Training Data Selection Methods for Deep Learning Based on Diversity-Aware Heuristic Strategies

刘子宁 雷纯熙 张蔚峻 伍思亦
2401110050 2401110045 2401110059 2401110071

1 Introduction

Efficient training of deep learning models often requires selective data sampling to accelerate convergence and enhance generalization. This research investigates data selection methods informed by heuristic strategies, which leverage model-specific metrics to prioritize training samples. The objective is to explore how different selection strategies impact model performance and efficiency in large-scale training tasks.

2 Problem Setting

Given a training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ and a model $f_\theta(x)$ parameterized by θ, the loss function $L(\theta, x, y)$ is used to measure the discrepancy between predictions and true labels. The problem can be defined as:

$$\min_{\theta} \frac{1}{|\mathcal{B}|} \sum_{(x,y)\in\mathcal{B}} L(\theta, x, y),$$

where $\mathcal{B} \subseteq \mathcal{D}$ is the selected batch of training samples. The challenge is to construct $\mathcal{B}$ based on criteria that optimize training efficiency while ensuring diversity in the selected data.

Key considerations:

- Samples should be selected to maximize their contribution to the training process.
- The selection strategy must balance focus on challenging examples and coverage of the dataset's diversity.

3 Methods

We propose to evaluate the following training data selection strategies:

1. **High-Loss Selection**[1]: Samples are chosen based on their loss values $L(\theta, x, y)$. Let $\mathcal{B}$ be the top k samples ranked by $L(\theta, x, y)$.
2. **High-Gradient Selection**[2]: Samples are prioritized based on the norm of the gradient of the loss with respect to model parameters:

$$\|\nabla_\theta L(\theta, x, y)\|.$$

3. **High-Influence Selection**[3]: Samples are selected based on the influence of the input on the loss, calculated as:
$$\left\| \frac{\partial L(\theta, x, y)}{\partial x} \right\|.$$
4. **Random Sampling (Baseline)**: Samples are randomly selected from $\mathcal{D}$ without considering loss or gradient metrics.

4 Experimental Setup

- **Datasets**: CIFAR-10 and MNIST for image classification tasks.
- **Model Architectures**: CNNs with varying depths and complexity.
- **Evaluation Metrics**:
  - Final model accuracy on test data.
  - Computational cost: Time and resources required for training.

5 Expected Outcomes

- High-Loss and High-Gradient Selection are expected to accelerate convergence by focusing on challenging examples.
- High-Influence Selection may improve generalization by emphasizing samples that shape the loss landscape.
- Random Sampling serves as a baseline to quantify the benefits of heuristic strategies.

References

[1] Angela H Jiang et al. "Accelerating deep learning by focusing on the biggest losers". In: *arXiv preprint arXiv:1910.00762* (2019).

[2] Krishnateja Killamsetty et al. "Glister: Generalization based data subset selection for efficient and robust learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 9. 2021, pp. 8110–8118.

[3] Garima Pruthi et al. "Estimating training data influence by tracing gradient descent". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 19920–19930.