Imbalance Robust Softmax for Deep Embeeding Learning

Hao Zhu[1*], Yang Yuan[2*], Guosheng Hu[2,4], Xiang Wu[3], and Neil Robertson[4]

[1] Australian National University, Canberra, Australia
[2] Anyvision, Belfast, United Kingdom
[3] Reconova, Beijing, China
[4] Queens University of Belfast, Belfast, United Kingdom
[1]Hao.Zhu@anu.edu.au,[2]bengouawu@gmail.com

Abstract. Deep embedding learning is expected to learn a metric space in which features have smaller maximal intra-class distance than minimal inter-class distance. In recent years, one research focus is to solve the open-set problem by discriminative deep embedding learning in the field of face recognition (FR) and person re-identification (re-ID). Apart from open-set problem, we find that imbalanced training data is another main factor causing the performance degradation of FR and re-ID, and data imbalance widely exists in the real applications. However, very little research explores why and how data imbalance influences the performance of FR and re-ID with softmax or its variants. In this work, we deeply investigate data imbalance in the perspective of neural network optimisation and feature distribution about softmax. We find one main reason of performance degradation caused by data imbalance is that the weights (from the penultimate fully-connected layer) are far from their class centers in feature space. Based on this investigation, we propose a unified framework, Imbalance-Robust Softmax (IR-Softmax), which can simultaneously solve the open-set problem and reduce the influence of data imbalance. IR-Softmax can generalise to any softmax and its variants (which are discriminative for open-set problem) by directly setting the weights as their class centers, naturally solving the data imbalance problem. In this work, we explicitly re-formulate two discriminative softmax (A-Softmax and AM-Softmax) under the framework of IR-Softmax. We conduct extensive experiments on FR databases (LFW, MegaFace) and re-ID database (Market-1501, Duke), and IR-Softmax outperforms many state-of-the-art methods.

1 Introduction

Recently, convolutional neural networks (CNNs) have significantly boosted the state-of-the-art performance in many computer vision tasks especially in image classification [1,2,3,4,5,6]. Not surprisingly, CNNs have achieved great success in the field of biometrics, in particular, face recognition (FR) [7,8,9,10] and person re-identification (re-ID) [11,12,13]. This success is derived from the fact that CNNs are able to encode images into rich, semantic and discriminative representations (features) which can be used to effectively measure the similarity between two identity-related images. These two tasks

*indicates equal contribution.

(FR and re-ID) differ from general image classification in terms of two challenges: open-set setting and data imbalance in the training set.

Open-set setting is much more widely applied than close-set for FR and re-ID. For open-set setting, the identities of test set are disjoint with those of training set. In the real world, FR and re-ID system train the CNN (feature extractor) using images collected from one specific group of people, e.g. celebrities from IMDb in CASIA WebFace [14] database. During test stage, however, the FR and re-ID systems work in places, such as one police station, where the gallery (blacklist) and the probe (people appear in this police station) are mostly likely disjoint with training set (e.g. those celebrities). In contrast, classical image classification (e.g. ImageNet Challenge) uses the close-set setting where training and test sets share the same classes. Traditionally, both open-set and close-set problems adopt the softmax function because of its simplicity and probabilistic interpretation. Together with the cross-entropy loss, they form arguably one of the most commonly-used components in CNN architectures.

Under open-set setting, however, softmax suffers from one drawback: deep learning with softmax loss only learns separable features that are not discriminative enough for 'unseen' classes in testing. It results from the fact that softmax loss does not explicitly optimise the intra- and inter-class distances. To address this, some methods combine the softmax loss with metric learning [9,15,10] to enhance the discrimination power of features. Metric learning based methods commonly suffer from the way of building mini-batches by sampling. Other methods try to add new constraints (e.g. center loss [16], large-margin term [17,18], L2 normalization [19,20]) that make features more compact and thus more discriminative.

Data imbalance is another challenge for FR and re-ID. Unlike those popular datasets MNIST [21], CIFAR-10 [22] and ImageNet [23], FR and re-ID datasets are commonly highly imbalanced. As shown in Fig.1, only a limited number of identities appear frequently (more than hundreds), while most of the others appear relatively rarely (fewer than ten times) in the popular face database CASIA-Webface [14] and re-ID database Market-1501 [24]. Surprisingly, very little research explores the problem of data imbalance in FR and re-ID. In this paper, we show that deep embedding learning with the most widely used softmax (and its variants such as A-Softmax [18]) encounters difficulty in the presence of imbalanced training data even using either metric learning or other regularizations. Although some softmax variants such as A-Softmax [18] can solve the open-set problem by learning compact features, they do not perform well when the training data is imbalanced. To our knowledge, the only work exploring the data imbalance problem for FR is the range loss [25]. However, range loss does not deeply investigate the reason why this imbalance impacts the deep embedding learning.

In this work, we aim to learn deep embeddings which can achieve two targets: 1) being discriminative for open-set and 2) being robust to data imbalance. As existing works [16,17,18,19,20], target 1) can be achieved by learning compact features (i.e. reduce intra-class variance). To achieve target 2), we have to first investigate why data imbalance influences the performance of softmax-based deep classification. In this work, we explore the reason. *During the back-propagation training with imbalanced data, two strengths, which determine the update of the weights (usually the penultimate fully-connected layer), are imbalanced (see Eq. 2): the one keeping the weights at their class*

centers is much smaller than that pushing them away. This imbalance causes the weight of the class with minor samples being far away from its class center, leading to degraded classification performance. Based on this analysis, target 2) can be achieved by making the weight from the class with minor samples close to its class center.

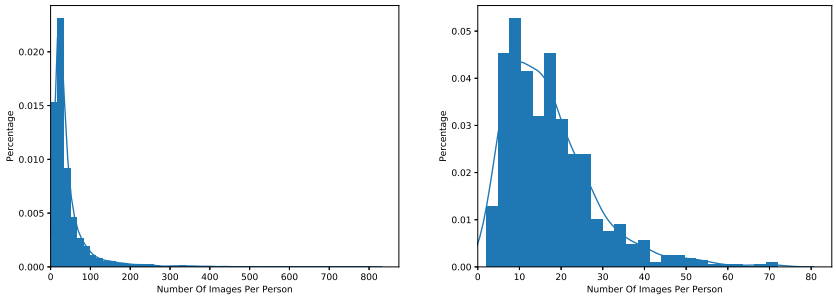
To simultaneously achieve targets 1) and 2) , we propose a uniformed framework, Imbalance-Robust Softmax (IR-Softmax). First, IR-Softmax solves the open-set problem by being compatible with the softmax variants (e.g. A-Softmax [18], AM-Softmax [26] which can learn discriminative embeddings. Second, motivated by the aforementioned analysis on data imbalance, IR-Softmax alleviates the influence of data imbalance by setting the weights as their class centers in the feature space instead of updating with back-propagation. In this way, IR-Softmax effectively avoids the shift between the weights and their centers, which is the main reason of performance degradation caused by data imbalance detailed in Section 3.1.

Our contributions can be summarised as:

1. We deeply investigate the reason why data imbalance degrades the performance of softmax-based classifications in the perspective of neural network optimisation (Section 3.1) and feature distribution (Section 3.2).

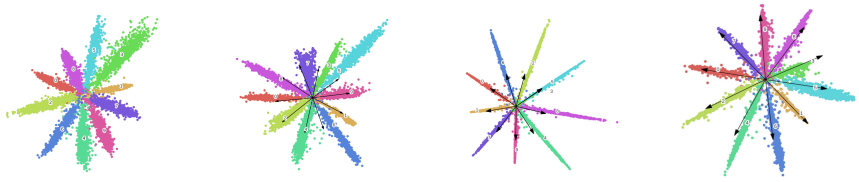
2. IR-Softmax can learn embeddings which are discriminative under open-set protocol. In particular, IR-Softmax is a unified framework, e.g. it can generalise to softmax and its variants (e.g. A-Softmax [18], AM-Softmax [26]) to achieve discriminative feature learning. More importantly, IR-Softmax can effectively reduce the influence of data imbalance by bridging the gap between weights (the penultimate fully-connected layer) and their class centers in feature space.

3. Our extensive experiments demonstrate the effectiveness and generalisation of the proposed IR-Softmax, and we achieve state-of-the-art performance on challenging FR (LFW [27], MegaFace [28]) and re-ID (Market-1501 [24], DUKE-MTMC [29]) benchmarks. The code will be made publicly available.



(a) WebFace-CASIA (b) Market-1501

Fig. 1. Long-tailed distribution on WebFace-CASIA [14] and Market-1501 [24] database. The number of images per person drops drastically, and only a few identities have a large number of images.



(a) Softmax, bal. (b) Softmax, imbal. (c) A-Softmax, imbal. (d) **IR-Softmax**, imbal.

Fig. 2. The distribution of deeply learned features. 'bal.' (balanced) setting contains 10 classes, which all has 6,000 images from MNIST [21]. In contrast, 'imbal.' (imbalanced) setting contains 6000 images for all classes but class '3' with 300 images. A-Softmax in (c) refers to [18]. The label of each class is plotted on its center. In addition, we also plot the weights (from the fully-connected penultimate layer) to each class with an arrow in (b)-(d). Note that our fully-connected layer consists of only 2 neurons to facilitate visualisation.

2 Related Work

In this section, we briefly review the methods of discriminative feature learning in the field of face recognition (FR) and person re-identification (re-ID). Recently, two popular ways of deep embedding learning are: (1) metric learning and (2) discriminative softmax (softmax's variants which are more discriminative for open-set problem). Apart from these two strategies, we discuss the training data imbalanced problem in the field of FR and re-ID.

Metric Learning Metric learning is widely used for FR and re-ID. In practice, to learn more discriminative features, many works combine softmax loss and deep metric learning loss (contrastive [9,30] loss or triplet loss [10]). Unlike softmax, contrastive and triplet losses accept image pairs or triplets (3 or a multiple of 3) as input respectively. For contrastive loss, if the input pair belongs to the same class, their features are required to be as similar as possible. Otherwise, the contrastive loss would require their distance larger than a particular margin. Similar to contrastive loss, the triplet loss [10] encourages a similar distance constraint. Specifically, the triplet loss minimises the distance between an anchor sample and a positive sample (of the same identity) and maximises the distance between the anchor sample and a negative sample (of different identity). Clearly, the contrastive and triplet losses can encourage intra-class compactness and inter-class separability, making the learned feature more discriminative. However, both contrastive and triplet losses require a carefully-designed pair/triplet selection procedure. For example, using contrastive loss, it is hard to build training pairs from a mini-batch, especially for the training set with many classes. Normally the mini-batch size is not more than 256, while the number of categories is far more than 256 in the application of FR and re-ID. Clearly the online selection only produces a few positive pairs and much more negative ones.

Discriminative Softmax Apart from metric learning, some softmax variants are proposed, aiming to learn more discriminative features to solve the open-set problem. Wen

et al.[16] add a new supervision signal, called center loss, to softmax loss for face recognition. Specifically, the center loss simultaneously learns a feature center for each identity and penalises the distances between the deep features of examples and their corresponding feature centers. With the joint supervision of softmax loss and center loss, this method can easily obtain inter-class dispersion and intra-class compactness. Large-Margin Softmax loss [17] proposes a new perspective to softmax and optimises the angles between weights and features. However, the magnitude of weights are also considered, and thus it is also sensitive to data imbalance just the same as softmax. By contrast, A-Softmax loss [18] controls the magnitude of weights (i.e. $\|w\|_2 = 1$) and thus make the weights optimised in an angular space. Although A-Softmax is theoretically suitable for deep embedding learning, it actually does not work well in the setting of data imbalance detailed in Section 3. [31,26,32] relax the margin with more efficient and effective ways. Some works [19,20,32] try to optimise the features on a hyper-sphere to make features more discriminative.

