Deep Learning Based on an Improved Minimum Classification Error Criterion Algorithm: A Case Study of Typhoon Satellite Cloud Imagery (Postprint)

Authors: Zheng Zongsheng, Hou Qian, Zou Guoliang, Lu Qi

Date: 2018-08-13T00:00:00+00:00

Abstract

To address the gradient reversal problem that occurs in the network when samples are misclassified in the objective function established based on the traditional Minimum Classification Error (MCE) criterion, the Minimum Classification Error criterion is introduced, and an FMCE objective function with a correction term is defined. Using cross-entropy with higher accuracy as the base function and FMCE as the correction function, an improved cross-entropy objective function CE-FMCE is proposed, which enhances the probability of label class output during network backpropagation. CE-FMCE not only overcomes the gradient reversal problem of the traditional MCE objective function, but also remedies the deficiency of the cross-entropy function in its indiscriminate handling of gradients for non-label sets. Comparative experiments were conducted on CE-FMCE against MSE, cross-entropy, MCE, and M3CE on both a self-constructed typhoon cloud image dataset and the universal MNIST dataset, and the experimental results demonstrate that CE-FMCE outperforms other objective functions.

Full Text

Preamble

Research on Deep Learning Based on Improved Minimum Classification Error Criterion Algorithm: A Case Study of Typhoon Satellite Cloud Images

Zheng Zongsheng, Hou Qian, Zou Guoliang†, Lu Qi
(College of Information, Shanghai Ocean University, Shanghai 201306, China)

Abstract: The traditional objective function based on the minimum classification error criterion (MCE) suffers from gradient inversion problems in the network when samples are misclassified. This paper introduces the minimum classification error criterion and defines an FMCE objective function with a correction term. Using the high-precision cross-entropy as the base function and FMCE as the correction function, we propose an improved cross-entropy objective function CE-FMCE that enhances the probability of label class output during backpropagation. CE-FMCE not only overcomes the gradient inversion problem of traditional MCE objective functions but also compensates for the deficiency of cross-entropy functions in indiscriminately handling gradients for non-label sets. Comparative experiments between CE-FMCE and MSE, cross-entropy, MCE, and M3CE were conducted on both a self-built typhoon cloud image dataset and the general MNIST dataset. Experimental results demonstrate that CE-FMCE outperforms other objective functions.

Keywords: deep learning; convolutional neural network; cross-entropy; minimum classification error criterion; typhoon rating

0 Introduction

Deep learning is currently a popular machine learning algorithm that addresses the weak generalization capability of shallow neural networks for complex classification problems by simulating the human brain' s hierarchical learning process to extract deep abstract features from natural information, thereby improving generalization ability [1]. Convolutional neural networks (CNNs), one of the successful models in deep learning applications, were proposed by LeCun [7] in 1989. However, CNNs only achieved significant progress in image recognition applications after error rates were reduced by 9% [8]. Current research on CNN weight optimization methods primarily focuses on selecting appropriate network parameters (such as variable-size convolution kernels [9], parameter pooling [10], Dropout zeroing rates [11], etc.) and using activation functions with better sparsity characteristics (e.g., ReLU, Leaky ReLU, PReLU, etc.). Among weight optimization methods, the objective function is crucial as it represents the current state of the network and provides parameter gradients in the gradient descent algorithm during training. If the gradient in the output layer is too small, after attenuation through deep layers, lower layers essentially receive no effective training signals. Constructing objective functions has become a research hotspot for achieving optimal network weights and improving generalization capability.

Commonly used objective functions in convolutional neural network algorithms include mean square error (MSE) and cross-entropy. MSE is more suitable for regression problems. Papoulis et al. [13] argued that MSE estimates posterior probabilities in classification problems, and when the network' s output function is sigmoid, the network suffers from gradient vanishing. With CNN development, research has shown that cross-entropy loss functions have fewer flat regions than MSE, making it easier for networks to escape local optima [14, 15].

