Safeguarding Factories using Artificial Intelligence

Harsh Shah[1], Shubhankar Sawant[2], Kshitija Kale[3], Prof. Renuka Nagpure[4]

[1,2,3]*Student, Department of Information Technology, Atharva College of Engineering, Mumbai, Maharashtra India.*
[4]*Professor, Department of Information Technology, Atharva College of Engineering, Mumbai, Maharashtra India.*

Abstract -*The public is quick to embrace the use of cameras in a wide selection of locations and applications: Live traffic monitoring, parking, car surveillance, and smart spaces. These cameras provide information in a common area that needs careful analysis. Unfortunately, many visual surveys are still dependent on the operator to filter this video. We will analyze any text used with the aid of Artificial intelligence and divide the footage into two categories which are dangerous activities and safe activities. The advanced program uses effective methods such as Deep Learning. In particular, convolutional neural networks tend to detect features within video images, while Recurrent Neural Networks tend to analyze these features and predict potential activity within the video.*

Key Words**: Convolution Neural Network, Recurrent Neural Network, Long Short-Term Memory, Human Behaviour**, **Inception v3**

1. INTRODUCTION

Human Behavioral Analysis (HBA) is an important field to focus on artificial intelligence. It has many areas of use such as video surveillance, environment-assisted environment, smart shopping sites, etc. The availability of personal video data is increasing significantly, with the help of the leading companies in the area. From a Deep Learning perspective, this work is getting closer to HBA. Due to the growth of computer capacity, in-depth learning strategies have been a positive step in the context of fragmentation over the past few years. Convolutional Neural Networks (CNNs) are used for image recognition and RNNs for temporary understanding such as video or text. The following sections discuss them in detail

Problem Statement

In the Construction and Manufacturing Industry every year there are thousands of deaths due to the lack of surveillance events in the production/construction area. These industries are considered dangerous, with high occupational hazards leading to high temporary and permanent injuries and fatalities. By using this program, we can reduce the high number of deaths and accidents caused by factory workers. In India, there is no such type of system currently being used in factories that can make the transition to the industry as all life is important in today's world.

1.1 Deep Learning

Deep learning is a branch of machine learning that uses advanced neural networks, a variety of "building" systems to be able to access objects from a large amount of non-labeled training data. Due to better performance compared to conventional methods such as decision trees, vector support machines, Bayesian networks, etc., there is a growing interest in the last decade of intensive education use. Over the past few years, the proliferation of computer power has made it easier to process life-inspired techniques made over the past decades.

1.1 Convolutional Neural Network

Convolutional Neural Networks are a form of Artificial Neural Networks designed to process large amounts of input information (images, sound, or video). Due to a large number of input details, it would not be possible to remove this function via the standard Full Network (FC). In a broad sense, what CNN does is reduce data by looking at specific data regions to exclude certain features. CNN's are based on filters (characters) that serve as the weights of ANN Completely Connected. The only difference with FC weights is that a single convolutional filter is distributed across all input regions to produce a single effect. It's called Local Receptive Fields, and the weight reduction that CNN doesn't know is very beneficial. Road output is measured by sliding the filter above the input. The output between each kernel object and a separate input is calculated for each location. All items are then summarized to find the product in the current location.

Increasing the size of the feature maps is a common way to reduce the number of parameters and the calculation value on CNN. The integration layer works independently (usually following the convolutional layer) and uses the upper or middle role map feature to reduce the regions. The most effective way to find features is a combination of multiple layers of discussion and parallelism. This is because different kernel sizes can be customized, allowing basic and sophisticated functionality to be available in different parts of the network.

1.3 Recurrent Neural Network

The strategies outlined above are designed for planning

private details, however, what happened there Do we carry time information? Management of these needs, another type of neural system existed intended to show timeline details. These systems are called Recurrent Neural Systems (RNNs) also allow data to continue, by logging in it. In this case, we get xt input stream, hence the stream output, in all iteration, output ot, is another input of the next iteration. Basic RNNs help model a small temporary dependence. When dealing with the long-term sequence of information (in most real cases) an alternative replacement of RNN called Long Short-Term

Memory networks in use.

1.4 LSTM Networks

LSTM networks, presented by Hochreiter & Schmidhuber (1997), are RNN-based models that be able to learn long-term dependence. Vanilla RNNs have a very simple method structure, as one view per tanh launch layer. In contrast, LSTM cells have a more complex structure, in which there may be a single layer (tanh), four, are very interactive in a special way.

Video analysis provides more details about the work of recognition by inserting a temporary size from additional use that can be made by movement and so on details. At the same time, the work becomes more intense more challenging with processing is shorter video clips, as each video, can contain hundreds to thousands of frames, not all of which are useful. An ignorant person The solution would be to view video frames as still images, to use CNNs to identify each frame, and to rate Video-level predictions. Still, each video frame forms a small part of the story for video, such a method would use incomplete knowledge and thus can easily confuse groups, especially if well-organized divisions or parts of the video are not related to the action.

So we think it's a global learning a description of the video's temporary appearance Accurate classification of data is required. From the idea of modeling, this is as difficult as we should emulate fixed videos that have a certain number of parameters. The main purpose of this project is to design to use an in-depth effective learning solution that can be predicting and classifying human behavior into two categories which is a safe job and a dangerous job by using a combination of CNNs and RNNs architectures.

2. METHODOLOGY

2.1 ACTIVITY RECOGNITION SYSTEM

When we examine the technical nature of the various behavioral recognition systems, identify the tools we will use, and the expensive process of downloading two sets of data, we begin to present our proposal. This program is a combination of different DL models, CNN reads video frames and extracts features, while RNN reads those features, predicts performance. This DL model is based on Python, using Camera Frame (using Tensorflow frame as backend).

Before proceeding with this training step, the data should be processed in advance for the DL model to fit properly

2.2 MODEL

The state-of-the-art visual deep learning shows that the best way to deal with this problem is to use a model with Convolutional Neural Network, initially, to extract features of video frames, followed by Recurrent Neural Network which can model Frame sequence. Other Behavior Recognition DL models include 3D CNN using the FC network. In this way, the whole video is uploaded to 3D CNN simultaneously, and CNN is capable of extracting not only image elements but also motion or time elements. All of these items are included in the vanilla FCs network. The problem with this approach is that performance forecasting requires all diagrams.

However, activity can be predicted before the video ends, and this method is better because it can predict activity in real-time (predictive prediction).

2.3 CONVOLUTION PART

It is not an easy task to achieve a flexible 2D neural network by being efficient in understanding images and producing its features (vector that summarizes image details). This is due to the difficulty of finding a good model, and the amount of time and data required for training. As a result, the most common method of in-depth learning is to install a pre-trained model to extract features and submit a feature in the new model. Several models have previously been trained to see images. ImageNet is a database that organizes an annual challenge

(ILSVRC) since 2010, to test object detection and image classification algorithms. In this rival, several types of DL have emerged since 2012: Alex. Net (2012), ZF Net (2013), VGG Net (2014), GoogLeNet (2014), Microsoft ResNet (2015). All of these in-depth learning types (since 2012) have two key blocks: the image element extractor uses convolutions, number, and the structure of the layers determined by the model. The second block is connected to the splitting process, in this case, it will be a feedforward neural network that takes the element vector as its input and separates the object type (output size depends on the number of items to be included in the category). This second item is similar to any ILSVRC challenge pattern. In this project, we will use part of the release feature of a pre-trained model called transfer learning that focuses on storing information and applying it to a specific but related question. Model us which will use Inception v3 because it has excellent classification accuracy and low calculation costs. Some models gain better performance as Implementation of ResNet v2 but they have layers that require a lot of calculation and the increase in partitions is not so great. One of the models available within the Kera system is a model of Inception v3 (Figure 6). The manner in which initiation works is as follows. Instead of making a pyramid of convolutions (one after the other), Start has, they call the starting modules, layer groups where the flow is not sequential. Many combinations of different sizes are calculated separately for these modules then one line is added. This process allows for the release of additional functionality. It also helps 1 x 1 blend to reduce performance. The partition is a second part of the Startup network, made up of a fully connected layer and a softmax output layer. This method of dividing is it is appropriate if we need to split an image at the same time, but if we need to split an image stream, such as an image, then we need RNNs

2.4 RECURRENT PART

Recurrent Neural Network is the best way, depending on the state of the art, to learn in-depth to identify the sequence of inputs such as text, speech, or video composition. RNNs can show data sequences with internal loops that provide input to the network. Long

Short-term Memory Networks are a way for RNNs to be able to "remember" important parts of input sequences, regardless of where the time came from (simple RNNs only remember the latest parts of the sequence, with temporary memories). In this model, it is suggested that the LSTM network adopts part of the initial network function. The size vector size retrieved by the Inception v3 network is 2048, so it is suggested that the LSTM layer of the same size remember all aspects of the vector series (each LSTM cell will be fitted with one element). A fully integrated 512 Neuron layer is attached to the back of the LSTM layer.

3. IMPLEMENTATION

The proposed framework includes the use of two neural networks. They are Inception-v3 and LSTM.

The required steps are as follows:

Phase 1: Database structure

- Classification videos are dangerous and safe activities Work is collected from web sources, for example, Youtube, Dailymotion, and so on.
- Videos then need to be labeled. Since there are only two stages, it can be so is represented as 0 and 1.
- The video files in the above categories are renamed as 0_1, 0_2, 1_1, 1_2, etc.
- Video file for example 0_1, here '0 'stands for a dangerous category function, '1 'represents a video file hazardous category number and '_' used to divide those numbers.

Phase 2: Training

- 75% of the database is used for model training as well 25% of the data is used for model testing.
- Videos are converted to frames.
- Each video label is extracted and stored in sequence In each frame, the pixels of the image are converted to NumPy many redesigned lists at the beginning of the V3 model.
- The framework is then processed and features are developed extracted from the frame using the first V3 model.
- The elements of the transfer learning are used from independent to their labels are used as an input to the LSTM model.
- The LSTM model consists of 2 hidden layers with a recess sigmoid function sequence and discharge layer 2 neurons with softmax activation to produce video segmentation.

Phase 3: Testing

- Tests are performed on one video
- Frames are removed from the video
- In each frame, the pixels of the image are converted to NumPy list rebuilt in the early V3 model.
- The framework is then processed and features are developed extracted from the frame using the first V3 model.
- The elements are then stored in the correct order at the time converted to NumPy list and then reconstructed input LSTM status.
- Using the predict_class model of the LSTM model method Producing video file segmentation.

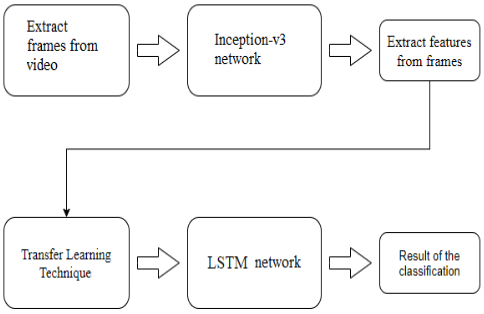
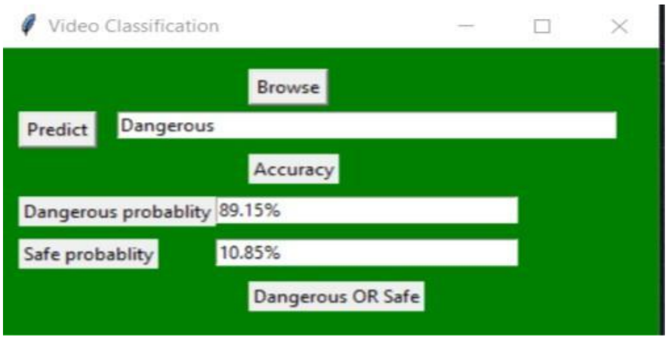

Extract frames from video
Inception-v3 network
Extract features from frames
Transfer Learning Technique
LSTM network
Result of the classification

Figure -1: Block Diagram

4. RESULTS

This section analyzes the results obtained for labeled databases built into the implementation phase.

Using the selected user interface file that needs to be split the following figure 2 shows the predictive effect when the installation video is for store hacking, which is classified by our system as a dangerous activity. Discussion diagrams are categorized by Safe Activity in Figure 3.

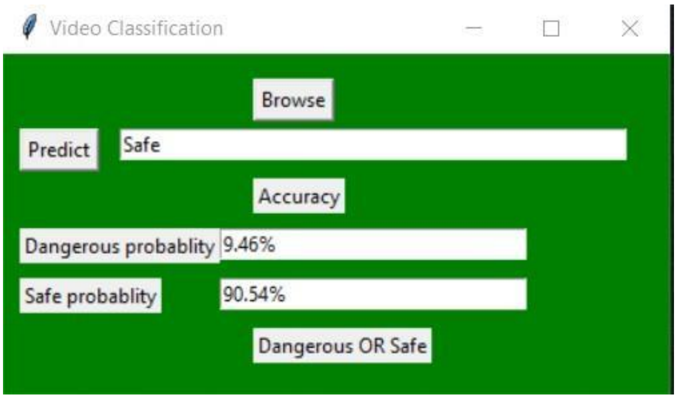


Predict: Dangerous

Dangerous probablity: 89.15%

Safe probablity: 10.85%

Figure -2: Dangerous Activity



Predict: Safe

Dangerous probablity: 9.46%

Safe probablity: 90.54%

Figure -3: Safe Activity

The fit() method which is used to train the model returns a History object. The history attribute of the History object is a dictionary that contains successive metrics and loss values.

Figure 4 shows the training and validation accuracy obtained while training the model at successive epoch.

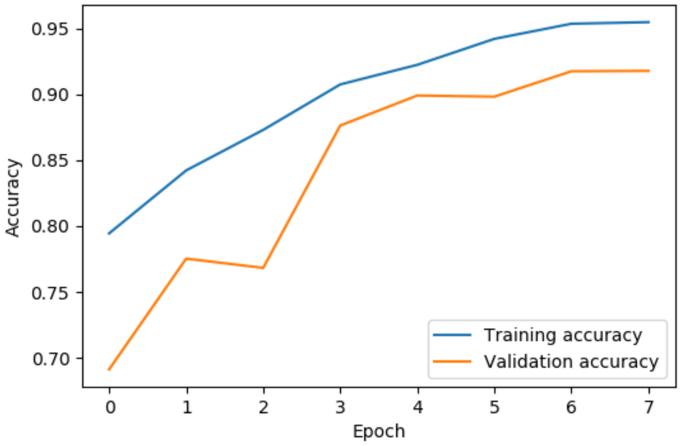


Figure -4: Training and Validation accuracy of the model

The training and validation accuracy obtained at last epoch is approximately 95 and 92 percent respectively.

5. CONCLUSIONS

The classification of video segmentation is problematic for many reasons, for example, lack of video databases, low accuracy, etc. The main points of this paper, the data management of different stocks is different from the Deep Learning model; transfer the learning of the pre-trained Deep Learning (Inception V3) model to our system; use of LSTM Recurrent Neural Networks. From day to day, daily inspections of recording are a problem that the framework may hold such a meeting and analyze the videos with high accuracy. Online video-based searches, security surveys, detection and deletion of copyrighted uploaded videos are part of its future rating.

As a future project, we aim to increase the accuracy of the system by taking advantage of the data diversification offered by Activitynet to obtain a more robust system of visual activity. A better model can be implemented by developing a fine-tuning process, which varies the number of layers, neurons, learning scale, etc.

ACKNOWLEDGEMENT

We express our deep gratitude to **Prof. Renuka Nagpure**, our project guide for her exemplary guidance, monitoring, and constant encouragement to push us beyond our limits

throughout this project. We would like to thank her for providing idealistic solutions and give fresh dimensions to the completion of this project. Regular virtual meetings and doubt solving sessions made it easy for us. We are overwhelmed to offer our gratitude to the Head of the Information Technology department **Prof. Deepali Maste** for her moral support which motivated us throughout the hardships to complete the project on time in this pandemic. Special thanks to the principal of Atharva of Engineering **Prof. Shrikant Kallurkar** and the entire administrative staff to provide us with well-furnished libraries along with necessary reference papers. Well-equipped IT labs and other infrastructure provided us with a perfect environment to deliver the best of us.

REFERENCES

[1] Wang, H.; Schmid, C. Action recognition with improved trajectories. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 3551–3558.

[2] Bilen, H.; Fernando, B.; Gavves, E.; Vedaldi, A.; Gould, S. Dynamic image networks for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3034–3042.

[3] Chen, C.; Liu, K.; Kehtarnavaz, N. Real-time human action recognition based on depth motion maps. J. Real Time Image Process. 2016, 12, 155–163. [CrossRef]

[4] Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neuralnetworks for human action recognition. IEEE Trans.Pattern Anal. Mach. Intell. 2013, 35, 221–231. [CrossRef][PubMed]

[5] Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Li, F.F. Large-scale video classification with convolutional neural networks. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.

[6] Simonyan, K.; Zisserman, A. two-stream convolutional networks for action recognition in videos. arXiv 2014, arXiv:1406.2199.

[7] Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks: Towards good practices for deep action recognition. arXiv 2016, arXiv:1608.00859.

[8] Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1933–1941.

[9] Wang, L.; Qiao, Y.; Tang, X. Action recognition with trajectory-pooled deep-convolutional descriptors. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.

[10] Lu, X.; Yao, H.; Zhao, S. Action recognition with multi scale trajectory-pooled 3D convolutional descriptors. Trans. Multimedia Tools Appl. 2017, 1–17. [CrossRef]

[11] Taylor, G.; Fergus, R.; LeCun, Y.; Bregler, C. Convolutional learning of spatiotemporal features. In Proceedings of the 11th European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010; pp. 140–153.

[12] Fernando, B.; Gavves, E.; Oramas, J.M.; Ghodrati, A.; Tuytelaars, T. Modeling video evolution for action recognition. In Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, The USA, 7–12 June 2015; pp. 5378–5387.

[13] Perronnin, F.; S´anchez, J.; Mensink, T. Improving the fisher kernel for large-scale image classification. In Proceedings of the 11th European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010; pp. 143–156.

[14] Zaremba, W.; Sutskever, I.; Vinyals, O. Recurrent neural network regularization. arXiv 2014, arXiv:1409.2329.

[15] Donahue, J.; Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 39, 677–691. [CrossRef] [PubMed]

[16] Veeriah, V.; Zhuang, N.; Qi, G.J. Differential recurrent neural networks for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Santiago, Chile, 7–13 December 2015; pp. 4041–4049.

[17] Yue-Hei, J.; Hausknecht, M.; Vijayanarasimhan, S. Beyond short snippets: Deep networks for video classification. In Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4694–4702.

[18] Wu, Z.; Wang, X.; Jiang, Y. Modelling spatial-temporal clues in a hybrid deep learning framework for video classification. In Proceedings of the 23rd ACM