VOWEL: A Local Online Learning Rule for Recurrent Networks of Probabilistic Spiking Winner-Take-All Circuits

Hyeryung Jang

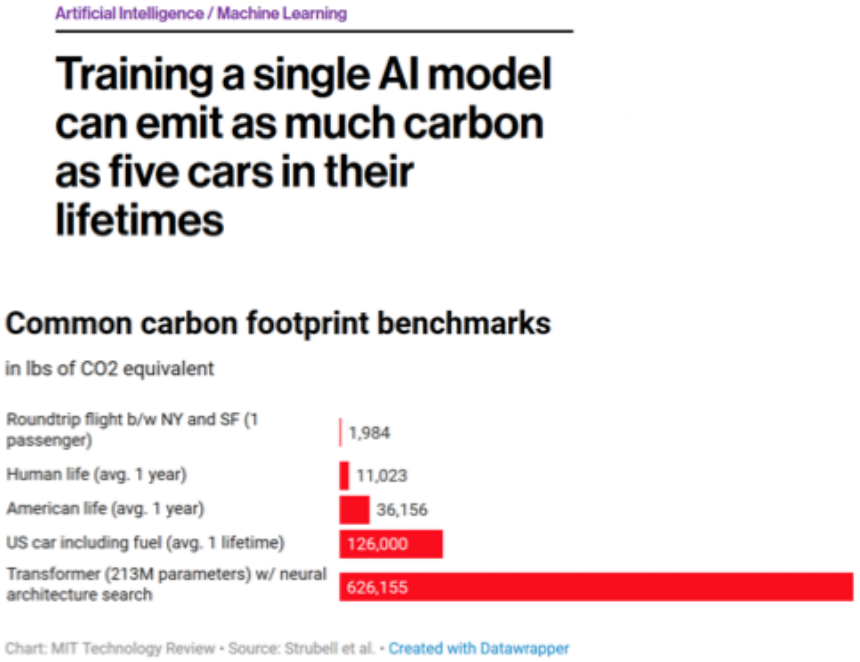
Joint work with Nicolas Skatchkovsky and Osvaldo Simeone

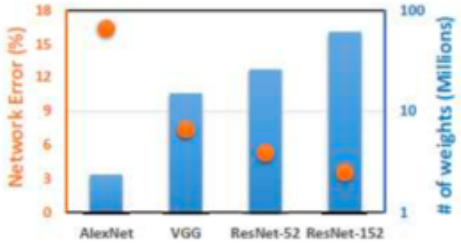
King's College London

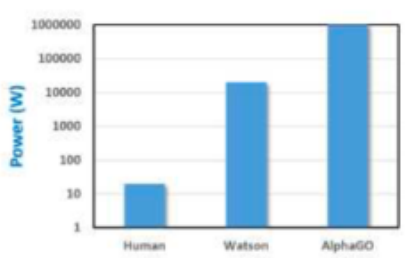
ICPR, Jan/12/2021

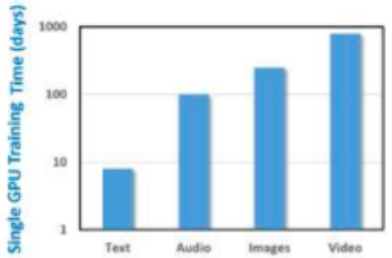
Machine Learning Today

- Breakthroughs in ML using deep Artificial Neural Networks (ANNs) have come at the expense of massive memory, energy, and time requirements:
  - Energy intensive, particularly for “always-on” edge intelligence applications
  - Many state-of-the-art solutions are not well suited for time-varying and non-stationary environments



