Imbalanced Thangka Image Classification research Based on the ResNet Network

Fuliang Zeng¹,², Wenjin Hu¹,²,*, Guoyuan He²,³, Chaoyang Yue²,³

¹School of Mathematics and Computer Science, Northwest Minzu University, Lanzhou 730000, China
²Key Laboratory of China’s Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University Lanzhou 730000, China
³National Languages Information Technology, Boudler, Northwest MinZu University, Lanzhou 730000, China
*Corresponding author’s e-mail: wenjin_zhm@126.com

Abstract. Aiming at the problem of performance degradation caused by ignoring the softmax score of the wrong class in the training process of the unbalanced data set of Thangka images and the problem of the loss of negative feature information in the propagation process of the ReLU activation function, a new loss calculation method is proposed. Firstly, the parameters of the pre-training model on the COCO data set are used as the initial parameters. Secondly, CE and CCE are used to calculate the loss during the calculation of loss in the back propagation. Finally, AReLU activation function is used and a weight assigned to CE and CCE is added as final loss to update the parameters. The experimental results show that this algorithm improves the convergence speed and accuracy of the model with respect to imbalanced data. Compared with other loss functions, ours method performance is state-of-the-art, such as complement cross entropy.

1. Introduction

Thangka is a unique painting art form in Tibetan culture. With distinctive national characteristics, strong religious color and unique artistic style, it depicts the holy Buddha's world with bright colors. The pigment is traditionally all gold, silver, pearl, agate, coral, pine, malachite, cinnabar and other precious mineral gems and plants such as saffron, rhubarb, indigo as the pigment to show its holiness. These natural materials guarantee the bright and brilliant color of Thangka, which is still bright and dazzling even after hundreds of years. Therefore, it is regarded as the treasure of Chinese national painting art and the "encyclopedia" of Tibetan nationality, as well as the precious intangible cultural heritage of Chinese folk art. According to hand-painted Thangka, they are divided into colored Tang (white background, painted in various colors), Golden Tang (golden background), Silver Tang (silver background), Red Tang (red background), and black Tang (black background). Based on this significance, the classification of Thangka images is of great value.

With the rapid development of artificial intelligence and the emergence of deep learning, researchers have applied deep learning to various fields such as NLP and CV. Domain of computer vision mainly include image classification[1,2,3], object detection[4,5,6,7] and image segmentation[8,9,10] tasks. In recent years, researchers have proposed a variety of CNN models for computer vision tasks, among which the most representative ones are GoogLeNet, VGG[11], ResNet[12]
and DenseNet[13]. The research shows that the deeper the network layer is, the better the performance will be. However, with the increase of the network layer, the problem of gradient vanishing becomes more and more obvious. To solve this problem, He Kaiming et al. proposed ResNet network. The DenseNet network proposed by Huang Gao et al. alleviates the gradient loss problem and increases the feature reuse. Image classification is one of the basic tasks in computer vision. In the classification task, researchers widely use cross entropy as the loss function of the network, but the existence of unbalanced data sets limits the generalization of these models. Based on this, we design a new loss calculation method, which combines the cross entropy and complement cross entropy[14], to classify Thangka data set with imbalanced class distribution. A large number of experiments show that our method has better classification effect.

2. Classification Method

2.1. ResNet network

ResNet, proposed by He Kaiming in 2015, is the most widely used CNN feature extraction network at present. It applies the concept of residual representation commonly used in computer vision to the construction of CNN model, so there is the basic block of Residual learning. Unlike the normal CNN network, which uses parameter layers to directly learn the mapping between input and output, ResNet uses multiple parameter layers to learn the residual representation between input and output. Research shows that this method converges faster and effectively improves the classification accuracy. The author adds an identity map to the residual block structure, as shown in Fig.1.

