HW1

Find and read three papers about Active Learning. For each paper, please write down:

1. What problem does this paper try to solve, i.e., its motivation
2. How does it solve the problem?
3. A list of novelties/contributions
4. What do you think are the downsides of the work?

Answer:

I have read 3 papers on Active Learning.

TiDAL: Learning Training Dynamics for Active Learning (ICCV- 2023)

https://openaccess.thecvf.com/content/ICCV2023/papers/Kye_TiDAL_Learning_Training_Dynamics_for_Active_Learning_ICCV_2023_paper.pdf

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

The purpose of active learning (AL) is to select the most useful data samples from an unlabeled data pool and annotate them to expand the labeled dataset under a limited budget. Even though existing AL methods are divided into two mainstream categories: diversity- and uncertainty-based methods, uncertainty-based methods are known to be effective in improving model performance as they select the most uncertain samples. But most uncertainty-based methods quantify data uncertainty based on static information from a fully-trained model "snapshot." Training Dynamics (TD)—the way the model changes throughout training—is neglected by these techniques, despite the fact that other fields of study have actually demonstrated that TD offers crucial hints for gauging data uncertainty. Thus, the authors aimed to check and improve uncertainty estimations by integrating TD.

2. How does it solve the problem?

The authors proposed TiDAL, a novel AL method that integrates TD into the uncertainty estimation process. Instead of relying solely on static snapshots, TiDAL captures the changes in model behavior during training. As tracking the TD of all the unlabeled data is computationally infeasible, they devised an efficient method to estimate the TD of unlabeled samples through a new module-TD prediction module. This module is trained on the readily available TD of labeled data and learns to predict the TD of unlabeled samples. Then, they employed the typical estimators, entropy or margin, based on the projected TD of each unlabeled sample, to identify the most uncertain sample; so that human annotators can label it. In this way, TiDAL provides more accurate uncertainty estimates without the computational cost of tracking every sample during training.

3. A list of novelties/contributions

The key contributions of the paper are as follows:

(1) This established a connection between training dynamics and active learning by presenting theoretical and experimental evidence supporting the usefulness of training dynamics in evaluating data uncertainty.

(2) This paper proposed a new method that efficiently predicts the training dynamics of unlabeled data to estimate their uncertainty.

(3) Compared to current active learning techniques, the proposed method performs better or comparably on both balanced and imbalanced benchmark datasets.

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

The downsides of TiDAL:

1. TiDAL is designed only for classification tasks, and thus it cannot be applied to AL targeting other tasks, such as regression.
2. The prediction of TD for unlabeled data is dependent on the model's learned dynamics from labeled data. If the labeled data is not adequately representative of the larger dataset, this could result in errors.
3. The TD prediction module adds extra complexity in the training phase.

Unlabeled data selection for active learning in image classification
https://www.nature.com/articles/s41598-023-50598-z

1. What problem does this paper try to solve, i.e., its motivation
Image classification plays a crucial role in various industries and sectors, but the need for large labeled datasets poses significant challenges in model development. On the contrary, a large amount of unlabeled data is typically available. So, this paper tried to improve image-classification models by employing “Active Learning” (AL). By using AL, the most informative unlabeled data points are selected. Once labeled, these data points can greatly enhance model performance, which eliminates the need for labor-intensive manual labeling and saves money. This paper also tried to employ new methods for unlabeled data selection because traditional AL methods like random sampling or uncertainty-based strategies may not capture the complex characteristics of image data.

2. How does it solve the problem?
The paper introduced three new data selection strategies within the active learning framework to select unlabeled images for labeling. These are:

1. Similarity-based Selection: This method evaluates the similarities between unlabeled and labeled image datasets. It ensures that the selected unlabeled data accurately represent the already labeled dataset, reducing the selection bias often present in uncertainty-based selection methods. This method ensures a more thorough representation by improving the coverage of the data distribution space.
2. Prediction Probability-based Selection: This method checks the prediction probabilities of the initial model on unlabeled data. Images with low confidence or high uncertainty are given priority for labeling as they are likely to be more informative.
3. Competence-based Active Learning: This method tailors the selection strategy to match the model’s learning progression and capacity. Because deep learning models are prone to stagnating in local optima, this is very important.

3. A list of novelties/contributions

1) The need for a balanced approach to active learning that matches data selection to the model's developing learning capacity was discussed in this paper.
2) All of the proposed methods work together to make training more productive and efficient while meeting the changing needs of image classification tasks.
3) As demonstrated by experiments on the Cifar10 and Cifar100 datasets, proposed new Active Learning methods outperform existing techniques in both efficacy and stability within the context of image classification tasks.

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

1. The proposed AL methods are tested only for image classification tasks. So, it is unclear how well it performs for more difficult tasks or datasets than image classification.
2. The initial model's predictions provide the basis for the prediction probability-based selection. Data selection quality may be compromised if the original model performs badly since it may give false uncertainty estimates for unlabeled data.

Active Learning is a Strong Baseline for Data Subset Selection (*NeurIPS 2022 Workshop HITY*)

https://openreview.net/pdf?id=PAgpyQ5rGS

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

When training models on big datasets, there is a significant computational cost involved. Thus, in the context of computational cost, identifying the ideal subset of training data that may roughly mimic the performance of models trained on whole datasets is crucial. But existing data subset selection methods are often complex and task-specific. Another closely related problem is the active learning problem developed for semi-supervised learning in which an important subset of unlabeled data is identified (for further labeling) by making use of the currently available labeled data. This paper's motivation comes from a simple observation: one can apply any off-the-shelf active learning algorithm in the context of data subset selection. So the paper tries to check whether simpler active learning techniques can outperform existing sophisticated data subset selection approaches.

2. How does it solve the problem?

The authors come up with a very simple idea. They pick a small random subset of data and pretend as if this random subset is the only labeled data, and the rest is unlabeled. Now, any off-the-shelf active learning algorithm can be applied to select samples from the unlabeled side. The labels of the chosen samples are disclosed and added to the subset following each stage of the sample selection process. Until the necessary data subset size is obtained, more iterations are carried out.

3. A list of novelties/contributions

1. The paper shows that active learning methods can be effectively used for data subset selection.
2. The authors find that it is crucial to find a balance between easy-to-classify and hard-to-classify examples when selecting a subset.
3. The approach outperforms all the existing data subset selection algorithms on benchmark tasks, according to the authors' results.

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

1. The paper focuses primarily on empirical evaluation regarding AL's performance.
2. Data subset selection using AL may affect the model's generalization, which is not studied in the paper.