TENN: A highly efficient transformer replacement for edge and event processing

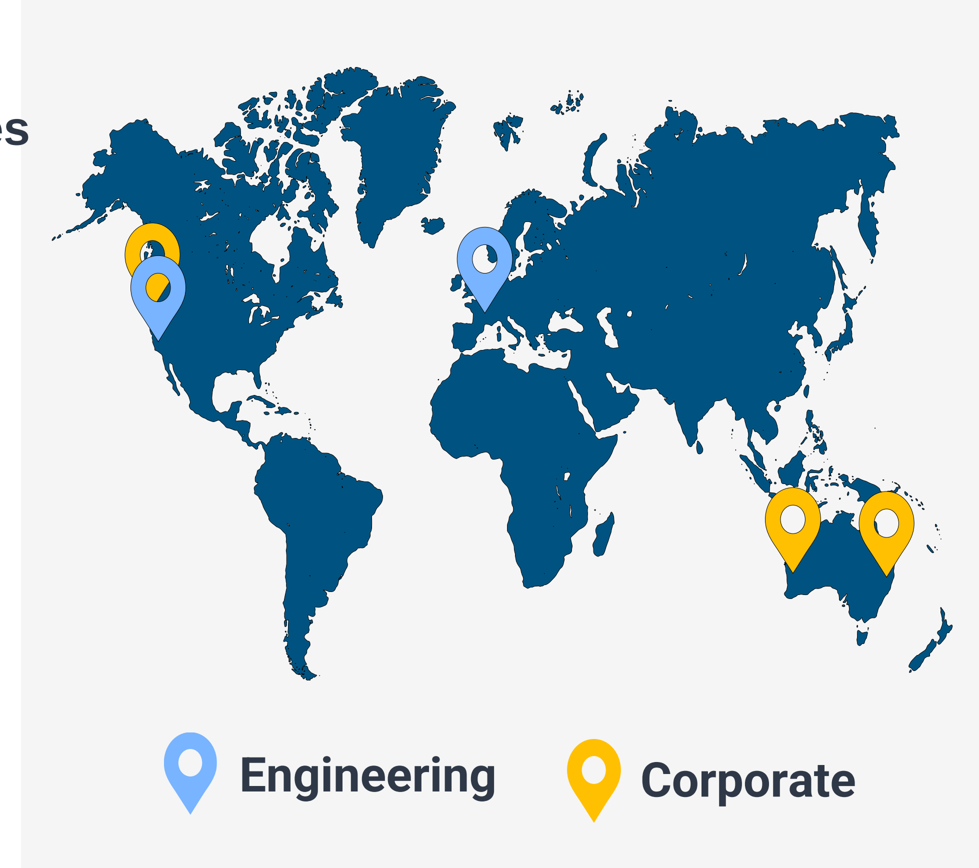
M Anthony Lewis, Yan Ru Pei and Olivier Coenen

April 25, 2024

brainchip
Essential AI

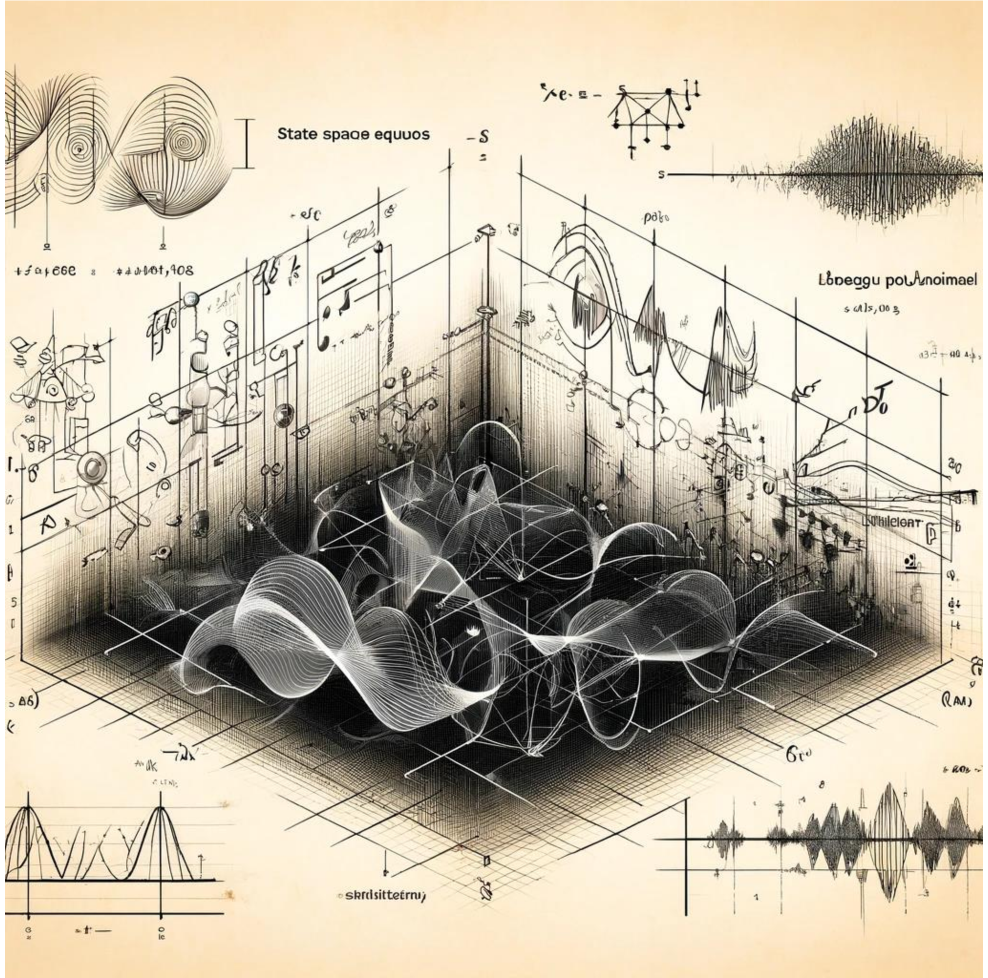
About BrainChip– Founded 2013

- **Business Model: IP License**
- **15+ yrs fundamental AI architecture research & technologies**
- **65+ data science, hardware & software engineers**
- **Publicly traded Australian Stock Exchange (BRN:ASX)**
- **10 Customers - Early Access, Proof of Concept, IP License**
  - Automotive
  - Consumer
  - Healthcare
  - Imaging
  - Transportation



TENN can reduce energy use by orders of magnitude

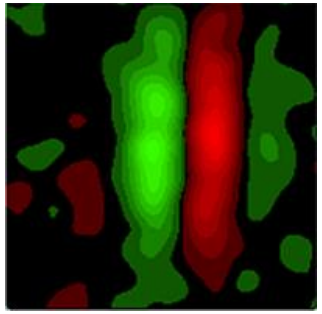
- **TENN** = TEMPORAL EVENT-BASED NEURAL NETWORK
- TENN is related to **State Space Models**
- Replacement for many Transformer tasks
  - Language Models
  - Time-series Data
  - Spatiotemporal Data
- Dramatically lowers energy requirements across all compute platforms



Kernel Representation Evolution

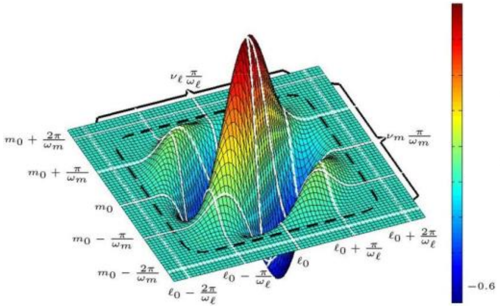
The journey from neurons to polynomials

Receptive Field of V1 Hubel & Wiesel, 1959, 1962



Receptive Field of a simple cell
DeAngelis et al., 1995)

Gabor filter: *continuous parametric models of receptive fields*
Popular in the 1990s.



Gabor filter

- A gabor filter is a combination of a gaussian filter and a sinusoidal term.

A gabor filter in 2 dimension is :

$$g(x,y;\lambda,\theta,\psi,\sigma,\gamma) = \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right)\cos\left(2\pi\frac{x'}{\lambda}+\psi\right)$$

Replaced by learnable kernels in deep learning.

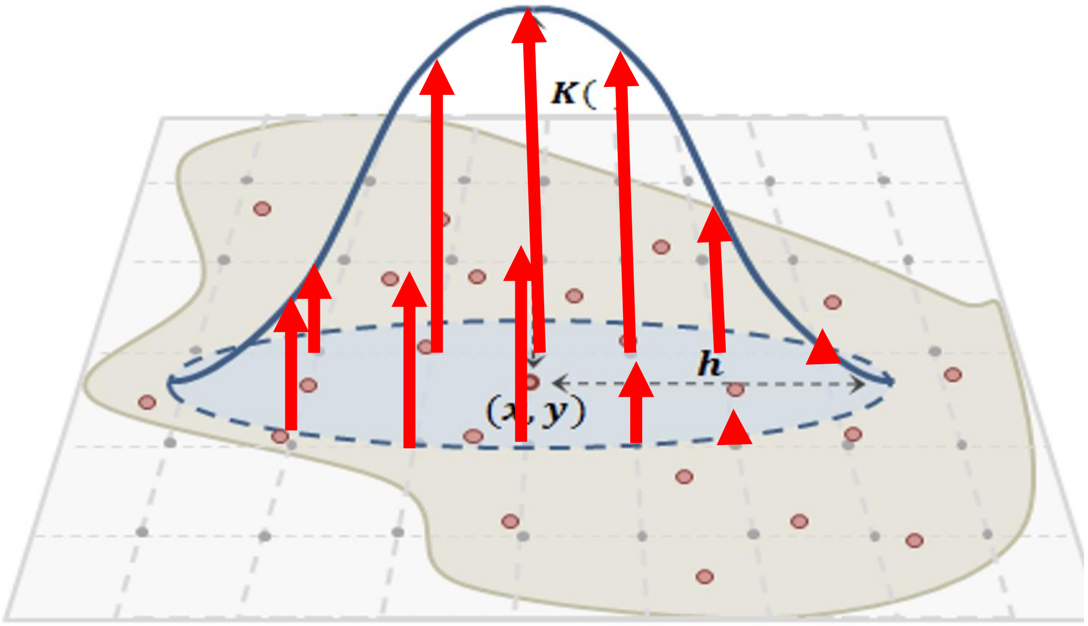
Brachmann & Redies
2016

brainchip
Essential AI

What price, Learnable Kernels?

- Explosion of parameters
- Discretization in time and space
- Time is particular problematic for event-based systems
- Learning is inextricably linked to a clock in conventional Deep Learning

Alternatives?

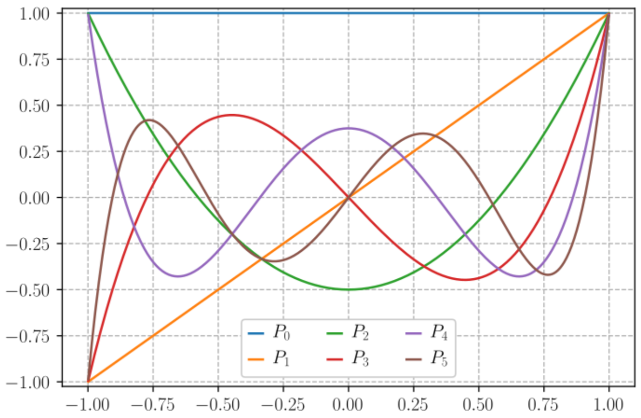


W_{ij}: red arrows

Representing time-series with orthogonal polynomials

BrainChip uses Chebyshev polynomial

Legendre polynomials

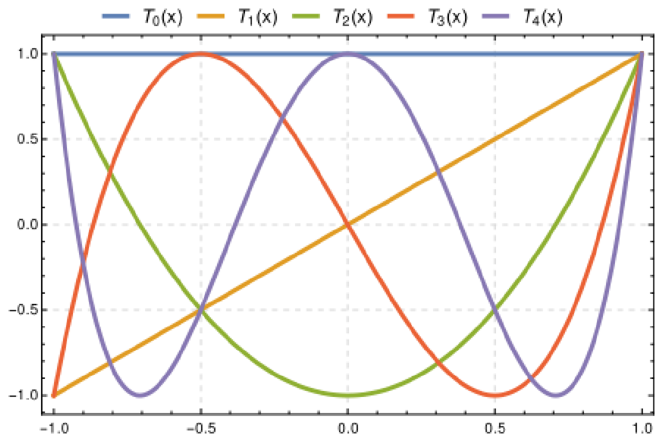


$$\frac{d}{dx}P_{n+1}(x) = (n+1)P_n(x) + x\frac{d}{dx}P_n(x)\,.$$

In Legendre polynomials basis can lead to exponential convergence for analytic functions.

Intolerant to discontinuities

Chebyshev polynomials



Chebyshev polynomial basis can lead to exponential convergence for a wide range of functions, including those with singularities or discontinuities.*

$$T_0(x) = 1$$
$$T_1(x) = x$$
$$T_{n+1}(x) = 2x\,T_n(x) - T_{n-1}(x).$$

*Lloyd N. Trefethen. 2019. Approximation Theory and Approximation Practice, Extended Edition. SIAM-Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

TENN has two modes: Convolution (Kernel) and Recurrent

Principles:

(1) Recurrence: Chebyshev and Legendre polynomials have recurrence relationship.

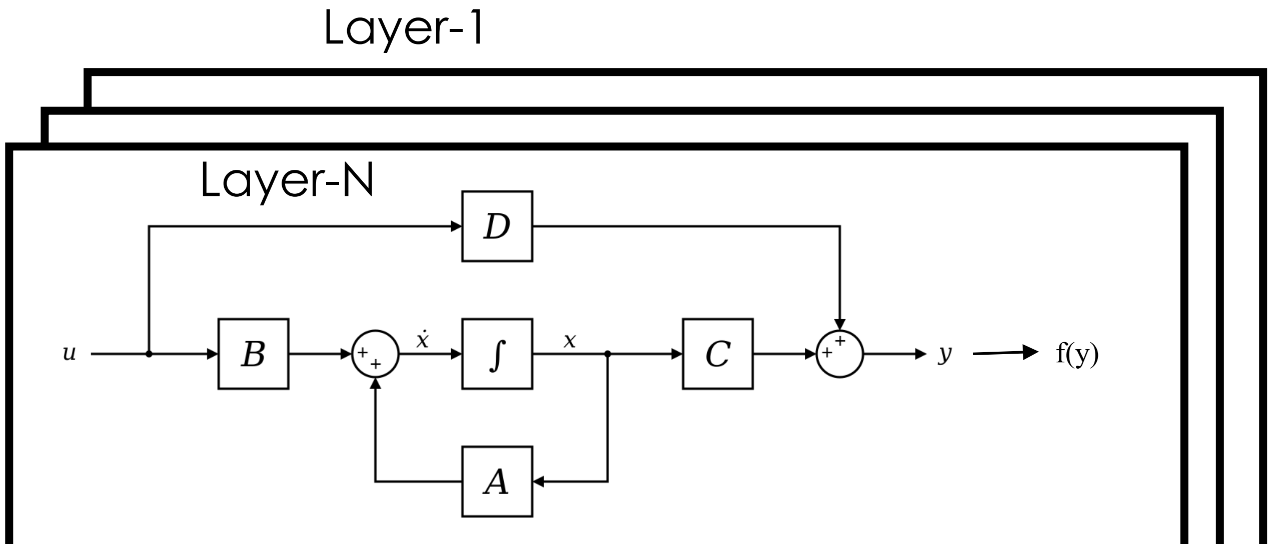
(2) Duality: Recurrence imputes duality: Convolutional form as well as recurrent form.

(3) Stable training: Train in Convolutional Domain

(4) Fast Running: Run in recurrent domain. Small foot-print

(5) Insight: TENNs and SSM are a stack of generalized Fourier filters running in a recurrent mode, with non-linearities between layers.

Surprise: Inspiration is from sophisticated signal processing but works with LLMs !!!



$$\dot{x} = Ax + Bu$$
$$y = Cx + Du$$
$$f(y)$$

brainchip
Essential AI

Recurrent to Convolution

Put it all together: recurrence, state space and kernel fine-tunning

A Matrix is initialized S.T. the resulting LTI system convolves the input U with polynomial basis.

A matrix leverages recurrence relationship of Chebyshev polynomials

$$x_n = Ax_{n-1} + Bu_n$$

$$y_n = Cx_n$$

where $x \in \mathbb{R}^p,\ u \in \mathbb{R}^h,\ y \in \mathbb{R}^q$

The recurrence relationship can be unfolded into a convolutional representation

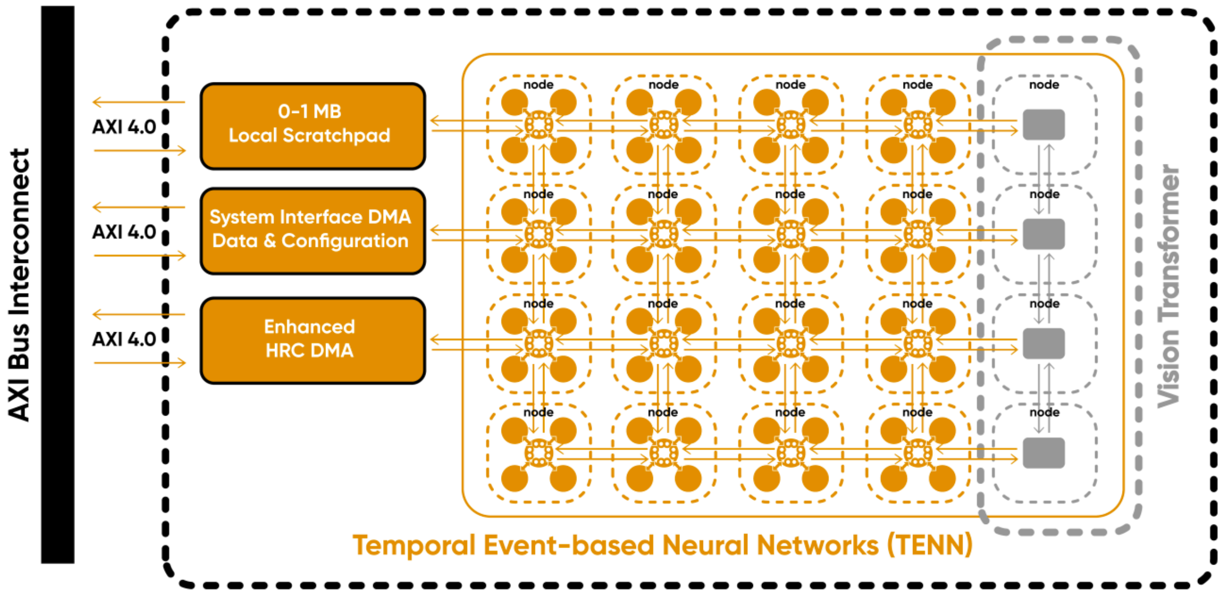
$$C[A^0, A^1, A^2, \dots, A^\infty]B$$

Parameterized by three matrices: A, B, C

We can now "fine-tune" the basis to create a better, low dimensional fit. Lose some of the time independence & orthogonality, however.

TENN Support in Akida 2.0

Akida 2.x Architecture and Benefits



Key hardware Features

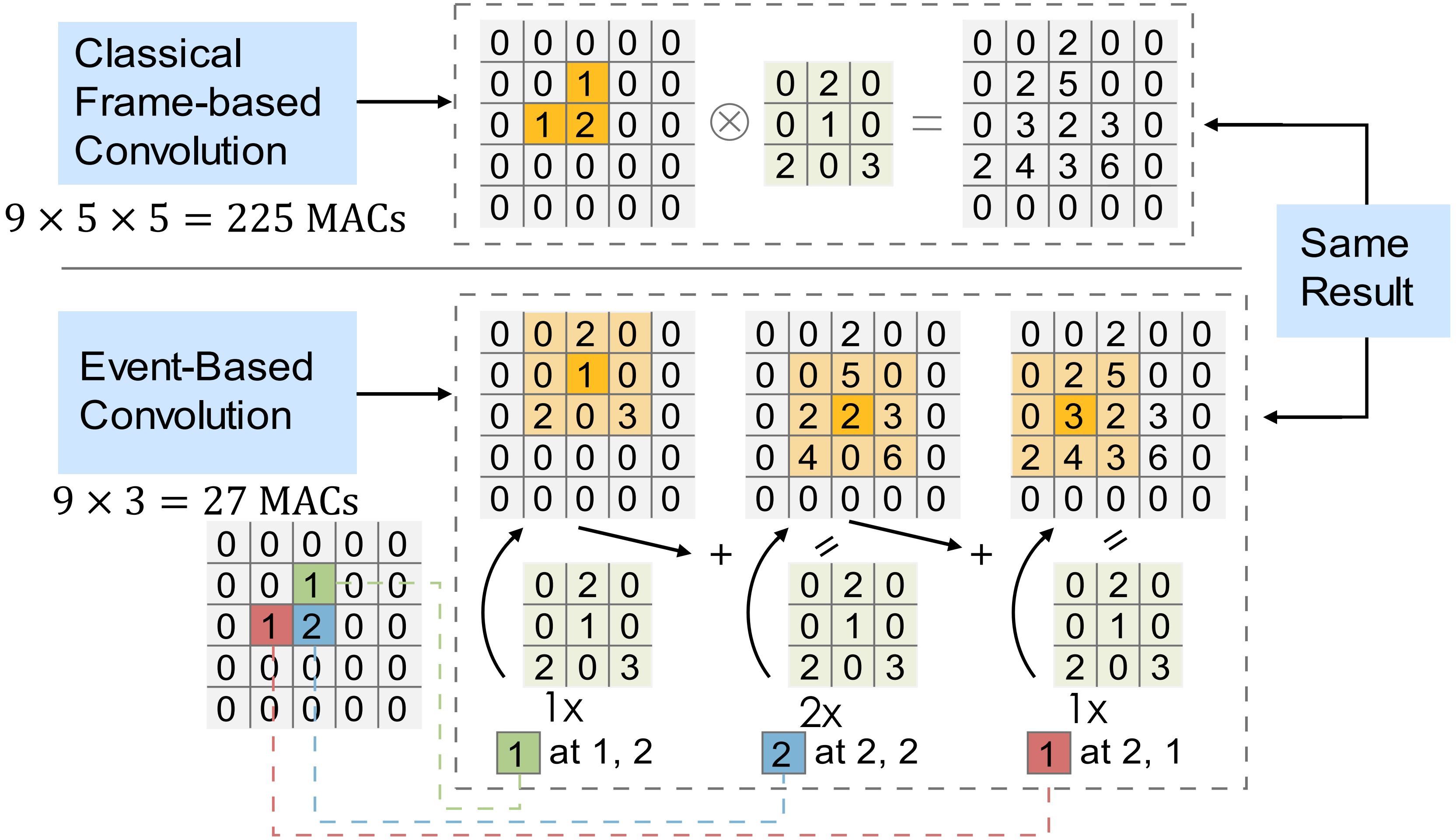
Digi

- Digital, Event-Based, at memory compute
- Highly Scalable
- Each Node connected by mesh network
- Inside each node is a event based TENN processing unit

brainchip
Essential AI

Event-Based Convolution, 2-D example

Benefits from Activation Sparsity



Research Roadmap for TENNs

One network: many uses

Audio

Denoising

Keyword spotting

Automatic Speech

Recognition

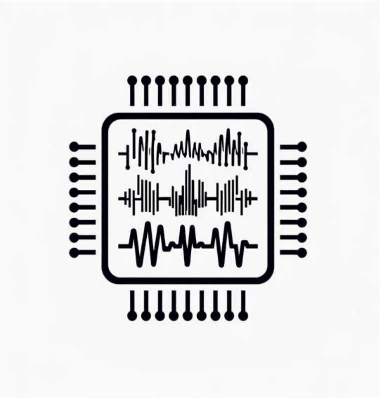
Raw Audio processing

Generative AI

- **Large Language Models**
- Intelligent Agents
- Primitive Reasoning
- LLama 1B Params equiv

Industrial AIoT

- **Condition Monitoring**
- Anomaly Detection
- Counting



BioMedical

- **EEG /EKG /EMG**
- Wearables for health
- Activity Monitoring
- VR/AR interface

TENN Performance

The following results are performance projections

brainchip
Essential AI

Task: Sentence generation

TENN is highly competitive with models of similar size

1. TENN trained on WikiText-103. 100M tokens
2. GPT models trained on open_web_text, Mamba trained on the Pile
3. TENN training time: ~3 days on (1) A100
4. Scores reported as negative entropy: $-log_2(1/VocabSize) - log_2(perplexity)$ (higher better)

Model	GPT2 Small	GPT2 Medium	TENN	Mamba 130M	GPT2 large	GPT2 full	Mamba 370M
Train_size	13 GB	13GB	0.1 GB	836GB	13GB	13GB	836GB
Score	9.7	10.2	10.3	10.4	10.4	10.8	10.9
Params (relative to TENN)	2	5.6	1	2.06	12.3	25	5.9
Energy (relative to TENN)	1700	5700	1	2.06	13000	27000	5.9

brainchip
Essential AI

TENNS generates tokens far faster than GPT-2 medium

Both models were prompted with the first 1024 words of the Harry Potter 1st novel

Inference done on a single CPU thread

TENN (ours):

gpt2-medium (theirs):

```
HARRY WAS COMPLETELY AFRAID
```

Task: Audio Denoising

Comparison of TENN versus SoTA

Model	Deep Filter Net V1	TENN	Deep Filter Net V2	Deep Filter Net V3
PESQ	2.49	2.61	2.67	2.68
Params (relative to TENN)	2.98	1	3.86	3.56
MACs (relative to TENN)	11.7	1	12.1	11.5

Noisy samples

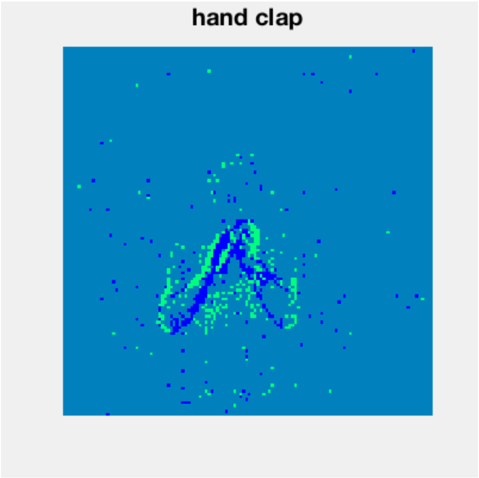
TENN (raw audios in and out)

Denoised

TENN can be extended to spatio-temporal dat

DVS Hand Gesture Recognition: IBM DVS128 Dataset

Network	Accuracy (%)	Parameters	MACs (billion) / sec	Latency* (ms)
TrueNorth-CNN	96.5	18 M	-	155
Loihi-Slayer	93.6	-	-	1450
ANN-Rollouts	97.0	500 k	10.4	1500
TA-SNN	98.6	-	-	1500
Akida-CNN	95.2	138 k	0.12	200
TENN-Fast	97.6	192 k	0.429	105
TENN	**100.0**	192 k	0.499	510



State of the Art SOTA

brainchip
Essential AI

Key Take aways

- **TENN**
  - Is highly power efficient
  - Can be mapped to Akida 2.0 IP
  - SoTA performance in areas explored to date
- **Future Work**
  - Enhance activation sparsity to take advantage of Akida 2.0 IP
  - Further Exploration of polynomial space