Temporal Event Neural Networks: A More Efficient Alternative to the Transformer

Chris Jones

Director Product Management

BrainChip Inc.

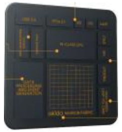
Brainchip AI – At a Glance

embedded VISION SUMMIT®

- **First to commercialize** neuromorphic IP platform and reference chip.
- **15+ yrs** fundamental research
- **65+** data science, hardware & software engineers
- **Publicly traded Austrialian Stock Exchange (BRD:ASX)**
- **10 Customers** – Early Access, Proof of Concept, IP License

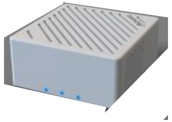
PRODUCTS

akida – IP



Reference SoC

metaTF – Software Tools



Edge Box*

TRUSTED BY

MegaChips, Mercedes-Benz, NASA, RENESAS, VORAGO TECHNOLOGIES (Opening up new possibilities), Valeo

PARTNERS

arm, intel foundry services, EDGE IMPULSE, TEKSUN (CULTIVATING TECHNOLOGY), PROPHESEE, Ai Labs, EMOTION3D, NVISO, SiFive

*Fulfillment through VVDN technologies

©2024 BrainChip Inc.

Key Focal Areas

- Provide path to run complex models on the Edge
- Reduce cost of training
- Reduce cost of inference

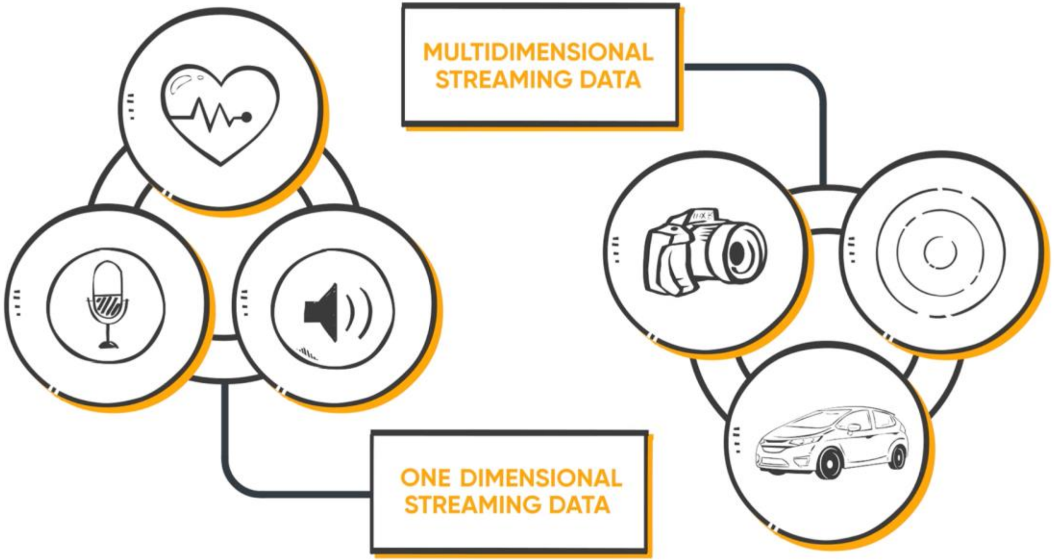
©2024 BrainChip Inc.
©2024 BrainChip Inc.

Temporal Event Neural Networks (TENNs)

©2024 BrainChip Inc.

Change the Game

Unleash Unprecedented Edge Devices



Up to 5000X
More Energy Efficient

Up to 50X
Fewer Parameters

10-30X
Lower Training cost vs. GPT-2

Same Or Better Accuracy

©2024 BrainChip Inc.

TENNs Application Areas

Spatiotemporal Integration

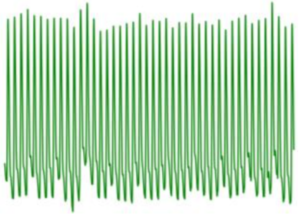
1. Multi-dimensional streaming requiring **spatiotemporal integration** (3D)
   - **Video object detection** – frames are correlated in time.
   - **Action recognition** – classifying an action across many frames
   - **Video frame prediction** – path prediction & planning
2. Sequence classification and generation in time:
   - **Raw audio classification:** keyword spotting without MFCC preprocessing
   - **Audio denoising:** generate contextual denoising
   - **ASR and GenAI**: compressing LLMs
3. Any other **sequence classification or prediction** algorithms
   - **Healthcare:** vital signs estimation
   - Anything that can be [transformed into a time-series/sequence prediction problem]()

Kinetics400

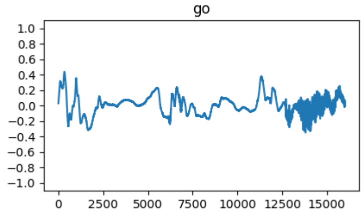
KITT

Sequence classification & generation

BIDMC Vital Signs



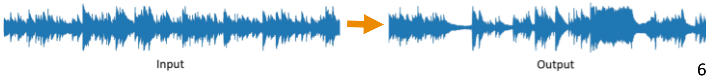
SC10 Raw Audio



Microsoft DNS Challenge

noisy raw speech

denoised speech

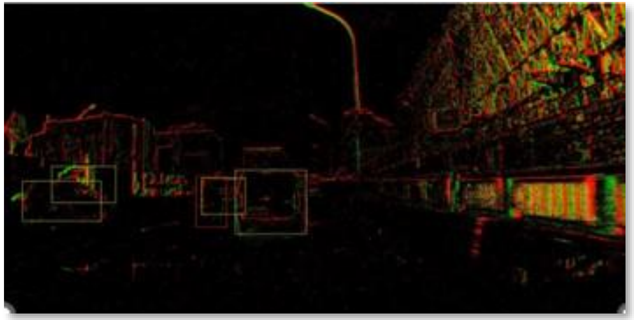


©2024 BrainChip Inc.

Improve Video Object Detection

Event Based Camera Comparison
(vs Gray Retinanet + Prophesee Road Object Dataset*)

Network	mAP (%)	Parameters (millions)	MACs / sec (Billions)
Akida TENN* + CenterNet	56	0.57	94
Resolution 1280 x 720	30% better precision	50x fewer parameters	30x fewer operations



Frame Based Camera Comparison
(vs SimCLR + ResNet50 using Kitti2D Dataset)**

Network	mAP (%)	Parameters (millions)	MACs / sec (Billions)
Akida TENN* + CenterNet	57.6	0.57	18
Resolution 1382 x 512	Equivalent precision	50x fewer parameters	5x fewer operations

< 20 mW

For 30 FPS in 7 nm***

* Gray Retinanet is the latest state of art in event-camera object detection

** SimCLR with a RESNET50 backbone is the benchmark in object detection -- Source: SiMCLR Review

*** Estimates for Akida neural processing scaled from 28 nm

©2024 BrainChip Inc.

TENN Can Be Extended to Spatio-Temporal Data

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

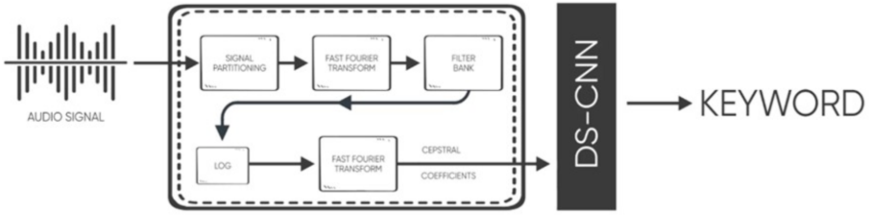
State of the Art

brainchip Essential AI

©2024 BrainChip Inc.

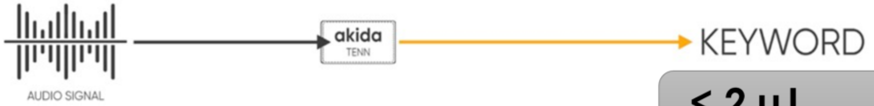
Enhance Raw Audio and Speech Processing

Without akida



Model	Accuracy	Total Memory (KB)	MACs (M/sec)
MFCC+DSCNN	92.43%	93.61	128

With akida



< 2 µJ
Per inference in 28nm

Model	Accuracy	Total Memory (KB)	MACs (M/sec)
Akida TENN*	97.12%	26	19

5% Better accuracy

Lower memory, BOM cost

7x fewer Ops

- No additional filtering or DSP hardware
- Much faster and power-efficient

* Estimates on 28nm process

TENNs also show substantial benefit for audio-denoising (Deep Noise Suppression), and other speech processing

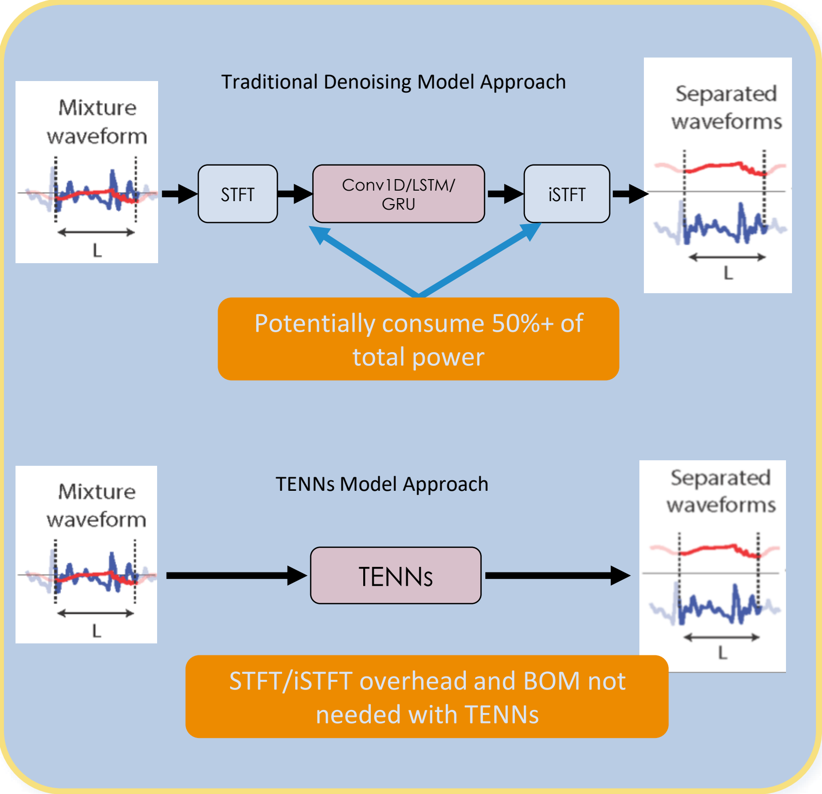
©2024 BrainChip Inc.

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

- **Audio denoising isolates a voice signal obscured by background noise**
- Traditional approach employs computationally intensive time domain to frequency domain transform and the inverse transform
- TENNs approach avoids expensive data transformations



©2024 BrainChip Inc.

TENN vs GPT2

Single thread CPU performance, 11th Gen Intel i7 - 3.00 GHz

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

ACROSS IT

HARRY HAD A SUSPICION SHE HAD BEEN JUST THAT WHEN SHE

> 2100 tokens/minute

< 10 tokens/minute

©2024 BrainChip Inc.
©2024 BrainChip Inc.

Task: Sentence Generation

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)	1.35	4.8	1	2.06	10.4	21.7	5.9
Energy (relative to TENN)	1700	5700	1	2.06	13000	27000	5.9
Training Time (relative to TENN)	~768 GPU hours 21x	~2264 GPU hours 62.8x	35 GPU hours				

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: ~1.5 days on (1) A100 (35 GPU hours)
4. GPT-2 Small training time: 4 days on (8) A100 (768 hours)
5. GPT-2 Medium estimated training time
6. Scores reported as negative entropy: $-log_2(1/VocabSize) - log_2(perplexity)$ (higher better)
7. Input (context) was 1024 tokens

©2024 BrainChip Inc.

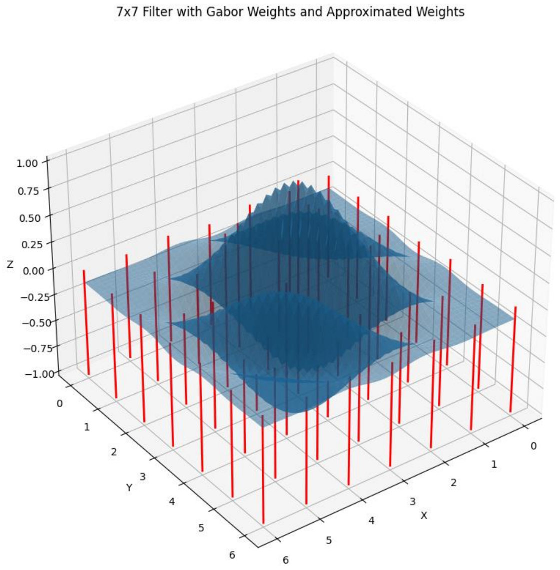
©2024 BrainChip Inc.

Technical Details

©2024 BrainChip Inc.

Learning Continuous Convolution Kernels

- Colored plane represents the continuous kernel we’re trying to learn
- Red arrows represent the individual weights in a 7x7 filter
- A large number of weights requires a large amount of computation
- Results in slow training and large memory bottlenecks

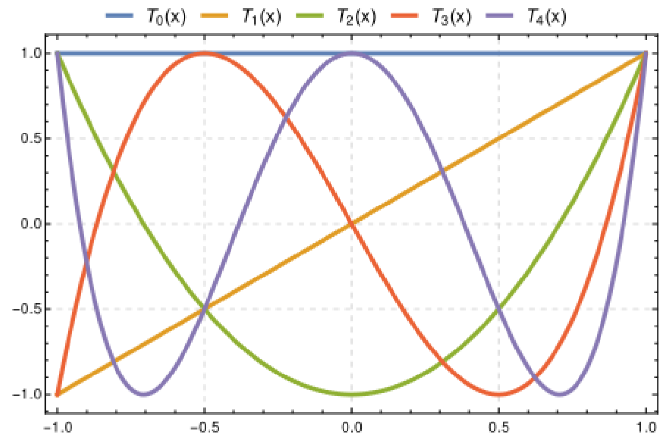


Representing Convolution Kernels with Orthogonal Polynomials

embedded VISION SUMMIT

- TENNs learns the continuous kernel directly through polynomial expansion.
- Learn coefficients for polynomials through backpropagation.
- Training is much faster because the polynomial coefficients (weights) converge independently and do not affect each other due to polynomials being orthogonal to each other.

Chebyshev polynomial

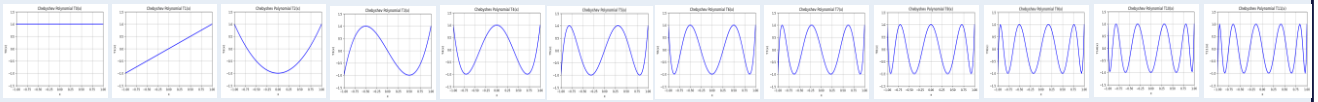


Chebyshev polynomial basis can lead to exponential convergence for a wide range of functions, including those with singularities or discontinuities.*

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

©2024 BrainChip Inc.

Visualizing the Computation

Time (t)	22	23	24	25
Input Buffer $I(t)$				
Polynomials $\boxed{C_{1-12}(\cdot)}$				
Coefficients $\boxed{a_l}$	[0.011, 0.871, 0.235, 0.678, 0.547, 0.298, 0.045, 0.945, 0.478, 0.284, 0.765, 0.199]			
Kernel h	$\mathrm{h}(t-\tau) = \sum_{l=0}^{L} a_l\, C_l(t-\tau)$ $\quad \mathrm{h}(\cdot) = a_1\, C_1(\cdot) + a_2\, C_2(\cdot) + a_3\, C_3(\cdot) + a_4\, C_4(\cdot) + a_5\, C_5(\cdot) + a_6\, C_6(\cdot) + a_n\, C_n(\cdot)$			
Convolution $\boxed{\chi}$	$\chi(t=25) = \sum_{k=22}^{25} \mathrm{h}(25-k)\, I(k) = h(3)\, I(22) + h(2)\, I(23) + h(1)\, I(24) + h(0)\, I(25)$			

Convolution: $\boxed{\chi(t)} = \mathrm{h} * I(t) = \int_{t-D}^{t} \mathrm{h}(t-\tau)\, I(\tau)\, d\tau \approx \sum_{k=22}^{25} \mathrm{h}(t-k)\, I(k)$

Nonlinear Output: $o(t) = f(\boxed{\chi(t)})$ $\quad f(\cdot)$: nonlinear activation function:

brainchip
Essential AI

©2024 BrainChip Inc.

Buffer Mode vs Recurrent Mode

Recurrence: Chebyshev polynomials have a recurrence relationship.
Duality: This particular recurrence imputes duality to buffer mode as well as recurrent mode.

Buffer (Convolutional) Mode
Overview
Buffering inputs over time
Benefit
Speed up training by reading the memory buffer in parallel
Training stability improved by orthogonality
Drawbacks
Higher memory usage

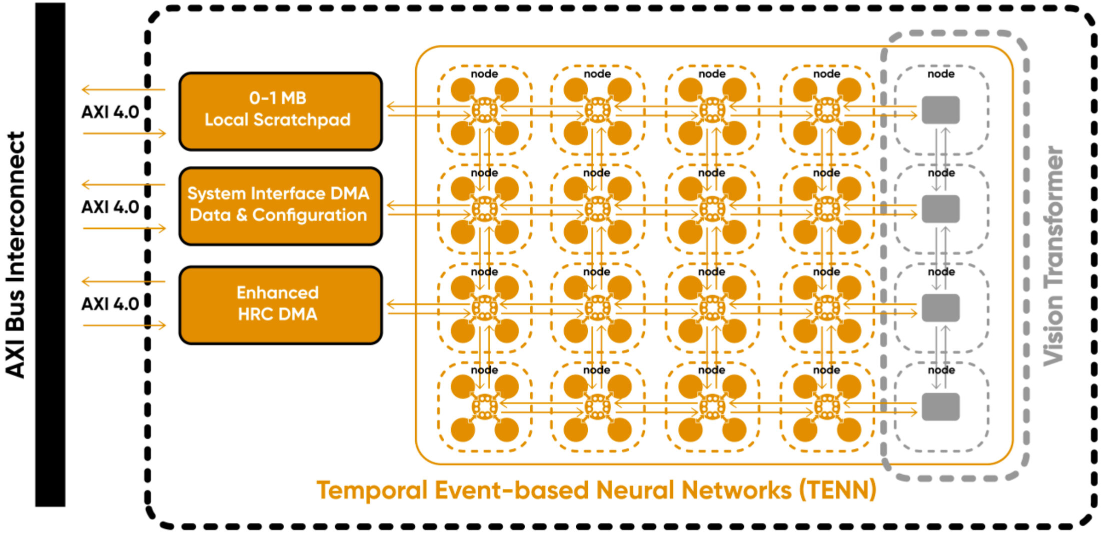
Recurrent Mode
Overview
Update previous state over time
Benefit
Save memory by generating polynomials recurrently, timestep-by-timestep
Lower memory usage benefits inference
Drawback
Training has to be done sequentially

©2024 BrainChip Inc.

Getting It to Market

©2024 BrainChip Inc.

Hardware IP to Run TENNs on the Edge



Key Hardware Features

- Digital, event-based, at memory compute
- Highly scalable
- Each node connected by mesh network
- Inside each node is an event-based TENN processing unit

©2024 BrainChip Inc.

Brainchip's Differentiation: Akida Technology Foundations

embedded VISION SUMMIT

Fundamentally different. Extremely efficient.

Silicon-Proven, Fully Digital Neuromorphic Implementation
Cost-effective, predictable design and implementation

Event-based Hardware Acceleration
Minimized compute and communication - Minimizes host CPU usage

At-Memory-Compute
Maximum throughput, Lowers latency and system bandwidth usage

On-chip Learning
One-shot/few-shot learning. Minimizes sensitive data sent. Improves security and privacy

Configurable And Scalable
Extremely configurable and post-silicon flexibility

Complex Models, High Accuracy
Unique spatial-temporal capabilities, accelerates Vision Transformers.

©2024 BrainChip Inc.

BrainChip Resources

Visit Us @ Booth #618

TENNs Paper “Building Temporal Kernels with Orthogonal Polynomials
https://bit.ly/brainchip_tenns

TENNs White Paper
https://brainchip.com/temporal-event-based-neural-networks-a-new-approach-to-temporal-processing/

Akida 2nd Generation
https://brainchip.com/wp-content/uploads/2023/03/BrainChip_second_generation_Platform_Brief.pdf

BrainChip Enablement Platforms
https://brainchip.com/akida-enablement-platforms/

©2024 BrainChip Inc.

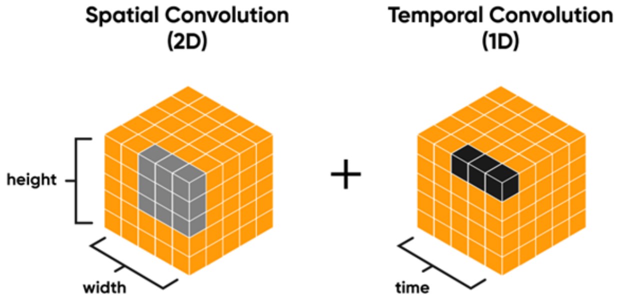
Backup Slides

©2024 BrainChip Inc.

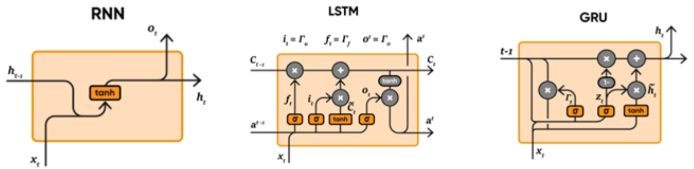
Improve Efficiency Without Compromising Accuracy

Temporal Event Based Neural Nets (TENNs)

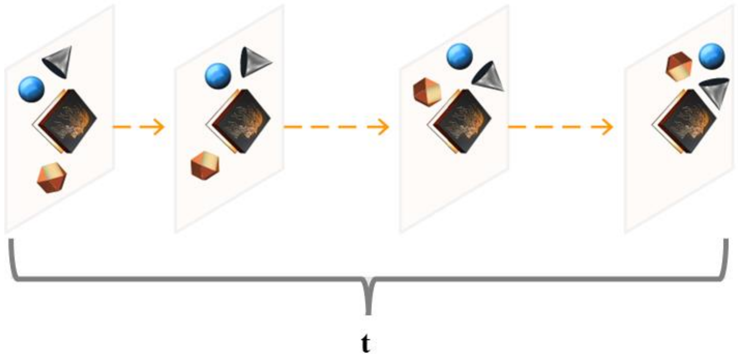
Extremely efficient 3D convolutions



TENNs deliver the benefits of and are much more efficient to train than RNNs



3D Time Series



- Simplifies solution to complex problems
- Reduces model size and footprint without loss in accuracy
- Easy to train (CNN-like pipeline)
- Supports longer range dependencies than RNNs

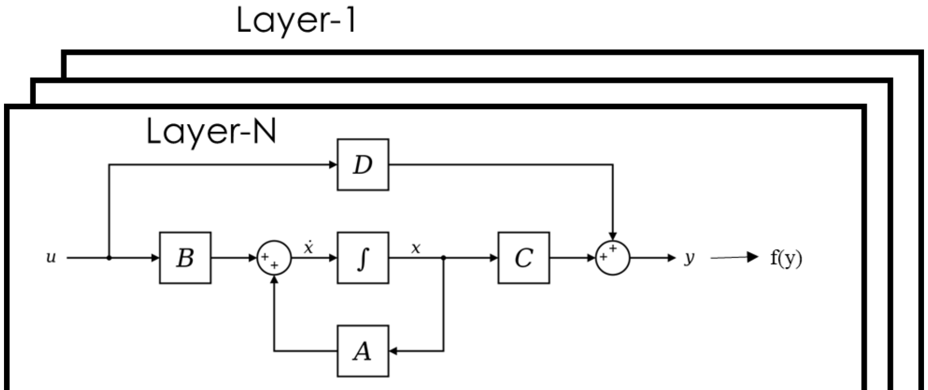
©2024 BrainChip Inc.

TENN Has Two Modes: Buffer and Recurrent Modes

Principles:

1. Recurrence: Chebyshev and Legendre polynomials have recurrence relationship.
2. Duality: Recurrence imputes duality: Buffer mode as well as recurrent mode.
3. Stable training: Train in buffer mode
4. Fast Running: Run in recurrent mode. 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.

Recurrent Mode



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

©2024 BrainChip Inc.

TENN Has Two Modes: Buffer and Recurrent Modes

embedded VISION SUMMIT

Buffer mode:

kernel $h(t) = \sum_{l=0}^{L} \boxed{a_l} \boxed{C_l(t)}$

convolution $\boxed{\chi} = h * I(t)$

buffer for h(t) & buffer for I(t)

convolution: dot product over 2 buffers

$$\boxed{\chi} = \tilde{\boldsymbol{h}} \cdot \boldsymbol{I} = \sum_k \tilde{\boldsymbol{h}}_k I_k$$

Recurrent mode:

kernel $h(t) = \sum_{l=0}^{L} \boxed{a_l} \boxed{C_l(t)}$

L convolutions over polynomials $\boxed{\chi_l} = \boxed{C_l} * I(t)$

kernel convolution $\boxed{\chi} = \sum_{l=0}^{L} \boxed{a_l} \boxed{\chi_l}$

Buffer mode for fast parallel training:

Entire kernel is stored in a memory buffer accessible at once

Convolution is computed in conventional way

Recurrent mode saves memory :

Polynomials generated recurrently, timestep by timestep & not stored in memory

Convolution of input over *L* polynomials computed timestep by timestep, accumulated over time; *L* separate convolutions

Kernel convolution is *L* polynomial convolutions weighted by the polynomial coefficients & summed

brainchip Essential AI

©2024 BrainChip Inc.