Name: Aishwarya Kulkarni
ID: 924430258

CSE 584: Machine Learning – Tools and Algorithms

Homework – 1

Paper 1: ACTIVE LEARNING FOR NEURAL PDE SOLVERS
arXiv:2408.01536v1 [cs.LG] 2 Aug 2024, 2408.01536v1 (arxiv.org)

1. **What problem does this paper try to solve, i.e., its motivation?**
   The motivation of this paper is to solve Partial Differential Equations. The issue with numerical solvers is that they need high temporal and spatial resolution to get good accuracy, but in turn lead to higher computational costs. Hence, it makes more sense to replace the numerical simulators with a more intelligent neural network model to predict the simulators outputs. Another advantage of this is that NNs are end-to-end differentiable. The authors claim that while neural PDE solvers are the obvious choice, they require huge amount of training data from the same simulator they're to replace, which is expensive! Instead, Active Learning is another option which could help reach the same accuracy using smaller training datasets by querying classical solvers with more informative initial conditions and PDE parameters thereby reducing the simulations to reach the same accuracy. But the authors also digress that there Is much left to be studied in Active Learning for neural PDE solvers.

2. **How does it solve the problem?**
   The authors introduce AL4PDE, a modular and extensible active learning benchmark. It provides a comprehensive framework that integrates multiple parametric PDEs, state-of-the-art surrogate models, and active learning methods within a solver-in-the-loop setting. By focusing on AL methods such as uncertainty- and feature-based approaches, the benchmark demonstrates how these methods can significantly enhance model performance in terms of error reduction and data efficiency. The AL4PDE benchmark is structured into three main components: (1) AL algorithms for selecting data points, (2) surrogate models that approximate the PDE solutions, and (3) the PDEs and corresponding simulators used for generating the training data. This modular design ensures flexibility, enabling users to easily add new AL approaches, surrogate models, or PDE problems to the framework. The authors use AL4PDE to investigate several key aspects, including the impact of different AL methods on average error, error distribution, and data reusability, along with performing an ablation study to understand design choices in batch active learning algorithms. The authors show that AL reduces the average error by up to 71% compared to random sampling, significantly reducing worst-case errors. Importantly, the datasets generated by AL are consistent across runs, producing similar distributions of PDE parameters and initial conditions. This consistency enhances the reusability of acquired datasets, benefiting other surrogate models not directly involved in data generation. Additionally, the benchmark reveals a temporal advantage of AL, indicating that active learning can accelerate the training process while maintaining accuracy. Overall, AL4PDE provides a robust and extensible platform for improving and evaluating active learning techniques in PDE solving.

3. **A list of novelties/contributions.**
   - AL4PDE is the first active learning (AL) framework for neural PDE solvers.
   - Supports the study of existing AL algorithms and development of new PDE-specific AL methods.
   - Includes various AL algorithms, differentiable simulators for PDEs (e.g., compressible Navier-Stokes), and neural surrogate models (e.g., U-Net).
   - Extensible for adding new algorithms, models, and tasks.
   - Experiments show that AL improves data efficiency and reduces worst-case errors.
   - LCMD and SBAL are identified as the top-performing AL algorithms.

- AL results in more accurate surrogate models trained faster.
- Data distribution remains consistent across random repetitions, initial datasets, and models, demonstrating reliable, reusable datasets.

4. **What do you think are the downsides of the work?**
   I think the AL4PDE might not be generalized to real-world applications as it may have limited considerations of complex PDEs. The results might not be accurate if the surrogate models have poor quality. Also, this framework might be complex to understand for new users.

Paper 2: Active Learning on a Budget: Opposite Strategies Suit High and Low Budgets arXiv:2202.02794v4 [cs.LG] 16 Jun 2022, Active Learning on a Budget: Opposite Strategies Suit High and Low Budgets (arxiv.org)

1. **What problem does this paper try to solve, i.e., its motivation?**
   The main purpose of the paper is to prove that the existing sampling methods in Active Learning (QBC, Uncertainty-Based, etc.) are best suitable when the budget is large. By budget, the authors mean number of labelled samples. The authors argue that many traditional active learning approaches are based on uncertainty sampling or diversity-based sampling, or a combination of them, which are known to require a large initial set of labelled examples. The authors call this a high budget regime. In a low budget regime, where the initial labelled dataset is small or absent at times, the authors present a study that shows random selection outperforms most deep AL strategies. A "cold start" problem is also explained as the poor ability of neural network models to capture the uncertainty which is often severe in case of a small budget regime. Overcoming this phenomenon is the motivation of this paper. The authors suggest that low and high budget regimes require the opposite querying strategies and that uncertainty sampling methods only support high budget regimes while the opposite, the selection of the least ambivalent points is suitable for low budget regime.

