Security Enhancement in ATMs through Helmet Detection using Inductive Transfer Learning

Madhusmita Dey[1], Swati Sucharita Barik[2]
[1](Department of Computer Science and Engineering,
Centurion University of Technology and Management, Bhubaneswar, Odisha
Email: madhusmitadey39@gmail.com, 160303110008@cutm.ac.in)
[2](Faculty of Computer Science and Engineering,
Centurion University of Technology and Management, Bhubaneswar, Odisha
Email: swati@cutm.ac.in)

ABSTRACT

Security Facilities are very important in the place of financial systems. Along with convenience, customers are facing many security problems in ATMs (Automated Teller Machine). As a part of security, surveillance cameras are fixed in each ATM which is monitored by human. ATM visitors who have wore helmets make it difficult to identify the person if any abnormal activity happens or violate the rules through the surveillance camera. Due to this problem users are not allowed to wear helmet in ATMs. For improving security in ATMs, an automated helmet detection using ATM surveillance camera is used for detecting the helmet. Here we used four pretrained models which are based on convolutional neural network such as AlexNet, VGG-16, ResNet-101 and Inception-v3. All models are trained on same conditions with ATM surveillance dataset. After Comparison of results we concluded that, the Google's inception-v3 model is giving 95% accuracy then other models. So, we can use the Inception-v3 model for helmet detection in ATMs for better security.

Keywords- *Convolutional neural network, Deep learning, Image classification, Transfer learning, Video surveillance: Helmet detection*

I. INTRODUCTION

Human brain plays an important role in Visualization or object detection. Visually understanding the objects and their features is a complex task for the brain. If we comparing the brain with the most powerful computers, simply we found that brain is most powerful than others. Brain observes the things and instantly and automatically divides the images in to meaningful simple and complex shapes, related objects and region of interest (ROI) [1]. Within milliseconds the brain makes decisions on the basis of different figures or shapes. It takes only those portions which will effect in decision making and discard all other portions of an image. The image in figure 1 shows that, here our interest area is the user, whereas the ATM and the posters are the part of the backend which are not considerable for decision making.

Similar thing happens in computer; here we are analysing the information from the digital images. The digital images are categories on the basis of different region of interest.

Fig. 1: ATM survillance image

Now a days intrusion detection(Helmet detection) is a very challenging problem for the system .For a real time example,a user entered in to the ATM with wore a helmet,a helmet detection system is analyse the information from digital images through camera or video camera.It is very helpful for identifying who violate the rules or does malicious activities (helmet detection). Because at that time the user present in that place and an emergency warning can be generate automatically.

The digital image store in the computer in the form of pixels.Each pixel has a specific number. From the number of the pixel, the computer understand the color of the each pixel of a digital image [2]. The recognition algorithm or method focus on the aspect of pixels which having some properties is used for comparing two nonlinear images [3]. A recognition method trained by training dataset and learning objects from input images which are real time images or offline images. A database is obtained by the recognition method which having different objects or classes of objects which are used for comparison between the images.Here the class is a type of object contains some common features like example different brands of cars .

II. RELATED THEORIES

Neural networks and deep learning [4] solves many complex problems in image recognition, speech recognition, and natural language processing and provide the best solutions to the problems. Deep learning help constructing machine learning models which are able to learn hierarchical representation of very large dataset.

Convolutional neural networks (**ConvNets** or **CNNs**) [5] are a category of neural networks which are playing an important role in the field of image recognition and classification. **ConvNets** is a feed forward neural network and have been successful in recognize faces, objects, characters and others. There are many drawbacks present in deep learning. First of all, we require lots of data, time, cost and memory to train and execute the model [6]. Second, most machine learning methods works perfectively for training and testing data which are taken from same feature space and distribution but if any changes occurred in distribution then the model must be rebuild to prevent the failure [7].

Overcome from this problem transfer learning approach is introduced. In transfer learning we can transfer knowledge from one task or domain to another related task or domain for solving problem [7]. Transfer learning focuses on storing knowledge gained while solving one problem and use gained knowledge from a pretrained model to a new task with minimum computation in less time. Convolutional network having more learning capabilities then other traditional networks. For getting better models in CNN, we can add more numbers of convolutional neural layers. In the field of Deep learning LeNet5 model was one of the very first convolutional neural network. This work was done by Yann LeCun. LeNet model was used for character recognition tasks such as reading zip codes, digits etc [8]. LeNet5 model is made up of a set of convolutional layers with normalized input, sub-sampling layers or max pooling layers and fully connected layers [9]. Train the model and then evaluate the network performance on the MNIST [10] dataset, CIFAR and ImageNet [11]. For getting proper output, we can iterate the model by adding or increasing the number and size of the layers until required outputs are observed .CNN consisting of convolution layers, ReLU layers (Normalization layers), pooling layer (Sub Sampling) and fully connected layers (Classification). Feature maps are generated by using the non linear activation function and reduced by the using maximum pooling. Network-in-Network [12] approach is proposed by Lin et al. This approach is used to increase the representational power of neural networks [13].

III. PROBLEM STATEMENT

ATMs are playing a vital role in our day to day life. Previously we were facing many problems to withdraw cash from a bank account. We had to stand in a queue for hours for withdrawing money. Overcome from this problem ATMs are introduced. ATMs are placed all over the world and still placing large number of ATMs every day. We know that, everything has some good or bad side. With convenience of ATM, it bought many security problems. The automated surveillance cameras are placed in every ATMs which are monitored by human. It is very difficult to identify the person does any abnormal activity if the user wore the helmet. So overcome from this problem the CNN taken as consideration which is using for object detection. Convolutional neural network (*CNN* or ConvNet) comes under a category of deep learning, feed-forward artificial neural network. By using CNN with transfer learning, we can solve the recognition or detection problem with fewer computations.

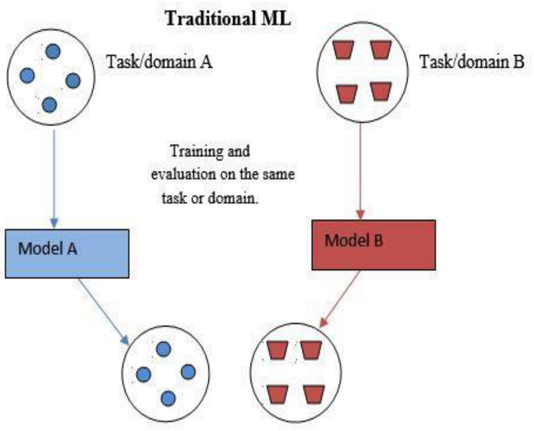

Fig. 2: Traditional Machine learning

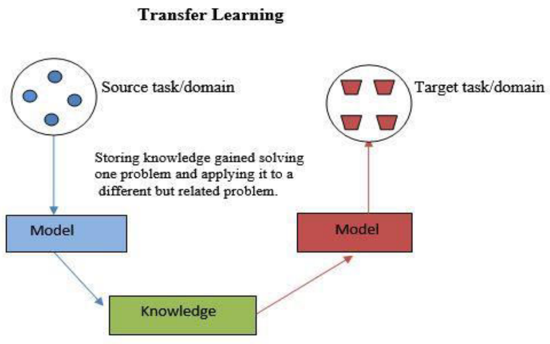
Transfer Learning
Source task/domain
Target task/domain
Storing knowledge gained solving one problem and applying it to a different but related problem.
Model
Model
Knowledge

Fig. 3: Transfer learning

IV. PROPOSED TECHNIQUE

Convolutional networks are core of the computer vision. In image recognition process fully connected networks are not used, because it requires huge numbers of neurons for huge number of parameters in each hidden layer. It is looks like a complex structure;

require huge memory for execution with high computational cost. The figure 4 shows that, in CNN the neuron in a layer will only be connected with the small region of the layer before it, instead of all the neurons connected in the fully connected manner.

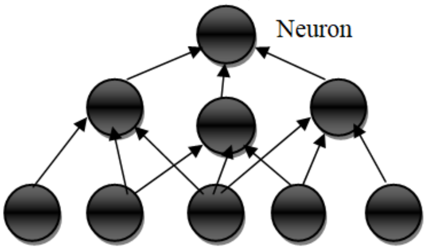

Fig. 4: One neuron connected with three other nearest neurons

The big image divided into small units. The units are known as filters. Each filter put on the same input image, if the filter matches with the image then the image is correctly classified. Here we are using Google's inception- v3 model or GoogLeNet. In ImageNet LSVRC2014 classification challenge the inception-v3 model or GoogLeNet is classify images of 1000 classes from ImageNet [13]. The model is build up on using 1.2million images for training, 50,000 for validation and 100,000 images for testing [13] which is used in this work. Transfer learning approach is applied on this model. The acquired knowledge of the model is transfer and used for solving other similar problems. So here we are using the acquired knowledge for detecting the helmet in ATMs from ATM surveillance dataset with two classes or categories which shows in figure 5.

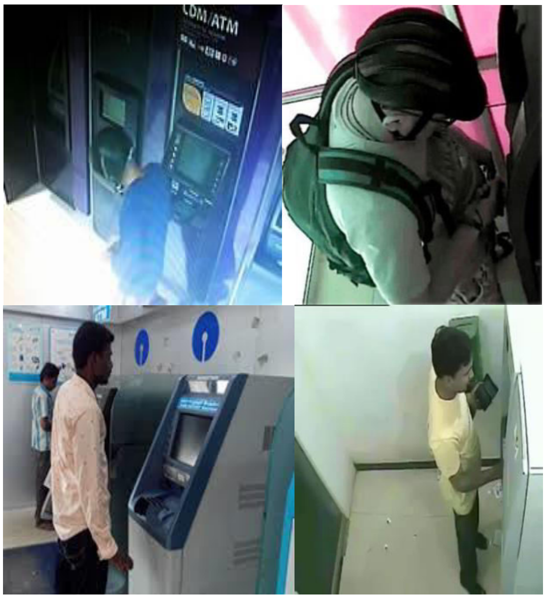

Fig. 5: ATM surveillance Dataset

The naive version inception module [13] contains one 1x1 convolution layer, one 3x3 convolution layer, one 5x5 convolution layer and one 3x3 max pooling layer. Then the architecture is combine all the layers with filter concatenation and produces a single output vector which is input for the next layer. The following figure 6 shows the naive version inception module.

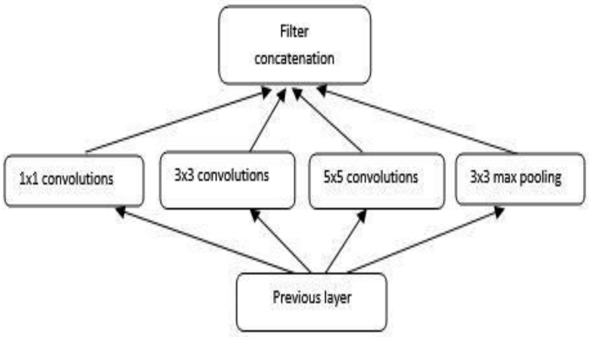

Fig. 6: Inception module, naive version

There is a problem with this above module; even a small number of 5x5 convolutions are computationally expensive on top of the convolutional layers with a large number of filters.

So overcome from this above problem discover the second idea of the proposed architecture is dimension reductions and projections, shows in figure 7. When we applied this approach in to convolutional layers, the approach is viewed as additional 1x1 convolutions. 1x1 convolutions are placed before each expensive 3x3 and 5x5 convolutions for remove the computational difficulties and also used for increasing the depth (number of stages) as well as width of the network without increasing the computational complexity. 1x1 convolutions or filters are placed in the projection layer after the 3x3 max pooling layers in the pool projection column. Rectified linear activations (*Rectified linear* unit) are used by these reduction and projection layers. This architecture results 2−3× faster performance than networks with non-Inception architecture. Still large number of difficulties occurred to make changes in the networks in the inception model. The module was difficult to scale up due to computational loss. Also the computational cost and number of parameters increases for simple transformation.

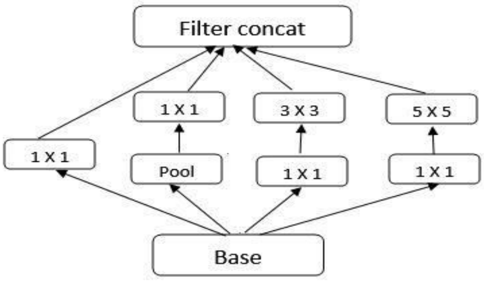

Fig.7: Inception module with dimension reductions

Larger filters or convolutions e.g. 5x5 or 7x7 operations are computationally more expensive. Using same number of filters, a 5x5 convolution is approximately 2.78 times more computationally expensive then a 3x3 convolution.

So overcome from this above problem, the larger convolutions are factorized in to smaller convolutions. The 5x5 convolution replaced by a multi-layer network with less parameter with the same input size and output size. The model was modified by replacing the 5x5 convolution with two 3x3 convolutions and 7x7 convolution with three 3x3 convolutions.

The following figure 8 shows the larger convolutions are factorized in to smaller convolutions

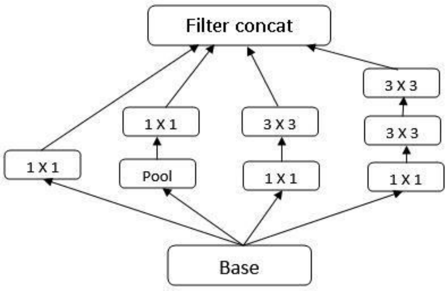

Fig. 8: Inception modules where each 5x5 convolution is replaced by two 3x3 convolution

The above technique, always not useful for reduce the larger convolution in to 3x3 convolutional layers. We also think that, one can factorize the larger convolution into smaller, e.g. 2x2 convolutions. A 11% saving of computation occurs when factorizing a 3x3 convolution into a two 2x2 convolutions. But we get better result using asymmetric convolutions [14] than 2x2 convolutions, e.g. nx1.The 3x1 convolution followed by a 1x3 convolution with the same receptive field as in a 3x3 convolution. If the numbers of input and output filters are equal, then the asymmetric convolution is 33% cheaper than the 2x2 convolution. So n x n convolution can replace by a 1 x n convolution followed by a n x 1 convolution. This module shows in figure 9.

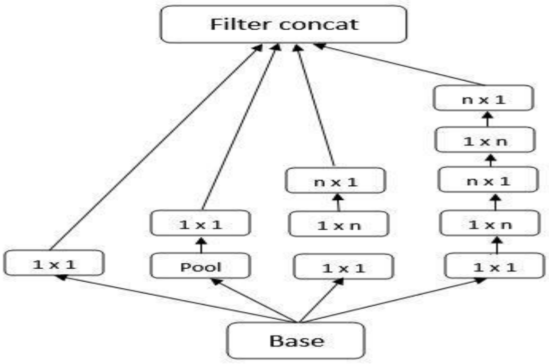

Fig. 9: Inception modules after the factorization of the n x n convolutions

The transfer values are generated after each image passed through 316 layers of the inception-v3 model which is trained with the ATM dataset. The transfer learning of deep neural network is used for feature extraction without recalculating again and again. So that computational requirements are extremely less by using this approach.

V. EXPERIMENTS AND RESULTS

Here all models are pretrained on the ImageNet database with more than a million images with 1000 object categories, such as keyboard, mouse, pencil, and many animals. Transfer learning is used on all over the pretrained models under same conditions with same ATM dataset. The Table1 shows the experimental results of the different models are used for helmet detection.

Table 1: Model accuracy

Networks	Accuracy
AlexNet	90%
VGG-16	85%
ResNet-101	90%
Inception-v3	95%

After comparison of transfer learned inception model with AlexNet, VGG-16 and ResNet-101, we concluded that the inception-v3 model is giving better accuracy than the other models. So, the inception-v3 model is an appropriate pretrained model which can use in ATMs for helmet detection. We used MathWorks MATLAB Neural Network Toolbox to train and validate the networks. All models are trained on Intel CORE i3 7th Gen CPU clocked at 2.40 GHz processor and 4GB RAM system.

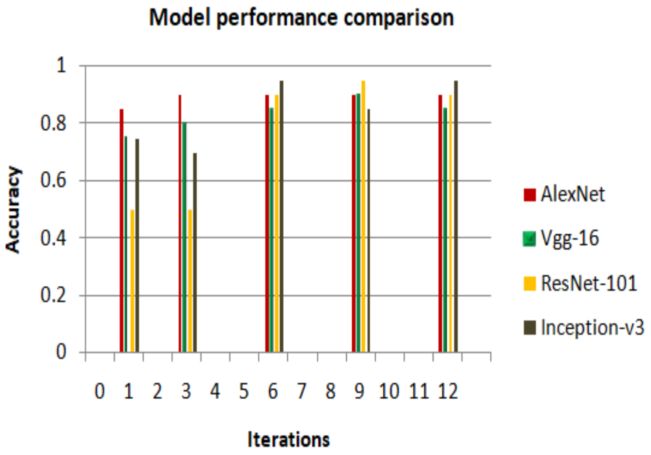

Fig. 10: Performance comparison

From the above performance comparison graph figure 10, we visualize that which model is giving better accuracy with in fixed number of iterations which are generated on the basis of the size of the dataset we have taken for this work. In first iteration, we compared all model's accuracy. From this, we identify that the AlexNet model gave higher accuracy then other model. But in last and final iteration the inception-v3 gave higher accuracy than others. Because of the final iteration shows the actual output, inception-v3 model is the most preferable one. The numbers of iterations are generated on the basis of the size of the dataset. It is not fixed. It varies according to the dataset size.

As the number of layers increases in CNN, the performance of image recognition task will be better. But adding more layers are not as simple as we think. One of the problems for increasing the number of layer is vanishing/exploding gradients which hamper the convergence. This can be overcome by normalizing initialization and normalizing intermediate layers. Another obstacle is the degradation problem, which can be overcome using residual learning.

VI. CONCLUSION

In our work we used transfer learning in convolutional neural network for helmet detection. We can use this model over videos, came from the surveillance camera as well as the system coupled with other sensors like motion sensor. In ATM we can add extra features like alarm, automated door lock along with the proposed system to enhance the security. Instead of train the model manually we can update the system such a way that the model can be train automatically on daily or weekly basis over data captured by surveillance camera.

In our model we achieved 95% accuracy in testing phase. Better result can be achieved by training the model on a larger dataset and improving the system specification.

REFERENCES

[1] M. Brown, Neuronal responses and recognition memory,in:Seminars in Neuroscience,Elsevier,Volume 8,Issue1,February1996, pp.23–32.
[2] S. Saha, C. pal ,R. paul,S. Maity ,S. Sau, A brief experience on journey through hardware developments for image processing and it's applications on Cryptography, arXiv preprint arXiv: 1212.6303,2012.
[3] J.Ma, L.Zheng, M.Dong, X. He, M. Guo, Y. Yaguchi, R.Oka, A segmentation-free method for image classification based on pixel-wise matching,Elsevier, Journal of Computer and System Sciences79,2013,256–268.
[4] J.Schmidhuber, Deep Learning in Neural Networks: An Overview, arXiv preprint arXiv: 1404.7828,2014.
[5] K.T. O'Shes, R. Nash, An Introduction to Convolutional Neural Networks, arXiv preprint arXiv: 1511.08458,2015.
[6] A. Ike,T. Ishihara,Y.Tomita, T.Tabaru, Technologies for Practical Application of Deep Learning, FUJITSU Sci.Tech.J.,Vol.53,No.5,pp.14-19,September 2017.
[7] K. Weiss, Email author, M.T. Khoshgoftaar, D. Wang, A survey of transfer learning. Weiss et al.Journal of Big Data 3:9, 3(1):1–40, 2016.
[8]Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Computation, 1(4):541-551,1989.
[9] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, "Gradient-based learning applied to document recognition", Proceedings of the IEEE86, 2278–2324, 1998b.
[10] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, EMNIST: an extension of MNIST to handwritten letters, arXiv preprint arXiv: 1702.05373,2017.
[11] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei, ImageNet: a Large-Scale Hierarchical Image Database. Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 20-25, 2009.
[12] M. Lin, Q. Chen, S. Yan, Network in network, arXiv preprint arXiv: 1312.4400, 2014.
[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, arXiv preprint arXiv: 1409.4842, 2014.
[14] D. Liang,Y. Zhang, AC-BLSTM: Asymmetric Convolutional Bidirectional LSTM Networks for Text Classification , arXiv preprint arXiv: 1611.01884, 2017.
[15] K. Grm, V. S truc, A. Artiges, M. Caron, H. K. Ekenel, Strengths and Weaknesses of Deep Learning Models for Face Recognition Against Image Degradations, arXiv preprint arXiv: 1710.01494,2017.
[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein etal., Imagenet large scale visual recognition challenge, arXiv preprint arXiv: 1409.0575, 2015.
[17] R. Pascanu, T. Mikolov, Y. Bengio, On the difficulty of training Recurrent Neural Networks, arXiv preprint arXiv: 1211.5063,2013.
[18] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, arXiv preprint arXiv: 1512.03385,2015.
[19] Anonymous authors, Avoiding Degradation in Deep Feed-Forward Networks by Phasing out Skip-Connections, under review as a conference paper at ICLR 2018.