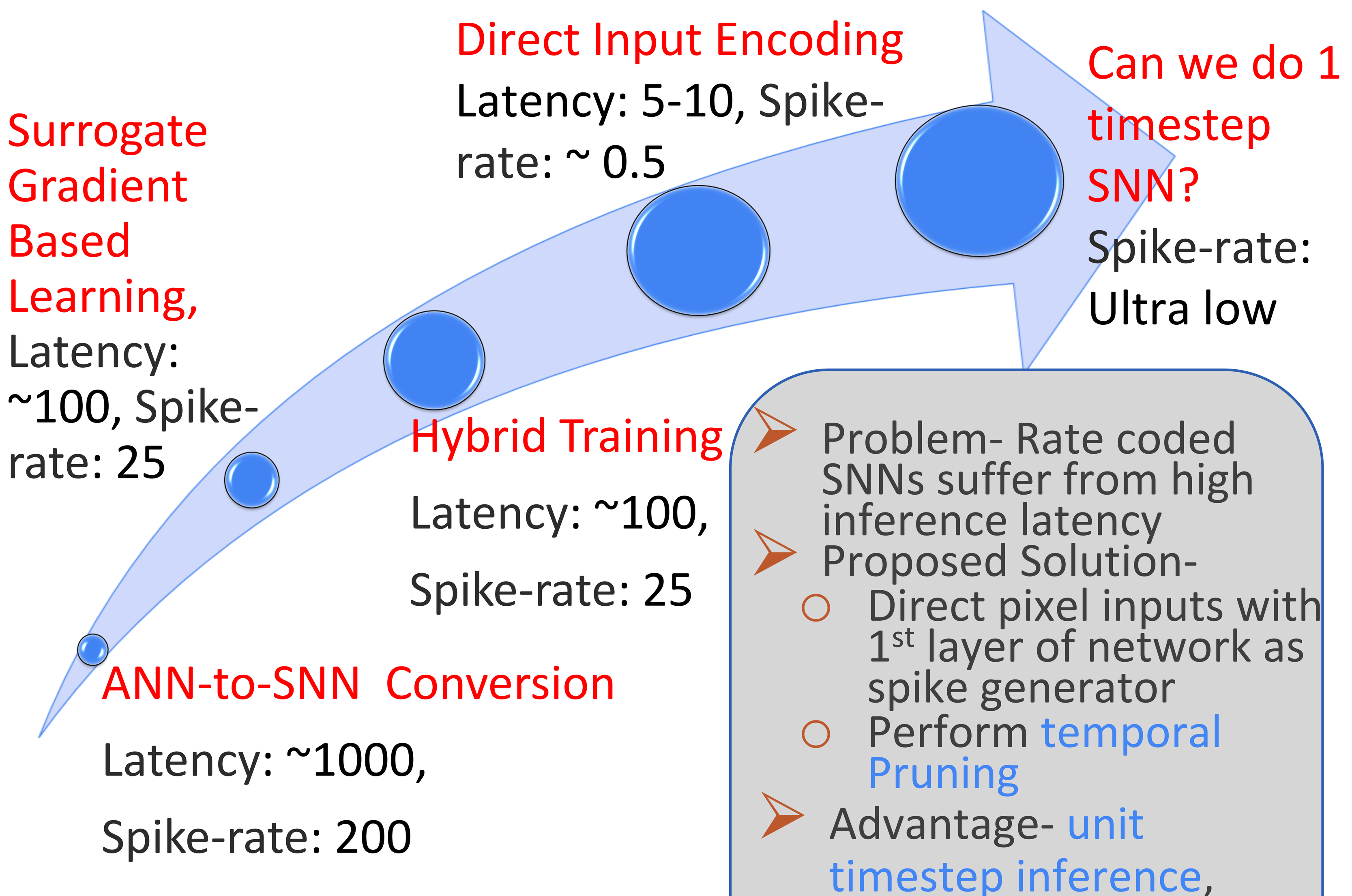
Towards Ultra Low Latency Spiking Neural Networks for Vision and Sequential Tasks Using Temporal Pruning

