WEAKLY SUPERVISED MULTISCALE-INCEPTION LEARNING FOR WEB-SCALE FACE RECOGNITION

Cheng Cheng[⋆]*, Junliang Xing*[†]*, Youji Feng*[⋆]*, Pengcheng Liu*[⋆]*, Xiaohu Shao*[⋆]*, Kai Li*[†]*, Xiang-Dong Zhou*[⋆]

[⋆] Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences
[†] Institute of Automation, Chinese Academy of Sciences

ABSTRACT

Supervised deep learning models like convolutional neural network (CNN) have shown very promising results for the face recognition problem, which often require a huge number of labeled face images. Since manually labeling a large training set is a very difficult and time-consuming task, it is very beneficial if the deep model can be trained from face samples with only weak annotations. In this paper, we propose a general framework to train a deep CNN model with weakly labeled facial images that are easily obtained and freely available on the Internet. Specifically, we first design a deep Multiscale-Inception CNN (MICNN) architecture to exploit the multi-scale information for face recognition. Then, we train an initial MICNN model with only a limited number of labeled samples. After that, we propose a dual-level sample selection strategy to further fine-tune the MICNN model with the weakly labeled samples from both the sample level and class level, which aims to skip outliers and select more samples from confusing class pairs during training. Extensive experimental results on the LFW and YTF benchmarks demonstrate the effectiveness of the proposed method.

Index Terms— Weakly Supervised Learning; Convolutional Neural Networks; Sample Selection; Face Recognition

1. INTRODUCTION

In recent years, deep neural networks, especially the convolutional neural networks (CNNs), have greatly advanced the development of high-performance face recognition models [1, 2, 3, 4, 5, 6]. The typical models include the FaceNet model [1], the DeepFace model [2], the DeepID nets [4, 5, 6]. The DeepID nets from the CUHK team in [4, 5, 6] are trained with around 290,000 face images from 12,000 identities. The FaceBook team in [2] trains the DeepFace model on a identity labeled dataset of four million facial images belonging to more than 4,000 identities. The FaceNet model trained by the Google team in [1] is trained on 200 Million photos of 8 Million people. All the above mentioned papers use the large scale datasets to train the Deep CNN models.

It is noteworthy that labeling a large set of facial images by hand is a laborious task and almost impractical in practice. Fortunately, due to the increasing pervasiveness of social network and mobile industry, nowadays tremendous amount of images are uploaded every day, and most of them can be easily collected from image search engines or photo sharing websites. A direct approach is to use the related labels that can be obtained easily, such as user tags from social networks, or keywords from image search engines. However, these noisy labels are not reliable, and contain much misleading information which will subvert the model during training.

In order to utilize the abundant weakly labeled data for deep model learning, we propose a general framework for training deep CNN models in a weakly supervised fashion in this paper. First, we design a deep Multiscale-Inception CNN (MICNN) architecture, which exploits multi-scale information for face recognition by integrating hierarchical representations. Then, an initial MICNN model is trained with a limited number of fully labeled data. Based on the initial model, a dual-level training sample selection strategy is employed to further fine-tune the MICNN model with the weakly labeled data. At the sample level, we model the relationships between the fully labeled samples and the weakly labeled samples, which is used for outlier detection. At the class level, a confusion matrix is estimated from current MICNN model, and is used to pick up more confusable class pairs for training. Extensive experimental evaluations on the LFW dataset and the YTF dataset demonstrate that the proposed method achieves the state-of-the-art performance for face recognition.

To summarize, in this work we have made the following three main contributions:

- We design a Multiscale-Inception modular within the CNN framework to form a deep architecture which can better incorporate the multi-scale information of the input image.
- We propose a dual-level training sample selection strategy which exploit the information from both fully labeled samples and weakly labeled samples to perform outlier removal and hard negative mining.
- We deploy our deep architecture for face recognition, and achieve state-of-the-art performances on two challenging face recognition datesets, the LFW dataset and the YTF dataset.

2. OUR APPROACH

We aim to build a general framework that enables deep CNN model to be trained more effectively with both a limited number of fully labeled samples and abundant weakly labeled data. The overall idea mainly consists of two steps. First, we obtain an initial Multiscale-Inception CNN (MICNN) model pre-trained with only a limited number of fully labeled samples (Section 2.1), which extensively exploit multi-scale information for face recognition. Then, we present a simple but effective dual-level sample selection strategy to further train / fine-tune the MICNN model with both weakly labeled and fully labeled face images (Section 2.2).

2.1. Deep Multiscale-Inception CNN Architecture

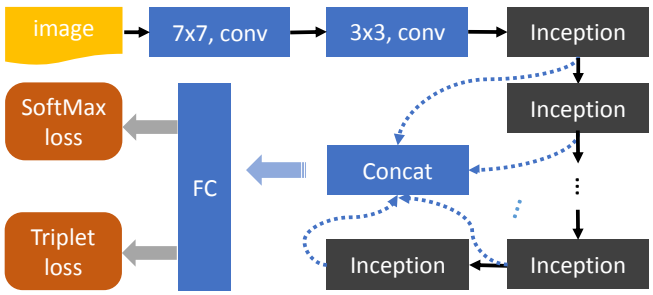

Fig. 1: The architecture of the proposed MICNN model. Multi-scale information are extensively exploited from multiple Inception modular.

It is well-known that the multi-scale information plays an important role in visual recognition [7]. Based on the Inception modular first proposed in GoogleNet [8], we design a deep Multiscale-Inception CNN architecture which extensively exploit multi-scale information for face recognition. Concretely, the proposed method learns an MICNN model by simultaneously minimizing the SoftMax loss for classification and the Triplet loss for ranking, which are complementary with each other for maximizing the inter-class variations and minimizing the intra-class variations. The multi-scale information is concatenated into a single layer and fed through both the identity classification signal and the similarity ranking signal as supervision. The MICNN architecture totally consists of 2 convolutional layers, 9 Inception layers, 5 pooling layers, 3 fully connected layers, 1 similarity ranking loss layer, and 1 SoftMax loss layer. The first four pooling layers use max operator and the last pooling layer is average operator. The outputs of the 9 Inception layers are added to the outputs of the last fully connected layer. Following [9], the batch normalization is used after each convolutional layer and before the ReLU activation layer. We train an initial MICNN model with a limited number of fully labeled samples.

Assuming that we have a set of training samples $\{d_i = (x_i, y_i)\}_{i=1}^{n}$, where $x_i \in R^m$ is associated with its class label $y_i \in \{1, 2, ..., c\}$, and c is the number of different classes. The hybrid identity classification and similarity ranking loss is formulated as:

$$\mathcal{L}(d_1, d_2, d_3) = \mathcal{L}_{cls}(d_1, d_2, d_3) + \lambda * \mathcal{L}_{rank}(d_1, d_2, d_3), \quad (1)$$

where λ is a weighting parameter between the identity classification $\mathcal{L}_{cls}$ signal and the similarity ranking signal $\mathcal{L}_{rank}$, which are respectively defined as follows,

$$\mathcal{L}_{cls}(d_1, d_2, d_3) = \text{cls}(x_1, y_1) + \text{cls}(x_2, y_2) + \text{cls}(x_3, y_3), \quad (2)$$

$$\mathcal{L}_{rank}(d_1, d_2, d_3) = \max(||x_1 - x_2||_2^2 - ||x_1 - x_3||_2^2 + \Delta, 0). \quad (3)$$

The identity classification loss $\mathcal{L}_{cls}(d_1, d_2, d_3)$ is defined on a triplet of samples, with the classification signal $\text{cls}(x_i, y_i) = -\log p_i$ as the standard cross-entropy / log loss, and p_i is the target probability distribution. Here $p_i = 0$ for all i except $p_i = 1$ for the target class y_i. The similarity ranking loss $\mathcal{L}_{rank}(d_1, d_2, d_3)$ is defined over a number of triplets [1]. Each triplet contains three image samples x_1, x_2 , and x_3. Among them, x_1 and x_2 are two face images samples from the same person while x_3 represents one face sample from a different person, Δ is a parameter that regularizes the gap between the distance of the two sample pairs, and is set to 0.2 in all our experiments. This loss encourages the distances between samples from the same person to take smaller values than that those from different persons.

2.2. Dual-level Training Sample Selection Strategy

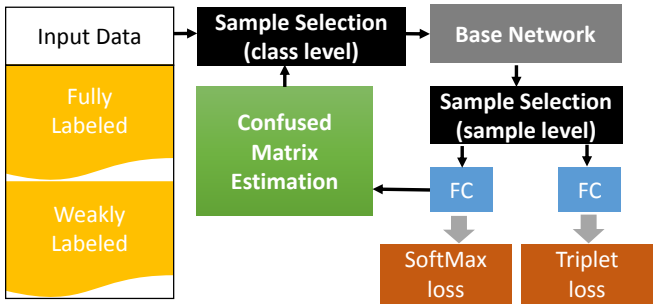

Fig. 2: The framework of the proposed dual-level sample selection strategy.

Since there are many noises in the weakly labeled training dataset, sample selection is a key step for effectively training the Deep CNN models with this kind of training samples. In addition, sample selection is also an effective way to speed up the convergence of the deep model learning. In this section, we propose a simple yet effective online high quality sample mining framework for further training / fine-tuning of our initial MICNN model with only a limited number of fully labeled samples and a large number of weakly labeled data. The overall sample selection strategy is illustrated in Figure 2. We conduct the high quality example selection process at both the sample level and the class level.

2.2.1. Sample level

In contrast to the traditional training strategies for the deep CNN models, which randomly select the training samples, we select M classes (each class contains N images, $M = M_c + M_w$) as input for the model training in each batch, where M_c is the number of classes belong to the fully labeled dataset, and M_w is the number of classes belong to the weakly labeled dataset. The dataset containing the selected images from class i is denoted as S_i. We set $M_c = 10, M_w = 40, N = 5$ in all the experiments.

Let p and q represent two face images, $f(p)$ is a 128-dimensional feature, which is obtained via the MICNN model in current training iteration. Let $d_t(p, q)$ denotes the Euclidean distance between $f(p)$ and $f(q)$, c_p denote the class label that image p belongs to. Then the *weight* of training sample p in the training data is defined as:

$$\omega(p) = \frac{1}{N} \sum_{q \in S_{c_p}} d_t(p, q). \tag{4}$$

After selecting the fully labeled data in each batch, we turn to calculate a threshold d_c, which is defined as:

$$d_c = \frac{1}{M_c * N} \sum_{i=1}^{M_c} \sum_{p \in S_i} \omega(p). \tag{5}$$

During the training process, these weakly labeled samples whose *weight* is bigger than $2 * d_c$ is ignored as an outlier. After that, suppose that each class remains N' images, then the number of possible triplets is $N' * M * (M-1) * N' * (N'-1)$, and we pick the top 20% hard part for training.

2.2.2. Class level

To improve the convergence speed of the model training, we enforce to put the easily confused class pairs into one batch. The confusion matrix ($f_{ij} \in R^{c*c}$) is estimated from the training data with the current MICNN model, and will be used in the second round training, as normally done in bootstrapping. The confusion degree between class i and class j is defined as:

$$f_{ij} = \begin{cases} 0 & i = j, \\ \dfrac{N_{i \rightsquigarrow j}}{N_i} & i \neq j, \end{cases} \tag{6}$$

where N_i is the number of samples in class i, and $N_{i \rightsquigarrow j}$ is the number of samples that come from class i but predicted into class j by the current MICNN model. In fact, since the top-1 error rate usually equals to zero, top-2 results are used for estimate the confusion matrix. It is worth noting that this sample selection strategy needs little extra operational cost, hence it is quite effective.

Table 1: Recognition rates (%) on the LFW dataset using the BLUFR protocol.

#	training data	sample selection	network	VR (%) @ FAR = 0.1%	DIR (%) @ Rank = 1, FAR = 1%
1	CIGIT	no	GoogleNet	95.55	76.89
2	CIGIT+MS-Celeb+	no	GoogleNet	fail	fail
3	CIGIT+MS-Celeb+	yes	GoogleNet	98.05	81.99
4	CIGIT+MS-Celeb+	yes	MICNN	98.44	88.48

3. EXPERIMENTS

In this section, we first introduce our employed datasets and experimental settings. Then we analyze several variations of the presented model. At last, we compare the performance of our proposed model with several state-of-the-art methods on the LFW dataset and the YTF dataset.

3.1. Datasets and Experimental Settings

The CIGIT-celebrity + MS-Celeb+ dataset is a hybrid dataset we construct from the Internet and the MS-Celeb-1M [10] to train the MICNN model. The CIGIT dataset contains 1,803,848 images of 21,506 celebrities, and labeled by humans. The MS-Celeb+ dataset is extended from MS-Celeb-1M [10], which contains 172,216 identities of 21,671,111 face images. The CIGIT dataset is fully labeled, while MS-Celeb+ is weakly labeled. The face samples in the dataset are aligned to the input size of 128x128 pixels. We train the MICNN model using the stochastic gradient descent (SGD) with momentum 0.9. We implement the network using Caffe [11] with mini-batch size 256. The learning rate is to set to 1e-1, and reduced to 1e-5 gradually. The models are learned from sketch, and trained on eight Titan X GPU for 300 hours. We learn the Joint Bayesian [12] model for face verification. The responses of the full-connection layer following the last pool layer is extracted to serve as the face representation. Face images from 2000 identities randomly sampled from CIGIT are used to learn the Joint Bayesian model.

The LFW dataset [13] contains 13,233 face images of 5,749 identities collected from the Internet. We conduct experiments on the LFW database using the both standard protocol [13] and the recently proposed BLUFR protocol [14]. The standard protocol divides the LFW dataset into 10 folds of mutually exclusive people set. For the unrestricted setting, performance is evaluated using the 10-fold cross-validation. Each split contains 600 image pairs which were predefined by LFW. Mean accuracy and standard error were reported. The BLUFR protocol defines 10-fold cross validation for face verification tests. In each trial, the test set contains 9,708 face images of 4,249 subjects, on average. As a result, over 47 million face comparison scores need to be computed. Following the protocol in [14], we report the verification rate (VR) at FAR = 0.1 % for the face verification and the detection and identification rate (DIR) at Rank-1 corresponding to an FAR of 1% for open-set identification.

Table 2: Results on the LFW dataset using standard protocol with unrestricted setting.

Method	Network Number	Mean accuracy
DeepFace [2]	1	95.92
DeepFace [2]	7	97.35
Li *et al.*. [16]	1	97.73
DeepID2 [5]	1	95.43
DeepID2 [5]	25	99.53
Deep Face Recognition [3]	1	98.95
FaceNet [1]	1	99.63
Proposed MICNN model	1	99.28

The YTF dataset [15] contains 3,425 YouTube videos of 1,595 subjects, and an average of 2.15 videos are available for each subject. In our experiments, we follow the standard protocol [15] on this dataset to report the average accuracy with cross validation.

3.2. Component Analysis

We evaluate the effect of different components within the proposed MICNN model on the LFW dataset using the BLUFR protocol [14], which is a more strict and effective metric as suggested in [14]. The experiment settings are listed in Table 1. Different methods employ different training data. We use only the fully labeled CIGIT data to get the baselines under strong supervisions. On the other hand, when all the data are used, the dual-level sample selection strategy is employed as discussed in Section 2.2.

We first study the effect of massive weakly labeled data. From row $\#1$, $\#2$ and $\#3$ we can see that training a DCNN model with only small amount of fully labeled data, the recognition performance is inferior. After adding the massive weakly labeled data, which are crawled from the Internet, the DCNN model could not get convergence. By using the proposed sample selection strategy to the model training, the face recognition performance is remarkably improved, which verify the efficacy of the proposed method. Finally, we also test the GoogleNet and the MICNN model by employing the sample selection strategy. By comparing row $\#3$ and $\#4$, we find that the proposed MICNN model is better than the baseline GoogleNet model.

3.3. Comparison with the State-of-the-arts

3.3.1. Results on the LFW dataset using standard protocol

We first compare the presented approach with several state-of-the-arts methods on the LFW dataset using the standard protocol under the unrestricted setting. These compared methods include FaceNet [1], DeepFace [2], Deep Face Recognition [3], and DeepID2 [5]. Based on the experimental results shown in Table 2, it can be observed that our single MICNN model achieve very competitive result to the state-of-the-arts, which either employs more fully labeled training samples [1] or performs multiple model fusion [5].

3.3.2. Results on the LFW dataset using the BLUFR protocol

Since the experimental results on the LFW dataset using standard protocol are already saturated, we also further compare our approach with the state-of-the-arts methods which are reported on the LFW database using the BLUFR protocol. The experimental results are listed in Table 3, from which we can see that the verification rate at FAR=0.1% of our model reaches 98.44%, the detection and identification rates at FAR=1% reaches 88.48%, which surpasses all the state-of-the-art algorithms (see Table 3).

Table 3: Experimental results on the LFW dataset using the BLUFR protocol.

Method	VR@FAR=0.1%	DIR@% Rank=1 FAR=1
Yi *et al.*.[14]	80.26	28.90
Wang *et al.*. [17]	87.65	46.31
Lv *et al.*. [18]	-	63.73
BJB [19]	93.05	-
Proposed MICNN model	**98.44**	**88.48**

3.3.3. Experiment results on the YTF dataset

To further evaluate the effectiveness and generalization ability of our proposed MICNN model, we finally compare the performance of the proposed approach to the state-of-the-art results reported on the YTF dataset, which are summarized in Table 4. Five methods, FaceNet [1], the DeepFace [2], Deep Face Recognition [3], DeepID2 [5] and NAN [20] report the best performances on the YTF dataset to date. From Table 4 it can be observed that our MICNN model achieves 97.41% recognition accuracy, which is the currently best results on this challenging dataset. These results well verify the superiority of our proposed MICNN model.

Table 4: Experimental results on the YTF dataset using the standard metric proposed in [15].

Method	Accuracy (%)	AUC
DeepFace [2]	91.40 ± 1.1	96.30
DeepID2 [5]	93.20 ± 0.2	-
FaceNet [1]	95.12 ± 0.39	-
NAN [20]	95.52 ± 0.06	98.70
Deep Face Recognition [3]	97.30	-
Proposed MICNN model	$\mathbf{97.41 \pm 0.06}$	**99.20**

4. CONCLUSIONS

In this work, we have presented a deep architecture for the task of face recognition. We have designed a deep Multiscale-Inception architecture that extensively exploit multi-scale information into an end-to-end deep learning system. We have also proposed a dual-level sample selection strategy, which is shown to be robust against noisy data. Experiments on the LFW dataset and YTF dataset have verified the efficiency and effectiveness of the proposed deep architecture.

5. REFERENCES

[1] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[2] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[3] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015, pp. 6–18.

[4] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation from predicting 10,000 classes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1891–1898.

[5] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation by joint identification-verification," in *Neural Information Processing Systems*, 2014, pp. 1988–1996.

[6] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang, "Deepid3: Face recognition with very deep neural networks," in *arXiv*, 2015.

[7] Pierre Sermanet and Yann LeCun, "Traffic sign recognition with multi-scale convolutional networks," in *International Joint Conference on Neural Networks*, 2013, pp. 1–6.

[8] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, and Jonathon Shlens, "Rethinking the inception architecture for computer vision," in *arXiv*, 2015.

[9] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 1875–1882.

[10] Yandong Guo, Lei Zhang, Yuxiao Hu, Yuxiao Hu, and Jianfeng Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European Conference on Computer Vision*, 2016, pp. 87–102.

[11] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM International Conference on Multimedia*, 2014, pp. 675–678.

[12] Dong Chen, Xudong Cao, Liwei Wang, Fang Wen, and Jian Sun, "Bayesian face revisited: A joint formulation," in *European Conference on Computer Vision*, 2012, pp. 566–579.

[13] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.

[14] Shengcai Liao, Zhen Lei, Dong Yi, and Dong Yi, "A benchmark study of large-scale unconstrained face recognition," in *International Conference on Biometrics*, 2014, pp. 1–8.

[15] Lior Wolf, Tal Hassner, and Itay Maoz, "Face recognition in unconstrained videos with matched background similarity," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 529–534.

[16] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li, "Learning face representation from scratch," in *arXiv*, 2015.

[17] Dayong Wang, Charles Otto, and Anil K. Jain, "Face search at scale: 80 million gallery," *Technical Report*, 2015.

[18] Jiang-Jing Lv, Cheng Cheng, Guo-Dong Tian, Xiang-Dong Zhou, and Xi Zhou, "Landmark perturbation-based data augmentation for unconstrained face recognition," *Signal Processing: Image Communication*, In Press.

[19] Cheng Cheng, Junliang Xing, Youji Feng, Deling Li, and Xiang-Dong Zhou, "Bootstrapping joint bayesian model for robust face verification," in *International Conference on Biometrics*, 2017.

[20] Jiaolong Yang, Peiran Ren, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua, "Neural aggregation network for video face recognition," *arXiv*, 2016.