UNO: Virtualizing and Unifying Nonlinear Operations for Emerging Neural Networks

Di Wu, Jingjie Li, Setareh Behroozi,
Younghyun Kim and Joshua San Miguel
University of Wisconsin–Madison

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Outline

- ❑ **Motivation**
- ❑ Algorithm
- ❑ Microarchitecture
- ❑ Evaluation
- ❑ Conclusion

Nonlinear operations are underrated

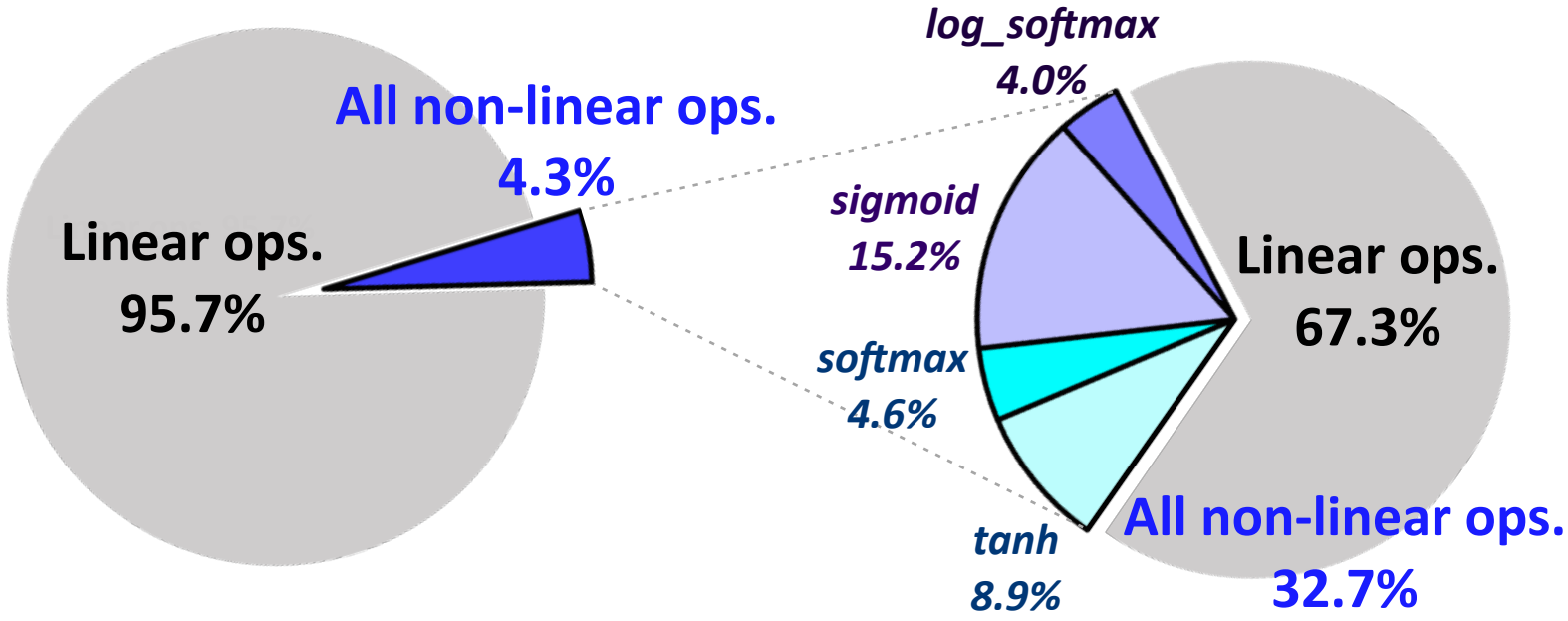
- Emerging deep neural networks have nonlinear operations with
  - High diversity: coexistence of multiple operations

Emerging NN	Nonlinear Operations
Capsule Neural Network (CapsNet)	*div, exp, log, sigmoid, softmax*
Graph Neural Network (GNN)	*div, exp, log, softmax*
Neural Machine Translation (NMT)	*log, sigmoid, softmax, tanh*

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Nonlinear operations are underrated

- Emerging deep neural networks have nonlinear operations with
  - High diversity
  - High complexity: beyond simple ReLU and max pooling



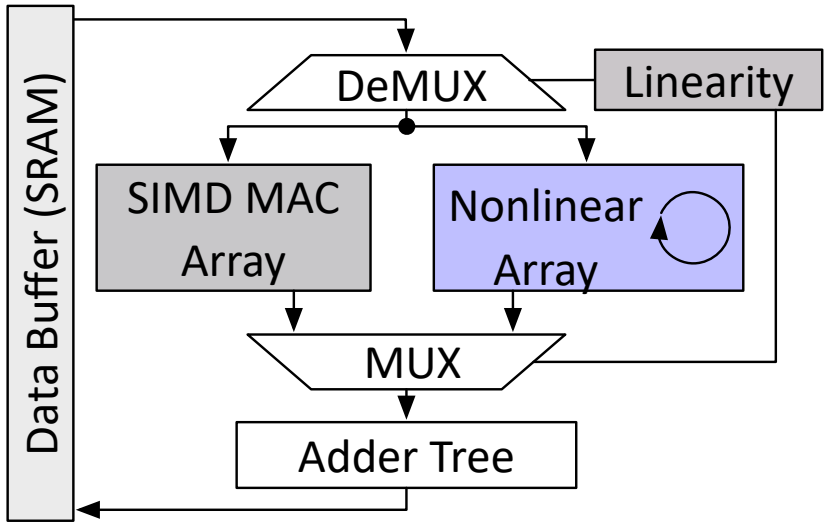
Operation count **Operation runtime**

Profiling NMT model on CPU

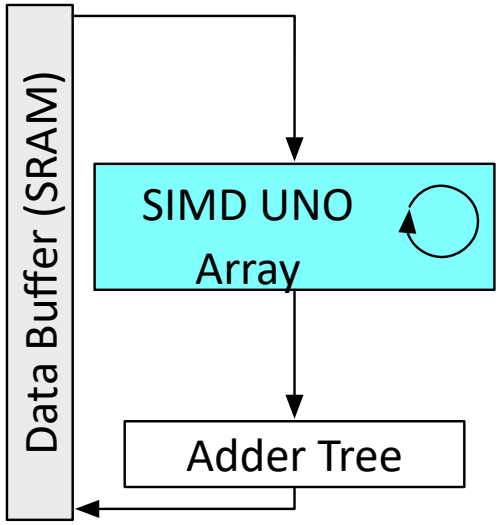
Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Nonlinear operations are underrated

- Emerging deep neural networks have nonlinear operations with
  - High diversity
  - High complexity
  - High cost: hardware overhead in accelerators



MAC array-based accelerators exhibit **low efficiency**



UNO array-based accelerators to **improve efficiency**

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Outline

- ❑ Motivation
- ❑ **Algorithm**
- ❑ Microarchitecture
- ❑ Evaluation
- ❑ Conclusion

Unify nonlinear operations (UNO)

- Leverage Horner's rule
  - Taylor approximation: unify the math expression

$$f_n(x) = \sum_{i=0}^{n} \frac{f^{(i)}(a)}{i!} \cdot (x-a)^i = \sum_{i=0}^{n} c_i \cdot (x-a)^i$$

  - MAC-compatible schedule: virtualize nonlinear operations using MAC

$$f_n(x) = offset + mac_n \cdot scale$$

$$mac_i = \begin{cases} |c_{n-i}| + mac_{i-1} \cdot var & \text{if } 1 < i \leq n, \\ |c_{n-1}| + |c_n| \cdot var & \text{if } i = 1. \end{cases}$$

  - Universal support: *div, exp, log, tanh, sigmoid, softmax, etc.*

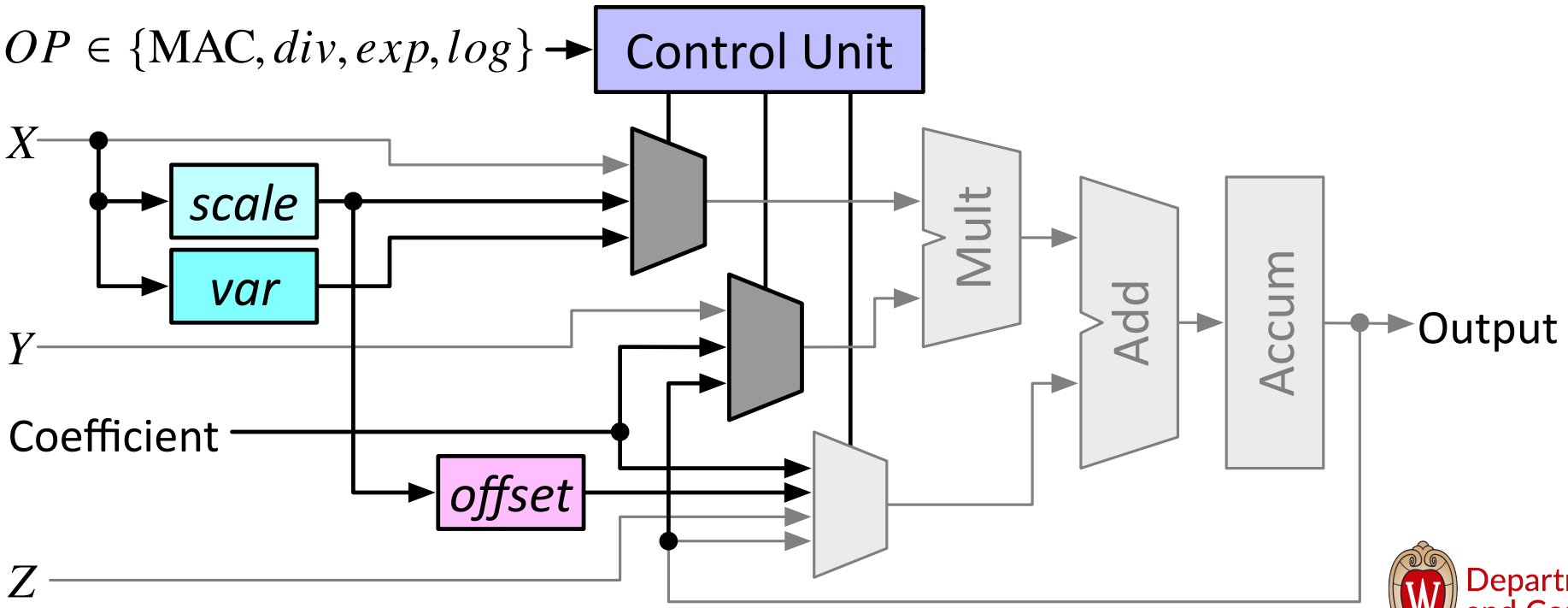
Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Outline

- ❑ Motivation
- ❑ Algorithm
- ❑ **Microarchitecture**
- ❑ Evaluation
- ❑ Conclusion

Extend a standard MAC unit

- Minimize overhead to support nonlinear operations
  - Gray blocks: original MAC unit components
  - Color blocks: extended components



Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Outline

- ❑ Motivation
- ❑ Algorithm
- ❑ Microarchitecture
- ❑ **Evaluation**
- ❑ Conclusion

Evaluate individual nonlinear operations

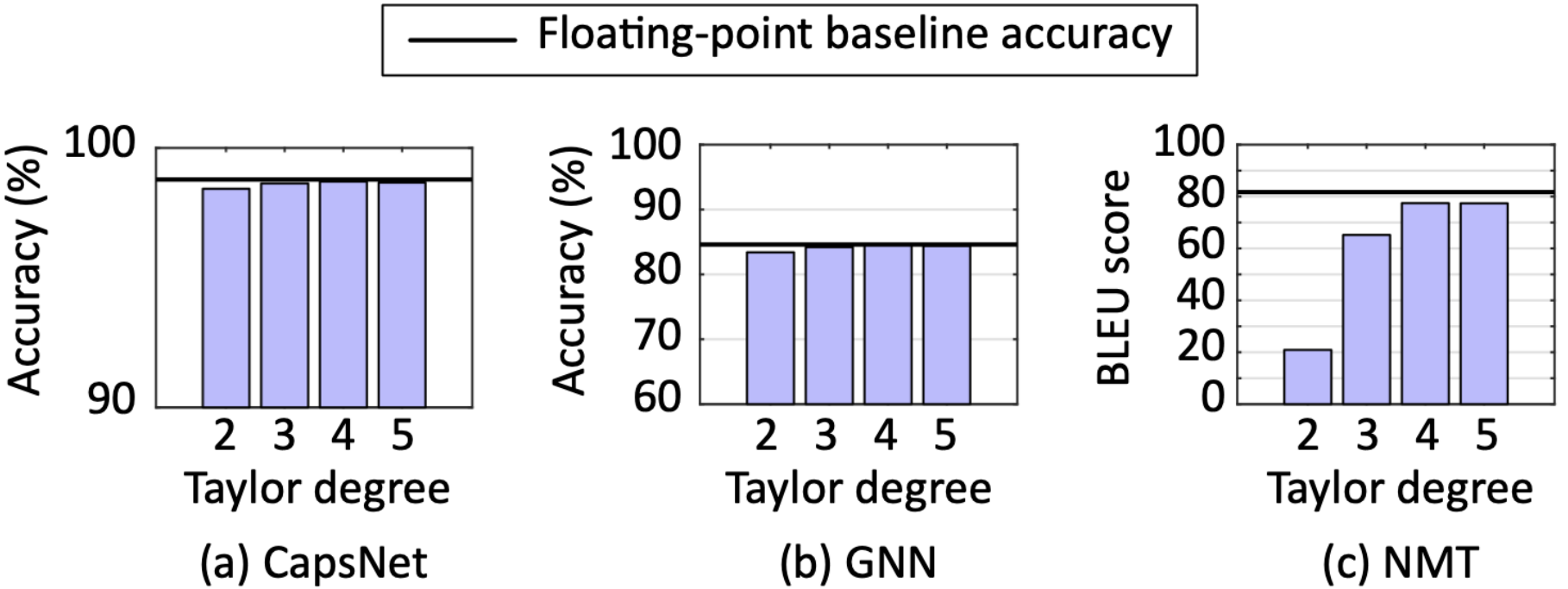
- Compare baseline and UNO processing elements
  - Similar accuracy
  - Lower area/power

div/exp/log refers to designs in [11,12,13] to compare against.

Design		Accuracy (%)	Area (μm²)	Power (mW)
Baseline (MAC+ div+ exp+ log)	div	99.93	3,068	1.09
	exp	99.70	1,118	0.22
	log	98.79	3,578	0.48
	Total	-	**9,323**	**2.35**
UNO	div	99.92	-	-
	exp	99.91	-	-
	log	99.15	-	-
	Total	-	**4,221**	**0.93**

