CSE 584, HOMEWORK 1

Name: Yogeshvar Reddy Kallam IdNum: 920385794 email: yvk@5381@psu.edu

Paper 1: Learning Algorithms for Active Learning

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

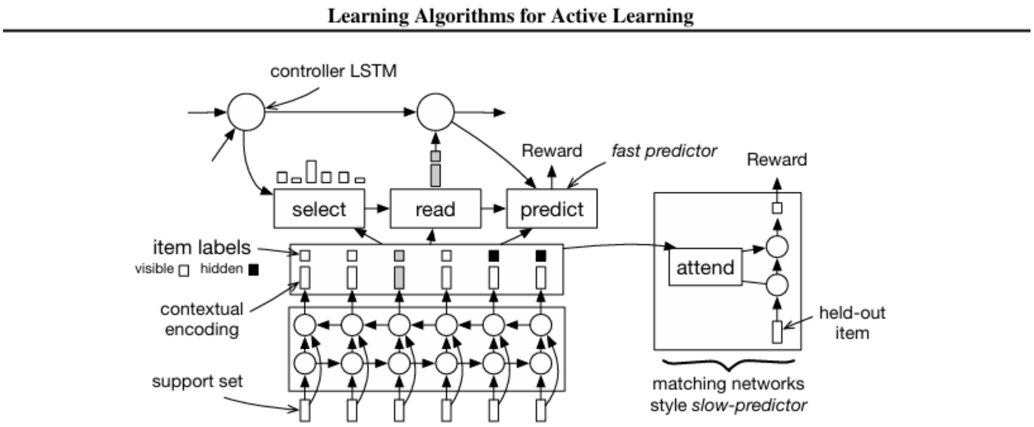
The paper addresses the challenge of active learning in the context of machine learning tasks where labeled data is limited, and the acquisition of new labels is costly or time-consuming. Traditional machine learning models often require large amounts of labeled data to achieve high performance, which can be impractical or infeasible in many real-world scenarios. For instance, in medical imaging, labeling data requires expert knowledge, which is both scarce and expensive. Similarly, in recommendation systems, obtaining preference information for new users can be difficult due to the cold-start problem, where little to no data is available for the system to make informed recommendations.

1.2 How does it solve the problem?

The paper proposes a novel approach to active learning that leverages metalearning to overcome these challenges. The authors introduce a model that can learn active learning algorithms end-to-end, meaning that the model learns to selectively query for labels in a way that maximizes its performance on a given task while minimizing the number of label requests. This is achieved by allowing the model to interact with a pool of labeled items across various related tasks, which informs the development of an active learning strategy tailored to the specific task at hand. The model's architecture is based on Matching Networks (MN), which are adapted to work in settings where labels are not initially available. The model operates by sequentially deciding which items to request labels for, adding these items to a labeled support set, and then using this support set to make predictions in a MN-style manner. The active learning process is framed as a sequential decision problem, where the model's goal is to request labels that will most effectively improve its predictive capabilities. The authors demonstrate the effectiveness of their model through empirical evaluation on the Omniglot dataset for one-shot learning tasks and the MovieLens dataset for bootstrapping a recommender system. These experiments show that the model can learn efficient label querying strategies and holds promise for practical applications, particularly in scenarios where data efficiency is crucial and labeled data is a scarce resource.

1.3 list of novelties/contributions

We introduce an end-to-end learning of active learning algorithms, unlike most previous attempts to devise active learning that relied on traditional handcrafted selection heuristics. Therefore, this model acquires a finer representation of the data, a strategy by which items get selected, and the predictive function for the task under consideration, outperforming heuristics oblivious to a particular task. It is a meta-learning model that exploits the availability of labeled instances from different but related tasks to learn how to select for a particular task; this lets active learning