Training Data Imbalance The aforementioned methods ignore the problem of training data imbalance which widely exist in FR and re-ID. In [33], researchers investigate many factors that influence the performance of fine-tuning for object detection with long-tailed distributions of samples. Their analysis and empirical results indicate that classes with more samples will achieve greater impact on the feature learning, and it is better to make the sample number more uniform across classes. In the field of FR and re-ID, unfortunately, the data imbalance problem is much worse than object detection [33]. Specifically, few identities have more than 1000 images and many identities have fewer than 10 images. Commonly a large-scale face dataset has more than 10,000 identities [14]. However, we still cannot simply discard these identities that only have few images. For face recognition, identities with few images cannot provide enough intra-class information for the model, but provide inter-class information which is more useful to open-set protocol. Many methods [25,34,35,36,37] have been proposed to solve the data imbalance in face recognition. However,these works do not deeply investigate the reason why the imbalance impacts softmax based deep embedding learning.

3 Methodology

In this section, we first provide insights into the influence of data imbalance on CNN performance by training a LeNet [21] on an imbalanced MNIST. Based on the conclusion drawn from the experiments, we propose a new loss function, Imbalance Robust Softmax (IR-Softmax), to reduce the influence of data imbalance while perform discriminative feature learning. Last, we discuss the relations between the proposed method and metric learning.

3.1 Motivation

Softmax regression (or multinomial logistic regression) is a generalisation of logistic regression to multi-class problem, therefore, softmax can handle $y_i \in \{1, \ldots, K\}$ (where K is the number of classes). Given a training set $\{(\mathbf{x}_i, y_i), \ldots, (\mathbf{x}_n, y_n)\}$, we learn an

embedding/projection $f(\mathbf{x})$, with which the softmax can be written as

$$J = -\frac{1}{n}\left[\sum_{i=1}^{n}\log\frac{\exp\left(f_{y_i}(\mathbf{x}_i)\right)}{\sum_{j=1}^{K}\exp(f_j(\mathbf{x}_i))}\right] \tag{1}$$

where f_j denotes the j-th dimension of the learned function $f(x)$, and n is the number of training samples. In CNNs, f is usually the output of a fully connected layer $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_k]$, so $f_j = \mathbf{w}_j^T\mathbf{x}_i + b_j$ and $f_{y_i} = \mathbf{w}_{y_i}^T\mathbf{x}_i + b_{y_i}$.

To analyse the influence of data imbalance, we come to the neural network optimisation process (we omit the bias term for simplicity):

$$\begin{aligned}\nabla_{\mathbf{w}_k}J &= -\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\left(1\{y_i = k\} - P(k|\mathbf{x}_i)\right) \\ &= \frac{1}{n}(\underbrace{\sum_{i=1}^{n}\mathbf{x}_i(P(k|\mathbf{x}_i) - 1)1\{y_i = k\}}_{term\ 1} + \underbrace{\sum_{i=1}^{n}\mathbf{x}_i P(k|\mathbf{x}_i)1\{y_i \neq k\}}_{term\ 2})\end{aligned} \tag{2}$$

where $P(k|\mathbf{x}_i) = \frac{\exp\left(f_k(\mathbf{x}_i)\right)}{\sum_{j=1}^{K}\exp(f_j(\mathbf{x}_i))}$, and $1\{\cdot\}$ is the indicator function: $1\{true\} = 1$, and $1\{false\} = 0$. It can be observed that the gradient of the parameter $\nabla_{\mathbf{w}_k}J$ contains two terms: term 1 (which is activated if $y_i = k$) and term 2 (if $y_i \neq k$). Thus the update of parameter $\mathbf{w}_k$ during optimisation depends on the samples not only from the k-th class but also from the other classes. Term 1 is actually the weighted center of the observed class; Term 2 can be viewed as the weighted centers of all the other classes if n is big enough. The update of $\mathbf{w}_k$ is determined by the balance of two strengths: one leads $\mathbf{w}_k$ to the center of class k (term 1), one 'pushes' w_k away from class k (weighted center of all the other classes). If the training data is imbalanced, the update of $\mathbf{w}_k$ corresponding to class k, which has much fewer samples than other classes, is fully dominated by term 2, making $\mathbf{w}_k$ being far away from center of class k.

To further analyse the influence of data imbalance on optimisation, we take one binary classification with softmax for example. As shown in Fig.3(a) and (b), there are both five samples for class 1 (blue points) and class 2 (red points) for balance setting; and nine samples for class 1 and one sample for class 2 for imbalance setting. Clearly, both $\mathbf{w}_1$ and $\mathbf{w}_2$ are influenced by all the samples from class 1 and 2. In imbalance setting (Fig.3(b)), $\mathbf{w}_2$ is determined by term 1 (1 sample from class 2) and term 2 (9 samples of class 1), where term 1 and 2 are detailed in Eq. (2). Clearly, the update of $\mathbf{w}_2$ is dominated by term 2, which pushes $\mathbf{w}_2$ far away from the center of class 2.

To explicitly show the influence of data imbalance on classification performance, we conduct a toy experiment on MNIST [21]. From Fig.2(b) and 2(c), not surprisingly, the data imbalance degrades the performance of the models trained with softmax and A-Softmax [18]. We can find the main issue caused by data imbalance: centers of relevant feature distributions being away from their weights (from penultimate fully-connected layer). For example, in Fig. 2(b), the feature center of class '3' (minor training data) and centers of '5' and '7' (the neighbours of '3') are all distant from their weights. Thus these biases (feature centers being far from its weights) caused by data imbalance

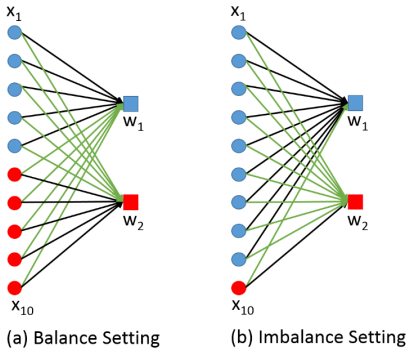


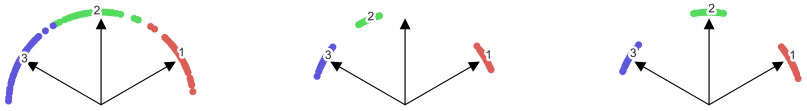
Fig. 3. One example of binary classification with softmax. (a) balance setting with five samples of class 1 (blue spots) and five samples of class 2 (red spots). (b) imbalance setting with nine samples of class 1 and one sample of class 2. Black and green lines indicate term 1 and term 2 in Eq. (2), influencing the update of $\mathbf{w}_1$ and $\mathbf{w}_2$ respectively.

will induce classification error for the corresponding categories. Though A-Softmax can learn discriminative features, it suffers from the same aforementioned bias problem as shown in Fig. 2(c). This observation provides the cue to solve the data imbalance problem and inspires our solution (Fig. 2(d)) detailed in Section 3.2.

3.2 Imbalance Robust Softmax (IR-Softmax)

In this work, we aim to learn features which can (i) improve the discriminative power of features in open-set protocol, and (ii) alleviate data imbalance problem. For (i), the desired open-set criterion is that the maximal intra-class distance is smaller than the minimal inter-class distance. However, softmax only maximises the the distance between weights rather then inter-class distance (Fig .4(a)). Derived from softmax, A-Softmax [18], however, only focuses on minimising intra-class distance, leading to compact features as shown in Fig. 4(b). For (ii), data imbalance can degrade the performance of deep CNNs. As analysed in Section 3.1, in particular Eq. (2), the data imbalance can cause imbalanced gradient updates during optimisation: the strength of making the weights close to feature center (term 1 in Eq. (2)) is much smaller than the strength of pushing away (term 2). This strength imbalance causes the weights being far away from their feature centers as shown in Fig. 4(b) , 2(b) and 2(c).

Based on the above analysis, to simultaneously solve the open-set and data imbalance problems, two criteria corresponding these two problems are: 1) minimising the intra-class distance by making the feature distribution with the same label more compact; 2) maximising the low bound of inter-class distance by making the center of features of each class being close enough (ideally equal) to its weights (usually the penultimate fully-connect layer). Criterion 1) can be achieved by learning compact features e.g. A-Softmax [18] and AM-Softmax [26]. To our knowledge, we are the first to investigate Criterion 2). To simultaneously achieve these two targets, we propose a



(a) Softmax, balanced (b) A-Softmax, imbalanced (c) IR-Softmax, imbalanced

Fig. 4. Feature distributions in angular space: (a) separable features under balanced data, feature center being close to its weight (the black arrow); (b) compact features under imbalanced data, and the feature center being far from its weight; (c) compact features under imbalanced data, the feature center being close to its weight.

novel framework, IR-Softmax*, which can achieve Criterion 1) by incorporating itself into discriminative softmaxs e.g. A-Softmax [18] and AM-Softmax [26].

Now we detail the way of meeting Criterion 2). As the analysis in Section 3.1, we can find that the imbalanced data causes the weights (from penultimate fully-connected layer) being away from their class centers after training as shown in Fig. 2(b), leading to degraded classification performance. Based on Criterion 2), the key idea of IR-Softmax is *setting the weights as their corresponding class centers in the feature space*, naturally avoiding the shift between the weights and their centers.

IR-Softmax is a unified framework which can be incorporated into softmax and its variants, leading to different forms of IR-Softmax. For classical softmax in Eq. (1), $f_j = \mathbf{w}_j^T \mathbf{x}_i + b_j$ is fed into softmax. In our IR-Softmax framework, $f_j^{'} = (\mathbf{c}_j^{'})^T \mathbf{x}_i + b_j$ replaces f_j, where $\mathbf{c}_j^{'}$, the center of features from class j, is defined as:

$$\mathbf{c}_j' = \frac{1}{\sum_i^n 1\{y_i = j\}} \sum_i^n \frac{1\{y_i = j\}\mathbf{x}_i}{\|\mathbf{x}_i\|_2} \tag{3}$$

Most importantly, in this work, we formulate two discriminative IR-Softmaxs derived from A-Softmax and AM-Softmax, respectively. To formulate A-Softmax and AM-Softmax which both normalise the weight w_j ($||\mathbf{w}_j||_2 = 1$), Eq. (1) can be modified as:

$$J = -\left[\sum_{i=1}^{m} \log \frac{\exp(\|\mathbf{x}_i\|\psi(\theta_{y_i}))}{\exp(\|\mathbf{x}_i\|\psi(\theta_{\mathbf{y}_i})) + \sum_{j\neq y_i}^{K} \exp(\|\mathbf{x}_i\|\cos(\theta_j))}\right] \tag{4}$$

From $f_j = \mathbf{w}_j^T \mathbf{x}_i + b_j$ (classical softmax) and $f_j^{'} = (\mathbf{c}_j^{'})^T \mathbf{x}_i + b_j$ (IR-Softmax version), we use $\mathbf{c}_j^{'}$ to replace $\mathbf{w}_j$. Similarly, we use $\mathbf{c}_j^T \mathbf{x}_i$ ($\mathbf{c}_j = \frac{\mathbf{c}_j'}{\|\mathbf{c}_j'\|_2}$) to replace $\|\mathbf{x}_i\|\cos(\theta_j)$

*code is available in https://github.com/allenhaozhu/IR-Softmax

for both A-Softmax and AM-Softmax. Thus, our IR-Softmax version of Eq. (4) is:

$$J = -\left[\sum_{i=1}^{m} \log \frac{\exp(\|\mathbf{x}_i\|\psi(\theta_{y_i}))}{\exp(\|\mathbf{x}_i\|\psi(\theta_{y_i})) + \sum_{j\neq y_i}^{K} \exp(\mathbf{c}_j^T \mathbf{x}_i)}\right] \tag{5}$$

For A-Softmax,

$$\psi(\theta_{y_i}) = ((-1)^k \cos(m\theta_{y_i}) - 2k), \theta_{y_i} \in [\frac{k\pi}{m}, \frac{(k+1)\pi}{m}] \tag{6}$$

where $k \in [0, m-1]$ and $m \geq 1$ is an integer that controls the size of angular margin. For original A-Softmax, $\theta_{y_i}(0 \leq \theta_i \leq \pi)$ is the angle between $\mathbf{w}_i$ and $\mathbf{x}_i$. Note that, in our IR-Softmax, θ_{y_i} is *the angle between the* $\mathbf{c}_{y_i}$ *and* $\mathbf{x}_i$, where $\mathbf{c}_{y_i}$ is the center of class i. For AM-Softmax,

$$\psi(\theta_{y_i}) = \cos(\theta_{y_i} + \alpha) \tag{7}$$

$$\psi(\theta_{y_i}) = \cos(m_1\theta_{y_i} - m_2) - m_3 \tag{8}$$

where $\theta_{y_i}(0 \leq \theta_{y_i} \leq \pi)$ of the original AM-Softmax is the angle between $\mathbf{w}_i$ and $\mathbf{x}_i$. Note that θ_{y_i} of IR-Softmax is *the angle between the* $\mathbf{c}_{y_i}$ *and* $\mathbf{x}_i$.

Now we can summarise the difference between IR-Softmax and softmax (and its variants). First, the weight $\mathbf{w}_i$ of softmax is updated via back-propagation, however, $\mathbf{c}_i$ of IR-Softmax can be computed directly from Eq. (3). Second, the update of $\mathbf{w}_i$ depends on samples of class i and samples from other classes as shown in Eq. (2). In contrast, the update of $\mathbf{c}_i$ of IR-Softmax only depends on the samples from class i, effectively avoiding the influence of data imbalance.

In practice, it is impossible to use all samples to calculate the centers as shown in Eq.(3). We have tried three different updating strategies for feature centres. i.replacing the weight with an instance feature (which makes the proposed method like docFace [34]). ii.memory bank: estimate the centre with last few (in a fixed window) samples in the same class (without BP). iii.a loss function to estimate centres on the unit sphere for different classes. The disadvantage of the first solution is that an additional softmax is necessary in case the convergence is slightly slow and unstable. The second solution relieves the unstable issue but no improvement in performance. The third method is equal to adding a new term $\|\mathbf{c}_i - \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|_2}\|_2^2, s.t. \|\mathbf{c}_i\| = 1$ to Eq.4 and use the corresponding BP process to update feature centres (weights in Eq.5) rather than Eq.4. After that $\ell 2$ normalization is used on feature centres to make sure new c_i on the unit sphere. We select the third one in our experiments because it works better than other two approaches.

3.3 Relation to Metric Learning

N-pairs loss [38] enforces softmax cross-entropy loss among the pairwise similarity in the *mini-batch*.

$$E = \frac{-1}{|P|} \sum_{(i,j)\in P} \log \frac{\exp(S_{i,j})}{\exp(S_{i,j}) + \sum_{k:y_k\neq y_j} \exp(S_{i,k})} \tag{9}$$

where $S_{i,j} = f(\mathbf{x}_i, \mathbf{\Theta})^T f(\mathbf{x}_j, \mathbf{\Theta})$ represents the inner product between two embeddings. and $|P|$ indicates the number of positive pairs (i,j). Comparing Eq. (4) and (9), we can see our method can be viewed as a special form of N-pair loss. However, there are two main differences: (1) Unlike softmax embedded in N-pairs, we employ A-Softmax [18] and AM-Softmax [26] to improve the discriminability of features. (2) The size of *mini-batch* (where N-pair works) limits the number of negative samples. In practice, it is usually difficult to make mini-batch more than 256 due to the memory limitation of GPU. In contrast, our method alleviates the problem by caching historical features. The positive and negative samples are equal to the number of categories.

4 Experiments

In this section, we first describe the experimental settings. We then evaluate our method on two different tasks, face recognition (FR) and person re-identification (re-ID), against four different benchmarks. For FR, we use CASIA-WebFace [14] as training set and evaluate our method on LFW [27] and MegaFace [28]. For re-ID, we evaluate on the Market-1501 [24] and Duke [29] datasets.

4.1 Face Verification

All the faces and their landmarks are detected by MTCNN [39]. We use the detected 5 landmarks (two eyes, nose and two mouth corners) to perform similarity transformation. When the detection fails, we simply discard the image if it is in the training set, but use the provided landmarks if in the test set.

We use the publicly available training dataset CASIA-WebFace [14] (excluding the images of 59 identities appearing in testing sets [26]) to train our CNN models. CASIA-WebFace has 494,414 face images belonging to 10,575 (in fact, 10,516 after removing) different individuals. As shown in Fig. 1, CASIA-WebFace is an imbalance dataset. Some identities have very few images (e.g, only one image), while some have more than 300 images. These face images are horizontally flipped for data augmentation in the training process. Note that the number of samples in training set (0.49M) is relatively small compared to other private datasets used in DeepFace [2] (4M), VGGFace [40] (2M) and FaceNet [10] (200M). In the testing process, we extract the deep features from the output of the FC1 layer and do not employ any pre-processing (such as PCA and flipped features). The cosine distance between two features is applied. A nearest neighbor classifier and thresholding are used for face identification and verification, respectively.

