Neural Network Panning

Screening the Optimal Sparse Network Before Training

Authors: Xiatao Kang
Ping Li
Jiayi Yao
Chengxi Li

Introduction:

A. Pruning the neural network before training can accelerate the training phase, and has interpretability and strong application value.

B. We generalize the pruning process as the weight expressive force transfer process, where any weight has an equal opportunity to express.

C. We propose a multi-metric, multi-process iterative pruning method before training, called Panning.

D. We design a Panning environment and obtain the optimal expressive force transfer policy based on reinforcement learning.

Methods:

A. Artificial Panning

- Iteratively prune T times.
- Allow the removed weights to repeat the activity.
- Gradually increase the compression ratio.
- Screen the weights based on the fusion metrics.

B. Panning Based on Reinforcement Learning

- Panning Environment

The State Space

$s_t = (\mathcal{L}, \Delta\mathcal{L}, \mathcal{L}_s, \Delta\mathcal{L}_s, \rho_t, \rho_e, t)$

Contains the current state loss, loss change, effective compression ratio, and pruning times.

The Action Space

$a \in [-1,1], p_i = \frac{a_i + 1}{2}$

Actions are used to modify the metric.

Reward Function

$R_t = -\left|\mathcal{N}(\mathcal{L}) - \mathcal{N}(\mathcal{L}_s)\right| - \left|\mathcal{N}(\Delta\mathcal{L}) - \mathcal{N}(\Delta\mathcal{L}_s)\right| - \alpha\left|\rho_e - \rho_t\right| - r_{\text{done}}$

Our goal is to ensure that pruning has minimal impact on network loss and gradients, and that the model compresses efficiently.

- TD3 Agent

Learning Panning policy by sampling spatial actions.

C. Select and optimize metrics

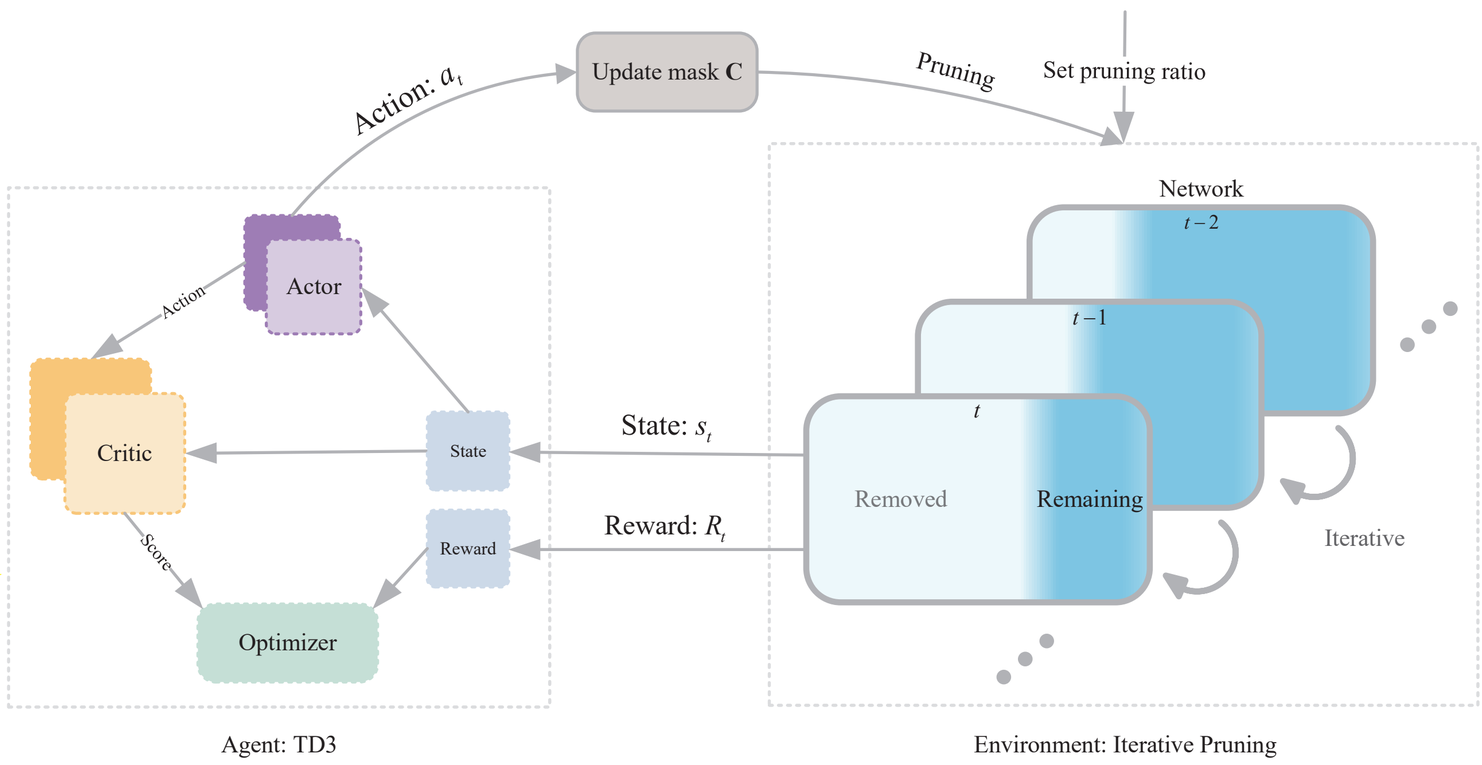
Results:

A. For pruning before training, Panning has outstanding performance at various compression ratios, with more pronounced improvements in the underfitting state.

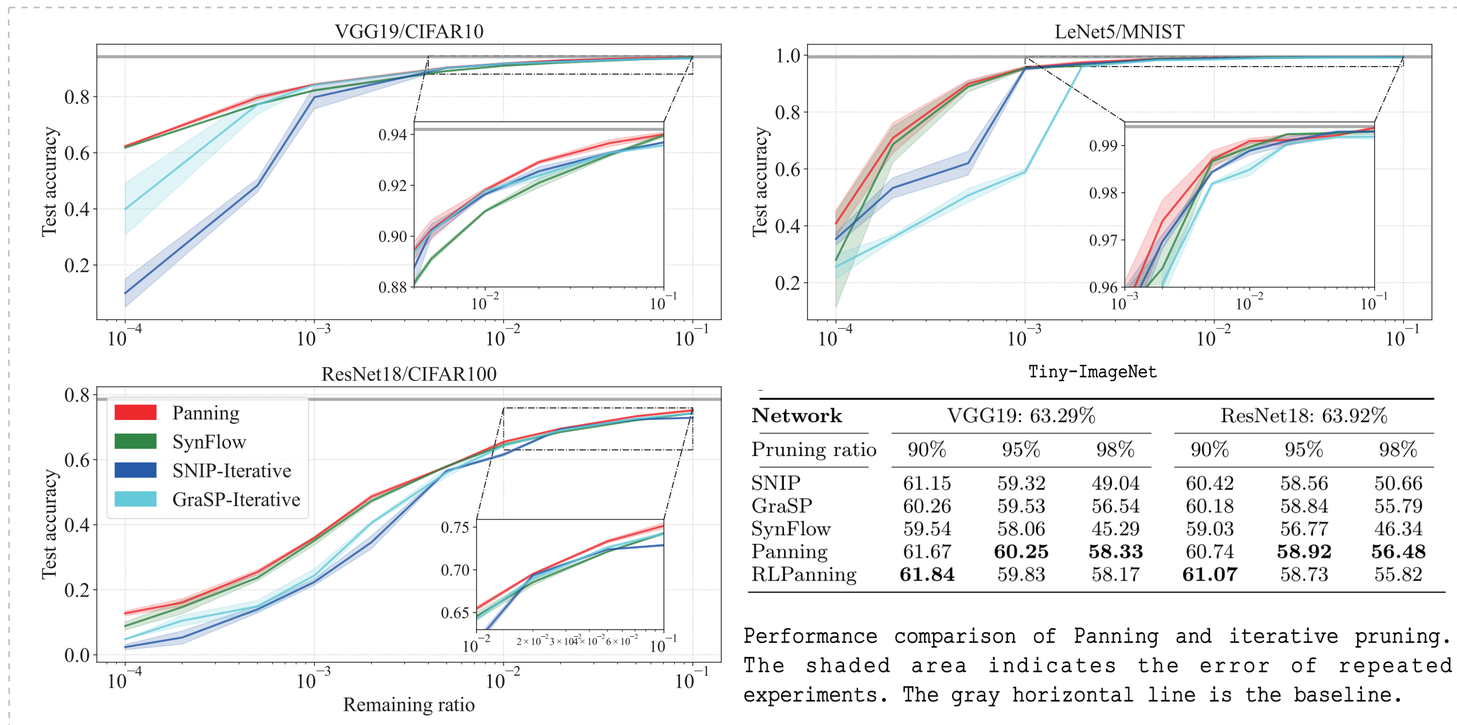
B. RLPanning can suppress the interference caused by changes in dataset size and network scale, and can more profoundly reflect the inherent nature of weight screening.

