Dynamic Attention Loss for Small-Sample Image Classification

Jie Cao*, Yinping Qiu*, Dongliang Chang†, Xiaoxu Li* and Zhanyu Ma†
*Lanzhou University of Technology, Lanzhou, China
E-mail: xiaoxulilut@gmail.com
†Beijing University of Posts and Telecommunications, Beijing, China
E-mail: mazhanyu@bupt.edu.cn

***Abstract*— Convolutional Neural Networks (CNNs) have been successfully used in various image classification tasks and gradually become one of the most powerful machine learning approaches. To improve the capability of model generalization and performance on small-sample image classification, a new trend is to learn discriminative features via CNNs. The idea of this paper is to decrease the confusion between categories to extract discriminative features and enlarge inter-class variance, especially for classes which have indistinguishable features. In this paper, we propose a loss function termed as Dynamic Attention Loss (DAL), which introduces confusion rate-weighted soft label (target) as the controller of similarity measurement between categories, dynamically giving corresponding attention to samples especially for those classified wrongly during the training process. Experimental results demonstrate that compared with Cross-Entropy Loss and Focal Loss, the proposed DAL achieved a better performance on the LabelMe dataset and the Caltech101 dataset.**

I. INTRODUCTION

The tremendous progressing of deep learning brings Convolutional Neural Networks (CNNs) a widely using in computer vision fields [1][2], such as data mining [3], face recognition [4], image classification [5][6], video tracking [7], target detection [8][9][10], and so on. However, for image classification tasks, there still exist many challenges when dealing with small-sample datasets, the most important cause of this is the large variance intra-class and inter-class, and this leads to the categories' confusion problem, influence the accuracy of the classification results. Therefore, how to obtain the discriminative features between categories and decrease the variance of intra-class and inter-class, is a key to increase the accuracy of small-sample classification tasks.

Many methods such as data augmentation, adjusting optimizer, networks ensemble, are trying to decrease the confusion between categories, and loss function is an effective way to solve this problem.

Softmax cross-entropy loss is the most popular used loss function in CNNs, which combines softmax and Cross- Entropy (CE) loss. Many studies are proposed by modifying CE loss, such as focal loss [11], center loss [12], dual-cross entropy Loss [13], triplet loss [14], large-margin softmax loss [15], angular softmax loss [16], and large-margin regularized softmax cross-entropy loss [17], etc. Focal loss [11] reshapes the CE loss by using two regularization term to balance the imbalanced classes and down-weight easy classified examples to focusing on hard negatives. Center loss [12] learns a discriminative center from deep features for each class, and minimizes the quadratic sum of distance between the center and its intra-class samples to make them more compactable. Dual cross-entropy loss proposed a new loss which adds a regularization term to CE loss, and use the regularization term to constraint on the probability of a data point which is assigned to the class except its ground-truth[13].

Triplet loss [14], using a three tuple which is consists of an anchor x_a, its intra-class sample x_p and its inter-class sample x_n, targeting at decreasing the distance between similar categories and enlarging the distance between different categories, it improves the accuracy but with the price of huge time costing. Large-margin Loss (L-softmax) [15] uses an integer variable M multiplied by the margin angle value between samples, to make the training more difficult and the classification margin more larger, and make the learning of objective be harder. Angular softmax loss [16] adds limitative terms on L-softmax, which constraints weights and bias as a fixed value, making the prediction only depends on the angle between weights and feature vector. Large-margin regularized softmax cross-entropy loss [17] adds a quadratic regularization term to CE loss, enlarge the decision boundary of classes, and the regularization term makes loss function to be easily optimized. All of these loss functions mentioned above alleviate the confusion problem of small-sample classification tasks from different aspects.

