خلاصه مقاله

"Coverage-centric coreset selection for high pruning rates"

* We first propose a novel metric to measure the coverage of a dataset on a specific distribution by extending the classical geometric set cover problem to a distribution cover problem.

* This metric helps explain why coresets selected by SOTA methods at high pruning rates perform poorly compared to random sampling because of worse data coverage.

* Different from SOTA methods that prune unimportant (easy) examples first, CSS is inspired by stratified sampling and guarantees the sampling budget across importance scores to achieve better coverage at high pruning rates.

★ To compare the data coverage of different coreset methods, we need to quantitatively measure how well a dataset S covers a distribution P. In classical geometric set cover setting, we say a set S' is a r-cover of another set S, when a set of r-radius balls centered at each element in S' covers the entire S. The radius r can be used as a metric to measure coverage of S' on S

★ To measure how well a set covers a distribution, we extend the classical geometric set cover to the density-based distribution cover. Instead of covering a set, we study the covering on a distribution P_μ and consider probability density in different areas of the input space. Instead of taking a complete cover, we introduce the cover percentage p to describe how a set covers different percentages of a distribution to better understand the trade-off between cover radius r and cover percentage p

* To make a more obviously quantitative comparison, we propose to use area under the p-r curve, AUC_{pr}, as a proxy metric to assess the quality of a coreset selection strategy.

* lower values of AUC_{pr} suggest better coverage by the coreset. we note that AUC_{pr} is the expected minimum distance between examples following the underlying distribution P_μ to those in the coreset, as stated

in the following proposition.

* In practice, we can assess AUC_{pr} with test set ~~D_{test}: $AUC_{pr}(S)$=~~

$$AUC_{pr}(S) = \underset{x \in D_{test}}{E} \left[\min_{x' \in S} d(x', x) \right]. \quad \cancel{\underset{x \in D_{test}}{E}}$$

* If the sampling budget is limited, easy examples in the high-density area provide more coverage than hard examples in the low-density area.

* Compared to SOTA methods, CSS still assigns the sampling budget to the high-density area containing easy examples at high-pruning rates, which provide larger coverage to the underlying distribution

* Compared to random sampling, CSS assigns a larger sampling budget to the low-density area, where hard examples are informative for training.

* CSS first divides the dataset into different non-overlapping strata based on importance scores.

* Each stratum has a fixed-length score range, but may include different numbers of examples

* we fix an initial budget on the number of examples to be chosen from each strata, based on the desired pruning rate, but, if a stratum has fewer examples than the budget, remaining budget is evenly assigned to other strata.

* with a low pruning rate, stratified sampling tends to first discard data from low-importance strata.

* Before CCS assigns budgets to strata, CCS prunes β percent "hard" examples based on importance scores. This is based on two insights: 1 Mislabeled examples often also have higher importance scores but do not benefit accuracy.

2 when data is really scarce, we need more data from higher-density areas to get better coverage.

* In practice, we use grid search to find β as a hyperparameter.

* We combine CCS with AUM score.