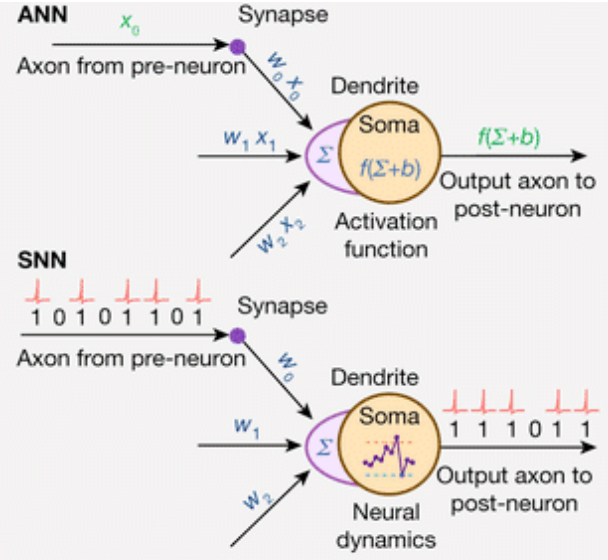


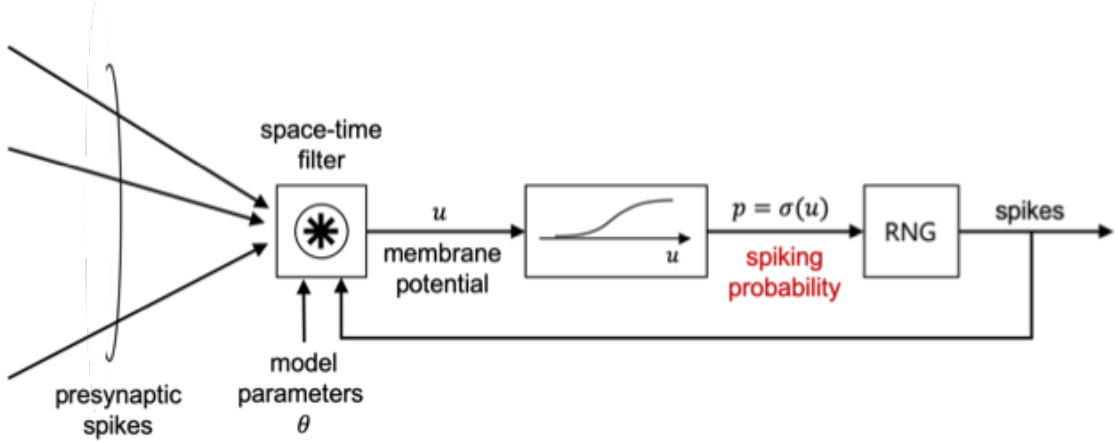




Neuromorphic Computing: Spiking Neural Networks (SNNs)

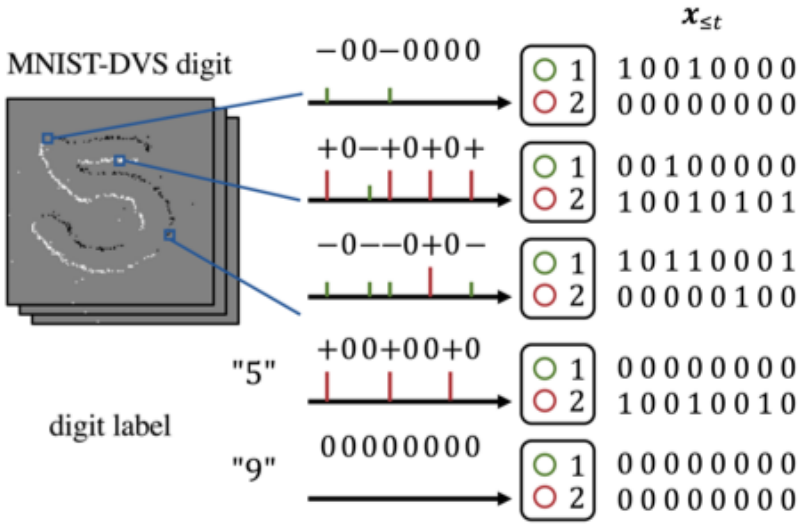
- Spiking Neural Networks (SNNs) take inspiration from the dynamic, temporally sparse, event-driven learning and inference operations of the human brain
- ANNs to SNNs: From static scalar activation functions to dynamic neuronal models
  - Neurons in the brain sense, process, and communicate over time using sparse binary signals (=spikes)
- Probabilistic SNN Models based on Generalized Linear Model (GLM)
  - More flexible and enable principled differentiable learning rules





