Fixed point (Q5,2)
$u = 3.875$
$\hat{u} = 3.75$
$|\epsilon_{\hat{u}}| = 0.125$

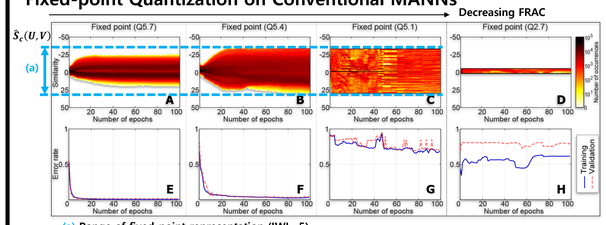


overflow

Fixed point (Q1,6)
$u = 3.875$
$\hat{u} = 1.984375$
$|\epsilon_{\hat{u}}| = 1.890625$



Fixed-point Quantization on Conventional MANNs



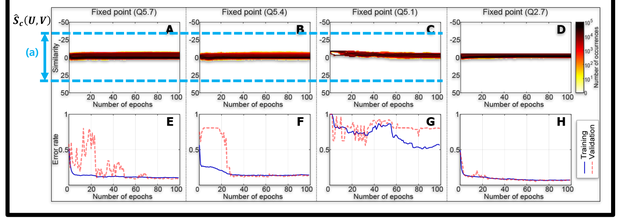
(a) Range of fixed-point representation (IWL=5)

A,E: *FRAC*↑→ $\epsilon_{\hat{v}_i}, \epsilon_{\hat{u}_i}$ ↓ → # of overflows $(\hat{S}_c)$ ↓ → $\epsilon_{\hat{S}_{c,max}}$ ⇓ → $\epsilon_{\hat{c}}$ ⇓ → *Error rate*⇓

C,G: *FRAC*↓→ $\epsilon_{\hat{v}_i}, \epsilon_{\hat{u}_i}$ ↑ → # of overflows $(\hat{S}_c)$ ↑ → $\epsilon_{\hat{S}_{c,max}}$ ⇑ → $\epsilon_{\hat{c}}$ ⇑ → *Error rate*⇑

Proposed Methods - Quantized MANN (Q-MANN)

	Similarity	Similarity measure function
Previous	Cosine similarity (dot product)	$S_C(U,V) \approx \sum_i u_i v_i$
Proposed (Q-MANN)	Hamming similarity (XNOR)	$S_H(\hat{U},\hat{V}) = \sum_i S_{\hat{u}_i} S_{\hat{v}_i} \sum_{k=0}^{n-2} W_k XNOR(\hat{u}_{ik}, \hat{v}_{ik})$
	Manhattan distance (subtraction)	$\frac{\partial S_H(\hat{U},\hat{V})}{\partial \hat{u}_i} \approx S_{\hat{u}_i} 2^{\alpha} (S_{\hat{u}_i} - S_{\hat{v}_i}) - \sum_{k=0}^{n-2} S_{\hat{v}_i} 2^{\alpha} (\hat{u}_{ik} - \hat{v}_{ik})$

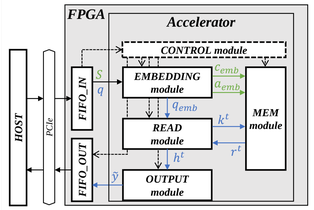


Topic 2

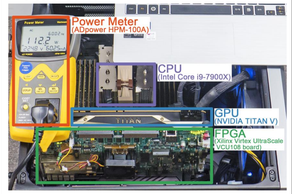
Application-specific inference accelerator

Memory-augmented Neural Networks + Question Answering Accelerator on FPGA (DATE-19, TVLSI-21)

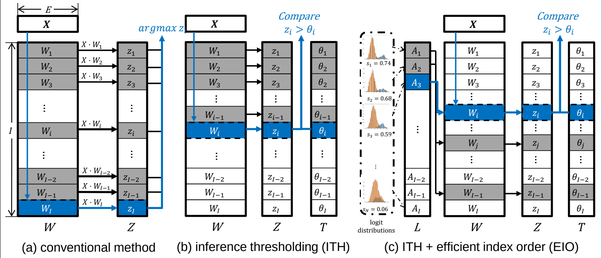
Accelerator Architecture



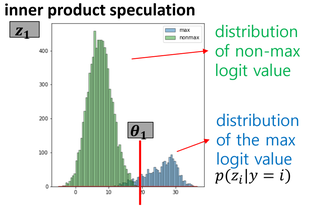
Heterogeneous Computing Platform



Fast Inference Methods on the Proposed Accelerator



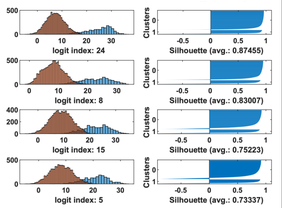
Data-based maximum inner product speculation



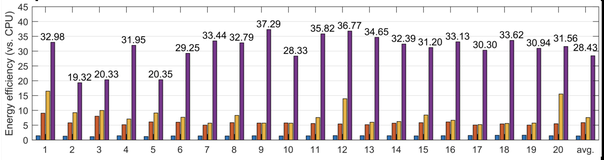
distribution of non-max logit value

distribution of the max logit value $p(z_i|y=i)$

Efficient index ordering (EIO)



Experimentall Results on bAbI dataset

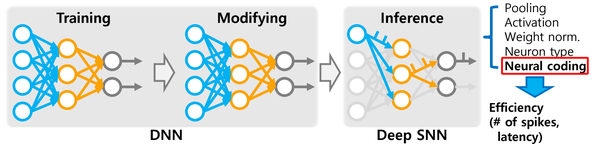


Topic 3

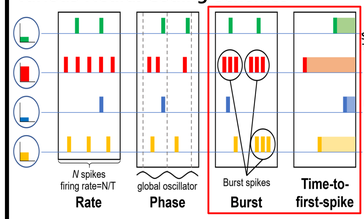
Spiking Neural Networks

Improving Efficiency of Inference on Deep SNNs - Burst coding and Time-to-first-spike coding (DAC-19, DAC-20)

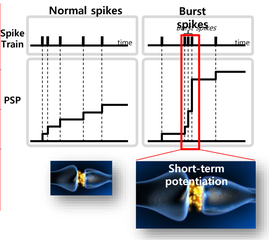
Indirect Training Method - DNN-to-SNN Conversion



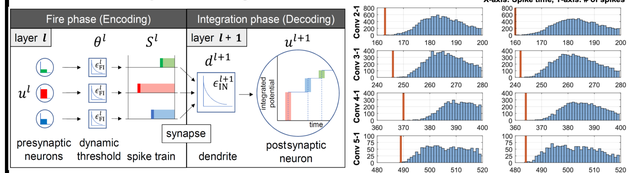
Various Neural Coding Schemes



Burst Coding

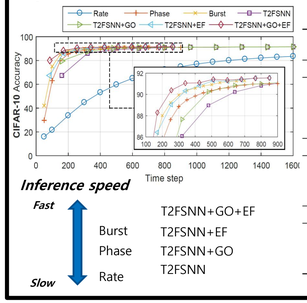


Time-to-first-spike Coding



$$z_j(t) = \sum_i w_{ij}\,\epsilon_{\text{IN}}(t - t_{\text{ref}}) + b_j \qquad \epsilon(t) = \exp(-(t-t_d)/\tau) \quad \theta(t) = \theta_0 \epsilon_{\text{FI}}(t - t_{\text{ref}})$$

Experimental Results



TN: TrueNorth SN: Spinnaker

Neural Coding	Accuracy (%)	Latency	Spikes (10^6)	Energy[a] TN	Energy[a] SN
MNIST					
Rate	99.10	200	0.100	1.000	1.000
Phase	99.20	**16**	3.000	12.048	19.228
Burst	99.25	87	0.251	1.265	1.763
Reverse	99.08	-	-	-	-
Our Method	**99.33**	40	**0.002**	**0.128**	**0.085**
CIFAR-10					
Rate	91.14	10,000	61.949	1.000	1.000
Phase	91.21	1,500	35.196	0.317	0.418
Burst	91.41	1,125	6.920	0.112	0.112
Our Method	**91.43**	**680**	**0.069**	**0.041**	**0.025**
CIFAR-100					
Rate	66.50	10,000	81.525	1.000	1.000
Phase	68.66	8,950	258.408	1.805	2.351
Burst	68.77	3,100	25.074	0.309	0.308
Our Method	**68.79**	**680**	**0.084**	**0.041**	**0.025**

Inference speed

Fast
Burst
Phase
Rate
Slow

T2FSNN+GO+EF
T2FSNN+EF
T2FSNN+GO
T2FSNN