Therefore, using cross-entropy as the objective function yields better results for multi-classification problems. Current research on objective functions mostly focuses on introducing relevant parameter terms based on cross-entropy for specific problems. For instance, Gui Zhe added an intra-individual difference loss function as a regularization term for face recognition, enabling the network to learn feature vectors belonging to the same person to be as similar as possible in space [12]. However, cross-entropy does not distinguish gradients for non-label dimensions during gradient descent, treating them with identical training. This prevents the classifier from effectively distinguishing the label class from the most easily confused class, thereby reducing model accuracy.

If a classifier can distinguish the correct class from the most easily confused category during training, the error rate will inevitably decrease. Juang et al. [16] first introduced the minimum classification error (MCE) method into shallow neural network training. Their proposed objective function considered the most confusing category in the non-label set to reduce classification error rates. However, the logistic objective function based on minimum classification error construction suffers from gradient vanishing. When deep CNNs adjust network weights through backpropagation, gradient saturation in the topmost objective function affects network training. Therefore, traditional convolutional neural networks typically use a softmax layer as the output function.

To overcome gradient saturation problems, Feng et al. [17] proposed establishing an objective function based on max-margin minimum classification error (M3CE). However, when samples are misclassified, M3CE exhibits partial non-label dimensions where the gradient direction is opposite to that of cross-entropy, leading to insufficient training information and reduced convergence speed during backpropagation, thus limiting its widespread application.

To overcome the gradient direction problems in traditional MCE objective functions, this paper introduces the minimum classification error criterion, defines an FMCE objective function with a correction term, and proposes a corrected cross-entropy objective function (CE-FMCE) based on the minimum classification error criterion. CE-FMCE not only overcomes the gradient direction problem of traditional MCE objective functions but also compensates for the deficiency of cross-entropy in indiscriminately handling gradients for non-label dimensions. This objective function is applied to both a self-built typhoon satellite cloud image dataset and the MNIST handwritten digit database. Comparative experiments under the CNN framework demonstrate the effectiveness of the proposed CE-FMCE objective function.

1 Traditional Convolutional Neural Networks

Convolutional neural networks consist of alternating convolutional and sampling layers. The network training process includes forward propagation and backpropagation. In forward propagation, convolutional layers perform implicit feature extraction on input images through convolution operations with layer

kernels, while sampling layers downsample feature maps to reduce image resolution, decrease computation, and improve network convergence speed. The network employs residual backpropagation rules [18] during backpropagation, transmitting from the output layer to input layers layer by layer, continuously reducing the model objective function and updating each neuron' s weights to obtain optimal network weights. These two processes cycle iteratively until the objective function reaches a specified threshold or maximum iterations are reached.

The residual represents the error signal between the network' s actual output and expected output. For a network with depth L, the output layer residual can be expressed as $\delta_L = \frac{\partial \ell}{\partial z_L}$, where ℓ is the loss function and z_L is the output layer' s pre-activation.

Let z_j be the input to the softmax function. The value of the j-th neuron in the output layer is expressed as:

$$p_j = \frac{\exp(z_j)}{\sum_{i=1}^{m} \exp(z_i)}$$

where m represents the number of sample classes.

For a training set containing N samples, the cross-entropy function is expressed as:

$$L = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} y_{nc} \log(p_{nc})$$

Since the training error is the sum of errors from all samples, for convenience of analysis, we consider only a single sample here. The output layer residual δ_L can be expressed as:

$$\delta_{L,j} = \begin{cases} p_j - 1, & j = k \\ p_j, & j \neq k \end{cases}$$

where k denotes the sample' s label dimension. According to this formula, when the dimension being solved is the sample' s label dimension ($j = k$), the cross-entropy function does not distinguish gradients for all non-label dimensions during backpropagation but trains them identically, reducing network convergence speed and training accuracy.

2 Improved MCE Objective Function Algorithm and Proof

2.1 Improved MCE Objective Function Algorithm

To address the problems with cross-entropy functions in traditional CNNs, we can optimize CNNs by establishing an objective function using the minimum error criterion based on cross-entropy. By defining a misclassification measure and applying different training methods to different non-label dimensions, we compensate for cross-entropy' s deficiencies.

