CSE 584 HW 1

Kade E. Carlson

September 15, 2024

1 Paper 1: Active Learning with Statistical Models, Cohn et al.

1. What problem does this paper try to solve?

In many contexts of machine learning, active learning can be used to choose a statistically "optimal" point for the next observation in training data. This makes for a more efficient use of computational resources especially when data is difficult or expensive to obtain. The authors acknowledge that in a neural network context, active learning can become prohibitively expensive. The motivation is to show that there are machine learning architectures where active learning does not become expensive and considerably helps certain models learn.

2. How does it solve the problem?

To solve this problem, the authors discuss a heuristic that has better computational speed and performance on two different types of machine learning models: Gaussian Mixture models and Locally Weighted Regression models. These two models both have their own mathematical foundations for the proposed heuristic. The authors compared the variance of the data and the mean-squared error between a random learning approach and an active learning approach. Both models showed considerable increases in performance in both the variance and the mean-squared error when using active learning versus when using the random observation selection.

3. A list of novelties and contributions

1. Mathematical model for the approximation of variance for two machine learning models that is to be minimized
2. Proposal of using active learning on Mixture of Gaussian and Locally Weighted Regression models instead of neural networks
3. Providing a statistical analysis to the benefit of minimizing data variance in an active learning scenario

4. Applying active learning to an interesting robotics problem

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

I think by far the biggest downside is that other active learning statistical heuristics were not used to compare against the one being proposed. It's good that the authors compared against random sampling, but how can we be sure that this is truly an improvement to the state of the art without knowing how it compares to other active learning methods? The other downside is that I think that not enough effort was put into analyzing what types of problems that this would be the most effective for. I think providing experiments in a context where they say active learning is best (difficult to obtain data) would have been beneficial. Of course, this isn't always feasible given hardware and time constraints but at least some sort of mention of it could have made the paper stronger.

2 Paper 2: Deep Bayesian Active Learning with Image Data, Gal et. al

1. What problem does this paper try to solve?

Active learning is a good way to optimally determine the next point in a dataset to receive a label. This way, computational resources are not wasted and obtaining data becomes more efficient especially when data is hard to obtain. This paper notes that active learning suffers from two problems: reliance on small data and use of model uncertainty. The problem is that deep learning models require large amounts of data and don't represent uncertainty at all. This paper attempts to solve the problem of using active learning in a deep learning context by extending it into high dimensions and capturing model uncertainty as well.

2. How does it solve the problem?

To solve this problem the authors propose using a Bayesian Convolutional Neural Network which is like a regular CNN but each model parameter has a Gaussian prior fitted to it. This helps incorporate model uncertainty. The authors also use various acquisition functions such as max value entropy search to help determine where to observe next. The authors used random sampling as a baseline to compare, and each acquisition function chosen performed better than random sampling while also requiring fewer images from the dataset.

3. A list of novelties and contributions

1. A collection of various acquisition functions that will perform best for a BCNN as well as a computationally tractable form for each acquisition function.

2. The proposal to use Bayesian Convolutional Neural Networks to do image data recognition so that the model uncertainty can be captured by the acquisition function.

3. A noticeable improvement to current image recognition techniques, at least for a skin cancer dataset

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

I think the biggest downside of the work is that this active learning approach is not generalizable to different types of deep learning models. This method is ad hoc in the sense that it solves this image classification problem by fitting a BCNN into it but it does not really answer the question of how can active learning be used for many types of deep learning models. I also think if they ran more experiments on many different datasets and then compared it would have made for a stronger paper.

3 Paper 3: From Theories to Queries: Active Learning in Practice, Settles

1. What problems does this paper try to solve?

Many active learning approaches in literature assume that a single annotator is used and can always be trusted and that each query is expensive. In real-world situations these conditions do not always hold, so it is important to determine solutions to active learning in practice. This paper surveys the many solutions to active learning in practice.

2. How does it solve the problem?

There are six research directions that tackle the challenges in active learning. The first is querying in batches. Querying in batches allows us to find informative and diverse labels to choose next. A popular approach at the time was to cluster queries according to a utility function and then choose representative data from each cluster. A second research direction is noisy oracles. Annotations in datasets are going to have noise due to human error and fatigue. Some work has looked at letting learners repeat queries, but work still needs to be done here. A third direction is variable labeling costs. Many active learning approaches have become cost aware and it has been shown in previous work that when labeling costs are unknown, regression models can be constructed to predict costs. These have been shown to be more accurate than known cost aware models. A fourth direction is alternative query types. It has become common for queries to be chosen as "bags" where a bag contains at least one positive query to the desired instance. A fifth direction is multi-task active

learning. Here the idea is that a query will be used for multiple tasks and its "informativeness" will be decided for each task. The final direction is changing models. There are not many promising solutions here as of 2011 but some have suggested parsing models to train other parsing models.

3. A list of novelties and contributions

1. A complete survey of the time of the writing of active learning problems and solutions for using active learning in practice
2. The author mentions previous work done by him where he proposed a novel technique for cost sensitive active learning

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

For a survey paper, I think it does the job well, however, the references are pretty jumbled. It would be better if the author organized the references so that they were much easier to find. I also think maybe a discussion on how companies like IBM, Google, and Microsoft are incorporating active learning into their work is warranted considering it is mentioned but never discussed in detail.