Biologically Plausible Variational Policy Gradient with Spiking Recurrent Winner-Take-All Networks

Zhile Yang¹, Shangqi Guo†², Ying Fang³, Jian K. Liu†¹
¹University of Leeds, ²Tsinghua University, ³Fujian Normal University

Keywords: reinforcement learning, spiking neural networks

Message: + SVPG, a spiking-based variational policy gradient method with RWTA network and R-STDP. + Experiment results reveal its potential for solving RL tasks and to have inherent robustness.

1. Background & Problem

- **Background**
  - Human-level? Adaptability, robustness ...
  - ANN to SNN conversion
  - Surrogate gradient-based BP
  - Modulated STDP

- **Problem**
  - LIF neuron model
  - Local learning rules
  - Global RL target
  - Policy function

- **Global RL target**
  \[ \nabla_{\pi} J(\pi) = E_{s \sim \rho(s)} \sum_{t=0}^{T-1} \sum_{i=0}^{T} \nabla \ln \pi(a_i|s_i) \]

- **Policy Inference**

  - Target function
    \[ p(\nu_a, \nu_b|s) = \frac{1}{Z(q)} \exp(E(\nu)) \]

  - Mean-field inference
    \[ \tilde{p}(\nu_a, \nu_b|s) = \tilde{p}(\nu_a|s)p(\nu_b|s) \]

  - Minimize KL-divergence
    \[ \mathcal{D}_{KL}(s)|\rho_b = 0 \| \rho = \frac{1}{Z(\rho)} \exp(E(\nu)|q + w^T \nu + b) \]

- **In RWTA network**
  - Let \[ \rho_l = \tilde{\rho} \]
    \[ \Rightarrow \] RWTA net is suitable for policy inference

- **In neuron model**
  - Let \[ \int \kappa(y) = 1/g \]
    \[ \Rightarrow \] The designed STDP rules can do approximated policy optimization

2. Designs

- **Winner-take-all circuits**
  - Environment
  - Fully connected
  - Action Selection
  - Unidirectional FC
  - Bidirectional FC
  - WTA

- **Policy function**
  - Policy \( \pi \) based on an energy function of the firing states.
    \[ \pi(s) = \sum_{a} \tilde{p}(\nu_a|s) \cdot p(\nu_b|s) = \frac{\tilde{p}(\nu_a|s) \exp(E(\nu))}{Z(\rho)} \]

- **Input noises**
  - MNIST (RL)
  - Gym InvertedPendulum-v0

3. Policy Inference & Optimization

- **Policy Inference**

  - Target function
    \[ p(\nu_a, \nu_b|s) = \frac{1}{Z(q)} \exp(E(\nu)) \]

  - Mean-field inference
    \[ \tilde{p}(\nu_a, \nu_b|s) = \tilde{p}(\nu_a|s)p(\nu_b|s) \]

  - Minimize KL-divergence
    \[ \mathcal{D}_{KL}(s)|\rho_b = 0 \| \rho = \frac{1}{Z(\rho)} \exp(E(\nu)|q + w^T \nu + b) \]

- **In RWTA network**
  - Let \[ \rho_l = \tilde{\rho} \]
    \[ \Rightarrow \] RWTA net is suitable for policy inference

- **In neuron model**
  - Let \[ \int \kappa(y) = 1/g \]
    \[ \Rightarrow \] The designed STDP rules can do approximated policy optimization

4. Experiments

- **Tasks & Variations**

  - MNIST (RL)
  - Gym InvertedPendulum-v0

  - Input noises
  - Network noises
  - Pendulum length
  - Pendulum thickness

- **Results**

  - SVPG
  - BP
  - BPTT
  - EP
  - ANN

  - MNIST: 0.920 ± 0.001
  - GymIP: 199.87 ± 0.27

- **Graph accuracy**
  - MNIST: Graph accuracy
  - GymIP: Graph accuracy

- **Pendulum length**
  - MNIST: Pendulum length
  - GymIP: Pendulum length

- **Pendulum thickness**
  - MNIST: Pendulum thickness
  - GymIP: Pendulum thickness