DATASET OPTIMIZATION FOR CONVOLUTIONAL NEURAL NETWORKS BASED ON ACTIVE LEARNING

$Liguo\ Zhou^1, Zhongyuan\ Wang^1,\ Shu\ Wang^2, Yimin\ Luo^3$

1. School of Computer, National Engineering Research Center for Multimedia Software, Wuhan University, Wuhan, China
Collaborative Innovation Center of Geospatial Technology, Wuhan, China
2. Biomedical Engineering & Imaging Science, King's College London, London, UK
3. Remote Sensing Information Engineering School, Wuhan University, Wuhan, China

ABSTRACT

Convolutional neural networks (CNNs) have shown significant advantages in computer vision fields. For the optimizations of CNNs, most research works focus on feature extraction, which creates deeper structures and more diverse activations, but the optimization on dataset is rarely discussed. Due to the booming of data, most CNNs suffer from serious problems of dataset redundancy and resulting high computational burden. To this end, this paper brings in an informativeness ranking idea and proposes a new methodology for dataset refinement based on active learning (AL). Particularly, for classification problems with a large number of classes, this paper further proposes entropy ranking (ER), a new active learning method, to enhance the optimization ability. Extensive experiments on MNIST and CIFAR-10 prove its effectiveness in terms of the higher classification accuracy for CNNs at a less training cost.

Index Terms**—**dataset optimization, CNN, active learning, entropy ranking.

1. INTRODUCTION

Deep learning is a hot topic in computer vision [1]. Convolutional neural networks (CNNs), a typical kind of deep learning algorithms, have been widely used in visual classification tasks. Compared to traditional models of computational structures, like support vector machines (SVMs) [2], CNNs can automatically learn more abstract features [3]. Despite of their strong learning ability, CNNs are subjected to considerable expenditure of computing resource and difficulties of parameter adjustment caused by the booming of data. Therefore, the optimization on dataset is required, and this paper intends to achieve this optimization based on active learning (AL).

AL is a machine learning technique for dataset optimization, which achieves this optimization by selecting the most informative samples for training [4]. Therefore, sample selection strategies are the key to AL, and these strategies are all based on the paradigm of informativeness ranking from classification outcomes. Compared to random sampling (RS), AL significantly improves the efficiency of classification, whose informativeness criterion has been widely applied to many machine learning tasks [5,6,7]. However, most applications for AL are aimed at solving the problem of sample scarcity and high labeling cost. For dataset redundancy, actually, AL can also be applied as a mean of optimization. Specifically, to reduce training costs of CNNs, AL can work as a filter which establishes a new dataset with the most informative samples selected from the original dataset.

From a measuring standpoint of informativeness, there are two main kinds of AL algorithms: probability-based AL and distance-based AL. Breaking Ties (BT) [8] is a typical probability-based AL algorithm, which ranks samples' informativeness according to their two highest posterior class probabilities. As CNNs' output of the last fully-connected layer can be processed samples' class probabilities by Softmax [1], BT can be directly applied to CNNs as a tool of dataset optimization.

However, for classification tasks which focus on a large number of classes, dealing with two highest posterior class probabilities is of little representativeness. Considering that entropy is a classical measure of informativeness, alternatively, this paper proposes a new AL algorithm based on CNNs' output, named entropy ranking (ER). Compared to BT, ER takes care of all the class probabilities provided by CNNs. The maximization of entropy gives a naturally multiclass heuristic, and samples with high entropy are of high uncertainty. Accordingly, selecting samples with high uncertainty is helpful to increase the informativeness of dataset, which promotes CNNs' training.

This paper mainly makes the following two contributions. Firstly, CNNs are improved from a perspective of dataset optimization based on AL. Secondly, AL family is enriched by the proposed ER method. Experiments on MNIST [9] and CIFAR-10 [10] prove the effectiveness of this dataset optimization, as higher classification accuracy can be achieved with smaller dataset for training. Moreover, it is also proved that the proposed ER outperforms BT in classification tasks whose class number is huge.

2. RELATED WORKS

The key to deep learning is back propagation training, and CNN improves it by reducing the number of parameters which obtained from spatial relationships. Therefore, CNN can be regarded as a choice of topology or architecture [11]. A typical CNN consists of three types of layers: convolutional layer, pooling layer and fully-connected layer. In most CNNs, Softmax outputs samples' class probabilities for classification. Based on it, Softmax loss function, regarded as s a supervision component for deep neural network model's training [12], dedicates to continuously increase the discernibility of sample characteristics based on the output of the last fully-connected layer with labels. For CNNs' optimization, most research works focus on network architecture [13, 14], activation [15, 16] and loss function [12,17].

