Discriminative Active Learning

1. What problem does this paper try to solve, i.e., its motivation

Active learning helps to reduce the code of labeling the massive datasets by selecting only the most important samples to label. Though there are many methods to do this, the existing methods have some problems.
1. Labeling one sample at a time is slow, especially for neural networks.
2. Many of these methods are designed to focus only on classification tasks, making it hard to apply them to other types of problems.
To solve the second issue DAL(Discriminative Active Learning) is used.

2. How does it solve the problem?

DAL solves this problem by not depending on specific labels of the data during the selection process. Instead of focusing only on the classification tasks, DAL looks at the differences between labeled and unlabeled data in a general way. The idea behind this method is that when choosing which examples to label from a large pool of unlabeled data, we want our labeled set to represent the entire dataset as accurately as possible. This means picking examples that cover all the different types of data in the pool. To do this, the method treats the problem like a **yes/no** decision i.e., binary classification between the labeled and unlabeled data. For each new example, the system checks if it's more likely to be unlabeled. The examples that are most likely to be unlabeled are then chosen for labeling, confirming they add useful new information to the labeled set.

3. A list of novelties/contributions

1. Relation to Domain Adaptation:

Just like in domain adaptation, where the aim is to make the source and target distributions similar, in DAL, the labeled data behaves as the source,and the unlabeled data is the target. The objective is to make the labeled and unlabeled sets as close as possible. The H-divergence measures the difference between these two datasets. DAL aims to reduce this difference by selecting the most distinct examples from the unlabeled set, labeling them, and thus making the two sets more alike. Formally, the target distribution refers to the uniform distribution over the unlabeled examples, and the

source distribution refers to the labeled examples.The H-divergence between these two distributions can then be calculated. The unlabeled set namely $\mathcal{D}_T = \frac{1}{|\mathcal{U}|} \sum_{x \in \mathcal{U}} \delta(x)$. Similarly, the “source” distribution will be $D_S = \frac{1}{|\mathcal{L}|} \sum_{x \in \mathcal{L}} \delta(x)$ .The H divergence becomes :

$$d_{\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) = 2 \sup_{h \in \mathcal{H}} \left| \frac{1}{|\mathcal{L}|} \sum_{x \in \mathcal{L}} h(x) - \frac{1}{|\mathcal{U}|} \sum_{x \in \mathcal{U}} h(x) \right|$$

2. *Relation to GAN:*

In GANs, the generator improves by getting feedback from the discriminator, but in this case, we can't adjust the data that easily. Instead, DAL selects the most distinct examples from the unlabeled set (U) and adds them to the labeled set (L). We can think of DAL as a simpler version of GANs, where we update the labeled set by including examples the discriminator finds most different from the labeled ones. This approach also helps prevent mode collapse, a common problem in GANs.

3. *Scalability to Large Batches:*

DAL performs well with large query batch sizes, something that poses a challenge to other active learning methods.

4. What do you think are the downsides of the work?

- The experiments are limited to image classification tasks alone, such as MNIST and CIFAR-10. Including more complex or non-vision tasks, might have strengthened the generalizability of the method.
- The paper states that uncertainty sampling works well when using large batches, which points out whether more complex methods like DAL are really necessary in those cases. This could make DAL less attractive in situations where simpler methods can get the job done just as effectively.

Active-Learning-as-a-Service: An Automatic and Efficient MLOps System for Data-Centric AI

1. What problem does this paper try to solve, i.e., its motivation

The paper mainly deals with the challenges of Active Learning in Data-Centric AI. Traditional Active Learning tools require us to select the strategy manually, which are often inefficient. They fail to:

- Automate the AL process.
- Select the optimal strategy for given datasets and/or budget constraints.
- Operate efficiently on large datasets, causing the Active Learning process to become slow and resource-intensive.

The goal of this paper is to create an easy to use and efficient Active Learning system that makes it simpler for anyone to improve AI models. The system is designed to automate, reducing the need for users to manually intervene or make complex decisions. By doing so, it helps lower the barriers to using Active Learning, so even those without deep knowledge of the process can benefit from it and get better results from their models.

2. How does it solve the problem?

To make Active Learning easier and more efficient, the paper has introduced Active Learning-as-a-Service(ALaaS). This automatically selects and runs Active Learning strategies based on the user's budget and target accuracy, without any manual intervention. It uses a server-client setup and advanced techniques like caching and batching to speed up the process. For users who aren't sure which AL strategy to use, ALaaS includes a predictive system (PSHEA) to choose the best strategy.

3. A list of novelties/contributions

- ALaaS uses smart techniques like pipeline processing, backend support, and caching to speed things up. These improvements help ALaaS handle large amounts of data much better than other tools, making it quicker and able to grow with bigger datasets.

- This has a smart AL agent that uses a predictive model called PSHEA. This agent quickly selects the best Active Learning strategies and drops the weaker ones as it goes. This agent will automatically choose the best strategy. The system takes care of most of the work for us. We just need to enter simple details like their budget and accuracy goals, and the AL agent will automatically choose the best strategy.
- Pipeline Optimization is the other strategy used in this paper. ALaaS splits the Active Learning process into clear steps, downloading, pre-processing, and learning. It also saves data along the way to avoid delays, which makes the whole process much faster. This setup makes ALaaS up to 10 times faster than other tools.

4. What do you think are the downsides of the work?

- Although ALaaS improves efficiency in certain cases, it may not generalize well across all datasets or application scenarios.
- Efficient use of ALaaS may require significant computational resources, which may not be accessible to all users.
- Though PSHEA improves automation, it introduces complexity in terms of model prediction and elimination rounds, which may not always yield significant performance gains over simpler methods.

Deep Bayesian Active Learning with Image Data

1. What problem does this paper try to solve, i.e., its motivation

Active learning works with small amounts of data, but deep learning usually needs a lot of data to work well. Also, active learning often relies on understanding the uncertainty in model predictions, but deep learning models don't usually handle uncertainty well.The goal of the paper is to make Active Learning more practical and scalable for deep learning, especially in tasks like image classification, where labeled data is scarce and expensive to obtain.

2. How does it solve the problem?

This paper combines new ideas from Bayesian deep learning to make active learning work better with deep learning, especially for complex data like images. It uses **Bayesian Convolutional Neural Networks (BCNNs)** to handle uncertainty and shows that this approach significantly improves active learning for tasks like recognizing handwritten numbers (MNIST dataset) and diagnosing skin cancer from images (ISIC2016 task).These BCNNs are combined with Active Learning techniques, where an acquisition function selects the most informative data points for labeling based on model uncertainty.

3. A list of novelties/contributions?

- This paper introduces us with Bayesian Convolutional Neural Networks into active learning, helping to handle uncertainty in deep networks.
- It's also providing a system that makes it easier to handle image data when it's expensive and time-consuming to label.
- These method significantly reduce the number of labeled examples needed to create accurate models. This was demonstrated through experiments on MNIST and skin cancer diagnosis data.
- There are real world applications which use this approach for diagnosing skin cancer from lesion images, showing its practical importance in medical tasks.

4. What do you think are the downsides of the work?

- The suggested approach is computationally expensive because it has high demands on resources both for training and inference.

- The Bayesian methods integrated into deep learning models make the processes of their training more complex, thus harder to implement and scale.
- Although the method is very encouraging in images, it has not yet had any testing with other kinds of high dimensional data, which may pose some challenges.