HIERARCHICAL TRAINING FOR LARGE SCALE FACE RECOGNITION WITH FEW SAMPLES PER SUBJECT

Yuhao Ma[1,2] *Meina Kan*[1,3] *Shiguang Shan*[1,3] *Xilin Chen*[1]

[1] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences(CAS),
Institute of Computing Technology, CAS
[2]University of Chinese Academy of Sciences
[3]CAS Center for Excellence in Brain Science and Intelligence Technology
yuhao.ma@vipl.ict.ac.cn,{kanmeina,sgshan,xlchen}@ict.ac.cn

ABSTRACT

Recent progress of face recognition benefits a lot from large-scale face datasets with deep Convoluitonal Neural Networks(CNN). However, when dataset contains a large number of subjects but with few samples for each subject, conventional CNN with softmax loss is heavily prone to overfitting. To address this issue, we propose a hierarchical training schema to optimize CNN with coarse-to-fine class labels, referred to as Hit-CNN. Firstly trained with coarse class labels and then refined with fine class labels, Hit-CNN is enabled the to capture the distribution of data from major variations to fine variations progressively, which can effectively relieve the overfitting and lead to better generalization. In this work, the hierarchical coarse-to-fine class labels are obtained via hierarchical k-means clustering according to the face identities. Evaluated on two face datasets, the proposed Hit-CNN provides better results compared with the conventional CNN under the circumstances of large-scale data with few samples per subject.

***Index Terms*—** face recognition, large-scale training, few samples

1. INTRODUCTION

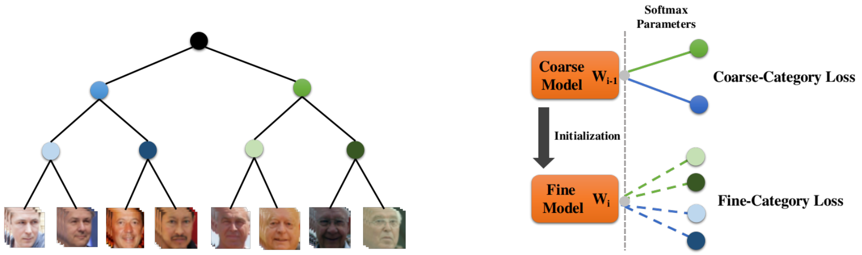
In recent years, the Convolutional Neural Networks(CNN) have achieved great success in various computer vision tasks such as image classification [1],object detection [2], etc. In the field of face recognition, CNNs have also made a breakthrough and even surpassed human-level performance [3, 4]. This remarkable advancement benefits from the high non-linear capacity of CNN models and its ability to learn from massive training data. Large-scale datasets are crucial to the effectiveness of CNN models. However, for face recognition task, collecting such large-scale datasets with a great number of identities and adequate samples for each identity is time consuming and financial challenging. In many cases, it is more likely to collect a dataset that consists of enough subjects, but with few samples for each subject. In this work, we focus on the training issues on such large scale face dataset with few samples per subject.

Applying CNN with softmax loss is the most common way for face or the general object recognition. It works well when the number of classes is relatively small(e.g.1K $\sim$ 10K), but when the scale grows to tens or even hundreds of thousands, training may encounter some difficulties. Firstly, GPU memory footprint can be unbearable, especially when large batch size is preferred for better convergence. Secondly, the size of parameters for softmax classification grows linearly to the number of classes, the massive weight parameters in the fully-connected layer can easily lead to overfitting. The overfitting becomes more serious when there are only several samples per subject, which implies less intra-class variations.

Another disadvantage of flat softmax classifier is that it doesn't fully utilize the inter-class similarity. It is ordinary to find that some categories are more similar thus harder to distinguish. As the number of classes become larger, visually similar pairs are more likely to appear. Treating them evenly with flat softmax may push the solution away from the global optimum.

In this work, we propose a hierarchical training schema for face recognition tasks with few samples per subject. Firstly, all training samples are clustered hierarchically via k-means clustering w.r.t. the identities. According to this hierarchy, the CNN is firstly trained with the coarse category labels, then with finer category labels. As more samples are contained in each coarse category, more variations are brought against overfitting, resulting in a pre-trained model with better generalization. Thus the optimization of the succeeding fine training is well regularized to achieve better generalization as well. Experiments are performed on two datasets, MegaFace [5] containing about 200K subjects with 3 samples per subject, and a newly collected dataset consisting of about 94K subjects with 5 samples per subject. The proposed Hit-CNN significantly outperforms conventional CNN, demonstrating its effectiveness of handling the problem of large-scale training with few samples per subject.

978-1-4799-7061-2/18/$31.00 ©2018 IEEE



(a) Hierarchical Clustering

(b) Hierarchical Training of CNN

Fig. 1. Overview of Hit-CNN. (a) The hierarchical clustering to obtain coarse-to-fine categories, nodes marked with similar colors contain subjects with higher visual similarity. (b) Training with coarse category labels for better generalization and fine category labels for accurate optimization. Hierarchical training strategy can induce a CNN with better generalization in the scenario of large scale dataset with few samples per subject.

2. RELATED WORKS

In this section, we review the works relate to large-scale training for face recognition and hierarchical learning.

Large-scale training for face recognition. For face recognition tasks, large-scale datasets are crucial. The works of triplet loss [3] and pair-wise loss [6] are two representative metric learning based methods. They bypass softmax by optimizing embeddings directly, which makes them proper alternatives to learn face representation when the number of classes is large. However, how to select informative triplets or pairs still remains challenging. Besides, the huge number of candidates usually causes higher or even unbearable cost of training time [7]. For specific scenario of large scale subjects but with few samples per subject, a straightforward solution is to apply data augmentation, which usually requires domain specific knowledge such as expression altering and pose warping [8]. Our method has no conflicts with this approach and can be applied as complement.

Hierarchical learning. Hierarchical structure is a straightforward choice to organize data based on their relationships. The hierarchy can be either predefined [9] or learned by different approaches [10, 11]. Various tasks such as image classification [12, 13] and image retrieval [14] benefit from the prior knowledge in the hierarchy. However, training with flat softmax is incapable of making use of it. Some recent works have introduced category hierarchy into CNN-based methods. In [11], the flat softmax classifier is replaced with several subnets with fine-grained softmax classifiers according to the hierarchy. This approach has successfully improved the classification performance, however it can't be simply applied to face recognition problem. In face recognition, usually the identities for testing and training do not overlap, thus similarity of two faces is generally computed as the distance of the feature from the fully-connected layers rather than the output of softmax layer in [11].

3. METHODLOGY

3.1. Overview

Our hierarchical training method includes two stages as shown in Fig.1. The first stage is the initial stage which achieves the category hierarchy of subjects by using the simplest hierarchical k-means clustering. According to the hierarchical clustering results, the coarse-to-fine class labels of each identity is obtained.

In the second stage, CNN model is firstly trained with coarse category labels, then with finer and finer category labels, forming a progressive manner. All the steps of CNN training except the last one employs the coarse category labels, leading to a pre-trained model which is not accurate but with much better generalization. Finally, the CNN is fine-tuned with the intrinsic class labels, leading to an accurate model still with favorable generalization.

Definition of notations: N denotes the number of the levels of the hierarchy, C_i denotes the number of categories at the i-th level. C_N equals the number of intrinsic labels.

3.2. Learning Category Hierarchy

In this work, the category hierarchy is established according to the visual similarity to characterize the semantic relationship between subjects. Simply, a top-down k-means clustering is adopted. For each subject, the average feature of all samples, as an unified representation of this subject, is used for clustering. The feature for face samples can be either RGB values or features extracted from a CNN face descriptor. In this work, the CNN we use is a model trained either on the original training set with few samples per subject(i.e. the baseline), or an auxiliary set with many samples per subject.

Firstly, the coarse categories at the first level are obtained by applying k-means on the C_N subject features with $k = C_1$.

Then the same clustering is performed within every coarse category recurrently. For each cluster at the $(i-1)$-th level, k-means is applied to get several sub-clusters at the i-th level(C_i in total). At the N-th level, the leaf node contains one single identity. Formally, for a training sample x_k, its class label is denoted as y_k, its coarse class label at i-th level is denoted as $y_k^i \in \{1, 2, \cdots, C_i\}$, then in the N-th level, $y_k^N = y_k$.

The learning of hierarchy is straightforward and simple, without elaborate design. This is because that all the coarse class labels of the hierarchy are only used to obtain a coarse model as a guidance or initialization, so a rough hierarchy can achieve comparable results as an elaborate one.

3.3. Hierarchical Training

In our hierarchical training schema, the CNN is firstly trained from scratch with the coarsest class labels with softmax loss:

$$\min_{W_1} \sum_k -\log p(y_k^1|x_k) \tag{1}$$

where W_1 are the parameters of the CNN optimized by using the coarse class labels at the 1-st level, and $p(y_k^1|x_k)$ is the output of x_k on the y_k^1 class node.

Unlike intrinsic categories containing only few samples, the coarse categories contain a large collection of visually similar samples with richer variations. The visual similarity makes it possibleto distinguish different coarse classes, meanwhile the adequate samples and the intra-class variety somehow relieve the overfitting problem, providing a better initialization for further training.

Recurrently, with the pre-trained model W_{i-1} as initialization, the model W_i is further fine-tuned with the finer coarse class labels at the i-th level as below:

$$\min_{W_i} \sum_k -\log p(y_k^i|x_k) \tag{2}$$

Finally, the CNN model is fine-tuned with the intrinsic class labels (i.e.$y_k^N = y_k$) based on the model pre-trained with the hierarchical coarse labels as initialization:

$$\min_{W_N} \sum_k -\log p(y_k|x_k) \tag{3}$$

As a result, the parameter of the CNN is updated from W_1, to W_2, then to W_i and finally to W_N . In this process, the category labels used for the optimization are from coarse to fine, which leads the parameter of CNN from a random point to more and more accurate points. This hierarchical training strategy enables the CNN to capture the distribution of samples from major variations to fine variations progressively. The preceding training with coarse class labels well relieves the overfitting problem as more samples are included in one class, providing a good initialization for succeeding training. This initialization can be considered as a kind of progressive regularization which can induce better generalization of the fine-tuned model.

The coarse-to-fine hierarchical training process is illustrated in Fig.1(b). While fine-tuning from model W_{i-1}, parameters before softmax (on the left side of the gray dashed line) of W_i are directly copied from model W_{i-1}, as they share the same structure. However, the size of parameters in the softmax does not match. The softmax parameters of the fine categories at $i-th$ level are copied from its corresponding super class at the $(i-1)-th$ level (the dashed lines are copied from the corresponding solid line with the same color).

4. EXPERIMENTS

To evaluate the effectiveness of the proposed method, we compare the Hit-CNN to the vanilla CNN on two large scale dataset, the MegaFace with 200K subjects and a privately collected dataset with 94K subjects for training. Commonly, the testing set contains a target set and a query set for verification. The ROC curve w.r.t true positive rate(TPR) and false positive rate(FPR) is employed to report the performance.

4.1. Evaluation on MegaFace

We firstly evaluate our method on the challenging MegaFace [5], which is recently released for large scale face recognition. These photos are with wide variations such as pose and age. The training set of MegaFace contains 672K identities with about 4 million images, but not all the subjects satisfy the condition of few images. So, we re-sample it to formulate a novel protocol for large scale training with few samples per subject. Specifically, 200K identities with 3 images per subject are randomly selected as training set, 2K subjects are randomly selected as the testing set, including 2K images as the target set and 4K images as the query set. Besides, two more benchmarks are also used as alternatives for evaluation. The first is LFW [15] with 6K face pairs for verification. The second is MegaFace Challenge [16], in which 1 million images are used as distraction set to identify the images from the probe set FaceScrub [17].

Experiment Settings. On this dataset, the baseline CNN model is a 34-layer ResNet [18], which is directly trained on 200K identities. For our proposed Hit-CNN, the hierarchical clustering is performed to get a 4-level hierarchy with 0.2K,2K,20K,200K clusters at each level respectively. The feature extractor for clustering is an auxiliary CNN trained on MS-Celeb-1M [19]. For intensive evaluation of the coarse-to-fine strategy, Hit-CNN with several different settings are evaluated: 1) Hit-CNN(0.2K-200K), in which the CNN is firstly trained with 200 coarse categories and then refined with the intrinsic 200K subjects; 2) Hit-CNN(0.2K-2K-200K); 3) Hit-CNN(0.2K-2K-20K-200K). For fair comparison, the baseline CNN and Hit-CNNs are optimized with the same learning rate

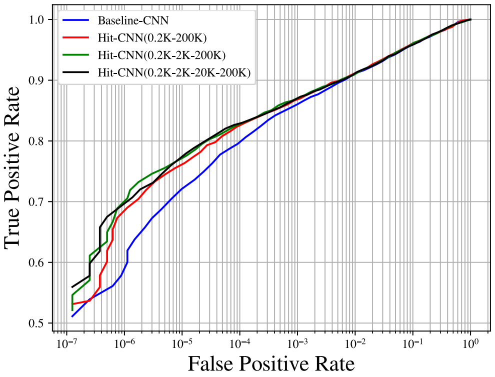


Fig. 2. The evaluation of baseline CNN and our Hit-CNN on MegaFace in terms of ROC curves.

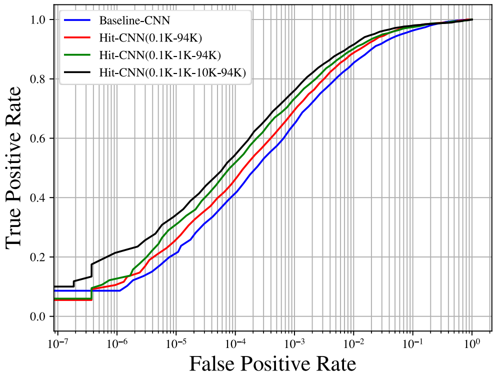


Fig. 3. The evaluation of baseline CNN and our Hit-CNN on TCID in terms of ROC curves.

policy and iterations. On LFW benchmark the performance is reported in terms of mean average precision of verification following the standard protocol. On MegaFace Challenge benchmark, rank-1 identification accuracy for probe images is evaluated.

Results. The evaluation results are displayed in Fig.2 and Table 1. As can bee seen, Hit-CNNs perform much better than the baseline CNN on both identification and verification tasks. Moreover, Hit-CNN with deeper hierarchy performs better, which demonstrates that the progressive training with coarse-to-fine labels can effectively improve the performance and generalization of the baseline CNN in the case of large scale dataset with few samples per subject.

Model	Accuracy(%)	
	LFW	MegaFace
Baseline-CNN	96.25	26.57
Hit-CNN(0.2K-200K)	97.55	34.27
Hit-CNN(0.2K-2K-200K)	97.80	39.22
Hit-CNN(0.2K-2K-20K-200K)	97.92	40.26

Table 1. The evaluation of the baseline CNN and our Hit-CNN on LFW in terms of mAP and on MegaFace Challenge in terms of rank-1 accuracy.

4.2. Evaluation on a New Dataset for E-payment

Dataset. In recent year, face recognition is widely used in many fields such as e-payment to verify identity. For this practical scenario where usually only several images are available for each subject, we collect a private dataset to evaluate our method. This dataset, referred to as TCID, consists of 94K subjects with 5 samples per subject. For each subject, two images are certificate photos and three images are collected from real-world scenarios with variations of pose, illumination and expressions. For evaluation, about 2K subjects are used as testing set, with one certificate photo as target, and one or two wild photos as query.