To make fair comparison, we use two widely used CNN architectures for face recognition: 9-layer Light CNN [41] and 20-layer ResNet-20 [18]. Note that the faces are cropped to two different sizes (128x128 and 112x96) to fit the setting in [41] and [18] respectively. In the training process, our IR-Softmax is appended after the feature layer, i.e. the second last inner-product layer. The networks are trained in an end-to-end way.

For simplicity, we denote IR-Softmax (A) as our IR-Softmax instance derived from A-Softmax, and IR-Softmax (AM) from AM-Softmax in the whole experiment section.

LFW The LFW dataset [27] contains 13,233 images from 5,749 identities, with large variations in pose, expression and illumination. All the images are collected from the internet. We evaluate our methods on two protocols: (1) official protocol [27] and (2) BLUFR protocol [42]. For (1), LFW is divided into 10 predefined splits for cross validation. We follow the standard 'Unrestricted, Labeled Outside Data' protocol. Because the performance of face recognition is almost saturated on this protocol, researchers propose a more challenging BLUFR protocol [42]. For (2), BLUFR utilises all 13,233 images to evaluate the performance in the open-set setting. The Verification Rate (VR) at False Accepted Rate (FAR) 0.1% (VR@FAR=0.1%) and Detection and Identification rate (DIR) at FAR 1% (DIR@FAR=1%) are reported under BLUFR. It is noteworthy that not only three identities exist in both CASIA-Webface [14,26] and LFW [27]. We removed them according to [26] during training to build a complete open-set validation.

LFW Official Protocol, Light CNN As shown in Table 2, the performance is evaluated by six methods. The proposed IR-Softmax(A) and IR-Softmax(AM) greatly outperform their original versions (A-Softmax and AM-Softmax). Compared with the baseline (i.e. Softmax), IR-Softmax (A) improves the verification accuracy from 97.15% to 98.38%, and IR-Softmax (AM) from 97.15% to 98.63%.

LFW Official Protocol, ResNet-20 The evaluation results of ResNet-20 are listed in Table 1. Other state-of-the-art results of A-Softmax and AM-Softmax using ResNet-20 are also presented for comparison. Compared with the baseline (i.e. Softmax), IR-Softmax (A) loss improves the verification accuracy from 97.08% to 99.23% on LFW. From the results, we can see that the proposed methods IR-Softmax(A) and IR-Softmax(AM) can outperform the corresponding original versions.

BLUFR, Light CNN From Table 2, we can observe that the proposed method significantly outperforms the other methods [17,18]. Specifically, IR-Softmax(A) beats the softmax baseline (which we finetune our model from), and improves the VR@FAR=0.1% from 83.32% to 94.61%, while DIR@FAR=1% from 60.64% to 75.12%. Both versions of IR-Softmax are able to outperform their counterparts. It means that the proposed method can significantly enhance the discriminability of deeply learned features in the open-set protocol, demonstrating the effectiveness of the proposed method.

BLUFR, ResNet-20 Since ResNet-20 models [18,19] are also widely used for face recognition, we make comparisons based on ResNet-20 in Table 1. IR-Softmax with ResNet-20 keeps the similar superiority compared with other models in the BLUFR protocol of LFW. Note that our approach is better than range loss, which is proposed to solve the problem of data imbalance in face recognition. Though range loss uses a larger training set (MS-celeb [43]) and a deeper network (ResNet-50), our method still outperforms it with VR@FAR=0.1% from 93.72% to 97.08% (IR-Softmax(A)) or 98.09% (IR-Softmax(AM)) while DIR@FAR=1% from 71.11% to 81.52% (IR-Softmax(A)) or 85.00% (IR-Softmax(AM)).

MegaFace One of the most challenging datasets for face recognition is MegaFace [28]. The MegaFace dataset contains a gallery set and a probe set. The gallery set contains

Table 1. Performance on ResNet with various loss functions. CenterLoss, NormFace model and sphereface model are provided by authors. NormFace and CenterLoss use ResNet-28 like [16], another two methods use ResNet-20 [18].

loss function	LFW [27] 6000 pairs	BLUFR [42] VR@FAR=0.1%	BLUFR [42] DIR@FAR=1%	MegaFace [28] rank1@1e-6	MegaFace [28] VR@FAR=1e-6
Softmax	97.08%	78.26%	50.85%	45.26%	50.12%
CenterLoss [16]	99.00%	94.50%	65.46%	63.38%	75.68%
NormFace [19]	98.98%	96.16%	75.22%	65.03%	75.88%
A-Softmax [18]	99.08%	96.58%	79.97%	67.41%	78.19%
IR-Softmax(A)	99.23%	97.08%	81.52%	69.48%	80.32%
AM-Softmax [26]	98.98%	97.69%	84.82%	72.47%	84.44%
IR-Softmax(AM)	99.21%	98.09%	85.00%	75.28%	85.67%

Table 2. Performance on Lighten CNN with various loss functions. All Results are derived under the same settings used in [41].

loss function	LFW [27] 6000 pairs	BLUFR [42] VR@FAR=0.1%	BLUFR [42] DIR@FAR=1%	MegaFace [28] rank1@1e-6	MegaFace [28] VR@FAR=1e-6
Softmax	97.15%	83.32%	60.64%	47.31%	54.86%
Large-Margin [17]	98.35%	91.62%	64.76%	59.03%	70.57%
A-Softmax [18]	98.20%	91.16%	66.55%	54.87%	60.75%
IR-Softmax(A)	98.38%	94.61%	75.12%	64.71%	75.94%
AM-Softmax [26]	98.58%	94.67%	72.80%	65.33%	78.76%
IR-Softmax(AM)	98.63%	95.36%	79.92%	66.71%	78.83%

more than 1 million images from 690K identities; The probe set consists of two existing datasets: Facescrub [44] and FGNet. MegaFace has multiple testing scenarios including identification, verification and pose-invariance under two protocols i.e. large or small training sets. The training set is considered small if it is less than 0.5M. We evaluate our IR-Softmax under the small training set protocol.

Lighten-CNN Table 2 shows that our IR-Softmax(A) outperforms A-Softmax result by a margin (almost 10% for rank-1 identification rate and 15% for VR at 1e-6 FAR) on the small training dataset protocol while IR-Softmax(AM) outperforms AM-Softmax result by a margin (7% for rank-1 identification rate and 1.4% for VR at 1e-6 FAR). Compared to the softmax baseline, our method performs significantly better: 15% from IR-Softmax (A) and 19% IR-Softmax(AM) for identification, 21% from IR-Softmax (A) and 24% from IR-Softmax (AM) for verification.

ResNet-20 Table 1 shows that our IR-Softmax (A) outperforms A-Softmax result by a margin (almost 2% for rank-1 identification rate and 2% for VR at 1e-6 FAR) on the small training dataset protocol while IR-Softmax(AM) outperforms AM-Softmax result by a margin (almost 3% for rank-1 identification rate and 1.2 % for VR at 1e-6 FAR). Compared to the softmax baseline, our method performs significantly better: 24% from IR-Softmax (A) and 30% from IR-Softmax (AM) for identification, 30% from IR-Softmax (A) and 35% IR-Softmax (AM) for verification.

Note that the performance of any testing methods on Megaface is intimately linked to the quality of face alignment. Thus we do not compared with other methods with different alignments. The results in Table 2 are therefore computed under the same setting of face alignment and are directly comparable. These results demonstrate that our IR-Softmax is well designed for open-set face recognition especially when the training set is imbalanced. One can also see that, smaller intra-class distance is not the only important issue for learning features, but larger and evenly inter-class angular margin can significantly improve face recognition performance.

4.2 Person Re-identification

For the evaluation of re-ID, we focus on two well-known re-ID databases: Market-1501 [24] and DUKE [29] datasets. As shown in Fig. 1, we demonstrate the distribution of market-1501 database. Although there are no identities with more than 100 images like WebFace dataset, the number of images per person ranges from 5 to 80. The DUKE [29] also has the similar imbalance pattern. We use the standard evaluation metrics for both datasets, namely the mean average precision score (mAP) and the cumulative matching curve (CMC) at rank-1. We follow common practice by using random crops and random horizontal flips during training. Specifically, we resize all images to 256×128, of which we take random crops of size 224×112. Many methods for re-ID rely on pre-trained models (e.g. ResNet). Indeed, these models usually lead to impressive results. However, pre-trained models reduce the flexibility to make task-specific changes in a network. For example, some application scenarios need compact models rather than large ones pre-trained on Imagenet. Our method clearly suggests that it is also possible to learn deep models from scratch and achieve state-of-the-art performance. We use a Lighten CNN [41] based on the ResNet Architecture, which is faster than the current ResNet-50 used by many works [45]. Compared with other methods, we do not use the corresponding pretrained models in ImageNet for finetuning. Thus we use the softmax to train a baseline model with the re-ID dataset directly. And other methods (e.g. large-margin and the proposed method) employ the baseline model as the pre-trained model and finetune this model further.

Market-1501 The Market-1501 dataset contains 1,501 identities, 19,732 gallery images and 12,936 training images captured by 6 cameras. All the bounding boxes are generated by the DPM detector [46]. The dataset uses both single and multi-query evaluation, we report the results for both. Table 3 compares our IR-Softmax (A) to other approaches. For Market-1501, the improvements achieved by IR-Softmax are significant: (1) Compared with softmax, the Rank-1 accuracy rises from 81.47% to 91.87%, and the mAP from 57.42% to 76.72% in the setting of single query; (2) In the setting of multi query, the Rank-1 accuracy rises from 86.40% to 94.33%, and the mAP from 65.97% to 82.22%. IR-Softmax (A) significantly outperforms not only the softmax baseline but also other state-of-the-art methods [13].

DukeMTMC-reID The DukeMTMC-reID dataset is collected via 8 cameras and used for cross-camera tracking (handover). Table 4 compares our IR-Softmax to other approaches. For DukeMTMC-reID, IR-Softmax(A) works much better than softmax: the

Table 3. Comparison with the state-of-the-art methods on the Market-1501 dataset. The rank-1 accuracy and mAP on single and multiple query are reported respectively.

Method	Single Query		Multi. Query	
	rank-1	mAP	rank-1	mAP
BoW +KISSME [24]	44.42	20.76	-	-
MR CNN [47]	45.58	26.11	56.59	32.26
DSN [48]	55.43	29.87	71.56	46.03
Gate Reid [49]	65.88	39.55	76.04	48.45
SOMAnet [50]	73.87	47.89	81.29	56.98
DeepTransfer [30]	83.70	65.50	**89.60**	73.80
Basel+LSRO [13]	**83.97**	**66.07**	88.42	**76.10**
SVDNet [45]	82.30	62.10	-	-
Softmax	81.47	57.42	86.40	65.97
Large-margin [17]	90.08	72.22	92.75	78.79
IR-Softmax(A)	**91.87**	**76.72**	**94.33**	**82.88**

Table 4. Comparison with state-of-the-art methods on DukeMTMC-reID. Rank-1 accuracy and mAP are reported.

