CSE 584: Machine Learning: Tools and Algorithms

Assignment-1

Submitted by: Lakshmi Chandrika Yarlagadda (lvy5215)

Submission date: 15th September 2024

Paper1: Active-Learning-as-a-Service: An Automatic and Efficient MLOps System for Data-Centric AI - Yizheng Huang, Huaizheng Zhang, Yuanming Li, Chiew Tong Lau, Yang You

Problem:

This paper addresses the problem that the use of Active Learning in real-world machine learning workflows, especially within a data-centric AI setting, is inefficient and complex. AL is critical as it reduces the need for manually labeling large amounts of data while keeping the performance of the model intact. However, several challenges concern the application of AL:

- Current AL tools require users to manually select which AL strategy is appropriate for their task. This might not be a trivial decision, mostly because different strategies work best under different conditions. Users in such instances are most often left with trial and error to identify the right strategy, wasting valuable time and resources.
- Active learning means running different models and selecting the most informative points to label. This can be computationally expensive, considering cases where large datasets are dealt with.
- AL requires the implementation of back-end systems, integrating machine learning models with data storage and infrastructure which might involve a lot of boilerplate code, thus, engineering costs are very high.

Solution:

Active Learning as a Service (ALaaS) simplifies this often-complex, active learning process with automation of its key steps and makes it both faster and more user-friendly. The system is initiated by users providing basic inputs such as their dataset, budget, and desired accuracy. The system then takes over, using a smart agent-called the **Predictive-based Successive Halving Early-Stop (PSHEA)**. This agent automatically chooses the best active learning strategy by testing in parallel multiple approaches that are avaiblae in the AI zoo, predicting their outcomes, and discarding the underperforming approaches as early as possible, to speed up the selection process. ALaaS is built on a flexible server-client architecture, which means it can easily scale across various devices and cloud environments, handling large amounts of data without any issues.

To enhance this efficiency, the system further breaks down the workflow into various stages like downloading data, pre-processing of data, and selection of the most useful data to label. The system then runs these concurrently, which in essence is **stage-level parallelism**. It also utilizes techniques, like data caching and batching, to reduce the wait times while utilizing these resources at full capacity. By providing automation for decision-making, optimizations of performance, along with a minimalist user interface, ALaaS casts the normally complex active learning pipeline into a smooth, end-to-end experience accessible to diverse users, from experts to non-expert users.

Novelties:

- PSHEA automatically chooses the best strategy of AL on predictive accuracy and available budget to greatly reduce the requirement of human-setting configuration.
- The system uses stage-level parallelism, data caching, and batching to accelerate the AL process, making it 10 times faster opposed to classic methods.
- It adopts a service-based approach, reducing the complexity of applying AL in real-world applications. The deployment of ALaaS can be easily done on cloud as well as on the local system with minimal configuration.
- ALaaS is open-sourced, modular, and it can be easily adapted to be integrated into most other machine learning workflows.

- Non-expert users can easily engage with the system, reducing the barrier to entry for active learning applications.

Downsides:

- Even though ALaaS improves efficiency, the performance on very large datasets could still be bounded by server-client architecture and network bandwidth limitations, especially in a distributed environment.
- While the system automates many processes, it still requires users to provide accurate budget and target accuracy inputs, which may be challenging for those unfamiliar with the domain.
- While the system automates the selection of an active learning strategy, it gives very limited insights into why one strategy has been chosen over another; which might be useful in further optimization.
- While the system supports various AL strategies, users may find limitations in customizing these strategies beyond what is provided in the AL strategy zoo.

Paper2: ACTIVERAG: Revealing the Treasures of Knowledge via Active Learning - Zhipeng Xu, Zhenghao Liu, Yibin Liu, Chenyan Xiong, Yukun Yan, Shuo Wang, Shi Yu, Zhiyuan Liu, Ge Yu

Problem:

The paper addresses the limitations of current RAG models. Traditional RAG models regard LLMs as passive receivers of external knowledge, which seriously limits their ability to deeply understand and learn from that knowledge. This has brought about a series of challenges such as hallucinations and outdated memories making it hard for LLMs to deliver reliable answers on knowledge-intensive tasks. The motivation for this lies in moving away from this passive learning towards active learning paradigms where the LLM constructs and solidifies knowledge by connecting the information retrieved with previously acquired knowledge.

Solution:

ACTIVERAG enhances RAG models by enabling active engagement with retrieved knowledge, allowing language models to process and integrate information deeply. This is done in three stages:

1. In the Retrieval Stage, ACTIVERAG begins by retrieving relevant passages, this stage is similar to the normal RAG models.

2. The Knowledge Construction Stage is ACTIVERAG's key innovation, where the LLM actively engages with the retrieved data, linking it to prior knowledge and building a deeper understanding. Four agents guide this process:

- Semantic Association - connects new information with existing knowledge.
- Epistemic Anchoring - helps process unfamiliar concepts.
- Logical Reasoning - improves problem-solving through structured knowledge.
- Cognitive Alignment - prevents errors by aligning new data with known facts.

3. The Cognitive Nexus Stage integrates this newly constructed knowledge with the LLM's reasoning, allowing it to refine and correct its output, ensuring more reliable, informed answers.

Novelties:

- Introduces an active learning paradigm for LLMs, allowing them to construct knowledge actively rather than passively receiving it.
- The different kinds of agents developed in the model for constructing knowledge included Semantic Association, Epistemic Anchoring, Logical Reasoning, and Cognitive Alignment, which helped in developing a better understanding of the model from multiple viewpoints.
- It integrates knowledge construction outcomes into the process of reasoning within the LLM to enhance the model's ability in providing correct answers.
- Improved performance by more than 5% on question-answering datasets from the current RAG models.

Downsides:

- The model will require multiple calls to the API for an initial chain-of-thought generation, processing knowledge construction, and the execution of the cognitive nexus. This may result in increased time latency and heightened costs by calling the API more frequently.

- Including long retrieved passages and knowledge construction results can make the input to LLMs very long, possibly creating processing complications that may lower performance.
- The performance of ActiveRAG heavily depends on the passages retrieved. Any noisy or irrelevant information upon retrieval may affect performance in an adverse manner.

Paper3: Class-Balanced Active Learning for Image Classification - Javad Zolfaghari Bengar, Joost van de Weijer, Laura Lopez Fuentes, Bogdan Raducanu

Problem:
The paper addresses active learning in the context of imbalanced datasets, specifically the problem of the long-tail distribution commonly experienced in real-world datasets. This is driven by the observation that most current research in active learning focuses on balanced datasets, while real world datasets often tend to be imbalanced, resulting in suboptimal classifier performance. This forces active learning algorithms to oversample the majority classes and under sample the minority classes, which affects the performance of classifiers trained on these data.

Solution:
The authors address the problem of class imbalance in active learning by proposing an approach that makes sure the selected samples will not only be informative but also aim for a more balanced class distribution. The key problem they deal with here is the absence of true class labels among the pool of unlabeled samples.

To solve this, the class distribution is estimated using the predicted probabilities from the model's softmax output. A matrix is created to show how likely each sample belongs to different classes, and the difference between the current and desired class balance is measured. This helps guide the selection process to pick samples that will create a more balanced class distribution, with a vector Ω(c)\Omega(c)Ω(c) showing how many samples are needed from each class.

In addition to balancing, the authors use entropy to measure uncertainty, with higher entropy indicating more informative samples. These uncertain samples can improve the model's performance. They combine the goals of selecting the most uncertain samples and maintaining class balance in one optimization problem. This can be formulated as minimizing $\mathbf{z \in [zT\ P \odot log(P)1C \times 1 + \lambda\ \ell1(\Omega(c) - Pz)],}$ where z is a binary indicator vector that selects which samples are ultimately chosen, and λ is the regularization parameter that trades off informativeness against class balance.

The first part of the cost function maximizes the entropy of selected samples to ensure they are informative, while the second part minimizes deviation from the desired class balance. This optimization ensures each batch is both informative and balanced. The regularization parameter \(\lambda\) allows adjusting the focus between class balance and informativeness, based on the dataset and active learning needs.

Novelties:
- A new active learning approach for class balancing in sample selection to deal with active learning issues caused by imbalanced datasets.
- The framework is shown to improve performance not only on imbalanced datasets but also on balanced datasets by mitigating the sampling bias introduced by active learning.
- Developed an optimization method that can be combined with existing active learning strategies, making it versatile and applicable to various scenarios.

Downsides:

- The proposed optimization problem, although effective, has extra computational complexity than that of simpler active learning methods. This can make the approach not quite practical on large-scale or real-time applications, where computational efficiency is a matter of concern.
- The method relies on an initial labeled dataset to kickstart the active learning process, which may not always be available or may be limited in size.
- The method requires fine-tuning of the regularization parameter (λ) that controls the trade-off between informativeness and class balance, which might not be straightforward for different datasets or applications.