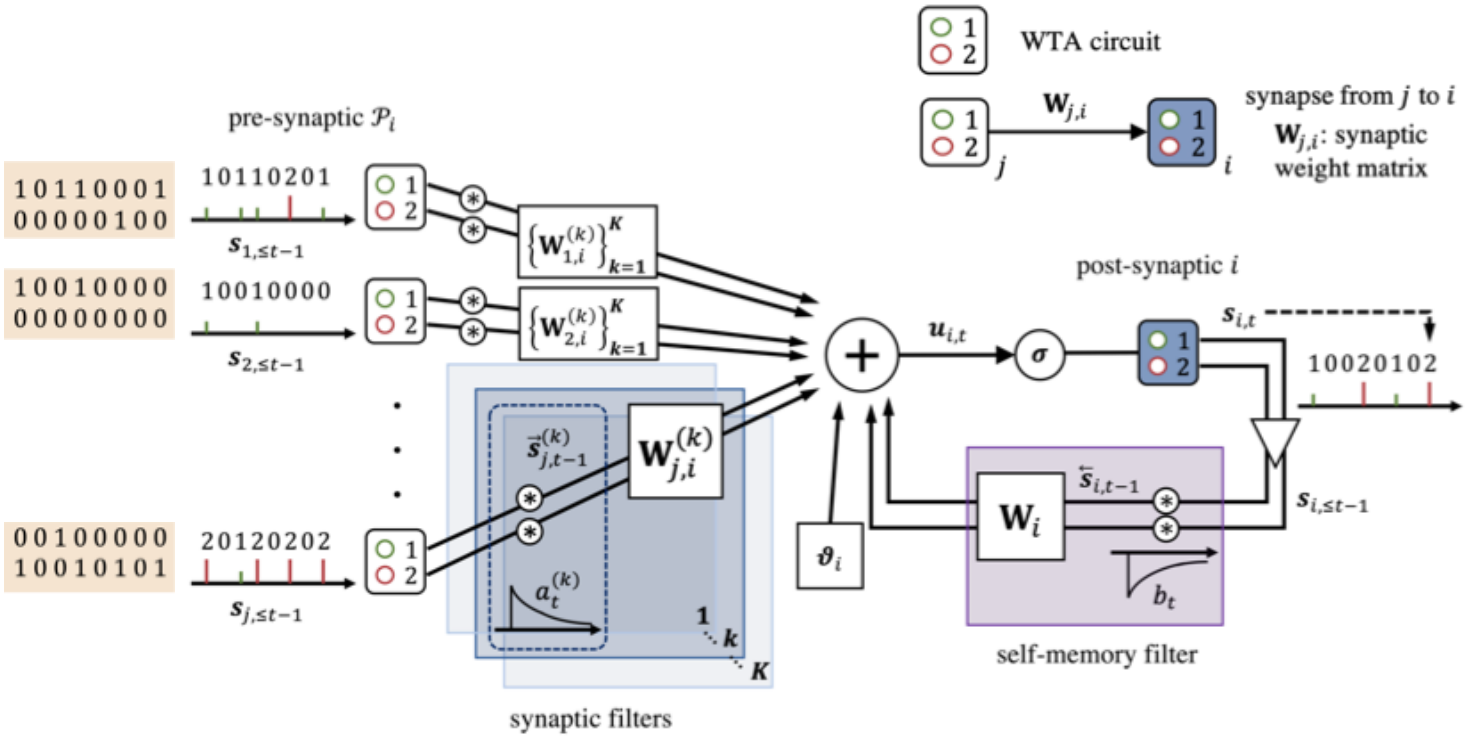
Multi-Valued Spikes: Beyond Binary Spikes

- Time-encoded data may assign spikes discrete, categorical, values
  - e.g., polarity $(+, -)$ for data from neuromorphic cameras
- With standard SNN models, this information is either discarded or encoded by increasing the number of signals
  - e.g., producing a binary signal (spike or not), treating as two separate binary signals
- How to model and train an SNN that can process with multi-valued spikes?



Probabilistic (GLM) Winner-Take-All SNNs (WTA-SNNs)

- Discrete data can be encoded via one-hot vector
- It can be processed and produced by Winner-Take-All (WTA) spiking circuit:
  - Group of correlated spiking units, with at most one of the spiking units emitting a spike at any given time



- WTA circuit: categorical one-hot GLM
- Synaptic weights defined by matrices
- Membrane potential

$$\boldsymbol{u}_{i,t} = \sum_{j \in \mathcal{P}_i} \mathbf{W}_{j,i} \vec{\boldsymbol{s}}_{j,t-1} + \mathbf{W}_i \overleftarrow{\boldsymbol{s}}_{i,t-1} + \boldsymbol{\vartheta}_i$$

- Probability of WTA i to emit a spike at unit c

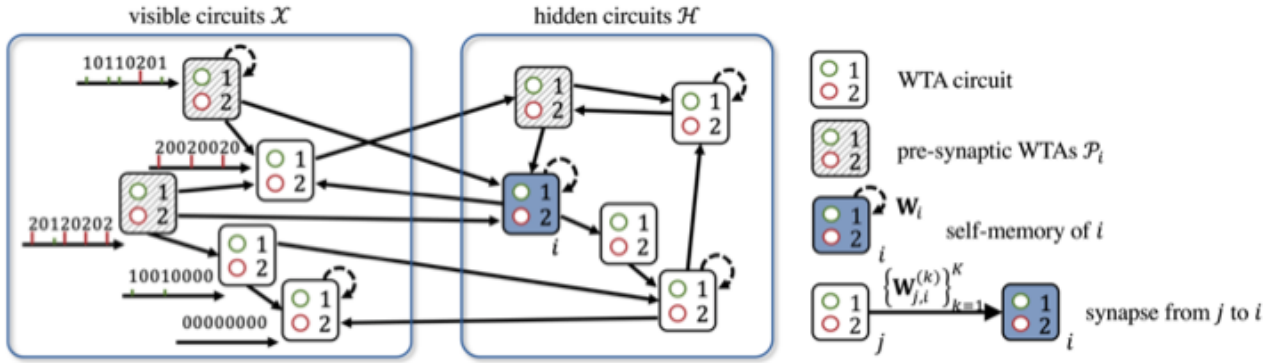
$$p_{i,t}(c; \boldsymbol{\theta}_i) = \boldsymbol{\sigma}_{\mathrm{SM}}^{c}(\boldsymbol{u}_{i,t})$$

- Log-prob: categorical (neg) cross-entropy

$$\log p_{i,t}(\boldsymbol{s}_{i,t}; \boldsymbol{\theta}_i) = \bar{H}\left(\boldsymbol{s}_{i,t}, \boldsymbol{\sigma}_{\mathrm{SM}}(\boldsymbol{u}_{i,t})\right)$$

Training Probabilistic WTA-SNNs

- The visible circuits encode the desired behavior $\boldsymbol{x}_{\leq T}$ of the network as specified by training data
- The spiking behavior $\boldsymbol{h}_{\leq T}$ of the hidden circuits is not specified by data

