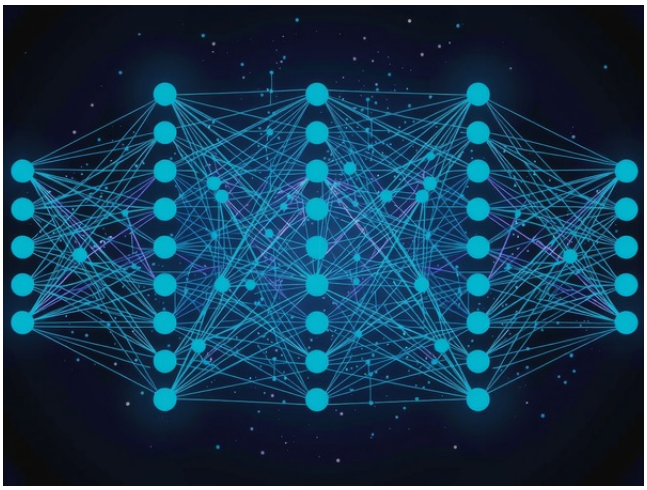
INNOVATION PARTNERSHIPS
UNIVERSITY OF MICHIGAN

Over-Paramterized Model Optimization with Polyak-Łojasiewicz Condition

TECHNOLOGY NUMBER: 2023-461



OVERVIEW

Enhances deep neural network efficiency and performance via advanced structured pruning

- Improves upon existing pruning by dynamically minimizing model condition number during training
- Image recognition, speech processing, natural language, real-time AI on constrained devices

BACKGROUND

Deep neural networks have become pivotal in computing tasks such as image and speech recognition. Historically, building high-performing models relied on increasing parameter count—so-called over-parameterization—to boost accuracy. However, large models incur significant computational and memory costs, limiting real-world adoption, especially on resource-constrained devices. Traditional pruning methods, which remove unnecessary parameters or nodes for efficiency, often fail to optimally balance model size with accuracy because they use heuristic or static criteria, sometimes sacrificing generalization and training speed. These shortcomings highlight the critical need for pruning approaches that not only reduce computational load but also expressly target the underlying factors affecting both training efficiency and test-time performance.

INNOVATION

Technology ID
2023-461

Category
Software
Software & Content

Inventor
Mingzhi Dong
Li Shang
Yixuan Chen
Yubin Shi
Qin Lv
Robert Dick

Further information
Ashwathi Iyer
ashwathi@umich.edu

Innovation Partnerships Tech Marketing Team
IPInventions@umich.edu

View online

This work introduces a dynamic structured pruning method guided by theoretical insights into over-parameterized deep models, particularly focused on minimizing the convergence and generalization gaps. It leverages the condition number—defined as the ratio of the Lipschitz constant to the Polyak-Łojasiewicz (PL) constant—as the central criterion for pruning. A novel gating network is trained alongside the main model to identify and mask suboptimal nodes during training, directly minimizing the model's condition number via an added regularization loss. This technical advance enables more precise and adaptive model pruning, preserving nodes essential for efficient learning and strong generalization. Experimental results demonstrate improved training speed and test performance across various deep network architectures, making this approach broadly applicable in fields like healthcare diagnostics, autonomous vehicles, mobile AI, and beyond.

ADDITIONAL INFORMATION

PROJECT LINKS:

DEPARTMENT/LAB:

- Robert Dick, Electrical Engineering and Computer Science (EECS)

LICENSE: