类脑计算

百倍能效的下一代人工智能

2026.8

中山大学理学院

华为2012实验室—前沿技术与创新中心

- 源于《2012》同名电影
- 打造数据洪流中的诺亚方舟
- 构建公司未来的生存能力

中央研究院

引领5-10年前沿技术

先进计算与存储实验室 (ACS Lab)：
量子计算，类脑计算，光计算

2019-2026

循忆星海（上海）科技有限公司

- 类脑算法，端侧智能解决方案
- 存算一体忆阻器AI芯片
- 智能穿戴、具身智能、脑机接口、卫星天基AI算力

2026-

目录

背景：当前人工智能面临问题

传统AI在具身智能场景能力不足、大模型场景能耗问题凸显，面临瓶颈

- AI在算力中心、智能消费电子、具身智能、柔性制造等领域中发挥着关键作用。然而，其**能力**和**能效**的不足已成为制约这些领域发展的瓶颈。

能力

海量人工筛选数据+离线训练

训练GROK3: 12.8 万亿 token

能效

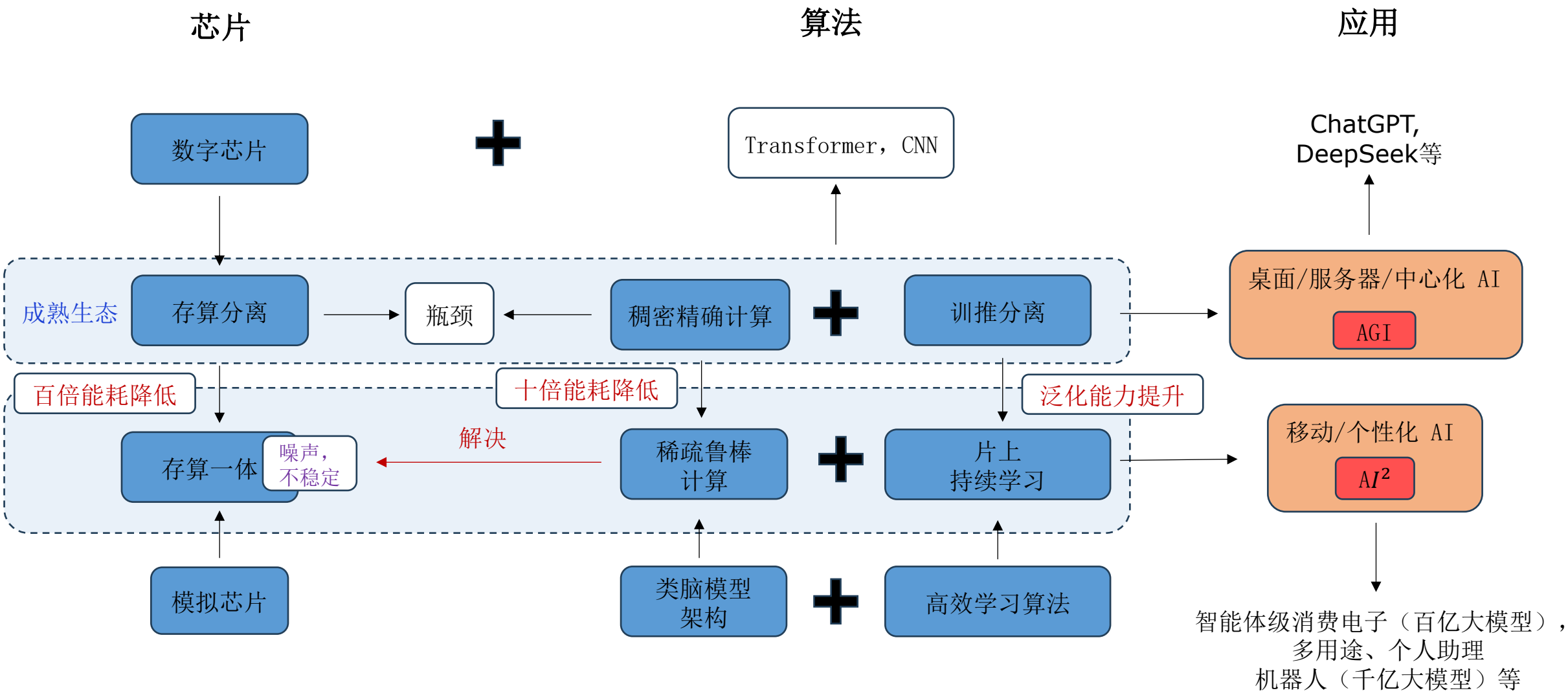
Nvidia A100: 1.5TOPS/W

人脑: ~10000 TOPS/W

- **算法能力受限**-训推分离范式：海量数据->巨量算力离线训练->部署后在线推理。难以适应新任务、新环境，泛化性差。
- **芯片能效不足**-存算分离架构：片上数据搬运的时间和功耗是计算的百倍以上。当前主流AI芯片能效比人脑低3~4个数量级

下一代AI要求以更高能效实现高级智能，类脑智能是潜在解决路径

类脑智能：类脑算法+存算一体芯片，实现百倍能效的下一代AI



目录

类脑算法，如何类脑？

形式类脑：基于大脑中已发现的机制

如脉冲神经网络，局部学习等

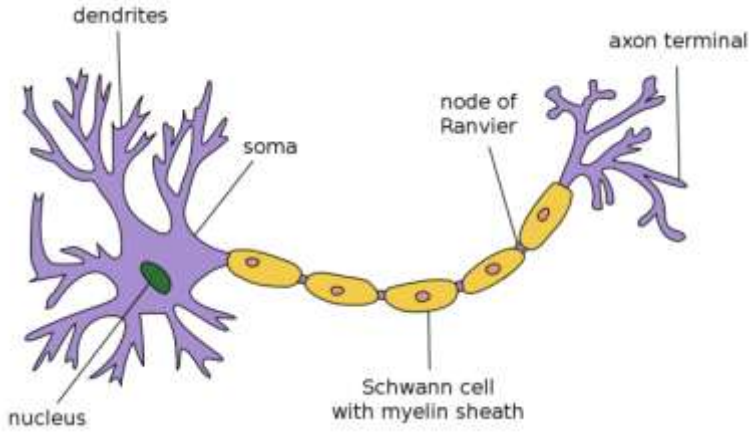
功能类脑：从已知的大脑功能出发

如强化学习，自监督学习，预测编码，记忆机制等

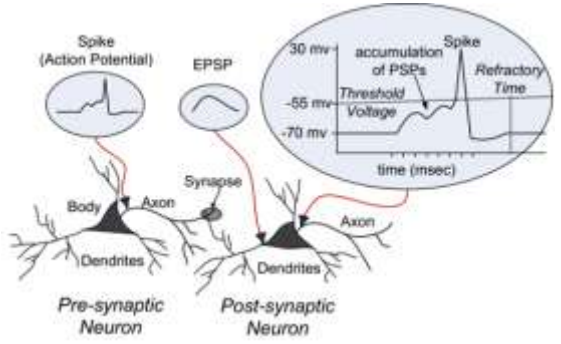
必要性反思，“飞机不需要扇动翅膀”！

生物神经元与数学模型

一个神经元在树突（dendrite）接受输入脉冲，导致细胞体(soma)的膜电压升高，当达到特定阈值时，发出一个输出脉冲到轴突(axon)，并通过突触发送神经递质与后续神经元树突上的接受体相结合来传递信号。

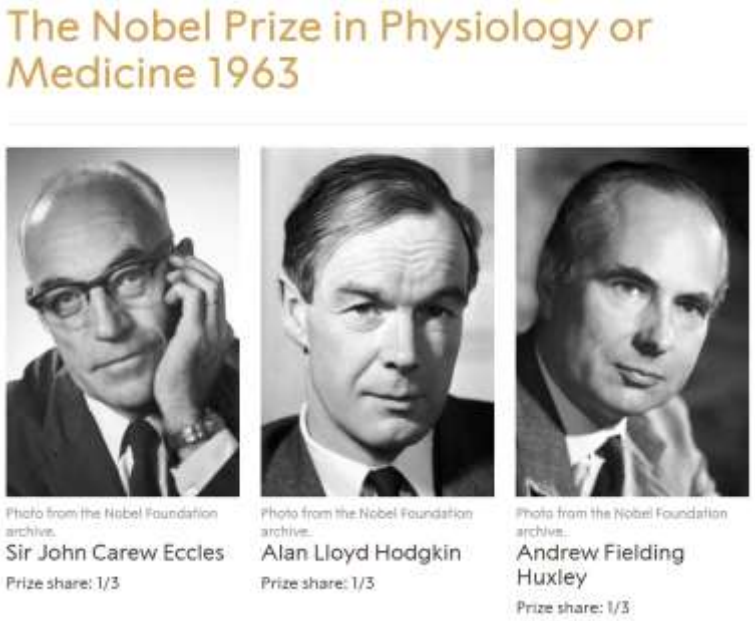


M. Petrovici, 2016. Modified from http://en.wikipedia.org/wiki/File:Neuron_Hand-tuned.svg.



Hodgkin-Huxley (HH) Neuron Model

A. Hodgkin and A. Huxley the 1963 Nobel Prize.



$$\langle I_{\mathrm{Na}^+} \rangle = g_{\mathrm{Na}^+} m^3 h (u - E_{\mathrm{Na}^+})$$
$$\langle I_{\mathrm{K}^+} \rangle = g_{\mathrm{K}^+} n^4 (u - E_{\mathrm{K}^+}).$$

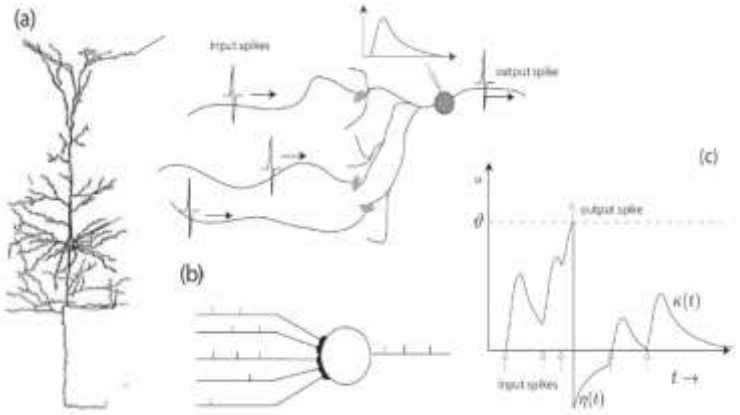
$$\dot{x} = -\frac{1}{\tau_x(u)}[x - x_0(u)], \quad x \in \{m, h, n\}$$

$$C_{\mathrm{m}} \dot{u} = -g_{\mathrm{Na}^+} m^3 h (u - E_{\mathrm{Na}^+}) - g_{\mathrm{K}^+} n^4 (u - E_K) - g_{\mathrm{l}} (u - E_{\mathrm{l}}) + I$$

树突神经元模型：精细但复杂

脉冲神经网络（Spiking neuron network, SNN)

点神经元模型：简单但高效



Abstract Neuron Model

Leaky-integrate and fire (LIF) model

$$C_{\mathrm{m}}\frac{du}{dt} = g_{\mathrm{l}}(E_{\mathrm{l}} - u) + I$$

Spike response model (SRM)

$$u(t) = \sum w_i \, (\varepsilon * s_i)(t) + (\nu * s)(t)$$

Discretized model

$$u_t' = \tau u_{t-1} + I_t$$
$$s_t = H(u_t' - v_{\mathrm{th}})$$
$$\text{Soft reset}: \; u_t = u_t' - s_t v_{\mathrm{th}}$$
$$\text{Hard reset}: \; u_t = u_t'(1 - s_t) + s_t u_r$$

事件驱动神经动力学：
神经元通过累积脉冲，改变膜电位，当膜电位超过阈值，神经元会发放脉冲，并且膜电位降低到静息状态。

- 只有脉冲发放时才会进行下一步计算（低能耗）
- 每次更新只针对特定的神经元，是一种异步计算方式（低时延）

这种事件驱动的计算方式，可以极大降低能耗和时延。

- 神经元内部，由树突接受信号，经过复杂动力学处理后汇总到胞体，再由轴突发出信号。

加法计算特性：
神经元基于脉冲通信，每次接收脉冲的过程是神经元膜电压进行一次加法的过程

- 将乘加计算转化为加法计算，大幅降低能耗
- 利用脉冲发放时间表征信息，保证网络的表达能力

这种基于加法计算进行信息传递的方式，在保证网络精度的同时，大幅降低网络的计算能耗

SNN特点：时空动力学，稀疏加法计算

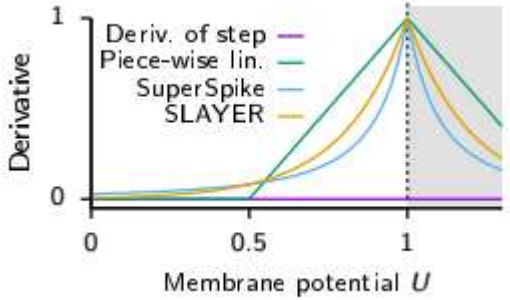
深度脉冲神经网络的训练

基于代理梯度（Surrogate gradient）的直接训练

$$u'_t = \tau u_{t-1} + I_t$$
$$s_t = H(u'_t - v_{\text{th}})$$
$$\text{Soft reset}: \ u_t = u'_t - s_t v_{\text{th}}$$
$$\text{Hard reset}: \ u_t = u'_t(1 - s_t) + s_t u_r$$

$$u^{t,n} = \tau u^{t-1,n}(1 - y^{t-1,n}) + I^{t,n}$$

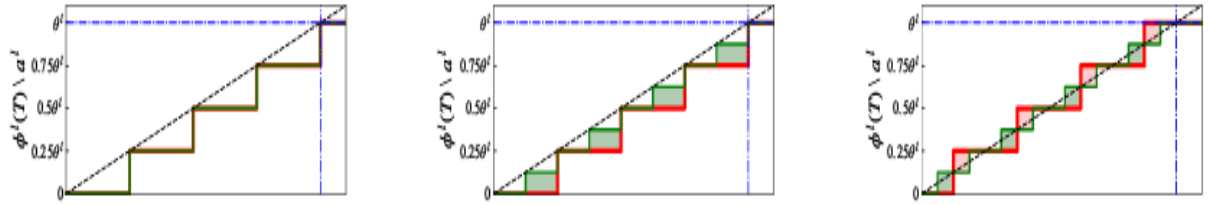
$$\frac{\partial L}{\partial w} = \sum_t \frac{\partial L}{\partial y^t}\frac{\partial y^t}{\partial u^t}\frac{\partial u^t}{\partial I^t}\frac{\partial I^t}{\partial w}$$



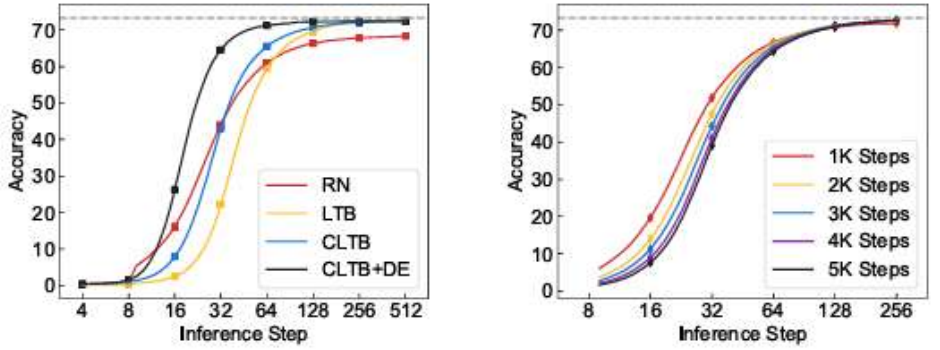
平滑连续函数逼近真实梯度

Surrogate gradient Learning, E. Neftci, 2019.
Slayer, Sumit, 2019.
Superspike, F. Zenke. 2017.
STBP, Y. Wu, 2017

Training free, ANN转SNN



基于脉冲发放率逼近ANN激活函数



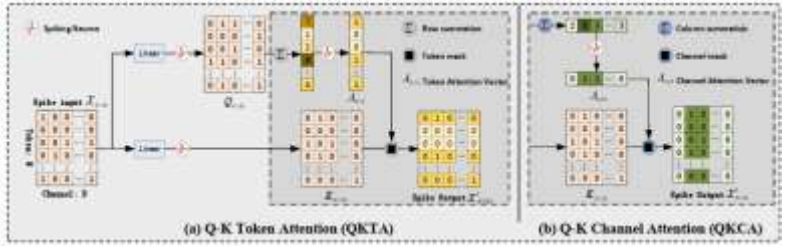
脉冲发放窗口越长，逼近误差越小，精度越高

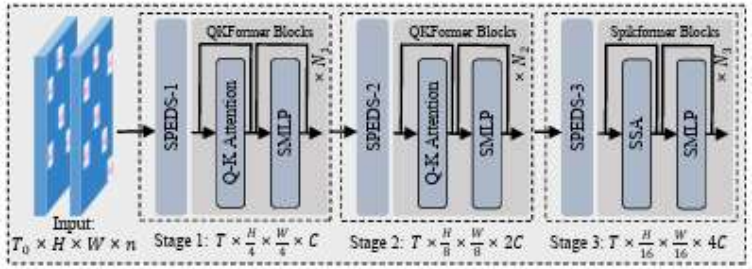
Tong Bu, et.al, ICLR 2022
Zecheng Hao, et.al, ICLR 2023
Tong Bu, et.al, CVPR, 2025

深度脉冲神经网络最新进展

QKFormer [1]（图像任务）
鹏城实验室&北京大学

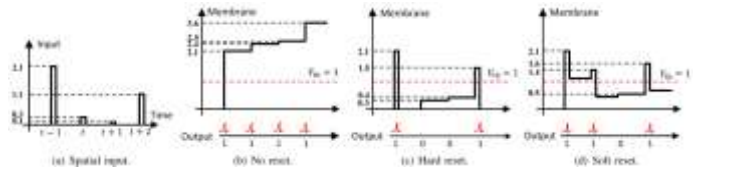
- 提出QKFormer实现线性复杂度的脉冲注意力机制，显著降低注意力机制能耗。基于多尺度脉冲表征和脉冲块编码方法（SPEDS），首次实现SNN在ImageNet上Top-1 85%精度。

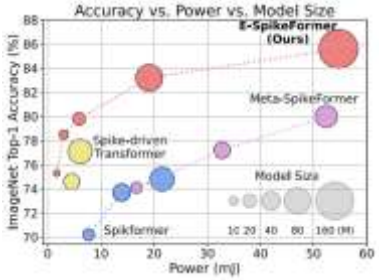


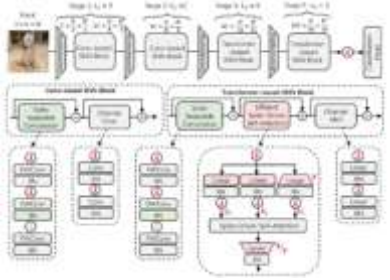


E-SpikeFormer [2]（图像任务）
中科院自动化所&华为ACSLab

- 提出脉冲发放逼近（SFA）方法，显著降低脉冲神经网络的训练存储开销和推理时延。在ImageNet数据集上实现173M Spiking Transformer，达到86.2%Top-1精度，相对同等精度ANN Transformer计算量降低一半。

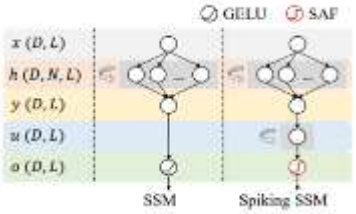






SpikingSSM [3] (长序列任务)
华为ACSLab

- 提出支持SNN长序列并行计算的代理动力学网络（SDN）。相比BPTT训练方法在8K序列长度提速100倍。实现90%全网稀疏度，相比原SSM计算量低18倍，精度误差1%。在WikiText-103数据集任务（1亿token）以1/3模型参数量 (70M) 精度超过SpikeGPT（200M）。



Model	SNN	WikiText-103	Parameters
Transformer	No	20.51	231M
S4	No	20.95	249M
SpikeGPT	Yes	39.75	213M
SpikingSSM	Yes	33.94	75M

Table 4: Performance comparison of SpikingSSMs with previous works on WikiText-103 dataset.

Model	Accuracy (%)	Spiking Rate (%)	Speed (ms)
LIF	85.45	12.08	1480
SDN	85.57	11.92	230
SDN-S	81.52	18.30	285

Table 5: Performance comparison on the sCIFAR10 dataset.

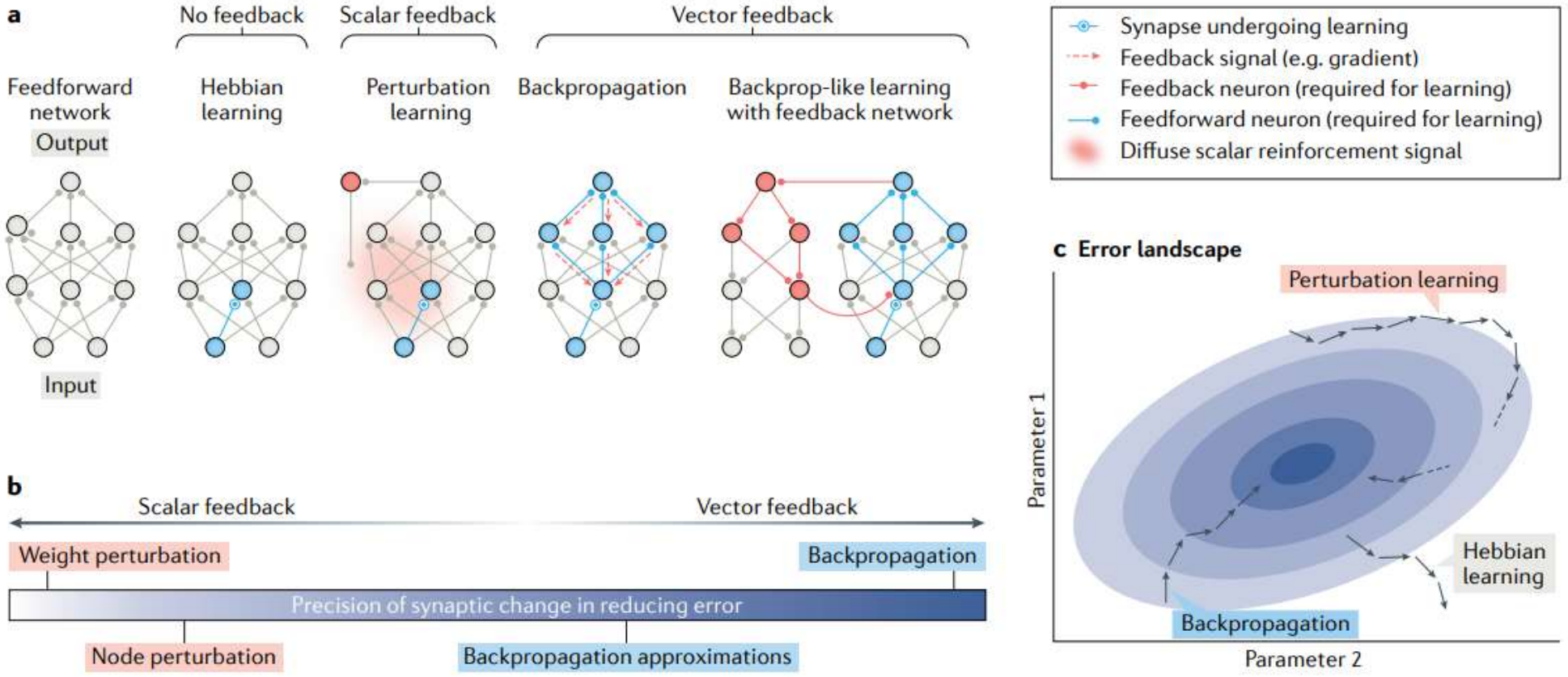
Model	SNN	LISTOPS	TEXT	RETRIEVAL	IMAGE	PATHFINDER	Path-X	AVG
Transformer	No	36.37	64.27	57.46	42.44	71.40	—	53.66
LMUFormer	No	34.43	68.27	78.65	54.16	69.9	—	59.24
S4D-Lin	No	**60.52**	**86.97**	**90.96**	**87.93**	**93.96**	**92.80***	**85.52**
Spiking LMUFormer	Yes	37.30	65.80	79.76	55.65	72.68	—	60.2
Binary S4D	Yes	54.8	**82.5**	85.03	82	82.6	61.2	74.69
S6-based SNN	Yes	55.70	77.62	88.48	80.10	83.41	—	72.55
SpikingSSM-VF	Yes	59.93	82.35	88.2	86.81	**93.68**	94.80	84.30
(spiking rate)		(0.13)	(0.1)	(0.06)	(0.22)	(0.07)	(0.1)	(0.11)
SpikingSSM-VT	Yes	**60.23**	80.41	**88.77**	**88.23**	93.31	**94.82**	**84.33**
(spiking rate)		(0.14)	(0.06)	(0.06)	(0.15)	(0.08)	(0.1)	(0.08)

Table 3: Performance comparison of SpikingSSM and previous works on the LRA dataset. *Since the original S4D-Lin failed in the Path-X task, we instead present the result of another close variant S4D-lin. -VF and -VT denote fixed and trainable threshold, respectively. Furthermore, we take the 50% accuracy for the absence of Path-X accuracy as did in the work of S4D, then compute the overall average metrics across all tasks as AVG. The spiking rate for each task have also been calculated, which is indicated by blue font.

[1] QKFormer: Hierarchical Spiking Transformer using Q-K Attention, NeurIPS (Spotlight) 2024
[2] Scaling Spike-Driven Transformer With Efficient Spike Firing Approximation Training, IEEE TPAMI 2025
[3] SpikingSSMs: Learning Long Sequences with Sparse and Parallel Spiking State Space Models, AAAI (Oral) 2025

类脑学习：反向传播算法之外的可能

反向传播算法作为当前深度学习模型主流学习算法，有权重搬运、前向反向更新锁定等约束，对存储、数据搬运、算力有很高的要求，导致大模型训练计算量大，速度受限。受大脑启发的局部学习算法具备计算、存储开销小等优势。



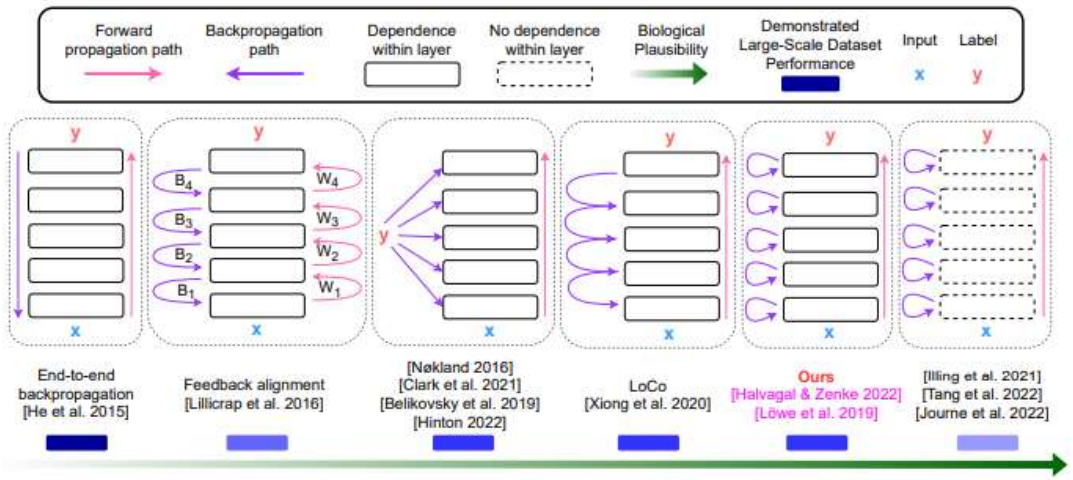
深度神经网络的学习算法

Backpropagation and the brain, Lillicrap T.P, Hinton G. et.al, 2019 Nature review neuroscience

局部学习，前向学习：高效计算实现端侧在线学习

反向传播算法作为当前深度学习模型主流学习算法，有权重搬运、前向反向更新锁定等约束，对存储、数据搬运、算力有很高的要求，导致大模型训练计算量大，速度受限。受大脑启发的局部学习算法[1-6]具备计算、存储开销小等优势。

Learning from scratch



全局反向传播算法的相关约束与局部学习的逐步替代

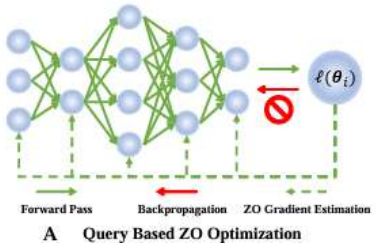
局部、前向学习优势：学习时延小，计算存储低，端侧友好

挑战：可拓展性，收敛速度，如何构建监督信号?

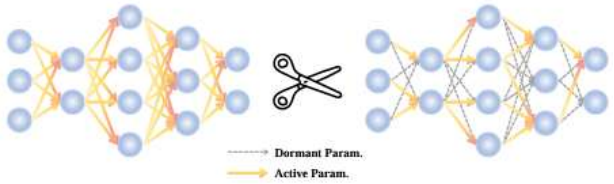
[1] Hebbian Deep Learning Without Feedback, ICLR, 2023 (Spotlight)
[2] Unsupervised learning by competing hidden units, PNAS, 2019
[3] Blockwise Self-Supervised Learning at Scale, Yann LeCun, et.al. 2023

Fine tuning of LLM with ZO

$$\hat{\nabla} f(\mathbf{x}) = \frac{1}{q}\sum_{i=1}^{q}\left[\frac{f(\mathbf{x}+\mu\mathbf{u}_i) - f(\mathbf{x}-\mu\mathbf{u}_i)}{2\mu}\mathbf{u}_i\right]$$



A Query Based ZO Optimization



B Model Pruning Guided Sparsity

SST2	Roberta-Large				OPT-1.3B			
	FT	LoRA	Prefix	Prompt	FT	LoRA	Prefix	Prompt
FO-SGD	91.4	91.2	89.6	90.3	91.1	93.6	93.1	92.8
Forward-Grad	**90.1**	89.7	89.5	87.3	90.3	90.3	90.0	82.4
ZO-SGD	89.4	90.8	90.0	87.6	**90.8**	90.1	**91.4**	84.4
ZO-SGD-MMT	89.6	90.9	90.1	88.6	85.2	91.3	91.2	**86.9**
ZO-SGD-Cons	89.6	**91.6**	90.1	88.5	88.3	90.5	81.8	84.7
ZO-SGD-Sign	52.5	90.2	53.6	86.1	87.2	91.5	89.5	72.9
ZO-Adam	89.8	89.5	**90.2**	**88.8**	84.4	**92.3**	**91.4**	75.7

[4] DEEPZERO: SCALING UP ZEROTH-ORDER OPTIMIZATION FOR DEEP MODEL TRAINING, ICLR, 2024
[5] Fine-Tuning Language Models with Just Forward Passes, 2023
[6] Revisiting Zeroth-Order Optimization for Memory-Efficient LLM Fine-Tuning: A Benchmark, 2024

华为类脑算法团队工作

团队已在NeurIPS，ICML，AAAI，ICLR，CVPR，TPAMI，TNNLS等顶会顶刊发表类脑计算领域论文数十篇，并获得NeurIPS 2022、ICLR 2023 Spotlight以及AAAI 2025 Oral。专利6项，成果转化4项，技术成果已应用于相关工业场景。

-SpikingSSM: 稀疏脉冲线性注意力网络，在16K长序列任务精度超过Transformer，计算量相比降低20倍（2025 AAAI (Oral) 2025），相关技术已落入HW产线，在无线信道预测场景计算量相比原方案降低120倍。

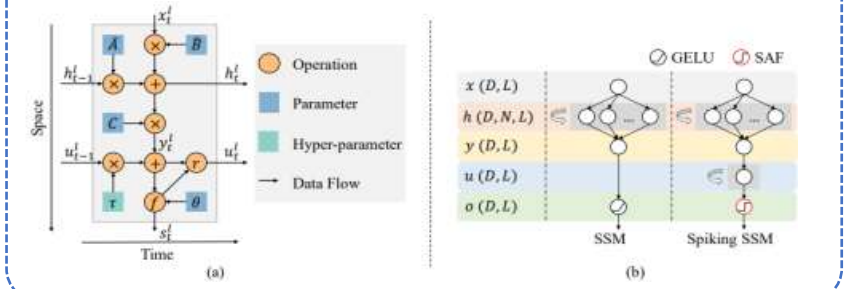
-SpikingTransformer: 基于视觉任务的脉冲注意力网络，在ImageNet达到86.2%Top-1精度，相对同精度ANN Transformer计算量降低一半。（2025 IEEE TPAMI）

-SpikingMamba: 业界首个RNN架构线性注意力脉冲大语言模型（>1B参数量），相比同参数量ANN大模型同等精度下计算量降低四倍。落地公司端侧大模型等项目。

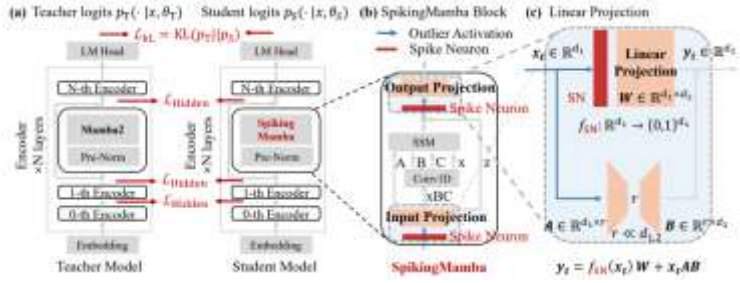
-DeepHebb: 基于赫布法则的无监督局部学习深层卷积网络，实现前向推理学习，精度达到业界最优（2023 ICLR (Spotlight)），相关技术入选HW2024年报亮点工作。

-ZO-LoRA: 基于零阶优化的大模型局部学习算法，在7-13B大模型精度达到业界最优，逼近全局反向传播算法，训练存储量降低10倍。

SpikingSSM



SpikingMamba



SpikingSSM: 可并行计算的树突动力学SNN

针对脉冲神经元串行计算与大模型并行计算不兼容的问题，开发可替代LIF动力学的并行计算代理动力学网络（SDN），实现SNN在长序列任务的并行计算。相比传统BPTT训练方法在8K序列长度提速100倍。稀疏度比原SSM提高10倍（90%全网稀疏度），同时在16K长度Path-X任务精度提高2%。在WikiText-103数据集任务（1亿token）以1/3模型参数量(70M)精度超过SpikeGPT（200M），相比原SSM理论计算能耗降低18倍。

Approximation of SDN

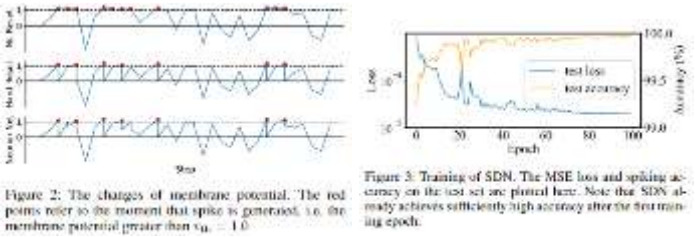
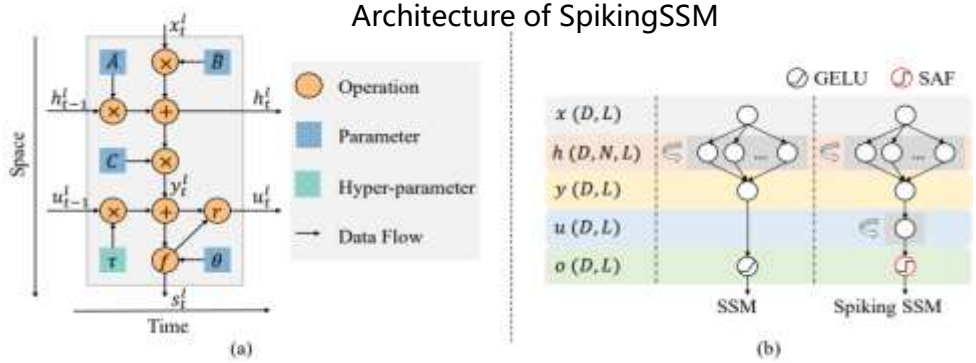


Figure 2: The changes of membrane potential. The red points refer to the moment that spike is generated, i.e. the membrane potential greater than $v_{th} = 1.0$

Figure 3: Training of SDN. The MSE loss and spiking accuracy on the test set are plotted here. Note that SDN already achieves sufficiently high accuracy after the first training epoch.

Architecture of SpikingSSM



SSM

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t$$
$$y_t = Ch_t$$
$$y_t = \sum_{k=1}^{t} C\bar{A}^{t-k}\bar{B}x_k$$

LIF neuron

$$u'_t = \tau u_{t-1} + I_t$$
$$s_t = H(u'_t - v_{\text{th}})$$
$$\text{Soft reset}: \ u_t = u'_t - s_t v_{\text{th}}$$

SpikingSSM

$$u'_t = \tau u_{t-1} + y_t$$
$$s_t = H(u'_t - v_{\text{th}})$$

Model	SNN	LISTOPS	TEXT	RETRIEVAL	IMAGE	PATHFINDER	Path-X	AVG
Transformer	No	36.37	64.27	57.46	42.44	71.40	—	53.66
LMUFormer	No	34.43	68.27	78.65	54.16	69.9	—	59.24
S4D-Lin	No	**60.52**	**86.97**	**90.96**	**87.93**	**93.96**	**92.80***	**85.52**
Spiking LMUFormer	Yes	37.30	65.80	79.76	55.65	72.68	—	60.2
Binary S4D	Yes	54.8	**82.5**	85.03	82	82.6	61.2	74.69
S6-based SNN	Yes	55.70	77.62	88.48	80.10	83.41	—	72.55
SpikingSSM-VF	Yes	59.93	82.35	88.2	86.81	**93.68**	94.80	84.30
(spiking rate)		(0.13)	(0.1)	(0.06)	(0.22)	(0.07)	(0.1)	(0.11)
SpikingSSM-VT	Yes	**60.23**	80.41	**88.77**	**88.21**	93.51	**94.82**	**84.33**
(spiking rate)		(0.14)	(0.06)	(0.06)	(0.15)	(0.08)	(0.1)	(0.08)

Table 3: Performance comparison of SpikingSSM and previous works on the LRA dataset. *Since the original S4D-Lin failed in the Path-X task, we instead present the result of another close variant S4D-Inv. -VF and -VT denote fixed and trainable threshold, respectively. Furthermore, we take the 50% accuracy for the absence of Path-X accuracy as did in the work of S4D, then compute the overall average metrics across all tasks as AVG. The spiking rate for each task have also been calculated, which is indicated by blue font.

Model	SNN	WikiText-103	Parameters
Transformer	No	20.51	231M
S4	No	20.95	249M
SpikeGPT	Yes	39.75	213M
SpikingSSM	Yes	33.94	75M

Table 4: Performance comparison of SpikingSSMs with previous works on WikiText-103 dataset.

Model	Accuracy (%)	Spiking Rate (%)	Speed (ms)
LIF	85.45	12.08	1480
SDN	85.57	11.92	230
SDN-S	81.52	18.30	285

Table 5: Performance comparison on the sCIFAR10 dataset.

Method	Speed (ms)			
	$L = 1K$	$L = 2K$	$L = 4K$	$L = 8K$
BPTT	1370	2900	8040	25600
SLTT	1210	2720	7740	25600
Ours	183	196	200	253
Ratio	7.5×	15.0×	40.2×	101.2×

Table 1: Comparison on training speed of different methods. The input has a batch size of 64. Training with SDN achieves significant acceleration, the speed up ratio amplifies with increasing sequence length.

SpikingSSMs: Learning Long Sequences with Sparse and Parallel Spiking State Space Models, accepted by AAAI (Oral) 2025

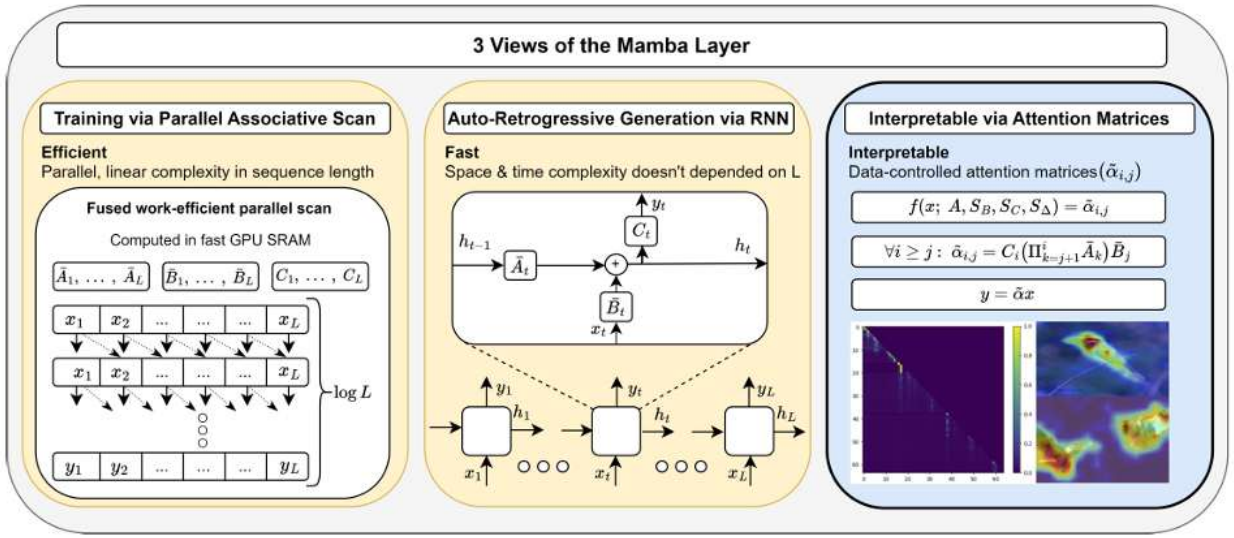
线性注意力RNN: Mamba，Transformer alternative?

Algorithm 1 SSM (S4)

Input: $x : (\mathtt{B}, \mathtt{L}, \mathtt{D})$
Output: $y : (\mathtt{B}, \mathtt{L}, \mathtt{D})$
$\boldsymbol{A} : (\mathtt{D}, \mathtt{N}) \leftarrow$ Parameter
▷ Represents structured $N \times N$ matrix
$\boldsymbol{B} : (\mathtt{D}, \mathtt{N}) \leftarrow$ Parameter
$\boldsymbol{C} : (\mathtt{D}, \mathtt{N}) \leftarrow$ Parameter
$\Delta : (\mathtt{D}) \leftarrow \tau_\Delta(\text{Parameter})$
$\overline{\boldsymbol{A}}, \overline{\boldsymbol{B}} : (\mathtt{D}, \mathtt{N}) \leftarrow \text{discretize}(\Delta, \boldsymbol{A}, \boldsymbol{B})$
$y \leftarrow \text{SSM}(\overline{\boldsymbol{A}}, \overline{\boldsymbol{B}}, \boldsymbol{C})(x)$
▷ Time-invariant: recurrence or convolution
return y

Algorithm 2 SSM + Selection (S6)

Input: $x : (\mathtt{B}, \mathtt{L}, \mathtt{D})$
Output: $y : (\mathtt{B}, \mathtt{L}, \mathtt{D})$
$\boldsymbol{A} : (\mathtt{D}, \mathtt{N}) \leftarrow$ Parameter
▷ Represents structured $N \times N$ matrix
$\boldsymbol{B} : (\mathtt{B}, \mathtt{L}, \mathtt{N}) \leftarrow s_B(x)$
$\boldsymbol{C} : (\mathtt{B}, \mathtt{L}, \mathtt{N}) \leftarrow s_C(x)$
$\Delta : (\mathtt{B}, \mathtt{L}, \mathtt{D}) \leftarrow \tau_\Delta(\text{Parameter} + s_\Delta(x))$
$\overline{\boldsymbol{A}}, \overline{\boldsymbol{B}} : (\mathtt{B}, \mathtt{L}, \mathtt{D}, \mathtt{N}) \leftarrow \text{discretize}(\Delta, \boldsymbol{A}, \boldsymbol{B})$
$y \leftarrow \text{SSM}(\overline{\boldsymbol{A}}, \overline{\boldsymbol{B}}, \boldsymbol{C})(x)$
▷ Time-varying: recurrence (*scan*) only
return y



$$\bar{A} = f_A(A, \Delta), \quad \bar{B} = f_B(A, B, \Delta), \quad h_t = \bar{A}h_{t-1} + \bar{B}x_t, \quad y_t = Ch_t$$

$$B_i = S_B(\hat{x}_i), \quad C_i = S_C(\hat{x}_i), \quad \Delta_i = \text{softplus}(S_\Delta(\hat{x}_i))$$

$$f_A(\Delta_i, A) = \exp(\Delta_i A), \quad f_B(\Delta_i, A, B_i) = \Delta_i B_i,$$

$$\bar{A}_i = f_A(\Delta_i, A), \quad \bar{B}_i = f_B(\Delta_i, A, B_i)$$

$$y = \tilde{\alpha}x, \quad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_L \end{bmatrix} = \begin{bmatrix} C_1\bar{B}_1 & 0 & \cdots & 0 \\ C_2\bar{A}_2\bar{B}_1 & C_2\bar{B}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ C_L\Pi_{k=2}^{L}\bar{A}_k\bar{B}_1 & C_L\Pi_{k=3}^{L}\bar{A}_k\bar{B}_2 & \cdots & C_L\bar{B}_L \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_L \end{bmatrix}$$

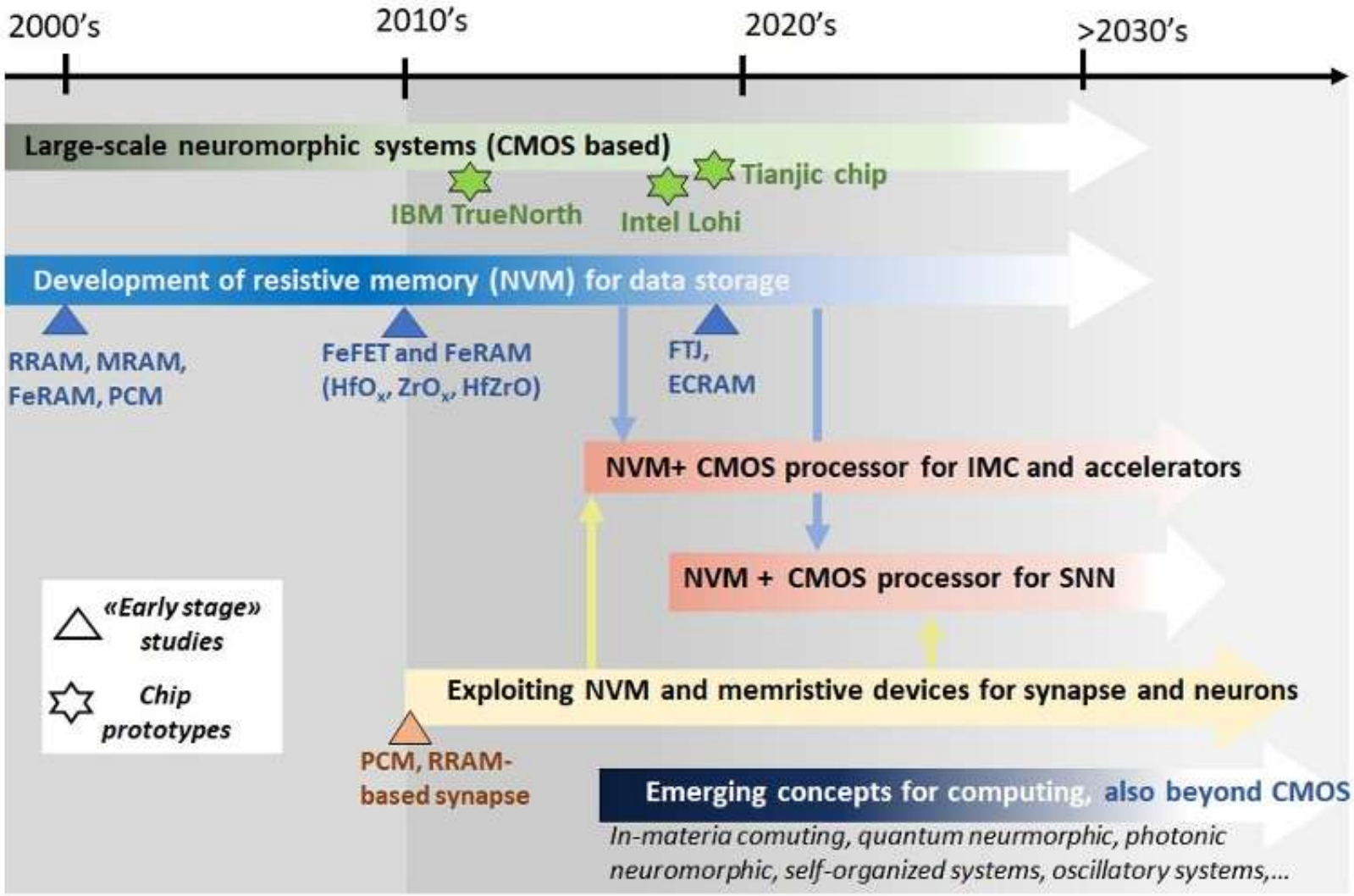
$$\tilde{\alpha}_{i,j} = C_i\left(\Pi_{k=j+1}^{i}\bar{A}_k\right)\bar{B}_j$$

基于有限容量记忆的注意力机制

Mamba: Linear-Time Sequence Modeling with Selective State Spaces, 2023
The hidden attention of mamba models, 2024

目录

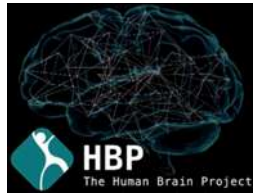
类脑芯片技术发展



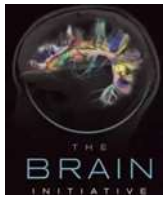
Roadmap to Neuromorphic Computing with Emerging Technologies, 2024

神经形态芯片(Neurormorphic Chips) – 模拟与数字ASIC芯片

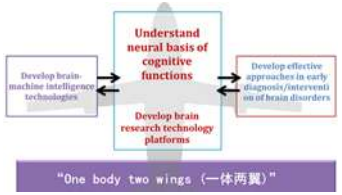
下一代人工智能



欧洲 $10.5亿



美国 $30亿



中国 $70亿

类脑技术在全世界范围已成为重大战略投资项目

大脑

神经动力学
小样本学习
高效、低功耗
容错能力强

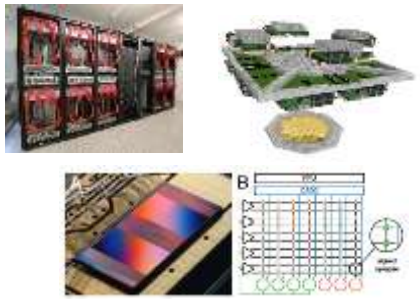
类脑智能

存算一体、模拟计算、稀疏计算、
事件驱动、神经可塑、分布并行、

人工智能

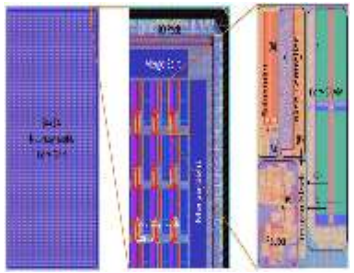
统计学
海量数据
暴力计算
鲁棒性差

目前的主要参与者



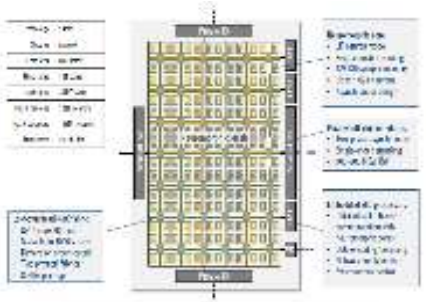
德国：海德堡大学BrainScales

2013年启动，模拟数字混合芯片512神经元，系统~400万神经元，65nm



美国：IBM TrueNorth

数字芯片，12.8万神经元，0.32亿突触, 系统0.32亿神经元，28nm



美国：Intel Loihi2

数字芯片， 100万神经元，1.2亿突触，系统100亿神经元，7nm

英国: 曼彻斯特大学 SpiNNaker

数字芯片， 15.3万神经元，1.53亿突触（外挂），ARM核，22nm

中国：清华大学天机

数字芯片， 25万神经元，0.25亿突触，12nm

浙江大学达尔文二代

14.7万神经元，1000万突触，系统最大1亿神经元（联合研究）

华为类脑芯片

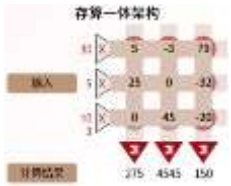
数字芯片， 25.6万神经元，0.64亿突触，系统支持25.6亿神经元

忆阻器：存算一体实现高能效AI芯片

传统芯片

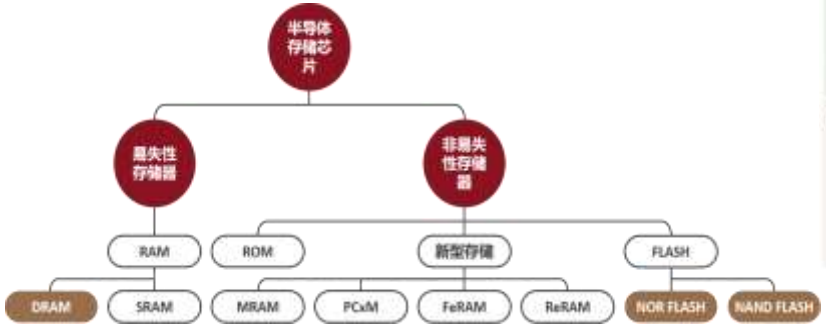
冯诺伊曼架构：计算与存储分离，数据搬运功耗和时延是计算的百~千倍，极大限制芯片能效。主流AI芯片能效<2TOPS/W。

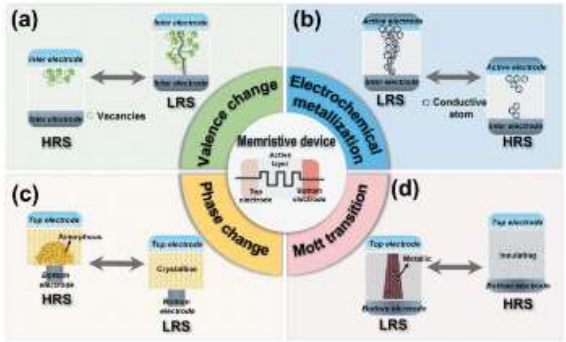
存算一体芯片

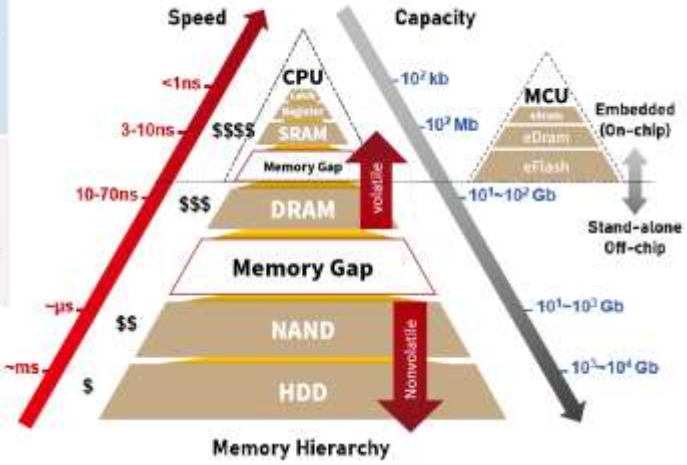


数据存储与计算融合在芯片的同一片区中，无需数据搬运，彻底解决冯诺依曼架构瓶颈。

忆阻器（memristor）属于新一代非易失性存储介质，在物理器件层面模拟人脑突触实现存算一体，根据原理主要分为四种器件：PCM, MRAM, ReRAM, FeRAM。算力可达 100TOPS/W，相对传统芯片提升50-100倍。



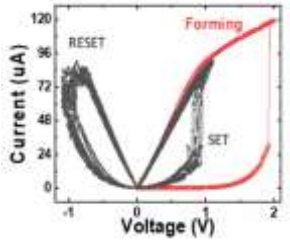


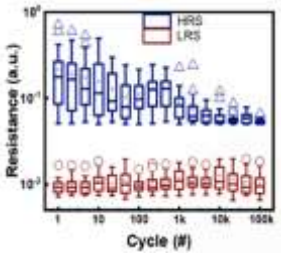


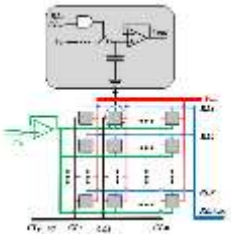
趋势：类脑算法与忆阻器的结合，实现高能效、鲁棒、端侧在线学习系统

忆阻器存算一体AI芯片的挑战

- 模拟器件的**热噪声、不稳定性、 器件差异、工艺良率**等问题影响忆阻器芯片的可拓展性、高精度训练和推理稳定性。
- 传统AI模型的**并行计算流程、训推分离算法**不适合在模拟器件上实现。

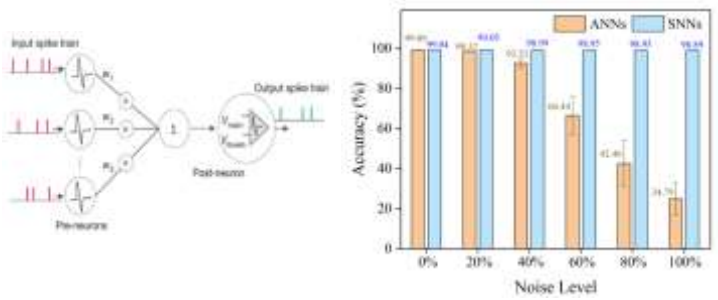






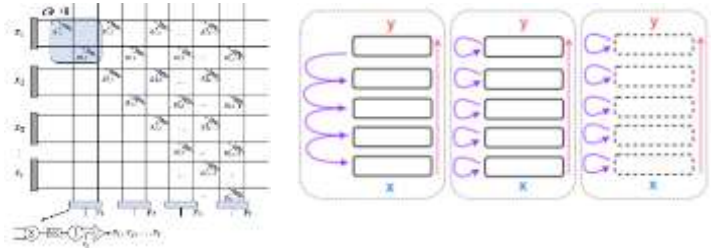
类脑计算的解决方案

稀疏类脑架构 ⟶ 噪声环境鲁棒计算



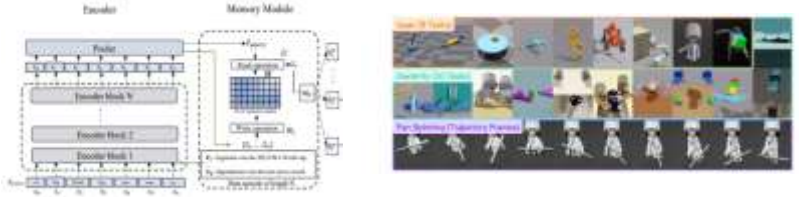
- 脉冲神经网络（SNN）通过**稀疏**计算**百倍降低**计算能耗，简化忆阻器电路；**二值脉冲信号**实现噪声环境下鲁棒计算。
- 类脑循环动力学网络相比传统Transformer架构在端侧实现更**高效部署**。

局部学习⟶ 端侧在线学习



- 局部学习**大幅降低（1-2个数量级）**网络训练开销，降低学习时延与能耗，突破传统训推分离范式，在端侧实现**数据流中在线学习**。
- 动态权重调节实现噪声环境中的稳定推理与学习。

记忆机制⟶端侧持续学习



- **短期记忆机制**高效利用实时数据，降低数据冗余和采集成本，缩短训练周期。
- **结构化记忆模型**在新任务下调用并更新旧知识，达到“稳定-可塑”平衡，实现**终身持续学习**。

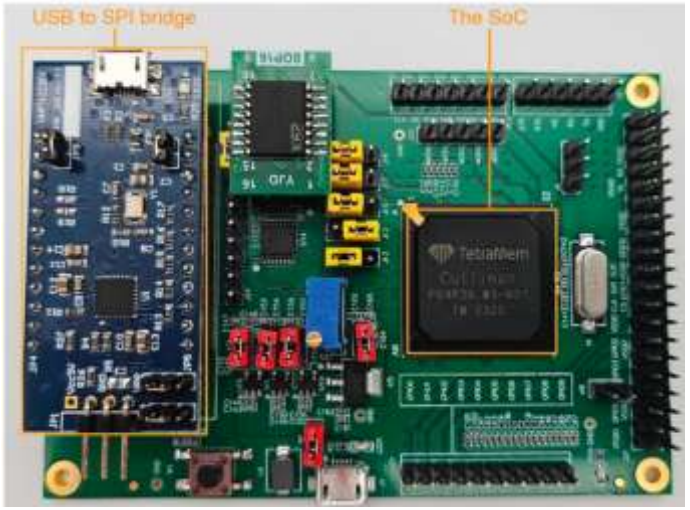
目录

忆阻器最新进展

TetraMem公司（硅谷），用ReRAM实现ANN加速

美籍华人教授，忆阻器学界泰斗创办
Joshua Yang、Mike Chen, USC教授，Qiangfei Xia, UMASS教授

2023 Nature正刊，通过高分辨率电导实现高精度ANN计算



TetraMem MX100 SoC 芯片， 内置存内计算NPU 和 RISC-V SoC。
可搭配 TetraMem SDK 实现边缘AI计算。2025年，小型网络实现 21TOPS/W

忆阻器最新进展

基于忆阻器芯片的脑机接口解码，实时四自由度无人机控制（天津大学，清华大学，2025.1）

nature electronics

Article

https://doi.org/10.1038/s41928-025-01340-2

A memristor-based adaptive neuromorphic decoder for brain–computer interfaces

Received: 18 February 2024

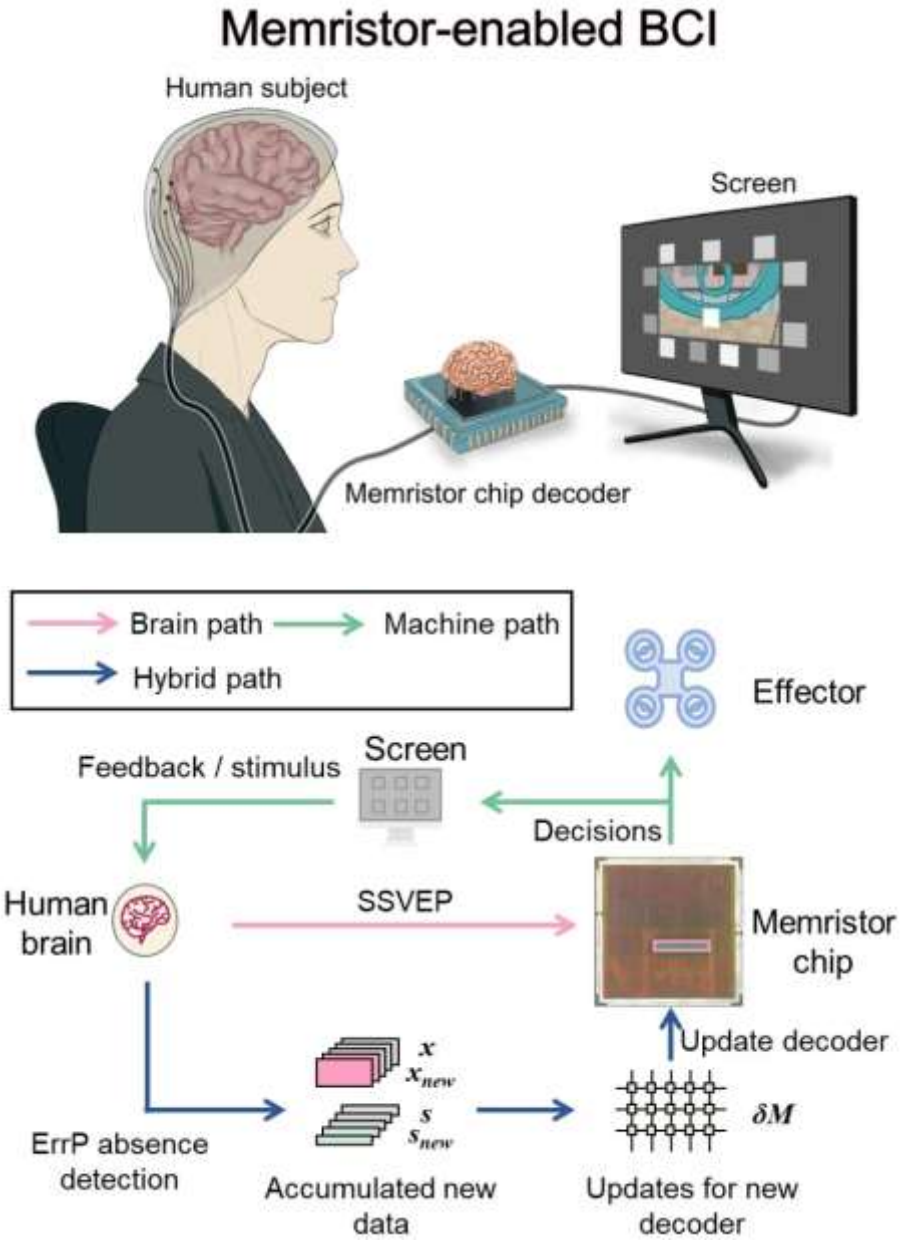
Accepted: 4 January 2025

Published online: 17 February 2025

Check for updates

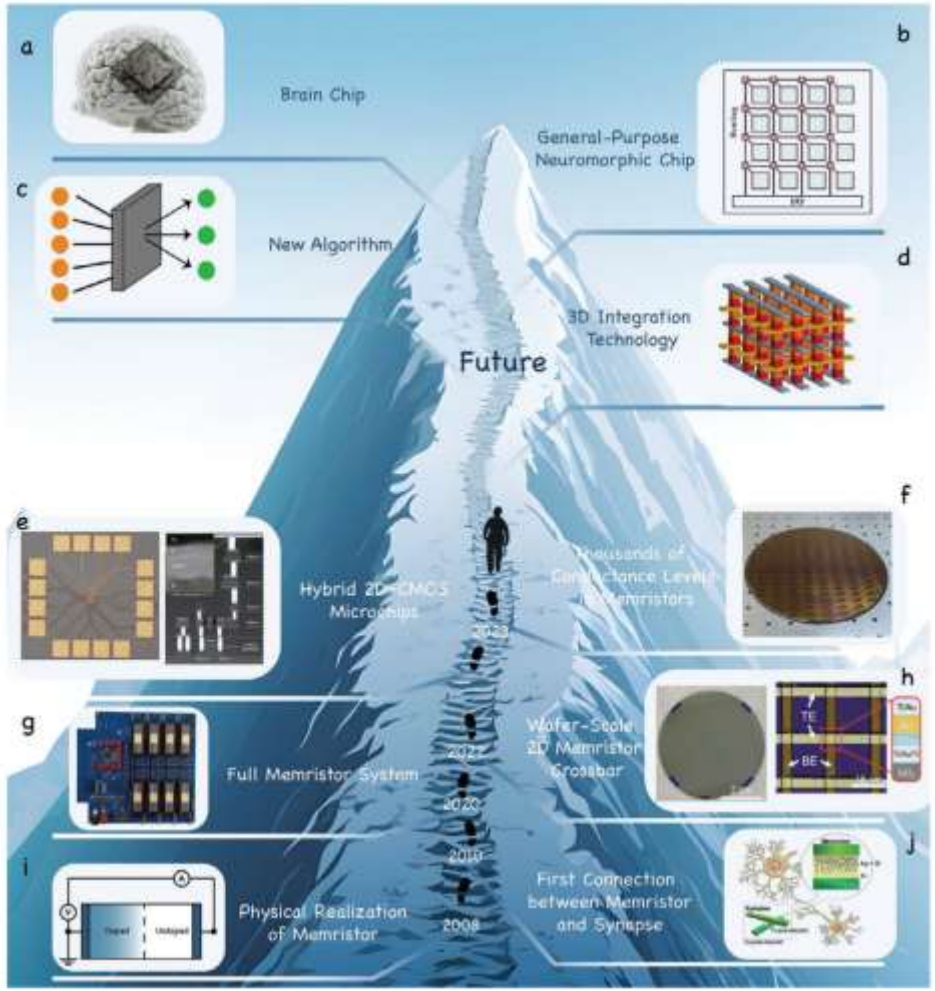
Zhengwu Liu[1,2,9], Jie Mei[3,4,9], Jianshi Tang[1,5] ✉, Minpeng Xu[3,6] ✉, Bin Gao[1,5], Kun Wang[3,6], Sanchuang Ding[1], Qi Liu[1], Qi Qin[1], Weize Chen[3,4], Yue Xi[1], Yijun Li[1], Peng Yao[1], Han Zhao[1], Ngai Wong[2], He Qian[1,5], Bo Hong[7], Tzyy-Ping Jung[3,8], Dong Ming[3,6] ✉ & Huaqiang Wu[1,9] ✉

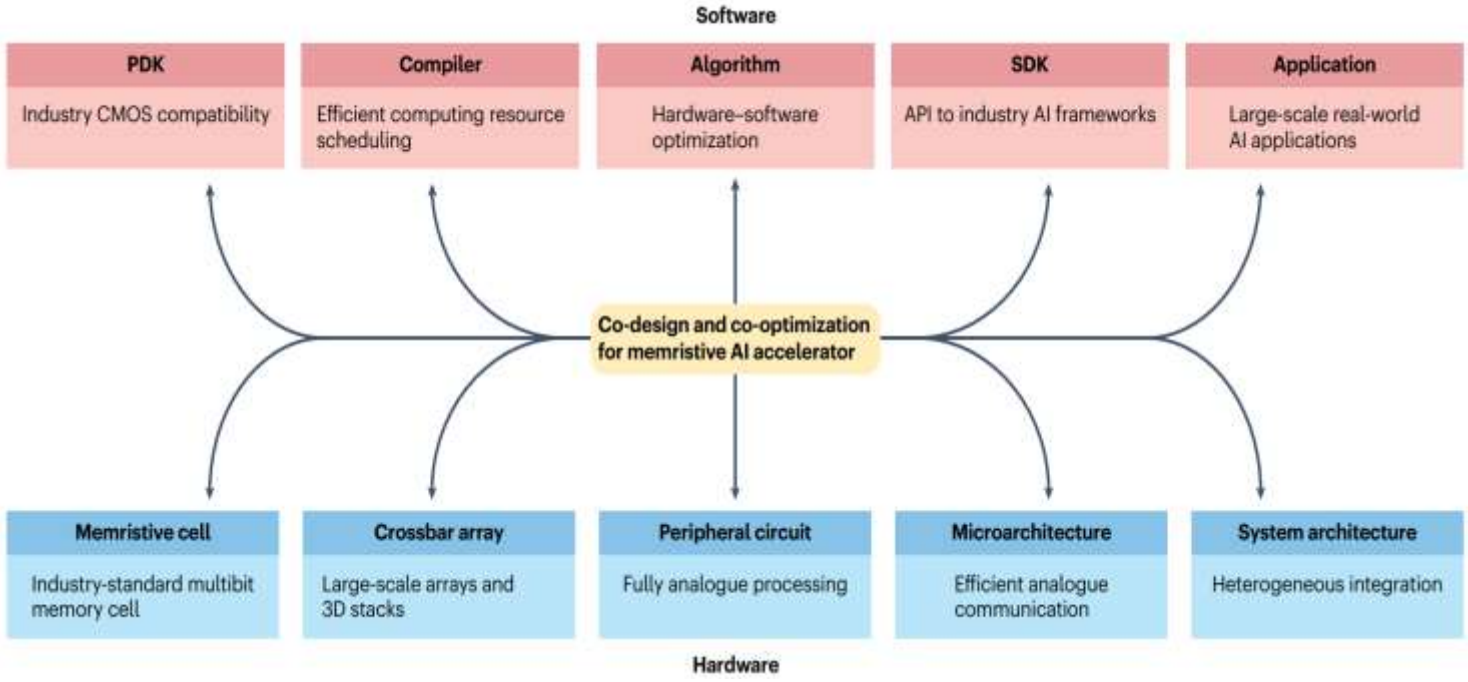
（来源：Nature Electronics）



- 首创了"**双环路脑机协同演进框架**"，忆阻器不断适应和学习脑电信号的变化-大脑不断优化输出的脑电特征，两者相互促进，形成了协同进化。
- 打破了传统脑机接口只能维持1小时左右的限制，实现了6小时的持续稳定操控，解码准确率提升20%。脑机接口的解码速度提升了216倍，能耗降低1000倍以上

忆阻器AI芯片的挑战：软硬协同开发，交互迭代





Memristor-Based Neuromorphic Chips, 2024
Memristor-based hardware accelerators for artificial intelligence, 2024

冷卢子未，电子科技大学应用物理系本科，德国海德堡大学博士，师从类脑计算先驱Prof. Karlheinz Meier（Google学术引用32+万次，H指数241，**欧盟人类大脑计划神经形态计算项目主任，清华大学、华为类脑计算顾问**），前华为主任工程师（2012实验室-中央研究院-先进计算与存储实验室）。**科技创新2030重大项目-脑科学与类脑研究-“支持在线学习的类脑芯片架构”项目（上海交大、华为） 华为负责人。**参与、主导华为多代类脑芯片网络架构、算法设计。深圳市海外高层次人才，南科大-华为合作项目企业导师，IEEE Senior Member。主持多项华为-国内高校、科研院所合作项目。领导华为类脑计算团队在人工智能顶会顶刊发表数十篇论文，主导类脑算法在华为无线、终端、云计算等场景的落地。**循忆星海（上海）科技有限公司创始人**

群聊：类脑计算交流群（5）

实习咨询请联系（发送个人简历、意向研究方向，硕博优先）

lengluziwei@braingalaxy.ai

谢谢大家，欢迎入群交流！

（将定期发送类脑前沿资料、讲座信息等）