CMPSC 584 – Fall 2024 – Homework 1

Sandhya Somasundaram
939157171
sms9723@psu.edu

September 15 2024, Sunday

1 Learning Loss for Active Learning

- Donggeun Yoo and In So Kweon

Link to paper: https://arxiv.org/abs/1905.03677v1

Motivation:

Many real world tasks - like object detection, medical imaging, human pose estimation - now resort to solutions using deep neural networks. As mentioned in the paper, the performance of these neural networks improves drastically with more annotated data. But, the cost of annotating data is high and in the use cases such as analysis of medical images, a good expertise is required to do the annotations. This depicts the data hungry nature of neural networks and how annotated data improves performance. Active learning comes picture where it tries to reduce the number of data points that need labelling while still maintaining the high performance. In the traditional approach of active learning, we select the most uncertain samples for labelling. However, these methods might be task-specific or task specific when used in combination with the large neural networks. This paper, "Learning Loss for Active Learning" by Donggeun Yoo and So Kweon tackles this very problem and tries to create an efficient and task-agnostic active learning method that works well with deep networks without having to make expensive task-specific adjustments. Hence, it reduces the cost of annotations while maintaining the high performance of various tasks.

Solution & Implementation:

The paper introduces a novel solution of implementing a Loss Prediction Module(LPM). This solutions contains a set of models which comprises of the target model and the loss prediction model. The target model is the main model performing the necessary tasks. The loss prediction module imitates the loss defined in the target model. This loss prediction model predicts the loss on unlabeled data and does not require the true labels as well. The goal is to select the samples providing most information to make the prediction, thus reducing the manual effort required for annotation. Initially, a small portion of randomly selected labelled data is used to train both the models.

Once trained, the LPM predicts the loss for the unlabelled data with the help of intermediate feature maps obtained target model to predict the loss for each data point. These intermediate feature maps are reduced using a Global Average Pooling (GAP) which are followed by fully connected layers with ReLU activations. The final output is a scalar value which predicts the loss of the unlabelled data point. Then, the LPM ranks these data points according to the predicted loss and selects the top-k samples. Now, these are considered the most difficult data points to annotate, which are then annotated manually and this data is now added to the training set, thus increasing the size of annotated dataset. The target model and loss predicting model are retrained jointly using a combined loss function that balances the target model's performance and the accuracy of LPM's predictions. The LPM is suitable for a wide range of deep learning tasks without any need for specific modifications and performs well across tasks.

Novelties and Contributions:

There are several novelties introduced in this paper to improve active learning strategies. Firstly, the usage of a loss prediction model to guide active learning is a very novel idea. Predicting loss directly from data rather than relying on traditional inefficient ways is a major breakthrough. Secondly, this method doesn't need to customized for specific tasks like traditional active learning. This model works for any use case without modifications, ensuring its versatility. In addition, LPM is very lightweight and trained along with the target model. This ensures efficient usage of computational resources. The validation is also provided on various tasks, thus proving the claims. This paper has paved a great way for further research in this domain of active learning.

Limitations:

Though this work introduces a novel method of active learning, it is still in the infancy stages and has limitations to address before it could be considered perfect for all scenarios. Few limitations are accepted by the authors and there are more question which can be posed before accepting the solution provided. Some of the limitations are:

- The authors note that for more complex tasks such as object detection and human pose estimation, the LPM is less accurate in predicting the losses.

- The data samples considered for validation may not be diverse enough. This could lead to the training a model which overfits.

- LPM is claimed to be relatively lightweight. But if we scale it to a relatively larger dataset using large models, training an additional LPM model would incur a higher memory requirement and also a need for greater computational resources.

- In cases where the labelled dataset might contain noise, the LPM might incorrectly learn to prioritize these datapoints thus leading to a degradation in the performance.

2 Can Active Learning Preemptively Mitigate Fairness Issues?

- Frederic Branchaud-Charron, Parmida Atighehchian, et. al.

Link to paper: https://arxiv.org/abs/2104.06879v1

Motivation:

The data collected and used to train models in machine learning play a very important role in the outcome and predictive power of these models. A significant challenge in the field machine learning is dataset bias, which arises from imbalanced datasets. This could be due to the fact that while training models on large datasets, it inadvertently reflects some societal biases resulting in unfair outcomes. To reduce this bias, fairness interventions are applied during training or post-training steps. In this paper, an innovative approach to this is proposed where we address the fairness at the data collection stage through active learning which selects the most informative data. In this paper, the authors try to explore if active learning can be used to mitigate fairness issues preemptively. In specific, they use BALD (Bayesian Active Learning by Disagreement) an active learning heuristic which can help reduce biases in the model's predictions by creating balanced datasets.

Solution & Implementation:

For this solution, the authors make use of an active learning heuristic, BALD, which selects the most informative samples which are to be labelled by measuring the amount of uncertainty in the prediction which can be reduced with more data. This uncertainty in the model's prediction that can be reduced with more data. BALD allows the model to focus on data points which will have greatest impact on reducing this uncertainty by maximizing the information shared between the model parameters and predictions. This paper tests the hypothesis that BALD reduces biases by focusing on underrepresented or ambiguous data which are generated using the Synbols tool. Two types of biases are stimulated - Minority Group Bias, where minority group is underrepresented, and sensitive attribute bias, where sensitive attribute has a strong correlation to target label. Further they also evaluate if

BALD can reduce predictive parity without having information on the sensitive attributes. Interaction between BALD and GRAD(Gradient reversal) is explored where GRAD prevents the usage of sensitive attributes by having an adversarial network. BALD and GRAD are combined to ensure both performance improvement and fairness.

Novelties and Contributions:

There are several novelties introduced in this paper to improve fairness in machine learning models. Firstly, the authors explore if biases can be preemptively addressed by focusing on informative data points that might induce bias in the dataset. Second, GRAD and BALD are combined to reduce the epistemic uncertainty and reliance on sensitive attributes which leads to improved fairness outcomes in the model. Additionally, by generating the data using Synbols, one can isolate and control the bias induced in the system. The authors have also tested how BALD interacts in such scenarios to ensure fairness. Finally, the paper uses predictive parity as the main metric to measure the fairness which also evaluates the difference in accuracy and thus assess the performance of BALD effectively.

Limitations:

Though this work introduces a novel method to ensure fairness in ML, there are some issues to be address as discuused below:

- The experiments completely rely on data generated by Synbols tool. They may not completely represent a real world dataset.
- BALD assumes no knowledge of sensitive attributes. However, in real world scenarios, we might have more information about these attributes. How can we make good use of it?
- There might be a need to trade off accuracy with fairness sincea higher regularization used in GRAD might impact the accuracy.

3 Active Learning for NLP with Large Language Models

- Xuesong Wang

Link to paper: https://arxiv.org/abs/2401.07367

Motivation:
Supervised deep learning requires vast amounts of labeled data and annotations done by humans which can be very expensive and time consuming. In NLP, in specific, labeling can be more difficult as it requires understanding context, nuances and domain knowledge. This paper ”Active learning for NLP with Large Language Models” by Xuesong Wang addresses the challenge of high annotation costs and data labeling for NLP tasks. The authors try to combine active learning with Large Language Models (LLM) to automate the annotation process. The author tries to investigate if LLMs can be used a annotators.

Solution & Implementation:
For this solution, the author proposes a mixed astrategy where labeled data is used in combination with human annotation. The study is conducted in two parts -

- Annotation using LLMs.
  GPT-3.5 and GPT-4 are used to label samples from the NLP datasets. The LLMs can make accurate predictions with few labeled datasets. A comparison of performance is evaluated by varying the demonstration examples to these LLMs. Different demonstration selection strategies like selection random examples, examples consuming fewer tokens, most similar input example are chosen for the tests. A consistency based strategy iss proposed to detect the uncertain samples. LLM is run with different temperature settings and the inconsistent predictions across multiple runs is flagged for a human review. Hence, they ensure majority of annotation is done by the LLMs.

- Active learning with mixed annotations.
  In this part, active learning is used to further reduce the labeling costs. The model is trained iteratively starting with a small labeled dataset and selects the most informative sample to label in each round. For the uncertain samples flagged by the previous step is sent for human annotations while the rest are annotated using the LLMs. Three active learning strategies are used for the evaluation - Random sampling, least confidence and Breaking ties.

Novelties and Contributions:
The paper introduces a novel approach of a mixed anotation strategy with LLMs. These annotate most of the data except few uncertain ones that are annotated manually. In addition, the paper introduces a consistency-based strategy which is a novel approach given the recent versions of GPTs no longer provide confidence scores. This also helps flag the samples if the outputs are inconsistent across runs with a varied temperature settings. Thus ensuring only the most reliable predictions are used for training. The authors also provide a detailed analysis of the cost saving achieved by the GPTs. Different strategies are evaluated, thus displaying the versatality of the model.

Limitations:
Though this work introduces a novel method to annotate data in NLP domain, there are some issues to be address as discused below:

- The experiments were only conducted on three datasets and this might not work on other datasets. Further, depending upon the model used, the performance for various tasks is different.

- Only single experimental runs were conducted by the author and this might give us a biased output.

- The cases of incorrect predictions by LLMs are not considered. The output looks for inconsistencies across runs, but it is so possible that in each run, a wrong or incorrect data is being generated which would result in lower accuracies of the model.