strategies generalize across tasks. They evaluate the efficiency of the model in querying labels using "active" variants of the one-shot learning tasks on the Omniglot dataset. The algorithm further learns the MovieLens dataset in order to bootstrap a recommender system, testing the model for effectiveness. This shows practical applicability and the potential for solving the cold-start problem of recommender systems. The model is designed with an adaptive selection policy that relies on context-free and context-sensitive encoding, a controller module, and a selection module to adaptively select the most informative items to label, moving away from fixed heuristic policies. Trained through reinforcement learning, the model refines its behavior over many training episodes to maximize its performance during test episodes. It proposes a novel architecture composed of modules for encoding, controlling, selecting, reading, and predicting, integrated in such a way as to have a synergic interaction in an active learning process. Empirical validation shows that this model can learn strong active learning algorithms; it achieves results close to an optimistic baseline on Omniglot and outperformed several baselines on MovieLens. These contributions cumulatively advance the state of the art in active learning by proposing a model that can learn to query labels in a very selective, efficient, and effective manner, with applications across domains where labeled data are scarce or expensive to acquire.

1.4 What do you think are the downsides of the work

The model does contain a lot of moving parts: several encoding modules, control, selection, read, and predict. This may limit scalability and computational resources on larger datasets or more involved tasks. That means performance for the model in a 1-shot setting, especially in a 10-way classification problem, could point to a limitation when there is an increase in the number of classes in the support set but when the shots (labels) are fixed. Probably, the model may not generalize well when the unseen classes are increased with very few examples. In this case, an oracle policy trains the model that knows all the labels. This could be the reason for the failure of the model in real-world applications because no such oracle can exist there, and the model needs to make decisions without full information. Inactive learning for the model, it was tested on certain data and tasks; for instance, the Omniglot dataset was used for one-shot learning, and MovieLens was used for recommender systems. It is not tested against a larger range of tasks, and hence it is hard to predict the functioning of the model in larger domains. These ablation tests support that some streamlining can be done to the architecture because the removal of certain components did not affect performance much in particular settings, for instance, the attention temperature or context-sensitive encoder. In that respect, this is indicative that the current architecture might be unnecessarily complicated, probably at the cost of model interpretability and efficiency. While the model is trained to balance the task performance and labeling cost, in real-world scenarios, the actual cost-benefit trade-off is apt to be nuanced and difficult to capture. The ability of the model to make optimal decisions when faced with various costs for labeling and prediction errors remains underexplored. While the paper argues that the model is promising for practical applications, it also gives the impression that further research is needed in order to adapt the approach to more realistic and diversified settings. From these possible downsides and directions for future work, scalability, generalization, and real-world applicability of the proposed active learning model obviously require further research.

CSE 584, HOMEWORK 1

Name: Yogeshvar Reddy Kallam IdNum: 920385794 email: yvk@5381@psu.edu

Person 2: Improving Generalization with Active Learning

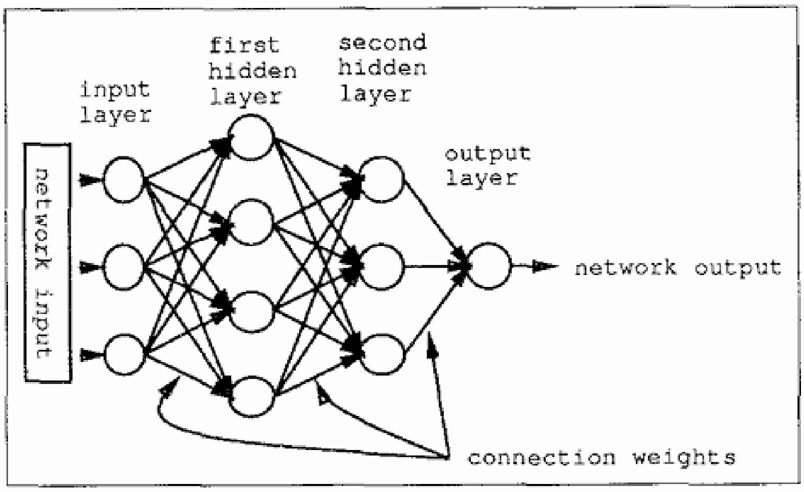
2.1 What problem does this paper try to solve, i.e., its motivation

Papers such have taken up the challenge of improving the generalization capability of machine learning models, especially neural networks, through a variant active learning called selective sampling. This study was motivated by the need to address the inefficiencies that are common in traditional passive learning, which depends on random sampling of training data. These methods are often slow to converge and would normally require a large number of labeled examples to achieve good generalization performance.

2.2 How does it solve the problem?

The authors present a method known as selective sampling to solve a problem: the improvement of generalization capability of machine learning models, especially neural networks. Active learning is the key to this problem, where instead of passively learning from randomly drawn examples, the learning algorithm takes control of the information in the input domain, selects the most informative examples, and learns from these examples. The authors present a neural network implementation called the SG-network that keeps track of two concepts: the "most specific" and the "most general" that are consistent with the training data. These define a region of uncertainty within which the network can query for new examples. With such a region of uncertainty, the algorithm refines the notion it has of the target concept in a more efficient manner. To do so, the network is trained on a mix of actual training examples and "background" points drawn from the input distribution. The background points introduce a bias that biases the network towards a "most specific" configuration. As background examples are usually many more in number compared to the actual training data, variable learning rates for the actual training examples and background examples have been proposed by the authors for creating a balance between the impact of background examples and real data. The paper illustrates the effectiveness of selective sampling through experiments in three domains: a simple boundary-recognition problem, learning a 25-input real-valued threshold function, and recognizing the secure region of a small power system. In each case, networks that were trained using selective sampling showed significant improvements in generalization error over those that had been trained using random sampling or naive querying algorithms. In short, this work outlines a method of improving the generalization of neural networks with an active learning strategy by the selective sampling of data points, which turn out to be the most informative, hence increasing effectiveness and efficiency in its learning process.

IMPROVING GENERALIZATION WITH ACTIVE LEARNING



2.3 list of novelties/contributions

The authors present the SG-network: a neural network incorporating a selective sampling method. The network continuously monitors "most specific" and "most general" concepts that define a region of uncertainty and serve as a guide for querying new, informative examples. The authors have conducted experiments in three diverse domains: a simple boundary-recognition problem, learning a 25-input real-valued threshold function, and power system security analysis to show the effectiveness of selective sampling. It is shown through these experiments that networks trained with selective sampling outperform those trained with either random sampling or naive querying algorithms. This paper discusses various practical and theoretical limitations selective sampling faces, such as the maintenance of correct uncertainty regions for complex concepts and the possibility of oversampling. The authors claim that Bayesian analysis could refine these utility estimates for queries about different parts of this uncertainty region, yielding a more sophisticated active learning approach. The paper then places its contributions into the bigger context of active learning research by discussing related work, pointing out the ways in which selective sampling extends and complements prior work in the field. All these contributions put together further advance the knowledge of active learning and its ability to enhance the generalization of the machine learning model, especially for those occasions when labeled data is hard to obtain or too expensive.

2.4 What do you think are the downsides of the work

The greater the length of the description of the concept, the higher the likelihood that the area of uncertainty blankets the whole domain, which reduces selective sampling's effectiveness to no better than that of random sampling. The problem is accentuated in high-dimensional spaces, or when the underlying representation complexity associated with a concept is very high. This could lead to the model being less effective at selective sampling and, by extension, less performing well overall. If the region of uncertainty becomes very small, it might result in oversampling concerning certain regions. This may result in biased learning outcomes. The bias in this respect mainly comes from the inductive bias of the training algorithm and may not be representative concerning the large domain. Hence, the model over-concentrates on those areas and probably misses out on other prospective important parts of the data. In the case of relatively simple concept classes, it is practical to implement selective sampling. However, as the complexity of the concept class increases, computation and maintenance of an accurate approximation of the region of uncertainty become difficult. This, in practice, considering more complex tasks, does not make selective sampling feasible in such cases. The paper does not fully discuss the theoretical limitations that surround the selective sampling approach; limitations that could impede generalizability to different learning scenarios or concept classes. Without being well-informed about such theoretical constraints, it cannot be easy to predict how well the method will perform in diverse settings. Another direction might be connected with inductive backpropagation biases, which are not studied that well and might affect the reliability and predictability of the proposed method. Backpropagation will try to fit the data using the fewest number of units possible, which may not be in line with the goals of selective sampling. The possible results of such misalignment could be suboptimal performance on occasion. The effectiveness of selective sampling is dependent upon knowledge of the input distribution. Without it, it becomes difficult to select points from the domain without making assumptions, which may yield poor representations of the concept. Indeed, this dependence upon knowledge about the distribution is a serious limitation since it clamps flexibility of the model. The paper mentions that the models can run into pathological behavior; the use of transition area between 0 to 1 in a neural network can model uncertainty. In these cases, the region of uncertainty no longer characterizes the true uncertainty in the domain, which can mislead the selective sampling process and degrade performance. While perhaps implicit in the excerpts given, active learning is computationally more expensive than passive learning. In particular, when querying and providing a label are complex or expensive, then active learning requires additional computational overheads that need to be factored into considerations of the practical feasibility of active learning methods. It has been designed especially for neural networks executing binary concept learning. In other learning paradigms or data types, there is no generalization. One of the drawbacks of this method is a lack of generalization to other learning paradigms. This limits the wide applicability of the method. This is the delicate balance between the influence of background examples and actual training data that, in practice, is done mainly by careful adjustment of learning rates. This balance is critical for the performance of the model but may be very unobvious in practical situations. Poor adjustments can lead to either underfitting or overfitting and, therefore, improper effectiveness. Nevertheless, there are possible ways out of these drawbacks-that is, findings and further research areas discussed in the paper. For example, there is the Bayesian analysis that refines utility estimates for querying different parts of the uncertainty region in more effective ways, and it may be fruitful to pursue committee-based methods for improving the accuracy of utility estimates. These are solutions to patch up some of these limitations and help the model be more realistic in application.

CSE 584, HOMEWORK 1

Name: Yogeshvar Reddy Kallam IdNum: 920385794 email: yvk@5381@psu.edu

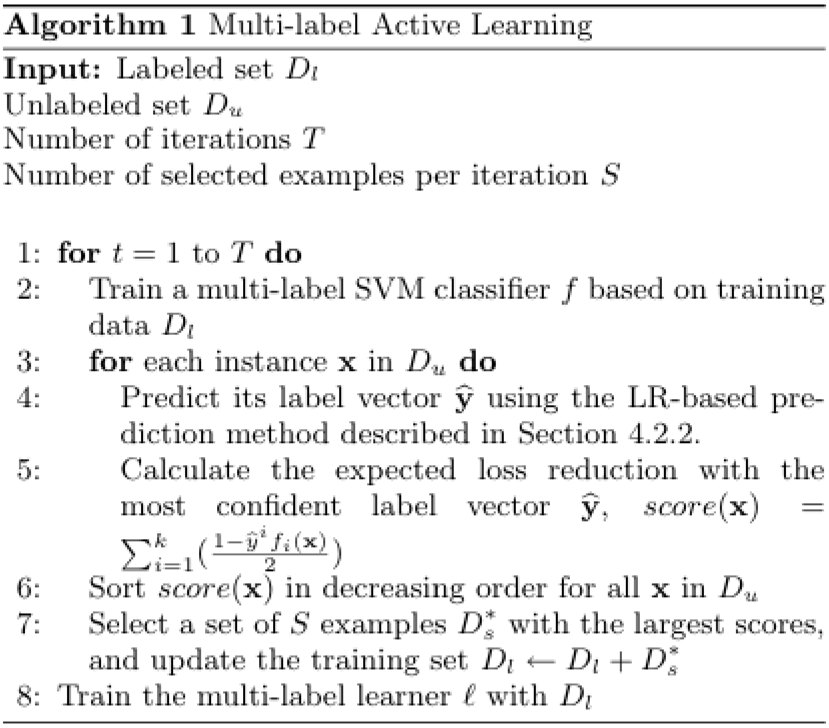
Person 3: Effective Multi-Label Active Learning for Text Classification

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

This paper addresses the problem of multi-label active learning in text classification with a view to reducing human effort and the cost involved in labeling large volumes of text data for use in training multi-label text classifiers. Conventionally, active learning algorithms address only single-label problems, in which each instance of data is associated with no more than one label. However, in real-world scenarios, most text classification tasks are intrinsically multilabelled, with a document being capable of belonging to more than one category.

1.2.1 How does it solve the problem?

The novelty in the following paper presents multi-label active learning in text classification to reduce human effort in labeling text data with high classification performance. The contribution can be highlighted in a sample selection strategy considering multi-label information for optimal reduction of the expected model loss. In their approach, the base classifier considered for the multilabel problem is a Support Vector Machine. The authors follow the one-versus-all approach where a binary classifier is trained for each class, which then decides the final labels of the given data. This method borrows strengths of SVMs in their natural habitat of binary classification and extends it to multi-label scenarios. It performs the pool-based active learning methodology where a learner selects, from a pool of unlabeled data iteratively, the most informative data for labeling while updating the classifier. This will ensure that the data points which contribute most valuable information will be selected for labeling and improve efficiency in the process of learning. Fundamentally, the method has an optimization problem to select a subset of data from the unlabeled pool such that the expected loss reduction of the multi-label classifier is maximum. This ensures the selected samples contribute highly towards enriching the model's performance. Authors in the presented approach measure the loss reduction by drawing on the concept of version space of SVM. They approximate the loss by computing the SVM margin and heuristics simplifying the computation. This allows a practical estimation of how much benefit can be achieved by labeling specific data points. It proposes a new prediction strategy for the number of labels estimation for each data point by logistic regression to predict from probabilities output by binary classifiers. The result is better estimates of label distribution for each instance. Instead of estimating the conditional probability for all possible combinations of labels, the authors approximate the expected loss function with the loss function on the most probable combination of labels. In this way, simplification on the estimation process is achieved. The approach remains effective but will be more feasible in most computations. Empirical evaluation was performed on real datasets. Its effectiveness showed that the developed method outperformed state-of-the-art active learning algorithms for multi-label text classification. These results can reveal practical benefits and robustness of the proposed method. The paper investigates a method for multi-label active learning in text classification, which integrates effective data selection for labeling. The method aims to minimize the expected loss in the multi-label classifier and human effort spent in labeling, while its results for the text classification task are promising.



Algorithm 1 Multi-label Active Learning

Input: Labeled set D_l
Unlabeled set D_u
Number of iterations T
Number of selected examples per iteration S

```
for $t = 1$ to $T$ do
    Train a multi-label SVM classifier $f$ based on training data $D_l$
    for each instance $\mathbf{x}$ in $D_u$ do
        Predict its label vector $\hat{\mathbf{y}}$ using the LR-based prediction method described in Section 4.2.2.
        Calculate the expected loss reduction with the most confident label vector $\hat{\mathbf{y}}$, $score(\mathbf{x}) = \sum_{i=1}^{k}(\frac{1-\hat{y}^i f_i(\mathbf{x})}{2})$
    Sort $score(\mathbf{x})$ in decreasing order for all $\mathbf{x}$ in $D_u$
    Select a set of $S$ examples $D_s^*$ with the largest scores, and update the training set $D_l \leftarrow D_l + D_s^*$
Train the multi-label learner $\ell$ with $D_l$
```

1.3 list of novelties/contributions

The authors have developed a new active learning approach that is specifically aimed at a text classification task involving multiple labels, as an extension of traditional active learning algorithms that are usually applicable to only single-label problems. The approach selects data to be labeled in a manner that maximizes the expected reduction rate of model loss. It strategically enhances this for the efficient learning process by targeting the most informative data. The authors are presenting an extended Support Vector Machine framework for multi-label data, with which traditionally the model has been applied to single-label classification problems. In this paper, they estimate the notion of version space to approximate the loss reduction in the multi-label context, increasing the capacity of the model to cope with more than one label simultaneously. A method of prediction for possible labels of those data points that are not labeled to estimate the expected loss is introduced. Then, the expected loss is approximated based on the most confident prediction results so that the model can keep a focus on the most relevant data points. Extensive empirical evaluations using seven real-world datasets evidence that the proposed method outperforms the current state-of-the-art active learning methods in the literature for multi-label text classification. These have shown the practical effectiveness of the approach in diverse scenarios. It will thus reduce the amount of labeled data by a large margin while sustaining high classification accuracy, hence reducing labeling cost and human effort. This reduction in labeling effort makes the method particularly valuable for applications where labeled data is scarce or expensive to obtain. Taken together, these contributions try to solve the problem of effectively training multi-label text classifiers using as little labeled data as possible, making this document quite useful for both researchers and practitioners in machine learning and text classification.

1.4 What do you think are the downsides of the work

CSE 584, HOMEWORK 1

Name: Yogeshvar Reddy Kallam IdNum: 920385794 email: yvk@5381@psu.edu

While the document provides insight into the novelty of the approach and advantages of the multi-label active learning method for text classification, it is silent on the downsides. However, from the context and challenges commonly faced in the field, a number of potential downsides can be surmised. In cases such as multi-label text classification, where some methods such as one-versus-all in SVM, this may lead to increased computational complexity with an increase in the number of labels. As a result, resource consumption and the training time will become longer, a price to be probably high for big applications. The method of label prediction proposed seems to perform better but might still face difficulties when the number of labels is huge or the different labels' distribution is highly imbalanced. That could be a source of complexity that would really influence the accuracy and efficiency of the process of label prediction. Empirically, the method is evaluated on text classification datasets, and its effectiveness on other domains and types of data, such as images or audio, would have to be validated. This neglect of generalization to other domains might limit the wider generalizability of the approach. The use of heuristics to simplify the process of version space size estimates may further be less accurate approximations, therefore facing consequences on the choice of most informative instances, affecting the overall performance. These heuristic approximations might introduce biases that affect the effectiveness of the model. In general, active learning performance is prejudged by the quality and representativeness of an initial labeled dataset. Poor choice of the initial set may lead to convergence issues for the active learning process to an optimum solution. This is another important dependence-the dependence on an initial labeled set-which could be a critical factor affecting the success of the method. While dealing with very large datasets that feature many labels and instances, scalability is not well taken care of. In this respect, considering the future trend of increasing dataset sizes, the efficiency and practicality of the current methodology may be challenged, limiting its use to large-scale scenarios. Although the methodology is proposed for cost reduction in labeling, actual cost savings are dependent a lot on the context of the application. In some scenarios, even with a reduction, the actual cost of expert labeling might still not be affordable, thus making the approach infeasible. It must be noted that the proposed approach might involve model assumptions, on which its applicability may get limited. For example, the proposed SVM-based approach assumes the data is linearly separable or at least in a kernelized form, a fact that may not hold true for real-world, complex data. In fact, these model assumptions have to be carefully considered when applying the method to a wide range of datasets. Fully understanding the downsides would involve an in-depth look at how well the methodology performs based on varying conditions: datasets used, label distributions, and computational resources. Such an analysis would provide a much fuller understanding of the method's limitations and areas that might benefit from further work.

References:

1. Learning Algorithms for Active Learning, Philip Bachman, Alessandro Sordoni, Adam Trischler. https://arxiv.org/abs/1708.00088
2. Cohn, D., Atlas, L. & Ladner, R. Improving generalization with active learning. Mach Learn 15, 201–221 (1994). https://doi.org/10.1007/BF00993277
3. Effective multi-label active learning for text classification ,Bishan Yang, Jian-Tao Sun, Tengjiao Wang, Zheng ChenAuthors Info & Claims. https://doi.org/10.1145/1557019.1557119