AirFace: Lightweight and Efficient Model for Face Recognition

Xianyang Li  Feng Wang  Qinghao Hu  Cong Leng
Nanjing Artificial Intelligence Chip Research, Institute of Automation, Chinese Academy of Sciences (AIRIA)
China
qq1052308104@gmail.com  {wangfeng, huqinghao, lengcong}@airia.cn

Abstract

With the development of convolutional neural networks, significant progress has been made in computer vision tasks. However, the commonly used loss function softmax loss and highly efficient network architectures for common visual tasks are not as effective for face recognition. In this paper, we propose a novel loss function named Li-ArcFace based on ArcFace. Li-ArcFace takes the value of the angle through a linear function as the target logit rather than through cosine function, which has better convergence and performance on low dimensional embedding feature learning for face recognition. In terms of network architecture, we improved the performance of MobileFaceNet by increasing the network depth, width and adding attention module. Besides, we found some useful training tricks for face recognition. Under all the above effects, we won the second place in the deepglint-light challenge of L-FR2019.

1. Introduction

The development of deep convolutional networks (DCNN) has made remarkable progress in a series of computer vision tasks. However, it is not so effective while using the common method for face recognition. Softmax loss, which is commonly used in classification, can’t maximize inter-class variance and minimize intra-class variance of embedding feature vectors. In order to obtain highly discriminative embedding features (See Figure 1), a series of novel loss functions have been proposed in recent years, such as A-Softmax [12], CosFace/AM-Softmax [21, 19], ArcFace [5]. Among them, ArcFace achieved state-of-the-art performance by adding additive margin between classes in the angle space. But, when ArcFace is used to train some efficient networks with small(128) embedding size, it’s hard to train from scratch (for example, the embedding feature size of MobileFaceNet [4] is 128, it can’t converge when trained from scratch with ArcFace). As a result, to guarantee the convergence of training, pre-training is always required with softmax loss. To overcome this problem, we carefully designed a novel loss function named Li-ArcFace based on ArcFace, which performs better in convergence.

Face recognition technology is now widely used on mobile devices, which requires that the computational cost of the model should not be too large. In recent years, some highly efficient neural network architectures have been proposed, such as MobileNetV1 [10], ShuffleNet [23], and MobileNetV2 [16], but they’re all designed for common visual recognition tasks instead of face recognition. Their performance is really general when used for face recognition directly. MobileFaceNet is designed for face recognition based on MobileNetV2, achieved remarkable accuracy on LFW [11], AgeDB [15]. And it is even comparable to state-of-the-art big DCNN model on MegaFace [14] Challenge 1 under the much smaller computational resources. In this paper, with limited amount of computation, we carefully designed a higher performance network architecture based on MobileFaceNet.

Figure 1. Schematic diagram of discriminative embedding features. W1 refers to the center of the class 1, f1 refers to the embedding feature of class 1, f2 refers to the embedding feature of class 2. During training, the variance of the same person becomes smaller, and the variance of different people becomes larger.
2. Related Work

**Loss function.** Softmax loss is the most commonly used loss function in classification, which is presented as:

\[
L_1 = -\frac{1}{N} \sum_{i=1}^{N} \log \left( \frac{e^{\theta_{y_i} \mathbf{x}_i}}{\sum_{j=1}^{n} e^{\theta_{y_j} \mathbf{x}_i}} \right)
\]

where \(x_i\) denotes the embedding feature of the \(i\)-th sample belonging to the \(y_i\)-th class, and the dimension of the embedding feature (hereinafter referred to as embedding size) is set as \(d\). \(W_j \in \mathbb{R}^{d \times n}\) denotes the \(j\)-th column of the weight \(W \in \mathbb{R}^{d \times n}\) and \(b_j \in \mathbb{R}\) is the bias term. The batch size is \(N\) and the class number of training data is \(n\). However, the traditional softmax loss lacks the power to supervise the embedding feature to minimize inter-class similarity and maximize intra-class similarity. In SphereFace [12] and NormFace [20], the bias term is being removed at first and then fixing the \(\|W_j\| = 1\), \(\|x_i\| = s\) by \(l_2\) normalisation, such that the logit is:

\[
W_j^T x_i = \|W_j\| \times \|x_i\| \times \cos \theta_j = s \times \cos \theta_j
\]

Where \(\theta_j\) denotes the angle between \(x_i\) and \(W_j\). Thus the logit is only depend on the cosine of the angle. The modified loss function can be formulated as:

\[
L_2 = -\frac{1}{N} \sum_{i=1}^{N} \log \left( \frac{e^{s \cos \theta_{i, j}}}{\sum_{j=1}^{n} e^{s \cos \theta_j}} \right)
\]

Although \(L_2\) guarantees a high similarity of features of the same person, it can’t separate different classes very well. In this paper, we use N-Softmax denotes \(L_2\). In ArcFace [5], the authors added an additive angular margin \(m\) within \(\cos \theta_{y_i}\), which can enhance the intra-class compactness and inter-class discrepancy, the ArcFace is formulated as:

\[
L_3 = -\frac{1}{N} \sum_{i=1}^{N} \log \left( \frac{e^{s \cos (\theta_{y_i} + m)}}{e^{s \cos \theta_{y_i}} + \sum_{j=1 \neq y_i}^{n} e^{s \cos \theta_j}} \right)
\]

When using ArcFace to train models with 512-dimensional embedding feature, it has well convergence and state-of-the-art performance. However, it will be difficult to converge while training some highly efficient networks with 128-dimensional embedding feature from scratch.

**network architectures.** Face recognition is being used more and more on mobile devices. So it’s really important to optimize the trade-off between accuracy and computational cost when designing deep neural network architecture. In recent years, some highly efficient neural network architectures have been proposed for common visual tasks. MobileNetV1 [10] used depthwise separable convolution instead of traditional convolution to reduce computational cost and improve network efficiency. MobileNetV2 [16] introduced inverted residuals and linear bottlenecks to further improve network efficiency. However, these lightweight network architectures are not so accurate when using these unchanged for face recognition. The author of MobileFaceNet [4] found the weakness of common mobile networks for Face recognition, and solved it by replacing global average pooling with global depthwise convolution(GDC). And the network architecture of MobileFaceNet is specifically designed for face recognition with smaller expansion factors in bottlenecks and more output channels at the beginning of the network architecture.

3. Proposed Approach

3.1. Li-ArcFace

In ArcFace, the authors added an angular margin \(m\) within \(\cos \theta_{y_i}\), which takes \(s \times \cos (\theta_{y_i} + m)\) as the target logit. The loss function proposed by us takes the angle after a linear function as the logit rather than cosine function. In the same way, we remove the bias term and then fix the \(\|W_j\| = 1\), \(\|x_i\| = s\) by \(l_2\) normalisation, the \(\theta_j\) denotes the angle be-
between \( x_i \) and \( W_j \), thus \( \theta_j = \arccos \left( \frac{W_j^T x_i}{s} \right) \in [0, \pi] \). At first, we constructed a linear function \( f(x) = (\pi - 2x) / \pi \), and we have \( f(\theta_j) \in [-1, 1] \). Then we add an additive angular margin \( m \) in the target logit. In the end we have \( s \times (\pi - 2(\theta_{ij} + m)) / \pi \) as the target logit. We call this novel loss function Li-ArcFace, the prefix Li refers to the linear function. The whole Li-ArcFace can be formulated as follows

\[
L_4 = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{s(\pi - 2(\theta_{ij} + m)) / \pi} + \sum_{j=1, j \neq y_i} e^{s(\pi - 2\theta_j) / \pi}}{e^{s(\pi - 2\theta_{ij}) / \pi}}
\]

(5)

There are two main advantages of using this linear function to replace cosine function. Firstly, it is monotonic decreasing when the angle is between 0 and \( \pi + m \), which will have better convergence, especially for model with small embedding size. For example, when training MobileFaceNet with ArcFace, it will lead to divergence (NaN). Therefore, softmax loss must be used for pre-training before convergence. The proposed loss does not require these two-stage training. Secondly, the penalty of the proposed loss function increases linearly as the angle between embedding feature \( x_i \) and center \( W_{yi} \) increasing, so that the target logit decreases linearly which is more intuitive (See Figure 2). It will not decline rapidly in some angles, but slowly in others (corresponding to the gradient value of the target logit curve), which makes the proposed loss function have better performance in the end. In terms of geometric decision margins, ArcFace will have a part of overlap area when the \( x_i \) deviates too much from the center \( W_{yi} \), because of its non-monotonicity of the target logit curve. This area can be distinguished as class 1 and class 2 with ArcFace. The proposed loss function wouldn’t have this overlap area (See Figure 3).