2. **How does it solve the problem?**
   The authors propose a new strategy – TypiClust, or Typical Clustering. It is a strategy for low budget active learning regimes. Firstly, they establish theoretical foundations that support their claim that uncertainty sampling is suitable only in high budget regimes and an opposite sampling strategy is suitable for low budget regimes. The authors observe a phase-transition-like phenomenon: in the low budget regime, over-sampling the "easier" region, which can be learned from fewer examples, improves the outcome of learning, whereas in the high-budget regime, over-sampling the alternative region is more beneficial. However, estimating prediction certainty is difficult, and cannot be reliably accomplished in the low-budget regime with access to very few labelled examples. Hence, the authors try replacing the notion of certainty with the notion of typicality in which a point is typical if it lies in a high-density region of the input space, irrespective of labels. TypiClust aims to pick a diverse set of typical examples, which are likely to be representative of the entire dataset. To this end, TypiClust employs self-supervised representation learning and then estimates each point's density in this representation. Diversity is obtained by clustering the dataset and sampling the densest point from each cluster. On comparing TypiClust to various other AL strategies in the low budget regime, it consistently improves generalization by a large margin, reaching the SOTA results in many problems. This new strategy is more beneficial for semi-supervised learning.

3. **A list of novelties/contributions**
   A list of contributions by the authors is:
   i. A novel theoretical model analyzing Active Learning (AL) as biased sampling strategies in a mixture model.
   ii. Prediction of the cold start phenomenon in AL.
   iii. Prediction that opposite strategies suit AL in the low budget and high-budget regimes.
   iv. Empirical support of these theoretical principles.

v. TypiClust, a novel strategy that significantly improves active learning in the low-budget regime.
vi. Large performance boost to SOTA semi-supervised methods by TypiClust.

4. **What do you think are the downsides of the work?**
Some downsides I can think of are that TypiClust relies heavily on semi-supervised learning and its ability to extract meaningful features from the unlabelled dataset. I also think that this adds some level of complexity to the strategy. There could also be some issues of overfitting in the early stages, even though the strategy aims to overcome it. There might also be a few struggles with noisy data.

Paper 3: ALiPy: Active Learning in Python
arXiv:1901.03802v1 [cs.LG] 12 Jan 2019, 1901.03802v1 (arxiv.org)

1. **What problem does this paper try to solve, i.e., its motivation?**
The author's motivation is to reduce the labelling cost of data used in supervised machine learning methods, which need a huge set of labelled data for training. This task is usually very expensive. There is limited labelled data but abundant unlabelled data in the real-world. A solution to this is Active Learning which reduces the labelling cost by iteratively selecting the most valuable data to query their labels from a human user.

2. **How does it solve the problem?**
The authors introduce a python toolbox named ALiPy. It is designed to simplify the implementation, evaluation, and analysis of active learning (AL) methods. It offers a modular framework, allowing users to conveniently manage various components of active learning, such as data processing, active selection, label querying, and result visualization. The toolbox comes with more than 20 state-of-the-art AL algorithms, along with tools for configuring and creating custom approaches tailored to specific needs, like multi-label data, noisy annotators, or varying query costs. With comprehensive documentation and an open-source repository on GitHub, ALiPy is easy to install via PyPI and supports a wide range of active learning scenarios.

3. **A list of novelties/contributions**
→ Modular Design: ALiPy decomposes the active learning process into independent components, allowing users to modify or replace any part of the framework without constraints.
→ Wide Range of AL Strategies: It provides implementations of more than 20 active learning algorithms, offering users many options for experimentation.
→ Support for Novel AL Settings: The toolbox supports multi-label data, noisy annotators, and cost-sensitive settings, expanding the range of possible use cases.
→ Flexibility: Users can experiment with different data splitting strategies, oracles, and evaluation criteria. The toolbox supports parallelization for running multiple experiments efficiently.
→ Easy Usability for Different Users: ALiPy accommodates users of varying expertise, from those unfamiliar with active learning to advanced researchers wanting to develop new methods. It provides simple interfaces and detailed documentation.

4. **What do you think are the downsides of the work?**
Although ALiPy supports multi-label and noisy data, it seems to be specific for typical classification problems. The framework's efficiency maybe less with huge datasets. It relies heavily on other python libraries.