Utilization of Resnet in RGB-D Facial Recognition Problems

Xi Xiong

The Ohio State University, Columbus Ohio, USA
xiong.319@osu.edu

Abstract. Resnet, from its emergence, has always been a state-of-the-art model for facial recognition problems. The 2019 Bench Council posted several challenges, including an International 3D Face Recognition Algorithm Challenge, which aims at soliciting new approaches to advance the state-of-the-art in face recognition. We focus on utilizing a 4-channeled Resnet on this new problem and achieve 90% validation set accuracy resulting in second prize on the Bench-19 International Artificial Intelligence System Challenges.

Keywords: Resnet, Facial recognition, RGB-D

1 Introduction

1.1 Motivation

International Open Benchmark Council (Bench Council) is a non-profit research institute, which aims to promote the standardization, benchmarking, evaluation, incubation, and promotion of Chip, AI, and Big Data techniques. In 2019 Bench Council posted several challenges, including the Cambircon track, RISC-V track, X86 track, and 3D face recognition track. The track we chose was 3D face recognition, which aims at soliciting new approaches to advance the state-of-the-art in face recognition. The source code of AIBench is publicly available from this website {http://www.benchcouncil.org/benchhub/AIBench/} (Sign up to get access). An industry-leading internet service AI Benchmark Suite [1] is used in this competition. This paper depicts the effort made towards the utilization of Resnet of this RGB-D facial recognition problem. The problem is a classic face recognition task given traditional RGB face image plus depth information. The dataset consists of over 20000 faces of 1212 distinct personnel, many of them being celebrities, presented with jpeg images and Nd arrays representing depth information accordingly. The dataset includes faces from various races and also images from different ages of the same person. We try to approach the problem with different methods. Initially, we try to start with a 2D face recognition system which register 2d face landmarks and recognize faces by comparing face landmarks with registered face landmarks. This method achieves fairly good performance but lacks utilization of depth data, remaining a space of improvement. Further experiments show that the depth channel is extremely noisy for use in the face recognition system. Based on the fact that

convolutional neural networks outperform other models in many image recognition tasks, we decided to approach the problem with deep convolutional neural networks.

1.2 Challenge

There are two main challenges that we face in this problem, being the inconsistency of data and difficulty in optimization. The first problem exists because the data provided are all of the different resolutions and sizes. To encounter this problem, we padded the images by 90 and used a center crop to format the images to 224*224 for training consistency. The padding was well-considered to compensate for different size of images. After cropping, the images are of the same size to feed into the network; this is a very conservative optimization to the dataset but judging from the results some minor additional noise is added to the dataset. The other challenge is the noise in data: when inspecting the image dataset, we find out that there are noise data that not belong to the personnel identity. Though we try to delete some of those data, the dataset is too large to be cleaned by hand, and this is very likely to affect the result of the training process. For efficiency and computing power limits, we only utilized an 18 layered version of Resnet, and might optimize the model to Resnet-50 for better performance.

1.3 Contribution

Traditional 2D face recognition method divide face recognition into face registration, face detection and face verification these three steps. In order to obtain good performance on such system, each part must function well at the same time. This is extremely hard since for each new situation encountered, we have to modify three parts to adapt the new case. Therefore, we merge all three parts into one deep Resnet and simplify the face recognition problem to a classification problem where our model predicts a label for input face image. We also take inspirations from a paper [2] that focused on transfer learning from a pretrained 2D network and [3] another paper that has a different implementation on a similar Resnet model.

2 Background and Related Work

As the universal approximation theorem implies, a feedforward network with a few layers is enough to approximate any functions [5]. However, the network would become prone to overfitting issues with the data, hence going for a deeper neural network is necessary for a better result in this problem. Since AlexNet, having only 5 layers, the state-of-the-art CNN has grown deeper, with the VGG network [6] and GoogleNet (also codenamed Inception_v1) [7] had 19 and 22 layers respectively. Since the infamous vanishing gradient problem exists in the back-propagation process due to repeated multiplication making the gradient negligibly small, merely piling up layers does not work anymore. For the deeper it goes, the performances would be bottlenecked by this problem, and the results can even start degrading. Before Resnet came out, various methods are being carried out by different researchers in vail to solve the problem effectively.

To mitigate this problem. As addressed in [6], they attempted to solve the problem by adding an auxiliary loss in a middle layer as extra supervision, but still in vain to solve the problem. Resnet [4] incorporates identity shortcut connections, which essentially skip the training of one or more layers creating a residual block. The residual block is a pre-activation variant of residual block [8] in which the gradients can pass through any shortcuts unimpededly. Because of its compelling results in various image recognition benchmarks, we chose Resnet as the building block of the model. As Resnet was implemented for RGB datasets, we optimized the network to have a fourth input depth information channel feeding into the network in addition to the 3 RGB channels, fitting the RGB-D problem.

3 Method

We deploy the dataset provided by Bench2019 for training and testing. Input images are either padded or cropped to 224x224 to feed into ResNet-18. There are 23,140 valid RGB-D images collected from 1212 identities. For the test set, we sample a subset uniformly over 1212 identities from original data.

Moreover, we employ accuracy as an evaluation metric for this classification problem setup. Trained by an Nvidia RTX 2080 for 40 epochs in less than an hour, our model achieves 90% accuracy on the validation set without parameter fine-tuning. However, the model received no further improvement if we increase training epochs. In further experiments we swap ResNet-18 for ResNet-50 while leaving hyperparameters unchanged and the result is disappointing. The generalization for ResNet-50 in this problem setup is significantly worse than ResNet-18 that the validation accuracy drops to around 60%.

The network used to solve this problem is a modified version of Resnet-18, which follows the Resnet model from the 2015 Resnet academic publication, Deep Residual Learning for Image Recognition by He et al. [4]. The authors of this paper argue that stacking layers shouldn't degrade the network performance, because if we simply stack identity mappings upon the current network, and the resulting architecture would perform the same. This indicates that the deeper model should not produce a training error higher than its shallower counterparts. Kaiming He’s team hypothesize that letting the stacked layers fit a residual mapping is more straightforward than letting them directly fit the desired underlying mapping. We then modified the network to have 4 input channels as our model to fit the problem with an additional depth layer. The core idea of Resnet is introducing a so-called “identity shortcut connection” that skips one or more layers, as shown in figure 1:

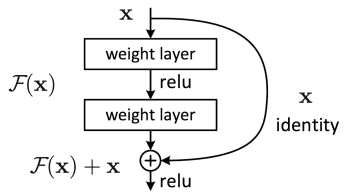

Fig. 1. Residual learning: a building block

Figure 2 shows the whole model:

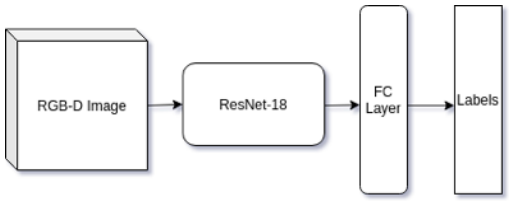

Fig. 2. General Network Architecture based on Resnet [1]

The abbreviated ResNet-18 block we implemented is shown in figure 3:

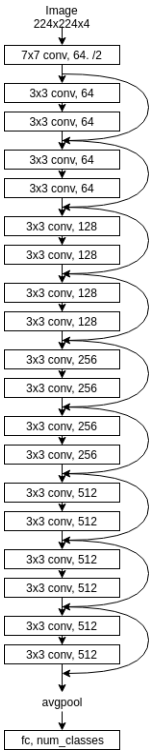

Fig. 3. General architecture of Resnet-18

4 Conclusion

In this paper, we apply the ResNet-18 network for large scale face recognition in a classification setup. Benefitting from the simplified problem setup, our solution combines face registration and face recognition module into a simple neural network that can be tuned by simply adjusting hyperparameters and modifying the dataset. However, this simplified problem setup comes with severe drawbacks. First of all, we can improve performance by utilizing multi-node computing platforms and optimize the training process for more efficiency, as mentioned in this paper [9], which emphasizes methods of scalable and comprehensive datacenter AI benchmarking. Moreover, since the face registration step is merged into a neural network with the recognition part, it is hard for our model to generalize to new faces. Excessive training on new faces makes the model in favor of new faces, while insufficient fine-tune on new data reduces accuracy on new face class. Moreover, our solution lacks the capability to adjust behavior in the face registration step due to the big neural network setup. A potential solution to such problems of intractable face registration can be one-shot learning, which effectively computes a feature representation for new faces based on existing data.

Acknowledgments. We want to give our sincerest thanks to Heming Sun for his advice on face recognition methods as well as to Yujie Hui for his advice in revising this paper.

References

1. Gao, Wanling and Tang, Fei and Wang, Lei and Zhan, Jianfeng and Lan, et al.: AIBench: An Industry Standard Internet Service AI Benchmark Suite (AIBench). arXiv preprint arXiv:1908.08998 (2019).
2. Xiong, Xingwang and Wen, Xu and Huang, Cheng: Improving RGB-D face recognition via Gong, Tongyan, and Niu Huiqian: An Implementation of ResNet on the Classification of RGB-D Images (Bench'19). (2019)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016).
4. Anastasis Kratsios: Universal approximation theorems (stat.ML). (2019)
5. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, (2014).
6. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1–9, (2015).
7. K. He, X. Zhang, S. Ren, and J. Sun. Identity Mappings in Deep Residual Networks. arXiv preprint arXiv:1603.05027v3, (2016).
8. Gao, Wanling and Luo, Chunjie and Wang, Lei and Xiong, Xingwang et al.: AIBench: towards scalable and comprehensive datacenter AI benchmarking. In Proceedings of 2018 BenchCouncil International Symposium on Benchmarking, Measuring and Optimizing (Bench18), pages 3–9, (2018)