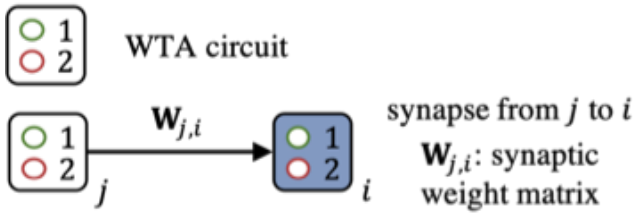


- Maximum likelihood: $\max_{\theta} \log p_{\boldsymbol{\theta}}(\text{visible})$
  - Maximize the probability that the model produces the desired outputs for the visible circuits
  - $\log p_{\boldsymbol{\theta}}(\text{visible}) = \log \sum_{\text{hidden}} p_{\boldsymbol{\theta}}(\text{visible}, \text{hidden}) \geq \mathbb{E}_{q(\text{hidden})}\left[\log \frac{p_{\boldsymbol{\theta}}(\text{visible},\text{hidden})}{q(\text{hidden})}\right]$
  - Forward distribution $\log q(\text{hidden}) = \sum_t \sum_{i:\text{hidden}} \bar{H}\left(\boldsymbol{h}_{i,t}, \boldsymbol{\sigma}_{\text{SM}}\left(\boldsymbol{u}_{i,t}\right)\right)$
  - We specifically optimize the **lower bound (ELBO)**: $L(\boldsymbol{\theta}) = \sum_t \sum_{i:\text{visible}} \bar{H}(\boldsymbol{x}_{i,t}, \boldsymbol{\sigma}_{\text{SM}}(\boldsymbol{u}_{i,t})) - \alpha \cdot \text{reg}$

VOWEL: **V**ariational **O**nline learning rule for spiking **W**inner-tak**E**-al**L** circuits

- Use an unbiased estimate (via REINFORCE) of the gradient of the bound $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$
- For visible circuits, we have two-factor rule:

$$\mathbf{W}_{j,i} \leftarrow \mathbf{W}_{j,i} + \eta \cdot \underbrace{\left(\boldsymbol{x}_{i,t} - \boldsymbol{\sigma}_{\mathrm{SM}}\left(\boldsymbol{u}_{i,t}\right)\right)}_{\text{post-synaptic predictive error}} \cdot \underbrace{\left(\vec{\boldsymbol{s}}_{j,t-1}\right)^{\mathsf{T}}}_{\text{pre-synaptic trace}}$$

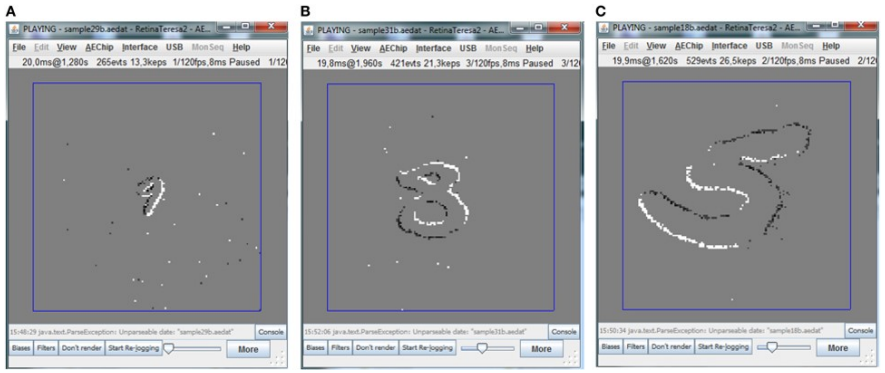


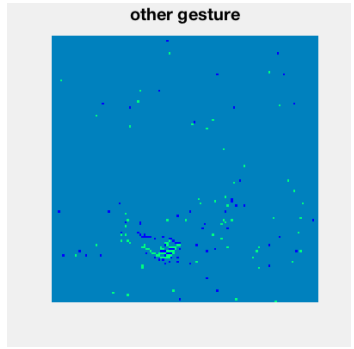
- For hidden circuits, we have three-factor rule:

