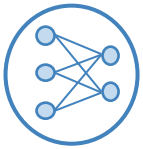
SpikingBrain: Spiking Brain-inspired Large Models

Brain-inspired Mechanisms

Construct



Model Architecture

Training

Development Pipeline

Multi-scale sparsity

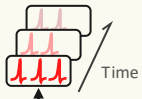
- ✓ Event-driven
- ✓ Adaptive Sparsity
- ✓ Firing Activity Control

- ✓ Modular Sparsification
- ✓ Functional Specialization

- ✓ Compressed Memory
- ✓ Continuously Updated
- ✓ Markov Properties
- ✓ Dendritic Dynamics

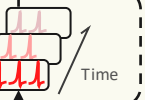
(Hybrid) Linear Models

Spike Encoding



MoE / FFN

- Adaptive threshold
- Simplified computation
- Both integer & spike



Hybrid Efficient Attention

Open-source Transformer models

Conversion-based Training

- CPT ~ 150B tokens
- Long-context
- General data

Generality & Efficiency
Performance Comparable !

Train from scratch

- PT >10T tokens
- Limited context
- Date quality

Next-generation neuromorphic chip design

Support

Adaptation on Non-NVIDIA GPU Clusters

- Training Framework
- CUDA / Triton Operators
- Communication Primitives
- Parallelism strategies

- Spiking sparsity **> 69%**
- **< 2%** data resource
- **> 100x** 4M TTFT speedup