LETTER

Gradient-Enhanced Softmax for Face Recognition

Linjun SUNy, Student Member, Weijun LIy, Xin NINGy, Liping ZHANGy, Xiaoli DONGy, and Wei HEy, Nonmembers

SUMMARY This letter proposes a gradient-enhanced softmax supervisor for face recognition (FR) based on a deep convolutional neural network (DCNN). The proposed supervisor conducts the constant-normalized cosine to obtain the score for each class using a combination of the intra-class score and the softmax of the inter-class scores as the objective function. This mitigates the vanishing gradient problem in the conventional softmax classifier. The experiments on the public Labeled Faces in the Wild (LFW) database denote that the proposed supervisor achieves better results when compared with those achieved using the current state-of-the-art softmax-based approaches for FR.

key words: convolutional neural network, face recognition, softmax classifier, vanishing gradient

1. Introduction

Recently, the face recognition (FR) performance in unconstrained conditions has significantly improved because of the emergence of deep convolutional neural network (DCNN). Figure 1 presents the common framework of the DCNN-based methods for FR. During the training phase, the supervisor is fed with features extracted from the training data through DCNN, which provides feedback to update the DCNN. After training, the trained DCNN becomes a feature extractor and evaluates the facial similarity based on the distance between features.

The supervisor plays a significant role in this framework. In the early development of DCNN-based methods for FR, DeepID1 employs the conventional softmax classifier to guide the feature extractor and evaluate the face similarity based on Joint Bayesians. FaceNet2 uses the metric learning to learn the mapping of face images to a compact hypersphere space, where Euclidean distances directly correspond to the face similarity. The current prevailing supervisors are based on softmax classifier, which can be improved using two methods. The addition of an external auxiliary loss to the conventional softmax classifier is the first method. DeepID2 uses contrastive loss to enhance the discriminative power of the features. A previous study adds center loss, and VGGFace employs triplet loss. The second method is to seek self-improvement. NormFace adds an L2-constraint with respect to the features and weights in the softmax classifier to allow the supervisor to perform metric learning. SphereFace and AM-Softmax introduce the margin to a conventional softmax classifier to boost the FR performance.

In the present study, we demonstrate that the conventional softmax classifier supervisory signals weaken drastically when the training samples are accurately classified. With the vanishing gradient problem, the supervisor failed to further decrease or increase the intra-class variation or the inter-class variation of face features, respectively. To solve the vanishing gradient problem, we propose a modified softmax-based supervisor called gradient-enhanced softmax (GESoftmax). In FR, the LFW evaluation shows that the proposed supervisor achieves better results than the current state-of-the-art softmax-based approaches.

2. Proposed Approach

In this section, we describe vanishing gradient problem in the conventional softmax classifier, which describes the GESoftmax in detail. We perform a simple experiment on Modified National Institute of Standards and Technology database (MNIST) to illustrate the effectiveness of the proposed supervisor.

2.1 Vanishing Gradient in Softmax Classifier

As shown in Fig. 2, the conventional softmax classifier con-
sists of two parts: score calculator and softmax loss. The blue dotted box indicates the score calculator revealing the sample’s score for each category. The higher the score, the more likely the sample to that category. The score is calculated as:

\[ s_i = w_i^T f \]  

(1)

where \( f \) is the feature vector extracted from the input face image through DCNN, and the corresponding score for the \( i \)-th class is \( s_i \). Moreover, \( w_i \) is also a vector of the same dimension as \( f \), which denotes the agent of the \( i \)-th category in softmax classifier. In Fig. 2, the orange dotted box indicates the softmax loss, evaluating the difference between the soft maximum scores for all categories and the score for the target category of the input face image. The formula is as follows:

\[ L_S = \log \left( \sum_{j=1}^{C} e^{s_j} \right) - s_t \]  

(2)

where \( C \) is the number of categories, and \( t \) indicates the target category of the input face image. To minimize \( L_S \), the target score \( s_t \) will be the highest among scores of all categories. The gradient of \( L_S \) with respect to scores is formulated as:

\[ \frac{\partial L_S}{\partial s_i} = \begin{cases} q_t - 1, & i = t \\ q_i, & i \neq t \end{cases} \]  

