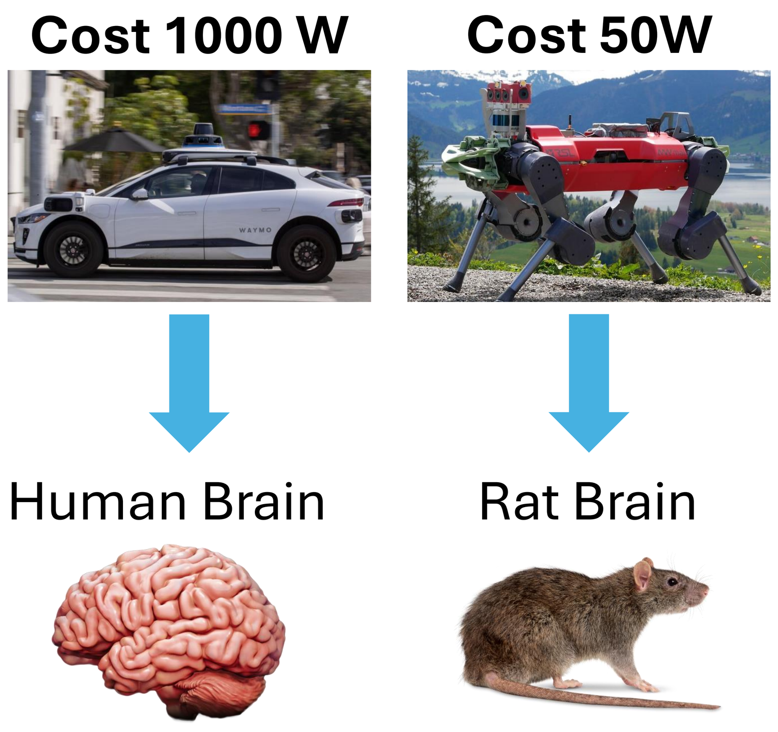
Optimizing Event-based Neural Networks on Digital Neuromorphic Architecture: A Comprehensive Design Space Exploration

Yingfu Xu[1], Kevin Shidqi[1], Gert-Jan van Schaik[1], Refik Bilgic[2], Alexandra Dobrita[1], Shenqi Wang[1], Roy Meijer[1], Prithvish Nembhani[1], Cina Arjmand[1], Pietro Martinello[1], Anteneh Gebregiorgis[3], Said Hamdioui[3], Paul Detterer[1], Stefano Traferro[1], Mario Konijnenburg[1], Kanishkan Vadivel[1], Manolis Sifalakis[1], **Guangzhi Tang[4]**, Amirreza Yousefzadeh[5]

*[1] imec the Netherlands, [2] imec, [3] Delft University of Technology, [4] **Maastricht University**, [5] University of Twente,*

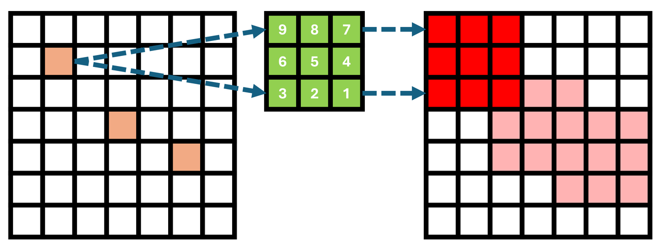
Neuromorphic Computing and Event-based Neural Networks

Self-Driving Cars	Robots
Cost 1000 W	**Cost 50W**



Human Brain	Rat Brain
10^{11} neurons	10^{8} neurons
10s of W	**10s of mW**

Neuromorphic Computing develops **Energy Efficient AI systems** inspired by the key computing paradigms of the brain

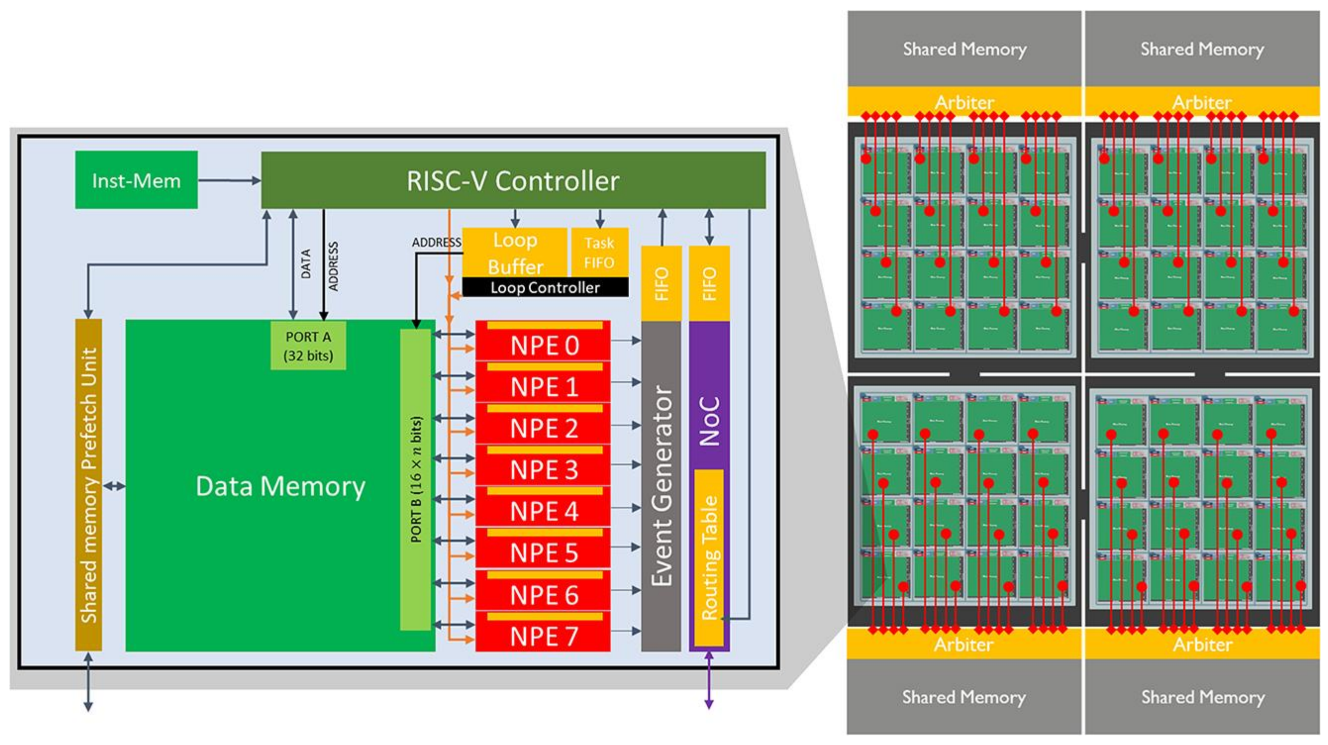


P1: Event-based Computing
- High activation sparsity
- Compute non-zero event

Core 1: Compute Data → Data → Core 2: Compute Data

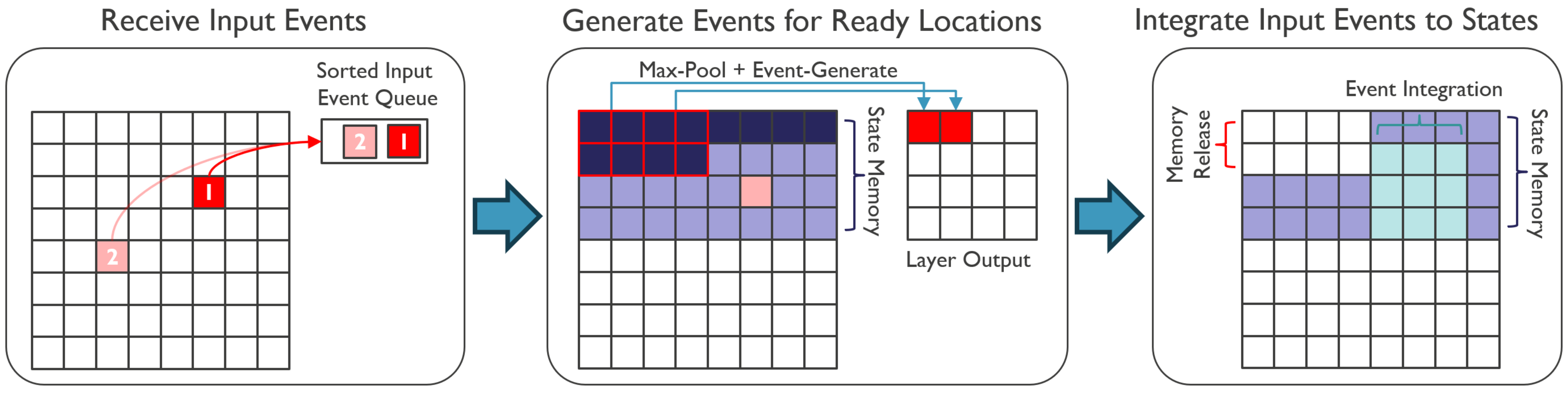
P2: Data-flow Computing
- Instant data consumption
- Near-memory computation

SENECA Neuromorphic Architecture *[1]*



- **Scalable and flexible** architecture design
- Programable **controller** and **hierarchical memory**

Event-driven Depth First Convolution *[2]*



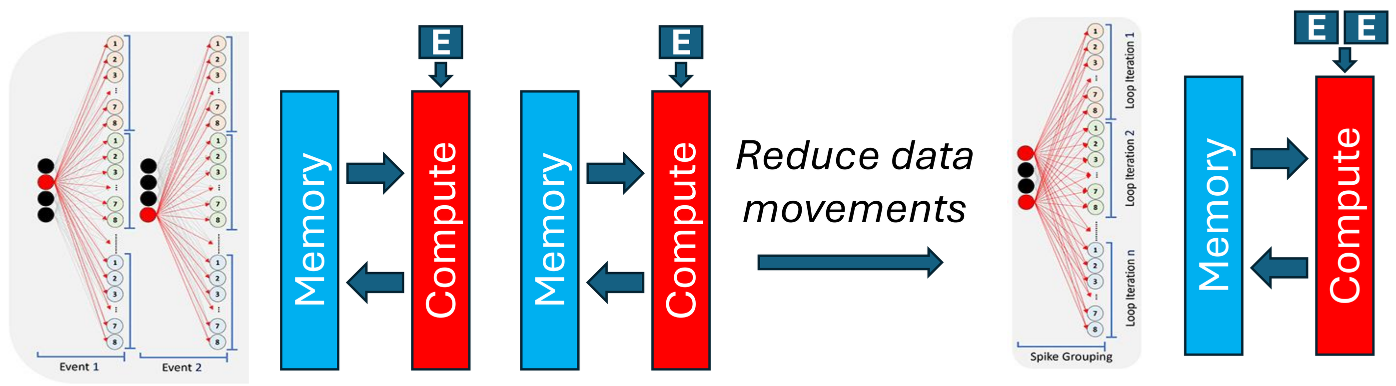
Reduce state memory cost X*Y*C to **min(X, Y)*C*4**

Reduce layer latency Full layer to **First event**

Improve vs	Energy	Latency	Area
Loihi	>50x	>3x	>3x
TrueNorth	>9x	>3x	>100x

** 5-layer CNN: (8c3k2p)-(16c3k2p)-(32c3k2p)-128-class*

Spike-Grouping for Reducing Memory Access

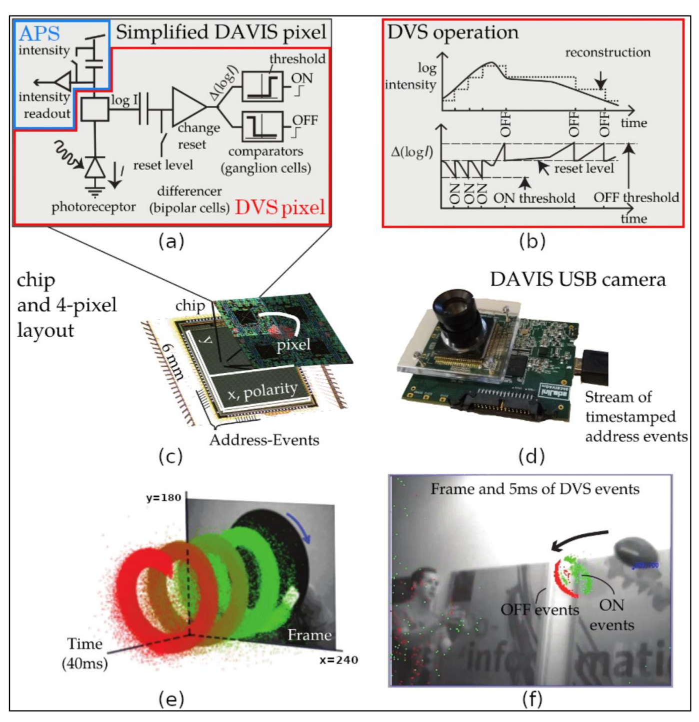


Event-by-event	Spike-grouping
2 state mem access / event	2 state mem access / group

- Reduce **neural state** memory movements
- Combine with Convolution for events in **same pixel**