As can be seen from Figure 1, ResNet has an identity map for every two layers of convolution, and the output is:

\[ H(x) = F(x) + x \]

(1)

However, when the number of \( F(x) \) and \( x \) channels is different, the identity mapping cannot be connected to the next convolutional layer. The author uses two methods to solve this problem. The first is to use zero-padding to increase the number of channels. The second is to use 1\( \times \)1\( \times \)n convolution of one to increase the number of channels, where \( n \) is the number of channels for the next residual block. This algorithm doesn’t add additional parameters and computational complexity, and solves the problem of deep network degradation. When the back propagation of the network is carried out, the obtained gradient is:

\[ \frac{\partial L(X)}{\partial x} = \frac{\partial L(x)}{\partial H(x)} \times \frac{\partial H(x)}{\partial x} \]

(2)
where $L(x)$ is the object function. It can be seen from Eq.(3) that the gradient will not disappear under the condition of 1, which makes the optimization of deep network more simple.

### 2.2. Improved Classification Network

#### 2.2.1. Cross entropy loss

Cross entropy loss is one of the most commonly used loss functions in deep learning classification tasks, which is derived from the definition of KL divergence. Assuming that $P$ is the true probability distribution and $Q$ is the predicted probability distribution, the cross entropy formula is denoted as:

\[
H(P,Q) = -\sum_{i=1}^{n} P(x_i) \times \log(Q(x_i))
\]

It can be seen from Eq.(4) that cross entropy is actually a measure of the similarity degree of two probability distributions. In the classification task, the lower the cross entropy is, the closer the predicted value is to the real label, and the better the model performance is. However, in the case of one-hot label, only one probability value is used in the calculation of the cross entropy formula, and this probability value will gradually increase with the optimization of the network, while other probability values will decrease correspondingly. On the whole, the goal of optimizing the entire probability distribution is realized. In multi-classification tasks, cross entropy can be defined as:

\[
H(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} y_{ij} \log \hat{y}_{ij}
\]

where $N$ is the number of batch samples, $K$ is the number of categories, $g$ behalf of the real label index, and $\hat{y}_{ij}$ is model to predict the probability of the sample class. Although the cross entropy loss function is most widely used in deep learning classification tasks, it can be seen from Eq.(5) that it has certain limitations: softmax function calculates the probability of incorrect class as zero because $\hat{y}_{ij}^g$ is always zero. This means that the cross entropy ignores $\hat{y}_{ij}^g$ in the calculation of the loss value, so the probability predicted by the model may generate cumulative errors. In order to avoid such errors, Complement Objective Training (COT) [15] was proposed by Chen et al., where the core idea is evenly suppressing softmax probabilities on incorrect classes during training.

#### 2.2.2. Complement cross entropy

To address the problem of imbalanced datasets, Yechean Kim et al. inspired by complement objective training proposed a new loss named complement cross entropy (CCE). CCE Loss is the combination of cross entropy and complement entropy. The formula of complement is defined as follows:

\[
C(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1,j \neq g}^{K} \frac{y_{ij} \log \hat{y}_{ij}}{1 - \hat{y}_{ij}^g} - \frac{1}{1 - \hat{y}_{ij}^g}
\]

where $\frac{1}{1 - \hat{y}_{ij}^g}$ is the regularization factor. First, the author defines the definition of equilibrium supplementary entropy $C'(y, \hat{y})$ as follows:
\[
C'(y, \hat{y}) = \frac{1}{K-1} C(y, \hat{y})
\]

where \( \frac{1}{K-1} \) is the equilibrium factor. In order to balance cross entropy and complement entropy, they add to the original complement as Eq. (8):

\[
\tilde{C}(y, \hat{y}) = \frac{\gamma}{K} - 1
\]

When \( \gamma = -1 \), the CCE definition is given:

\[
D(y, \hat{y}) = H(y, \hat{y}) + \tilde{C}(y, \hat{y})
\]

This algorithm trains the minority classes better by neutralizing the highest softmax score of the wrong class. It not only prevents the possible overfitting of the samples on the main class, but also overcomes the performance degradation caused by the imbalanced data set in the training class. Although the softmax score of the main error class is suppressed during neutralization, the information obtained for the minority class is very little. Based on this problem, this article combines CE and CCE, and defines a super parameter \( \beta \) to neutralize CE and CCE. A large number of experiments show that the performance of the network on the imbalanced data set is the best when \( \beta = 0.5 \). We add CE and CCE as final loss to calculate the loss value, which is defined as:

\[
\beta D(y, \hat{y}) + H(y, \hat{y})
\]

The classification process is shown in Fig. 2, which inhibits the score of error classes in the data set and increases the probability of training samples of a few classes.
3. Experimental Analysis

3.1. Datasets
In this paper, three types of thangka images collected manually are used as datasets, which are Buddha Mother, King Kong and Sakyamuni, as shown in Figure 3. Since thangka images are rare and of different types, we have collected a total of 1,000 thangka images, and their distribution is uneven. To facilitate training, we normalized the data set to pixels and shuffled all the order.

![Samples](image)

3.2. Experimental Environment
All experiments were carried out under the Linux system using Pytorch deep learning framework, and 100 epochs were trained with the environmental parameters as follows: GPU [GeForce RTX 2080Ti 11Gb, Intel(R) Xeon(TM) E5-2620 CPU@2.10GHz, 32 Gb Memory.

3.3. Evaluation Indicators
In order to evaluate the algorithm performance in this paper, the following indexes are adopted as measurement standards:

\[
\text{accuracy} = \frac{TR + TN}{TP + TN + FP + FN},
\]

where TP, TN, FP and FN respectively represent true positive, true negative, false positive and false negative, that is, the accuracy rate represents the percentage of the number of correctly classified samples in the total samples. Eq.(11) is used as the performance index of dichotomy, while the laboratory in this paper belongs to multi-classification, so we take the average accuracy of each class as the performance index of this paper, as shown in Eq.(12).

\[
\sum_{i=1}^{K} \text{accuracy}^{(i)} \quad (12)
\]

where \(K\) is the number of sample types, and \(\text{accuracy}^{(i)}\) represents the accuracy of the \(i^{th}\) .

3.4. Experimental Result
We comparison experiments were conducted with other loss functions, and then experiments were conducted after adding AReLU activation function to the algorithm, which showed the effectiveness of the proposed algorithm.

| Table 1 Comparison of accuracy (%) among different loss functions |
|-------------|---|---|---|
|             | CE | CCE | Ours   |
| ResNet-18   | 91.33 | 91.35 | **92.01** |
| ResNet-34   | 92.10 | 92.25 | **92.55** |
| ResNet-50   | 92.01 | 93.36 | **93.50** |
| VGG-16      | 89.60 | 90.11 | **91.25** |
| Inception-V3 | 90.12 | 90.21 | **91.87** |
The accuracy comparison of the algorithm in this paper with that of CE and CCE is shown in Table 1, where the algorithm in bold is the algorithm in this paper. It can be seen from Table 1 that the classification accuracy of Thangka images of CE and CCE on the data set in this paper is lower than that of the algorithm in this paper, indicating the robustness and generalization of the algorithm in this paper. It can be seen that we compare five network models for feature extraction, and the algorithm in this paper has the best performance on each network. In particular, it reached 93.50% on ResNet-50.

When we train, it seems to converge more slowly when we use the pre-training model. Therefore, this paper attempts to change the ReLU activation function into AReLU activation function for all the networks tested in this paper, and the effect is shown in Fig.4.

![Figure 4](image_url)

**Comparison of convergence rates**

It can be clearly seen from Fig.4 that when there is no AReLU activation function training, the training accuracy and verification accuracy converge slowly, and the verification accuracy fluctuates greatly. When we use the algorithm in this paper and the AReLU activation function, we can find that the training and verification convergence is faster. After 25 epochs, the network converged, and even the fluctuation was smaller and smoother during verification. And you can see that the accuracy has improved. Here is a comparison of accuracy experiments.

|   | CE+ARelu | CCE+ARelu | Ours +ARelu |
|---|----------|-----------|-------------|
| ResNet-18 | 91.73    | 90.65     | **92.82**   |
| ResNet-34 | 92.14    | 92.69     | **93.09**   |
| ResNet-50 | 92.28    | 93.22     | **94.57**   |
| VGG-16    | 89.50    | 91.63     | **92.15**   |
| Inception-V3 | 90.85 | 90.96     | **91.89**   |

Comparison of experimental results of the AReLU activation function is added, as shown in Table 2. It can be seen from Table 2 that after the addition of AReLU activation function, our algorithm is still more accurate than other algorithms, and the experimental results are better than all experiments without AReLU, which indicates that AReLU activation function effectively solves the problem that ReLU activation function loses negative image features in back propagation.

|   | Ours   | Ours+ARelu |
|---|--------|------------|
| ResNet-18 | 92.01  | **92.82**  |
| ResNet-34 | 92.55  | 93.09      |
| ResNet-50 | 93.50  | **94.57**  |
| VGG-16    | 91.25  | 92.15      |
| Inception-V3 | 91.87  | 91.89      |
4. Conclusion
In this paper, in order to solve the unbalanced class while training data set limit model generalization to degrade the performance of the model of the problem, the existing cross entropy loss function and complement cross entropy loss function weighted additive and use the attention - based rectified linear unit activation function to improve the convergence speed and accuracy in the process of the trained network. In addition, it overcomes the problem that ReLU activation function loses negative characteristics in feature propagation. Experimental results show that this algorithm is superior to other algorithms in solving the classification problem of unbalanced classes. When resnet-50 network is used as the trunk network, its performance improves by 2.56% compared with that of COMPLEMENT cross entropy loss function. In the future, we will focus on the classification and detection of unbalanced data sets and the detection of small sample data.

Acknowledgments
This work was supported in part by The Nature Science Foundation of China under Grant No. 62061042, 61561042, 61862057, Fundamental Research Funds for the Central Universities under Grant No.31920200066, program for innovative research for innovative research team of SEAC (No.2018[98]) and Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education and by the special fund for talent introduction of northwestern nationalities university.

References
[1] Li Q, Cai W, Wang X, et al. Medical image classification with convolutional neural network[C]//2014 13th International Conference on Control Automation Robotics & Vision (ICARCV). IEEE, 2014: 844-848.
[2] Paoletti M E, Haut J M, Plaza J, et al. A new deep convolutional neural network for fast hyperspectral image classification[J]. ISPRS journal of photogrammetry and remote sensing, 2018, 145: 120-147.
[3] Xiao T, Xu Y, Yang K, et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 842-850.
[4] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.
[5] Szegedy C, Toshev A, Erhan D. Deep neural networks for object detection[C]//Advances in neural information processing systems. 2013: 2553-2561.
[6] Alghamdi M. Algorithm for Object Detection and Obstacle Avoidance for Medication Delivery Robot[D]. The Catholic University of America, 2020.
[7] Tian Z, Shen C, Chen H, et al. Fcos: Fully convolutional one-stage object detection[C]//Proceedings of the IEEE international conference on computer vision. 2019: 9627-9636.
[8] Ghosh S, Das N, Das I, et al. Understanding deep learning techniques for image segmentation[J]. ACM Computing Surveys (CSUR), 2019, 52(4): 1-35.
[9] Hesamian M H, Jia W, He X, et al. Deep learning techniques for medical image segmentation: Achievements and challenges[J]. Journal of digital imaging, 2019, 32(4): 582-596.
[10] Zhao A, Balakrishnan G, Durand F, et al. Data augmentation using learned transformations for one-shot medical image segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2019: 8543-8553.
[11] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.
[12] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
[13] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

[14] Kim Y, Lee Y, Jeon M. Imbalanced Image Classification with Complement Cross Entropy[J]. arXiv preprint arXiv:2009.02189, 2020.

[15] Chen H Y, Wang P H, Liu C H, et al. Complement objective training[J]. arXiv preprint arXiv:1903.01182, 2019.