Method	Rank-1 (%)	mAP (%)
BoW + KISSME [24]	25.13	12.17
LOMO + XQDA [51]	30.75	17.04
Basel + LSRO [13]	67.68	47.13
ACRN [52]	72.58	51.96
PAN [53]	71.59	51.51
SVDNet [45]	**76.70**	**56.80**
Softmax	61.98	41.17
Large-margin [17]	75.58	56.25
IR-Softmax(A)	**76.84**	**57.47**

Rank-1 accuracy: 76.84% vs 61.98%, and the mAP 57.47% vs 41.17%. Beyond that, the imbalance robust softmax also outperforms other state-of-the-art methods[45].

5 Conclusion

In this paper, we investigated thoroughly the potential effects of data imbalance on the deep embedding learning and proposed a new framework, Imbalance Robust Softmax (IR-Softmax). IR-Softmax can simultaneously solve the open-set problem and reduce the influence of data imbalance. Extensive experiments on FR and re-ID are conducted, and the results show the effectiveness of IR-Softmax. In Future work, we plan to extend this framework to more softmax based methods and other applications like few-shot learning.

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012)
2. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229 (2013)
3. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR. (2015)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016)
5. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: ICCV. (2015)
6. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR. (2015)
7. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: CVPR. (2014)
8. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: CVPR. (2014)
9. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: NIPS. (2014)
10. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR. (2015)
11. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 152–159
12. Zheng, L., Zhang, H., Sun, S., Chandraker, M., Tian, Q.: Person re-identification in the wild. CoRR **abs/1604.02531** (2016)
13. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: ICCV. (2017)
14. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. CoRR (2014)
15. Sun, Y., Wang, X., Tang, X.: Deeply learned face representations are sparse, selective, and robust. In: CVPR. (2015)
16. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: ECCV. (2016)
17. Liu, W., Wen, Y., Yu, Z., Yang, M.: Large-margin softmax loss for convolutional neural networks. In: ICML. (2016)
18. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. (2017)
19. Wang, F., Xiang, X., Cheng, J., Yuille, A.L.: Normface: L_2 hypersphere embedding for face verification. In: ACM MM. (2017)
20. Ranjan, R., Castillo, C.D., Chellappa, R.: L2-constrained Softmax Loss for Discriminative Face Verification. ArXiv e-prints (2017)
21. LeCun, Y., Cortes, C., Burges, C.J.: The mnist database of handwritten digits (1998)
22. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. (2009)
23. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. (2009)
24. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: CVPR. (2015)

25. Zhang, X., Fang, Z., Wen, Y., Li, Z., Qiao, Y.: Range loss for deep face recognition with long-tail. CoRR **abs/1611.08976** (2016)
26. Wang, F., Liu, W., Liu, H., Cheng, J.: Additive Margin Softmax for Face Verification. ArXiv e-prints (2018)
27. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst (2007)
28. Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The megaface benchmark: 1 million faces for recognition at scale. In: CVPR. (2016)
29. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: ECCV. (2016)
30. Geng, M., Wang, Y., Xiang, T., Tian, Y.: Deep transfer learning for person re-identification. CoRR **abs/1611.05244** (2016)
31. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 5265–5274
32. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 4690–4699
33. Ouyang, W., Wang, X., Zhang, C., Yang, X.: Factors in finetuning deep model for object detection with long-tail distribution. In: CVPR. (2016)
34. Shi, Y., Jain, A.K.: Docface+: Id document to selfie matching. IEEE Transactions on Biometrics, Behavior, and Identity Science **1** (2019) 56–67
35. Zhong, Y., Deng, W., Wang, M., Hu, J., Peng, J., Tao, X., Huang, Y.: Unequal-training for deep face recognition with long-tailed noisy data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 7812–7821
36. Khan, S., Hayat, M., Zamir, S.W., Shen, J., Shao, L.: Striking the right balance with uncertainty. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 103–112
37. Yin, X., Yu, X., Sohn, K., Liu, X., Chandraker, M.: Feature transfer learning for face recognition with under-represented data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 5704–5713
38. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. In: NIPS. (2016)
39. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. SPL (2016)
40. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: BMVC. (2015)
41. Wu, X., He, R., Sun, Z., Tan, T.: A light CNN for deep face representation with noisy labels. CoRR **abs/1511.02683** (2015)
42. Liao, S., Lei, Z., Yi, D., Li, S.Z.: A benchmark study of large-scale unconstrained face recognition. In: ICB. (2014)
43. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: MS-Celeb-1M: A dataset and benchmark for large scale face recognition. In: European Conference on Computer Vision, Springer (2016)
44. Ng, H.W., Winkler, S.: A data-driven approach to cleaning large face datasets. In: ICIP. (2014)
45. Sun, Y., Zheng, L., Deng, W., Wang, S.: Svdnet for pedestrian retrieval. CoRR **abs/1703.05693** (2017)
46. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR. (2008)
47. Ustinova, E., Ganin, Y., Lempitsky, V.S.: Multiregion bilinear convolutional neural networks for person re-identification. CoRR **abs/1512.05300** (2015)

48. Zhang, L., Xiang, T., Gong, S.: Learning a discriminative null space for person re-identification. In: CVPR. (2016)
49. Varior, R.R., Haloi, M., Wang, G.: Gated siamese convolutional neural network architecture for human re-identification. In: ECCV. (2016)
50. Barbosa, I.B., Cristani, M., Caputo, B., Rognhaugen, A., Theoharis, T.: Looking beyond appearances: Synthetic training data for deep cnns in re-identification. CoRR **abs/1701.03153** (2017)
51. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: CVPR. (2015)
52. Schumann, A., Stiefelhagen, R.: Person re-identification by deep learning attribute-complementary information. In: CVPRW. (2017)
53. Zheng, Z., Zheng, L., Yang, Y.: Pedestrian alignment network for large-scale person re-identification. CoRR **abs/1707.00408** (2017)