MLKD 2024
The First International Conference on
Machine Learning and Knowledge Discovery
Amirkabir University of Technology, December 18-19, 2024

Optimizing Data Curation through Spectral Analysis and Joint Batch Selection (SALN)

Mohammadreza Sharifi*

Abstract

In modern deep learning models, long training times and large datasets present significant challenges to both efficiency and scalability. Effective data curation and sample selection are crucial for optimizing the training process of deep neural networks. This paper introduces SALN, a method designed to prioritize and select samples within each batch rather than from the entire dataset. By utilizing jointly selected batches, SALN enhances training efficiency compared to independent batch selection. The proposed method applies a spectral analysis-based heuristic to identify the most informative data points within each batch, improving both training speed and accuracy. The SALN algorithm significantly reduces training time and enhances accuracy when compared to traditional batch prioritization or standard training procedures. It demonstrates up to an 8x reduction in training time and up to a 5% increase in accuracy over standard training methods. Moreover, SALN achieves better performance and shorter training times compared to Google's JEST method developed by DeepMind. The code and Jupyter notebooks are available at `github.com/rezasharifi82/SALN`.

Keywords: Spectral Analysis, Data Curation, Joint Batch Selection

*Department of Computer Science, Ferdowsi University of Mashhad, `sharifi.mohammadreza@mail.um.ac.ir`