Event-based Vision and Optical Flow

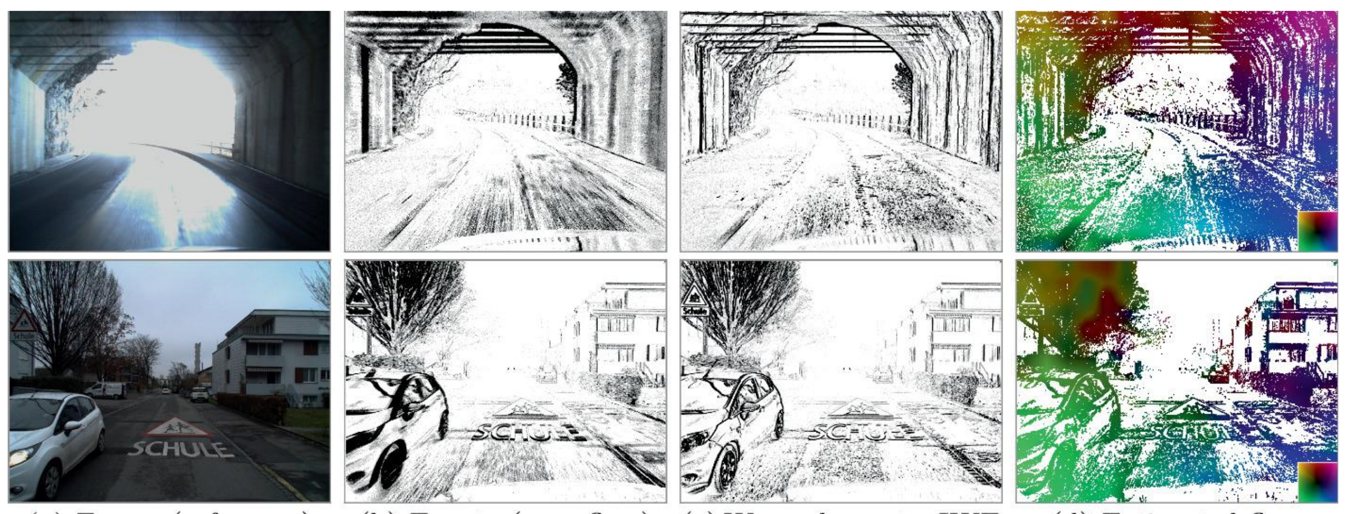
Event-based Camera



Sense Light Intensity Changes

Emulate — Optical Flow — Predict

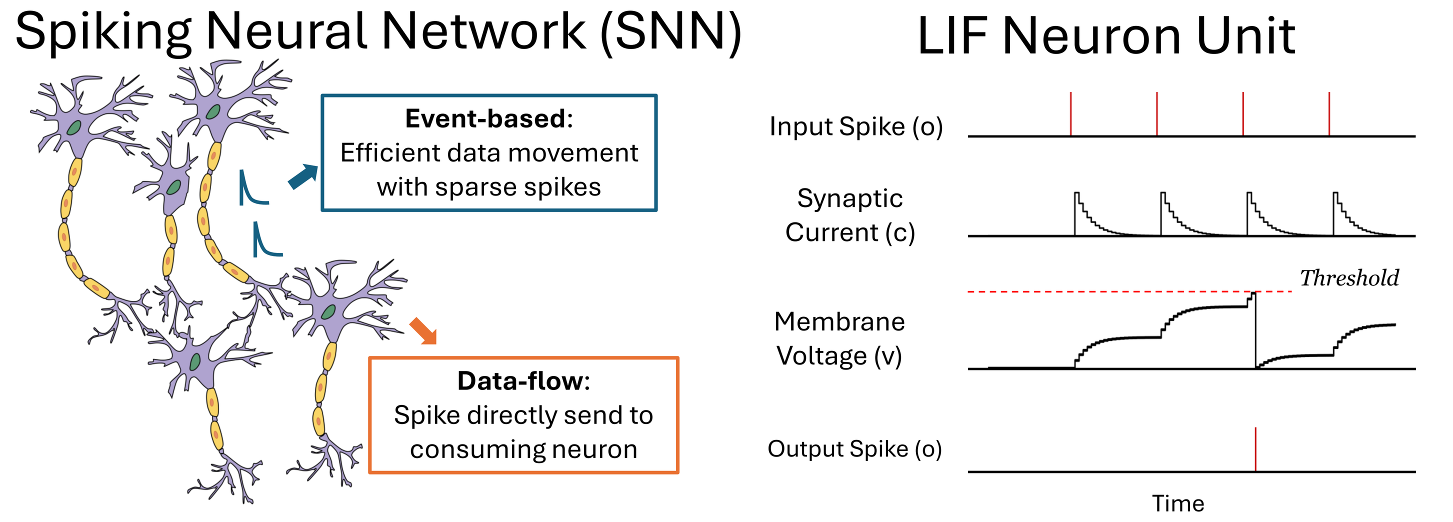
Motion → Intensity Change → Events



(a) Frame (reference) (b) Events (zero flow) (c) Warped events, IWE (d) Estimated flow

Spiking Neural Network and Event-based ANN

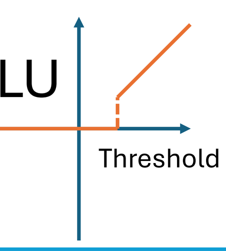
Spiking Neural Network (SNN) | LIF Neuron Unit



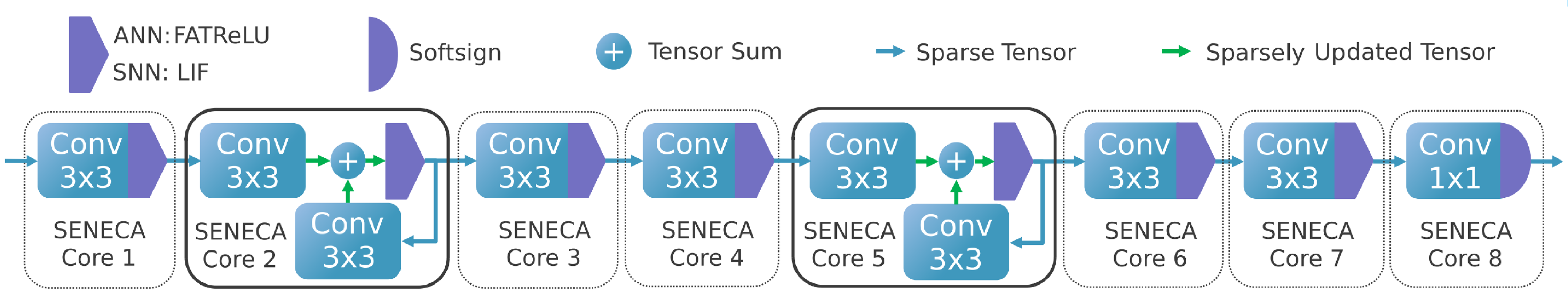
Event-based Recurrent Neural Network (ANN)