In this paper, we find that the CE loss only places attention to the samples which assigned to their true categories, but place no attention to the samples misclassified, and this leads to that CE loss is deficient on discriminating samples, intuitively, if we can give more attention to those samples easily misclassified with others, the final performance will improve. Following this idea, we propose the *Dynamic Attention Loss* (DAL), which adds a dynamically updated rate-weighted term as a soft label on CE loss, putting more attention to correctly classified samples and misclassified samples, and decreases the confusion between categories.

The contributions of this paper are summarized as follows: (1) we proposed Dynamic Attention Loss which improves CE loss by giving dynamic attention to misclassified samples. The proposed loss can alleviate that CE loss only focus on correctly classified samples, and (2) the proposed loss function promotes the network to extract discriminative features by

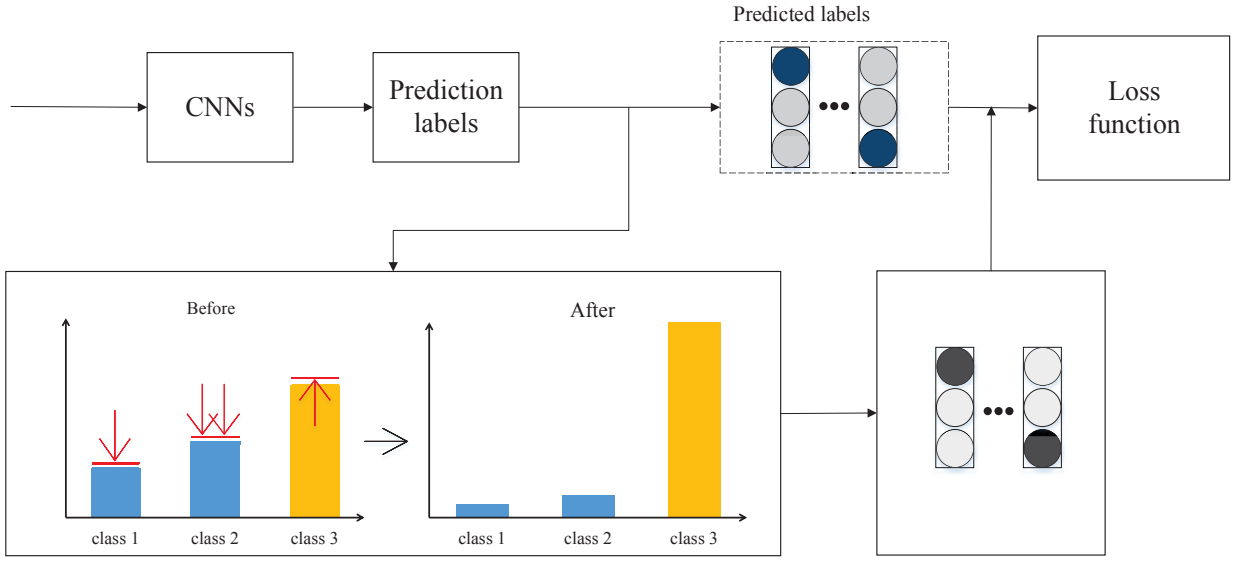

Fig. 1. An overview of the proposed *Dynamic Attention Loss* (DAL).

decreasing the confusion between similar classes and increasing the intra-class compacts and inter-class differences.

The experimental results on the LabelMe dataset and the Caltech101 dataset demonstrate that comparing with CE loss and focal loss, the proposed loss has a better generalization ability, effectively improves the classification performance.

II. Dynamic Attention Loss

Before introducing DAL loss, we first retrospect the CE loss, which is expressed as equation (1).

$$L_{CE} = -\frac{1}{N}\sum_{i}^{N} y_i^{T} \log(P_i), \tag{1}$$

Where N is the total number of samples in training data, one-hot vector y_i is the corresponding label of sample x_i, P_i is the probability of x_i.

Based on CE loss, we proposed Dynamic Attention Loss (DAL) which is showing in Figure 1. The proposed loss pays attention to the misclassified samples, and uses a soft label to control the confused categories. The formal definition of DAL followed as (2) and (3):

$$L_{DA} = -\frac{1}{N}\sum_{i}^{N} (Q_i + y_i^{T}) \log(P_i), \tag{2}$$

$$Q_i = -\frac{1}{a}(1 - y_i)^{T}, \tag{3}$$

where Q_i is the confusion rate-weighted we calculate every epoch during the training process to control the confusion rate of classes, decimal a ($a > 1.0$) is an inhibiting factor to decrease the probability of misclassified categories, N, y_i and P_i are same as in equation (1). About the selection value of $a > 1.0$, the reason is explained as follows: if we taking $y_i =$ $[0,0,1]$ and $P_{i3} < 0.5$, when $a = 1.0$, $Q_i = [-1,-1,0]$, $Q_i + y_i$ will equal as $[-1,-1,1]$, then $(P_{i1} + P_{i2}) > P_{i3}$, the computation of loss will be negative, for this reason we set $a >$ 1.0.

The proposed loss is a dynamic process in training, Algorithm1 describes the training process of DAL. After training started, we compute the confusion matrix between prediction labels L and corresponding probability P_n, use it with the regulatory factor a to update soft labels Q_i in every epoch, the labels L which computes loss are updated followed, then we use the updated L with P_n to calculate the loss.

Algorithm1 training proceduce of a network with DAL

Input : Trainset $X = \{(x_i, y_i) \mid i \in \{1,2,...,I\}\}$,

labels L_da, epoch N, parameter a.

Steps :

Initialize $n \leftarrow 0$, $P_n \leftarrow 0$, $a \leftarrow 0$.

repeat

$n = n + 1$

train network and got prediction P_n.

compute confusion matrix between P_n

and L_da, got Q_n.

update $Q_n = -Q_n / a$, with diagonal $Q_{ii} = 1$

update $L\_da = Q_n$.

$loss = -\sum L * \log(P_n)$.

Until $n = N$

TABLE I
Comparation of classification performances of different methods on the LabelMe dataset and Caltech101 dataset.

Dataset		CE	FL	DAL
LabelMe	Mean.	0.8813	0.8597	**0.8893**
	Std.	0.704	0.989	0.595
Caltech101	Mean.	0.5935	0.6152	**0.6224**
	Std.	1.873	5.645	1.722

III. EXPERIMENTAL RESULTS

In this section, we select two small-sample datasets, the LabelMe dataset and the Caltech101 dataset, compare DAL with Focal Loss and Cross-Entropy loss, to evaluate the performance. The data and its preprocessing are introduced in the first part, the second part introduces the network structures, parameters setting and other experimental details.

A. Data and data Preprocessing

LabelMe dataset (LM): A subset from [18]. This dataset contains 8 classes natural scene images respectively are coast, forest, highway, inside city, mountain, open country, street, and tall building. The dataset contains 1600 images with the size of 256×256, and there are 200 images for each class, we average each class into training dataset and testing dataset for experiments.

Caltech101 dataset [19]: A dataset contains objects from 101 categories, most of these pictures are about 300×200 pixels size, the amount of each category is different about 40 to 80, and the total number images of the dataset is 9146, we randomly divided it into equal training dataset and testing dataset.

We first use Python Imaging Library to resize the images into 256×256, and then use VGG16 network pre-trained on the ImageNet dataset to extra the features, the outcome of the pre-trained network is the size of 512×8×8=32768 pixels.

B. Implementation Datils of The Mentioned Methods

To evaluate the performance of the proposed loss DAL, we select two loss functions for comparison: Cross-Entropy (CE) and Focal Loss (FL). CE is the Cross-Entropy loss which we simply use as the loss function after two fully connected networks. Focal loss is implemented with the two fully connected networks, and we set the parameter gamma respectively as 5 and 15 for the LabelMe dataset and the Caltech101 dataset with no alpha. For all experiments, we use Pytorch[2] as the framework and choose python[2] to be the language, and the parameters including initial learning rate, L_2-norm, epoch, etc., are set to be the same in each experiment.

A two-layer fully connected network is used to be the base internet of our implementation, which is consists of input layers, hidden layers, and output layers. The number of hidden neurons is 32. The first layer uses Rectified Linear Unit function (ReLU) as activation function, and Stochastic Gradient Descent (SGD) is the optimizer which has the initial learning rate as 0.001, the momentum is 0.9, the epoch number is 300.

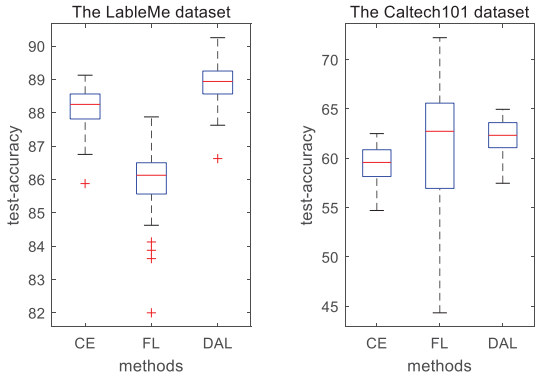

Fig. 2. Boxplot of the testing accuracy on the LabelMe dataset and the Caltech101 dataset. The red central marks denote median, the top and bottom edges of blue box denote the 25^{th} and 75^{th} percentiles, the blacklines outside the box denote the maximum and minimum, the red cross dots denote outliers.

C. Classification Performances

We run CE, FL, and DAL with the fully connected network mentioned above on the LabelMe dataset and the Caltech101 dataset 60 times for each method, and the mean value, standard variance of classification results are shown in Table I, in particular, Table I shows that DAL obtains the mean accuracy of 88.31% on the LabelMe dataset and 62.24% on the Caltech101 dataset, the corresponding standard deviation are 0.60 and 1.77. The accuracy of DAL performs better than FL and CE loss, it shows that the proposed loss has a better performance on classification accuracy and network stability.

To show more details of the experiments, in addition, we use the boxplot of all methods' performances on two datasets. From figure 2, we can see that the boxplot of DAL is the most compact compared with CE, FL, and DAL also obtains the best medians both on the LabelMe dataset and the Caltech101 dataset, that indicates DAL has the most stability and accuracy among 3 methods.

D. Accuracy for different a Values

The parameter a is a key affection factor to accuracy results of the proposed loss. Different values of a leads to different results of experiment. To demonstrate this influence, we choose some representative values a to implemented with the proposed loss. Since the results is very instability and unsatisfactory when $a < 3.0$, we set $a > 3.0$. Table II shows the experimental results of DAL with different value of a on the LM dataset. The test accuracy increases with the increasing of a and reaches the peak at $a = 15.0$. then the accuracy declines with the increasing of a. The maximum and minimum of accuracy almost unaffected by a, and this leads to that the standard deviation of accuracy keeps in a relatively stable interval.

Figure 3 is the boxplot of Table II, for convenience, we only use integer as the values of a. We can see that the test accuracy got the highest value at $a = 15.0$ on the LabelMe dataset. We believe there exists other decimals a to have better behaviors on DAL waiting studiers to discover.