Traditional AL algorithms, developed on support vector machines (SVMs), focus on the interaction between the user and the classifier [7]. Based on the class probabilities provided by CNNs, BT method, a classical AL algorithm, achieves informativeness ranking though the minimum strategy. BT considers that samples whose two highest class probabilities are very close are of high uncertainty. Previously, class probabilities for BT method are calculated by fitting a sigmoid function to the SVM decision function [8]. In most tasks, especially in remote sensing image classification, AL returns the classification outcome for next selection of informative unlabeled samples. For BT's optimization, most research works focus on its cooperation with spatial information for selecting unlabeled sample [18]. Conversely, in this paper, AL is applied to reduce the redundancy of labeled samples with CNNs' classification results.

3. DATASET OPTIMIZATION

In previous works, as summarized in [19], the classical flowchart of dataset optimization based on AL consists of five components: unlabeled set, labeled set, sample selective strategies, classifier and supervisors. As depicted in Fig. 1, informative unlabeled samples should be selected out by sample selective strategies and labeled by supervisors iteratively until the labeled set is strong enough to train out a good classifier. AL significantly relieves the difficulties of sample scarcity, therefore, it has been widely used in remote sensing image classification tasks where the human annotation cost is extremely high [20,21].

Actually, dataset optimization is not just what sample scarcity calls for. In the case of sample redundancy, refining the redundant dataset is also needed. In this data explosion era, data redundancy is a common phenomenon. Despite of the great learning ability provided by CNNs, facing a tremendous dataset for training, complexity of network architecture and waste of computing resources should be also taken into consideration. Therefore, refinement is of great significance to both network architecture's simplification and performance. Here, AL provides a good idea for achieving this refinement. As depicted in Fig. 1, a small amount of samples is selected from the original dataset randomly for the establishment of initial CNN model, and then, AL's sample selective strategies can be applied to the original dataset's refinement based on the obtained CNN's outcomes. A new dataset consisting of highly-informative samples is established, and trained with it enables us to obtained a stronger classification model.

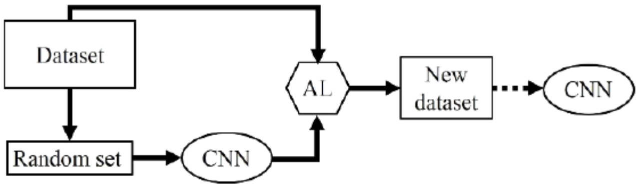

Fig. 1. AL for dataset optimization based on CNNs.

According to our flowchart in Fig. 1, the first CNN model’s role in the flowchart is providing samples’ class probability for AL selective strategies. Why can CNN output samples’ class probability? The answer is its Softmax. Most CNN architectures have convolutional layers, pooling layers and fully-connected layers. Softmax, which can provide class probabilities of samples in the last fully-connected layer of CNNs, and Softmax loss function works as a supervision component which controls back propagation training of CNNs. Softmax loss can be formulated as (1), where $p_{y_i}(x_i)$represents the probability of sample x_i belonging to its labeled class y_i , and N represents the number of samples for training.

$$\mathrm{L_S}= -\frac{1}{N}\sum_{i=1}^{N}\log[p_{y_i}(x_i)] \qquad (1)$$

This probability of which samples belong to each class is output by Softmax and is calculated by equation (2), where j represents class number, and W_k represents the k-th column of weight.

$$p_k(x_i) = \frac{e^{W_k^T x_i}}{\sum_{j=1}^{class} e^{W_j^T x_i}} \qquad k \in \{1,2,\dots,class\} \qquad (2)$$

In this paper, class probabilities obtained by equation (2) are the key to dataset refinement via AL based on CNNs.

4. ACTIVE LEARNING BASED OPTIMIZATION FOR CNNS

In this section, two sample selective strategies of probability-based AL (BT and ER) are presented and applied to CNNs for dataset optimization.