The objective function based on minimum classification error is generally defined through three steps: a) For a certain class c, define a discriminant function $g_c(z)$ b) For each sample feature in class c, given a misclassification measure $d_c(z)$ c) Construct an objective function $\ell(d_c(z))$

Traditional methods only consider the influence of $d_c(z)$ when establishing $\ell(d_c(z))$. For example, using logistic objective functions leads to gradient vanishing as network depth increases. M3CE defines $\ell(d_c(z))$ such that when samples are misclassified, only the misclassified dimension has a positive gradient. Comparing with the cross-entropy formula reveals this gradient direction is opposite to cross-entropy, causing decreased convergence speed and insufficient training. To prevent erroneous signals from propagating to network 底层 and causing unpredictable errors, this paper improves MCE by proposing the FMCE objective function.

The definition process is as follows: a) For multi-classification problems, networks use softmax as the output layer. Therefore, this paper adopts the softmax function as the discriminant function $g_k(z) = p_k$.

b) If a classifier can separate the correct class from the most easily misclassified class, the network' s recognition rate will improve. Therefore, the misclassification measure $d_k(z)$ is defined as:

$$d_k(z) = -g_k(z) + \left(\frac{1}{C-1} \sum_{j \neq k} g_j(z)^{\xi} \right)^{1/\xi}$$

where k is the sample' s label class and r represents the most easily misjudged class for the sample' s label class in the softmax function output.

When $d_k(z) > 0$, the model produces misclassification. Traditional MCE objective functions increase the influence of erroneous information on the network, reducing model robustness. Therefore, when $d_k(z) > 0$, we add a correction term α. For a single sample, the expression of $\ell(d_k(z))$ is:

$$\ell(d_k(z)) = \begin{cases} \ln(1 + \exp(d_k(z))), & 0 < d_k(z) < 1 \\ \ln(1 + \exp(-d_k(z))), & -1 < d_k(z) \leq 0 \end{cases}$$

The relationship between the gradient direction and $d_k(z)$ can be summarized in .

Table 1. Objective Function Gradient Analysis

Since softmax output can represent posterior probability, $d_k(z) = -p_k + p_r$. When a sample changes from misclassification to correct classification, $d_k(z)$ changes from 1 to -1. Analyzing Table 1, equations (11)-(12), and (4), we find that when $d_k(z) > 0$ (sample misclassified), our method achieves differentiated processing of non-label dimension gradients while ensuring the gradient direction of $\ell(d_k(z))$ is consistent with cross-entropy. When $d_k(z) < 0$ (sample

correctly classified), the gradient direction also remains consistent with cross-entropy. Therefore, our method ensures improved output for the correct class while applying different training to non-label classes, proving theoretical feasibility.

In summary, $\ell(d_k(z))$ can serve as a supplement to L_{CE}, meaning FMCE can optimize CNNs as a cross-entropy supplement. Therefore, we introduce FMCE into cross-entropy to form the CE-FMCE objective function, with the final expression:

$$L_{CE-FMCE} = L_{CE} + \alpha L_{FMCE}$$

2.2 Proof of Improved MCE Objective Function

We provide theoretical derivation of the output layer residual for the objective function to prove FMCE' s feasibility. For analysis convenience, we consider only a single sample, simplifying equation (8) to:

$$L_{FMCE} = \begin{cases} \ln(1+\exp(d_k(z))), & 0 < d_k(z) < 1 \\ \ln(1+\exp(-d_k(z))), & -1 < d_k(z) \leq 0 \end{cases}$$

The output layer residual in gradient descent is $\delta_{FM} = \frac{\partial L_{FMCE}}{\partial z_j}$. Since $d_k(z)$ relates to p_k and p_r, we discuss them separately.

For $j = k$ (label dimension):

$$\frac{\partial L_{FMCE}}{\partial z_k} = \begin{cases} (1-p_k)\sigma(d_k(z)), & 0 < d_k(z) < 1 \\ -(1-p_k)\sigma(-d_k(z)), & -1 < d_k(z) \leq 0 \end{cases}$$

For $j = r$ (most confusing class):

$$\frac{\partial L_{FMCE}}{\partial z_r} = \begin{cases} -p_r\sigma(d_k(z)), & 0 < d_k(z) < 1 \\ p_r\sigma(-d_k(z)), & -1 < d_k(z) \leq 0 \end{cases}$$

For $j \neq k$ and $j \neq r$ (other non-label dimensions):

$$\frac{\partial L_{FMCE}}{\partial z_j} = \begin{cases} -p_j\sigma(d_k(z)), & 0 < d_k(z) < 1 \\ p_j\sigma(-d_k(z)), & -1 < d_k(z) \leq 0 \end{cases}$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$ is the sigmoid function.

Since $p_k, p_r \in [0,1]$ and $d_k(z) \in [-1,1]$, we have $0 < \sigma(d_k(z)) < 1$ and $0 < \sigma(-d_k(z)) \leq 1$. The analysis shows that when $0 < d_k(z) < 1$ (misclassification), our method achieves differentiated processing of non-label dimension gradients while ensuring gradient direction consistency with cross-entropy. When $-1 < d_k(z) \leq 0$ (correct classification), gradient direction also remains

consistent with cross-entropy. Therefore, the proposed method theoretically corrects the gradient reversal problem in M3CE [17] when misclassification occurs, while improving correct class output and applying different training to non-label classes, proving theoretical feasibility.

3 Experimental Results and Analysis

3.1.2 Results Comparison

All comparative experiments in this paper were conducted on a Windows 10 system with Intel Core i5-6500M 3.2 GHz CPU and 4 GB memory, using the TensorFlow-based Keras deep learning framework. Experiments were divided into two groups: the first group used our self-built typhoon satellite cloud image dataset to prove CE-FMCE' s feasibility for typhoon level classification, while the second group used the general MNIST handwritten digit database to verify CE-FMCE' s universality. Both groups employed the network model described in Section 2.3 and compared CE-FMCE with widely used objective functions, achieving expected results.

Model accuracy was evaluated through correctness rate. Assuming the true class label for the n-th sample among N samples is y^n_{true} and the predicted class label is y^n_{pred}, the model classification accuracy is:

$$ACC = \frac{1}{N}\sum_{n=1}^{N}\mathbb{I}(y^n_{true} = y^n_{pred})$$

In the first experiment, we compared our proposed objective function with MSE and cross-entropy. The network accuracy over 200 iterations is shown in [Figure 3: see original paper].

Figure 3. Training Results with Different Objective Functions

Comparison reveals that cross-entropy improves network optimization over MSE, achieving 96.25% accuracy after 200 iterations. CE-FMCE achieves better results than other objective functions, reaching 98.0% network accuracy with the fastest convergence speed, while MSE and cross-entropy show unsatisfactory convergence.

We compared CE-FMCE with widely used multi-classification objective functions on both training and test sets, with results shown in .

Table 3. Model Accuracy with Different Objective Functions

Objective Function	Training Set (%)	Test Set (%)
MSE	95.4	82.0
Cross-entropy	-	83.38
M3CE [17]	97.0	84.59

Objective Function	Training Set (%)	Test Set (%)
CE-FMCE	98.0	86.78

MSE achieved 95.4% and 82.0% accuracy on training and test sets respectively. Cross-entropy reached 83.38% test set accuracy. M3CE [17] improved both training and test set accuracy to 97.0% and 84.59% respectively, a 1.21% improvement over cross-entropy. Our CE-FMCE achieved 98.0% training set accuracy and 86.78% test set accuracy after 200 iterations, 1.75% higher than cross-entropy on the training set and 3.4% higher on the test set. CE-FMCE theoretically corrects M3CE' s gradient reversal problem during misclassification, which is well validated in Table 3, proving CE-FMCE' s feasibility with a 2.19% test set accuracy improvement over M3CE.

3.2 CE-FMCE Comparison Experiments on MNIST Dataset

To verify CE-FMCE' s universality, the second group of experiments employed the general MNIST handwritten digit database. The MNIST dataset contains 10 classes of handwritten digits (0-9) with 42,000 training samples and 10,000 test samples at 28$×$28 resolution, with partial samples shown in [Figure 4: see original paper].

Figure 4. Partial MNIST Dataset Samples

The second group compared CE-FMCE with other objective functions on MNIST using the same network model as the first group. MNIST samples were converted to 28$×28×$1 format as network input, with models trained for 100 iterations. Results are shown in .

Table 4. Model Accuracy with Different Objective Functions

Objective Function	Training Set (%)	Test Set (%)
M3CE [17]	99.87	99.06
CE-FMCE	99.92	99.11

CE-FMCE achieved 99.92% training set accuracy and 99.11% test set accuracy, a 0.05% improvement over M3CE on the test set and superior performance to other objective functions. However, the accuracy improvement is relatively small compared to other multi-classification objective functions, because MNIST digits have minimal noise interference and low recognition difficulty, resulting in high baseline accuracy.

Comparison between Tables 3 and 4 shows that CE-FMCE provides more significant accuracy improvements for images with complex features.

4 Conclusion

Deep learning achieves complex function approximation through deep nonlinear networks to express deep abstract features of natural images, overcoming the weak generalization capability of traditional shallow neural networks for complex classification problems. Current deep learning research primarily focuses on optimizing network parameters. As a crucial component of deep learning algorithms, the objective function not only characterizes the current network state but also provides parameter gradients in the network' s gradient descent algorithm. Therefore, this paper studies objective functions and proposes CE-FMCE, a corrected cross-entropy objective function based on MCE, to address gradient problems in traditional MCE-based objective functions. Comparative experiments on our self-built typhoon cloud image dataset and the general MNIST handwritten digit database demonstrate that CE-FMCE achieves expected results compared with widely used objective functions (MSE, cross-entropy, MCE, and M3CE), fully proving our method' s feasibility and providing a new research direction.

The deep learning algorithm based on the improved minimum classification error criterion differentiates gradients in the non-label set, resulting in longer batch processing times while not significantly improving recognition accuracy for simple feature samples. Therefore, future research will focus on parallelization to improve efficiency and further optimization.

References

[1] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521(7553): 436-444.
[2] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks [C]// Proc of European Conference on Computer Vision. Cham: Springer, 2014: 818-833.
[3] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
[4] Xiao Jianxiong, Hays J, Ehinger K A, et al. SUN database: large-scale scene recognition from abbey to zoo [C]// Proc of IEEE Computer Vision and Pattern Recognition. 2010: 3485-3492.
[5] Huang Kaiqi, Ren Weiqiang, Tan Tieniu, et al. A review on image object classification and detection [J]. Chinese Journal of Computers, 2014, 37(6): 1225-1240.
[6] He Xiping, Zhang Qionghua, Liu Bo. Deep learning model of target classification features based on HOG [J]. Computer Engineering, 2016, 42(12): 176-180.
[7] LeCun Y. Generalization and network design strategies [C]// Proc of Connectionism in Perspective. 1989.
[8] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]// Proc of International Conference on Neural Information Processing Systems. [S.l.]: Curran Associates Inc, 2012: 1097-1105.
[9] Xu Liyang, Fan Chunxiao. Multi-scale tracker based on improved kernel cor-

relation filter [J]. Journal of Computer Application, 2015.
[10] Jia Tao, Zhou Lili, Chen Jian, et al. Via and pad detection in PCB CT images based on convolutional neural network [J]. Application Research of Computer, 2018, 35(2): 637-640.
[11] Finn C, Hendricks L A, Darrell T. Learning compact convolutional neural networks with nested dropout [J]. Eprint Arxiv, 2015.
[12] Gui Zhe. Facial feature extraction and matching based on deep learning [D]. Chengdu: University of Electronic Science and Technology of China, 2016.
[13] Papoulis A, Saunders H. Probability, random variables and stochastic processes (2nd edition) [J]. A Papoulis, 1989, 33(6): 1637-1637.
[14] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]// Proc of International Conference on Neural Information Processing Systems. [S.l.]: Curran Associates Inc, 2012: 1097-1105.
[15] Szegedy C, Wei Liu, Jia Yangqing, et al. Going deeper with convolutions [C]// Proc of IEEE Computer Vision and Pattern Recognition. 2015: 1-9.
[16] Juang B H, Katagiri S. Discriminative learning for minimum error classification [pattern recognition] [J]. IEEE Trans on Signal Processing, 1992, 40(12): 3043-3054.
[17] Feng Ziyong, Sun Zenghui, Jin Lianwen. Learning deep neural network using max-margin minimum classification error [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing.
[18] Bouvrie J. Notes on convolutional neural networks [J]. Neural Nets, 2006.

Note: Figure translations are in progress. See original paper for figures.