Comparative Analysis of Different Loss Functions for Deep Face Recognition

Aman Pathak
Department of Computer Science & Engineering
Medi-Caps University (MU)
Indore Madhya Pradesh India
a.pathak4892@gmail.com

Ritu Maheshwari
Department of Computer Science & Engineering
Medi-Caps University (MU)
Indore Madhya Pradesh India
ritu.nitttr@gmail.com

ABSTRACT

Face Recognition has been one the fastest emerging field in the last ten years. Convolutional neural network (CNN) or Deep convolutional neural network (DCNN) have significantly developed the extraordinary state-of-the-art solution for Face Recognition. This promising development results from the enhanced learning and representation of the discriminative features. The learning depends extensively on the loss function employed in the model. The loss function plays a vital role in the training of CNN and its job is to evaluate the performance of the model, i.e. bad performance results in a huge loss and vice versa. The gradients of this loss function are further used in the back propagation of errors which in turn enables the model to improve its learning from the given data. The objective of this paper is to have a comparative analysis of different loss functions available for the Deep Face Recognition.

CCS CONCEPTS

• Computing methodologies~ Artificial intelligence~ Computer vision~ Computer vision representations~ Image representations

KEYWORDS

Face Recognition, Convolutional neural network, Discriminative features, Loss function.

ACM Reference format:

Aman Pathak and Ritu Maheshwari. 2019. Comparative Analysis of Different Loss Functions for Deep Face Recognition. In *Proceedings of 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence (ACAI'19). Sanya, China, 8 pages.* https://doi.org/10.1145/3377713.3377779

ACAI '19, December 20–22, 2019, Sanya, China

ACM ISBN 978-1-4503-7261-9/19/12…$15.00
https://doi.org/10.1145/3377713.3377779

1 Introduction

Over the past few years, the remarkable progress on classification tasks such as object recognition [1][2][3], scene recognition [4][5], and action recognition [6][7] is exclusively due to deep convolutional neural networks. The robust layered learning structure coupled together with convolution and pooling layers, that aims to learn low-level features as well as high-level features, generates a strong visual representation capability of CNN. The great success of many vision challenges is attributed to the availability of plentiful datasets and robust loss functions. Deep CNNs, one of the most influential factors that have been under continuous research and enhancement, is also responsible for the evolutionary facial verification, clustering, identification and recognition task.

Under supervised learning, Face recognition is one of the most practically exploited application. Systems such as face unlock for mobile devices and biometric attendance system used in the industries, both leverage from the fundamentals of face recognition. Innovations such as finding missing persons or pets, guiding the blinds, preventing retail crime, safeguarding the law enforcement [8] benefit from the recent developments on face representation. Face identification is responsible for classifying an input face to a specific identity. Face verification determines whether the input face resembles a group of other faces of the same identity. The process of face recognition can be understood as face identification followed by face verification. The state-of-the-art method for *face representation* is deep convolution neural network embedding. A pose normalization step is often employed and the model maps a *face* to an embedding feature vector. The discriminative characteristics come from the point that in the feature space, the feature vectors of the same person have a negligible distance while that of different identity have considerable distance among themselves.

There are primarily three attributes of a face recognition system and these change the method of learning the deep convolutional network embedding [9]. First and foremost is the dataset chosen for the task. The dataset acts as a fuel for the training of CNN and it directly influences the accuracy of the model. Currently, many publicly available large scale training datasets such as Labeled Faces in the Wild (LFW) [10], YouTube Faces (YTF) [11], Cross-Age Celebrity Dataset (CACD) [12],

Age database (AgeDB) [13] and MegaFace [14], MS-Celeb-1M [15], VGGFace2 [16], CASIA Webface [17] are considered for the face recognition task. The number of images provided in these datasets typically range from thousands to a few million. A table providing a brief statistics of the above-mentioned datasets is presented in Table 1.

The second attribute is the network architecture. The high capacity and scalable deep convolutional neural networks such as ResNet [18][19] and Inception-ResNet [20] are the state-of-the-art architecture for image-classification. Recently, many complex architectures specifically for the task of face recognition have been proposed. CNN's such as DeepID3 [21], VGGFace [22], Baidu [23], DeepFace [24], FaceNet [25] and so on, have successfully defeated humans on some levels.

Table 1. Brief statistics for publically available large scale Face datasets

Dataset	*Number of Identities*	*Number of Images*	*Year*
LFW	5,749	13,233	2007
YTF	1,595	3425 (videos)	2011
CACD	2,000	163,446	2014
AgeDB	568	16,488	2017
MegaFace	3,311,471	4,753,320	2016
MS-Celeb-1M	100,000	8,200,000	2016
VGGFace2	9131	3.3 M	2018
CASIA Webface	494,114	10,575	2014

The loss functions can be divided into two types [26]: The Euclidean distance based loss function and the Cosine distance based loss function. The two differ only in terms of how the maximum intra-class compactness and inter-class separability is achieved. The Euclidean distance based loss functions aim at improving the discriminative ability either by minimizing the intra-class variation and by maximizing the inter-class distance on Euclidean space. The loss functions including in this category are Triplet Loss [25], Marginal Loss [27], Center Loss [28], Range Loss [29]. The Cosine distance based loss function firstly transforms the output of the last layer of CNN from Euclidean distance to Cosine distance, followed by the addition of an angular constraint term, that maximizes the inter-class angular margin distance. The loss function including in this category are Soft-Margin Softmax [30], Large-Margin Softmax [31], Angular Margin Softmax [32], ArcFace [33] and many more.

The paper is divided into the following sections. Section II provides a detailed understanding of different categories of loss functions for Deep Face Recognition. Section III is Literature Review. The section describes a broad summary of the researches performed and the achieved results. Section IV provides inferences drawn from each loss function. Section V presents a comparative analysis of the loss functions. This is based on the scored accuracies, solving methodologies, limitations, required computational power. Section VI concludes the paper with the findings of the analysis.

2 Categories of Loss Functions for Deep Face Recognition

Alessandro Calefati et. al [9] theorized that loss functions for Deep Face Recognition can be broadly classified into four major categories (1) Deep Metric Learning Approaches (2) Joint Supervision with Softmax (3) Imbalanced Classes-Aware loss functions and (4) Angle-Based loss function.

2.1 Deep Metric Learning Approaches

This approach aims at learning a similarity metric from the given training data. A similarity metric is a method by which a neural network represents a similarity between two things. Common measures used are Euclidean Distance, Cosine Similarity, Pearson Coefficient, and so on. The obtained similarity metric is then used on a new face to determine a match from the previously learned face. This is done by finding a function that maps input face pattern such that simple distance approximates the semantic distance.

2.2 Joint Supervision with Softmax

Softmax Loss is undoubtedly the most popular loss function on CNN. Softmax loss is actually derived from the combination of ***Cross-Entropy Loss*** (also called Multinomial Logistic Loss) and the ***Softmax Function.***

$$S(x_j) = \frac{e^{x_i}}{\sum_i e^{x_i}} \; ; \; \mathcal{L}_{CE} = -\frac{1}{N}\left(\sum_{i=1}^{N} y_i \log(\hat{y}_i)\right) \tag{1}$$

$$\mathcal{L}_S = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\left(S(x_j)\right) = -\frac{1}{N}\sum_i \log\left(\frac{e^{x_i}}{\sum_i e^{x_i}}\right) \tag{2}$$

The output of the last layer of the neural network is called logit scores. These are raw scores generated before activation. The softmax function activates these logits and outputs a vector that measures the probability distribution of the predictions. Then Cross-Entropy measures the consistency of the probabilities and the true vector of the input which generates a loss. For the backpropagation of this loss, the gradients of cross-entropy loss are obtained and subsequently, the parameters are tuned in the direction of loss minimization.

The Softmax Loss in (2) does not explicitly emphasize the discriminative learning of the features i.e. it takes no measurable steps to maximize intra-class compactness and inter-class separability. Hence, the Softmax Loss function performs poorly in face recognition tasks.

2.3 Class Imbalance-Aware Loss Function

Class Imbalanced dataset or *long-tailed* dataset is the kind of dataset where just a handful of classes appear more frequently while most of the other classes occur rarely. This property is noticeably visible in the MS-Celeb-1M [15] and MegaFace [14] dataset. Training a neural network on such a long-tailed dataset would result in a poor biased learning and inaccurate

representation of the long-tailed classes. As a result, the intra-class compactness would be abnormally dispersed and loose, at the same time the inter-class separability would be compromised. The flaw in such a dataset is pretty discernible. However, replicating the existing data and filtering out such classes is generally considered, but that results in improper learning and omission of many of such classes from our study.

2.4 Angle-based Loss function

Angle based Loss function are fundamentally based on introducing a *decision boundary* in the angular space. These loss function leverages the modified softmax loss in which the ***logit*** (the raw score obtained by last layer of neural network) is transformed from $W_i^T x_i$ to $\|W_i\|\|x_i\|\cos\theta_j$ where $W_i \in \mathbb{R}^d$ denotes the j^{th} column of weight $W \in \mathbb{R}^{dxn}$, $x_i \in \mathbb{R}^d$ denotes the feature of the i^{th} sample and θ_j is the angle between the W_j and x_i.

$$\mathcal{L}_s = -\log\left(\frac{e^{w_{y_i}^T x_i}}{\sum_j^K e^{w_j^T x_i}}\right) = -\log\left(\frac{e^{\|w_{y_i}\|\|x_i\|\cos(\theta_{y_i})}}{\sum_j^K e^{\|w_j\|\|x_i\|\cos(\theta_j)}}\right) \quad (3)$$

3 Literature Review

Florian Schroff et al. [2015] [25] proposed FaceNet: A Unified Embedding for Face Recognition and Clustering. The system proposes a Triplet Loss function that aims at maximizing the intra-class compactness and inter-class separability by training on triplets of input images. The exceptional accuracy of 98.87% on the LFW [10] dataset using the fixed center crop method achieved the state-of-the-art solution.

Weiyang Liu et al. [2016] [31] proposed Large-Margin Softmax Loss for Convolutional Neural Networks. It achieves a flexible classification angle margin (m) between the classes and simultaneously avoids overfitting of the network. Experiments on MNIST [34], CIFAR10 [35], CIFAR100 [35] reveal the least recognition error rate of 0.31%, 7.58%, 29.53% respectively. The outstanding accuracy of 98.71% on LFW shows the clear success of Large-Margin Softmax Loss.

Yandong Wen et al. [2016] [28] proposed Center Loss for Deep Face Recognition. In joint supervision with Softmax Loss, Center Loss achieves remarkable intra-class compactness. The CNNs trained with Center Loss can be easily optimized by standard SGD [36]. Experiments on LFW and YTF dataset using the Caffe [40] library suggests a remarkable performance of the Center Loss with an accuracy of 99.28% and 94.9%.

Xiao Zhang et al. [2016] [29] proposed Range Loss for the long-tailed datasets. Range Loss aims at learning discriminative features within mini-batch when facing an imbalanced class dataset. A comparison of accuracies with the Contrastive Loss [37] and Triplet Loss on the VGG Network [38] proves that Range Loss optimizes the poor class more efficiently. Performances on the renowned LFW [10] and YTF [11] datasets and performances on other CNN structures like DeepId-2+ [39], Baidu [23], fairly indicates the success of Range Loss.

Jiankang Deng *et al*. [2017] [27] proposed Marginal Loss for the Deep Face Recognition. In joint supervision with Softmax Loss, the Marginal Loss introduces a threshold margin distance θ to effectively maximise the intra-class compactness and inter-class separability. Experiments on several large scale face datasets such as Cross-Age Celebrity Dataset (CACD) [12], Age database (AgeDB) [13] reveal the exceptional performance.

Xuezhi Liang et al. [2017] [30] proposed Soft-Margin Softmax Loss for Deep Classification. The loss function introduces an angular margin (m), a non-negative real number that maximizes the intra-class compactness and inter-class separability. Using the Caffe [40] library along with the MNIST [34], CIFAR10 [35], CIFAR100 [35] dataset, the experiment concludes with Soft-Margin Loss outperforming Softmax Loss function.

Hao Wang et al. [2018] [32] proposed CosFace: Large Margin Cosine Loss for Face Recognition. CosFace proposes a novel loss function, Large-Margin Cosine Loss (LMCL), which increases the angular margin of the learned features by introducing the cosine margin. Using the Caffe library, LMCL achieved state-of-the-art accuracy if 99.33% on LFW and 96.10% on YTD [11] datasets.

Alessandro Calefati et al. [2018] [9] proposed Git loss with joint supervision with Softmax Loss and Center Loss. A novel function is added to the supervision signals that specially target at achieving inter-class separability. The experiment includes the LFW [10], YTF [11], dataset associated with the Inception ResNet-V1 [41] network architecture. Git Loss overtakes Softmax Loss with a significant margin of 0.9% and 1.90% in the LFW and YTF dataset respectively.

Xin Wei et al. [2018] [26] proposed Minimum Margin Loss for Deep Face Recognition. It specifies a Minimum Margin that targets in optimizing the inter-class separability. In joint supervision with softmax Loss and Center Loss, Minimum Margin displays remarkable performance on the LFW and YTF dataset trained on Inception-Res-Net-V1 with an accuracy of 99.63% and 95.5%.

Jiankang Deng et al. [2019] [33] proposed ArcFace, an Additive Angular Margin Loss function for Deep Face Recognition. The geometric interpretation of the proposed Angular Margin shares the exact correspondence to the geodesic distance. The extraordinary and state-of-the-art performance on LFW, YTD, MegaFace [14] outperforms all existing loss functions with an accuracy of 99.83%, 98.02%, and 97.91%.

4 Inferences Drawn

4.1 Triplet Loss

Triplet Loss became famous when FaceNet [25] (developed by Google) used it as their loss function and won the 2014 ImageNet challenge. It uses Euclidean distance to measure the distance between the deeply learned feature in the feature space. As the name suggests the triplet loss uses three images at a time as an input, a positive sample, a negative sample, and an anchor sample. Although the positive sample and anchor sample belongs to the

same identity, the feature vector of the anchor belongs closer to that of the negative identity in the feature space.

$$\mathcal{L} = \sum_i^N \left[\left\| f(x_i^a) - f(x_i^p) \right\|_2^2 - \| f(x_i^a) - f(x_i^n) \|_2^2 + \alpha \right]_+ \quad (4)$$

where x_i^a, x_i^p, x_i^n denote the images of anchor sample, positive sample and negative sample, respectively.

In order to improve the discriminative abilities, the Euclidean distance between the face embeddings of the same identity must be minimized and that of distinct identities must be maximized. Therefore, during training, the Triplet Loss minimizes the distance of anchor-positive pair and maximizes the distance of anchor-negative pair. This optimizes the intra-class compactness and inter-class separability. The Triplet Loss forces a margin between the two identities and thereby allowing the discriminability from other identities.

4.2 Large-Margin Softmax Loss

Softmax Loss is the most popular loss function used in the classification tasks, yet, it does not emphasize particularly about the discriminative learning of the inputs. Inspired by the Contrastive Loss [37] and Triplet Loss [25], the Large-Margin Softmax [31] introduces an angular decision margin that encourages better intra-class compactness and inter-class separability. The Softmax Loss function undergoes a margin constraint which results in a Large Margin Softmax Loss function.

$$\mathcal{L}_i = -\log\left(\frac{e^{\|W_{y_i}\|\|x_i\|\psi(\theta_{y_i})}}{e^{\|W_{y_i}\|\|x_i\|\psi(\theta_{y_i})} + \sum_{j \neq y_i} e^{\|W_j\|\|x_i\|\cos(\theta_j)}} \right) \quad (5)$$

$$\psi(\theta) = \begin{cases} \cos(m\theta), & 0 \leq \theta < \pi/m \\ \mathcal{D}(\theta), & \pi/m \leq \theta < \pi \end{cases} \quad (6)$$

For a sample x of class 1, the inequality holds $\|W_1\|\|x\|\cos(\theta_1) > \|W_2\|\|x\|\cos(\theta_2)$. Instead of comparing with $\|W_1\|\|x\|\cos(\theta_1)$ the parameter m (a positive integer) forces a rigorous classification inequality, $\|W_1\|\|x\|\cos(m\theta_1) > \|W_2\|\|x\|\cos(\theta_2)$. As we increase the value of m, we obtain a stricter and discriminative neural network. It is worth noting that forward and backward propagation becomes a lot complex on the introduction of the angular margin. This is because the Taylor series expansion is possible only if m is positive.

4.3 Center Loss

In order to emphasize the discriminatory power of the features, the Center Loss [28] particularly aims at enhancing the intra-class compactness.

$$\mathcal{L}_c = \frac{1}{2} \sum_{i=1}^{m} \left\| x_i - c_{y_i} \right\|_2^2 \quad (7)$$

$$\mathcal{L} = \mathcal{L}_S + \lambda \mathcal{L}_c \quad (8)$$

$$\mathcal{L} = -\sum_{i=1}^{m} \log \frac{e^{w_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{w_j^T x_i + b_j}} + \frac{\lambda}{2} \sum_{i=1}^{m} \left\| x_i - c_{y_i} \right\|_2^2 \quad (9)$$

where c_{y_i} is the y_ith class center of deep features.

Firstly, class centers for all the classes are randomized. They are repeatedly updated by averaging the deep features that train the network. While training the network the loss function penalizes in proportion to the distance of a feature from its class center. That means greater the distance from the class center greater the penalty incurred. As a result, the center loss, targets in maximizing the intra-class compactness of the deeply learned features.

4.4 Range Loss

The Range Loss [29] significantly aims at refining the learning capability of the network in cases of long-tailed dataset. This function enhances learning by generalizing all the long-tailed classes. The keyword *range* represents the maximum Euclidean distance between the two farthest features in a class. This loss treats the intra-class compactness and inter-class separability individually and aims at optimizing both of them.

$$\mathcal{L}_R = \alpha \mathcal{L}_{R_{intra}} + \beta \mathcal{L}_{R_{inter}} \quad (10)$$

$$\mathcal{L}_R = \alpha \sum_{i \subseteq I} \frac{k}{\sum_{j=1}^{k} \frac{1}{\mathcal{D}_j}} + \beta \, max\left(m - \left\| \bar{x}_Q - \bar{x}_R \right\|_2^2, 0 \right) \quad (11)$$

$$\mathcal{L} = \mathcal{L}_S + \lambda \mathcal{L}_R \quad (12)$$

The intra-class loss $\mathcal{L}_{R_{intra}}$, targets at minimizing the dispersed features and this is achieved by calculating k ($k = 2$ performs well) greatest range's harmonic mean. The inter-class loss $\mathcal{L}_{R_{inter}}$, targets at the maximization of the distance between the two nearest class centers. This is repeatedly updated over mini-batches until convergence.

4.5 Marginal Loss

Marginal Loss function [27] as the name suggest, forces a rigorous margin distance threshold in order to increase the intra-class compactness and inter-class separability. The concept lies in considering the distance between the feature and the farthest feature of the same class to be less than threshold θ, which ensures the compactness. Similarly considering the distance between the feature and nearest feature of a different class to be more than the threshold θ, which ensures the separability.

$$\mathcal{L}_m = \frac{1}{m^2 - m} \sum_{i,j,i \neq j}^{m} \left(\xi - y_{ij} \left(\theta - \left\| \frac{x_i}{\|x_i\|} - \frac{x_j}{\|x_j\|} \right\|_2^2 \right) \right)_+ \quad (13)$$

$$\mathcal{L} = \mathcal{L}_s + \lambda \mathcal{L}_m \quad (14)$$

where $y_{ij} \in \{ \pm 1 \}$, x_i and x_j are two input feature vectors.
The loss function considers all the sample pairs in the batch and forces them under the threshold θ. Thus by enforcing a rigorous threshold on Marginal Loss successfully achieves intra-class compactness and inter-class separability.

4.6 Soft-Margin Loss

The Soft-Margin Loss function [30] is an advancement to the Large-margin Softmax Loss. The angular margin imposed by Large-Margin Softmax m, eliminates numerous possibilities of different margins. This is because to expand $\cos(m\theta_1)$ into the

Taylor series, m is bounded to be a positive integer. This restriction also causes the propagation on the forward and backward side to be heavily intensive. To overcome such shortcomings, a soft margin which is a positive real number term $\overline{m}$ is introduced.

$$\mathcal{L}_i = -\log\left(\frac{e^{w_{y_i}^T x_i - \overline{m}}}{e^{w_{y_i}^T x_i - \overline{m}} + \sum_{j \neq y_i} e^{w_j^T x_i}}\right) \quad (15)$$

The changes employ that $W_1^T x$ changes to $W_1^T x - \overline{m}$ and thus the inequality for a sample x in class 1 becomes $W_1^T x > W_1^T x - \overline{m}$. Soft-Margin Loss function is easy to implement and provides various angular margin decision boundaries thus maximizing the intra-class compactness and inter-class separability. Unlike Large Margin loss, the soft margin $\overline{m}$, does not make the forward and backward propagation intensive. Thus it includes the merits of both the Softmax Loss and the Large-Margin loss function.

4.7 Large-Margin Cosine Loss

The face-embeddings learned from the Softmax Loss exhibit an ingrained angular distribution, and the Large-Margin Cosine Loss [32] aims at leveraging the consistency of the angular margin and the cosine of the angle [42]. After the L_2 normalization of features x and weight W we obtain the scaling parameter s, and hence the predictive probability lies only on the cosine of angle $(\cos\theta)$. To this a cosine margin term m is introduced to maximize the discriminative abilities of the deeply learned feature.

$$\mathcal{L}_{LMCL} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{s\left(\cos\left(\theta_{y_{i,i}}\right) - m\right)}}{e^{s\left(\cos\left(\theta_{y_{i,i}}\right) - m\right)} + \sum_{j \neq y_i}^{n} e^{s\cos\left(\theta_{j,i}\right)}} \quad (16)$$

To guarantee a huge hyperspace for feature learning with a large angular margin, the value of the scaling parameter s which translates as the radius of the hypersphere, must be sufficiently big. A small value lead to poor convergence. For a sample x of class 1, the CosFace holds the inequality $\cos(\theta_1) - m > \cos(\theta_2)$. The optimised value of the hyper-parameter m, ensures better intra-class compactness and inter-class separability.

4.8 Git loss

The Center loss doesn't emphasize much about the inter-class separability. Inspired by this motive, Git Loss [9], a loss function inspired by a famous version control software Git, was introduced. It works in association with the Softmax Loss and the Center Loss. In this, a new function that especially enhances the inter-class separability is added to the loss function equation. This function penalizes in inverse proportion to the distance between a feature and all $c-1$ class centers. The greater the distance the lesser the penalty. As a result, it pushes the feature away thus increasing the separability.

$$\mathcal{L}_G = \frac{1}{2}\sum_{i=1}^{m} \frac{1}{1 + \left\|x_i - C_{y_i}\right\|_2^2} \quad (17)$$

$$\mathcal{L} = \mathcal{L}_S + \lambda_C \mathcal{L}_C + \lambda_G \mathcal{L}_G \quad (18)$$

Concluding, Git loss includes the merit of the Center Loss $\mathcal{L}_C$ and Softmax loss $\mathcal{L}_S$ to create a loss function that improves the discriminative power of the feature by *pushing* and *pulling*.

4.9 Minimum Margin Loss

The Minimum Margin Loss [26] uses Euclidean distance for its loss computation and is an advancement to the Marginal loss. The motivation for the Minimum Margin Loss comes from the strict enforcement of the distance threshold θ in the Marginal Loss. It realizes the possibility of a practical case where the distance between the two farthest samples in the class is greater than the two samples of a different class. In this case, the loss value would spike and would make the training procedure hard to converge. The total loss $\mathcal{L}$ is given as:

$$\mathcal{L} = \mathcal{L}_S + \alpha\mathcal{L}_c + \beta\mathcal{L}_M \quad (19)$$

$$\mathcal{L}_M = \sum_{i,j=1}^{K} max\left(\left\|c_i - c_j\right\|_2^2 - \mathcal{M}, 0\right) \quad (20)$$

where α and β are the hyper-parameters $\mathcal{L}_c$denotes the Center loss (5), $\mathcal{L}_M$ denotes the Marginal Loss (10) , c_i and c_j are the class centers of ith and jth classes respectively.

To tackle this problem, a novel function $(\mathcal{L}_M)$ is added which utilizes the center values obtained from center loss. he f and applies a penalty if the corresponding distance between the two class centers is greater than margin threshold $(\mathcal{M})$. Therefore, in joint supervision with Softmax loss and Center Loss [28], it ensures maximized intra-class compactness and with Minimum Margin Loss, it achieves inter-class separability.

4.10 Additive Angular Margin Loss

The ArcFace or Additive Angular Margin Loss [33] is a specialized loss function developed for the Deep Face recognition task. ArcFace encourages discriminative learning by introducing an additive angular margin. The correspondence between the angle and arc in a normalized hypersphere is used to optimize the geodesic distance margin. By applying an arc-cosine function to obtain the angle θ between W and x we add the additive angular margin m to it.

$$\mathcal{L}_{ArcFace} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{s\left(\cos\left(\theta_{y_i} + m\right)\right)}}{e^{s\left(\cos\left(\theta_{y_i} + m\right)\right)} + \sum_{j=1, j \neq y_i}^{n} e^{s\cos\theta_j}} \quad (21)$$

During training, the feature vectors undergo normalization and are re-scaled to a parameter s, thus eliminating other parameters. As a result, making the prediction dependent only on the angle θ. ArcFace is easily to implement with several lines of code and

during the training, it just adds a negligible computational complexity.

5 Comparative Analysis

Undoubtedly, **Triplet Loss** achieves remarkable accuracy, but the process is extremely tedious and laborious. It is slower than all other loss functions as it suffers from a huge Data Expansion $O(N^2)$ where N is the total number of training samples. Generating all the possible triplets would be inefficient as many triplets easily satisfy the condition, however, if the process of triplet selection is optimized, such that only those are selected that would violate the restriction, the convergence rate would be much faster. Similarly, Center Loss would also converge faster if the number of classes is less and the number of samples within classes is limited, as evaluating the averages for the new center would be memory efficient.

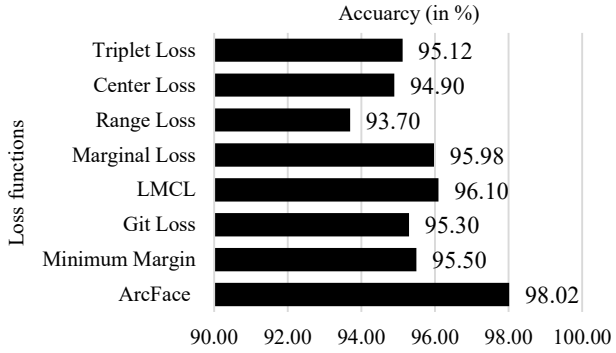

Fig. 1. A comparative face verification accuracy scored by various loss functions on the YouTube Faces (YTF) dataset.

Although Center Loss enhances the discriminative power of the deeply learned feature, it primarily increases just the intra-class compactness with a little emphasis on the inter-class separability Contrary to **Range Loss,** the intra-class compactness and inter-class separability both are achieved with the $\mathcal{L}_{R_{intra}}$ and $\mathcal{L}_{R_{inter}}$ functions. **Git Loss** fails in achieving optimum intra-class compactness simultaneously with optimum inter-class separability, as the functions $\mathcal{L}_C$ and $\mathcal{L}_G$ have opposite nature. Therefore, the trade-off between intra-class compactness and inter-class separability is inevitable. Undoubtedly, both the Center Loss Range Loss performance better than the Softmax Loss and the Git Loss outperforms Softmax Loss + Center Loss.

Marginal Loss sets an overstrict distance threshold θ which forces the distance of the two closest features belonging to different class to be greater than the distance of the two farthest sample of the same class. This condition result in a much harder convergence. **Minimum Margin Loss** provides much faster convergence than the Marginal Loss and the function also results in a better representation of the deeply learned features which is proven by the state-of-the-art accuracies.

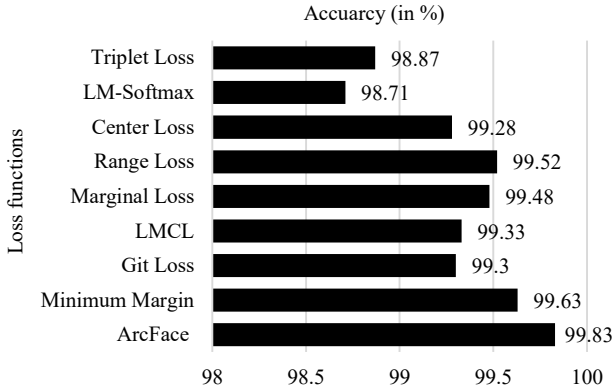

Fig. 2. A comparative face verification accuracy scored by various loss functions on the Labeled Faces in the Wild (LFW) dataset.

Range Loss brilliantly reduces the negative effect posed by the long-tailed classes and thereby providing a solution to such datasets. Using the harmonic mean for the intra-class loss is remarkable. Similar to the Triplet Loss effective, it requires a lot of computation power as updates are performed on a mini-batch of just four samples. This also costs a much slower convergence of the network.

CosFace is better than Soft-Margin Softmax and the Large-Margin Softmax, as the cosine margin term produces a larger angular margin and this enhances the discriminative power of the deeply learned face embeddings. The **ArcFace** is far better than Softmax based loss function because in Softmax based loss function as the number of identity (N) increases, it linearly increases the size of the Weight matrix. ArcFace is unquestionably better than the Triplet Loss function as the accuracy is comparable without the significant disadvantage of Data Explosion. ArcFace is the most advanced Angle-Based Loss since the additive angular margin m maps to the exact geodesic distance on a hypersphere. Clearly, ArcFace produces noticeable discriminative learning of deeply learned face embeddings.

Large-Margin Loss function certainly increases the discriminative power of the deeply learned function but at the expanse of heavy computation. The forward and backward propagations involve fairly complex calculations. **Soft Margin Loss** is significantly better than Softmax Loss function and slightly better than Large-margin Loss function in terms of accuracy because of the *soft continuous* margin. Also, Soft-Margin Loss just changes the forward computation of the loss equation, thus easier to implement. One of the biggest flaws about the Large-Margin is the constraint of *m* being a positive integer, as a result, this chops off many options for the different angular margins. Clearly proven by facts Large-Margin Loss outshines the Softmax Loss with appreciable intra-class compactness and inter-class separability.

6 Conclusion

In this paper, a comparative analysis of different loss functions for Deep Face Recognition is presented. Beginning from the Softmax Loss, it is a proven fact that the discovery of newer loss functions has enhanced the discriminative learning of the deep face embedding. The study of various loss functions concludes that the key to better discriminative learning is the maximization of the intra-class compactness and inter-class separability of the features. Fig,1 and Fig.2 presents a comparative face verification accuracy scored by various loss functions on the two standard testing datasets, LFW and YTF respectively. The loss functions from top to bottom are arranged in chronological order. It can be concluded that as newer loss functions have emerged, the increase in the accuracies is evident. Among the various approaches to loss functions, the Angle-Based Loss Functions have emerged as best in terms of training accuracy, testing accuracy and convergence rate. The loss functions which are in joint supervision with the Softmax loss do not achieve stable discrimination since intra-class loss function and inter-class loss function behave in opposition to one another. For the long-tailed datasets, Range Loss demonstrates an effective choice. Among all the functions, ArcFace or Additive Angular Margin Loss has undoubtedly achieved the state-of-the-art solution for Deep Face Recognition. Evaluating the accuracies on standard face datasets like LFW, YTF and MegaFace verify its superiority over other loss functions. In this paper, several loss functions were analyzed thoroughly and a comparative analysis is presented to demonstrate the performance under various practical situations.

REFERENCES

[1] [1] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229, 2013.

[2] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012).

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1–9, 2015.

[4] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In Advances in neural information processing systems, pages 487–495, 2014.

[5] E. Zhou, Z. Cao, and Q. Yin. Naive-deep face recognition: Touching the limit of lfw benchmark or not? arXiv preprint arXiv:1501.04690, 2015.

[6] E. Zhou, Z. Cao, and Q. Yin. Naive-deep face recognition:Touching the limit of lfw benchmark or not? arXiv preprint arXiv:1501.04690, 2015.

[7] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 1725–1732, 2014.

[8] Jesse Davis. 21 AMAZING USES FOR FACE RECOGNITION – FACIAL RECOGNITION USE CASES, 2019. https://www.facefirst.com/blog/amazing-uses-for-face-recognition-facial-recognition-use-cases/

[9] A. Calefati, M. Kamran Janjua, S. Nawaz, I. Gallo. Git Loss for Deep Face Recognition arXiv preprint arXiv:1807.08512v4, 2018

[10] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[11] L.Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 529–534. IEEE, 2011.

[12] Chen, B.C., Chen, C.S., Hsu, W.H.: Face recognition and retrieval using cross-agereference coding with cross-age celebrity dataset. IEEE Trans. Multimedia 17(6), 804–815 (2015)

[13] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. Agedb: The first manually collected in-the-wild age database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2017.

[14] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4873–4882, 2016.

[15] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In European Conference on Computer Vision, pages 87–102. Springer, 2016.

[16] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In FG, 2018.

[17] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. arXiv preprint arXiv:1411.7923, 2014.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016.

[19] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In European Conference on Computer Vision, pages 630–645. Springer, 2016.

[20] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inceptionv4, inception-resnet and the impact of residual connections on learning. arXiv preprint arXiv:1602.07261, 2016.

[21] Y. Sun, L. Ding, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. CoRR, abs/1502.00873, 2015.

[22] Y. Sun, L. Ding, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. CoRR, abs/1502.00873, 2015.

[23] J. Liu, Y. Deng, and C. Huang. Targeting ultimate accuracy: Face recognition via deep embedding. arXiv preprint:1506.07310, 2015.

[24] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," in Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, ser. CVPR '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 1701–1708.

[25] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 815–823, 2015.

[26] X. Wei, H. Wang, B. Scotney, and H. Wan, "Minimum margin loss for deep face recognition," arXiv preprint arXiv:1805.06741, 2018.

[27] J. Deng, Y. Zhou, and S. Zafeiriou, "Marginal Loss for Deep Face Recognition," 2017, pp. 60–68.

[28] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A Discriminative Feature Learning Approach for Deep Face Recognition," in Computer Vision – ECCV 2016, ser. Lecture Notes in Computer Science. Springer, Cham, Oct. 2016, pp. 499–515.

[29] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao, "Range Loss for Deep Face Recognition with Long-Tailed Training Data," in 2017 IEEE International Conference on Computer Vision (ICCV), Oct. 2017, pp. 5419–5428.

[30] Liang, Xuezhi, Wang, Xiaobo, Lei, Zhen, Liao, Shengcai, and Li, Stan Z. Soft-margin softmax for deep classification. In International Conference on Neural Information Processing, pp. 413–421. Springer, 2017.

[31] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-Margin Softmax Loss for Convolutional Neural Networks," in International Conference on Machine Learning, Jun. 2016, pp. 507–516.

[32] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Zhifeng Li, Dihong Gong, Jingchao Zhou, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In CVPR, 2018.

[33] J. Deng, J. Guo, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," arXiv:1801.07698 [cs], Jan. 2018, arXiv: 1801.07698.

[34] LeCun, Yann, Cortes, Corinna, and Burges, Christopher JC. The mnist database of handwritten digits, 1998.

[35] Krizhevsky, Alex. Learning multiple layers of features from tiny images. Technical Report, 2009.

[36] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE 86(11), 2278–2324 (1998).

[37] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 1, pages 539–546. IEEE, 2005.

[38] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[39] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In Advances in Neural Information Processing Systems, pages 1988–1996, 2014.

[40]] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. arXiv:1408.5093, 2014.

[41] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[42] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In European Conference on Computer Vision (ECCV), pages 499–515, 2016.