DEEP ACTIVE LEARNING FOR IMAGE CLASSIFICATION

Hiranmayi Ranganathan, Hemanth Venkateswara, Shayok Chakraborty, Sethuraman Panchanathan

Center for Cognitive Ubiquitous Computing (CUbiC)
Arizona State University

ABSTRACT

In the recent years, deep learning algorithms have achieved state-of-the-art performance in a variety of computer vision applications. In this paper, we propose a novel active learning framework to select the most informative unlabeled samples to train a deep belief network model. We introduce a loss function specific to the active learning task and train the model to minimize the loss function. To the best of our knowledge, this is the first research effort to integrate an active learning based criterion in the loss function used to train a deep belief network. Our extensive empirical studies on a wide variety of uni-modal and multi-modal vision datasets corroborate the potential of the method for real-world image recognition applications.

Index Terms— Computer vision, deep learning, deep belief networks, active learning, entropy

1. INTRODUCTION AND BACKGROUND

In recent years, deep learning has emerged as a dominant machine learning tool for a wide variety of domains [13].Deep architectures have been widely explored in computer vision and have achieved tremendous improvement in several vision tasks including image recognition [12], object detection [6], and image segmentation [15] among others. The surge of deep learning started in 2006 when Hinton *et al.* [8], introduced Deep Belief Networks (DBNs). DBNs and its variants have been shown to depict excellent performance in several applications including visual object recognition [19], emotion recognition [17], speech phone recognition [4] and image denoising [18].

A fundamental challenge in training a deep neural network is the requirement of large amounts of labeled training data. While gathering large quantities of unlabeled data is cheap and easy, annotating the data (with class labels) is an expensive process in terms of time, labor and human expertise. Thus, developing algorithms to minimize human effort in training deep models is of paramount practical importance. Active learning algorithms automatically identify the salient and exemplar samples from large amounts of unlabeled data and reduce human annotation efforts in inducing a classification model.A comprehensive review of several active learning algorithms developed over the last several years can be found in [21]. Specifically, batch mode active learning (BMAL) algorithms have been widely used in computer vision applications with promising empirical results [3, 9, 7, 2].

Even though both deep learning and active learning have been extensively studied, research on combining the two is still in a nascent stage. Wang and Shang [24] proposed AL-DL, an active labeling method for deep learning with DBNs. Stark *et al.* [23] presented an active learning algorithm using CNNs for CAPTCHA recognition. Along similar lines, Zhou *et al.* [25] proposed the active deep network (ADN) framework for sentiment classification.

All these algorithms treat active learning and deep model training as two independent problems. A deep model is first learned using a conventional loss function; the active sampling condition is then defined based on the posterior probabilities obtained from the last layer or the distance of a sample from the decision boundary. However, the merit of a deep model lies in its ability to learn a discriminating set of features for a given task; this property has not been leveraged in the existing algorithms combining deep learning and active learning. In this paper, we propose a novel deep active learning algorithm which is designed to exploit this property and study its performance on a wide variety of computer vision applications. We now describe the proposed framework.

2. PROPOSED FRAMEWORK

The core idea of this research is to leverage the feature learning capability of deep models to identify the most informative unlabeled samples for active learning. To achieve this, we append an entropy based term to the conventional softmax loss term and train the network to optimize the joint loss function.

Formally, let $X = \{x_1, x_2, \dots, x_n\}$ be the training set containing n samples. The subset of labeled samples is represented as $X_l = \{x_1, x_2, \dots, x_{n_l}\}$. The corresponding labels for X_l are denoted by $Y_l = \{y_1, y_2, \dots, y_{n_l}\}$. Let the subset of unlabeled data points be, $X_u = \{x_{n_l+1}, x_{n_l+2}, \dots, x_{n_l+n_u}\}$. $X = X_l \cup X_u$, is the union of the disjoint subsets X_l and X_u. Therefore, $n = n_l + n_u$. The goal is to estimate a classifier function $f(x)$, using the labeled data $\mathscr{D} = \{X_l, Y_l\}$. Here, $f(x_i)$, $\forall i \in [1, \dots, n_l]$ is the conditional probability that the classifier assigns x_i to label y_i. The classifier function $f(.)$ is then applied on the unlabeled data $f(x_i)$, $\forall i \in [n_l{+}1, \dots, n]$, to predict the label $\hat{y}_i$. The accuracy of the classifier is tested by comparing the predicted labels of the unlabeled data with the ground truth labels for the unlabeled data. Since the classifier $f(.)$ is implemented using a deep belief network, we use the standard cross-entropy loss to estimate the empirical classification error $E(\mathscr{D}; f)$, which is given by,

$$\operatorname*{argmin}_{f \in \mathscr{F}} E(\mathscr{D}; f) = \frac{1}{n_l} \sum_{i=1}^{n_l} L(f(x_i), y_i), \quad (1)$$

where the cross-entropy loss is given by,

$$L(f(x_i), y_i) = -\sum_{j=1}^{C} 1\{y_i = j\} \text{log} f_j(x_i), \; \forall i \in [1, \ldots, n_l]. \quad (2)$$

Here, C is the total number of label categories and $1\{.\}$ is the indicator function. $f_j(x_i) = e^{h_{ij}^N} / \sum_{j'} e^{h_{ij'}^N}$ is the softmax function defined on the activation h_{ij}^N, where h_{ij}^N is j-th component of the i-th data point in the N-th (final) layer of the network. The softmax function ensures $f(x_i) = [f_1(x_i), f_2(x_i), \ldots, f_C(x_i)]^\top$ is a probability vector with $f_j(x_i)$ being the probability that data point x_i is assigned to category C. During the process of active learning the labeled dataset is repeatedly augmented with newly obtained labeled data taken from the unlabeled dataset. The classifier $f(.)$ is in turn updated by retraining with the updated labeled dataset.

From the unlabeled set X_u, an oracle is given a batch of points B to be labeled. The labeled batch B is then combined with the labeled set X_l along with the corresponding labels (added to Y_l) that are provided by the oracle, i.e. $X_l \rightarrow X_l \cup B$. These data points are removed from the unlabeled set X_u in order to ensure X_l and X_u are disjoint, i.e. $X_u \rightarrow X_u \backslash B$. A new and improved classifier is estimated using the augmented labeled sets $\{X_l, Y_l\}$. This procedure is repeated until we run out of budget to get labeled data from the oracle. Given the probability of label assignment for a data point, entropy (from Information Theory) can be used to obtain a measure of uncertainty regarding its label assignment. The set B can therefore be chosen by selecting the data points with the largest uncertainty. Entropy based measures have been applied previously to select a set B with the most informative (highest uncertainty) data points in order to estimate an active learning based classifier [3]. The entropy can be expressed in terms of assigned label probabilities for the unlabeled data as,

$$H(f(x_i)) = -\sum_{j=1}^{C} f_j(x_i) \text{log} f_j(x_i), \; \forall i \in [n_l + 1, \ldots, n] \quad (3)$$

where $f_j(x_i)$ is the probability of assigning x_i to category j. We define a probability vector for x_i as $p_i := [f_1(x_i), f_2(x_i), \ldots, f_C(x_i)]^\top$. In standard active learning settings, a classifier is first trained on the labeled data and used to obtain the predictions for the unlabeled data. Entropy is then applied to obtain the uncertainty of such a classifier prediction. In this two-step approach, the unlabeled data does not play a role in training the classifier. We propose an active learning model where we combine the entropy measure along with the cross-entropy loss during training. We discuss the benefits of this joint loss in the following section.

2.1. Joint Loss for Active Learning

The DBN is trained by combining both the labeled and unlabeled data with the aim to obtain least entropy on the unlabeled data and least cross-entropy on the labeled data. The network restructures its weights while minimizing cross-entropy (for labeled data) and minimizing entropy (for unlabeled data)in one step. The joint loss ensures that the data points with the largest entropy that are selected to form B, are the most uncertain and informative unlabeled data points with respect to the classifier $f(.)$. Over successive iterations, the positive effects of this joint training get enhanced. Based on this intuitive reasoning, we combine the cross-entropy loss in Equation (2) and the entropy in Equation (3) to formulate a classifier with a joint loss that is given by,

$$\begin{aligned} \operatorname*{argmin}_{f \in \mathscr{F}} E(\mathscr{D}; f) = & \frac{1}{n_l} \sum_{i=1}^{n_l} L(f(x_i), y_i) \\ & + \frac{\lambda}{n_u} \sum_{i=n_l+1}^{n} H(f(x_i)). \end{aligned} \quad (4)$$

where λ controls the relative importance of the entropy loss.

The output of the N-th layer of the network (before the loss) for a data point x_i, is given by the vector h_i^N. We define $p_{ij} := f_j(x_i) = e^{h_{ij}^N} / \sum_{j'} e^{h_{ij'}^N}$, the probability that data point x_i belongs to class j. The loss in terms of probabilities is given by,

$$\begin{aligned} E(X_l, X_u, Y_l) = & -\frac{1}{n_l} \sum_{i=1}^{n_l} \sum_{j=1}^{C} 1\{y_i = j\} \text{log} p_{ij} \\ & - \frac{\lambda}{n_u} \sum_{i=n_l+1}^{n} \sum_{j=1}^{C} p_{ij} \text{log} p_{ij}. \end{aligned} \quad (5)$$

We outline the derivative of the loss $E(.)$ with respect to h_{pq}^N, which is the q-th component of the p-th data point in the output of the N-th layer as,

$$\frac{\partial E}{\partial h_{pq}^N} = \begin{cases} \frac{1}{n_l} \Big(p_{pq} - 1\{y_p = q\} \Big), & p \in [1, \ldots, n_l] \\ \frac{\lambda}{n_u} p_{pq} \Big(\sum_j^C p_{pj} h_{pj}^N - h_{pq}^N \Big), & p \in [n_l + 1, \ldots, n]. \end{cases} \quad (6)$$

During the training procedure, the derivative $\partial E / \partial h^N$ is backpropagated through the network in order to update the weights of the network.

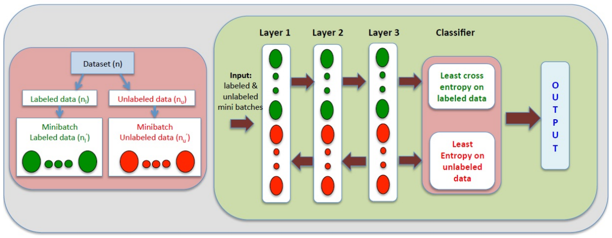


Fig. 1. Deep Active Learning Network Architecture. Best viewed in color.

2.2. Active Learning Network Architecture and Training

Figure 1 illustrates the network architecture of our Deep Active Learning model. Our model is a three layer DBN [8]. Since the number of data points $n = n_l + n_u$, is usually very large, we use mini-batch based gradient descent to train the network. We present the network with a mini-batch of n' data points which consists of n'_l labeled data points (green circles) and n'_u unlabeled data points (red circles), i.e. $n' = n'_l + n'_u$ with $n'_l \leq n_l$ and $n'_u \leq n_u$. The cross-entropy loss is computed over the labeled data in the mini-batch and entropy loss is computed over the unlabled data in the mini-batch. The negative gradient of the joint loss function with respect to the mini-batch is backpropagted in order to train the DBN. When the network has seen all the data points in the training set (labeled and unlabeled), we consider it as one epoch. We repeat the training procedure over multiple epochs until convergence and consider this as one training iteration t of the active learning algorithm. At the end of every iteration t, we sample the most informative batch of unlabeled data points (using Equation 3) to form B. We obtain the labels for B using an oracle and update the labeled and unlabeled datasets, as discussed earlier. We iterate until we run out of unlabeled data to be labeled or we run out of budget to get them labeled.

3. EXPERIMENTS AND RESULTS

Our DBN model consists of three hidden layers, with 500 units in each hidden layer. In the unsupervised learning stage, we used a learning rate of 0.05 and number of epochs as 100. We employed minibatch ($n' = 100$), momentum (0.5 for first 5 epochs and 0.9 later) and weight decay (0.0002). In the supervised finetuning stage, we used learning rate = 0.05, number of epochs = 50 and minibatch size = 100 (with 50 labeled and 50 unlabeled data samples). The same deep architecture and parameters were used for all the competing methods for fair comparison. The weight parameter λ was selected as 1 based on preliminary experiments.

We studied the performance of the algorithm on a variety of uni-modal and multi-modal datasets from different application domains. We used four uni-modal datasets in our experiments: (i) the **VidTIMIT** face recognition dataset [20]; (ii) the **Cohn-Kanade (CK)** AU-Coded Expression Database [10]; (iii) the **MNIST** database of handwritten digits [14]; and (iv) the **CIFAR 10** dataset [11] for object recognition. We also validated the performance of our algorithm on two multi-modal datasets for emotion recognition: (i) **emoFBVP** [17] and (ii) **MindReading** [5].

Our objective was to test the performance of the proposed active sampling framework for deep learning and not to outperform the best accuracy results on these datasets; so, we did not follow the precise train/test splits given for many of these datasets. Each dataset was divided into an initial training set, an unlabeled set and a test set. For a given batch size k, each algorithm selected k instances from the unlabeled pool to be labeled in each iteration. After each iteration, the selected points were removed from the unlabeled set, appended to the training set and the performance was evaluated on the test set. The goal was to study the improvement in performance on the test set with increasing sizes of the training set. The experiments were run for 20 iterations. This setup is similar to previous work [24]. The dataset details are summarized in Tables 1 and 2.

Dataset	Training	Unlabeled	Testing	Batch Size
VidTIMIT	500	20000	8000	100
CK	500	10000	5000	100
MNIST	1000	50000	10000	200
CIFAR	1000	45000	10000	200

Table 1. Uni-modal Dataset Details.

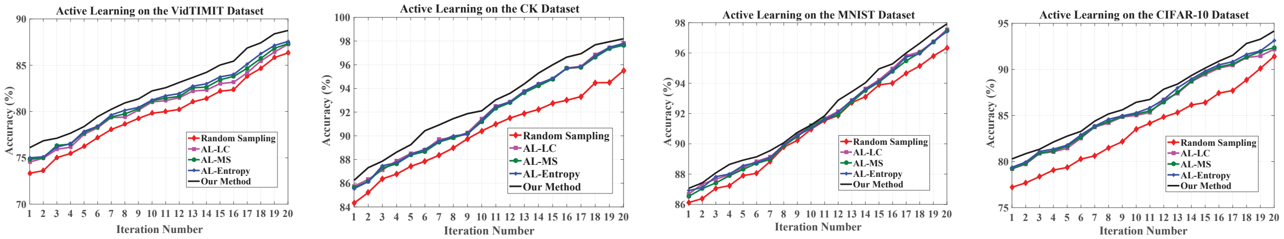
Dataset	Training	Unlabeled	Testing	Batch Size
emoFBVP	400	20000	10000	80
MindReading	1000	70000	10000	200

Table 2. Multi-modal Dataset Details.