Recurrent Unit → Event Generator →

FATReLU

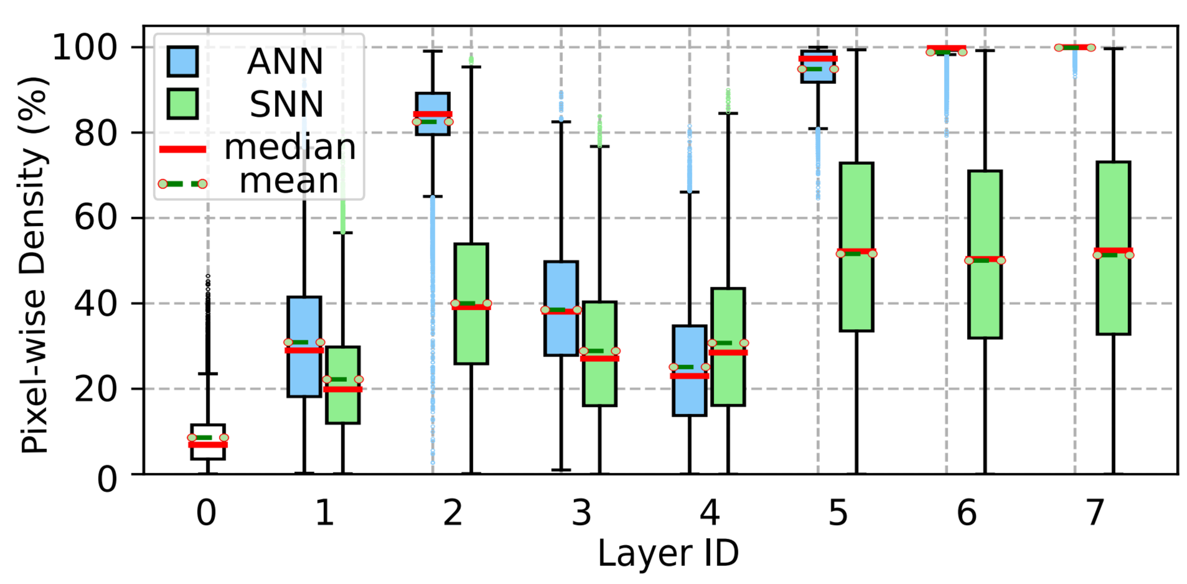


FireNet with **Sparse ANN or SNN** for Event-based Optical Flow Prediction *[3]*



- **Event-based Optical Flow Prediction**: Estimation optical flow using event camera
- **Fair Comparison of ANN and SNN**: Similar architecture, sparsity, deploy hardware
- **Hardware-aware Training**: Novel activation sparsity finetuning for ANN and SNN
- **State-of-the-art Accuracy**: Maintain low prediction error with >90% activation sparsity

Where does **SNN work better than** ANN?



- SNN having **higher Pixel-wise sparsity** than ANN
- More events in pixels increase **data reuse** chances
- Result in **lower energy and latency** on hardware

[1] Tang, Guangzhi, et al. "SENECA: building a fully digital neuromorphic processor, design trade-offs and challenges." Frontiers in Neuroscience, 2023.
[2] Xu, Yingfu, et al. "Optimizing event-based neural networks on digital neuromorphic architecture: a comprehensive design space exploration." Frontiers in Neuroscience, 2024.
[3] Xu, Yingfu, et al. "Event-based Optical Flow on Neuromorphic Processor: ANN vs. SNN Comparison based on Activation Sparsification." arXiv preprint, 2024.

Maastricht University

Contact: **Guangzhi Tang**
Assistant Professor
DACS, Maastricht University

We are actively hiring Postdoc or PhD positions!

*This work was partially funded by research and innovation projects **REBECCA** (KDT JU under grant agreement No. 101097224), **NeuroKIT2E** (KDT JU under grant agreement No. 101112268)*

Explore Activation Sparsity in Recurrent LLMs for Energy-Efficient Neuromorphic Computing

Ivan Knunyants[1,2], Maryam Tavakol[2], Manolis Sifalakis[1], Yingfu Xu[1], Amirreza Yousefzadeh[3], **Guangzhi Tang[4]**,

[1] imec the Netherlands, [2] Delft University of Technology, [3] University of Twente, [4] ***Maastricht University***

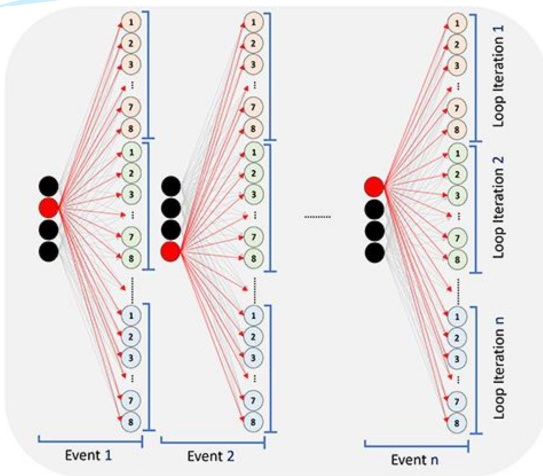
Energy Efficient Neuromorphic Computing

Large Language Models (LLMs)
Cost more than 100000 W (>500 GPUs)

Human Brain
10^{11} neurons

10s of W

Neuromorphic Computing develops **Energy Efficient AI systems** inspired by the key computing paradigms of the brain

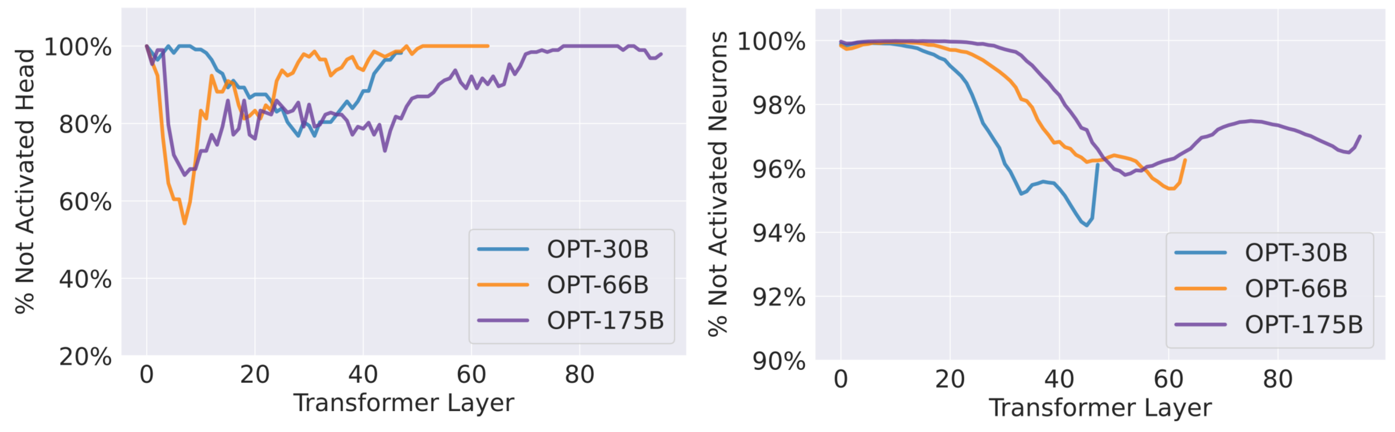


- Neuromorphic Chip mimics the brain's computing paradigm
- Process activation event-by-event to exploit sparsity

Energy efficiency linearly correlates with activation sparsity

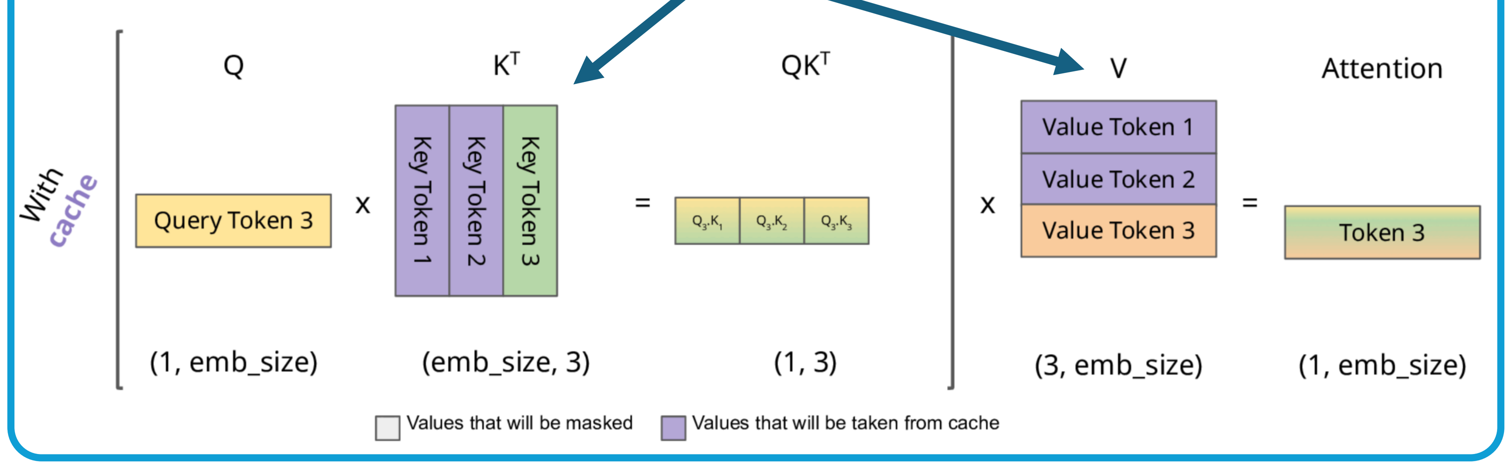
Activation Sparsity in LLMs

Theoretical token-wise activation sparsity in LLMs (ICML 2023)



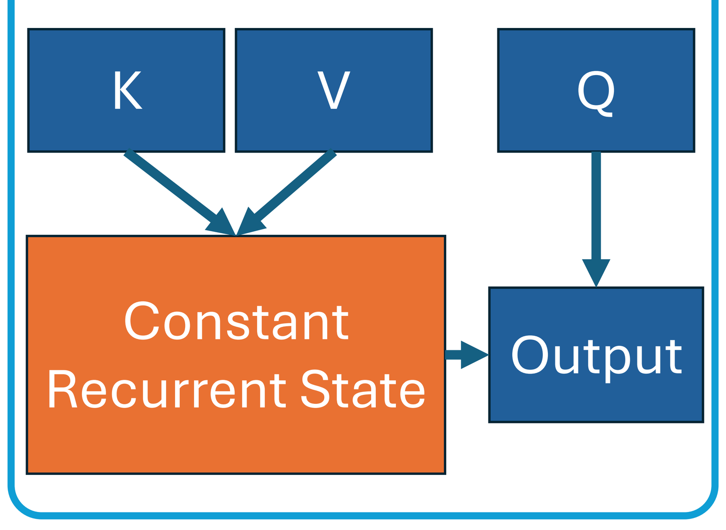
- Activation sparsity exists in LLMs and the natural sparsity ReLU activation in MLP block is >95%
- Explore sparsity in other linear projection blocks require threshold searching
- SOTA threhsold searching require **costly training**

From Costly Self-Attention LLM to **Efficient Recurrent LLM**



KV Cache for Self-Attention LLM

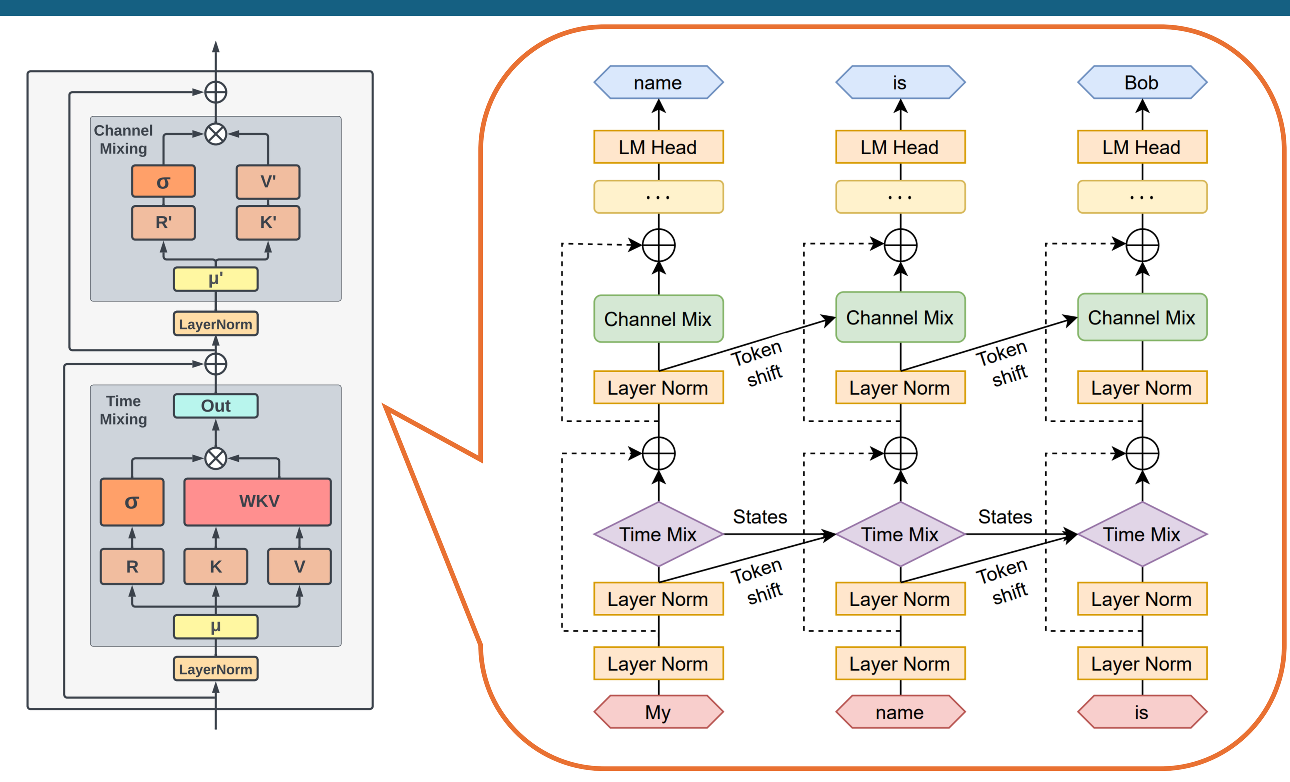
Memory of tokens stored in growing blocks



Recurrent LLM

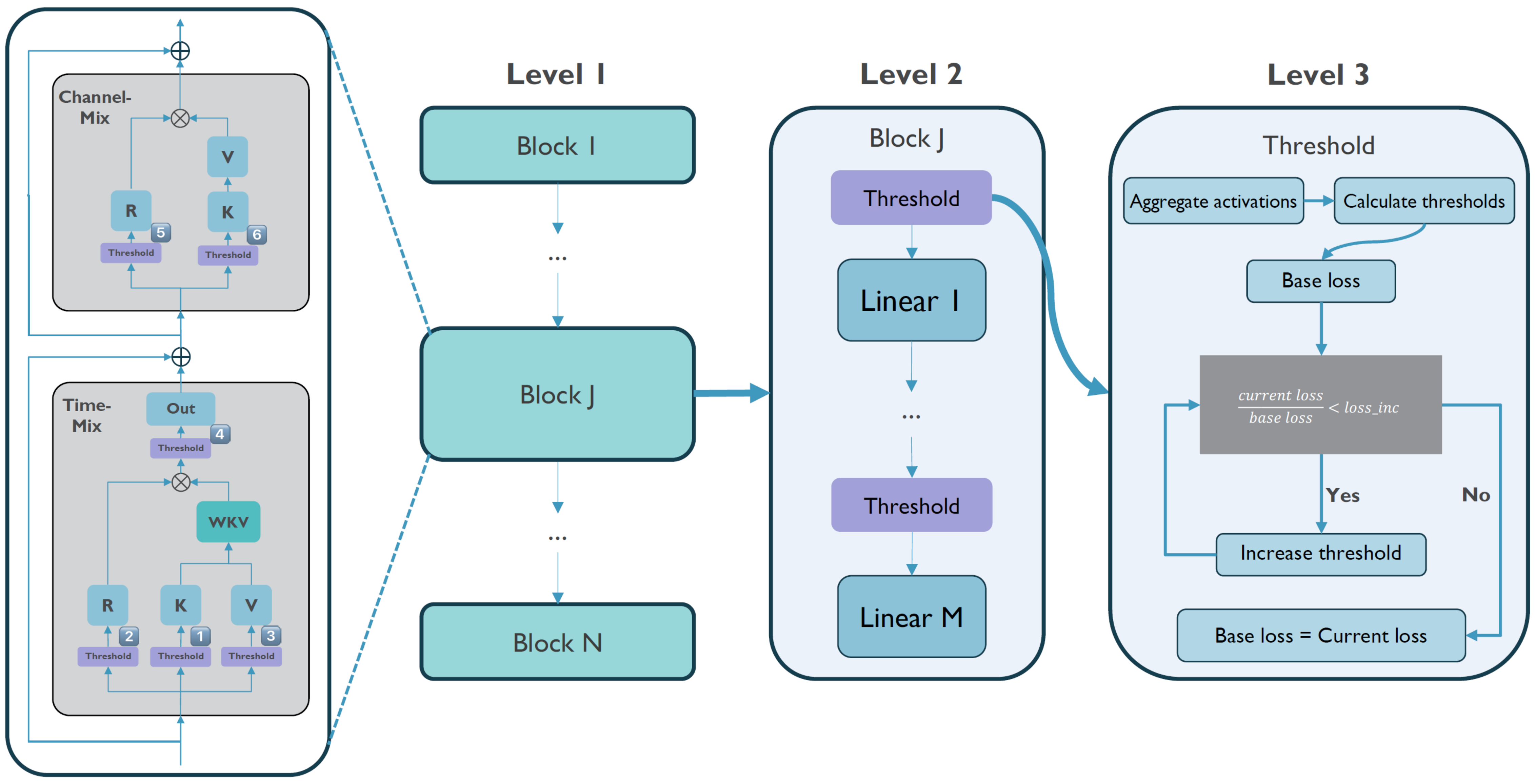
Implicit recurrent memory

RWKV Architecture (EMNLP, 2023)



Comparable performance with self-attention LLM

Training-free Threshold Initialization for Sparse Recurrent LLM



Threshold initialization algorithm. **Level 1**: iterate over LLM blocks. **Level 2**: in each block iterate over thresholding functions following the predefined order. **Level 3**: search the optimal threshold for each thresholding function by performing R-LLM inference.

Neuromorphic Hardware Simulation Study

Average Energy Cost (μJ) per token

	Time-Mix		Channel-Mix		Overall	
	Sparse	Dense	Sparse	Dense	Sparse	Dense
Computation	5.0	11.9	9.3	15.5	14.3	27.4
Memory	7.4	17.6	13.9	23.1	21.3	40.7
Total	**12.4**	29.5	**23.2**	38.6	**35.6**	68.1

Average Latency (ms) per token

	Time-Mix		Channel-Mix		Overall	
	Sparse	Dense	Sparse	Dense	Sparse	Dense
Computation	0.9	2.1	1.7	2.8	2.6	4.9
Memory	1.3	3.1	2.5	4.1	3.8	7.2
Total	**2.2**	5.2	**4.2**	6.9	**6.4**	12.1

Perform realistic neuromorphic hardware simulation study on the SENECA neuromorphic processor using real hardware measurements

You can check **Poster 12: Optimizing Event-based Neural Networks on Digital Neuromorphic Architecture** for a detailed overview on the SENECA neuromorphic processor

Benchmarking with Baseline RWKV using MiniPile Dataset

Model size	Model type	Sparsity (%)	Test loss	Loss Increase (%)
430M	Baseline [3]	28.01	2.2377	
	Our approach	57.03	2.3377	4.47
1.5B	Baseline [3]	28.38	2.0222	
	Our approach	59.99	2.1111	4.40
3B	Baseline [3]	28.65	1.9297	
	Our approach	63.16	2.0510	6.29

Double activation sparsity with minimal loss increase on RWKV LLMs

Extension to self-attention OPT on Zero-shot Benchmarks

Model	Activation sparsity (%)			Overall sparsity	AVG Benchmark Accuracy (%)
	QKV	UpProj	DownProj		
2.7B Base [17]	0	0	96	48	60.3
2.7B Training-based [14]	50	35	96	71.125	58.5
2.7B Our (loss_inc = 1.0003)	46	35	97	70.125	59.8
2.7B Our (loss_inc = 1.0004)	48	38	97	71.25	58.6
2.7B Our (loss_inc = 1.0005)	50	39	97	72.125	58.3

- Our training-free approach can also extend to self-attention LLM
- Our method achieves same performance as the training method (ICLR 2024) while 30x faster than its training on GPUs using large dataset

Maastricht University

Contact: **Guangzhi Tang**
Assistant Professor
DACS, Maastricht University

We are actively hiring Postdoc or PhD positions!

*This work was partially funded by research and innovation projects **REBECCA** (KDT JU under grant agreement No. 101097224), **NeuroKIT2E** (KDT JU under grant agreement No. 101112268)*