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

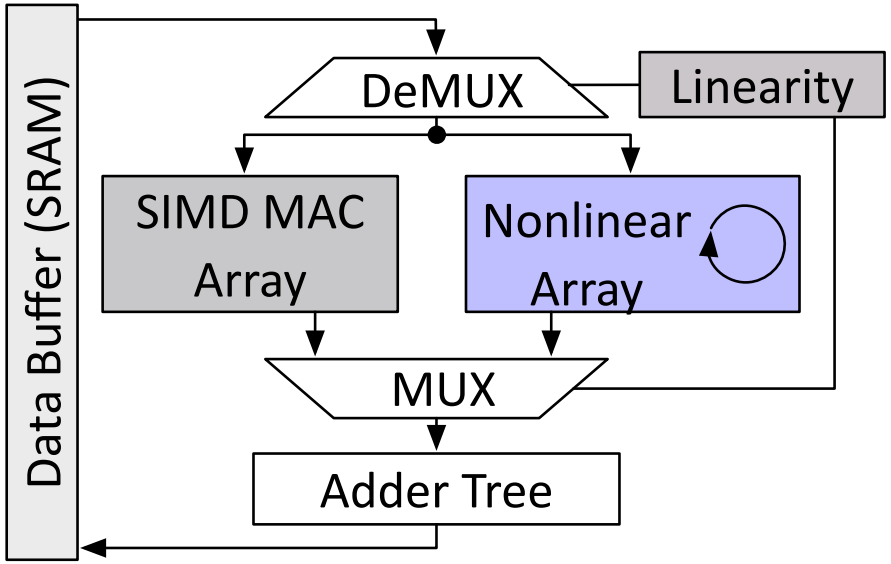
Evaluate emerging neural networks



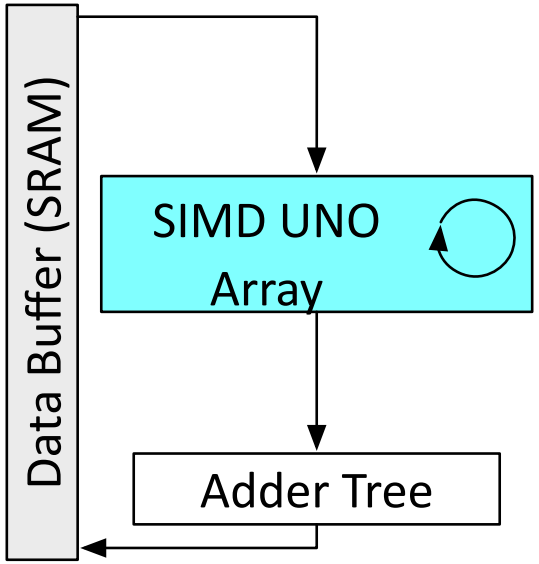
- Accuracy evaluation
  - UNO attains low accuracy loss even with heavy approximation across different neural networks

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Evaluate emerging neural networks



MAC-based SIMD array



UNO-based SIMD array

- Efficiency evaluation (SIMD 64)
  a) Baseline: MAC array based
  b) Proposed: UNO array based

*Nonlinear array in (a) refers to designs in [11,12,13].

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Evaluate emerging neural networks

	Model	Baseline	UNO	Increase (%)
Area (mm^2)	-	0.659	0.283	**-57.0**
Power (mW)	-	205.5	66.5	**-67.6**
Throughput (samples/s)	CapsNet	470.3	567.1	**+20.6**
	GNN	6.0	6.0	**+0.0**
	NMT	132.5	160.6	**+21.2**
Energy Efficiency (samples/s)	CapsNet	2288	8527	**+272.7**
	GNN	29	91	**+209.1**
	NMT	645	2415	**+274.5**

- Efficiency evaluation (SIMD 64)
  - UNO at least triples the energy efficiency for all evaluated models

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Evaluate emerging neural networks

	Model	Baseline	UNO	Increase (%)
Area (mm^2)	-	0.659	0.283	**-57.0**
Power (mW)	-	205.5	66.5	**-67.6**
Throughput (samples/s)	CapsNet GNN NMT	470.3 6.0 132.5	567.1 6.0 160.6	**+20.6** **+0.0** **+21.2**
Energy Efficiency (samples/s)	CapsNet GNN NMT	4648 68 1538	10953 116 3036	**+135.6** **+70.3** **+97.9**

- Efficiency evaluation (SIMD 64)
  - UNO still doubles the energy efficiency with power gating

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Outline

- ❑ Motivation
- ❑ Algorithm
- ❑ Microarchitecture
- ❑ Evaluation
- ❑ **Conclusion**

Conclude this work

- UNO addresses the inefficiency of nonlinear operations in emerging neural networks.
- UNO unifies nonlinear operations with Taylor approximation.
- UNO virtualizes Taylor approximation on the MAC unit to minimize hardware overhead and boost efficiency.

Department of Electrical and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

Thank you!
Q & A

Di Wu, Jingjie Li, Setareh Behroozi,
Younghyun Kim and Joshua San Miguel
University of Wisconsin–Madison

Department of Electrical
and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON