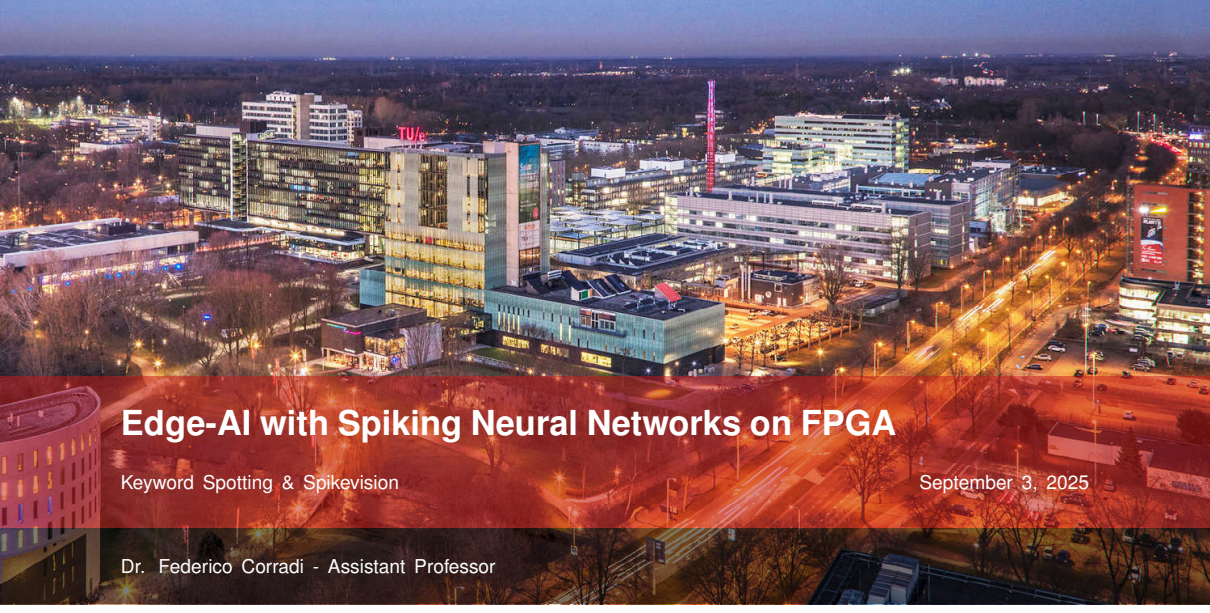


Edge-AI with Spiking Neural Networks on FPGA

Keyword Spotting & Spikevision

September 3, 2025

Dr. Federico Corradi - Assistant Professor

Electrical Engineering, Neuromorphic Edge Computing Systems Lab

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Outline

Motivation: Dynamic Sensing at the Edge is the Next Big Challenge!


Face Recognition

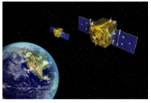

Navigation


Monitoring


Fast autonomy


Smart Home


Surveillance


Space Awarness


Mobile Devices

Edge-AI requires

- Real-time operation

Motivation: Dynamic Sensing at the Edge is the Next Big Challenge!


Face Recognition

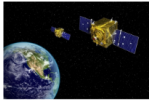

Navigation


Monitoring


Fast autonomy


Smart Home


Surveillance


Space Awarness


Mobile Devices

Edge-AI requires

- Real-time operation
- Battery operated (low-power)

Motivation: Dynamic Sensing at the Edge is the Next Big Challenge!


Face Recognition

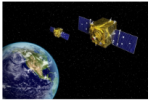

Navigation


Monitoring


Fast autonomy


Smart Home


Surveillance


Space Awarness


Mobile Devices

Edge-AI requires

- Real-time operation
- Battery operated (low-power)
- Consume data locally (privacy)

Motivation: Dynamic Sensing at the Edge is the Next Big Challenge!


Face Recognition

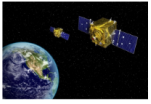

Navigation


Monitoring


Fast autonomy


Smart Home


Surveillance


Space Awarness

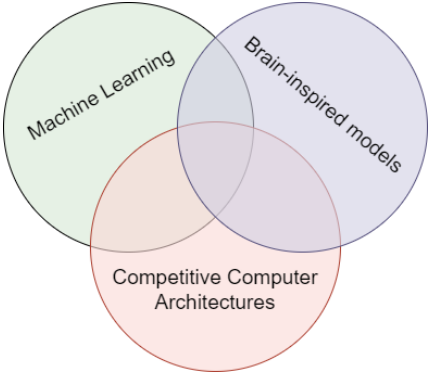

Mobile Devices

Edge-AI requires

- Real-time operation
- Battery operated (low-power)
- Consume data locally (privacy)
- Efficient execution of AI algorithms (intelligence, autonomy)

Edge Artificial Intelligence

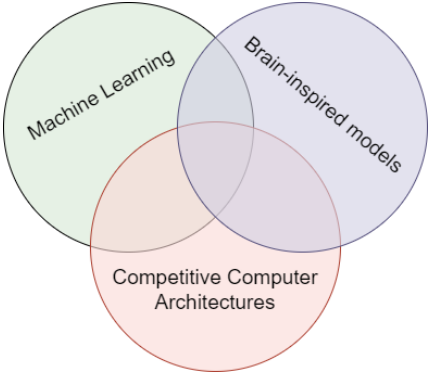
Lessening the computational burden for running AI



Edge Artificial Intelligence

Lessening the computational burden for running AI

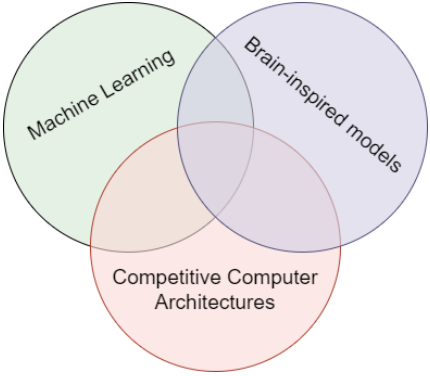
- Hardware accelerators



Edge Artificial Intelligence

Lessening the computational burden for running AI

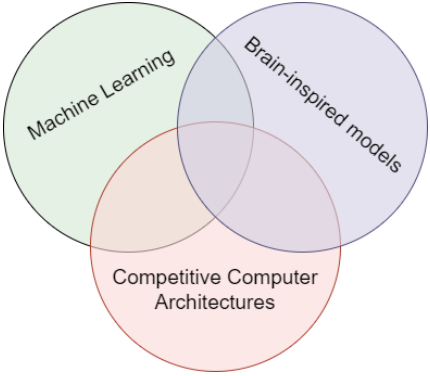
- Hardware accelerators
- Network compression



Edge Artificial Intelligence

Lessening the computational burden for running AI

- Hardware accelerators
- Network compression
- New neural network architectures

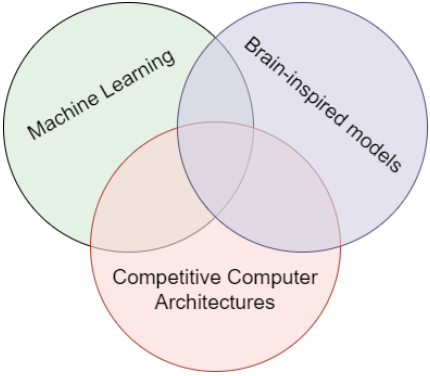


TU/e

Edge Artificial Intelligence

Lessening the computational burden for running AI

- Hardware accelerators
- Network compression
- New neural network architectures
- New devices, materials, novel design styles

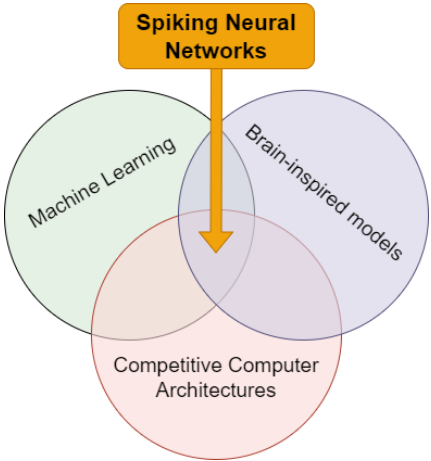


TU/e

Edge Artificial Intelligence

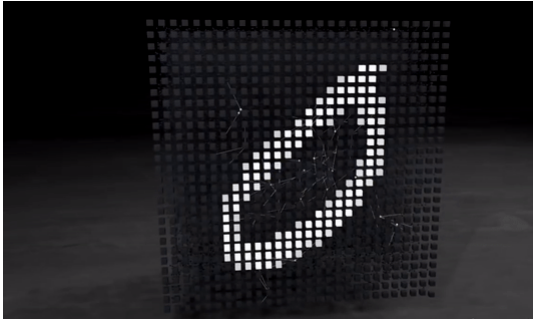
Lessening the computational burden for running AI

- Hardware accelerators
- Network compression
- New neural network architectures
- New devices, materials, novel design styles
- Spiking neural networks (i.e., brain-inspired computing)



TU/e

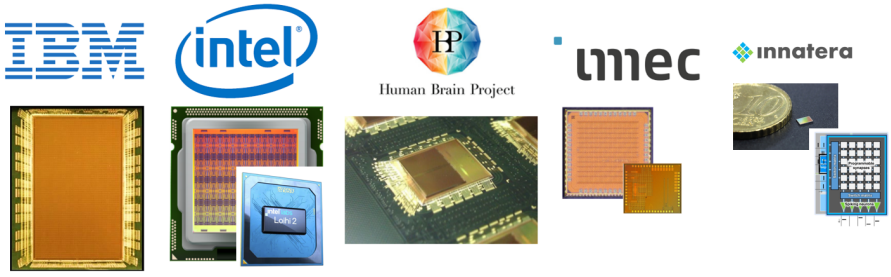
Why Edge-AI with SNNs?



SNN properties

- Fully parallel $\Rightarrow$ asynchronous
- Spatiotemporal processing $\Rightarrow$ rich dynamics
- Event-driven sparsity $\Rightarrow$ fewer MACs or no MACs at all
- Synaptic Plasticity $\Rightarrow$ Local Learning

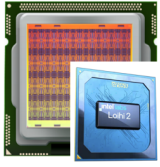
Neuromorphic Computing Hardware, many approaches!

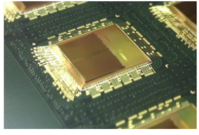


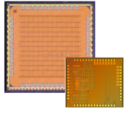
Modern Neuromorphic Computing

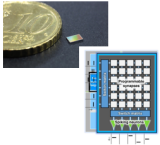
- Run spiking neural networks as *models* of computation

Neuromorphic Computing Hardware, many approaches!





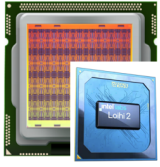


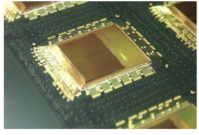


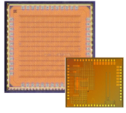
Modern Neuromorphic Computing

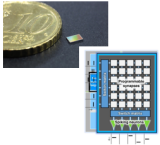
- Run spiking neural networks as *models* of computation
- Leverage advanced semiconductor node technologies

Neuromorphic Computing Hardware, many approaches!





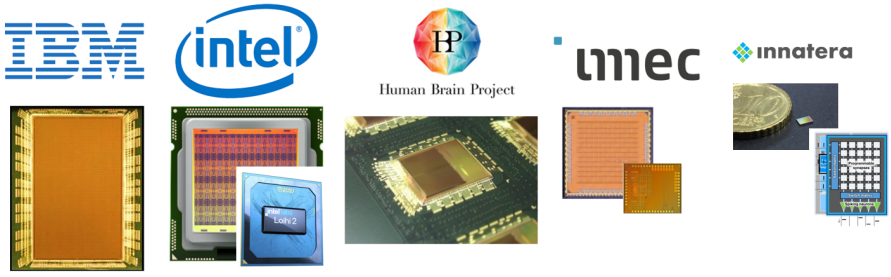




Modern Neuromorphic Computing

- Run spiking neural networks as *models* of computation
- Leverage advanced semiconductor node technologies
- Combines analog, asynchronous, and digital logic circuits

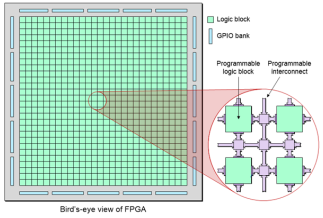
Neuromorphic Computing Hardware, many approaches!

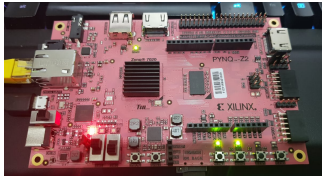


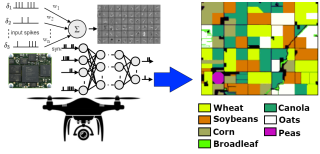
Modern Neuromorphic Computing

- Run spiking neural networks as *models* of computation
- Leverage advanced semiconductor node technologies
- Combines analog, asynchronous, and digital logic circuits
- Neuromorphic computing systems for edge-AI applications

Neuromorphic Architectures in Field Programmable Gate Arrays







FPGA

- Easy development-deployment cycle
- Exploit parallelism
- Event-driven designs
- Validate design before tape-out

- **F. Corradi**, G. Adriaans, S. Stuijk *Gyro: A Digital Spiking Neural Network Architecture for Multi-Sensory Data Analytics*, DroneSE and RAPIDO, **2021**
- A. Sankaran, P. Detterer, N. Alachiotis, **F. Corradi** *An Event-driven Recurrent Spiking Neural Network Architecture for Efficient Inference on FPGA*, ICONS, **2022**
- **F Corradi**, Z. Shen, H. Zhao, N. Alachiotis *Accelerated spiking convolutional neural networks for scalable population genomics*, HEART, *2024*
- Y. Zhang, D.M. Gomony, H. Corporral, **F. Corradi**, *A scalable hardware architecture for efficient learning of recurrent neural networks at the edge*, VLSISOC, **2024**
- Z Shen, and **F Corradi**, *SpikeVision: A Fully Spiking Neural Network Transformer-Inspired Model for Dynamic Vision Sensors, IEEE Asilomar, **2024***

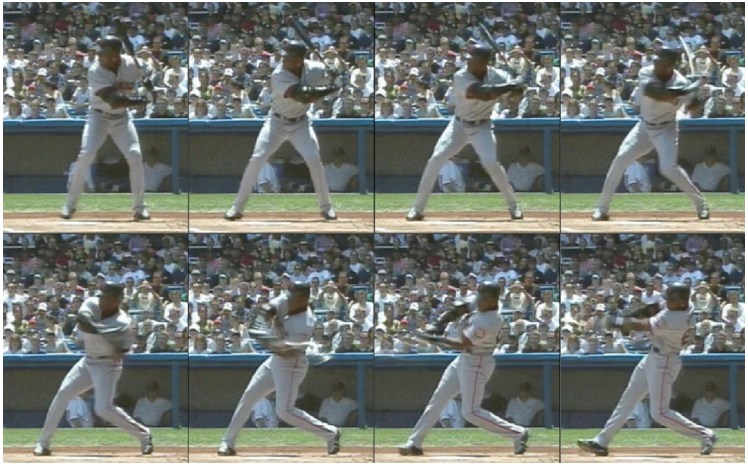
Outline

Edge Artificial Intelligence

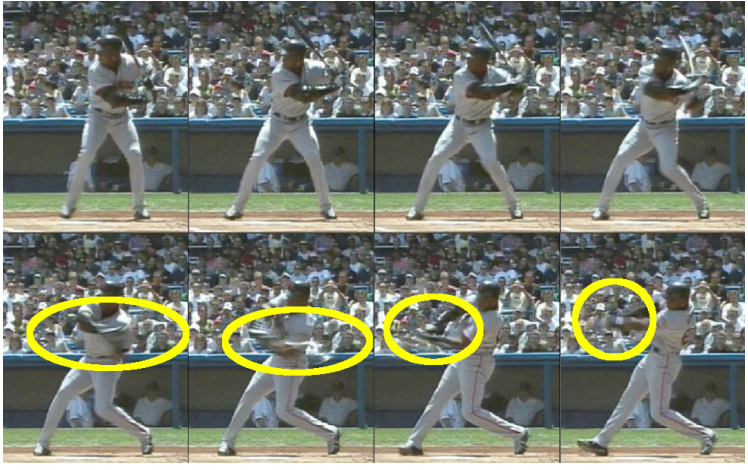
SpikeVision: MATMUL Free Transformer-Inspired SNN on FPGA

Conclusions and Outlook

Conventional Image Sensors: Time-based Sampling



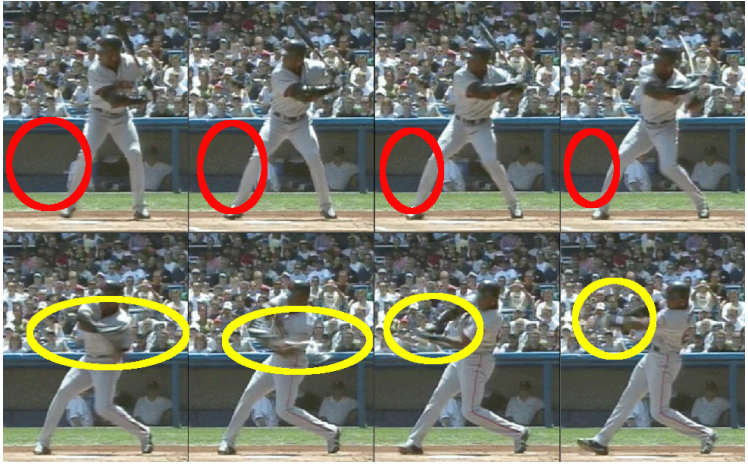
Conventional Image Sensors: Time-based Sampling



under-sampling

- motion blur
- displacement between frames

Conventional Image Sensors: Time-based Sampling



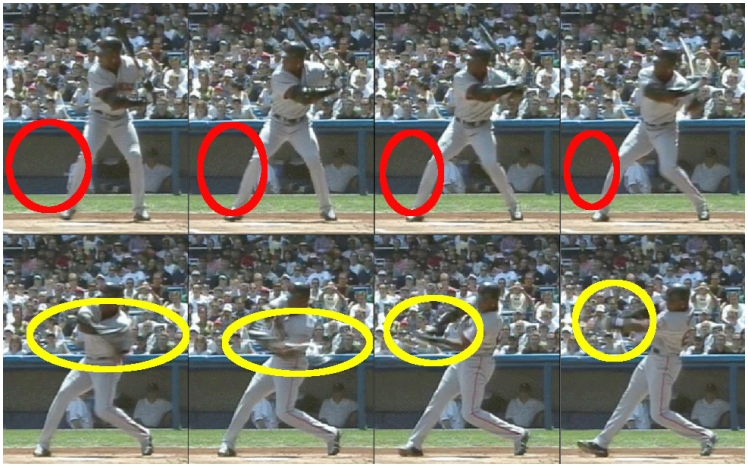
under-sampling

- motion blur
- displacement between frames

over-sampling

- redundant useless data
- power and resource hungry

Conventional Image Sensors: Time-based Sampling



under-sampling

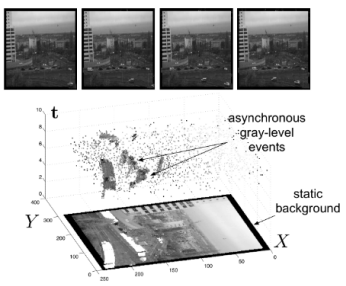
- motion blur
- displacement between frames

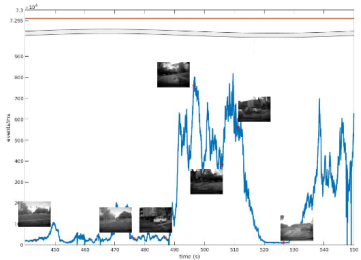
over-sampling

- redundant useless data
- power and resource hungry

Conventional Sampling: Impossible Trade-off of Power vs "Frame Rate"!

Sensing in the Time Domain: Event-Sampling



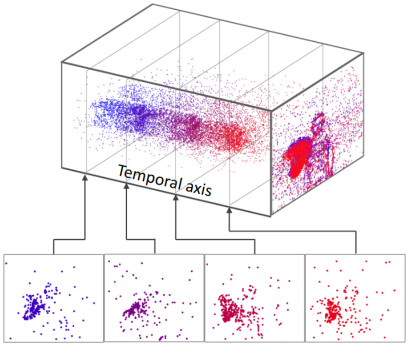


- Events and no frames
- Data is acquired asynchronously
- High-temporal resolution (<3μs)
- Low-power
- High-dynamic range

- Low-data load
- 2-3 orders of magnitude compared to standard cameras
- Adapted to dynamic scenes analysis

TU/e

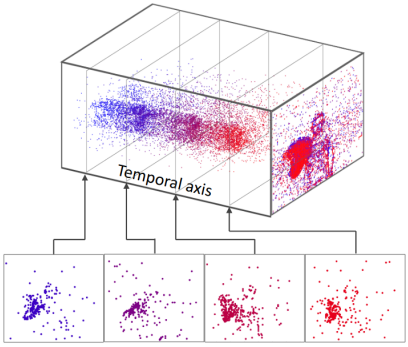
Event-Sampling is Not Yet Fully Exploited!



Creating fixed-interval frames from events loose properties of event-based data.

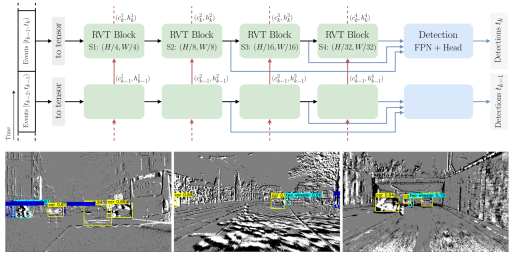
[He, Weihua, et al., Neural Networks 2020]

Event-Sampling is Not Yet Fully Exploited!



Creating fixed-interval frames from events loose properties of event-based data.

[He, Weihua, et al., Neural Networks 2020]

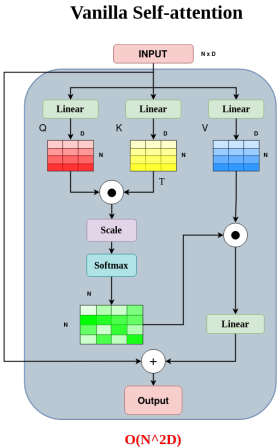


Deep learning models and associated GPUs hardware are: inadequate, power hungry, and carry many wasteful operations.

[Gehrig, M., & Scaramuzza, D. 2023]

TU/e

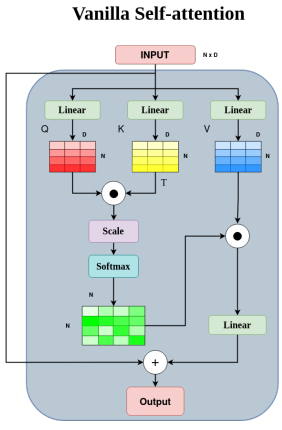
Recent Advancements: Vision Transformer (ViT)



$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

[Dosovitskiy A. et al, 2021] [Liu Y. et al, IEEE TNNLS, 2023]

Recent Advancements: Vision Transformer (ViT)



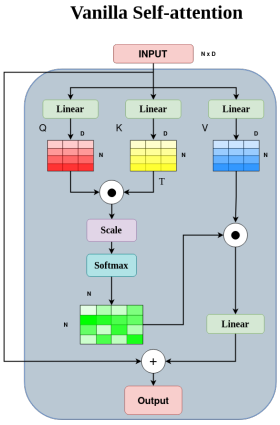
$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Pros:

- Long-distance (temporal) relations
- High accuracy across tasks

[Dosovitskiy A. et al, 2021] [Liu Y. et al, IEEE TNNLS, 2023]

Recent Advancements: Vision Transformer (ViT)



$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Pros:

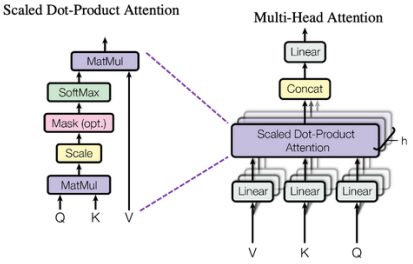
- Long-distance (temporal) relations
- High accuracy across tasks

Cons:

- High computational requirements

[Dosovitskiy A. et al, 2021] [Liu Y. et al, IEEE TNNLS, 2023]

State-of-the-Art: SNN + self-attention



MODEL	Advantages	Disadvantages
SpikeGPT [2]	Generative language model	Activation layer
Spikeformer [3]	Spiking Q, K and V	Matrix multiplications
TE-Spikformer [4]	Balance the neuron activation	Matrix multiplications
SpikingViT [5]	Time extension embedding	Matrix multiplications
Spike-driven Transformer [6]	No matrix multiplications in encoder blocks	Head-classification requires real-value MATMUL & Encoder is an MLP

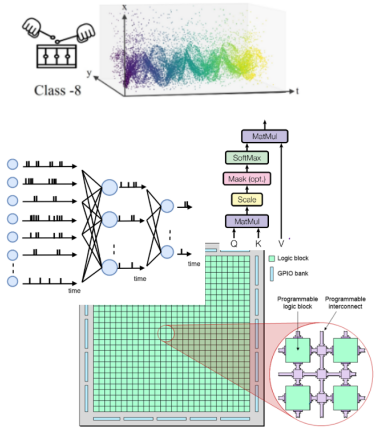
These models are all using "Time based frames".

GAP: Event-driven, no MATMUL, edge-compatible

The Focus of This Work

Research question

Is it possible to obtain a highly accurate attention-inspired computational efficient (**SNN-based**) model that can take advantage of both: spatial sparsity and high-temporal resolution of event-based cameras?



[Z. Shen, F. Corradi *SpikeVision: A Fully Spiking Neural Network Transformer-Inspired Model for Dynamic Vision Sensors*, IEEE Asilomar, 2024]

TU/e

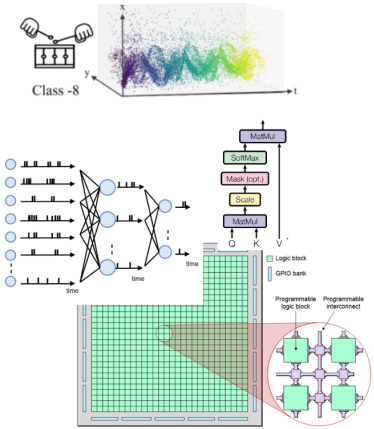
The Focus of This Work

Research question

Is it possible to obtain a highly accurate attention-inspired computational efficient (**SNN-based**) model that can take advantage of both: spatial sparsity and high-temporal resolution of event-based cameras?

Requirements

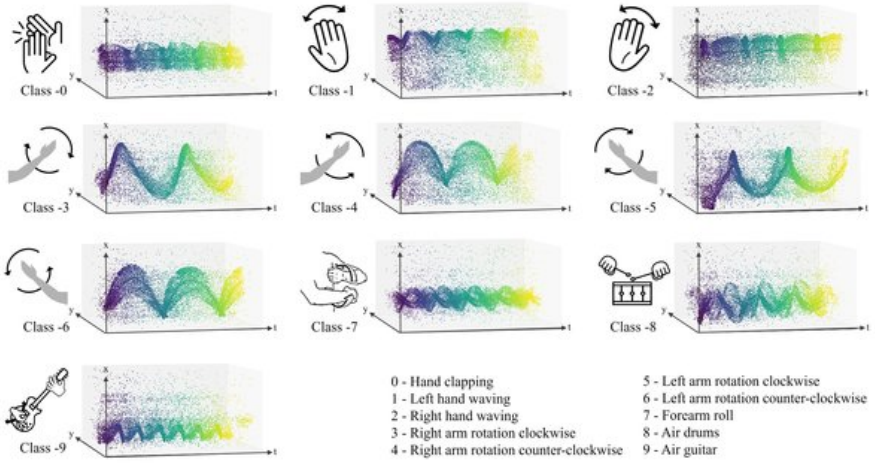
- Fully Event-based
- Fully-Binary Communication
- No MATMUL
- Hardware Friendly (FPGA)



[Z. Shen, F. Corradi *SpikeVision: A Fully Spiking Neural Network Transformer-Inspired Model for Dynamic Vision Sensors*, IEEE Asilomar, 2024]

TU/e

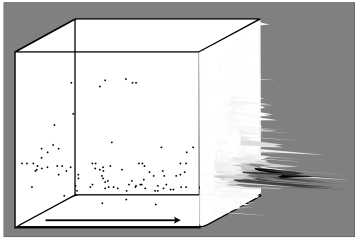
Dataset IBM DVS128

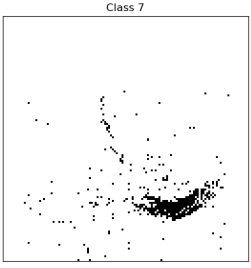


[A. Amir et al, *A Low Power, Fully Event-Based Gesture Recognition System*, IEEE CVPR 2017]

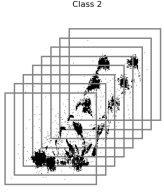
TU/e

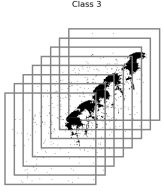
Idea 1: Encoding Event-based Data while Preserving Time Resolution

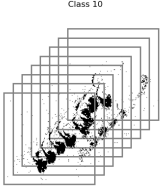


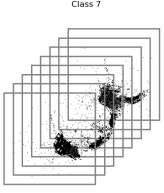


× No fixed time interval. ✓ **Use fixed event amount.**

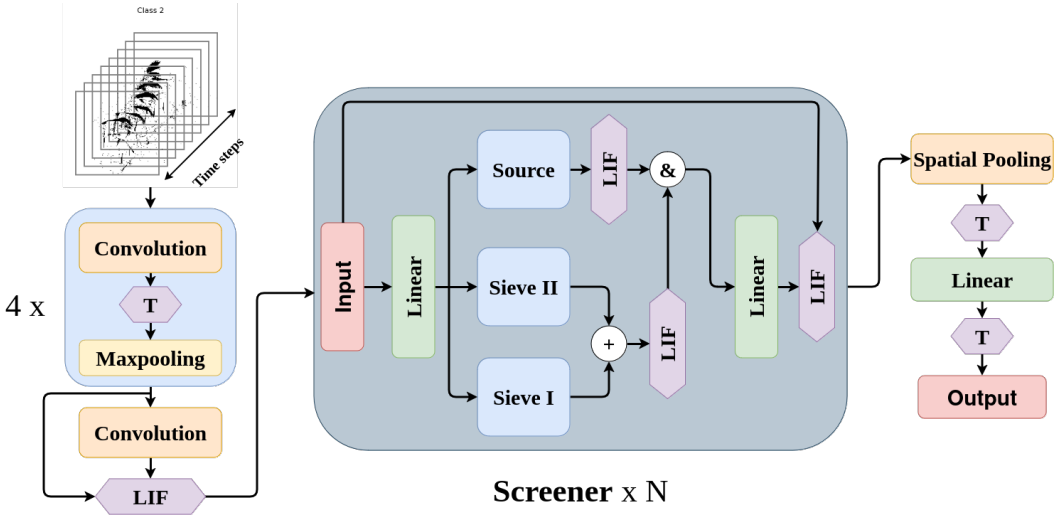






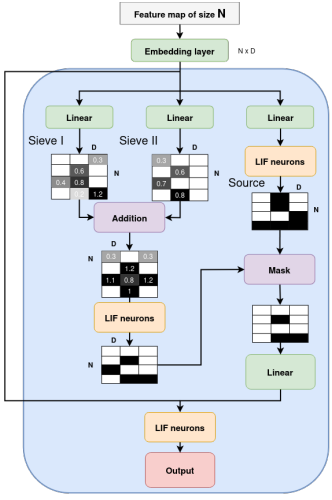


Idea 2: Fully SNN Transformer-inspired Model for DVS



TU/e

Attention-Inspired: Screener

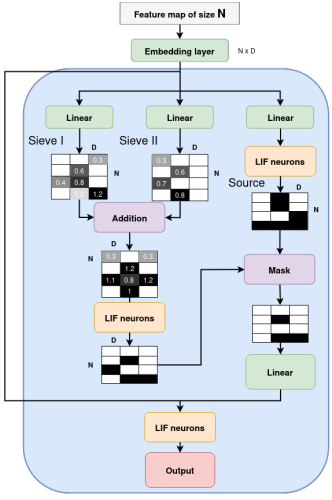


Attention-inspired Screener

- Fully Event-Driven.

TU/e

Attention-Inspired: Screener

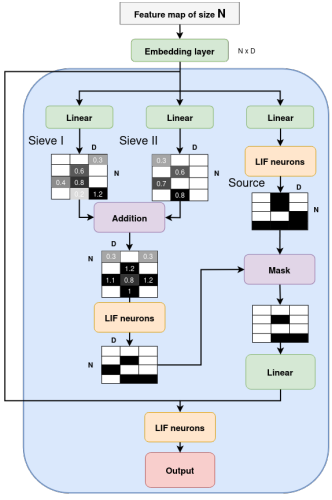


Attention-inspired Screener

- Fully Event-Driven.
- Keep fine grained temporal information.

TU/e

Attention-Inspired: Screener

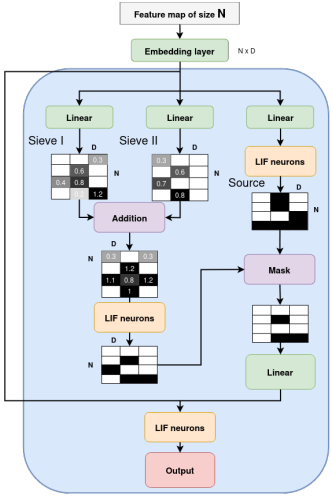


Attention-inspired Screener

- Fully Event-Driven.
- Keep fine grained temporal information.
- Track spatial and temporal dependencies.

TU/e

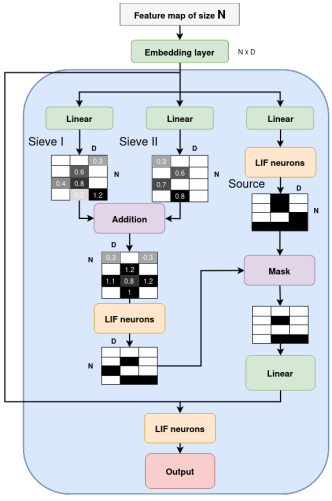
Attention-Inspired: Screener



Attention-inspired Screener

- Fully Event-Driven.
- Keep fine grained temporal information.
- Track spatial and temporal dependencies.
- The Sieves act like filters to the Source: they come up with an agreement and drop the unimportant pixels in the Source matrix.

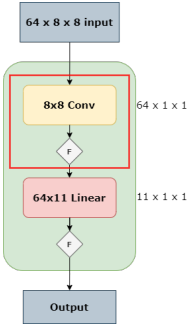
Attention-Inspired: Screener



Attention-inspired Screener

- Fully Event-Driven.
- Keep fine grained temporal information.
- Track spatial and temporal dependencies.
- The Sieves act like filters to the Source: they come up with an agreement and drop the unimportant pixels in the Source matrix.
- Skip connection to improve training stability and gradient flow.

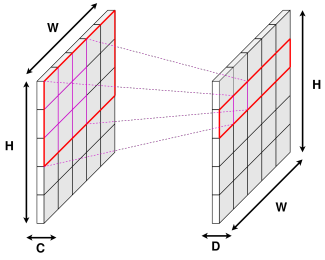
Head-Classification



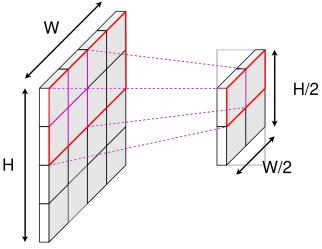
Head-Classification

- Fully Event-Driven.
- The output that emit most of the events $\rightarrow$ chosen class.

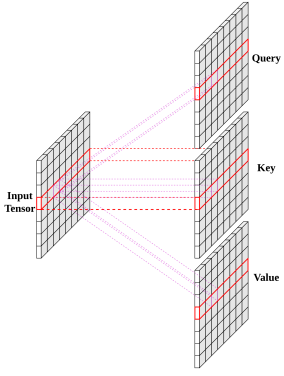
Idea 3: Dataflow Optimization



Convolution



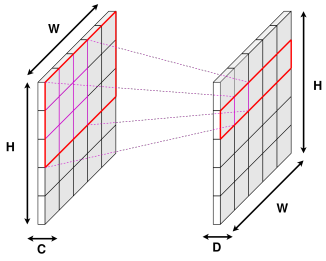
Max pooling



Screener

TU/e

Dataflow Optimization



Convolution

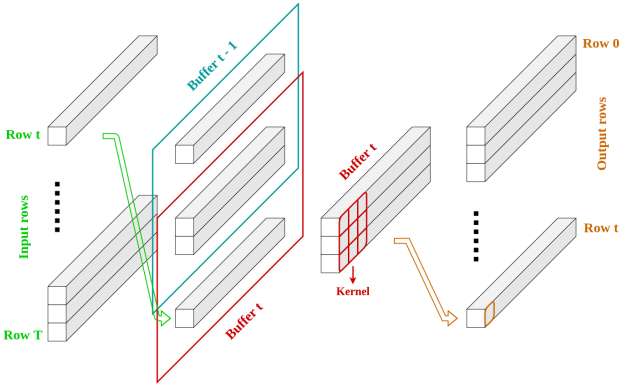
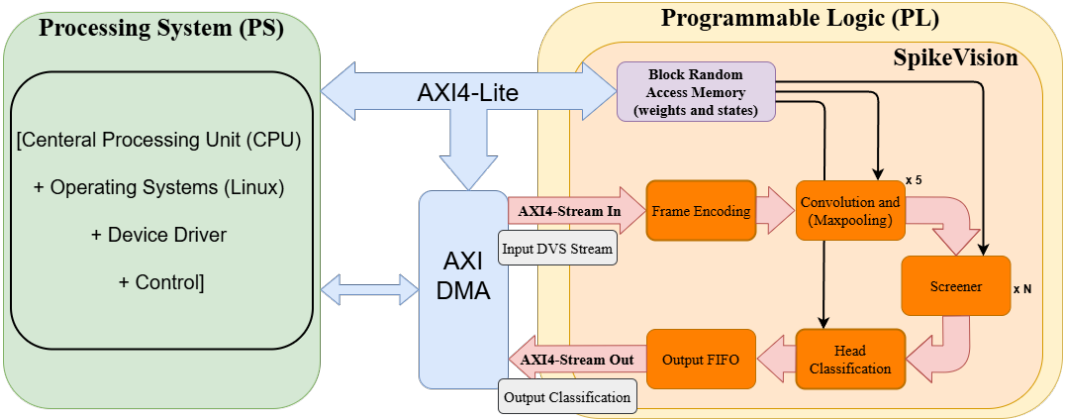


Figure: Row-by-row convolution as an example

TU/e

Hardware Design

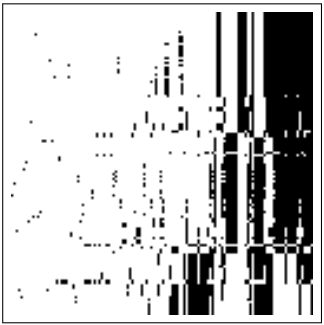


TU/e

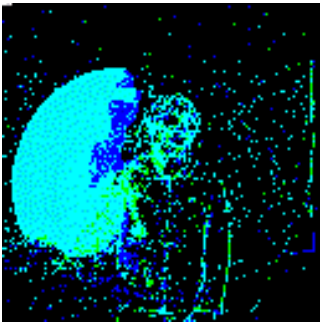
Benchmarks & Datasets

MNIST

Genomics



DVS gesture



Permuted-MNIST

Results: Accuracy & Model Generalization & Scaling

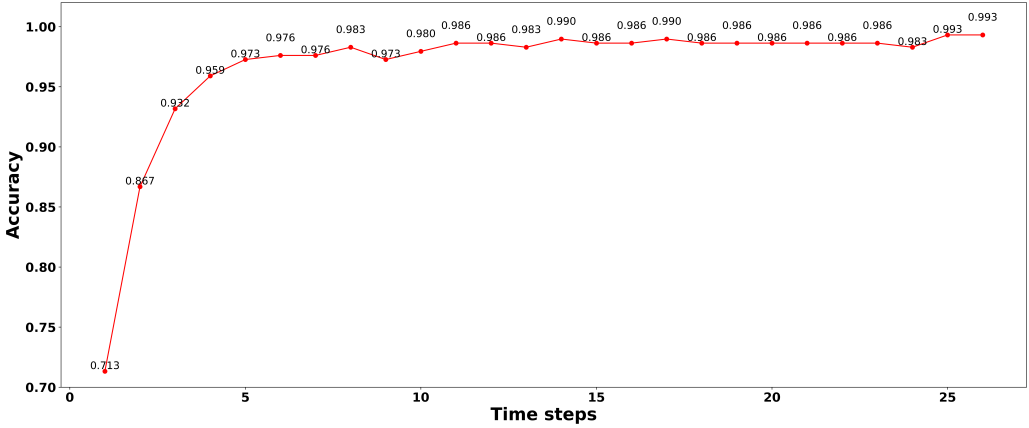
Table: Performance on MNIST, PSMNIST and genomics

Model	MNIST	PMNIST	genomic1	genomic2	genomic3	genomic4	genomic5	genomic6	genomic7	genomic8
SweepNet [11]	-	-	100%	92%	99%	89%	89%	99%	79.5%	99%
SCNN [12]	-	-	100%	90.7%	99%	87.2%	84.35%	94.35%	74.05%	98.6%
SpikeVision	99.3%	95.7%	99.9%	92.35%	99.2%	90.15%	89.65%	98.7%	80.3%	99%

Table: Performance on DVS gesture

Model	embedded dimension	layers of encoders	accuracy
SDT (Spike-driven Transformer) [6]	256	2	98.3%
SpikeVision(ours)	64	1	93.86%
SpikeVision(ours)	64	2	94.2%
SpikeVision(ours)	128	1	96.59%
SpikeVision(ours)	128	2	97.61%
SpikeVision(ours)	256	1	96.93%
SpikeVision(ours)	256	2	99.3%

Results: Latency & Accuracy (Trade-off)



TU/e

Results: Resource Utilization & Latency (Trade-off)

Block design	BRAM	BRAM (%)	FIFO	FIFO (%)	LUT	LUT (%)	LUTRAM	LUTRAM (%)
ZCU102	912	-	548160	-	274080	-	144000	-
8-bit Row-by-row	445	48.79	210351	38.37	114885	41.92	1920	1.33

	Conv1 (ns)	MP1 (ns)	Conv2 (ns)	MP2 (ns)	Conv3 (ns)	MP3 (ns)	Conv4 (ns)	MP4 (ns)	Conv5 (ns)	Enc (ns)	Head (ns)	Latency (s)	FPS	bs	Power (W)
Flatten	9.44×10^{7}	5.24×10^{6}	8.10×10^{8}	2.62×10^{6}	7.82×10^{8}	1.31×10^{6}	7.69×10^{8}	6.55×10^{5}	3.82×10^{8}	1.29×10^{8}	2.07×10^{5}	3.11	(7.75)	-	4.404
Row-by-row	8.21×10^{4}	2.05×10^{4}	6.56×10^{5}	2.05×10^{4}	2.62×10^{6}	2.05×10^{4}	5.24×10^{6}	2.05×10^{4}	5.55×10^{6}	1.09×10^{7}	1.92×10^{5}	0.138	(7.28)	-	4.164
GPU (A100)	-	-	-	-	-	-	-	-	-	-	-	0.157	638.64	20	300
CPU	-	-	-	-	-	-	-	-	-	-	-	3.98	25.11	20	180

Zhanbo Shen, and Federico Corradi, *SpikeVision: A Fully Spiking Neural Network Transformer-Inspired Model for Dynamic Vision Sensors, IEEE Asilomar, 2024*

Real-time Demonstrator

Live demonstrator

Show accurate real-time performance on Human Activity and Gesture Recognition (HAR): robust to several light conditions, exploits the high-temporal resolution, fully event-based, and easily implemented in FPGA.

PROPHESEE
METAVISION FOR MACHINES

Zhanbo Shen, and Federico Corradi, *SpikeVision: A Fully Spiking Neural Network Transformer-Inspired Model for Dynamic Vision Sensors, IEEE Asilomar, 2024*

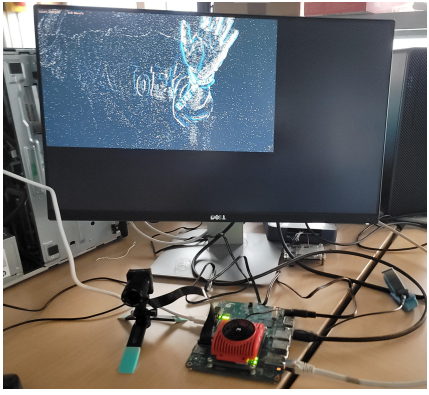


Figure: Prophesee event-based camera & Xilinx FPGA

Outline

Edge Artificial Intelligence

SpikeVision: MATMUL Free Transformer-Inspired SNN on FPGA

Conclusions and Outlook

Summary & Discussions

In this work, we proposed

- A new SNN transformer-inspired model with high accuracy, using feature from spatially sparse and high-temporal resolution event-based data.
- A new dataflow implementing a finer pipeline.
- A digital hardware SNN architecture for FPGAs.

Summary & Discussions

In this work, we proposed

- A new SNN transformer-inspired model with high accuracy, using feature from spatially sparse and high-temporal resolution event-based data.
- A new dataflow implementing a finer pipeline.
- A digital hardware SNN architecture for FPGAs.

Discussions

- Training SNNs is still a complex task
  - Vanishing gradients
- Exploring multiple Sieves stages
- Ablation studies (ANN vs SNN)

Summary & Discussions

In this work, we proposed

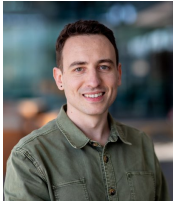
- A new SNN transformer-inspired model with high accuracy, using feature from spatially sparse and high-temporal resolution event-based data.
- A new dataflow implementing a finer pipeline.
- A digital hardware SNN architecture for FPGAs.

Discussions

- Training SNNs is still a complex task
  - Vanishing gradients
- Exploring multiple Sieves stages
- Ablation studies (ANN vs SNN)

Main Contributor

- Zhanbo Shen (PhD student)



Key Enabling Technology NWO: IMAGINE project.

Reference

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is All You Need,” 2017.

[2] R.-J. Zhu, Q. Zhao, G. Li, and J. K. Eshraghian, ‘SpikeGPT: Generative Pre-trained Language Model with Spiking Neural Networks’. arXiv, Jun. 26, 2023. Accessed: Dec. 20, 2023. [Online]. Available: http://arxiv.org/abs/2302.13939

[3] Z. Zhou et al., ‘Spikformer: When Spiking Neural Network Meets Transformer’. arXiv, Nov. 22, 2022. Accessed: Jan. 02, 2024. [Online]. Available: http://arxiv.org/abs/2209.15425

[4] S. Gao, X. Fan, X. Deng, Z. Hong, H. Zhou, and Z. Zhu, ‘TE-Spikformer: Temporal-enhanced spiking neural network with transformer’, Neurocomputing, vol. 602, p. 128268, 2024. doi: 10.1016/j.neucom.2024.128268.

[5] L. Yu et al., "SpikingViT: a Multi-scale Spiking Vision Transformer Model for Event-based Object Detection," in IEEE Transactions on Cognitive and Developmental Systems, doi: 10.1109/TCDS.2024.3422873.

[6] M. Yao et al., ‘Spike-driven Transformer’. arXiv, Jul. 04, 2023. Accessed: Jan. 02, 2024. [Online]. Available: http://arxiv.org/abs/2307.01694

[7] A. Amir et al., ‘A Low Power, Fully Event-Based Gesture Recognition System’, in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI: IEEE, Jul. 2017, pp. 7388–7397. doi: 10.1109/CVPR.2017.781.

[8] X. She, S. Dash, and S. Mukhopadhyay, ‘SEQUENCE APPROXIMATION USING FEEDFORWARD SPIKING NEURAL NETWORK FOR SPATIOTEMPORAL LEARNING: THEORY AND OPTIMIZATION METHODS’, 2022.

[9] S. U. Innocenti, F. Becattini, F. Pernici, and A. Del Bimbo, ‘Temporal Binary Representation for Event-Based Action Recognition’. arXiv, Oct. 18, 2020. Accessed: May 02, 2024.

[10] M. Schöne, N. M. Sushma, J. Zhuge, C. Mayr, A. Subramoney, and D. Kappel, ‘Scalable Event-by-event Processing of Neuromorphic Sensory Signals With Deep State-Space Models’. arXiv, Apr. 29, 2024. Accessed: May 02, 2024. [Online]. Available: http://arxiv.org/abs/2404.18508

[11] S. Yoo, E. Y.-J. Lee, Z. Wang, X. Wang, and W. D. Lu, ‘RN-Net: Reservoir Nodes-Enabled Neuromorphic Vision Sensing Network’. arXiv, May 29, 2023. Accessed: May 02, 2024.

[12] M. Yao et al., ‘Temporal-wise Attention Spiking Neural Networks for Event Streams Classification’, in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada: IEEE, Oct. 2021, pp. 10201–10210.

Reference

[11] H. Zhao, P. Pavlidis, and N. Alachiotis, 'SweepNet: A Lightweight CNN Architecture for the Classification of Adaptive Genomic Regions', in Proceedings of the Platform for Advanced Scientific Computing Conference, Davos Switzerland: ACM, Jun. 2023, pp. 1–10. doi: 10.1145/3592979.3593411.

[12] Federico Corradi, Zhanbo Shen, Hanqing Zhao, and Nikolaos Alachiotis. 2024 (in press). Accelerated Spiking Convolutional Neural Networks for Scalable Population Genomics. In . ACM, New York, NY, USA, 15 page.

Interested? Reach out!

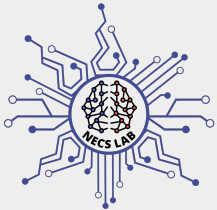
Neuromorphic Edge Computing Systems Lab

Federico Corradi

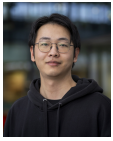
Eindhoven University of Technology

f.corradi@tue.nl, +31(0)402 472 556

Neuromorphic Edge Computing Systems Lab



Neuromorphic Edge Computing Systems Lab

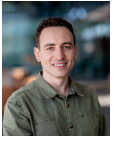

Zhanbo Shen


Shimeng Ye


Dr. Roel Jordans


Lorenzo Pes


Yvonne Vullers


Dr. Federico Corradi


Dr. Bojian Yin


Stefano Chiavazza