Name: Kush Harish Vora
Course: CSE584
Homework-1
14th September 2024

PAPER 1: ACTIVE LEARNING FOR CONVOLUTIONAL NEURAL NETWORKS: A CORE-SET APPROACH

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

Training deep learning networks, specifically CNN, requires large amounts of data. However, obtaining such data is not only expensive but very time consuming. Moreover, CNNs don't start saturating when trained on large amounts of data. Hence, a new paradigm called active learning was introduced where the model learns on a small set of labeled data, and then for every iteration queries an oracle (database) for additional unlabeled data points and then tries to predict them. There are various active learning heuristics in the literature. However, these traditional active learning heuristics do not apply well on convolutional neural networks due to various reasons. Classical active learning heuristics function by querying 1 data-point at a time from the unlabeled pool. This is not enough for large networks like CNNs, because one data-point may not have any statistical impact on the learning performance of the network due to local optimization algorithms– leading to poor performance. Hence, there was a need for new heuristics that work well for CNNs. With this motivation in mind, O. Sener et al. proposed an active learning heuristic that works by querying batches of data from the unlabeled pool. They claim that their approach has resulted in state of the art results on various image classification tasks.

How does it solve the problem?

To solve the limitation of classical active learning heuristics, one needs to query large amounts of records per iteration. The authors correlate this heuristic to the core-set selection problem, where for each iteration, the goal is to query a batch of unlabeled data points such that the model that was trained on this subset performs competitively to those models that are trained on the entire dataset. To solve the problem of core-set selection, O. Sener et al. provide a bound that lies between the average loss over any subset queried from the dataset and the remainder of the points. This is done through the geometry of the data points. As a good active learning heuristic for convolutional neural networks, one must try choosing a subset that minimizes this bound. The authors also correlate this problem to the famous k-center geometric problem. The authors propose a core-set loss function, which is the difference between the average empirical loss over the labeled data-points and average empirical loss over the whole dataset (including unlabeled data-points). Since the k-center problem is NP-hard, the authors use a greedy algorithm to approximate the solution to the k-Center problem. This algorithm iteratively selects points that maximize the minimum distance to the current set of selected points, ensuring a 2-OPT solution, which is within twice the optimal solution.

List of novelties/contributions

1. **Core-Set Heuristic for CNN:** The authors introduce a new heuristic for applying the active learning training for CNN's. This heuristic works by choosing a batch of records for every iteration from the unlabeled pool.
2. **Core-Set Loss to train the CNN:** The authors introduce a new loss function to train the CNN's using this heuristic. They coin this as the core-set loss function. The aim is to minimize the difference between the average empirical loss over the labeled data-points and average empirical loss over the whole dataset including the unlabeled data points.
3. **Theoretical Guarantees:** The authors not only show empirical results, but they also provide theoretical bounds on the core-set loss using the geometry of data points.
4. **Greedy Approximation Algorithm:** The authors correlate the core-set problem to the geometric k-center problem. Since this problem is NP-hard in nature, the authors come up with a greedy approximation algorithm to solve the problem.
5. **Model agnostic active learning:** The authors claim that their approach is model agnostic, and can work with any convolutional neural network. They also provide empirical results validating their claims by achieving SOTA results on a number of tasks.
6. **Experimental results:** The approach proposed by the authors was tested on common benchmarking datasets— MNIST (digit classification), CIFAR-10, and CIFAR-100. Their results were SOTA.

What do you think are the downsides of the work?

1. The authors have showcased empirical results on only image classification tasks and have claimed that it is a model-agnostic approach. There are no results that demonstrate the working of this active learning heuristic for other tasks such as natural language processing, object detection, or image segmentation.
2. The active learning heuristic proposed in this work, functions on an assumption that there is zero training error for the core-set. It is not necessary that this assumption will hold true in a real world scenario.
3. The K-center problem is NP-hard and hence they solved the problem by a greedy approximation algorithm. The computation intensity for solving this problem was not explored in the paper.
4. The mathematical formulation and optimizations discussed in the paper were complex and may require mathematical and geometric expertise for any researchers that want to replicate the work in other domains of machine learning and deep learning.

PAPER 2: LEARNING LOSS FOR ACTIVE LEARNING

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

D. Yoo et al. claim that most of the active learning approaches that are presented in the literature are designed for the specific task and do not extend to other deep learning tasks. This is because either their approach does not support other tasks or it becomes computationally impractical to implement them for other tasks. Experimental results from weakly supervised or semi-supervised learning demonstrate that more samples of annotated data yields better performance— however, the costs associated with annotating this data is very high. Moreover, some tasks like object detection or medical image segmentation require trained and board-certified scientists or doctors to annotate. Due to these budget constraints, a lot of deep learning tasks have not achieved human-like accuracy. Moreover, the active learning approaches in the literature like uncertainty-based approach, diversity-based approach, and expected model change— all suffer from some or other limitations. For example: for task-agnostic uncertainty-based approaches, numerous models form a consensus, however, this construction becomes very expensive and hence infeasible. Moreover, they do not scale to large datasets. For distribution-based approaches, though they are task-agnostic, they require extra engineering to design location invariant features and hence are infeasible. While there are some good active learning solutions in the literature that have helped reduce the annotation cost— they however are task–specific and can not be extended to other tasks. Hence, the authors of the paper have come up with a task-agnostic loss prediction module that can be used to train deep learning models, specifically convolutional neural networks on a limited set of annotated data.

How does it solve the problem?

D. Yoo et al. propose a novel active learning framework for deep neural networks. The motivation behind this work stems from the fact that a normal loss/cost function which is used to train a neural network is task-agnostic, i.e. the same loss function is used to train the network on a variety of tasks, regardless of the network architecture. The authors propose a loss prediction module that learns to predict the loss of an input data point. At first, a network is trained on an initial labeled dataset which learns both the target loss and the initial loss prediction module. After this initial set of training, all the unlabeled data points in the pool are evaluated by the learnt loss prediction module to obtain data loss pairs. The human annotator then annotates the top K pairs with highest loss where K can be set according to budget constraints. This loop is iterated upon until the budget exhausts or the model reaches the desired performance. The loss is calculated in parallel to the target loss, making it a 2-layered pipeline. The loss prediction module not only minimizes the annotation and engineering costs, but also, minimizes the computation overhead of learning the new loss.

List of novelties/contributions

1. **Task-Agnostic Design:** Inspired by how the loss function that is used to train a neural network does not depend upon the architecture or the task, the authors developed the loss prediction module to be task-agnostic, allowing it to be incorporated into any neural network and for any task easily.
2. **Comprehensive evaluation:** D. Yoo et al. have provided empirical evidence of their loss prediction module performing state of the art results across three different tasks— image classification, object detection, and human pose estimation.
3. **Loss prediction module:** They propose a new loss module that can be attached to any neural network to predict the loss of an unlabeled data point. This loss is then used to pick top-K elements for the human annotator to annotate thereby reducing the workload on an expert annotator and also the costs associated with it.
4. **Joint training:** The loss prediction module learns in conjunction with the standard loss in a neural network. It does not take any extra computation requirements.

What do you think are the downsides of the work?

1. The paper acknowledges that while they have considered the uncertainty of the data, the diversity or the density (how dense the data is) is not taken into account. In tasks where diversity is very important when it comes to selecting the training data, this can be a major drawback of their approach.
2. While the approach has proven to be working with all 3 tasks— image classification, human pose estimation, and object detection, one can see that the loss prediction module has deteriorating performance as the task starts getting complex. It is possible that on more complex tasks the module will not work as is, and will require modifications.
3. Although the authors claim that the 2-staged training is computationally light. They have not provided empirical evidence proving their claims.
4. The authors have showcased empirical results on only image specific tasks and have claimed that it is a task-agnostic approach. There are no results that demonstrate the working of this for other modalities like language, speech, or other unstructured data.

PAPER 3: DEEP BAYESIAN ACTIVE LEARNING WITH IMAGE DATA

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

Y. Gal et al. addresses the challenge of scaling active learning approaches to high-dimensional data, specifically image data, using deep learning (neural) networks. Classical active learning approaches work on small datasets and model uncertainty. However, this does not work for deep learning because most models require vast amounts of labeled data to perform well. Since classical active learning approaches are focused on being efficient with small datasets, we can not directly employ them for deep learning use cases. Another reason is the lack of uncertainty representation in deep learning. Most active learning methods rely on acquisition functions that are based on uncertainty to determine which unlabeled samples are the most informative. However, deep learning networks are not known for representing their uncertainty, making it difficult to query samples in an active learning setting. Hence, the paper aims to overcome these limitations of existing active learning techniques by introducing a framework for deep bayesian active learning that scales to high-dimensional data and deep learning networks.

How does it solve the problem?

Y. Gal et al. proposes a solution that involves using bayesian convolutional neural networks (BCCN) that extend normal CNNs by incorporating bayesian inference. This is done to represent the uncertainty in the model's prediction. They use Monte Carlo (MC) dropout as an approximate Bayesian inference method, allowing the deep learning model to estimate uncertainty. With this, they are able to overcome one of the limitations of the active learning techniques (uncertainty). Once the model can represent uncertainty using the MC dropouts, the next step is to choose an appropriate acquisition function that can leverage the uncertainty to select the most informative datapoint for labeling. The authors discuss several functions such as Bayesian active learning disagreement to focus on the most uncertain data points. The proposed solution has been demonstrated effectively on high-dimensional data such as images. The authors also showed the practicality of the work by implementing it on a skin cancer diagnosis task, showing improvements over other traditional active learning methods.

List of novelties/contributions

1. **Bayesian CNNs:** The authors introduced the use of BCNNs to represent uncertainty, enabling the application of active learning with deep learning networks for high dimensional data, which is a novel application in the context of active learning.
2. **Bayesian Active Learning Disagreement:** The authors have developed an acquisition function that can be used to query the most informative data points by leveraging the model uncertainty.
3. **Empirical results on real world applications:** The authors demonstrated their framework's effectiveness by diagnosing melanoma from a small set of labeled lesion images

What do you think are the downsides of the work?

1. The Bayesian approach, especially with MC dropout, can be resource and computation intensive, especially if the models are being reset at each acquisition step (as specified in the melanoma skin cancer experiment). This may render the approach infeasible when working with very large datasets or real-time applications.
2. The performance of the proposed framework was heavily reliant on the chosen model architecture and the set of hyperparameters. For other tasks, getting an ideal architecture and hyperparameter, may require extensive tuning.
3. The work focuses on high-dimensional data, specifically images. The work does not touch upon if the approach will generalize to other modalities like speech, and natural language, or even mixed datasets.
4. The work has been extensively tested on small datasets. It is possible that the approach does not fare well when we are dealing with larger datasets, which is particularly the case in various real-time applications.