Active learning HW: read three papers about active learning and review

To be clarify, before I begin this homework, I know nothing/ heard little bit about active learning. So I decide to use this homework with a learning blog style to collect the way of how do i learn it and how to use active learning with possible related research.

Usually I do research with bottom way: first find first paper of a concept, then find follow up. But when I look at a early neurips back in 2010, I find its kinda hard to read with idea before the deep learning. It might be good for background knowledge but has nothing to do with my current research. Therefore for this home work, I want to use top down method: 1. To find a interesting application with Active learning related to modern problem, then basic on the application to understand what is AL and what can it helped on my research.

So for this homewor
Lates usage with LLM I found those two:
1.https://arxiv.org/abs/2404.08078
2. https://aclanthology.org/2023.findings-emnlp.334.pdf
3. https://arxiv.org/pdf/1910.01177

For 1 is just fresh out of arxiv. So might not be that useful but I can consider about fun and easy background reading to begin with.

Then I will read 2 which is from famous group and published on EMNLP

3 is actually related with my field of research. The second author is famous in our field and graduate from Penn State. But I found out its actually quite boring.

SQBC: Active Learning using LLM-Generated Synthetic Data for Stance Detection in Online Political Discussions

Motivation: Stance detection is an important task for many applications that analyze or support online political discussions, but stance detection lacks annotation and data. Stance detection dataset: a question q and response stands for positive for q or negative for q. Authors use an active learning method called SQBC based on the "Query-by-Committee" approach + LLM providing synthetic data.

Problem to solve: Not enough data for annotation on political discussions, and it's expensive to label this kind of data.

Method: The authors decided to use two methods: 1) using LLM synthetic data for fine-tuning, and 2) using a new active learning method (called Synthetic Data-driven Query By Committee (SQBC)) using LLM-generated synthetic data. They then use the synthetic data as an oracle to

infer the most interesting unlabelled data points by checking the similarity between the embeddings of the unlabelled data to the synthetic data.

How active learning plays a part in this research: They use active learning to solve the data labeling problem. This is achieved by letting the model choose the most interesting samples, commonly called most informative samples, from a set of unlabelled data points, which are then passed to, e.g., a human annotator for labelling. The most informative samples are those about which the model is most uncertain. By actively choosing samples and asking for the correct labelling, the model is able to learn from few labelled data points, which is advantageous especially when annotated datasets are not available.

Due to the nature of hard labeling, the authors decided to use synthetic data generated by LLM for the "most informative" sample and "human labeling" part in active learning.

They use Mistral-7B-Instruct-v0.1 with prompts to generate positive and negative responses based on given unlabeled q. The dataset they generated is called D.

Then they try to apply the active learning approach: "Query by Committee": basically where an ensemble of models is trained on the labelled data D and then chooses the subset Dch of Dunlabel for labelling where the most informative samples is Dcu.

In order to use active learning, the goal is 1) to find out which is the "most informative" sample and 2) human/LLM label.

To find more informative samples: Since it's an NLP task, they can use some of the NLP properties which involve the embedding of the sentence and finding nearest neighbors. They find K nearest neighbors of given unlabeled data in labeled synthetic data. Then use a score function s to define # of nearest neighbors that would be label 1. The authors define the most informative as having a score between 1 and 0, which means that the unlabeled data would be contradictory and can be easily positive or negative. That's how the authors define choosing the most informative sample.

The label would be LLM-labeled or using the majority vote of the k nearest neighbours.

The result, of course, is that this synthetic method would outperform the old way; otherwise, they wouldn't have published this paper.

My personal thoughtsof short-coming:

1. Author has bias: half 0 label and half 1 label, when letting KNN choose the contradicted one, it's actually falling into the authors' manual label but losing some label distribution.
2. Evaluation only on F1 score.
3. Maybe involved with some training data for KNN instead of just synthetic?

Paper 2

Active Learning Principles for In-Context Learning with Large Language Models

Problems to solve:

1. When faced with tasks where there is only unlabeled data available, how can we select the most appropriate samples to label and then use as in-context learning?
2. When we have labeled data for a given task, how can we efficiently identify the most informative combination of demonstrations for in-context learning?

Method: They consider ICL with AL setting where we have a large pool of unlabeled data for demonstration. We want to sample a batch of k data points using a data acquisition algorithm and we only perform a single iteration to find the most informative examples from the pool.

They build few-shot data acquisition algorithms inspired by the most prevalent AL algorithmic families: uncertainty sampling, diversity sampling, and similarity.

Authors assume that these k samples are subsequently labeled by humans.

Approaches:

1. **Baseline**: Random sampling from ICL pool
2. **Diversity method**:
    - Pool of unlabeled data with Sentence-BERT embeddings
    - Perform k-means clustering (number of clusters = k)
    - Select one data point from each cluster
    - Principle: A diverse set of in-context examples can offer greater advantages compared to random sampling
    - This selection strategy ensures chosen demonstrations likely encompass a broad range of information
3. **Uncertainty method**:
    - Instead of traditional AL methods using baselines such as maximum entropy or least confidence, they use loss
    - Use off-the-shelf LLM to score each candidate example from the pool by perplexity
    - E.g., a high perplexity set of in-context examples can yield greater advantages compared to randomly sampling from the dataset
4. **Similarity method**:
    - Based on KATE, a kNN-augmented in-context example selection method
    - Retrieves examples from the pool that are semantically similar to a test query sample

- Uses Sentence-BERT representations of both the pool and the test set to find the k-nearest neighbours
- Rationale: The most similar demonstrations to the test example will best help the model answer the query
- Limitation: Each test example will have a different prompt, as the k most similar demonstrations will be different
- Assumes ability to acquire labels for any in-context example selected from the pool

Conclusion: These AL approaches on ICL perform better than random selection.

What AL is doing here: For ICL, it's clearer that AL can also be used to choose the most "informative" samples in a dataset for selection based on various criteria (diversity, uncertainty, similarity).

Paper 3

IMPROVING DIFFERENTIALLY PRIVATE MODELS WITH ACTIVE LEARNING

Problem to solve: Selecting samples from the public dataset. Then label those samples and finetuning a DP model while preserving privacy.

Author use active learning to decide # of labels need to be sampled by select the most informative samples in unlabeled dataset.

Two method:

— DIVERSEPUBLIC and NEARPRIVATE

1. DIVERSEPUBLIC. Use trained model to get embedding that the activation before logits throught public data. Then perform PCA on embedding. Then authors select number of uncertain points according to lohit entropy of private model, project their embeddings onto the top few principal components, and cluster those projections into groups. Finally, they pick a number of samples from each representative group up to N labels (N most imformative) in total and fine-tune the DP model with these labeled data. It can be applied even to models for which we cannot access the original training data and 0 privacy cost
2. NEARPRIVATE Require privacy cost by apply DP-PCA of training data to get embedding & training data's PC. Then project Project k most 'uncertain' private/public data onto PCs. Then use nearest neibors to get the most informative public data points for labeling.

Those methods are pretty lame and didnt have any proof of how those methods are privated. Other that using existing AL method for DP, it has not too much contribution. I would think It belongs to AAAI instead of ICLR and It got rejected for reason.

So in conclusion, from those three papers, my understand of AL is that it can be use to deal with unlabel data problem and only need to unlabel the “hard” and “most” useful data for better fine-tuning a given model. Also how to choose/ define those “hard”/”most informative” data would also be another interesting questions. Usually they use given metric or use something like PCA/KNN to project the distribution of training data into the unlabel data and find out similarity between train and unlabel data. Thats how they usually choose to calculate.