IoT Based Attendance System using Face Recognition

S. K. Verma[1], Ashwin Raj[2], Anurag Shrivastava[3], Gourav Kumar[4], Murari Kumar[5], Nilesh K Tiwari[6]

Department of Electronics & Communication Engineering, B.I.T Sindri
Email: [1]santoshverma.ece@bitsindri.ac.in; [2]rajashwin167@gmail.com;
[3]shrivastavaanurag316@gmail.com; [4]grvkmr0712@gmail.com; [5]murarikumarcrj@gmail.com;
[6]nileshkumartiwari050@gmail.com.

Abstract:

In this work, the idea of face recognition is investigated and executed to develop an attendance system. The application (here a web app) is powered with a camera-enabled device to learn and recognise the subject. The main contribution is implementing object recognition algorithm, MobileNet SSD, and connecting the program with an external camera device. These findings are the basis that inspired to revolutionise the attendance system to some extent. A remarkable development throughout the process is observed and this idea has immense possibilities to be implemented widely to develop online-proctored assessments, IoT based securities, surveillance system, etc. A range of industries/institutes is identified where it could be deployed to ease work and provide robust solution to many problems.

Keywords: *Machine Learning, Face Recognition, IoT*

1. Introduction

Convolution Neural Networks have revolutionised by creating a kingdom of opportunities to experiment and build innovative things with the power of vision. Self-driving cars, mobile robots in industries/factories, online-proctoring, security surveillances, etc have well utilised this breakthrough to drive innovation and growth in their respective industries. Its ubiquity and cross industry application can revolutionise the way things work for good.

In this paper we present a unified system for face recognition (who is this person) extended to support the mechanism of smart attendance system. Our implementation is based on transfer learning from a specialized, pre-trained, high accuracy deep convolution network, MobileNet [1] developed by Google. On the top of this convolution network, we train images to recognise it later when required. This real time training of a person's face turned up to be fairly accurate in recognising the subject.

The pioneer work on modern convolutional neural networks occurred in the 1990 by LeCun et. al. [2]. Here, the author has discussed a program designed to recognise handwritten digits. Decades of research helped CNN go popular during 2012 when Alex et. al. published his research entitled as "*ImageNet Classification with Deep Convolutional Neural Networks*" [3]. This work was an outbreak in the science of computer vision which received significant accuracy in classification of images. Our model uses Stochastic Gradient Descent Algorithm [4] for optimizing the cost function. Further significant development to improvise the accuracy of a convolution network are found in [5]-[7] and covers IoT devices convergence with Edge AI. Several open datasets have been developed for research purposes and further advancements of Computer Vision.

This work presents detailed study of the idea of face recognition through one of the convolution networks, MobileNet. This low-latency, fairly-efficient model shows its excellent performance when there are a fewer audience to train. Moreover, this work presents the detail study of the machine learning process from journey of learning, implementing, correcting and learning back.

The paper is further organised as: Section 2 describes the basic understanding of MobileNet. Section 3 describes on-the-top training of images. Section 4 describes an integrated setup of our project. Section 5 explains the elaborated study of the model using an experiment. Section 6 closes with a conclusion.

2. MobileNet

MobileNet uses depth-wise separable convolution to reduce complexities in calculations and the size of the

model. It particularly achieves significant accuracy in embedded vision applications to detect object(s) in an image. With certain tweaks, the architecture of MobileNet supports real-time classification. As a part of its specialization, i.e., classification, our idea of face recognition is completely based on this. We use a pre-trained MobileNet with efficiently tuned hyperparameters that makes this architecture an exceptional deep convolution network for classification. Its credibility is backed by the research team at Google [1].

The depth-wise separable convolution is a form of factorized convolution which factorise a standard convolution at any level into depth-wise and a 1 x 1 convolution called a point wise convolution. Factorisation at every level reduces the cost of computation drastically by forming two separate layers for filtering (depth-wise convolution) and combining the outputs (point wise convolution). MobileNet uses both batch normalization (BN) and ReLU (Rectified Linear Units) nonlinearities for both layers.

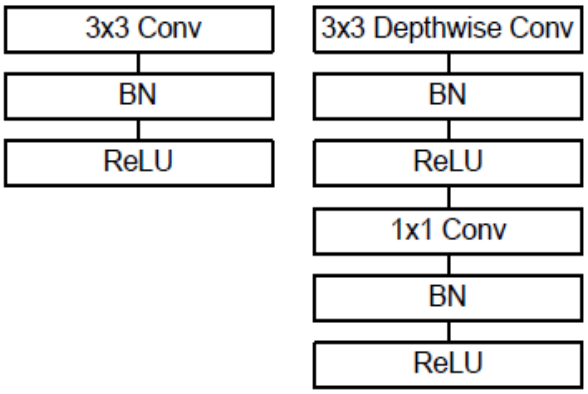


Fig 1: Standard convolution (L) vs Depth-wise separable convolution (R)

MobileNet uses 3 x 3 convolution which uses 8-9 times lesser computation than standard convolution with significant lesser parameters and delivers very less difference in accuracy than the standard classifiers.

The following design (Table 1) of architecture has been proven to outperform many classifiers architecture like GoogleNet [8] and VGGNet [9]. There have been many other comparisons where this architecture with relatively faster computation has out-performed without trading off with the accuracy.

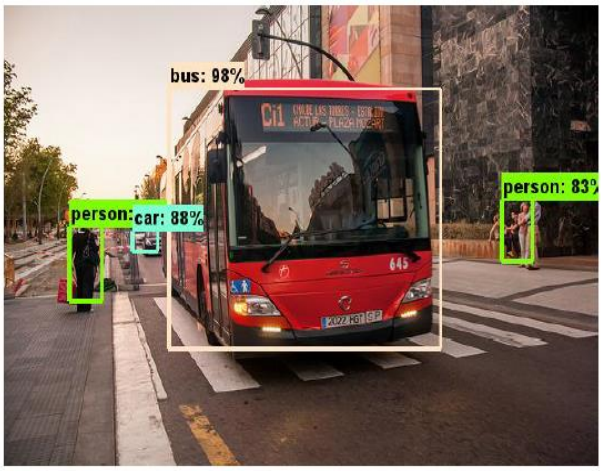


Fig 2: Object detection using MobileNet SSD (Single Shot Detection)

Table 1: MobileNet architecture

Type/Stride	Filter Shape	Input Size
Conv / s2	3 × 3 × 3 × 32	224 × 224 × 3
Conv dw/s1	3 × 3 × 32 *dw*	112 × 112 × 32
Conv/s1	1 × 1 × 32 × 64	112 × 112 × 32
Conv dw/s2	3 × 3 × 34 *dw*	112 × 112 × 64
Conv/s1	1 × 1 × 64 × 128	56 × 56 × 64
Conv dw/s1	3 × 3 × 128 *dw*	56 × 56 × 128
Conv / s1	1 × 1 × 128 × 128	56 × 56 × 128
Conv dw/s2	3 × 3 × 128 *dw*	56 × 56 × 128
Conv/s1	1 × 1 × 128 × 256	28 × 28 × 128
Conv dw/s1	3 × 3 × 256 *dw*	28 × 28 × 256
Conv /s1	1 × 1 × 256 × 256	28 × 28 × 256
Conv dw/s2	3 × 3 × 256 *dw*	28 × 28 × 256
Conv /s1	1 × 1 × 256 × 512	14 × 14 × 256
5×(Conv dw/s1)	3 × 3 × 512 *dw*	14 × 14 × 512
Conv /s1	1 × 1 × 512 × 512	14 × 14 × 512
Conv dw/s2	3 × 3 × 512 *dw*	14 × 14 × 512
Conv /s1	1 × 1 × 512 × 1024	7 × 7 × 512
Conv dw/s2	3 × 3 × 1024 *dw*	7 × 7 × 1024
Conv/s1	1 × 1 × 1024 × 1024	7 × 7 × 1024
Avg Pool /s1	*Pool* 7 × 7	7 × 7 × 1024
FC/s1	1024 x 1000	1 × 1 × 1024
Softmax/s1	Classifier	1 × 1 × 1024

3. On-the top training

It was a great deal for us to take MobileNet into account for face recognition as this convolution network

specializes classification. Understanding more on this enlightened us that different person can be considered as different objects and thus the requirement of face recognition can be served with this model. Further, the paper [1] says that a little distillation with the architecture achieved accuracy close to one of the great architecture available for face recognition, FaceNet [10]. The experiment shows MobileNet accuracy 79.4% against FaceNet accuracy 83% [1]. By far, MobileNet is the most accurate classifier or a detector with such less computational cost.

From the available API of MobileNet, transfer learning i.e., imported the topology of MobileNet architecture [1] is done and added a couple of layers and modified them as per requirement (number of classifications).

4. Setup and Working

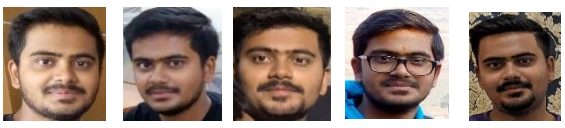
In this section the working of the model is discussed. In a camera-enabled device, this application takes images of the subject as samples. The more the samples captured, the better the results. These samples are now trained on the architecture of MobileNet. Persons are labelled accordingly and training is done. Further, the real-time capturing predicts who the person is in the frame. This project uses a web browser with an active web server(local) to bring everything under one umbrella. For its heavy-duty deployment and production, this setup would require high-end hardware. Though we proposed this system and built for limited use, it can be scaled to a higher level (as per requirement).

5. Example

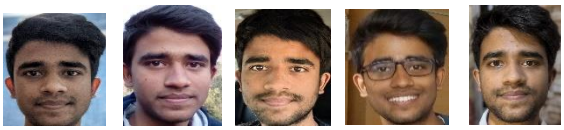
An experiment is carried out to determine the credibility of the proposed model. For this, samples of 5 people are collected. This example elaborates everything stated by far.

Person 1: Ashwin Raj

Person 2: Anurag Shrivastava



Person 3: Gourav Kumar



Person 4: Murari Kumar

Person 5: Nilesh K Tiwari

Table 2: Dataset created using samples

Sample Name	Number of Samples
Person 1 (Ashwin)	40
Person 2 (Anurag)	40
Person 3 (Gourav)	40
Person 4 (Murari)	40
Person 5 (Nilesh)	40

The samples are stacked and labelled accordingly. Training of the model with *5* distinct persons each of one having *40* images as samples is carried out. Following table states the hyperparameter used to tune this model and minimise loss.

Table 3: *Hyperparameters used in the model*

Hyperparameters	Value
Learning Rate (α)	0.002
Epoch	30

Hyperparameters are tuned to achieve higher accuracy and minimal loss in a model.

Learning Rate (α): It is rate by which cost function is optimised. It is generally rate at which algorithm learns the values of the parameters.

Epoch: One pass through all data is called epoch. The more the epoch, the better the model. It is simply the number of iterations run an algorithm through.

Finally, the model is trained giving the aforementioned inputs to the convolution network, MobileNet. The loss of ***0.025*** is obtained after completing all the iterations. Lower the loss, better the model fits with the dataset. Now the prediction part of the model is performed. It is found that the model predicted majority of the test samples correctly [11].

The experiment is further carried out to analyse the results against number of samples per person. We tabulated the loss of the model in Table 4 considering different number of samples per person. Further we tested the model with test data and recorded accuracy of the model.

Table 4: *Recording loss and accuracy with different numbers of samples.*

No. of samples (per person)	Loss measured	Correct guesses (out of 10 test-data tested)
10	0.01945	6/10
20	0.01416	7/10
40	0.01258	7/10
60	0.00845	8/10

Reading the above table clearly helps us drawing the fact that, *greater the number of samples lower is the loss*. Similarly, it's observed that the *greater the number of samples, the better the model performs*. The data in the table is *not concrete and absolute* and merely draws the aforementioned observations. ***It may vary with a different set of datasets.***

6. Conclusion

A face recognition system has been built using a classification specialized model and demonstrated with five different samples. Function of the proposed model has been explained through understanding of network architecture. Additionally, integration of camera with network, training, and prediction in a web browser to make the project fully-functional is also explained.

The proposed model presents an open-ended solution to IoT based attendance system using face recognition. It also reflects that the computer vision is the future for any technological advancements and its incredible integration with IoT.

7. References

[1] A. G. Howard, M. Zhu, B. Chen, Dmitry K., W. Wang, T. Weyand, Marco A. H. Adam. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv: 1704.04861, 2017

[2] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in Proceedings of the *IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791*

[3] A. Krizhevsky, I. Sutskever, Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in neural information processing systems 25, 1097-1105.*

[4] Diederik P. Kingma, Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv: 1412.6980v9 [cs.LG]*

[5] Sergey Ioffe, Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv:1502.03167 [cs.LG]

[6] N Srivastava, G Hinton, A Krizhevsky, I Sutskever, R Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research 15 (1), 1929-1958

[7] Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge Computing: Vision and Challenges. IEEE Internet of Things Journal, 3, 637-646.

[8] B. Szegedy, W. Liu, et. al. Going Deeper with Convolution arXiv: 1409.4842v1,

[9] Karen S, Andrew Z. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv: 1409.1556, 2015.

[10] F. Schroff, Dmitry K., J. Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. arXiv: 1503.03832, 2015

[11] https://github.com/grvkmr07/project_claassifier/tree/main/attendanceSystem