(3)

in which

\[ q_i = \frac{e^{s_i}}{\sum_{j=1}^{C} e^{s_j}}. \]  

(4)

During training, the supervisory signal \( \frac{\partial L_S}{\partial s_i}(i = t) \) guides the neural network to reduce intra-class variation, whereas \( \frac{\partial L_S}{\partial s_i}(i \neq t) \) increases the inter-class variation. Note: \( 0 < q_i < 1 \) and \( \sum_{i=1}^{C} q_i = 1 \). Both \( q_t - 1 \) and \( \sum_{i=1}^{C} q_i \) will approach 0 when the loss \( L_S \) decreases. The intensity of the supervisory signal weakens as the training samples classify correctly. In this, the supervisor fails to further reduce the intra-class or increase the inter-class variation of the face features.

Figure 3(a) shows the trained distribution. We also illustrate the decay of the gradient with a randomly selected sample. Figure 3(b) shows the gradient of softmax loss with respect to scores. After 8 training epochs, the gradients become almost 0. In this, the supervisor fails to help the DCNN to get more compact features’ distribution.

2.2 Gradient-Enhanced Softmax

As shown in Fig. 4, the proposed supervisor consists of two parts: the score calculator and gradient-enhanced softmax loss. The blue dotted box indicates the score calculator, whereas the orange dotted box indicates the GESoftmax loss.

In our method, two differences are noted as compared with the conventional softmax classifier. The first difference
is marked with a green dotted box in Fig. 4. We use the constant-normalized cosine as the score for each category to prevent the over-enhancement caused by GESoftmax. The formulation is as follows:

\[
s_i = \alpha \frac{\mathbf{w}^T \mathbf{f}_i}{\|\mathbf{w}\| \|\mathbf{f}_i\|} \tag{5}
\]

where \(\|\cdot\|\) is the L2-norm of the input vector; \(\alpha\) is a constant. \(\alpha\) is called the temperature parameter, set to 30 in our experiments. More information about \(\alpha\) is in NormFace [6]. The second difference is the GESoftmax loss function computed as:

\[
L_{GES} = \lambda \cdot \log(\sum_{j=1,j\neq i}^{C} e^{s_j}) + (\lambda - 1)s_i, \tag{6}
\]

where \(s_i\) is the intra-class score, and \(\log(\sum_{j=1,j\neq i}^{C} e^{s_j})\) is the soft maximum of \(s_j(1 \leq j \leq C, j \neq i)\) that describes the inter-class score. \(\lambda\) is a hyperparameter with values ranging from 0 to 1, introduced to control the proportion of the two parts. Minimizing this loss will promote the intra-class score and restrain the inter-class scores. The gradient of GESoftmax loss with respect to scores is formulated as:

\[
\frac{\partial L_{GES}}{\partial s_i} = \begin{cases} 
\lambda - 1, & i = t \\
\frac{\lambda e^{s_t}}{\sum_{j=1,j\neq t}^{C} e^{s_j}}, & i \neq t.
\end{cases} \tag{7}
\]

In GESoftmax, the intra-class supervisory signal \(\partial L_{GES}/\partial s_i(i = t)\) is a negative constant. The inter-class signal \(\partial L_{GES}/\partial s_i(i \neq t)\) is positive and \(\sum_{j=1,j\neq i}^{C} \partial L_{GES}/\partial s_j = 1\). The supervisory signal intensity will not decay during training, and the supervised DCNN has sufficient momentum to reach better locations. In Fig. 5(a), we illustrate training sample features’ distribution and the gradient of GESoftmax loss with respect to scores in Fig. 5(b). Compared with Fig. 3, the gradients from GESoftmax loss are adequate during training, revealing that the trained intra-class divisions of features are more compact.

3. Experiments

In this section, we performed a series of experiments for FR to verify the effectiveness of GESoftmax. Current prevailing supervisors were reproduced for comparison, and the sensitivity of the hyperparameter \(\lambda\) was investigated.

3.1 Implementation Details

3.1.1 Dataset and Face Image Preprocessing

The training data were a small subset of publicly available VGGFace2 [11]. After removing the identities appearing in LFW [9], the training set consisted of 8529 identities. To avoid the long tail effect, we randomly selected 50 images for each identity, and the final training data were approximately 0.43M images. All faces in these images and their landmarks were detected by Multitask Cascaded CNN algorithm [12], and the detected faces were aligned with similarity transformation according to their landmarks.

3.1.2 Network Architecture

In our experiments, MobileFaceNet [13] was employed as the network architecture. The detailed architecture was shown in Table 1, in which the notations followed MobileNetV2 [14]. Each line described a sequence of operators, repeated \(n\) times and all layers in the same sequence consisted of the same number \(c\) of output channels. \(s\) was

![Fig. 5](image-url) (a) Training samples’ features distribution with GESoftmax loss. (b) The gradient of GESoftmax loss with respect to scores vs. training epoch.

| Input | Operator | t | c | n | s |
|-------|----------|---|---|---|---|
| 3 \times 112 \times 96 | Conv3 \times 3 | - | 64 | 1 | 2 |
| 64 \times 56 \times 48 | Depthwise Conv3 \times 3[14] | - | 64 | 1 | 1 |
| 64 \times 56 \times 48 | Bottleneck[14] | 2 | 64 | 5 | 2 |
| 64 \times 28 \times 24 | Bottleneck | 4 | 128 | 1 | 2 |
| 128 \times 14 \times 12 | Bottleneck | 2 | 128 | 6 | 1 |
| 128 \times 14 \times 12 | Bottleneck | 4 | 128 | 1 | 2 |
| 128 \times 7 \times 7 | Bottleneck | 2 | 128 | 2 | 1 |
| 128 \times 7 \times 7 | Conv1 \times 1 | - | 512 | 1 | 1 |
| 512 \times 7 \times 6 | GDConv7 \times 6[13] | - | 512 | 1 | 1 |
| 512 \times 1 \times 1 | Linear | - | 128 | 1 | 1 |
the stride used in the first layer of each sequence. All spatial convolutions used $3 \times 3$ kernels, and the expansion factor $t$ was applied to the input size.

3.1.3 Training Settings

We used the conventional mini-batch stochastic gradient descent to optimize the models, and all the models were trained from scratch. The batch size and weight decay parameter were set to 256 and $5e-4$, respectively. The learning rate began with 0.1, divided by 10 at the 21st, 36th, 46th, and 51st epoch. The complete training was finished at the 55th epoch.

3.2 Evaluation on LFW

We evaluate the trained models on LFW dataset [9]. Consisting of large variations in pose, expression, and illumination, LFW is a web-collected dataset that contains 13233 images from 5749 different identities. We extract the features of each face image and use the cosine distance as the similarity of two faces. All the models are evaluated using different protocols. The first is the standard unrestricted protocol of Large-scale Unconstrained Face Recognition (BLUFR) protocol [15]. Table 2 presents the test results with various supervisors. Compared with the current prevailing supervisors, GESoftmax achieves the best performance for FR.

3.3 Effect of $\lambda$

The hyperparameter $\lambda$ in Eq. (6) determines the proportion between intra-class and inter-class supervisory signal intensity. With $\lambda$ changes from 0 to 1, the inter-class signal increases gradually, whereas the intra-class signal decreases. In Table 2, we list the performance of GESoftmax where $\lambda$ varies from 0.45 to 0.55. From the table, we can obtain the best performance when $\lambda$ takes a certain value. However, a further study is required to investigate an appropriate method.

4. Conclusion

In this letter, we revealed the gradient vanishing problem in the conventional softmax classifier and proposed a novel softmax-based supervisor named as GESoftmax. The GESoftmax can effectively alleviate the gradient vanishing problem and promote the performance of the DCNN for FR. The evaluation on LFW shows that the proposed supervisor achieves better results than the current state-of-the-art softmax-based approaches.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 61572458 and 61901436).

References

[1] Y. Sun, X. Wang, and X. Tang, “Deep learning face representation from predicting 10,000 classes,” Proc. IEEE Conference on Comput. Vis. Pattern Recognit., pp.1891–1898, 2014.
[2] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” Proc. IEEE Conference on Comput. Vis. Pattern Recognit., pp.815–823, 2015.
[3] Y. Sun, Y. Chen, X. Wang, and X. Tang, “Deep learning face representation by joint identification-verification,” Advances in neural information processing systems, vol.2, pp.1989–1996, Dec. 2014.
[4] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” European conference on computer vision, pp.499–515. Springer, 2016.
[5] O.M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” British Machine Vision Conference, p.6, 2015.
[6] F. Wang, X. Xiang, J. Cheng, and A.L. Yuille, “NormFace: L2 hypersphere embedding for face verification,” Proc. 25th ACM international conference on Multimedia, pp.1041–1049, ACM, Oct. 2017.
[7] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “SphereFace: Deep hypersphere embedding for face recognition,” Proc. IEEE Conference on Comput. Vis. Pattern Recognit., pp.212–220, 2017.
[8] F. Wang, J. Cheng, W. Liu, and H. Liu, “Additive margin softmax for face verification,” IEEE Signal Process. Lett., vol.25, no.7, pp.926–930, July 2018.
[9] G.B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: a database for studying face recognition in unconstrained environments,” 2008.
[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proc. IEEE., vol.86, no.11, pp.2278–2324, Nov. 1998.
[11] Q. Cao, L. Shen, W. Xie, O.M. Parkhi, and A. Zisserman, “VGGFace2: A dataset for recognising faces across pose and age,” 2018 13th IEEE Int. Conf. Automatic Face & Gesture Recognition (FG 2018), pp.67–74, IEEE, 2018.
[12] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and

| Supervisor         | LFW 6000pairs | LFW BLUFR VR@FAR=0.1% | LFW BLUFR DIR@FAR=1% |
|--------------------|---------------|-----------------------|-----------------------|
| Softmax[1]         | 95.05% ± 1.25%| 78.41% ± 3.69%        | 41.89% ± 3.13%        |
| Softmax-Triplet[5] | 98.15% ± 0.73%| 88.21% ± 2.57%        | 48.85% ± 3.98%        |
| Softmax+Contrastive[3] | 98.37% ± 0.71% | 90.89% ± 2.00% | 48.55% ± 3.89% |
| Softmax+Center[4]  | 97.58% ± 0.92%| 92.57% ± 1.26%        | 61.01% ± 3.60%        |
| L2-Softmax[6]      | 97.47% ± 0.80%| 90.48% ± 1.64%        | 63.38% ± 2.95%        |
| A-Softmax[7]       | 98.43% ± 0.58%| 96.15% ± 1.15%        | 76.56% ± 1.97%        |
| AMSoftmax[8]       | 98.62% ± 0.65%| 97.29% ± 1.07%        | 80.20% ± 2.52%        |
| GESoftmax ($\lambda$ = 0.45) | 98.67% ± 0.65% | 98.22% ± 0.70% | 84.75% ± 1.37% |
| GESoftmax ($\lambda$ = 0.475) | 98.97% ± 0.57% | 98.52% ± 0.71% | 85.73% ± 1.11% |
| GESoftmax ($\lambda$ = 0.5) | 98.78% ± 0.42% | 98.31% ± 0.66% | 86.68% ± 1.20% |
| GESoftmax ($\lambda$ = 0.525) | 98.83% ± 0.54% | **98.61**% ± 0.53% | **88.16**% ± 1.35% |
| GESoftmax ($\lambda$ = 0.55) | 98.68% ± 0.45% | 98.53% ± 0.68% | 86.45% ± 1.58% |
alignment using multitask cascaded convolutional networks,” IEEE Signal Process. Lett., vol.23, no.10, pp.1499–1503, Oct. 2016.

[13] S. Chen, Y. Liu, X. Gao, and Z. Han, “Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices,” Chinese Conference on Biometric Recognition, pp.428–438, Springer, 2018.

[14] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” Proc. IEEE Conference on Comput. Vis. Pattern Recognit., pp.4510–4520, 2018.

[15] S. Liao, Z. Lei, D. Yi, and S.Z. Li, “A benchmark study of large-scale unconstrained face recognition,” IEEE International Joint Conference on Biometrics, pp.1–8, IEEE, 2014.