Purdue University
Elmore Family School of Electrical and Computer Engineering

Sayeed Shafayet Chowdhury, Kaushik Roy, Purdue University

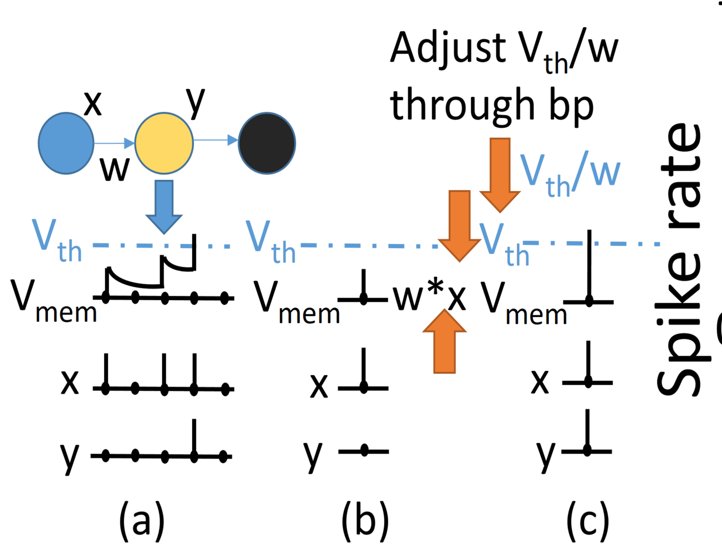
C-BRIC Center for Brain Inspired Computing

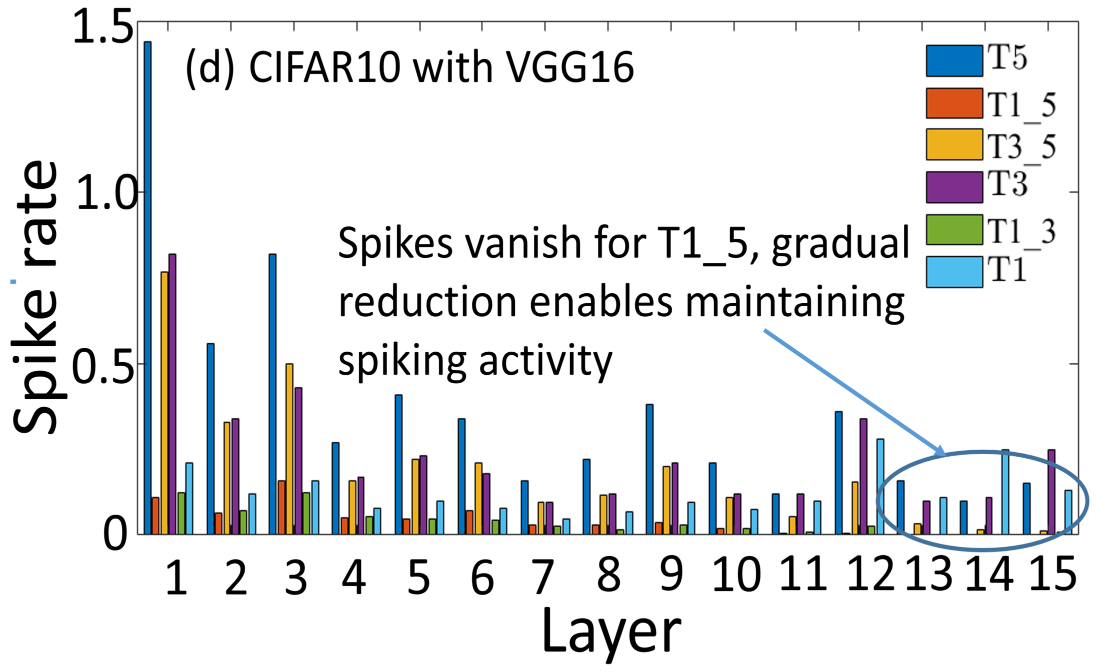
Background and Motivation



- Problem- Rate coded SNNs suffer from high inference latency
- Proposed Solution-
  - Direct pixel inputs with 1st layer of network as spike generator
  - Perform temporal Pruning
- Advantage- unit timestep inference, lower compute, high sparsity

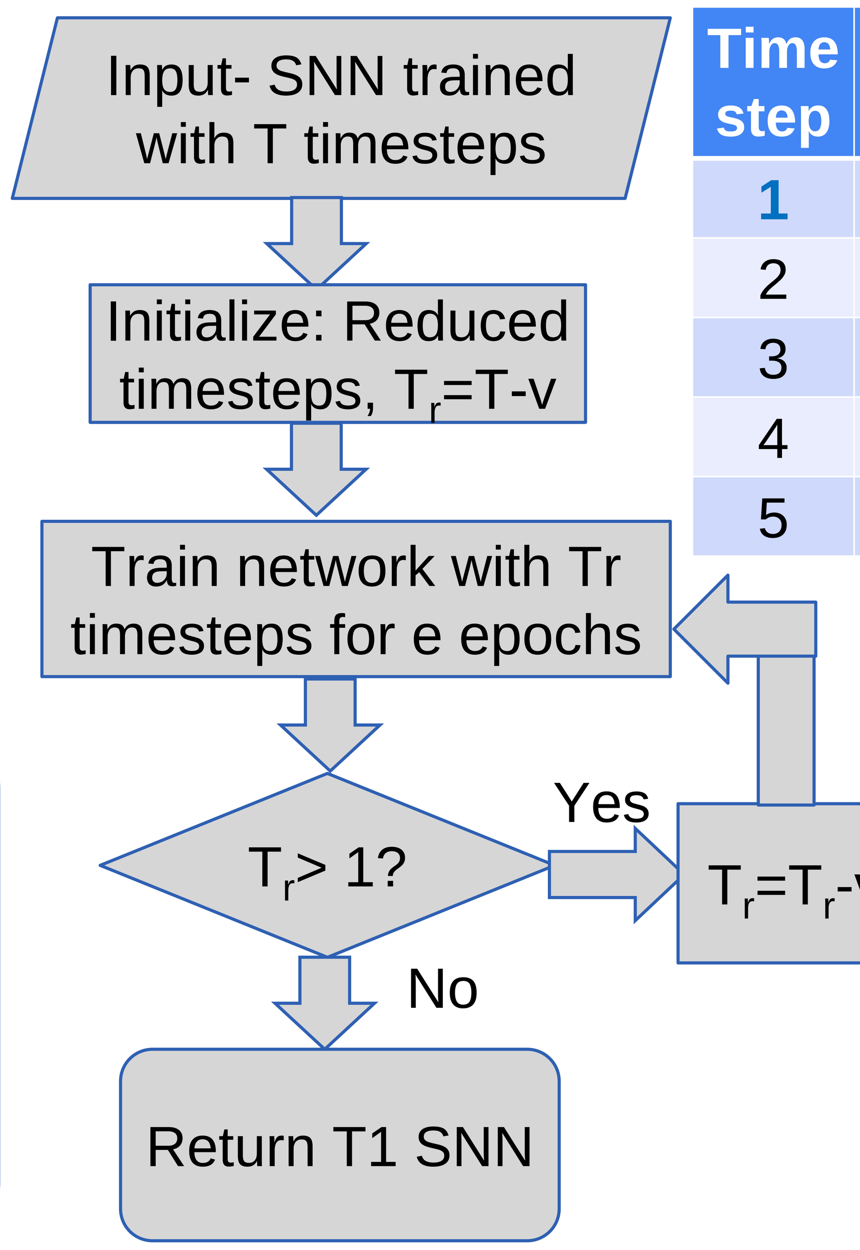
Training with Temporal Pruning





- Directly training with 1 timestep results in convergence failure due to spike vanishing at the deeper layers
- Divide and conquer approach- train an SNN with higher T and gradually reduce T till 1 ⇒ 'Temporal Pruning'
- At each pruning iteration, SNN trained previously with higher T is used as initialization; Leverage the temporal axis of SNNs for compression

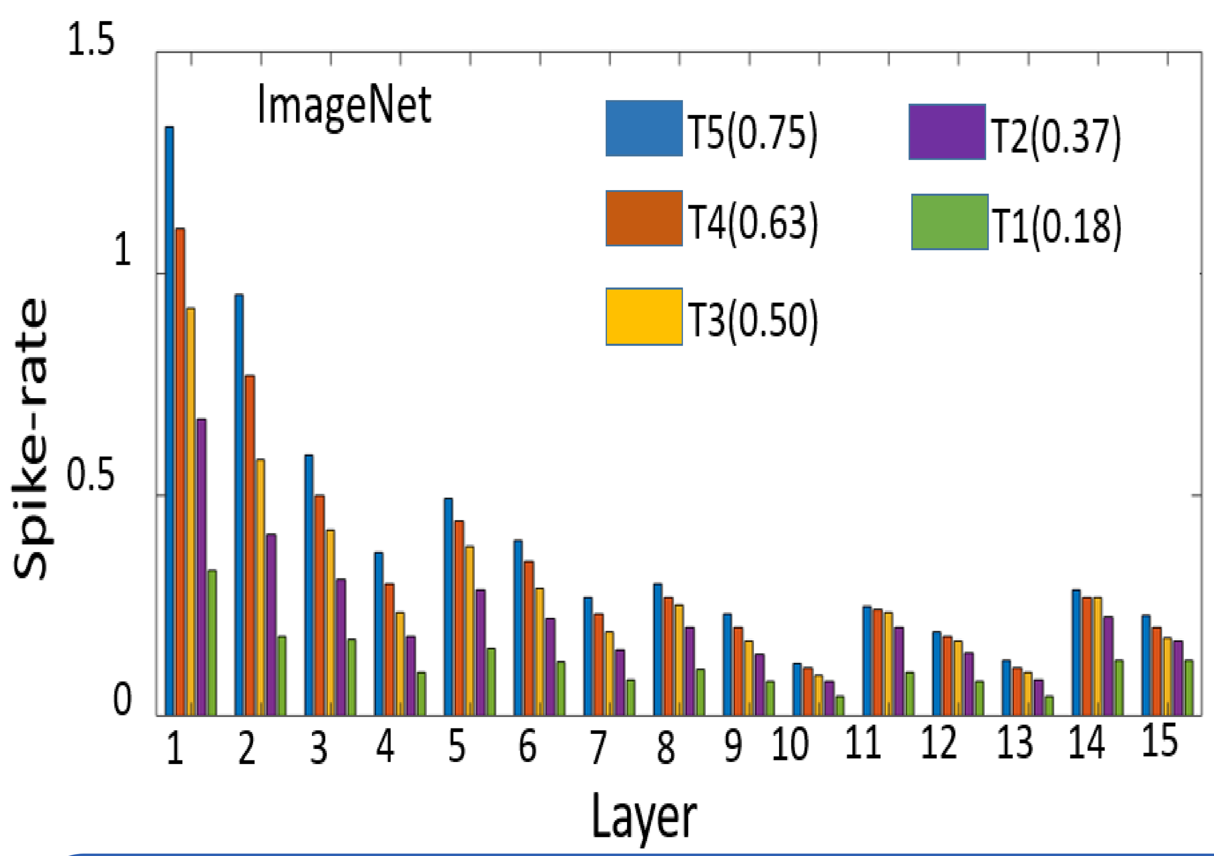
Proposed Approach and Results



Time step	CIFAR10 (top1 %)	CIFAR100 (top1 %)	ImageNet (top1 %)
1	**93.05**	**70.15**	**69.00**
2	93.72	71.43	69.00
3	93.85	71.46	69.01
4	93.87	71.51	69.03
5	93.90	71.58	69.05

- Hybrid training with sequential temporal pruning
- Reduces SNN latency to lowest possible limit
- Comparable performance to ANNs with iso-latency
- SOTA SNN performance on ImageNet with T=1

Efficiency Improvements



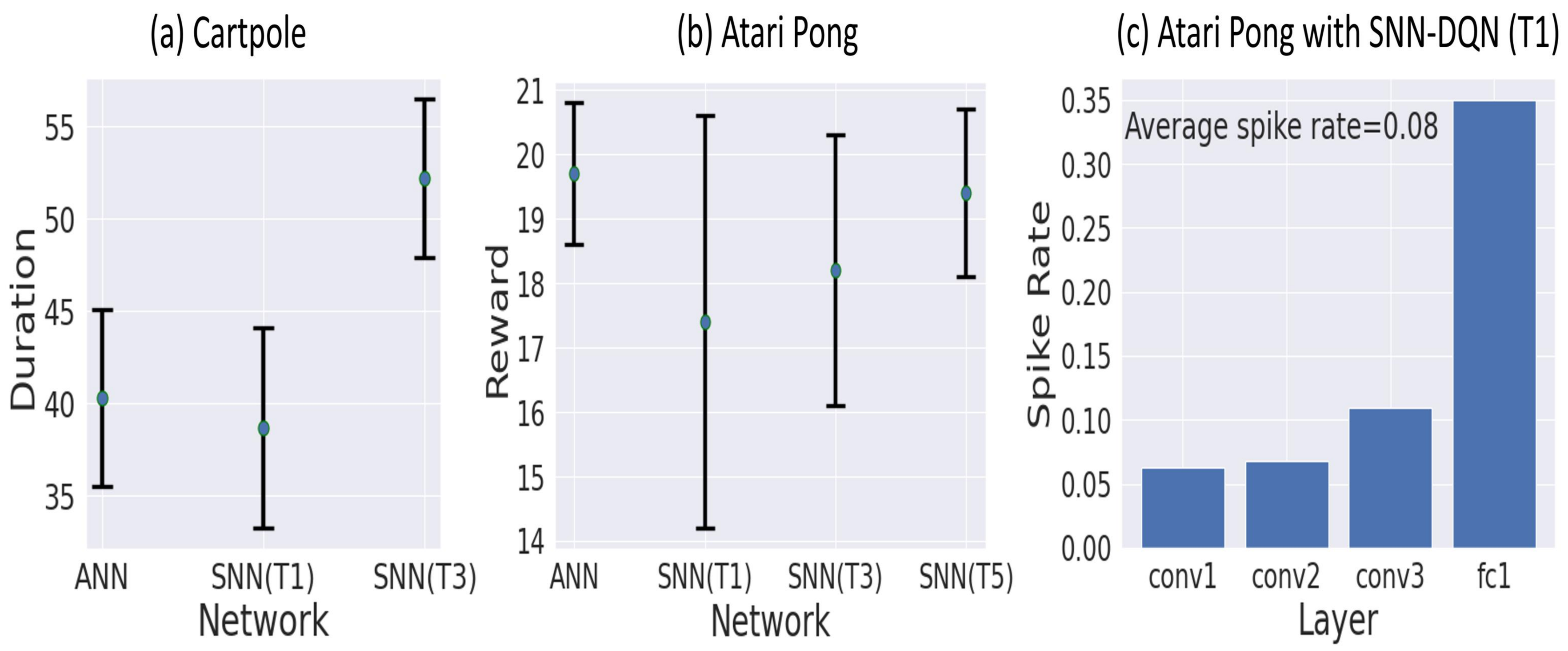
ANN vs T1 SNN compute energy

Network	Dataset	α
VGG16	CIFAR10	33.0
VGG16	CIFAR100	29.24
VGG16	ImageNet	24.61
ResNet20	CIFAR10	16.32
ResNet20	CIFAR100	15.35

- 5X reduction in spike-rate compared to prior SOTA
- 25-33X higher energy efficiency compared to ANNs
- 5-100X reduction in memory access cost of membrane potential compared to prior art

Key Takeaway: Temporal Pruning enables SNN inference with unit timestep providing ultra high efficiency. SNNs with inherent memory of membrane potential can enhance performance on RL tasks, demonstrating the suitability of SNNs for sequential learning.

Reinforcement Learning



- RL requires processing of sequential inputs- SNNs are suitable candidates due to inherent memory of membrane potential
- 1.3X reward on Cartpole compared to ANN using membrane potential
- Atari Pong inference with T=1 to 5, enhancement in reward with increase in T showing efficacy of recurrence of LIF neurons
- 5-7X higher energy efficiency compared to ANN-DQNs

Comparison with State-of-the-Art

Reference	Dataset	Accuracy (%)	Timesteps
Sengupta *et al.*	CIFAR10	91.55	2500
Wu *et al.*	CIFAR10	50.7	30
Rueckauer *et al.*	CIFAR10	90.85	400
Zheng *et al.*	CIFAR10	93.16	6
Rathi *et al.*	CIFAR10	92.70	5
This work	**CIFAR10**	**93.05**	**1**
Lu *et al.*	CIFAR100	63.2	62
Rathi *et al.*	CIFAR100	69.67	5
Park *et al.*	CIFAR100	68.80	680
This work	**CIFAR100**	**70.15**	**1**
Rathi *et al.*	ImageNet	69.00	5
Zheng *et al.*	ImageNet	67.05	6
Fang *et al.*	ImageNet	67.04	4
This work	**ImageNet**	**69.00**	**1**

Acknowledgments

This work was supported in part by C-BRIC, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA