BroadFace: Looking at Tens of Thousands of People at Once for Face Recognition

Yonghyun Kim [⋆1][0000−0003−0038−7850], Wonpyo Park [⋆2][0000−0003−0675−6362], and Jongju Shin[1][0000−0002−7359−5258]

[1] Kakao Enterprise, Seongnam, Korea
{aiden.d, isaac.giant}@kakaoenterprise.com
[2] Kakao Corp., Seongnam, Korea
tony.nn@kakaocorp.com

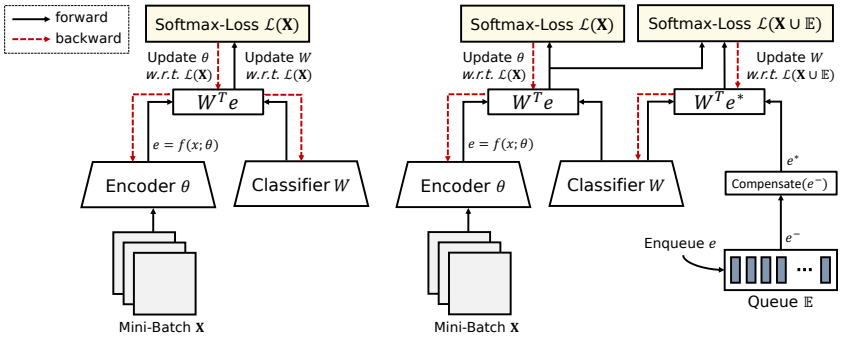
Abstract. The datasets of face recognition contain an enormous number of identities and instances. However, conventional methods have difficulty in reflecting the entire distribution of the datasets because a mini-batch of small size contains only a small portion of all identities. To overcome this difficulty, we propose a novel method called BroadFace, which is a learning process to consider a massive set of identities, comprehensively. In BroadFace, a linear classifier learns optimal decision boundaries among identities from a large number of embedding vectors accumulated over past iterations. By referring more instances at once, the optimality of the classifier is naturally increased on the entire datasets. Thus, the encoder is also globally optimized by referring the weight matrix of the classifier. Moreover, we propose a novel compensation method to increase the number of referenced instances in the training stage. BroadFace can be easily applied on many existing methods to accelerate a learning process and obtain a significant improvement in accuracy without extra computational burden at inference stage. We perform extensive ablation studies and experiments on various datasets to show the effectiveness of BroadFace, and also empirically prove the validity of our compensation method. BroadFace achieves the *state-of-the-art* results with significant improvements on nine datasets in 1:1 face verification and 1:N face identification tasks, and is also effective in image retrieval.

Keywords: face recognition, large mini-batch learning, image retrieval

1 Introduction

Face recognition is a key technique for many applications of biometric authentication such as electronic payment, lock screen of smartphones, and video surveillance. The main tasks of face recognition are categorized into face verification and face identification. In face verification, a pair of faces are compared to verify whether their identities are the same or different. In face identification, the identity of a given face is determined by comparing it to a pre-registered gallery of

⋆ Equal contribution



(a) Typical Mini-batch Learning (b) The Proposed Learning

Fig. 1. (a) In typical mini-batch learning, the parameter θ of encoder f and the parameter W of linear classifier are optimized on a small mini-batch $\mathbf{X}$. (b) In the proposed method, the parameter of the encoder is optimized on a small mini-batch, but the parameter of the classifier is optimized on both the mini-batch and the large queue $\mathbb{E}$ that contains embedding vectors e^- of past iterations.

identities. Many researches [1, 3, 4, 19, 27, 34, 38, 44, 47] on face recognition have been conducted for decades. The recent adoption [6, 7, 22, 32, 37, 39–41] of Convolutional Neural Networks (CNNs) has dramatically increased recognition accuracy. However, many difficulties of face recognition still remain to be solved.

Most previous studies focus on improving the discriminative power of an embedding space, because face recognition models are evaluated on independent datasets that include unseen identities. The mainstream of recent studies [6, 22, 39, 40] is to introduce a new objective function to maximize inter-class discriminability and intra-class compactness; they try to consider all identities by referring an identity-representative vector, which is the weight vector of the last fully-connected layer for identity classification.

However, conventional methods still have difficulty in covering a massive set of identities at once, because these methods use a small mini-batch (Fig. 1a) much less than the number of identities due to memory constraints. Inspecting tens of thousands of identities with the mini-batch of the small size requires numerous iterations, and this complicates the task of learning optimal decision boundaries in an embedding space while considering all of the identities, comprehensively. Increasing the size of the mini-batch may alleviate some of the problem, but in general, this solution is impractical because of memory constraints; it also does not guarantee improved accuracy [9, 12, 18, 48].

We propose a novel method, called *BroadFace*, which is a learning process to consider a massive set of identities, comprehensively (Fig. 1b). BroadFace has a large queue to keep a massive number of embedding vectors accumulated over past iterations. Our learning process increases the optimality of decision

boundaries of the classifier by considering the embedding vectors of both a given mini-batch and the large queue for each iteration. The parameters of the model are updated iteratively, so after a few iterations the error of enqueued embedding vectors gradually increases. Therefore, we introduce a compensation method that reduces the expected error between the current and enqueued embedding vectors by referencing the difference of the identity-representative vectors of current and past iterations. Our BroadFace has several advantages: (1) the identity-representative vectors are updated with a large number of embedding vectors to increase the portion of the training set that is considered for each iteration, (2) the optimality of the model is increased on the entire dataset by referring to the globally well-optimized identity-representative vectors, (3) the learning process is accelerated. We summarize the contributions as follows:

- We propose a new way that allows an embedding space to distinguish numerous identities in a broad perspective by learning identity-representative vectors from a massive number of instances.
- We perform extensive ablation studies on its behaviors, and experiments on various datasets to show the effectiveness of the proposed method, and to empirically prove the validity of our compensation method.
- BroadFace can be easily applied on many existing face recognition methods to obtain a significant improvement. Moreover, during inference time, it does not require any extra computational burden.

2 Related Works

Recent studies of face recognition tend to introduce a new objective function that learns an embedding space by exploiting an identity-representative vector. NormFace [39] reveal that optimization using cosine similarity between identity-representative vectors and embedding vectors is more effective than optimization using the inner product. To increase the discriminative abilities of learned features, SphereFace [22], CosFace [40] and ArcFace [6] adopted different kinds of margin into the embedding space. Futhermore, some works adopted an additional loss function to regulate the identity-representative vectors. RegularFace [52] minimized a cosine similarity between identity-representative vectors, and UniformFace [8] equalized distances between all the cluster centers. However, those methods can suffer from an enormous number of identities and instances because they are based on a mini-batch learning. Our BroadFace overcomes the limitation of a mini-batch learning, and, it can be easily applied on those face recognition methods.

In terms of preserving knowledge of model on previously visited data, the continual learning [13, 20] shares the similar concept with BroadFace. However, BroadFace is different from continual learning, as BroadFace preserves knowledge of previous data from the same dataset while continual learning preserves knowledge of previous data from different datasets.

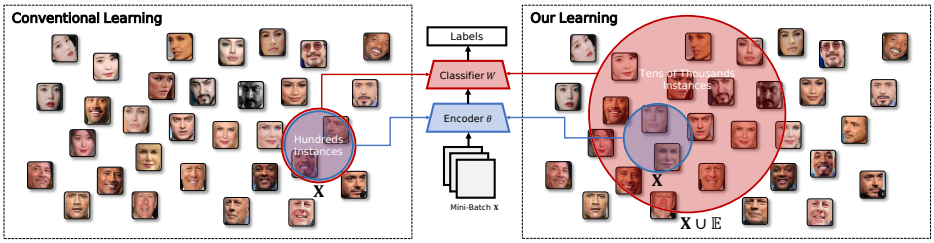


Fig. 2. Our learning refers more instances to learn the classifier in training stage.

3 Proposed Method

We describe the widely-adopted learning scheme in face recognition, and then illustrate the proposed BroadFace in detail.

3.1 Typical Learning

Learning of Face Recognition. In general, a face recognition network is divided into two parts: (1) an encoder network that extracts an embedding vector from a given image and (2) a linear classifier that maps an embedding vector into probabilities of identities. An evaluation is performed by comparing embedding vectors on images of unseen identities, so the classifier is discarded at the inference stage. Here, f is the encoder network that extracts a D-dimensional embedding vector e from a given image x: $e = f(x;\theta)$ with a model parameter θ. The linear classifier performs classification for C identities from an embedding vector e with a weight matrix W of $C \times D$-dimensions. For a mini-batch $\mathbf{X}$, the objective function such as a variant of angular softmax losses [6, 22, 39] is used to optimize the encoder and the classifier:

$$\mathcal{L}(\mathbf{X}) = \frac{1}{|\mathbf{X}|} \sum_{i \in \mathbf{X}} l(e_i, y_i), \tag{1}$$

$$l(e_i, y_i) = -\log \frac{\exp(\hat{W}_{y_i}^T \hat{e}_i)}{\sum_{j=1}^{C} \exp(\hat{W}_j^T \hat{e}_j)}, \tag{2}$$

where y_i is an labeled identity of x_i and $\hat{\cdot}$ indicates that a given vector is L_2 normalized (*e.g.*, $\|\hat{e}\|_2 = 1$). In Eq. 2, $\hat{W}_{y_i}$ acts as an representative instance of the given identity y_i that maximizes the cosine similarity with an embedding vector e_i. Thus, $\hat{W}_y$ can be regarded as the identity-representative vector, which is the expectation of instances belong to y:

$$\hat{W}_y = E_x \left[\hat{e}_i | y_i = y \right]. \tag{3}$$

Limitations of Mini-batch. The parameters of the model are updated in an iterative process that considers a mini-batch that contains only a small portion

of the entire dataset for each step (Fig. 2). However, use of a small mini-batch may not represent the entire distribution of training datasets. Moreover, in face recognition, the number of identities is very large and each mini-batch only contains few of them; for example, MSCeleb-1M [10] has 10M images of 100k celebrities. Therefore, each parameter update of a model can be biased on a small number of identities, and this restriction complicates the task of finding optimal decision boundaries. Enlarging the mini-batch size may mitigate the problem, but this solution requires heavy computation of the encoder, proportional to the batch-size.

3.2 BroadFace

We introduce BroadFace, which is a simple yet effective way to cover a large number of instances and identities. BroadFace learns globally well-optimized identity-representative vectors from a massive number of embedding vectors (Fig. 2). For example, on a single Nvidia V100 GPU, the size of a mini-batch for ResNet-100 is at most 256, whereas BroadFace can utilize more than 8k instances at once. The following describes each step.

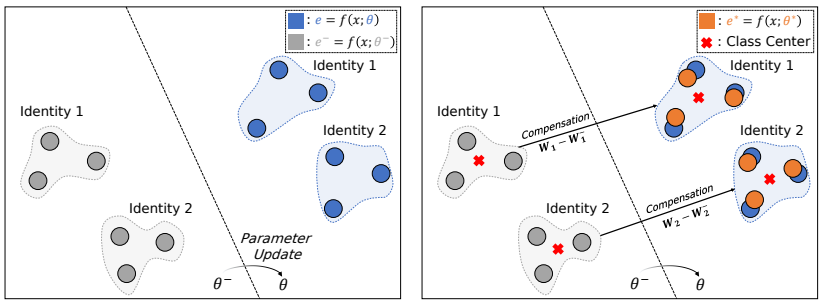
(1) Queuing Past Embedding Vectors. BroadFace has two queues of predefined size: $\mathbb{E}$ stores embedding vectors; $\mathbb{W}$ stores identity-representative vectors from the past. For each iteration, after model update, embedding vectors of a given mini-batch $\{e_i\}_{i\in\mathbf{X}}$ are enqueued to $\mathbb{E}$, and corresponding identity-representative vectors $\{W_{y_i}\}_{i\in\mathbf{X}}$ of each instance are enqueued to $\mathbb{W}$. By referring to the past embedding vectors in the queue to compute the loss $\mathcal{L}(\mathbf{X}\cup\mathbb{E})$, the network increase the number of instances and identities explored at each update.

(2) Compensating Past Embedding Vectors. As the model parameter θ of the encoder is updated over iterations, past embedding vectors $e^{\text{-}}\in\mathbb{E}$ conflict with the embedding space of the current parameter (Fig. 3a); $\epsilon = e - e^{\text{-}}$ where $\theta^{\text{-}}$ is the past parameter of the encoder and $e^{\text{-}} = f(x;\theta^{\text{-}})$. The magnitude of the error ϵ is relatively small when few iterations have been passed from the past. However, the error is gradually accumulated over iterations and the error hinders appropriate training. We introduce a compensation function $\rho(y)$ for each identity to reduce the errors as an additive model; $e_i^* = e_i^{\text{-}} + \rho(y)$ where e_i^* is a compensated past embedding vector to a current embedding vector (Fig. 3b). The compensation function should minimize an expected squared error J between the current embedding vectors and the compensated past embedding vectors that belong to y:

$$\begin{aligned}\text{minimize } J\big(\rho(y)\big) =& E_x\left[\big(e_i^* - e_i\big)^2 \big| y_i = y\right], \\ =& E_x\left[\big(e_i^{\text{-}} + \rho(y) - e_i\big)^2 \big| y_i = y\right].\end{aligned} \tag{4}$$

The partial derivative of J with respect to $\rho(y)$ is:

$$\frac{\partial J}{\partial \rho(y)} = E_x\Big[2\big(e_i^{\text{-}} + \rho(y) - e_i\big)\big| y_i = y\Big]. \tag{5}$$



(a) without Compensation (b) with Compensation

Fig. 3. (a) enqueued embedding vectors (gray circles) at the past are further away from the embedding vectors (blue circles) at the current iteration due to the parameter update and this indicates significant errors. (b) the compensated embedding vectors (orange circles) closely approach to the embedding vectors at the current iteration, by considering the difference between the identity-representative vectors (class centers) at the past and the current iteration.

Thus, the optimal compensation function is the difference between the expectations of current embedding vectors and past embedding vectors:

$$\begin{aligned}\rho(y) &= E_x\left[e_i \middle| y_i = y\right] - E_x\left[e_i^- \middle| y_i = y\right], \\ &\approx \lambda(W_y - W_y^-),\end{aligned} \tag{6}$$

where $W_y^- \in \mathbb{W}$ is an identity-representative vector, which is enqueued to the queue during the same iteration as $e_i^- \in \mathbb{E}$. As explained in Eq. 3, the identity-representative vector and the expected embedding vector point in the same direction when the vectors are projected onto a hyper-sphere, but the vectors are different in scale. Thus, we deploy a simple normalization term per each instance to adjust these scales: $\lambda = \|e_i^-\| / \left\|W_{y_i}^-\right\|$. Then the compensated embedding vector e^* is computed as:

$$e_i^* = e_i^- + \frac{\|e_i^-\|}{\left\|W_{y_i}^-\right\|}(W_{y_i} - W_{y_i}^-). \tag{7}$$

In empirical studies, the compensation function reduces the error significantly.
(3) Learning from Numerous Embedding Vectors. By executing the preceding two steps, BroadFace generates additional large-scale embedding vectors from the past. In our method, the encoder is trained on a mini-batch as before and the classifier is trained on both a mini-batch and the additional embedding vectors. The objective functions for the encoder and the classifier are defined as:

$$\mathcal{L}_{\text{encoder}}(\mathbf{X}) = \frac{1}{|\mathbf{X}|}\left\{\sum_{i \in \mathbf{X}} l(e_i)\right\}, \tag{8}$$

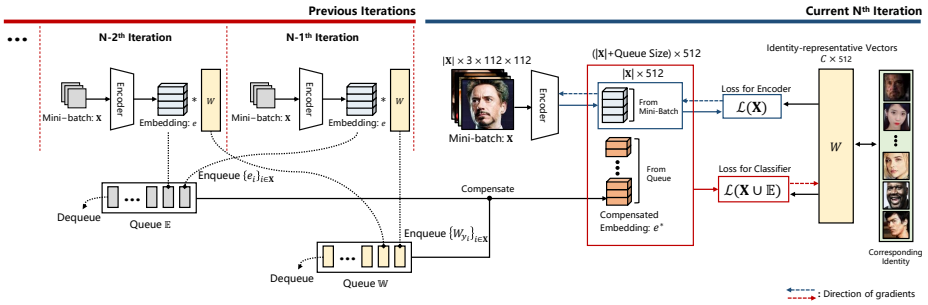


Fig. 4. Learning process of the proposed method. BroadFace deploys large queues to store embedding vectors and their corresponding identity-representative vectors per iteration. The embedding vectors of the past instances stored in the queues are used to compute loss for identity-representative vectors. BroadFace effectively learns from tens of thousands of instances for each iteration.

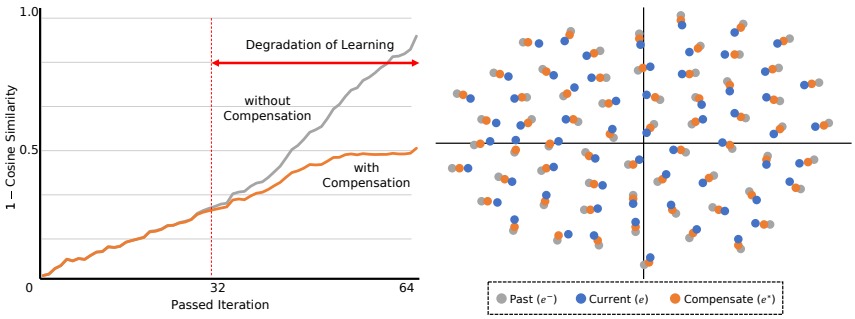
$$\mathcal{L}_{\text{classifier}}(\mathbf{X} \cup \mathbb{E}) = \frac{1}{|\mathbf{X} \cup \mathbb{E}|} \left\{ \sum_{i \in \mathbf{X}} l(e_i) + \sum_{j \in \mathbb{E}} l(e_j^*) \right\}. \tag{9}$$

The parameter θ of the encoder is updated *w.r.t.* $\mathcal{L}_{\text{encoder}}(\mathbf{X})$ while the parameter of the classifier W is updated *w.r.t.* $\mathcal{L}_{\text{classifier}}(\mathbf{X} \cup \mathbb{E})$. The large number of embedding vectors in the queue helps to learn highly precise identity-representative vectors that show reduced bias on a mini-batch and increased optimality on the entire dataset. The precise identity-representative vectors can accelerate the learning procedure. Moreover, our method can be easily implemented by adding several queues in the learning process (Fig. 4) and significantly improves accuracy in face recognition without any computational cost at inference stage.

3.3 Discussion

Effectiveness of Compensation. We show that the compensation method is empirically effective. After a small number of iterations, the error of the enqueued embedding vectors is also small and the compensation method is not necessary. However, after a large number of iterations, the error increases and the compensation method becomes necessary to keep a large number of embedding vectors (Fig. 5a). A large accumulated error may degrade the training process of the network (Fig. 7a and Fig. 7b). We illustrate how the compensation function reduces the difference between past and current embedding vectors in 2-dimensional space by t-SNE [24] (Fig. 5b). The past embedding vectors approach to current embedding vectors after applying compensation. This shows that the proposed compensation function works properly in practice.

Memory Efficiency. We compare BroadFace with enlarging the size of a mini-batch in terms of memory consumption (Fig. 6). A naïve mini-batch learning requires a huge amount of memory to forward and backward the entire network.



(a) Averaged Error (b) Scatter Plot

Fig. 5. (a) the average of the cosine errors between the embedding vectors at the current iteration and the past iteration with and without compensation. The errors are computed with randomly sampled instances over 64 iterations. (b) the scatter plot of the past (before 64 iterations), current and compensated embedding vectors for 64 instances.

The maximum size of a mini-batch is about 240 instances when a model based on ResNet-100 is trained on NVidia V100 of 32 GB. However, BroadFace requires only a matrix multiplication between the embedding vectors in $\mathbb{E}$ and the weight matrix W (Eq. 2). The marginal computational cost of BroadFace enables a classifier to learn decision boundaries from a massive set of instances, *e.g.*, 8192 instances for a single GPU. Note that, enlarging the size of a mini-batch to 8192 requires about 952 GB of memory which is infeasibly large for a single GPU.

4 Experiments

4.1 Implementation Details

Experimental Setting. As pre-processing, we normalize a face image to 112×112 by warping a face-region using five facial points from two eyes, nose and two corners of mouth [6, 22, 40]. A backbone network is ResNet-100 [11] that is used in the recent works [6, 15]. After the `res5c` layer of ResNet-100, a block of batch normalization, fully-connected and batch normalization layers is deployed to compute a 512-dimensional embedding vector. The computed embedding vectors and the weight vectors of the linear classifier are L_2-normalized and trained by the ArcFace [6]. Our model is trained on 4 synchronized NVidia V100 GPUs and a mini-batch of 128 images is assigned for each GPU. The queue of BroadFace stores up to 8,192 embedding vectors accumulated over 64 iterations for each GPU, thus the total size of the queues is 32,768 for 4 GPUs. To avoid abrupt changes in the embedding space, the network of BroadFace is trained from the pre-trained network that is trained by the softmax based loss [6]. We adopted

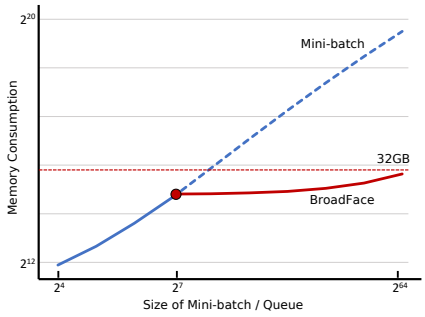


Fig. 6. Illustration on memory consumption of a conventional mini-batch learning (blue line) and the proposed BroadFace (red line) depending on the size of a mini-batch. The blue dotted line indicates memory consumption that is estimated for a large size of mini-batch by linear regression.

stochastic gradient descent (SGD) optimizer, and a learning rate is set to $5 \cdot 10^{-3}$ for the first 50k, $5 \cdot 10^{-4}$ for the 20k, and $5 \cdot 10^{-5}$ for the 10k with a weight decay of $5 \cdot 10^{-4}$ and a momentum of 0.9.

Datasets. All the models are trained on MSCeleb-1M [10], which is composed of about 10M images for 100k identities. We use the refined version [6], which contains 3.8M images for 85k identities by removing the noisy labels of MSCeleb-1M. For the test, we perform evaluations on the following various datasets:

- Labeled Faces in the Wild (LFW) [14] contains 13k images of faces that are collected from web for 5,749 different individuals. Cross-Age LFW (CALFW) [54] provides pairs with age variation, and Cross-Pose (CPLFW) [53] provides pairs with pose variation from the images of LFW.
- YouTube Faces (YTF) [43] contains 3,425 videos of 1,595 different people.
- MegaFace [17] contains more than 1M images from 690k identities to evaluate recognition-accuracy with enormous distractors.
- Celebrities in Frontal-Profile (CFP) [33] contains 500 subjects; each subject has 10 frontal and 4 profile images.
- AgeDB-30 [26], which contains 12,240 images of 440 identities with age variations, is suitable to evaluate the sensitivity of a given method in age variation.
- IARPA Janus Benchmark (IJB) [25, 42], which is designed to evaluate unconstrained face recognition systems, is one of the most challenging datasets in public. IJB-B [42] is composed of 67k face images, 7k face videos and 10k non-face images. IJB-C [25], which adds additional new subjects with increased occlusion and diversity of geographic origin to IJB-B, is composed of 138k face images, 11k face videos and 10k non-face images.

4.2 Evaluations on Face Recognition

We conduct experiments on the described various datasets to show the effectiveness of the proposed method.

Table 1. Verification accuracy (%) on LFW and YTF.

Method	LFW	YTF	Method	LFW	YTF
DeepID [36]	99.47	93.2	DeepFace [37]	97.35	91.4
VGGFace [28]	98.95	97.3	FaceNet [32]	99.64	95.1
CenterLoss [41]	99.28	94.9	RangeLoss [51]	99.52	93.7
MarginalLoss [7]	99.48	95.9	SphereFace [22]	99.42	95.0
RegularFace [52]	99.61	96.7	CosFace [40]	99.81	97.6
UniformFace [8]	99.80	97.7	AFRN [15]	99.85	97.1
ArcFace [6]	99.83	97.7	BroadFace	**99.85**	**98.0**

Table 2. Verification accuracy (%) on CALFW, CPLFW, CFP-FP and AgeDB-30.

Method	CALFW	CPLFW	CFP-FP	AgeDB-30
CenterLoss [41]	85.48	77.48	-	-
SphereFace [22]	90.30	81.40	-	-
VGGFace2 [2]	90.57	84.00	-	-
CosFace [40]	95.76	92.28	98.12	98.11
ArcFace [6]	95.45	92.08	98.27	98.28
BroadFace	**96.20**	**93.17**	**98.63**	**98.38**

Table 3. Identification and verification evaluation on MegaFace [17]. Ident indicates rank-1 identification accuracy (%) and Verif indicates a true accept rate (%) at a false accept rate of 1e-6.

Method	MF-Large		MF-Large-Refined [6]	
	Ident	Verif	Ident	Verif
RegularFace [52]	75.61	91.13	-	-
UniformFace [8]	79.98	95.36	-	-
SphereFace [22]	-	-	97.91	97.91
AdaptiveFace [21]	-	-	95.02	95.61
CosFace [40]	80.56	96.56	97.91	97.91
ArcFace [6]	81.03	96.98	98.35	98.49
BroadFace	**81.33**	**97.56**	**98.70**	**98.95**

LFW and YTF are widely used to evaluate verification performance under the unrestricted environments. LFW, which contains pairs of images, evaluates a model by comparing two embedding vectors of a given pair. YTF contains videos that are sets of images; from the shortest clip of 48 frames to the longest clip of 6,070 frames. To compare a pair of videos, YTF compares a pair of video-

Table 4. Verification evaluation with a True Accept Rate at a certain False Accept Rate (TAR@FAR) from 1e-4 to 1e-6 on IJB-B and IJB-C. † denotes BroadFace trained by CosFace [40].

Method	IJB-B			IJB-C		
	FAR=1e-6	FAR=1e-5	FAR=1e-4	FAR=1e-6	FAR=1e-5	FAR=1e-4
VGGFace2 [2]	-	0.671	0.800	-	0.747	0.840
CenterFace [41]	-	-	-	-	0.781	0.853
ComparatorNet [46]	-	-	0.849	-	-	0.885
PRN [16]	-	0.721	0.845	-	-	-
AFRN [15]	-	0.771	0.885	-	0.884	0.931
CosFace [40]	0.3649	0.8811	0.9480	0.8591	0.9410	0.9637
BroadFace†	0.4092	0.8997	**0.9497**	0.8596	**0.9459**	**0.9638**
ArcFace [6]	0.3828	0.8933	0.9425	0.8906	0.9394	0.9603
BroadFace	**0.4653**	**0.9081**	0.9461	**0.9041**	0.9411	0.9603

representative embedding vectors that are averaged embedding vectors of images collected from each video. Even though both datasets are highly-saturated in accuracy, our BroadFace outperforms other recent methods (Table 1).

CALFW, CPLFW, CFP-FP and AgeDB-30 are also widely used to verify that methods are robust to pose and age variation. CALFW and AgeDB-30 have multiple instances for same identity of different ages and CPLFW and CFP-FP have multiple instances for same identity of different poses (frontal and profile faces). BroadFace shows better verification-accuracy on all datasets (Table 2).

MegaFace is designed to evaluate both face identification and verification tasks under difficulty caused by a huge number of distractors. We evaluate our BroadFace on Megaface Challenge 1 where the training dataset is more than 0.5 million images. BroadFace outperforms the other top-ranked face recognition models for both face identification and verification tasks (Table 3). On the refined MegaFace [6], where noisy labels are removed, BroadFace also surpasses the other models.

IJB-B and IJB-C are the most challenging datasets to evaluate unconstrained face recognition. We report BroadFace with CosFace [40] and BroadFace with ArcFace [6] in verification task without any augmentations such as horizontal flipping in test time. Our BroadFace shows significant improvements on all FAR criteria (Table 4). In IJB-B [42], BroadFace improves 8.25 percentage points on FAR=1e-6, 1.48 percentage points on FAR=1e-5 and 0.36 percentage points on FAR=1e-4 comparing to the results of ArcFace [6].

4.3 Evaluations on Image Retrieval.

Both face recognition and image retrieval have the same goal that learn an optimal embedding space to compare a given pair of items such as face, clothes or industrial products. To show that BroadFace is widely applicable on other appli-

Table 5. Recall@K comparison with state-of-the-art methods. For fair comparison, we divide methods according to the dimension (Dim.) of an embedding vector. The numbers under the datasets refer to recall at K.

Methods	Dim.	In-Shop				SOP			
		1	10	20	30	1	10	10^2	10^3
Margin [45]	128	-	-	-	-	72.7	86.2	93.8	98.0
MIC+Margin [29]	128	88.2	97.0	-	-	77.2	89.4	95.6	-
DC [31]	128	85.7	95.5	96.9	97.5	75.9	88.4	94.9	98.1
ArcFace [6]	128	84.1	94.9	96.2	96.9	73.3	86.4	93.2	97.1
BroadFace	128	**89.8**	**97.4**	**98.1**	**98.4**	**79.7**	**90.7**	**95.7**	**98.4**
TML [49]	512	-	-	-	-	78.0	**91.2**	**96.7**	**99.0**
NSM [50]	512	88.6	**97.5**	**98.4**	**98.8**	78.2	90.6	96.2	-
ArcFace [6]	512	87.3	96.3	97.3	97.9	76.9	89.1	95.0	98.2
BroadFace	512	**90.1**	97.4	98.1	98.4	**80.2**	91.0	95.9	98.4

cations, we compare BroadFace with recently proposed metric learning methods for image retrieval.

Experimental Settings. We use ResNet-50 [11] that is pre-trained on ILSVRC 2012-CLS [30] as a backbone network. We use ArcFace [6] as a baseline objective function and set the size of the queue to 32k for BroadFace. We follow the standard input augmentation and evaluation protocol [35]. We evaluate on two large datasets with a large number of classes similar to face-recognition: In-Shop Clothes Retrieval (In-Shop) [23] and Stanford Online Products (SOP) [35].

In-Shop and SOP are the standard datasets in image retrieval. In-Shop contains 11,735 classes of clothes. For training, the first 3,997 classes with 25,882 images are used, and the remaining 7,970 classes with 26,830 images are split into query set gallery set for evaluation. SOP contains 22,634 classes of industrial products. For training, the first 11,318 classes with 59,551 images are used and the remaining 11,316 classes with 60,499 images are used for evaluation. Our baseline models that are trained with ArcFace [6] underperform comparing to the other state-of-the-art methods. BroadFace significantly improves the recall of the baseline models and the improved model even outperforms the other methods (Table 5).

4.4 Analysis of BroadFace

Size of Queue. BroadFace has only one hyper-parameter, the size of the queue, to determine the maximum number of embedding vectors accumulated over past iterations. Using the single parameter makes our method easy to tune and the parameter plays a very important role in determining recognition-accuracy. As the size of the queue grows, the performance increases steadily (Table 6a). Especially, without our compensation method, accuracy degradation occurs when the size of the queue is significantly large. However, our compensation method

Table 6. Effects of BroadFace varying the size of the queue and the type of the backbone network on IJB-B dataset in face recognition.

(a) Total Size of Queue

Size of Queue	TAR			
	FAR=1e-6		FAR=1e-5	
	without Compensation	with Compensation	without Compensation	with Compensation
0 (Baseline)	0.3828	0.3828	0.8933	0.8933
2048 (512 × 4 GPUs)	0.4310	0.4255	0.9061	0.9077
8192 (2048 × 4 GPUs)	0.4346	0.4394	0.9071	0.9085
32768 (8192 × 4 GPUs)	0.4259	0.4653	0.9078	0.9081

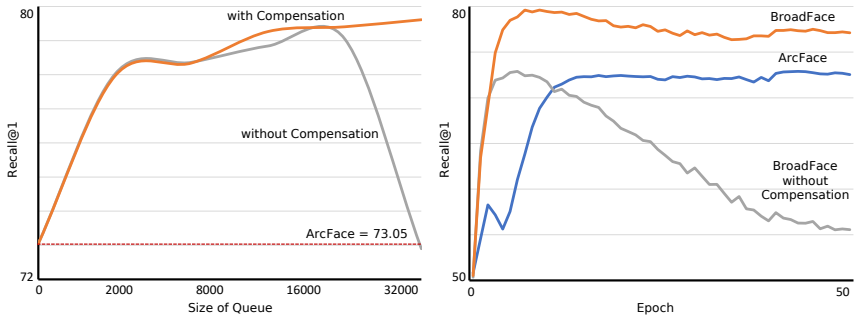
(b) Backbone Network

	Dim.	TAR				GFlops
		FAR=1e-6		FAR=1e-5		
		ArcFace	BroadFace	ArcFace	BroadFace	
MobileFaceNet [5]	128	0.3552	0.3665	0.8456	0.8458	0.9G
ResNet-18 [11]	128	0.3678	0.3808	0.8588	0.8638	5.2G
ResNet-34 [11]	512	0.3981	0.4325	0.8798	0.8828	8.9G
ResNet-100 [11]	512	0.3828	0.4653	0.8933	0.9081	24.1G

alleviates the degradation by correcting the enqueued embedding vectors. We show another experiment on the enormous size of the queue from 0 to $32,000$ in image retrieval (Fig. 7a). With the proposed compensation, the recall is consistently improved as the size of the queue is increased. However, without the proposed compensation, the recall of the models degrades when the size of the queue is more than 16k.

Generalization Ability. Our BroadFace is generally applicable to any objective functions and any backbone networks. We apply BroadFace to two widely used objective functions of CosFace [40] and ArcFace [6]. For both CosFace and ArcFace, BroadFace increases recognition-accuracy (Table 4). We also apply BroadFace to several backbone networks such as MobileFaceNet [5], ResNet-18 [11] and ResNet-34 [11]. We set the dimensions of embedding vector to 128 for light backbone networks such as MobileFaceNet and ResNet-18, and 512 for heavy backbone networks such as ResNet-34 and ResNet-100. BroadFace is significantly effective for all backbone networks (Table 6b). In particular, ResNet-34 trained with BroadFace achieves comparable performance to ResNet-100 trained only with ArcFace, even though ResNet-34 has much less GFlops.

Learning Acceleration. Our BroadFace accelerates the learning process of both face recognition and image retrieval. In face recognition, many iterations



(a) Size of Queue (b) Learning Acceleration

Fig. 7. (a) the recall depending on the size of the queue in BroadFace with and without our compensation function; the red line indicates the recall of ArcFace (baseline) on the test set. (b) the learning curve for the test set when the size of the queue is 32k; ArcFace reaches the highest recall at the 45^{th} epoch, our BroadFace reaches the highest recall at the 10^{th} epoch, and the learning process collapses without our compensation function.

are still needed to overcome a small gap of performance among the methods on the highly-saturated datasets. Thus, we experiment the acceleration of the learning process in image retrieval to clearly show the effectiveness (Fig. 7b). Our BroadFace reaches peak performance much faster and higher than the baseline model. Without our compensation method, the model gradually collapses.

5 Conclusion

We introduce a new way called BroadFace that allows an embedding space to distinguish numerous identities in a broad perspective by increasing the optimality of constructed identity-representative vectors. BroadFace is significantly effective for face recognition and image retrieval where their datasets consist of numerous identities and instances. BroadFace can be easily applied on many existing face recognition methods to obtain a significant improvement without any extra computational cost in the inference stage.

Acknowledgement

We would like to thank AI R&D team of Kakao Enterprise for the helpful discussion. In particular, we would like to thank Yunmo Park who designed the visual materials.

References

1. Ahonen, T., Hadid, A., Pietikäinen, M.: Face recognition with local binary patterns. In: Pajdla, T., Matas, J. (eds.) European Conference on Computer Vision Workshops (2004)
2. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: International Conference on Automatic Face and Gesture Recognition (2018)
3. Chen, D., Cao, X., Wen, F., Sun, J.: Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In: IEEE Conference on Computer Vision and Pattern Recognition (2013)
4. Chen, D., Cao, X., Wipf, D., Wen, F., Sun, J.: An efficient joint formulation for bayesian face verification. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
5. Chen, S., Liu, Y., Gao, X., Han, Z.: Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In: Chinese Conference on Biometric Recognition (2018)
6. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2019)
7. Deng, J., Zhou, Y., Zafeiriou, S.: Marginal loss for deep face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (2017)
8. Duan, Y., Lu, J., Zhou, J.: Uniformface: Learning deep equidistributed representation for face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2019)
9. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677 (2017)
10. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) European Conference on Computer Vision (2016)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
12. Hoffer, E., Hubara, I., Soudry, D.: Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In: Advances in Neural Information Processing Systems (2017)
13. Hou, S., Pan, X., Loy, C.C., Wang, Z., Lin, D.: Learning a unified classifier incrementally via rebalancing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 831–839 (2019)
14. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. rep., University of Massachusetts, Amherst (2007)
15. Kang, B.N., Kim, Y., Jun, B., Kim, D.: Attentional feature-pair relation networks for accurate face recognition. In: IEEE International Conference on Computer Vision (2019)
16. Kang, B.N., Kim, Y., Kim, D.: Pairwise relational networks for face recognition. In: European Conference on Computer Vision (2018)
17. Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The megaface benchmark: 1 million faces for recognition at scale. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)

18. Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P.: On large-batch training for deep learning: Generalization gap and sharp minima. International Conference on Learning Representations (2017)
19. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: IEEE International Conference on Computer Vision (2009)
20. Li, Z., Hoiem, D.: Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence **40**(12), 2935–2947 (2017)
21. Liu, H., Zhu, X., Lei, Z., Li, S.Z.: Adaptiveface: Adaptive margin and sampling for face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2019)
22. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
23. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
24. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research (2008)
25. Maze, B., Adams, J., Duncan, J.A., Kalka, N., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., Grother, P.: Iarpa janus benchmark - c: Face dataset and protocol. In: International Conference on Biometrics (2018)
26. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: Agedb: the first manually collected, in-the-wild age database. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (2017)
27. Nguyen, H.V., Bai, L.: Cosine similarity metric learning for face verification. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) Asian Conference on Computer Vision (2011)
28. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference (2015)
29. Roth, K., Brattoli, B., Ommer, B.: Mic: Mining interclass characteristics for improved metric learning. In: IEEE International Conference on Computer Vision (2019)
30. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (2015)
31. Sanakoyeu, A., Tschernezki, V., Buchler, U., Ommer, B.: Divide and conquer the embedding space for metric learning. In: IEEE Conference on Computer Vision and Pattern Recognition (2019)
32. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
33. Sengupta, S., Chen, J., Castillo, C., Patel, V.M., Chellappa, R., Jacobs, D.W.: Frontal to profile face verification in the wild. In: IEEE Winter Conference on Applications of Computer Vision (2016)
34. Simonyan, K., Parkhi, O., Vedaldi, A., Zisserman, A.: Fisher vector faces in the wild. In: British Machine Vision Conference (2013)
35. Song, H.O., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)

36. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: Advances in Neural Information Processing Systems (2014)
37. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: IEEE Conference on Computer Vision and Pattern Recognition (2014)
38. Turk, M.A., Pentland, A.P.: Face recognition using eigenfaces. In: IEEE Conference on Computer Vision and Pattern Recognition (1991)
39. Wang, F., Xiang, X., Cheng, J., Yuille, A.L.: Normface: L2 hypersphere embedding for face verification. In: ACM International Conference on Multimedia (2017)
40. 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: IEEE Conference on Computer Vision and Pattern Recognition (2018)
41. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) European Conference on Computer Vision (2016)
42. Whitelam, C., Taborsky, E., Blanton, A., Maze, B., Adams, J., Miller, T., Kalka, N., Jain, A.K., Duncan, J.A., Allen, K., Cheney, J., Grother, P.: Iarpa janus benchmark-b face dataset. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (2017)
43. Wolf, L., Hassner, T., Maoz, I.: Face recognition in unconstrained videos with matched background similarity. In: IEEE Conference on Computer Vision and Pattern Recognition (2011)
44. Wolf, L., Hassner, T., Taigman, Y.: Descriptor based methods in the wild. In: European Conference on Computer Vision Workshops (2008)
45. Wu, C.Y., Manmatha, R., Smola, A.J., Krahenbuhl, P.: Sampling matters in deep embedding learning. In: IEEE International Conference on Computer Vision (2017)
46. Xie, W., Shen, L., Zisserman, A.: Pairwise relational networks for face recognition. In: European Conference on Computer Vision (2018)
47. Yin, Q., Tang, X., Sun, J.: An associate-predict model for face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2011)
48. You, Y., Gitman, I., Ginsburg, B.: Scaling sgd batch size to 32k for imagenet training. arXiv preprint arXiv:1708.03888 (2017)
49. Yu, B., Tao, D.: Deep metric learning with tuplet margin loss. In: IEEE International Conference on Computer Vision (2019)
50. Zhai, A., Wu, H.Y.: Classification is a strong baseline for deep metric learning. arXiv preprint arXiv:1811.12649 (2018)
51. Zhang, X., Fang, Z., Wen, Y., Li, Z., Qiao, Y.: Range loss for deep face recognition with long-tailed training data. In: IEEE International Conference on Computer Vision (2017)
52. Zhao, K., Xu, J., Cheng, M.M.: Regularface: Deep face recognition via exclusive regularization. In: IEEE Conference on Computer Vision and Pattern Recognition (2019)
53. Zheng, T., Deng, W.: Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. Tech. rep., Beijing University of Posts and Telecommunications (2018)
54. Zheng, T., Deng, W., Hu, J.: Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. arXiv preprint arXiv:1708.08197 (2017)