INNOVATION PARTNERSHIPS
UNIVERSITY OF MICHIGAN

CCS: Coverage-centric Coreset Selection for High Pruning Rates

TECHNOLOGY NUMBER: 2024-366

Technology ID
2024-366

Category
Software
MOSS - Michigan Open Source Support

Inventor
Atul Prakash
Haizhong Zheng

Further information
Ashwathi Iyer
ashwathi@umich.edu

View online page

OVERVIEW

Selecting highly representative training data subsets for efficient, accurate model retraining

- Balances data coverage and importance, preserving accuracy even at high data pruning rates
- Supports resource-efficient model updates, dataset distillation, and continual learning scenarios

BACKGROUND

Managing large-scale machine learning datasets is challenging, especially when retraining models for evolving tasks or constrained by limited computation. One-shot coreset selection addresses this by identifying a subset of training data that allows future models to be trained with minimal loss in accuracy. Current state-of-the-art methods rely on picking data points with the highest importance, measured by influence or informativeness metrics, and perform well when only a small fraction of data is pruned. However, they suffer a drastic accuracy drop at higher pruning rates, sometimes underperforming even random sampling. This shortfall arises because high-importance selections often disregard coverage—the need to fairly represent the distribution of the original dataset—leading to subsets that miss crucial diversity and generalizability. As machine learning applications increasingly require scalable and efficient data management, especially for large or updating datasets, improved coreset selection methods that ensure both importance and adequate coverage are in high demand.

INNOVATION

Coverage-centric Coreset Selection (CCS) is a new method that addresses the shortcomings of prior coreset selection strategies, particularly under aggressive data pruning. CCS introduces a novel metric inspired by the geometric set cover problem, extending it to capture how well a subset covers the full data distribution—ensuring the selected subset remains representative and diverse. Unlike prior methods that focus solely on importance, CCS jointly optimizes for both data coverage and importance, resulting in subsets that better maintain critical dataset characteristics. Evaluated on five datasets, CCS demonstrates remarkable resilience at high pruning rates, achieving up to 19.56% higher accuracy than existing methods on scenarios like 90% pruning for CIFAR10, and consistently outperforming random selection while matching top performance at lower pruning rates. Practical applications include efficient model retraining, continual learning, scalable dataset distillation, and resource-constrained AI deployments where data efficiency and robustness are paramount.

ADDITIONAL INFORMATION

PROJECT LINKS:

- CCS Github

DEPARTMENT/LAB:

- Atul Prakash, Computer Science and Engineering (CSE)

LICENSE: