Cost-Based Budget Active Learning for Deep Learning

Patrick K. Gikunda
Computer Science Dept.
Paris8 University, France
kinyuagikunda@gmail.com

Nicolas Jouandeau
Computer Science Dept.
Paris8 University, France
n@up8.edu

Abstract

Majorly classical *Active Learning* (AL) approach usually uses statistical theory such as entropy and margin to measure instance utility, however it fails to capture the data distribution information contained in the unlabeled data. This can eventually cause the classifier to select outlier instances to label. Meanwhile, the loss associated with mislabeling an instance in a typical classification task is much higher than the loss associated with the opposite error. To address these challenges, we propose a *Cost-Based Bugdet Active Learning* (CBAL) which considers the classification uncertainty as well as instance diversity in a population constrained by a *budget*. A principled approach based on the min-max is considered to minimize both the labeling and decision cost of the selected instances, this ensures a near-optimal results with significantly less computational effort. Extensive experimental results show that the proposed approach outperforms several state-of -the-art active learning approaches.

1 Introduction

Active Learning (AL) considers that if a Learning Algorithm can choose an instance to label for training a model, then the instances choosen should maximize the learning performance under a fixed budget [16]. Typically this process involves randomly sampling large amount of data from underlying distribution for training a model. In *Deep Networks* (DN) acquiring labels for training set is very costly and time consuming even when using state-of-the-art computing resources. For example, its expensive to hire dermatologists to annotate 129,450 skin cancer images [5]. However, *Machine Learning* (ML) algorithms does not require all the training data to be labeled [12].

In AL there are three scenarios in which the ML algorithm will query the label of an instance, they include: a) *Membership Query Synthesis* that generates constructs of an instance from underlying distribution [2]. The queries of an instance are generated then labels are requested. In this scenario the quality of the randomly generated instances is not guaranteed; b) *Stream-Based Selective Sampling* that uses query strategy to determine whether to query the label of an instance or reject it based on some utility metrics [20]. The model sequencially picks data and checks the data one by one to determine whether to label the picked sample or not. In this scenario, selecting unlabeled instances comes at no or minimal cost; c) *Pool-Based Sampling* large pool of unlabled instances gathered at once and then examples are ranked in order of informativenes. The labels of most

In: A. Editor, B. Coeditor (eds.): Proceedings of the XYZ Workshop, Location, Country, DD-MMM-YYYY, published at http://ceur-ws.org

informative instances are queried [13]. The remainder of the paper is structured as follows: Section 2 discusses recent relevant literature on AL approaches. Section 3 describes the instance selection method used. Section 4 describes the datasets used, other AL methods and experimental results. Finally, Section 5 concludes the work by presenting some insights for further research.

2 Related Work

AL methods in deep networks use different strategies or a combination of the strategies to query for labels [7]. The strategies range from density estimation to multi-factor methods. The strategies can be categorized into two groups: *population-based* strategies and *pool-based* strategies [18]. In population-based AL, training and test sets are drawn from the same distribution with an assumption that training and test data both follow the same conditional distribution $p(y|x)$. In this type, the objective is to find the optimal training input density to generate the training input instances. In pool-based AL, the objective is to optimally select some unlabeled instances from a pool so that a model trained from them can best label the remaining samples. Regardless of whether it is population-based or pool-based, AL is an iterative process [14]. It first builds a base model from a small number of labeled training instances, and then using different or a combination of utility metrics it selects unlabeled instances and queries for their labels. The newly labeled instances are added to the training labeled set and the model is updated. This process iterates until a termination criterion is met, for example, when the labeling budget is exhausted or the maximum number of iterations is met. Based on the number of unlabeled instances to query at each iteration, AL methods can be grouped as either sequence-mode AL, where one instance is queried each time or batch-mode AL, where multiple instances are queried at each iteration [14].

This paper focuses on pool-based batch-mode AL for DN. Although numerous AL approaches have been proposed in the literature [16], a number of them are for standard pool-based AL problems e.g. Donmez *et al.* presents a dynamic approach that updated selection parameters based on estimated future residual error reduction after each actively sampled instance [3], Settles and Craven formulates the implementation of AL using information density [17] and Krempl *et al.* implements cost-sensitive probabilistic approach for binary classification [11]. Among those approaches limited to AL for DN include the work presented by Ducoffe and Precioso that uses margin theory to compute instances along the decision boundary, [4], [1], [8] and [9].

The emphasis in AL is to evaluate the informativeness of an instance, with an assumption that an instance with higher classification uncertainty is more crucial to label. This classical approach usually uses statistical theory such as entropy and margin to measure instance utility, however it fails to capture the data distribution information contained in the unlabeled data. This can eventually cause the classifier to select outlier instances to label. Therefore, its important to consider the classification uncertainty as well as instances diversity in a population while developing an AL solution. In our approach we consider both the uncertainty and correlation measure to calculate the most informative and representative instance, which we refer as a high confidence instance.

In AL where there is a pre-determined budget on labeling, it is important to estimate the objective function for data selection. This guarantees near-optimal results with significantly less computational effort. The aim is to maximize the objective function while minimizing data acquisition costs (or to remaining within a budget). To deal with this problem we formulate the most informative budget selection task as a continuous optimization problem. The aim is to determine possible queries that maximize the improvement to the classifiers strategy, without overspending the budget. The proposed model addresses the cost-sensitive learning problem based on learning algorithms that construct models for class probability estimation $p(y|x)$. The probability estimates provide an easy means for factoring in the misclassification losses in the classification decision making step. To address the labeling cost problem, we employ the same probability estimation techniques over the selected high confidence instances. The two methods are then combined into an algorithm that minimizes the combined cost. At each iteration (while the total sum of annotation cost is under a given budget) a high confidence instance [1] to label is selected.

3 Selecting Most Informative Instance

We consider the problem of actively selecting a batch of instances to label, where the contents of the batch must be constrained by some budget. We will use the following notation in this paper. Let x_i represents an instance and y_i where $y_i \in \{1, +1\}$ represents the class label for x_i, $D = D^L \cup D^U$, D^L denotes labeled instances where

[1] High confidence instances are the most informative and representative instances selected from the unlabeled set

$D^L = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, D^U denotes unlabeled instances where $D^U = \{(x_1, ?), (x_2, ?), ..., (x_n, ?)\}$, D^H denotes high confidence instances and Θ denotes the model defined by model parameters. For label space Y with K classes in D we use the class probability estimator $P(y|x)$ to compute the estimate of a label. In order to avoid the problem of generalization of unseen instances and to learn an accurate model, we present a robust approach that uses different utility metrics and a cost function. The utility metrics considered in this work are uncertainty, correlation and informativeness measure, thus we present three main components of our approach: a) uncertainty measure, b) correlation measure and the cost-based labeling.

3.1 Uncertainty Measure

Given a label space Y the uncertainty measure f_u of an instance considering both the features and the label can be defined as:

$$f_u(x) : \begin{cases} L^S \rightarrow R, & \text{(i) features view} \\ (D^U \times L^S) \rightarrow R, & \text{(ii) features-label view} \end{cases} \tag{1}$$

to a real number space R. From Equation 1 above: (i) the uncertainty measure is computed from instance-features only while (ii) the uncertainty measure is computed from both the instance-features and instance-label. In our method we consider the uncertainty measure computed from instance features and label view which is considered the most effective [10]. Out of the three common uncertainty measure criteria namely least confidence, sample margin and entropy, sample margin is prefered since it integrates the second most probable class label in the uncertainty metric hence able to reduce the error rate by defining the decision boundary. We therefore define uncertainty measure as:

$$f_u(x) = P(y_i = l_1|x_i; Y) - P(y_i = l_2|x_i; Y) \tag{2}$$

With l_1 and l_2 being the most likely and second most likely labels. High uncertainty value f_u implies current model have little knowledge of the instance, and including it into the training set can help improve to the prediction performance of the model.

3.2 Correlation Measure

When developing efficient AL methods, it is critical to consider samples distribution information [19]. The instance diversity information aids in selecting most representative instances. In order to have more information about the unlabeled instances it is appropriate to select a candidate instance in a more dense region. In addition, selecting an instance to label only based on uncertainty measure may lead to selecting an outlier instance, therefore exploiting sample instance diversity will provide the most informative instance to label. Our method is based on the fact that the trade off between instance uncertainty and correlation is an essential AL problem to address. Given a label space Y, we can define different groups of correlation of an instance x in a set of unlabeled set as;

$$f_c(x) : \begin{cases} D^U \times D^U \rightarrow R, & \text{feature view} \\ Y \times Y \rightarrow R, & \text{label view} \\ (D^U, y) \times (D^U, y) \rightarrow R, & \text{combined view} \end{cases} \tag{3}$$

to a real number space R. In Equation 3, the combination of feature and label correlation is called combined view. Different algorithms exist for exploiting this type of combination [10]. Majorly these algorithms are used in a multi-label learning tasks when an instance has more than one label. This setting is ideal for mining tasks on instances with complex structure. In this work we focus on exploiting the pairwise similarities of instances, therefore the informativeness of an instances is weighed by average similarity to its neighbours. Let x_i and x_j be a pair of instances. To cope with the drawback of uncertainty based selection, we then consider the diversity by evaluating the correlation of the instances. Given a label space Y the correlation measure $f_c(x_i, x_j)$ between a pair of instances x_i and x_j can be defined as:

$$f_c(x) = \frac{1}{D^U} \sum_{x_j \in D^U} (x_i, x_j) \tag{4}$$

The value of $f_c(x_i)$ represents the instance density of x_i in the unlabeled set. The larger the value, the more densely an instance is correlated with others. A low value of the correlation measure indicates an outlier

instance which should not be considered for labeling. Our motivation is that the most representative instances of a distribution can be very informative for improving the generalization performance. Therefore, given correlation measure $f_c(x_i)$ and uncertainty measure $f_u(x_i)$ the informativeness of an instance can be defined as:

$$f_i(x) = f_u(x_i) \cdot f_c(x_i) \tag{5}$$

It can be rewritten as:

$$x_i = \underset{i}{\operatorname{argmax}}(u_i \cdot c_i) \tag{6}$$

3.3 Cost-Based labeling

In our approach the high confidence instance evaluation is based on the instance informativeness which is computed from both uncertainty and correlation metrics. The model is trained on labeled instances: feature and label views. After querying for an high confidence unlabeled instance, a model prediction result is generated based on output probability distribution. Each instance $x_i = \{f_1^i, f_2^i, ...f_q^i, y^i\}$ in labeled set $D^L = \{x_1, x_2, ...x_s\}$ is represented in a feature space F consisting of a feature space and its class label y^i. The size of D^L is denoted by s and x_i denoted the ith instance in D^L. The prediction can be denoted as a mapping function from the feature space F to the class label space Y which can be expressed as;

$$P(x) : F \mapsto Y \tag{7}$$

The query strategy used in this work is based on the value of f_i discussed in equation 6. Instances are ranked based on the value f_i with top ranked instances being the most appropriate to label. Under the current distribution $P(y_i|x_i; Y)$ each possible instance $(x_i, ?)$ from the selected instances D^H will be labeled with label y_i. When $y_i = 1$, x_i is regarded as a high confidence sample. The model update strategy is to train a model based on the information provided by model weights computed from model validation of the performance. We employ the probability estimators in our approach to both minimize the labeling costs and the misclassification decisions. We make the assumption that the loss function associated with the decisions is represented as a static K and a loss matrix L available at learning time. The contents of $L(i, j)$ specify the cost incurred when an example is predicted to be in class i when in fact it belongs to class j. Therefore, high confidence instance selection criteria in this study will be based on probability of x_i belonging to K^{th} class which can be expressed as:

$$y_i = \underset{\xi \in D^H}{\arg\min}(C(\xi) + \sum_{k=1}^{K} P(k|\xi) \sum_{i} \sum_{j=1}^{K} P_{\xi,k}(j|x_i) L(y_i, j)) \tag{8}$$

The Algorithm 1 describes the *Cost-Based Budget Active Learning* (CBAL) with budget labeling.

Algorithm 1: Cost-Based Budget Active Learning (CBAL).

Input: labeled instance set D^L, unlabeled instance set D^U, loss matrix L, labeling cost C, empty set D^H, a budget m;
Output: model Θ;
$\Theta \leftarrow$ `getModel` (D^L);
while $|D^L| < m$ **do**
  for *each* x_i *in* D^U **do**
    $u_i \leftarrow f_u(x_i)$;
    $c_i \leftarrow f_c(x_i)$;
    $x^* \leftarrow \underset{i}{\operatorname{argmax}}(u \cdot c)$;
    $D^H \leftarrow D^H \cup \{x\}$;
  for each j learn $P(y|D^H)$;
  $y_i \leftarrow$ `getLabel using Eq.8`;
  $D^H \leftarrow D^H \setminus \{y_i\}$;
  $D^L \leftarrow D^L \cup \{y_i\}$;
  $\Theta \leftarrow$ `updateModel` (D^L);
return Θ;

In Algorithm 1, the labeling is defined by the budget m with model updates after each iteration (lines 4-14). At first the base model is trained using the initial set of labeled data D^L. Instance evaluation is done to identify the most informative and representative instance to label (lines 5-9). This evaluation returns the high confidence instances D^H selected from the unlabeled population (line 9). For each of the selected instance, its label is queried and consequently the labeled set is updated. The model selection strategy is updated with the learned parameters after every iteration. CBAL is designed to train a classification model using a small labeled population sample proportion.

4 Experiments

We conduct experiments with 12 real-world data sets (*wine*[2], *seeds*[3], *v2-plant seedling*[4], *liver*[5], *sonar*[6], *vehicle*[7], *breast*[8], *diabetic*[9], *heart*[10], *isolet*[11], *plant*[12], *svhn*[13]) previously considered by other authors in similar domain. For each data set, we split 80% of the instances as the training set, and the balance 20% as the test set to evaluate the prediction accuracy of the models. We select a subset of instances from the training data to query (100 instances per query) for labels and then construct a base classification model according to these labeled data. The goal is to pick out high confidence instances such that the constructed model maintains effective classificaion ability. The training is implemented in a batch-mode AL. We compare our method with other state-of-the-art methods: a). Random Sampling (RS) which selects a certain number of samples from a given set and quire labels; b). Core-Set AL (CSAL) [15] which defines the AL problem as a competitive sample core-set selection which is then applied to a CNN in a batch setting; c). Deep Bayesian Active Learning (DBAL): a Bayesian framework proposed in [6] for high dimensional data which considers Deep Learning problem of dependence on big amount of data; d). Adversarial AL for deep networks (AAL) a margin based approach proposed in [4] for deep networks with intention of reducing the number of queries to the oracle during training. The both the budget m and the initial labeled set is specified before start of iterations. A batch size of 64 was considered for all iterations for both training and testing selection. 100 queries were considered for each iteration.

4.1 Results and Discussion

Figure 1 shows the classification accuracy of different active learning approaches with varied number of queries. From the observation RS tends to yield better performance when the number of queries is small but as the number of queries increases it starts to slow its effectivenes in prediction. This observation might be as a result of sampling bias induced by an intelligent selection strategy. CSAL that define AL as a core-set problem, is not performing well at the start of training. As the number of queries increases, there is improvement and yields better performance. This is because with few training instances, the learned decision boundary tends to be inaccurate, and as a result, the unlabeled instances near the decision boundary may not be the most high confidence instances to label. The performance of DBAL is better on some datasets but performs poorly on others. This inconsistency might be as a result of identified cluster structure of unlabeled data that is not always consistent with the target classification model. The behavior of AAL is similar to that of DBAL. Finally, we observe that for most cases, CBAL is able to outperform the baseline methods significantly and we attribute this success to the principle selecting high confident samples at each iteration, and the specially aspect of minimizing the labeling and decision cost after each subsequent iteration.

5 Conclusion and Future Work

We propose a new near optimal AL approach called CBAL, that measure both the informative and representative of an instance using instance utility to get a high confidence instance to lable while minimizing the labeling and

[2]http://archive.ics.uci.edu/ml/datasets/Wine
[3]https://archive.ics.uci.edu/ml/datasets/seeds
[4]https://www.kaggle.com/vbookshelf/v2-plant-seedlings-dataset
[5]https://archive.ics.uci.edu/ml/datasets/Liver+Disorders
[6]https://archive.ics.uci.edu/ml/datasets/Connectionist+Bench+(Sonar,+Mines+vs.+Rocks
[7]https://archive.ics.uci.edu/ml/datasets/Statlog+Vehicle+Silhouettes
[8]https://www.kaggle.com/paultimothymooney/breast-histopathology-images
[9]https://www.kaggle.com/sovitrath/diabetic-retinopathy-224x224-gaussian-filtered
[10]https://archive.ics.uci.edu/ml/datasets/Heart+Disease
[11]https://archive.ics.uci.edu/ml/datasets/isolet
[12]https://www.kaggle.com/vipoooool/new-plant-diseases-dataset
[13]https://www.kaggle.com/stanfordu/street-view-house-numbers

decision cost. The proposed approach of minimizing cost is based on the minmax principled view. Our current work is based on budget constrained learning with pairwise similarities of instances. In the future, we plan to extend this work to multi-label learning tasks by considering instances with more than one label. In addition we plan to consider the expert knowledge in the training by allowing the user to control tradeoff between selection and labeling, this will lead to incoporating domain knowledge into AL

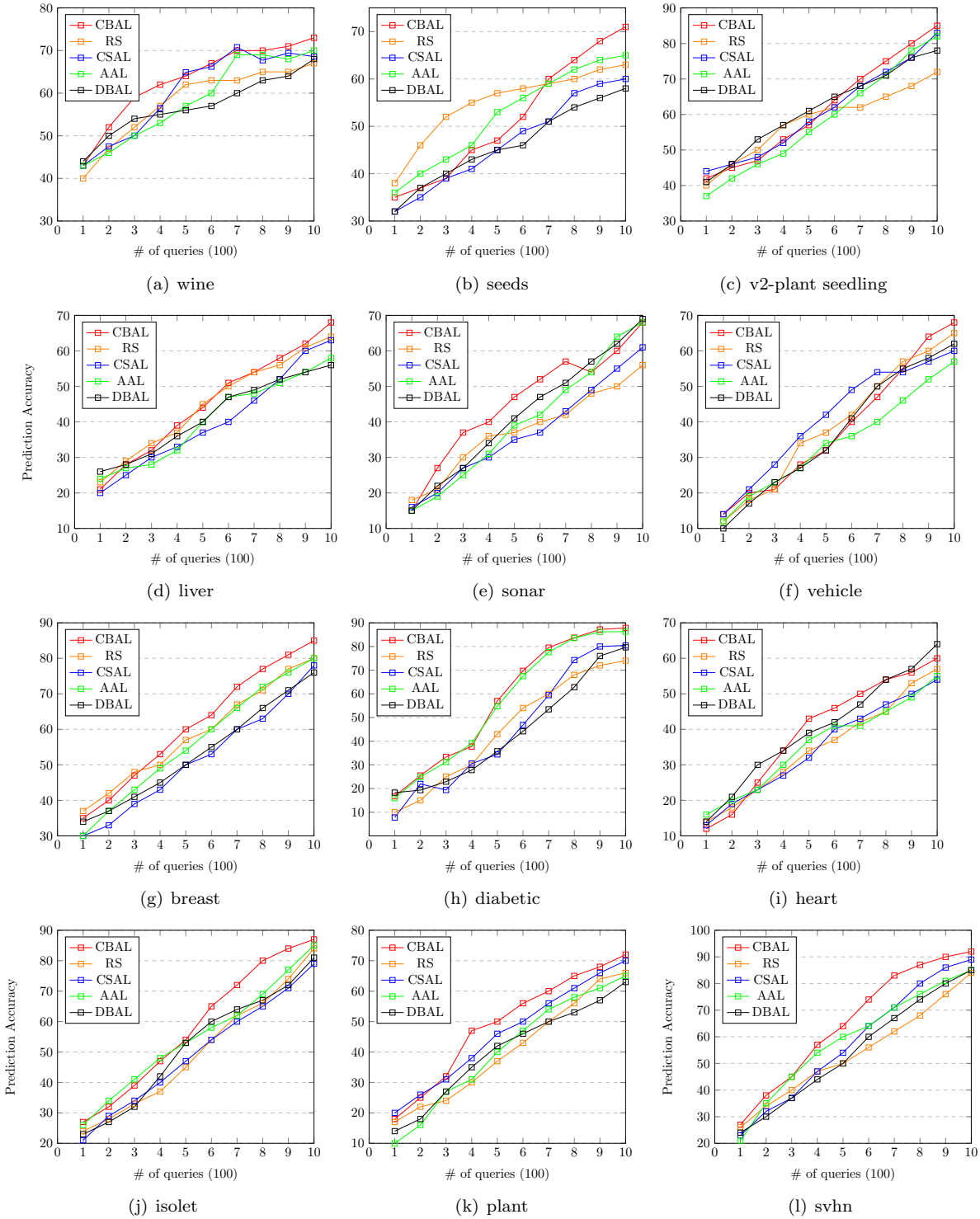

Figure 1: Classfication Accuracy on different datasets

References

[1] Hamed H Aghdam, Abel Gonzalez-Garcia, Joost van de Weijer, and Antonio M López, 'Active learning for deep detection neural networks', in *IEEE International Conference on Computer Vision (ICCV)*, pp. 3672–3680, (2019).

[2] Dana Angluin, 'Queries and concept learning', *Machine learning*, **2**(4), 319–342, (1988).

[3] Pinar Donmez, Jaime G Carbonell, and Paul N Bennett, 'Dual strategy active learning', in *European Conference on Machine Learning (ECML)*, pp. 116–127. Springer, (2007).

[4] Melanie Ducoffe and Frederic Precioso, 'Adversarial active learning for deep networks: a margin based approach', *arXiv preprint arXiv:1802.09841*, (2018).

[5] Andre Esteva and Brett Kuprel *et al*, 'Dermatologist-level classification of skin cancer with deep neural networks', *Nature*, **542**(7639), 115–118, (2017).

[6] Yarin Gal and Riashat Islam *et al*, 'Deep bayesian active learning with image data', in *International Conference on Machine Learning (ICML)*, pp. 1183–1192, (2017).

[7] Martina Hasenjäger and Helge Ritter, 'Active learning in neural networks', in *New Learning Paradigms in Soft Computing*, 137–169, Springer, (2002).

[8] Manuel Haussmann, Fred A Hamprecht, and Melih Kandemir, 'Deep active learning with adaptive acquisition', *arXiv preprint arXiv:1906.11471*, (2019).

[9] Steven CH Hoi, Rong Jin, Jianke Zhu, and Michael R Lyu, 'Batch mode active learning and its application to medical image classification', in *International Conference on Machine Learning (ICML)*, pp. 417–424, (2006).

[10] Sheng-Jun Huang, Nengneng Gao, and Songcan Chen, 'Multi-instance multi-label active learning.', in *International Joint Conferences on Artificial Intelligence (IJCAI)*, pp. 1886–1892, (2017).

[11] Georg Krempl, Daniel Kottke, and Vincent Lemaire, 'Optimised probabilistic active learning (opal)', *Machine Learning*, **100**(2-3), 449–476, (2015).

[12] Mingsheng Long and Han Zhu *et al*, 'Deep transfer learning with joint adaptation networks', in *International Conference on Machine Learning (ICML)*, pp. 2208–2217, (2017).

[13] Kamal Nigam and Andrew McCallum, 'Pool-based active learning for text classification', in *Conference on Automated Learning and Discovery (CONALD)*, (1998).

[14] Hiranmayi Ranganathan, Hemanth Venkateswara, Shayok Chakraborty, and Sethuraman Panchanathan, 'Deep active learning for image classification', in *IEEE International Conference on Image Processing (ICIP)*, pp. 3934–3938, (2017).

[15] Ozan Sener and Silvio Savarese, 'Active learning for convolutional neural networks: A core-set approach', *arXiv:1708.00489*, (2017).

[16] Burr Settles, 'Active learning literature survey', Technical report, University of Wisconsin-Madison, Dept of Computer Sciences, (2009).

[17] Burr Settles and Mark Craven, 'An analysis of active learning strategies for sequence labeling tasks', in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1070–1079, (2008).

[18] Masashi Sugiyama and Shinichi Nakajima, 'Pool-based active learning in approximate linear regression', *Machine Learning*, **75**(3), 249–274, (2009).

[19] Martin Szummer and Tommi S. Jaakkola, 'Information regularization with partially labeled data', in *Advances in Neural Information Processing Systems (NIPS)*, pp. 1049–1056, (2003).

[20] Xingquan Zhu and Peng Zhang *et al*, 'Active learning from data streams', in *IEEE International Conference on Data Mining (ICDM)*, pp. 757–762, (2007).