Pairwise Ranking Distillation for Deep Face Recognition

Mikhail Nikitin[1,2], Vadim Konushin[1], and Anton Konushin[2][0000−0002−6152−0021]

[1] Video Analysis Techonologies LLC, Moscow, Russia
{mikhail.nikitin,vadim}@tevian.ru
[2] M.V. Lomonosov Moscow State University, Moscow, Russia
ktosh@graphics.cs.msu.ru

Abstract. This work addresses the problem of knowledge distillation for deep face recognition task. Knowledge distillation technique is known to be an effective way of model compression, which implies transferring of the knowledge from high-capacity teacher to a lightweight student. The knowledge and the way how it is distilled can be defined in different ways depending on the problem where the technique is applied. Considering the fact that face recognition is a typical metric learning task, we propose to perform knowledge distillation on a score-level. Specifically, for any pair of matching scores computed by teacher, our method forces student to have the same order for the corresponding matching scores. We evaluate proposed pairwise ranking distillation (PWR) approach using several face recognition benchmarks for both face verification and face identification scenarios. Experimental results show that PWR not only can improve over the baseline method by a large margin, but also outperforms other score-level distillation approaches.

Keywords: Knowledge Distillation, Model Compression, Face Recognition, Deep Learning, Metric Learning.

1 Introduction

Face recognition systems are widely used today, and their quality keeps improving in order to better fit increasing security requirements. Nowadays majority of the computer vision tasks, including facial recognition, are solved with the help of deep neural networks, and there exists a clear dependency that in case of a fixed training dataset, a network with a lot of layers and parameters outperforms its lightweight version. As a result, the most powerful models use a large amount of memory and computational resources, and therefore their deployment is quite challenging. Indeed, switching to the model of higher capacity usually results in reducing of the inference speed, which is very important in some real-life scenarios. For example, if the model is supposed to run on a resource-limited embedded device or to be used in video surveillance system with thousands of queries per second, it is often necessary to replace a large network with smaller one for the purpose of satisfying the limitations of available computational resources. This creates a strong demand for methods that reduce model complexity while trying to preserve its performance as much a possible.

In general, there are two main strategies to reduce deep neural network complexity: one is to develop a new lightweight architecture [1–3], and another one is to compress already trained model. Network compression can be done in many different ways, including parameter quantization [4, 5], weights prunning [6, 7], low-rank factorization [8, 9], and knowledge distillation. All these compression methods, except for the knowledge distillation, focus on reducing model size in terms of parameters while keeping network architecture roughly the same. On the contrary, knowledge distillation, the main idea of which is to transfer knowledge encoded in one network to another, is considered to be a more general approach, since it doesn't impose any restrictions on the architecture of the output network.

Therefore, in this paper we propose a new knowledge distillation technique for efficient computation of face recognition embeddings. Our method utilizes the idea of pairwise learning-to-rank approach and applies it on top of the matching scores between face embeddings. Specifically, we consider scores' ranking produced by a teacher network as a ground truth label, and use it to detect and penalize mistakes in pairwise ranking of student's matching scores. Using LFW [32], CPLFW [33], AgeDB [34], and MegaFace [35] datasets, we show that the proposed distillation method can significantly improve face recognition quality compared to the conventional way of training the student network. Moreover, we found that our pairwise ranking distillation technique outperforms other scores-based distillation approaches by a large margin.

2 Related Work

In [14] a dichotomy of distillation approaches was proposed. It is based on the way how the knowledge is determined, and the authors distinguish *individual* and *relational* knowledge distillation methods.

2.1 Individual knowledge distillation

Individual knowledge distillation (IKD) methods consider each input object independently and force student network to mimic teacher's representation of that object. Let $F_T(x)$ and $F_S(x)$ represent the feature representations of teacher and student for input x respectively. Then, for training dataset $\chi = \{x_i\}_{i=1}^{M}$ the IKD objective function can be formulated as follow:

$$L_{IKD} = \sum_{x_i \in \chi} l(F_T(x_i), F_S(x_i)), \tag{1}$$

where l is some loss function that penalizes the difference between the teacher and the student. The knowledge in IKD methods is determined by the function $F(x)$, which can be defined in different ways. Some examples are presented below.

Authors of [10] and [11] describe the knowledge in terms of labels distribution, so that student uses output of teacher's classifier as a ground truth soft label vector. The motivation of such approach lies in observation that input image sometimes contains several objects in it and can be better described using a mixture of labels. Another approach was presented in [12], where authors propose to use hint connections, which go from teacher to student and transfer hidden layer activations. Depending on depth of network and spatial resolution of features where such distillation is applied, it makes student to mimic teacher at different levels of abstraction. However, over-reguralization of hidden layers can lead to poor quality, so usually hints are only used for embedding (pre-classification) layer [16, 21]. In order to successfully guide student even at initial layers, modification of hints idea was proposed in [13]. Transferring of activation was replaced there with transferring of spatial attention maps, i.e. instead of trying to reproduce teacher's feature representation as is, student only learns to analyze the same areas of input image.

Individual knowledge distillation methods utilize clear idea of imitating the teacher's output. However, due to the gap in model capacity between teacher and student, it may be difficult for the student to learn mapping function, which is similar or even identical to the teacher's one. Relational knowledge distillation approach refers to that problem and considers knowledge from another point of view.

2.2 Relational knowledge distillation

Relational knowledge distillation (RKD) methods define the knowledge using a group of objects rather than a single object. Each group of objects forms a structure in representational space, which can be used as a unit of knowledge. In other words, student in RKD methods learns to reproduce structure of teacher's latent space, instead of precise feature representations of objects. To describe relative structure of n input examples relational function ψ, which maps n-tuple of embeddings to a scalar value, is used. Putting $t_i = F_T(x_i)$ and $s_i = F_S(x_i)$, the objective function for RKD is defined as

$$L_{RKD} = \sum_{(x_1, x_2, ..., x_n) \in \chi^n} l(\psi(t_1, t_2, ..., t_n), \psi(s_1, s_2, ..., s_n)). \tag{2}$$

Accordingly to the above equation, the choice of relational function ψ defines certain RKD method. Easiest and the most obvious approach considers pairs of objects and encodes space structure in terms of Euclidean distance between two feature embeddings. Such approach with minor modifications is used in [14] and [16]. Similar idea was recently adapted in [18], where authors use correlation between teacher's and student's outputs as the pairwise relational function. Triplets-based RKD approach was proposed in [14]. Three points in representational space form an angle, and its value can be used to describe structure of the triplet. Another approach, which can also be considered as relational knowledge distillation, although it doesn't precisely follow the equation of RKD loss (2), was presented in [15]. Its main idea is to reformulate knowledge distillation problem as a list-wise learning-to-rank problem, where teacher's list of matching scores is used as ranking to be learned by student.

2.3 Knowledge distillation for Face Recognition

During the first several years of the development of knowledge distillation methods, experiments were carried out mostly on small classification problems. That is why the application of such techniques for face recognition problem hasn't been fully investigated yet, and only few studies have been published in this area.

Some recent works [21, 22] follow the idea of hint connections and impose constraints on the discrepancy between teacher's and student's embeddings. But in order to better fit angular nature of conventional losses used to train face recognition networks [28, 29, 31], authors put penalty on cosine similarity, instead of Euclidean distance. More specific approach, which is oriented to be used in metric learning tasks, was proposed in [19]. This approach utilizes the idea that high-capacity teacher network can better understand subtle differences between images, and uses this observation to adaptively choose margin value in triplet loss function. In [17] authors study knowledge distillation techniques in the context of fully convolutional networks (FCN). They notice that network inference effectiveness can be boosted not only by lowering model complexity, but also by decreasing the size of the input image. Following this idea, authors propose to keep the same FCN architecture and train student on a downsampled version of the original dataset with the help of distillation guidance from teacher's embeddings, computed on high-resolution input.

As can be seen, majority of existing distillation methods for face recognition problem utilize IKD approach, while the effect of RKD hasn't been yet investigated. In this paper, we propose a new relational knowledge distillation technique for face recognition. Our method is inspired by works [14], [16] and [15], and its main idea is to relax objective function (2) in a way that the loss is computed only for those pairs of relational function values, which violate teacher's ranking.

3 Pairwise ranking distillation

Facial recognition systems usually have a gallery of target face images as its component, and each incoming image is compared to it. Gallery image with maximum matching score is further considered to be a candidate for correct match. This leads to the idea that only relative positioning of matching scores is important, rather than their absolute values. In this paper we propose an approach that adapts pairwise ranking techniques for knowledge distillation problem. More specifically, our method considers pairs of relational function values and its goal is to minimize the number of their inversions.

Let $X^T = \{t_i\}_{i=1}^N$ and $X^S = \{s_i\}_{i=1}^N$ be the feature representations computed by teacher and student networks for input batch $X = \{x_i\}_{i=1}^N$ respectively. For both teacher and student we compute values $\Psi^T = \{\psi_i^T\}_{i=1}^M$ and $\Psi^S = \{\psi_i^S\}_{i=1}^M$ of relational function ψ for all possible input n-tuples of feature embeddings. Then the pairwise ranking (PWR) distillation loss is given by:

$$L_{PWR}(X^S, X^T) = \sum_{i,j} \mathbb{1}[\psi_i^T > \psi_j^T] l_{inv}(\psi_i^S, \psi_j^S), \tag{3}$$

where l_{inv} is the function that penalizes pairwise ranking inversions.

As can be seen from the above equation, pairwise ranking knowledge distillation is fully defined by the relational function ψ and the inversion loss function l_{inv}.

3.1 Relational function

In this work, we fix relational function ψ to be the function with two inputs and choose it in a way that value $\psi(x, y)$ characterizes similarity between objects x and y. To be precise, we examined Euclidean distance and cosine similarity as a relational function, and found that cosine similarity performs slightly better[3]. It is worth noting that one can choose any function which describes relationship of set of points in embedding space. For example, RKD-A [14] function, which measures the angle formed by the three objects, is also a valid choice.

3.2 Pairwise inversion loss function

Difference loss The most obvious way to keep the desired ranking of a pair of items is to penalize it as soon as the correct order is violated. For a pair of scalar values (x, y) with ground truth ranking $x > y$, wrong order can be detected by analyzing the difference of the elements: if $y - x$ is greater than zero, elements are misordered. Based on this observation we propose the **difference loss** as a simplest option of a pairwise inversion loss function:

$$l_{inv}(\psi_i^S, \psi_j^S) = max(\psi_j^S - \psi_i^S, 0). \tag{4}$$

In order to make the difference loss more flexible, we add non-linearity in the area of values where misranking happens ($\psi_j^S > \psi_i^S$). This let us to change behaviour of the loss function, and choose where to put more attention — to small or big mistakes. One easy way to add non-linearity to some function is to exponentiate it. This idea results in **power difference loss**:

$$l_{inv}(\psi_i^S, \psi_j^S) = max(\psi_j^S - \psi_i^S, 0)^p. \tag{5}$$

Setting $p > 1$ lowers penalty for marginal mistakes and increases penalty for large ones, while setting $p < 1$ results in opposite function behaviour (see Figure 1). Note that vanilla difference loss (4) is a special case of power difference loss ($p = 1.0$).

Another option to make difference loss non-linear is to put it into the exponential function. We define **exponential difference loss** as:

$$l_{inv}(\psi_i^S, \psi_j^S) = max(exp[\beta(\psi_j^S - \psi_i^S)] - 1, 0). \tag{6}$$

It is similar to power difference loss with $p > 1$, but its β parameter can be chosen so that the loss curve would be more flat (see Figure 2).

[3] It could be explained by the fact that we use angular margin loss function as a base loss to train our face recognition models. However, other RKD methods we compared with don't gain any advantage from cosine similarity relational function.

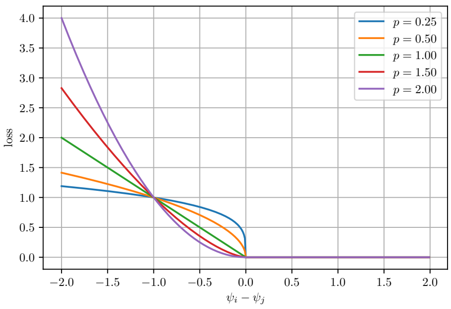

Fig. 1. Power difference loss.

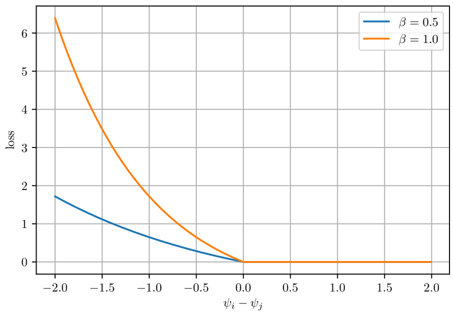

Fig. 2. Exponential difference loss.

Margin The next modification to difference loss we propose is to use a margin term, which is quite common in metric learning tasks [23, 24]. Introducing positive margin not only makes student to learn the same ranking for pairs of objects as teacher has, but also forces the distance between objects to be no less than the margin value. Such modification can be applied to any of the discussed above losses, but for simplicity we consider only the case of vanilla difference loss (4).

The most straightforward approach is to manually choose the margin value and use it throughout the whole training process:

$$\alpha = Const,$$
$$l_{inv}(\psi_i^S, \psi_j^S) = max(\psi_j^S - \psi_i^S + \alpha, 0). \tag{7}$$

Figure 3 depicts how loss curve would look like for different values of margin α.

However, in most cases it is difficult to set the margin so that it won't over-regularize training. To cope with this problem we propose to choose margin dynamically, taking into consideration the scale of objects to be ranked. Specifically, for each batch of objects X we estimate standard deviation of teacher's relational function values and use it as a margin:

$$\alpha_X = std(\Psi_i^T),$$
$$l_{inv}(\psi_i^S, \psi_j^S) = max(\psi_j^S - \psi_i^S + \alpha_X, 0). \tag{8}$$

One more option to choose margin we investigated is also adaptive, but now it's selected individually for each pair of objects. It is also based on values of teacher's relational function, and computed as their difference:

$$\alpha_{ij} = \psi_i^T - \psi_j^T,$$
$$l_{inv}(\psi_i^S, \psi_j^S) = max(\psi_j^S - \psi_i^S + \alpha_{ij}, 0). \tag{9}$$

The idea behind this approach is the following: student learns to preserve order of objects, while keeping the distance between them at least the same as teacher has. From some perspectives, it is similar to the RKD-D approach [14], but now we optimize lower bound of teacher and student difference, instead of forcing student to completely replicate teacher's output.

RankNet for knowledge distillation RankNet [25] is a classical learning-to-rank approach. It formulates ranking as a pairwise classification problem, where each pair is considered independently, and the goal of the method is to miniminize the number of inversions. That perfectly fits our formulation of pairwise ranking distillation, so we adapt RankNet to solve it. For each pair of objects RankNet defines probability of correct ranking and uses cross-entropy as a loss function:

$$P(\psi_i^S > \psi_j^S) = \frac{1}{1 + exp(-\beta(\psi_i^S - \psi_j^S))}, \tag{10}$$

$$l_{inv}(\psi_i^S, \psi_j^S) = -logP(\psi_i^S > \psi_j^S) = log(1 + exp(-\beta(\psi_i^S - \psi_j^S))). \tag{11}$$

As can be seen from Figure 4, RankNet loss function looks like a smooth version of difference loss with margin. Parameter β controls how sharp probability function is, and its increasing results in paying more attention to the area of values, which corresponds to ranking mistakes.

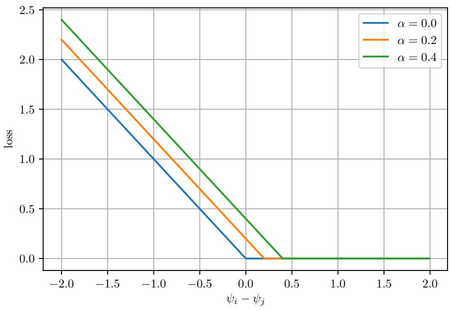

Fig. 3. Difference loss with margin.

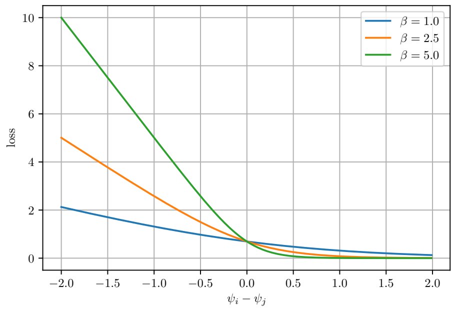

Fig. 4. RankNet loss.

4 Experiments

We evaluate proposed PWR distillation approach on the face recognition task. Throughout this section we refer to PWR with vanilla difference loss (4) as *PWR-Diff*, PWR with exponential difference loss (6) as *PWR-Exp*, and PWR based on RankNet (11) as *PWR-RankNet*. If margin is used, information about it is specified in parentheses. For example, pairwise ranking distillation based on exponential difference loss with adaptive margin computed for each pair of objects, would be named *PWR-Exp (teacher-diff)*.

To demonstrate robustness of the proposed approach, we compare it with other relational knowledge distillation methods. Namely, we consider *DarkRank* [15] and both RKD [14] approaches: distance-based (*RKD-D*) and angle-based (*RKD-A*). Note that knowledge distillation based on the equality of corresponding matching scores between teacher and student was investigated also in [16], but for the sake of simplicity we refer to this approach as *RKD-D* in this section. Regarding DarkRank method, it was noticed in [16] that soft version of DarkRank has numerical stability issues, which lead to severe limitations of batch size that can be used during training. At the same time, authors report that DarkRank-hard demonstrates similar results on a range of metric learning problems, while can be easily computed for any size of batch. That is why in our experiments we use hard version of DarkRank method.

4.1 Datasets

MS-Celeb-1M [30] is used to train all our models. Originally it contains 10 million face images of nearly 100,000 identities. However, due to the fact that the dataset was collected in a semi-automatic manner, significant portion of it includes noisy images or incorrect id labels. That is why we use cleaned version of MS-Celeb-1M provided by [31]. It consists of 5.8 million photos from $85,000$ subjects.

We evaluate trained models on LFW [32], CPLFW [33], AgeDB [34], and MegaFace [35]. The first three datasets employ face verification scenario, while MegaFace provides also evaluation protocol for face identificaion.

Labeled Faces in the Wild (LFW) consists of $13,233$ in-the-wild face images of 5749 identities. Besides images, the list of 6000 matching pairs (3000 positive and 3000 negative) is provided, together with their 10-fold split for cross-validation.

Cross-Pose LFW (CPLFW) uses similar to LFW evaluation protocol with the same total number of comparisons. However, its matching pairs are much more difficult. Faces in positive pairs show substantial pose variations, while negative pairs are constructed using identities of the same race and gender.

AgeDB dataset contains $16,488$ face images from 568 subjects, and also adopts 10-fold cross-validation protocol. This dataset is developed for age-invariant face verification, so all photos have not only identity, but also age labels. In our experiments we follow *AgeDB-30* protocol, where faces in matching pairs have age difference of 30 years. Besides age factor, other facial variations (i.e. pose, illumination, expression) are also included.

MegaFace is the most challenging benchmark in the area to date. It performs evaluation of face recognition algorithms at large-scale distractors. The gallery set of MegaFace includes 1 million images of $690,000$ identities, while the probe set consists of $100,000$ photos of 530 unique identities from FaceScrub [36] dataset. Results for both face identification and face verification are reported.

All faces are aligned by five facial landmarks detected using MTCNN [37] and then cropped to the size of 112×112.

4.2 Experimental setup

In all experiments we use ResNet18 [26] as a student model, and ResNet50 [26] as a teacher model. To obtain face embeddings we append a fully-connected layer on the top of the last convolutional layer. Both teacher and student models have embedding size of 512.

We conduct our experiments using MXNet [27] deep learning framework on a machine with 6 NVIDIA GeForce GTX 1080 Ti GPUs. Batch size is fixed to $552 (92 \times 6)$ for both reference models and student models during knowledge distillation. Stochastic gradient descent (SGD) optimizer is used in all experiments. Learning rate is initially set to 0.1, and during training it is divided by 10 each 2 epochs. The total number of epochs is 13. Baseline models are trained from scratch, while student models in all distillation experiments are initialized with pretrained weights of the baseline model.

Teacher model and baseline student model are trained using CosFace [28, 29] loss. CosFace is an angular margin classification loss, which is widely used in face recognition and other metric learning problems. We compare it with ArcFace [31], another popular angular loss function, and found that CosFace provides slightly better baseline results. Following [29], we set its parameters to be $margin = 0.35$ and $scale = 64.0$.

We found that investigated distillation losses have different convergence abilities for the face recognition task. Specifically, some of them can be used alone to successfully train student network, while others demonstrate sufficient performance only when combined with base classification loss. In addition, we examined whether student performance can be further boosted with the help of HKD (Hinton's Knowledge Distillation) [11] loss. As a result, overall objective function is defined as

$$L = \alpha L_{KD} + \beta L_{CosFace} + \gamma L_{HKD}, \tag{12}$$

where L_{KD} stands for relational distillation knowledge loss (RKD-D, RKD-A, DarkRank, PWR), and α, β and γ are the coefficients of corresponding loss terms.

When CosFace loss is used to stabilize training process of distillation, its weight β is always set to 1.0, and distillation weight α is chosen depending on the type of used L_{KD}. In case if HKD is used, its softmax temperature is set to 4.0, and combination with CosFace is done with $\beta = 0.7$ and $\gamma = 0.3$. Weight of the relational knowledge distillation term α was chosen empirically, following recommendations of original papers. Namely, we set $\alpha = 100$ for RKD-D, $\alpha = 200$ for RKD-A, and $\alpha = 1$ for DarkRank. Concerning our PWR distillation losses, we found $\alpha = 100$ to be a good option for PWR-Diff and PWR-Exp, while for PWR-RankNet it should be smaller (we use $\alpha = 15$).

4.3 Evaluation results

We follow standard protocols for all testing datasets. For LFW, CPLFW and AgeDB-30 verification accuracy, estimated on 10-fold cross validation, is reported. Evaluation on MegaFace dataset includes two protocols, verification and identification. We report *TPR@FPR = 1e-6* for verification, and accuracy at *rank-1* and *rank-10* for identification. Evaluation results are presented in Table 1.

In our experiments we found that RKD-D, RKD-A and DarkRank methods fail to achieve even baseline quality when used alone. That is why for these methods we only report results for experiments when they trained together with base classification loss. On the contrary, proposed PWR approach demonstrates quality improvement when used alone, while adding CosFace and HKD loss terms slightly degrades recognition quality. Therefore, the effect of inversion loss function used in PWR distillation was explored only in such setting.

As can be seen from Table 1, most of the methods demonstrate increase in student accuracy on LFW and AgeDB datasets, however its magnitude is different, especially on AgeDB, where proposed PWR approach beats all other distillation methods by a large margin. At the same time, only one distillation method, *PWR-Exp (teacher-diff)*, managed to boost accuracy on CPFLW dataset. It can be possibly explained by the fact that CPLFW contains images with large pose variations, while faces in the training dataset are mostly frontal, and even teacher's accuracy is relatively low on CPLFW.

Table 1. Experimental results.

Model	**LFW**	**CPLFW**	**AgeDB-30**	**MegaFace**		
				Ver	**Id(1)**	**Id(10)**
Teacher, CosFace	0.9962	0.9205	0.9775	0.9745	0.9698	0.9821
Student, CosFace	0.9943	0.8928	0.9665	0.9497	0.9380	0.9663
CosFace + RKD-D	**0.9947**	0.8895	0.9633	0.9430	0.9281	0.9632
CosFace + RKD-A	**0.9945**	0.8867	0.9652	0.9417	0.9287	0.9619
CosFace + RKD-DA	0.9940	0.8870	**0.9670**	0.9421	0.9280	0.9629
CosFace + DarkRank	**0.9947**	0.8738	**0.9672**	0.9421	0.9321	0.9631
CosFace + HKD	**0.9953**	0.8810	0.9635	0.9366	0.9239	0.9604
CosFace + HKD + RKD-D	**0.9953**	0.8868	**0.9675**	0.9486	0.9352	0.9648
CosFace + HKD + RKD-A	**0.9955**	0.8920	**0.9715**	**0.9526**	**0.9415**	**0.9676**
CosFace + HKD + RKD-DA	**0.9958**	0.8928	**0.9688**	**0.9528**	**0.9404**	**0.9680**
CosFace + HKD + DarkRank	**0.9953**	0.8737	**0.9690**	0.9400	0.9294	0.9618
PWR-Diff (0.1)	**0.9948**	0.8865	**0.9715**	**0.9502**	**0.9393**	**0.9675**
PWR-Diff (teacher-std)	**0.9945**	0.8885	**0.9727**	**0.9540**	**0.9433**	**0.9695**
PWR-Diff (teacher-diff)	**0.9958**	0.8898	***0.9742***	**0.9533**	**0.9433**	**0.9690**
PWR-Exp (0.1)	**0.9955**	0.8907	**0.9710**	**0.9497**	**0.9400**	**0.9681**
PWR-Exp (teacher-std)	**0.9958**	0.8807	**0.9710**	***0.9554***	***0.9439***	**0.9693**
PWR-Exp (teacher-diff)	***0.9963***	***0.8942***	**0.9728**	**0.9538**	**0.9433**	***0.9695***
PWR-RankNet	**0.9953**	0.8870	**0.9715**	**0.9508**	**0.9402**	**0.9674**

Considering MegaFace benchmark results, it's clear that *RKD-D* and *DarkRank* methods can not provide any recognition quality improvement, even when used together with auxiliary losses. Among methods under comparison besides PWR, only combination of *RKD-A* with *HKD* provides better results than the baseline model has. At the same time, all investigated pairwise ranking distillation approaches substantially improve MegaFace recognition quality.

Evaluation results demonstrate that the proposed family of PWR distillation techniques provides methods which outperform other relational knowledge distillation approaches on the face recognition task. However, some modifications of PWR are better than others. For example, considering loss function non-linearity, one can see that in most cases PWR-Exp shows slightly better results than PWR-Diff. As for the margin value, experiments where it was fixed (*PWR-Diff (0.1)*, *PWR-Exp(0.1)*, *PWR-RankNet*) perform worse than those where adaptive margin was used. As a result, we can conclude that the best relational distillation method for face recognition at the moment is *PWR-Exp* with adaptively chosen margin (*teacher-std* or *teacher-diff*).

5 Conclusion

We propose new relational knowledge distillation technique for deep face recognition, which is based on pairwise ranking of matching scores. During training of a student network our PWR approach considers pairs of relational function values and fixes those of them where values are ordered incorrectly, compared to the teacher's ranking. Experiments have proven, that the proposed method significantly outperforms other relational distillation approaches on a range of facial recognition benchmarks.

References

1. Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., Keutzer, K.: SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and $<$ 0.5 MB model size. arXiv preprint arXiv:1602.07360 (2016)
2. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
3. Zhang, X., Zhou, X., Lin, M., Sun, J.: ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6848–6856 (2018)
4. Han, S., Mao, H., Dally, W. J.: Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprint arXiv:1510.00149 (2015)
5. Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., Bengio, Y.: Quantized neural networks: Training neural networks with low precision weights and activations. The Journal of Machine Learning Research **18**(1), 6869–6898 (2017)
6. Han, S., Pool, J., Tran, J., and Dally, W.: Learning both weights and connections for efficient neural network. In: Advances in Neural Information Processing Systems, pp. 1135–1143 (2015)
7. Molchanov, P., Tyree, S., Karras, T., Aila, T., Kautz, J.: Pruning convolutional neural networks for resource efficient inference. arXiv preprint arXiv:1611.06440 (2016)
8. Denton, E. L., Zaremba, W., Bruna, J., LeCun, Y., Fergus, R.: Exploiting linear structure within convolutional networks for efficient evaluation. In: Advances in Neural Information Processing Systems, pp. 1269–1277 (2014)
9. Jaderberg, M., Vedaldi, A., Zisserman, A.: Speeding up convolutional neural networks with low rank expansions. arXiv preprint arXiv:1405.3866 (2014)
10. Ba, J., Caruana, R.: Do deep nets really need to be deep? In: Advances in Neural Information Processing Systems, pp. 2654–2662 (2014)
11. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
12. Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
13. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016)
14. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3967–3976 (2019)
15. Chen, Y., Wang, N., Zhang, Z.: DarkRank: Accelerating deep metric learning via cross sample similarities transfer. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)

16. Yu, L., Yazici, V. O., Liu, X., Weijer, J. V. D., Cheng, Y., Ramisa, A.: Learning Metrics from Teachers: Compact Networks for Image Embedding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2907–2916 (2019)
17. Karlekar, J., Feng, J., Wong, Z. S., Pranata, S.: Deep Face Recognition Model Compression via Knowledge Transfer and Distillation. arXiv preprint arXiv:1906.00619 (2019)
18. Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., Zhang, Z.: Correlation congruence for knowledge distillation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5007–5016 (2019)
19. Feng, Y., Wang, H., Yi, D. T., Hu, R.: Triplet distillation for deep face recognition. arXiv preprint arXiv:1905.04457 (2019)
20. Wang, M., Liu, R., Hajime, N., Narishige, A., Uchida, H., Matsunami, T.: Improved Knowledge Distillation for Training Fast Low Resolution Face Recognition Model. In: Proceedings of the IEEE International Conference on Computer Vision Workshops (2019)
21. Yan, M., Zhao, M., Xu, Z., Zhang, Q., Wang, G., Su, Z.: VarGFaceNet: An efficient variable group convolutional neural network for lightweight face recognition. In: Proceedings of the IEEE International Conference on Computer Vision Workshops (2019)
22. Duong, C. N., Luu, K., Quach, K. G., Le, N.: ShrinkTeaNet: Million-scale lightweight face recognition via shrinking teacher-student networks. arXiv preprint arXiv:1905.10620 (2019)
23. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 539–546 (2005)
24. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
25. Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G.: Learning to rank using gradient descent. In: Proceedings of the 22nd International Conference on Machine learning, pp. 89–96 (2005)
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
27. Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., Xiao, T., Xu, B., Zhang, C., Zhang, Z.: MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems. arXiv preprint arXiv:1512.01274 (2015)
28. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. IEEE Signal Processing Letters **25**(7), 926–930 (2018)
29. 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, pp. 5265–5274 (2018)
30. 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, pp. 87–102. Springer, Cham (2016)
31. 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, pp. 4690–4699 (2019)
32. Huang, G. B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In: Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition (2008)
33. Zheng, T., Deng, W.: Cross-pose LFW: A database for studying cross-pose face recognition in unconstrained environments. In: Beijing University of Posts and Telecommunications, vol. 5, pp. 4873–4882 (2018)

34. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: AgeDB: the first manually collected, in-the-wild age database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 51–59 (2017)
35. Kemelmacher-Shlizerman, I., Seitz, S. M., Miller, D., Brossard, E.: The megaface benchmark: 1 million faces for recognition at scale. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 4873–4882 (2016)
36. Ng, H. W., Winkler, S.: A data-driven approach to cleaning large face datasets. In: IEEE International Conference on Image Processing, pp. 343–347 (2014)
37. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)