Experiment Setting. The baseline CNN model is also the 34-layer ResNet [18] directly trained on the 94K identities. The hierarchical clustering is performed to get a 4-level hierarchy with 0.1K,1K,10K,94K clusters at each level respectively. The feature extractor for clustering is the baseline CNN. Hit-CNN with several different settings are evaluated: 1) Hit-CNN(0.1K-94K); 2) Hit-CNN(0.1K-1K-94K); 3) Hit-CNN(0.1K-1K-10K-94K). For fair comparison, the baseline CNN and Hit-CNNs are optimized with the same learning rate policy and iterations.

Results. The evaluation results are shown in Fig 3. As expected, Hit-CNN outperforms the baseline CNN with significant improvement. Similar as that on MegaFace, Hit-CNN with deeper hierarchy performs better, which demonstrates the effectiveness of our coarse-to-fine training strategy.

5. CONCLUSIONS AND FUTURE WORKS

In this work, we propose a coarse-to-fine hierarchical training method for CNN under the circumstance of large-scale subjects but with few samples per subject. Hit-CNN is trained with coarse-to-fine class labels progressively. The CNN is firstly trained with coarse category labels leading to an initialization with better generalization, which is further fine-tuned with the intrinsic category labels, inducing a more accurate model for face recognition. The evaluation on two challenging large-scale dataset demonstrates its effectiveness. In the future, we will consider to combine our hierarchical training with other strategies such as data augmentation.

6. ACKNOWLEDGEMENTS

This work was partially supported by National Key R&D Program of China under contracts No.2017YFA0700804 and Natural Science Foundation of China under contracts Nos.61650202,61772496,61402443 and 61532018.

7. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems(NIPS)*, 2012.

[2] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[3] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[4] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[5] Aaron Nech and Ira Kemelmacher-Shlizerman, "Level playing field for million scale face recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[6] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation from predicting 10,000 classes," in *The IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2014.

[7] Chong Wang, Xue Zhang, and Xipeng Lan, "How to train triplet networks with 100k identities?," *arXiv preprint arXiv:1709.02940*, 2017.

[8] Iacopo Masi, Anh Tun Trn, Tal Hassner, Jatuporn Toy Leksut, and Gérard Medioni, "Do we really need to collect millions of faces for effective face recognition?," in *European Conference on Computer Vision(ECCV)*, 2016.

[9] Yangqing Jia, Joshua T Abbott, Joseph L Austerweil, Thomas Griffiths, and Trevor Darrell, "Visual concept learning: Combining machine vision and bayesian generalization on concept hierarchies," in *Advances in Neural Information Processing Systems(NIPS)*, 2013.

[10] Ruslan Salakhutdinov, Antonio Torralba, and Josh Tenenbaum, "Learning to share visual appearance for multiclass object detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[11] Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis Decoste, Wei Di, and Yizhou Yu, "Hd-cnn: Hierarchical deep convolutional neural networks for large scale visual recognition," in *The IEEE International Conference on Computer Vision(CVPR)*, 2016.

[12] Jianping Fan, Xiaofei He, Ning Zhou, Jinye Peng, and Ramesh Jain, "Quantitative characterization of semantic gaps for learning complexity estimation and inference model selection," *IEEE Transactions on Multimedia*, vol. 14, no. 5, pp. 1414–1428, 2012.

[13] Jianping Fan, Yi Shen, Chunlei Yang, and Ning Zhou, "Structured max-margin learning for inter-related classifier training and multilabel image annotation," *IEEE transactions on image processing*, vol. 20, no. 3, pp. 837–854, 2011.

[14] Jia Deng, Alexander C Berg, and Li Fei-Fei, "Hierarchical semantic indexing for large scale image retrieval," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 785–792.

[15] 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.

[16] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *The IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2016.

[17] Hong Wei Ng and Stefan Winkler, "A data-driven approach to cleaning large face datasets," in *IEEE International Conference on Image Processing(ICIP)*, 2015.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[19] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao, "MS-Celeb-1M: A dataset and benchmark for large scale face recognition," in *European Conference on Computer Vision(ECCV)*. Springer, 2016.