### 3.2. Network Architectures

In this section, we introduce our proposed network architecture. Our network architecture is based on a deeper MobileFaceNet(y2) [5], so the residual bottlenecks proposed in MobileNetV2 are used as our main building blocks. Table 1 shows the details of our network architecture. We follow the MobileFaceNet, expansion factors for bottlenecks in our architecture are much smaller than those in MobileNetV2 and using PReLU as the non-linearity rather than ReLU. Nevertheless, we noticed the importance of network width in face recognition. MobileFaceNet is significantly wider than MobileNetV2 at the beginning of the network. But MobileFaceNet does not double the output channels of the network when sampling at last stage. We did that within limited amount of computation. At the same time, we carefully adjusted the depth of the network, and we introduced the attention module CBAM [22] into every bottleneck in the network. But we changed the last activation function in the attention module from sigmoid to \( 1 + \tanh \). The range of \( 1 + \tanh \) is 0 to 2. When a channel or spatial position needs to be enhanced, it is multiplied by a weight between one and two; when it needs to be weakened, it is multiplied by a weight between zero and one. Using \( 1 + \tanh \) instead of sigmoid is more intuitive and makes training converge faster.

### 3.3. Training Tricks for Face Recognition

During the competition of LFR, we found some useful training tricks for face recognition. Firstly, using a variety of loss functions to fine-tune the model will make the features more robust and improve the accuracy to some extent. In the competition, we used Li-ArcFace, ArcFace, combined loss to fine-tune our model. Secondly, in 512-dimensional embedding feature space, it is difficult for the lightweight model to learn the distribution of the features. It is an effective method to use some large models to guide the feature distribution of lightweight models [9, 13].

### 4. Experiments

#### 4.1. Ablation Experiment of Li-ArcFace

In this section, we mainly introduce some comparative experiments of different loss functions.

**Implementation Details.** We use the MobileFaceNet as the network architecture, in which the embedding size is set as 128. And batch size is set as 256 x 4. Training models on four NVIDIA TITIAN XP GPUs. We use SGD with momentum 0.9 to optimize models. The scale of the feature distribution of lightweight models [12, 11] is Age Database (AgeDB) [18] and Celebrities in Frontal Profile (CFP) [18]. In this paper, ArcFace (m=0.5, NS) refers to the attention module from sigmoid to \( 1 + \tanh \). The range of \( 1 + \tanh \) is 0 to 2. When a channel or spatial position needs to be enhanced, it is multiplied by a weight between one and two; when it needs to be weakened, it is multiplied by a weight between zero and one. Using \( 1 + \tanh \) instead of sigmoid is more intuitive and makes training converge faster.
Table 2. Verification performance (%) of different weight decay values. Wdml10 represents the weight decay multiplier (wd_mult) parameter of the last convolution layers is 10.

<table>
<thead>
<tr>
<th>Weight decay</th>
<th>LFW</th>
<th>CFP-FP</th>
<th>AgeDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>5e-4</td>
<td>99.17</td>
<td>94.09</td>
<td>92.95</td>
</tr>
<tr>
<td>5e-4 wdm110</td>
<td>99.18</td>
<td>93.86</td>
<td>93.18</td>
</tr>
<tr>
<td>4e-5 wdm110</td>
<td>99.05</td>
<td>92.07</td>
<td>92.00</td>
</tr>
</tbody>
</table>

Table 3. Verification performance (%) of different loss functions. ArcFace (m=0.5, NS) refers to the margin of ArcFace is set as 0.5, and pre-training via N-Softmax before using ArcFace.

<table>
<thead>
<tr>
<th>Loss</th>
<th>LFW</th>
<th>CFP-FP</th>
<th>AgeDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-Softmax</td>
<td>97.87</td>
<td>90.46</td>
<td>86.55</td>
</tr>
<tr>
<td>Li-ArcFace(m=0.35)</td>
<td>99.20</td>
<td>94.20</td>
<td>92.87</td>
</tr>
<tr>
<td>Li-ArcFace(m=0.4)</td>
<td>99.27</td>
<td>94.11</td>
<td>93.25</td>
</tr>
<tr>
<td>Li-ArcFace(m=0.45)</td>
<td>99.23</td>
<td>94.20</td>
<td>93.15</td>
</tr>
<tr>
<td>Li-ArcFace(m=0.5)</td>
<td>99.20</td>
<td>94.10</td>
<td>92.93</td>
</tr>
<tr>
<td>Li-ArcFace(m=0.4, NS)</td>
<td>99.23</td>
<td>94.51</td>
<td>92.90</td>
</tr>
<tr>
<td>Li-ArcFace(m=0.45, NS)</td>
<td>99.27</td>
<td>94.39</td>
<td>93.22</td>
</tr>
<tr>
<td>ArcFace(m=0.4, NS)</td>
<td>99.27</td>
<td>93.93</td>
<td>92.72</td>
</tr>
<tr>
<td>ArcFace(m=0.45, NS)</td>
<td>99.27</td>
<td>94.00</td>
<td>92.98</td>
</tr>
<tr>
<td>ArcFace(m=0.5, NS)</td>
<td>99.18</td>
<td>93.86</td>
<td>93.18</td>
</tr>
<tr>
<td>CosFace/AM-Softmax</td>
<td>99.20</td>
<td>93.39</td>
<td>92.55</td>
</tr>
</tbody>
</table>

Effect of Hyper-parameter m. In Table 3, we firstly explored the optimal setting for m of Li-ArcFace, and then we found it was between 0.4 and 0.45. We tried training model from scratch with ArcFace, but diverged after about 1600 iterations. Therefore, all the experiments with ArcFace were pre-trained by N-Softmax. Since the embedding size is smaller than 512, we also explored the optimal setting for m of ArcFace, and we found it was between 0.45 and 0.5.

Comparison with state-of-the-art loss functions. In Table 3, the difference between Li-ArcFace, ArcFace and Cosface is tiny on LFW, but all of them are obviously better than N-Softmax; On CFP-FP and AgeDB, Li-ArcFace is slightly better than ArcFace and CosFace. We have drawn the accuracy on CFP-FP and AgeDB during the training, which makes the contrast more obvious (See Figure 4). We also compared Li-ArcFace and ArcFace in the same situation pre-trained by N-Softmax. On CFP-FP, Li-ArcFace achieves the highest accuracy, and Li-ArcFace is still slightly better than ArcFace on AgeDB. In general, there is little difference in accuracy on the verification sets, but Li-ArcFace has better convergence and does not need pre-training stage when training the model with small embedding size.

4.2. Evaluation Results of Network Architecture and Training Tricks

We name our model that contains our network architecture and training tricks as AirFace. Under the same training data set MS1M-retina [2] and model constraints, (MS1M-retina is cleaned from [8] by [5], and [7] is the face detector and alignment tool used to pre-process the data) the accuracy of AirFace reached 88.415%@FPR=1e-8 in deepglint-light challenge of LFR19 [6, 3]. Meanwhile, we verified the performance of AirFace in MegaFace Challenge 1 compared with the previous state-of-the-art models. In Table 4, AirFace has reached incredible efficiency and performance.
5. Conclusions

In this paper, first of all, we propose a novel additive margin loss function for deep face recognition based on ArcFace. The proposed loss function solves the problem that ArcFace does not converge in training model with small embedding feature size. And it achieves the state-of-the-art results on several face verification datasets. Second, we have carefully designed an efficient network architecture and explored some useful training tricks for face recognition, which makes our model AirFace extremely efficient at both deepglint-light challenge and MegaFace Challenge 1.

References

[1] https://github.com/deepinsight/insightface. 4