$$\mathbf{W}_{j,i} \leftarrow \mathbf{W}_{j,i} + \eta \cdot \underbrace{\sum_{i:\text{visible}} \bar{H}\left(\boldsymbol{x}_{i,t}, \boldsymbol{\sigma}_{\mathrm{SM}}\left(\boldsymbol{u}_{i,t}\right)\right)}_{\text{global reward signal}} \cdot \underbrace{\left(\boldsymbol{h}_{i,t} - \boldsymbol{\sigma}_{\mathrm{SM}}\left(\boldsymbol{u}_{i,t}\right)\right)}_{\text{post-synaptic predictive error}} \cdot \underbrace{\left(\vec{\boldsymbol{s}}_{j,t-1}\right)^{\mathsf{T}}}_{\text{pre-synaptic trace}}$$

Experiments

- Neuromorphic dataset obtained by filming moving MNIST digits displayed on a screen or moving gestures with a neuromorphic camera
  - MNIST-DVS, DVS-Gesture

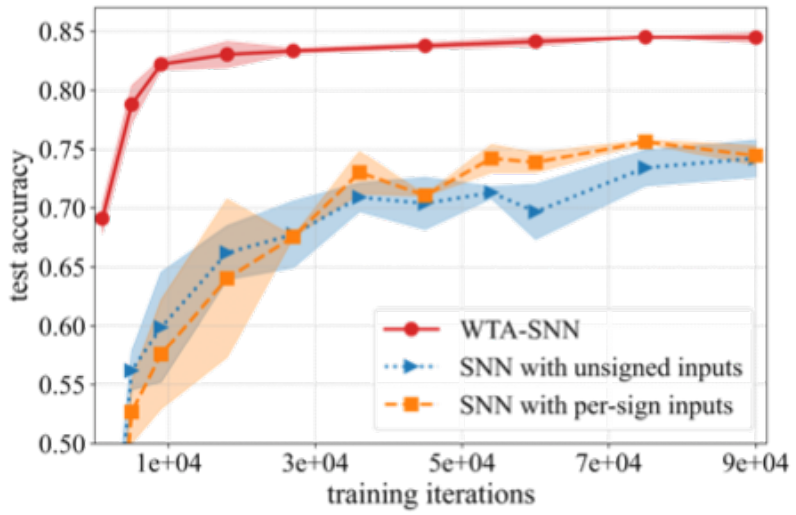




- WTA-SNN
  - All WTA circuits consist of two spiking units (for $+, -$ signs)
  - Consider a generic, non-optimized, network architecture with H fully connected hidden circuits
- Benchmarks: (binary) probabilistic GLM model and deterministic LIF model
  - With same number of spiking units ($= 2H$) and per-sign inputs
  - With ½ number of spiking units ($= H$) with unsigned (discarding the sign) inputs

Results

- MNIST-DVS
  - $H = 16$ hidden circuits ($2H$ hidden spiking units)



  - WTA-SNN ($C = 2$) trained with VOWEL outperforms conventional binary SNN solutions ($C = 1$)
  - The capability of WTA circuits to process information encoded both in the spikes' timings and their signs

- DVS-Gesture
  - Test accuracy of VOWEL and DECOLLE (deterministic SNN)

Model	Period	H	Accuracy
DECOLLE	1 ms	2048	61.42 ± 2.92 %
	1 ms	1024	57.75 ± 3.22 %
	1 ms	512	56.42 ± 2.03 %
	10 ms	512	34.72 ± 0.75 %
	20 ms	512	26.38 ± 0.28 %
VOWEL	20 ms	256	$60.26 \pm 0.91\%$
	20 ms	128	$57.96 \pm 0.11\%$
	20 ms	64	$57.89 \pm 0.35\%$

  - VOWEL can operate with a smaller number of hidden units and coarser sampling rate
  - The ability to directly distinguish patterns encoded in the values of the spikes

Thank you!

Any Questions?

hyeryung.jang@kcl.ac.uk