Active learning paper survey-CSE584

Varunsai Alaparthi
vsa5067@psu.edu

1.PT4AL: Using Self-Supervised Pretext Tasks for Active Learning

Motivation: The authors of this paper are mainly dealing with the problem of labeling data efficiently. Especially in the context of deep learning and machine learning. The task of labeling, especially for huge amounts of data, is costly and time-consuming, and already existing active learning methods are struggling with selecting the most informative & impact creating samples without relying heavily on some amount of labeled data. Additionally, these methods often face cold start issues, where the initial random selection of labeled samples can significantly impact overall performance. The authors aim to address respective issues by experimenting with a novel active learning approach that leverages self-supervised pretext tasks to improve data sample selection.

Solution: The authors propose Pretext Tasks for Active Learning (PT4AL), a novel framework that uses self-supervised learning to drive the sample selection for labeling. In PT4AL, a pretext task ,like predicting image rotation, is first trained on the unlabeled dataset, and the resulting loss is used to rank the samples. These samples are then divided into batches, and the most uncertain samples from each batch are selected for labeling. This approach allows for the sampling of both representative and difficult data points.This solution overall integrates uncertainty sampling with self-supervised learning, ensuring that the labeled set is diverse and informative.

Contribution:

1. **Self-supervised Pretext Tasks**: PT4AL introduces the use of pretext tasks to estimate the informativeness of data points, which improves both the diversity and difficulty of the selected samples.
1. **Batch Split and In-batch Sampling**: The unlabeled dataset is split into batches based on the pretext task loss. Samples are then selected using a combination of batch-based sampling and uncertainty sampling, giving a balance in diversity and difficulty.
2. **Cold-start Solution**: This approach solves the cold start problem by using pretext task losses to rank and select the most informative samples in the first iteration, without relying on random sampling

Cons/Improvements:

1. **Pretext Task Dependence:** The performance of PT4AL is highly dependent on the pretext task selected and certain tasks like contrastive learning may not correlate well with downstream performance, which makes it necessary to tailor the pretext task to the specific problem domain.
2. **Computational Overhead:** While the solution performs better in sample selection, the pretext task training adds an extra computational cost compared to simpler methods like random sampling. However, this overhead is compensated by better performance in the long run.

In conclusion this paper represents a significant advancement in active learning by incorporating self-supervised pretext tasks to guide sample selection. It far outperforms existing methods in terms of accuracy and addresses key issues like the cold start problem and redundant sampling.

2. Compute Efficient Active Learning

Motivation: The authors address a major issue in active learning: even though active learning methods aim to reduce labeling costs, they often require significant computational resources when dealing with very huge datasets. This limitation chokes scalability, making it difficult to implement in real-world scenarios where computation costs are high. The motivation behind this paper is to propose an efficient active learning method that reduces computational demands while maintaining the effectiveness of the model on large datasets.

Solution:

The authors propose a method-agnostic active learning framework designed to optimize sample selection in a computationally efficient manner. The central idea is that historical acquisition function values can predict future acquisition values, allowing for informed subsampling. The process involves selecting a small candidate pool from the entire unlabeled dataset based on past acquisition function values, evaluating the acquisition function only on this smaller pool, and then selecting the most informative samples for labeling. This reduces the number of samples needing evaluation at each turn , hugely reducing computational overhead.

Novelty/Contribution:

- **Subsampling Strategy:** This paper introduces a novel subsampling approach, which is driven by acquisition function history. Rather than recalculating acquisition values for all unlabeled samples, the system samples a smaller subset based on past data, ensuring that the computational load is less in comparison.
- **Flexible Integration:** The framework is flexible and can work with various acquisition functions, such as entropy and variation ratios. Additionally, it can be applied to both classification and regression tasks.

- **Computational Efficiency:** The proposed method significantly reduces compute load. For instance, in experiments on the CIFAR-10 dataset, the proposed method reduced compute time by up to 25% compared to baseline models while still giving improved accuracy.

Cons/Improvements:

- **Limited Exploration on Large Datasets:** While the method shows computational efficiency in medium-scale datasets like CIFAR-10, further testing is needed on larger datasets to validate its scalability.
- **Subsampling Limitation:** The subsampling technique, while efficient, risks missing some informative data points if the initial candidate pool is not diverse enough. The sampling may need further fine-tuning for more complex datasets.

Final Conclusion:

The framework introduced in this paper demonstrates a promising approach to reducing computational costs in active learning. It leverages historical acquisition function values to create a smaller candidate pool, focusing computational resources on the most important samples. Future work could further validate this method on larger, more diverse datasets

3. Active learning by Feature Mixing

Motivation: We know that Active learning aims to improve model performance by selecting the most informative data points for labeling,But traditional methods often face challenges such as high computational costs and the trade-off between selecting uncertain samples (near decision boundaries) and maintaining diversity in the dataset. Solving this is the main motivation behind the paper (ALFA-Mix) which is to enhance active learning by improving this trade-off, reducing redundancy, and selecting a diverse set of data points efficiently through feature mixing.

Solution:

The ALFA-Mix method introduces an interpolation-based approach that identifies the most informative samples for labeling by mixing features of labeled and unlabeled points in the latent space. Instead of relying solely on uncertainty or diversity-based sampling, the method computes a closed-form solution for feature interpolation between labeled and unlabeled data. The key idea is to use the gradients of unlabelled instances to identify points whose predictions would change most significantly when interpolated with labeled points. This allows ALFA-Mix to prioritize selecting both informative and diverse samples while reducing computational overhead by avoiding iterative optimization.

Novelty/Contribution:

- **Closed-form Optimization:** One of the major contributions of ALFA-Mix is its use of a closed-form solution for feature interpolation, which greatly reduces the computational complexity compared to gradient-based optimization approaches. This allows the framework to be faster and more scalable.
- **Balancing Uncertainty and Diversity:** The method effectively balances the selection of uncertain and diverse samples, ensuring that the labeled set covers both decision boundaries and a broad representation of the feature space. This leads to better model generalization, especially in small-budget scenarios.

Cons/Improvements:

- **Suboptimal in Some Cases**: While the closed-form solution is efficient, it might not always find the most optimal interpolation in cases where more fine-grained adjustments to the interpolation ratio are needed. In certain situations, iterative optimization methods might yield better results, albeit at a higher computational cost.
- **Application Beyond Vision Tasks**: ALFA-Mix is mainly tested on image classification tasks , and while the results are promising, more exploration is needed to validate its effectiveness across other tasks such as natural language processing or object detection or extending the method to image generation.

Final Conclusion:

In conclusion , ALFA-Mix provides a highly efficient and effective solution to active learning by introducing a closed-form feature mixing approach that combines the pros of uncertainty sampling and diversity-based sampling. The method consistently outperforms existing active learning techniques such as BADGE and CoreSet, especially in conditions with small labeling budgets. Its computational efficiency and versatility make it a practical solution for real-world active learning tasks, Although there's scope for future work to explore applications beyond image classification.