4. 1 BT

In terms of informativeness measuring, BT deals with the difference for the two most probable classes. Admittedly, samples whose two highest class probabilities are very close and are difficult to determine the classes they belong, and trained with these samples is thus beneficial to classification model's improvement. BT originates from this idea, and its heuristic can be formulated as (3), where $\mathrm{p}(x_i)$ represents the probability of sample x_i belongs to each class.

$$\hat{x} = argmin\, [p_{max}(x_i) - p_{submax}(x_i)] \quad (3)$$

Previously, the class probability used for BT is assessed by a sigmoid function (4) based on SVM decision function. In (4), according to [22], A and B are parameters to be estimated.

$$p(y_i = \omega | x_i) = \frac{1}{1+ e^{Af(x_i \omega)+B}} \quad (4)$$

Here, to apply BT to CNNs, samples' class probabilities can be obtained by (2).

4.2 ER

Despite of BT's effectiveness on sample selection, for classification problems with a large number of classes, two highest posterior class probabilities are unable to represent samples informativeness comprehensively. Therefore, taking all the class probabilities provided by the classifier into consideration is necessary. Entropy provides us with a good idea for solving this problem, whose maximization gives a naturally multiclass heuristic.

Previously, entropy has been applied to remote sensing image classification as an AL method, named entropy query-by-bagging (EQB) [21]. However, the probability calculation of EQB is complicated, which has to set up several classifiers for voting. Considering the cost of CNNs' training, bagging is unpractical and unnecessary as consequence of its computing cost. Although, it can hardly be applied to refine sample sets here, using entropy for informativeness ranking based on CNNs' outputs is of great significance. Accordingly, this new method is named ER.

ER evolves from EQB method but the samples' class probability is obtained by (2). With samples' class probability, the entropy of each sample can be calculated as equation (5).

$$\mathrm{H}\,(x_i) = \sum_{j=1}^{class} -p_j\,(x_i) log_{class} p_j(x_i) \quad (5)$$

According to the obtained entropy of all samples, those satisfying (6) can be selected out for the establishment of new training set.

$$\hat{x} = \arg\max\{H\,(x_i)\} \quad (6)$$

4.3. ALGORITHM

In terms of sample selection, selective strategies are an important part which should be taken into consideration, as it directly affects selected samples quality. On the other hand, the selected sample number is also of great significance to dataset refinement, in that it involves a proportional problem.

Intuitively, to achieve a higher classification accuracy with less samples, the more informative the dataset is, the greater the sample selective proportion should be chosen. Here, sample informativeness is a positive number which falls in the interval [0,1], therefore, we propose a calculation method (7) for sample redundancy.

$$\mathrm{R}\,(x_i) = 1 - \mathrm{H}\,(x_i) \quad (7)$$

For optimizations and application of AL, sample selective strategies has to account for selective proportion. Mostly, the number of sample selection is set empirically, hence, a systematic discussion on this proportion is needed. Here, we establish a simple relationship between selective proportion and dataset informativeness, which reaches a balance between redundancy and informativeness of dataset.

$$\hat{x} = max_{\{H(x_i)\}} min_{\{R(x_i)\}}\,\{x_i\} \quad (8)$$

Considering the simplification of this process, as illustrated in Algorithm 1, we provide a fast method for finding the optimal number of sample.

Algorithm 1 : Fast method for deciding selective number

```
1: Input: Sample entropy set {Ei};
   Output: Number of selective sample
2: {Ei'} ← descending sort {Ei}
   {Ri} ← {1-Ei}
   SE0 ← 0
   $SR_{N+1}$ ← 0
   for  i = 1 to N
        SEi ←     SEi 1 + Ei'
   end
   for  i = N to 1
        SRi ←     SRi 1 + Ri
   end
3: Find n which makes SEn ≈ Rn
4: Choose the top n from {Ei}.
```

5. EXPERIMENTS AND RESULTS

In this section, two well-known visual classification databases: MNIST [9] and CIFAR-10 [10] are used for validating the effectiveness of the proposed dataset optimization based on AL with CNNs.

5.1 DATASET

The MNIST is a handwritten digit database which has 60,000 training examples and 10,000 test examples in total. For its simple request for preprocessing and formatting, it has been widely used in experiments of pattern recognition and learning techniques. Moreover, CIFAR-10 dataset consists of 60000 32x32 color images in 10 classes, with 6000 images per class. There are 50000 training examples and 10000 test examples. We implement the CNN model using the Caffe [23] library with the proposed methodology, and details of the CNN architecture are given in Table 1 and 2. Here, 6,000 samples (600 for each class) are selected randomly from the original dataset for initial CNN models' establishment.

Table 1. The CNN Architecture of the Test on MNIST

Mnist					
Layer	Conv	Max Pool	Conv	Max Pool	FC
Num Output	20	-	50	-	500
Kernel Size	5×5	2×2	5×5	2×2	-
Stride	1	2	1	2	-
Pad	0	-	0	-	-

Table 2. The CNN Architecture of the Test on CIFAR-10

Cifar10								
Layer	Conv × 2	Max Pool	Conv × 2	Max Pool	Conv × 3	Max Pool	FC	FC
Num Output	64	-	128	-	256	-	2048	2048
Kernel Size	3× 3	2× 2	3×3	2× 2	3× 3	2× 2	-	-
Stride	1	2	1	2	1	2	-	-
Pad	1	-	1	-	1	-	-	-

5.2 RESULTS AND ANALYSIS

For fairness, we test proposed methods for 10 times for average. The classification results are compared in Fig. 2 and Fig. 3 for datasets MNIST and CIFAR-10, respectively.

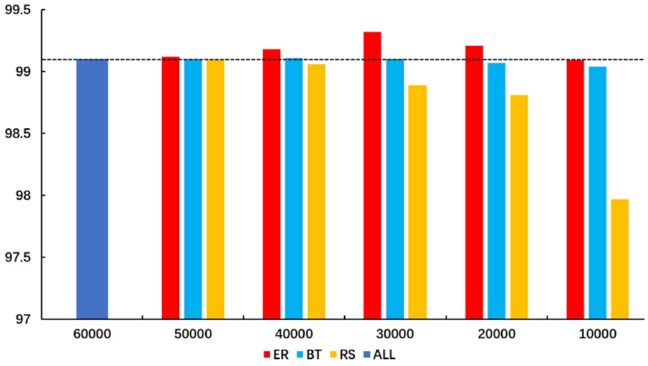

Fig. 2. Results on MNIST (using 30000 samples selected by ER achieves the highest accuracy).

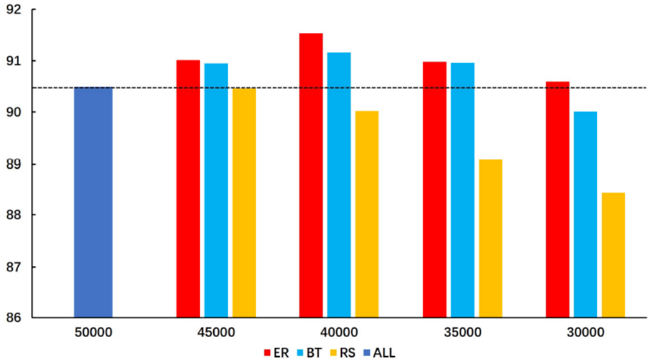

Fig. 3. Results on CIFAR-10 (using 40000 samples selected by ER achieves the highest accuracy).

Training the classification model with the whole 60000 samples, we suffer from high computing cost, and the classification accuracy is only 99.08%. Using BT and ER methods, the accuracy can climb to 99.21% and 99.29% respectively with only 30000 training samples (our fast method instructs us to select appropriate 30000 samples). In general, both of these two sample selecting strategies obviously outperform RS at same computing cost, which validates the effectiveness of this refinement. Similarly, training with the whole 50000 samples in CIFAR-10 dataset, we can only obtain a classification accuracy of 90.49%. However, with the help of BT and ER methods, the accuracy can climb to 91.16% and 91.53% respectively with 40000 training samples (our fast method instructs us to select appropriate 40000 samples). And the proposed ER method obviously outperforms BT method, not to mention RS. Overall, the proposed ER method shows advantage over BT method, as it takes all class probabilities into consideration.

6. CONCLUSION

As the consequence of data explosion, CNNs have to face an increasingly serious problem of dataset redundancy, which may cause high computational resources, difficulty of parameter tuning and sometimes decrease of classifier performance. This paper proposes a new paradigm of dataset refinement using AL method for CNNs. Moreover, we enrich AL family by proposing ER method, which selects samples with high uncertainty to increase the informativeness of dataset. Extensive experiments on public datasets prove the effectiveness of dataset optimization in terms of classification accuracy.

ACKNOWLEDGMENTS

The research was supported by National Natural Science Foundation of China (61671332, 61471274), Hubei Province Technological Innovation Major Project (2016AAA015), the National Key Research and Development Program of China (2016YFB0100901).

REFERENCES

[1] Sun Yi, Chen Yuheng, Wang Xiaogang and Tan Xiaoou, Deep learning face representation by joint identification verification, in Neural Information Processing Systems (NIPS), 2014, pp. 1988-1996.

[2] FuJie Huang and Yann Lecun, Large-scale Learning with SVM and convolutional for generic object categorization, in Computer Vision and Pattern Recognition (CVPR), 2006, pp. 284-291.

[3] Hongming Zhou, Guang-Bin Huang, Zhiping Lin, Han Wang and Yeng Chai Soh, Stacked extreme learning machines, IEEE Transactions on Cybernetics, 45(9):2013-2025, 2015.

[4] Lijun Zhang, Chun Chen, Jiajun Bu, Deng Cai, Xiaofei He and Thomas S. Huang, Active learning based on locally linear reconstruction, IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(10):2026-2038, 2011.

[5] Nicolás Garcia- Pedrajas, Javier Pérez Rodríguez and Aidade Haro-García, Oligo IS: Scalable Instance Selection for Class-Imbalanced Data Sets, IEEE Transactions on Cybernetics, 43(1):332-346, 2013.

[6] Peng Liu, Hui Zhang, and Kie B. Eom, Active deep learning for classification of hyperspectral images, IEEE Journal of Selected Topics in Applied Earth Observations and Remote,10(2):712-724, 2017.

[7] Bo Du, Zengmao Wang, Lefei Zhang, Liangpei Zhang, Wei Liu, Jialie Shen and Dacheng Tao, Exploring representativeness and informativeness for active learning, IEEE Transactions on Cybernetics, 41(7):14-26, 2017.

[8] Devis Tuia, Edoardo Pasolli and William J Emery, Using active learning to adapt remote sensing image classifiers, Remote Sensing of Environment, 115(9):2232-2242, 2017.

[9] LeCun Y., Cortes C. and Burges C.J., The mnist database of handwritten digits, 1998.

[10] Alex Krizhevsky, Learning multiple layers of features from tiny images, Technical report, U. Toronto, 2009.

[11] Itamar Arel, Derek Rose and Thomas P Karnowski, Deep Machine Learning—A New Frontier in Artificial Intelligence Research, IEEE Computational Intelligence Magazine, 5(4):13-18, 2010.

[12] Yandong Wen, Kaipeng Zhang, Zhifeng Li and Yu Qiao, A discriminative feature learning approach for deep face recognition, in European Conference on Computer Vision (ECCV), 2016, pp. 499-515.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun, Deep residual learning for image recognition, in Computer Vision and Pattern Recognition (CVPR), 2015, pp. 770-778.

[14] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, S Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke and Andrew Rabinovich, Going deeper with convolutions, in Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.

[15] Vinod Nair and Geoffrey E Hinton, Rectified linear units improve restricted Boltzmann machines, in International Conference on Machine Learning (ICML), 2010, pp. 807-814.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification, in International Conference on Computer Vision (ICCV), 2015, pp.1026-1034.

[17] Weiyang Liu, Yandong Wen, Zhiding Yu and Meng Yang, Large-Margin Softmax loss for convolutional neural networks, in International Conference on Machine Learning (ICML), 2016, pp.507-516.

[18] Edoardo Pasolli, Farid Melgani, Devis Tuia, Fabio Pacifici and William J Emery, Improving active learning methods using spatial information, in international geoscience and remote sensing symposium (IGARSS), 2011, pp.3923-3926.

[19] Mingkun Li and Ishwar K Sethi, Confidence-based active learning, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(8):1251-1261, 2006.

[20] Devis Tuia, Frederic Ratle, Fabio Pacifici, Mikhail Kanevski and William J Emery, Active learning methods for remote sensing image classification, IEEE Transactions on Geoscience and Remote Sensing, 48(6):2767-2767, 2006.

[21] Melba M Crawford, Devis Tuia and Hsiuhan Lexie Yang, Active learning: Any value for classification of remotely sensed data? Proceedings of the IEEE, 101(3):593–608, 2013.

[22] Devis Tuia, Edoardo Pasolli and William J Emeryy, Using active learning to adapt remote sensing image classifiers, IEEE Transactions on Geoscience and Remote Sensing, 115(9):2232-2242, 2011.

[23] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama and Trevor Darrell, Caffe: convolutional architecture for fast feature embedding, in Proceedings of the ACM International Conference on Multimedia, 2014, pp.675-678.