RL-Pruner: Reinforcement Learning Meets Neural Network Compression

A revolutionary approach to lightweight AI

Professor: Vijay Eranti

By: Mohibkhan Pathan
CMPE 255: Data Mining Section 49

Introduction

Why AI Needs Optimization?

- → AI and CNNs drive innovation in industries like healthcare and autonomous driving.
- → CNNs are powerful but bulky, computationally expensive, and difficult to deploy on edge devices.

The Problem

Challenges of CNN Deployment

- → Bulky models with high computational demands.
- → Resource limitations on edge devices.
- → Need for efficient, lightweight alternatives without losing accuracy.

The Solution

Enter RL-Pruner

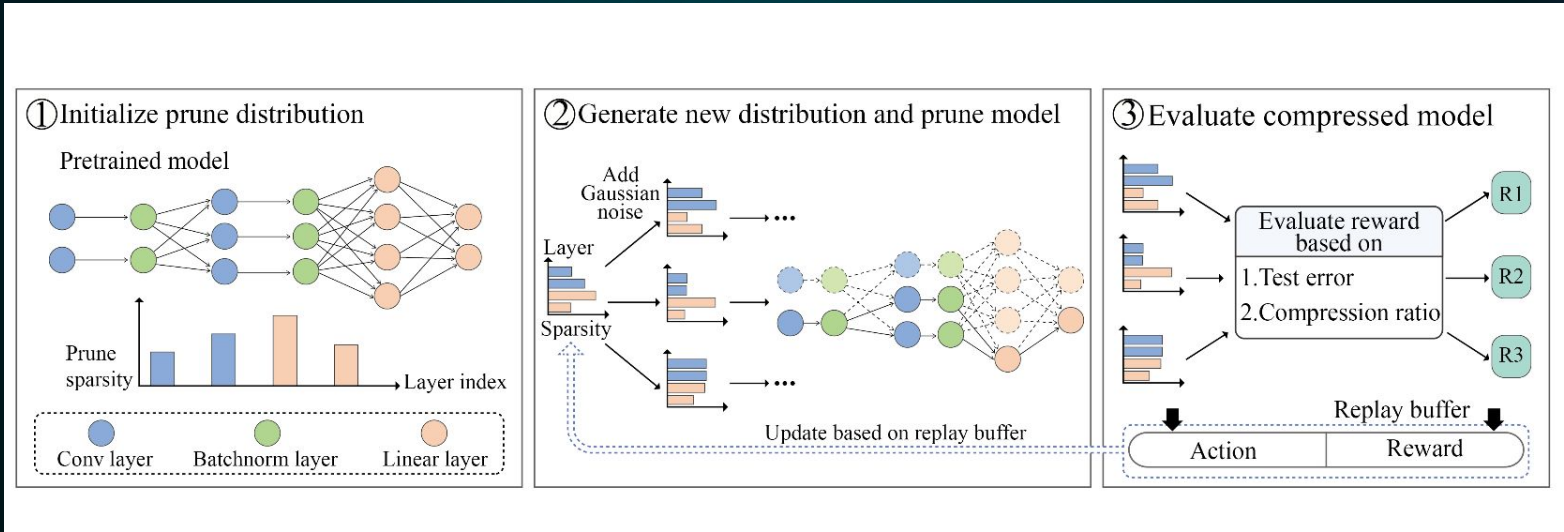
- → Uses reinforcement learning to dynamically prune unnecessary filters.
- → Creates smaller, faster CNNs with minimal accuracy loss.
- → Efficient for real-world applications like IoT, wearables, and autonomous vehicles.

How RL-Pruner Works

The Four-Step Process

- → Dependency Graph: Maps layer relationships to preserve critical connections.
- → Layer-Wise Sparsity: Dynamically adjusts pruning per layer using reinforcement learning.
- → Layer Pruning: Removes less critical filters using the Taylor criterion.
- → Knowledge Distillation: Fine-tunes the pruned model with guidance from the original model.

Working Diagram



Results That Matter

RL-Pruner’s Key Achievements

- → High Compression with Minimal Performance Loss:
  - VGG-19: 60% sparsity with <1% accuracy drop.
  - MobileNetV3-Large: 40% sparsity with minimal loss.
- → Superior to Other Methods:
  - Outperforms DepGraph and GNN-RL on accuracy and compression ratios.

Real-World Applications

Transforming Industries with RL-Pruner

- → Faster Image Recognition: Real-time processing for drones and cameras.
- → Wearables: Efficient AI for health monitoring and sensing.
- → Autonomous Vehicles: Reduces latency and ensures safer decision-making.

Conclusion

The Future of Lightweight AI

- → RL-Pruner combines reinforcement learning and structured pruning for efficient AI.
- → Enables powerful models on resource-constrained devices.
- → Opens new possibilities for sustainable and accessible AI solutions.

Thank You!