3.1. Comparison Baselines and Performance

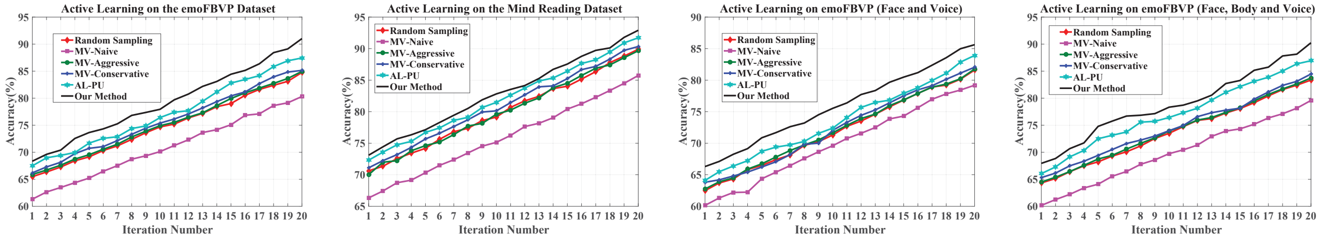
Uni-modal datasets: We used the following algorithms as baselines for comparison: (i) **Random Sampling**, which selects a batch of unlabeled samples at random from the unlabeled pool; (ii) **Active Labeling with Least Confidence (AL-LC)** [24] which selects the samples with the smallest of the maximum activations; (iii) **Active Labeling with Margin Sampling (AL-MS)** [24] which selects the samples with the smallest separation between the top two class predictions; and (iv) **Active Labeling with Entropy (AL-Entropy)** [24], which selects the unlabeled samples with the largest class prediction information entropy.

The results on the uni-modal datasets are depicted in Figure 2. In each graph, the x-axis denotes the iteration number and the y-axis denotes the accuracy on the test set. The proposed framework outperforms Random Sampling on all the datasets; the accuracy increases at a faster rate with increasing size of the labeled set. Our algorithm therefore identifies the salient and exemplar instances for manual annotation and attains a given level of performance with much reduced human labeling effort. The AL-LC, AL-MS and AL-Entropy methods depict better performance than Random Sampling,



(a) VidTIMIT (b) CK (c) MNIST (d) CIFAR-10

Fig. 2. Active Learning on the Uni-modal Datasets. Best viewed in color.



(a) EmoFBVP (b) Mind Reading (c) EmoFBVP: Face and Voice (d) EmoFBVP: Face, Body and Voice

Fig. 3. (a) and (b), Active Learning on the Multi-modal Datasets emoFBVP and MindReading. (c) and (d) Active Learning on Subsets of Modalities of the emoFBVP Multi-modal Dataset. Best viewed in color.

but are not as good as our method. This corroborates the merit of integrating an entropy based term in the loss function to train the deep belief network and learning the features accordingly, so that the unlabeled samples selected for annotation are maximally informative. The results unanimously lead to the conclusion that our algorithm depicts the best performance consistently across all the datasets.

Multi-modal datasets: Muslea *et al.* [16] proposed the *Multi-view (MV)* algorithm in which a separate classification model was trained for each modality (view). A set of *Contention Points* was identified from the unlabeled set, where at least two models produced different predictions; samples were queried from this set using three selection strategies: (i) **MV-Naive**, (ii) **MV-Aggressive** and (iii) **MV-Conservative**. Cebron and Berthold [1] used the same framework and incorporated sample diversity along with uncertainty (entropy) in batch selection. We compared our algorithm against all these methods, together with Random Sampling.

The results on the multi-modal emoFBVP and MindReading datasets are depicted in Figure 3(a) and 3(b). We note that Random Sampling depicts comparable performance as the MV-Aggressive and MV-Conservative methods. Thus, a simple method like random selection can sometimes depict good performance. The PU algorithm combining uncertainty and diversity depicts better performance than the Multi-view active learning algorithms. Our method demonstrates the best performance on both datasets; at any given iteration, it attains the highest accuracy on the test set. Thus, a deep belief network trained to minimize the cross-entropy loss on the labeled data together with the entropy loss on the unlabeled data succeeds in selecting the exemplar unlabeled samples for manual annotation in both uni-modal and multi-modal settings and achieves a given level of accuracy with the least amount of human effort.

We also studied the performance of our framework on different subsets of modalities of the emoFBVP dataset. Figure 3(c) and Figure 3(d) depict the results when using only the face and voice modalities and only the face, body and voice modalities respectively. The results depict a similar trend, further corroborating the generalizibility of our framework. A two-sided paired *t-test* at the significance level of $\alpha < 0.05$ reveals that the improvement in performance achieved by our method is statistically significant for all the datasets. From Figures 3(c), 3(d) and 3(a) we also note that the accuracy increases as more modalities are included in the dataset, which is intuitive.

4. CONCLUSIONS

In this paper, we proposed a novel algorithm to actively sample unlabeled instances that are most promising in training a deep belief network model. We introduced a loss function based on softmax and entropy losses and trained the deep model to optimize the loss function. To the best of our knowledge, this is the first research effort to incorporate an active learning based criterion in the loss function and train the deep network to optimize the objective. Our experimental results on a variety of uni-modal and multi-modal datasets from different application domains depict the promise and potential of the method for real-world image recognition applications.

5. REFERENCES

[1] N. Cebron and M. Berthold. Active learning in parallel universes. In *ACM International Conference on Information and Knowledge Management (CIKM)*, 2010.

[2] S. Chakraborty, V. Balasubramanian, and S. Panchanathan. Generalized batch mode active learning for face-based biometric recognition. In *Pattern Recognition Journal*, 2013.

[3] S. Chakraborty, V. Balasubramanian, Q. Sun, S. Panchanathan, and J. Ye. Active batch selection via convex relaxations with guaranteed solution bounds. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2015.

[4] G. Dahl, M. Ranzato, A. Mohamed, and G. Hinton. Phone recognition with the mean-covariance restricted boltzmann machine. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[5] R. El-Kaliouby and P. Robinson. Mind reading machines: Automated inference of cognitive mental states from video. In *IEEE International Conference on Systems, Man and Cybernetics*, 2004.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[7] Y. Guo. Active instance sampling via matrix partition. In *Advances of Neural Information Processing Systems (NIPS)*, 2010.

[8] G. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. In *Science*, 2006.

[9] S. Hoi, R. Jin, J. Zhu, and M. Lyu. Semi-supervised SVM batch mode active learning for image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[10] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.

[11] A. Krizhevsky. Learning multiple layers of features from tiny images. In *Technical Report*, 2009.

[12] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[13] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. In *Nature*, 2015.

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of IEEE*, 1998.

[15] Z. Liu, X. Li, P. Luo, C. Loy, and X. Tang. Semantic image segmentation via deep parsing network. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[16] I. Muslea, S. Minton, and C. Knoblock. Active learning with multiple views. In *Journal of Artificial Intelligence Research*, 2006.

[17] H. Ranganathan, S. Chakraborty, and S. Panchanathan. Multimodal emotion recognition using deep learning architectures. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016.

[18] M. Ranzato, J. Susskind, V. Mnih, and G. Hinton. On deep generative models with applications to recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[19] R. Salakhutdinov and H. Larochelle. Efficient learning of deep boltzmann machines. In *Artificial Intelligence and Statistics Conference (AISTATS)*, 2010.

[20] C. Sanderson. Biometric person recognition: Face, speech and fusion. In *VDM Verlag*, 2008.

[21] B. Settles. Active learning literature survey. In *Technical Report 1648, University of Wisconsin-Madison*, 2010.

[22] D. Shen, J. Zhang, J. Su, G. Zhou, and C. Tan. Multi-criteria-based active learning for named entity recognition. In *Association for Computational Linguistics (ACL)*, 2004.

[23] F. Stark, C. Hazirbas, R. Triebel, and D. Cremers. Captcha recognition with active deep learning. In *Workshop on New Challenges in Neural Computation*, 2015.

[24] D. Wang and Y. Shang. A new active labeling method for deep learning. In *International Joint Conference on Neural Networks (IJCNN)*, 2014.

[25] S. Zhou, Q. Chen, and X. Wang. Active deep networks for semi-supervised sentiment classification. In *International Conference on Computational Linguistics*, 2010.