C. Weight expressive force transfer and pruning before training can provide a more essential boost to the mechanistic analysis of overparameterized networks.

Evaluate the weight expressive force in terms of network structure, loss and gradient, etc., and prune the network by RL Agent before training for better performance.

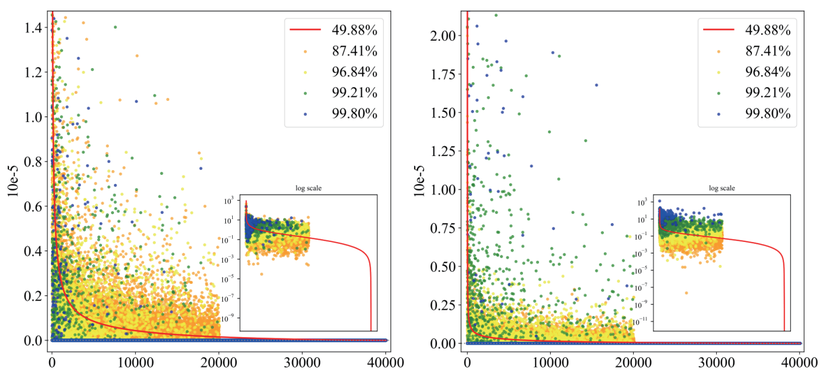


Overview of the RLPanning. On the right is the Panning environment, which iteratively prunes the network after setting the pruning rate, and the feedback includes the current state of the sparse network and the set reward. On the left is the TD3 agent, which acts through the feedback given by the environment, that is, it selects the optimal pruning strategy based on the network state during iterative pruning.



Network	VGG19: 63.29%			ResNet18: 63.92%		
Pruning ratio	90%	95%	98%	90%	95%	98%
SNIP	61.15	59.32	49.04	60.42	58.56	50.66
GraSP	60.26	59.53	56.54	60.18	58.84	55.79
SynFlow	59.54	58.06	45.29	59.03	56.77	46.34
Panning	61.67	**60.25**	**58.33**	60.74	**58.92**	**56.48**
RLPanning	**61.84**	59.83	58.17	**61.07**	58.73	55.82

Performance comparison of Panning and iterative pruning. The shaded area indicates the error of repeated experiments. The gray horizontal line is the baseline.



Comparison of loss-sensitive(left) and gradient-sensitive(right) weights at different pruning rates during iterative pruning. The subplots are the logarithmic changes in the vertical axis scale.

The weight expressive force is sensitive to the compression ratio. As the compression ratio increases, some of the less expressive weight metrics improve significantly (above the red line), and vice versa.

Affiliation:

School of Computer and Communication Engineering, Changsha University of Science and Technology

School of Computer, Xidian University

Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transp

Download Paper:

Download Poster:

GitHub Repository:

kangxiatao@gmail.com

cs.LG; cs.CV