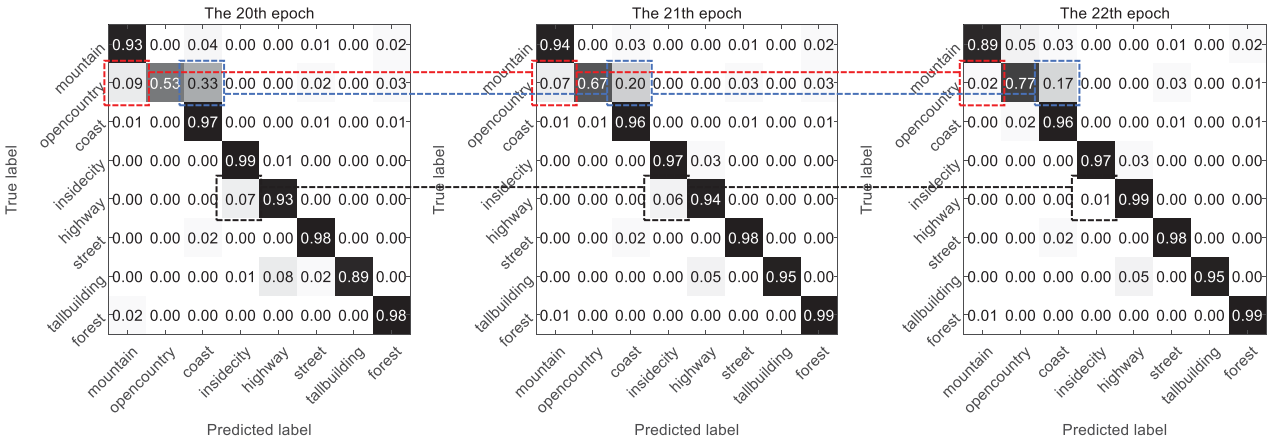

Fig. 4. The confusion matrix of DAL on the LabelMe dataset when epoch equals 20,21,22. The X-axis represents the real class and Y-axis the predicted class, the diagonal of confusion matrix shows the highest value in both rows or columns.

TABLE II
The accuracy of DAL with different a on the test data of the LabelMe dataset, the Mean., Max., Min. respectively represents the mean value, maximum value, and minimum value of test accuracy. When $a = 15.0$, then mean value of test accuracy is the highest.

a	Mean.	Max.	Min.
5.0	88.8417	90.2500	86.625
8.0	88.8563	90.2500	86.625
10.0	88.8875	90.3750	86.625
11.0	88.8604	90.2500	86.625
13.0	88.8666	90.2500	86.625
15.0	**88.8921**	90.2500	86.625
16.0	88.8875	90.3750	86.625
20.0	88.8604	90.2500	86.625

E. The Influence of DAL for Confused Categories

Since DAL uses the soft label dynamically decreases the confusion between classes, we randomly choose three epochs 20,21,22 from the total number of 300 epochs in training, and compute confusion matrixes of the three consecutive epochs to show this ability.

As we can see from Fig 4, the class *open country* is most easily confused with other classes. The *open country* has a misclassified value of 0.09 on *mountain* in 20th epoch, the value turns into 0.07 in 21th epoch and declines to 0.02 in 22th epoch (the red squares), and the misclassification value on *coast* is 0.33 in 20th epoch, and it turns down to 0. 20 and 0.17 in the next 21th and 22th epoch (the blue squares). The class *inside city* has a misclassification value of 0.07 on *highway*, in 20th epoch, it turns to 0.06 and 0.01 in 21th and 22th epoch (the black squares). These changes on misclassification values manifest that DAL is able to decrease the confusion between categories and make network to extract discriminative features to minimize the intra-class variance.

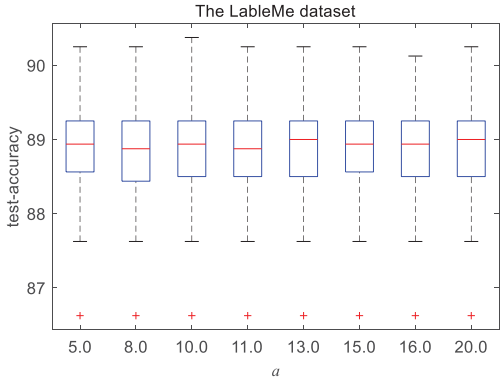

Fig. 3. Boxplot of the performance of DAL with different a implemented on the LabelMe dataset

IV. CONCLUSION

In this paper, we proposed a new loss function named as Dynamic Attention Loss (DAL), the loss uses a dynamic soft label to give attention to misclassified data, decrease the confusion between classes, and makes the network to obtain the ability of discriminating features correctly. The results of experiments show the superiority of the proposed loss.

The main reason why the proposed loss is effective is that the soft label paying dynamic attention to the confused categories. The large dataset like ImageNet or Cifar-10, also have the problem of categories are easily confused, and we can implement the dynamic process of DAL on large dataset to solve the problem. Besides, the way we use to give weights on samples to control the confusion of similar categories provides a new idea for exploring loss function from another aspect, which brings enlightenment to the future improvements of loss function.

ACKNOWLEDGEMENT

This work was partly supported by the National Key Research and Development Program of China under Grant 2018YFC0807205, the National Natural Science Foundation of

China (NSFC) grant No.61563030, No.61763028, No.61773071, No.61922015 and No.61906080, the Natural Science Foundation of Gansu Province, China, grant No.17JR5RA125, the Beijing Nova Program Interdisciplinary Cooperation Project No. Z181100006218137, Beijing Nova Program No. Z171100001117049 and by the Hong-liu Outstanding Youth Talents Foundation of Lanzhou University of Technology.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," pp. 1106–1114, 2012.
[2] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, and B. Shuai, "Recent Advances in Convolutional Neural Networks," pp. 1–38, 2006.
[3] N. V. Chawla, "Data Mining for Imbalanced Datasets: An Overview," Data Min. Knowl. Discov. Handb., pp. 875–886, 2009.
[4] P. Xu, Y. Huang, T. Yuan, K. Pang, and Y. S. Tao, "SketchMate : Deep Hashing for Million-Scale Human Sketch Retrieval," no. i, pp. 8090–8098.
[5] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," pp. 1–23, 2017.
[6] A. Fernández, S. del Río, N. V. Chawla, and F. Herrera, "An insight into imbalanced Big Data classification: outcomes and challenges," Complex Intell. Syst., vol. 3, no. 2, pp. 105–120, 2017.
[7] L. Wang, T. Liu, S. Member, G. Wang, K. L. Chan, and Q. Yang, "Video Tracking Using Learned Hierarchical Features," IEEE Trans. Image Process., vol. 24, no. 4, pp. 1424–1435, 2015.
[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, 2017.
[9] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," Proc. IEEE Int. Conf. Comput. Vis., vol. 2017-Octob, pp. 2999–3007, 2017.
[10] R. Girshick, J. Donahue, T. Darrell, J. Malik, and U. C. Berkeley, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2012.
[11] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," Proc. IEEE Int. Conf. Comput. Vis., vol. 2017-Octob, pp. 2999–3007, 2017.
[12] Y. Wen, K. Zhang, Z. L. B, and Y. Qiao, "A Discriminative Feature Learning Approach," Eccv, vol. 1, pp. 499–515, 2016.
[13] X. Li, L. Yu, D. Chang, Z. Ma, and J. Cao, "Dual Cross-Entropy Loss for Small-Sample Fine-Grained Vehicle Classification," IEEE Trans. Veh. Technol., vol. PP, no. XX, p. 1, 2019.
[14] A. Hermans, L. Beyer, and B. Leibe, "In Defense of the Triplet Loss for Person Re-Identification," 2017.
[15] W. Liu, "Large-Margin Softmax Loss for Convolutional Neural Networks," 2015.
[16] W. Liu and Y. Wen, "SphereFace : Deep Hypersphere Embedding for Face Recognition," pp. 212–220.
[17] C. Loss, "Large-margin Regularized Softmax," pp. 1–5, 2017.
[18] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and web-based tool for image annotation," Int. J. Comput. Vis., vol. 77, no. 1–3, pp. 157–173, 2008.
[19] T. Koda, "An Introduction to the Geometry of Homogeneous Spaces, Takashi Koda.pdf," Cviu, vol. 13, pp. 121–144, 2009.