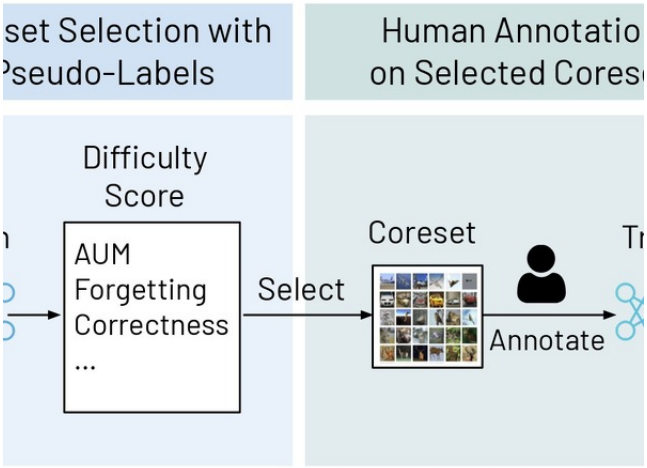
INNOVATION PARTNERSHIPS
UNIVERSITY OF MICHIGAN

ELFS: Enhancing Label-Free Coreset Selection via Clustering-based Pseudo-Labeling

TECHNOLOGY NUMBER: 2024-568



Technology ID
2024-568

Category
Software
MOSS - Michigan Open Source Support

Inventor
Yifu Lu
Atul Prakash
Elisa Tsai
Haizhong Zheng

Further information
Ashwathi Iyer
ashwathi@umich.edu

View online page

OVERVIEW

Label-free data selection method improves deep learning efficiency and reduces annotation costs

- Outperforms existing label-free approaches by estimating data importance without ground-truth labels
- Image classification, medical imaging, autonomous vehicles, large-scale dataset curation

BACKGROUND

Deep learning models owe much of their success to large, high-quality, human-annotated datasets. However, the annotation process is expensive and requires significant time and labor, especially for vast collections of data such as images. Selecting the most informative data points for annotation—known as coreset selection—can help maximize learning efficiency within a strict labeling budget. While state-of-the-art supervised coreset methods are effective, they rely on ground-truth labels for the entire dataset, which defeats the purpose of reducing annotation costs. Conversely, existing label-free coreset selection approaches struggle due to poor data representativeness and overly simplistic scoring. These limitations underscore the critical need for an accurate, label-free selection method that can prioritize data for annotation while maintaining strong downstream model performance.

INNOVATION

ELFS is a novel, label-free coreset selection method that addresses the shortcomings of existing approaches by leveraging deep clustering to estimate data difficulty without requiring ground-truth labels. ELFS introduces a double-end pruning mechanism, effectively mitigating biases in difficulty score calculations and resulting in the selection of more diverse and informative data subsets. Experiments on five visual benchmarks demonstrate its consistent superiority over other label-free baselines—sometimes even matching the performance of supervised coreset methods at moderate data reductions. This technical advance means that users can strategically annotate the most valuable examples, minimizing unnecessary labor and cost. ELFS is particularly applicable in areas like vision-based AI model training, medical image analysis, automated driving systems, and large-scale data curation.

ADDITIONAL INFORMATION

PROJECT LINKS:

DEPARTMENT/LAB:

- Atul Prakash, Computer Science and Engineering

LICENSE: