Retail Shrink Control by Shoplifting Detection

Naga Jyothirmayee Dodda
School of Computer Science
University of Windsor
Windsor, Canada
SID: 110083327

Prem Shanker Mohan
School of Computer Science
University of Windsor
Windsor, Canada
SID: 110036738

Roisul Islam Rumi
School of Computer Science
University of Windsor
Windsor, Canada
SID: 110080808

Sumaiya Deen Muhammad
School of Computer Science
University of Windsor
Windsor, Canada
SID: 110059602

***Abstract*—Retail Shrink is nothing but loss of inventory, which is one of the significant problems in today's business world. There are lot of factors that attributes to the inventory loss some of them are employee theft, administrator error, shoplifting, vendor fraud, damage etc., Our focus is to identify and control the shoplifting activity. Over the past decades, the area of shoplifting is attaining more attention as shoplifting accounts forover 35% of annual losses, according to the National Retail Federation's annual survey. Despite employing new technologies to supplement the traditional methods, the inventory loss is still prevalent in retail stores. Researchers are studying to improve the existing techniques and various deep learning and machine learning approaches are being applied to predict and identify the shoplifting activities. In this work, we have implemented Inception v3, Visual Geometry Group-16 (VGG-16) and Residual Neural Network-50 (ResNet50) to detect shoplifting and classify between a normal and a shoplifting event and received accuracy of 68%, 68.1% and 68.1% respectively.**

***Index Terms*—Shoplifting, Artificial Intelligence, Machine Learning, Image Classification, CNN, Deep Learning, videos and Data Augmentation.**

I. Introduction

Computer vision is a branch of computer science that allows computers to mimic the human visual system. It's an artificial intelligence subset that gathers data from digital photos or videos and analyses it to determine properties. The entire procedure entails gathering images, screening, analyzing, identifying, and extracting data. This in-depth processing enables computers to comprehend any visual input and respond appropriately.

To collect multi-dimensional data, computer vision projects convert digital visual input into clear descriptions. This information is subsequently converted into a computer-readable format to facilitate decision-making. The primary goal of this field of artificial intelligence is to teach robots how to get data from pixels.

When we look at a crowd image, our brain can tell who is a known face and who is a stranger, who is a male or a woman, who is a youngster or an adult, and generally what race someone belongs to. Depending on the foreground and lighting, we can also see what people are wearing, who looks put together and who doesn't, and what time of day or season it is.

Computer vision is defined as the provision of sight or visual skills to computers in order to assist them in making better judgments much faster, more efficiently, and consistently than humans could. Computer vision is being more widely used in a variety of businesses throughout the world. Theft detection, retail businesses, healthcare systems, autonomous automobiles, industrial sectors, quality evaluation, and other applications are increasingly relying on computer vision. And by incorporating Computer Vision into their operations, each sector has benefited greatly.

If we believe it, a computer can look at the same image and see nothing, but with computer vision, it can detect and identify all of the faces, give us the ages of everyone in the picture, and even properly tell we their ethnicity. Due to shadows, lighting, and forms, it may have a difficult time establishing the season and time of day, but crowd analytics, verification, and recognition are a breeze.

Deep Neural Networks (DNN) are commonly utilized in Computer Vision techniques because they have better image pattern recognition capabilities. Convolutional Neural Network (CNN, or ConvNet) is a type of deep neural network (DNN) that is most typically used to analyse image patterns.

Surveillance cameras are increasingly being used in public places. Monitoring capabilities of law enforcement agencies has not kept pace. It is difficult for a person to watch multiple monitors at a given instance and identify occurrence of multiple shoplifting events.

Though many machine learning and deep learning methods have been applied in recent many research works, the outcomes indicate that deep learning outperforms machine learning methods.

Convolutional Neural Networks is useful to develop intelligent system to automatically detect an anomaly without human intervention and raise a flag. Our project will process video input fed to the system and percepts the history of anomaly occurrence. Further, it will classify whether there is a suspicious activity taking place in the video that is passed as Input for validation.

In this research paper, we have implemented deep learning techniques and have conducted a comparative analysis to detect shoplifting event. The models we have used for image classification: Inception V3, VGG16, ResNet-50.

II. problem Statement

As per the survey conducted by National Retail Federation (NRF) in 2020, retail shrinkage has caused a loss of around

$62 billion [1]. Our focus is on shrinkage due to shoplifting - theft of goods from an open retail establishment, typically by concealing a store item in pockets, under clothes, or in a bag, and leaving the store without paying.

Shoplifting is a severe problem in the United States. According to Loss Prevention Media. More than 10 million people have been caught shoplifting in the past five years [2]. These stats are compiled by National Association for Shoplifting Prevention (NASP). Shoplifting accounts for over 35% of annual losses, according to the National Retail Federation's annual survey [3].

A. Motivation

With the increase in crime rate, surveillance cameras are increasingly being used in public places. It will be very easy to identify and detect any criminal activities taking place if the actions are being filmed. Monitoring capabilities of law enforcement agencies has not kept pace. Considering shoplifting as one of the criminal acts, it is difficult for a person to identify occurrence of multiple shoplifting events at a any instance if they are monitoring physically (manually). It is important to develop intelligent system with the help of Convolutional Neural Networks and some deep learning techniques to automatically detect an anomaly without human intervention and raise a flag if there is an occurrence of any suspicious activity.

B. Justification

Due to the surge in retail shrink because of shoplifting, efficiently detecting such activities as and when they are happening is much needed and has become critical. This study is highly beneficial as it can be used as a baseline for comparison with future machine learning methods using for shoplifting detection in any retail stores.

III. Literature Review

With the advancement of technology, Surveillance systems have become essential part of safeguarding business or home. Over the last couple of decades, researchers are investigating these systems to improve their functionality to detect any suspicious behavior to prevent shoplifting, robbery, pickpocketing etc. mischiefs without human intervention. The most prevailing techniques to support surveillance systems include tracing, motion detection, face recognition, and human activity recognition [4].

Martínez-Mascorro et al. [4] They suggested a new approach for extracting video segments in which people display behaviors important to the aim of reducing shoplifting crime. These behaviors encompass both regular and suspect actions. They implemented a 3D Convolutional Neural Network (3DCNN) model for extracting the features and classifying skeptical behavior. They presented a novel approach, the Pre-Crime Behavior (PCB) analysis, for processing the films and extracting the suspicious samples (video segments that indicate suspicious behavior). They extract samples from videos by analyzing a film that contains one or more shoplifting offenses and identifying the specific instant when the infraction is committed. Following that, they highlight the various suspicious situations in which a human observer has suspicions about what the individual in the video is doing. Finally, they choose the part in which the suspects are prepared to perform the crime. These segments serve as the Deep Learning (DL) model's training samples. They solely utilized PCB portions for experimenting. These parts are devoid of particular criminal action and include no information concerning a violation. Before attempting to steal from a store, they tried to mimic an aggressor's conduct. And, from their experimentation, they reached an accuracy of 75

An automated system has been proposed in [5] to identify any theft activity in the form of camera frames. Here, the authors implemented a Contouring strategy to distinguish between two adjacent frames of a video. The difference in frame intensities enables to result a movement detection. The only entity that has been considered for the contouring is a person. Also, CNN has been applied to form a suitable model for classifying theft activity.

In [6], the authors implemented CNN model along with transfer learning to identify in real-time three categories of mischiefs: 'Shoplifting', 'Robbery' and 'Break-in'. Pre-trained ImageNet weights have been applied to train the model and ResNet50 is employed to work as backbone of the model. The output frames are labelled as 'normal, 'shoplifting, 'robbery' or 'break-in'. This proposed system gives an alert to the owner whenever a shoplifter hides an item inside a bag or garments. Moreover, this system detects robbers as they cover their faces and use guns at shop staffs.

Ansari and Singh [7] proposed a model to discern shoplifting activities by instigating dual-stream fusion-based network which efficiently cohere with appearance and motion dynamics in the temporal domain. Also, deep inception V3 model is applied to extract activity-specific body position features from real-time video stream. Moreover, Long Short Term Memory (LSTM) is used to construct a relation between obtained features from sequential frames with the aim of differentiating human shoplifting movements. A remarkable accuracy has been achieved in this work which is 91.48%.

Another shoplifting detection system has been proposed in [8] which is based on deep neural network that uses inception V3 model for feature extraction and then uses LSTM network to investigate temporal sequences. The accuracy level gained in this work is 74.53%. [9] proposed an automatic shoplifting behaviour detection system from surveillance videos using Region of Interest (ROI) optical flow fusion network, instead of extracting features from the whole frames. Mask-R-CNN has been implemented to extract a person object as an ROI and the shoplifting activity has been marked by analysing the amount of change in the ROI person object using optical-flow.

The majority of security-support systems are centred on criminal incidence. [14] describe a snatching-detection system that uses backdrop removal and pedestrian monitoring to make a judgement. This method divides the frame into eight sections and looks for a speed shift in one of the monitored people.

[14]'s suggested algorithm can only notify when a person has already lost their items. [15] describe a real-world anomaly detection technique that involves training thirteen abnormalities such as burglary, fighting, shooting, and vandalism. They divide the samples into two categories: normal and abnormal, then utilise a 3DCNN to extract features. For decision making, their approach contains a ranking loss function and trains a fully-connected neural network. [16] offer a technique for detecting lingering individuals. The system incorporates many analyses for decision fusion and final detection, including distance, acceleration, direction-based, and grid-based analysis.

An automated surveillance system, sometimes known as a knight [17], is a system that is used for video surveillance and monitoring via various CCTV, and it operates in a self-contained manner. It can recognise and categorise targets effectively by monitoring the item coherently via several cameras using cutting-edge computer vision algorithms. It generates detailed textual descriptive summary information in the form of an orbit with Google Map tracking site position. This summary will guide police officers in their analysis and rapid reaction decisions. Surveillance's limits the system's shortcomings include the inability to recognise disguised items, detecting things among crowds, controlling crowds, and so on.

Among the early automated monitoring systems, functioning under harsh weather conditions. Background subtraction [18] [19] may also be used to solve problems such as traffic control, visual inspection, and computer-human interaction via moving object identification [20]. In these applications, we have certain objects of interest that may be detected and tracked for their behaviour.

IV. Methodology

A. Material and Data

In this paper, we applied our model on the large UCF Crime Data-set.The UCF-Crime data-set is a massive 128-hour surveillance video collection. It includes 1900 lengthy and uncut video footage of 13 actual irregularities. It includes irregularities such as Abuse, Arson, Arrest, Assault, Burglary, Explosion, Fighting, Robbery, Road-Accident, Shooting, Stealing, Shoplifting, and Vandalism. Additionally, there are a lot of combinations of these criminal activities which made this data-set diverse. It also contains the normal activity video footage to classify among the other criminal activities. These irregularities pose a serious threat to the safety of public. Among them we have prioritized the shoplifting activity to conduct our research.

Our chosen data-set contains video clips of two categories, they are shoplifting and normal videos. Before editing we have 50 video clips of the shoplifting data-set and the clip duration lies between 14 seconds to 37 minutes. On the other side, the duration of the normal videos lies between 7 seconds to 8 minutes and the number of the total clips is the same which is 50. All of the video clips have a resolution of 320 X 240 and frame per second (fps) 30. However, as the length of the shoplifting videos are quite lengthy and the theft occurs quite fast in some of the scenarios. For that reason, we have edited our shoplifting videos to label the shoplifting correctly

B. Proposed Method

The first hurdle we faced before starting to work on modeling, was preparing the data-set. Our data of the shoplifting can be considered as weakly labeled. It means they will have a lot of frames that are of no use to us. Also, some videos only contain suspicious activities which fall under comprehensive crime moment. That is why opted for editing the shoplifting videos by considering some criteria.

Fig. 1. Graphical representation of the proposed methods for editing the video.

- There are various concepts to detect a suspicious behavior [4] and we have utilized the concepts briefly explained below.
  1) **Normal Behavior:** In the CCTV footage most frames will contain this type of behavior where one might not be able to distinguish between crime and typical shopping behavior. Additionally, during this part of the segment the suspect might not be present as well.
  2) **Strict Crime Moment:** This is the part of the clip where actual theft occurs. It contains the decisive factors to identify the thief and the stolen product. Only using this one can be charged as a shoplifter.
  3) **Comprehensive Crime Moment:** The criminal is not always able to steal quickly. Most of the times a criminal wait for the best possible opportunity to steal a product, and during that time he/she shows suspicious behavior. There are quite a lot of unsuccessful attempts in some of the videos which do not help to charge a criminal. However, this is also an important segment to identify a criminal.
  4) **Crime Lapse:** If we merge both comprehensive crime moment and strict crime moment we get the resultant crime lapse. It is the entire scenario where the authority can identify the criminals plan of action, their ways of distracting the shop employees, stealing the product(s), stolen product(s), and escape plan. In a single video there might be more than occurrences of crime lapse segments.
  5) **Pre-Crime Behavior:** This segment contains the information from the first appearance of the suspect

till the comprehensive crime moment starts. There is a little difference between the normal behavior and pre-crime behavior. We have found quite a few videos that have frames that do not have the perpetrator, which we differentiated to be normal segment of a clip.

Fig. 2. Shoplifting Sample Images

- From the concepts above it can be understood that not every part of the videos are of any use to us. That is why in our dataset, we have only prioritized the strict crime moment and shed everything else. By doing that we made our dataset transform from weakly labeled to strongly labeled which means our data-set don't contain any garbage frames of the shoplifting activity. Also, it benefited our overall performance of the project by taking lesser storage, reducing the training time while improving the accuracy.
- **Video to Image:** After segmenting the strict crime moments we converted our videos to frames for training. As the videos were 30 fps, we conducted trial and error to reach a conclusion of picking a single frame in every 10 frames, so for every second we have 3 images.
- **Image Pre-Processing and Augmentation:** It is a part of image pre-processing which assists in altering the existing data by different factors to increase the total number data for model training. It also helps to reduce bias by creating a variety of data within a preferred range which gives the model new information to learn. To augment our images we applied zoom, altered brightness and flipped the images within a given range. To improve the accuracy of our model we have performed image pre-processing on the images that were extracted from the videos. We converted the color from BGR to RGB and resized the image to 224X224. Moreover, to increase the size of our dataset by considering to reduce bias at the same time we have incorporated an image data generator. The new generated images will have following attributes.
  1) Normalized
  2) Rotation range will be 40
  3) Height and width shifting range will be 0.2
  4) Shear range is 0.2
  5) Zoom range is 0.2
  6) Flipped horizontally
  7) Empty spaces of image shift will be covered by nearest pixel fill mode.

C. 3D CNN

Convolutional Neural Network has become famous for image classification tasks due to its groundbreaking layer containing a collection of kernels. It is a black box classifier which has five major components: structure, kernel, receptive field, number of layers and feature maps [23]. The feature extractor of CNN extracts different information of the images which depends upon the layers. Lower-level layers extracts information like edges, corners, etc. and higher the layer more complex features are extracted from the images. These features are then passed through the fully connected layers and flattened into a single dimensional array which afterwards are fed into the classifiers. The classifier outputs the probability score of the image of which class it belongs to.

In two-dimensional graphics Convolutions are applied to the two-dimensional maps produced from the feature to enumerate the features accessible from the geometrical dimension in convolutional neural networks. In the next steps of CNNs, we introduce to enumerate three-dimensional convolutions to gauge the features from both the temporal and geographical aspects must be considered. Convoluting a three-dimensional space yields the 3D convolution.The cube's kernel is created by putting together many spatial temporal patches in a continuous pattern.

The feature maps present in the convolution layer is linked with the multiple frames arranged contiguously in the previous layer in order to capture the motion related information. It is noted that 3D convolution kernel can select only one type of feature from the patch cuboid, provided the kernel weights are duplicated across the patch cube. A common design scheme of Convolution neural networks is the number of feature maps grows as the layers increases there by developing the various multiple types of features from the available lower level of maps. The 3D convolution is obtained by convolving a 3D filter kernel by stacking multiple contiguous frames together to produce the 3D cube. By this operation, the feature maps are connected to multiple contiguous frames.

3D CNN included an another dimension which is temporal, meaning time. Since the way we process the videos, in our project it is not possible to implement 3D CNN

D. Inception V3

Convolutional Neural Networks are used in the Inception V3 deep learning model for picture categorization. The Inception V3 is a more sophisticated version of the basic model Inception V1, which was initially published as GoogLeNet in 2014.

Inception v3 is a convolutional neural network for assisting in image analysis and object detection, and got its start as a module for Googlenet. It is the third edition of Google's Inception Convolutional Neural Network, originally introduced

during the ImageNet Recognition Challenge. The design of Inceptionv3 was intended to allow deeper networks while also keeping the number of parameters from growing too large: it has "under 25 million parameters", compared against 60 million for AlexNet.

Inception, like ImageNet, is a library of identified visual objects that aids object categorization in the realm of computer vision. Many other applications have utilized the Inceptionv3 architecture, which is frequently "pre-trained" from ImageNet. One such use is in the field of biological sciences, where it assists in the study of leukemia. After a famous "'we need to go deeper' internet meme" became widespread, referencing a line from Christopher Nolan's film Inception, the original name (Inception) was codenamed this manner.

By changing earlier Inception designs, Inception v3 primarily focuses on consuming less processing power. This concept was first introduced in the 2015 publication Rethinking the Inception Architecture for Computer Vision. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, and Jonathon Shlens collaborated on it.

Inception Networks (GoogLeNet/Inception v1) have shown to be more computationally efficient than VGGNet, both in terms of the amount of parameters created and the cost incurred (memory and other resources). When making modifications to an Inception Network, it's important to keep the computational advantages in mind. As a result, because of the uncertainty about the new network's efficiency, adapting an Inception network for multiple use cases becomes an issue. Several network optimization strategies have been proposed in an Inception v3 model to remove the restrictions and make model adaption simpler. Factorized convolutions, regularization, dimension reduction, and parallelized calculations are only a few of the approaches used.

Factorization is included in Google's Inception V3 [12], which is considered the greatest significant improvement over the previous version, Inception V2. Its 7x7 convolution may be decomposed into two one-dimensional convolutions (1x7, 7x1), while its 3x3 convolution can be reduced into two one-dimensional convolutions (1x7, 7x1) (1x3, 3x1). The RMSProp optimizer has also been included. The computation may be sped up in this method, and the convolution can be split into two, increasing network depth and non-linearity (ReLU [13] must be performed for each extra layer).

E. VGG16

In their publication "Very Deep Convolutional Networks for Large-Scale Image Recognition," K. Simonyan and A. Zisserman from the University of Oxford proposed the VGG16 convolutional neural network model. In ImageNet, a dataset of over 14 million pictures belonging to 1000 classes, the model achieves 92.7 percent top-5 test accuracy. It was a well-known model that was submitted to the ILSVRC-2014. It outperforms AlexNet by sequentially replacing big kernel-size filters (11 and 5 in the first and second convolutional layers, respectively) with numerous 33 kernel-size filters.

The figure illustrates the ConvNet settings. The nets are known by their names (A-E). All configurations follow the architecture's basic design and differ only in depth: the network A (8 conv. and 3 FC layers) has 11 weight layers, while the network E has 19 weight layers (16 conv. and 3 FC layers). The number of channels in the conv. layers is minimal, starting at 64 in the first layer and growing by a factor of two after each max-pooling layer until it reaches 512.

The convnets are fed a fixed-size 224 by 224 RGB picture during training. The only pre-processing done here is subtracting the mean RGB value derived on the training set from each pixel. The picture is processed through a stack of convolutional (conv.) layers, where filters with a very narrow receptive field are utilized, such as 3 3 (which is the lowest size to capture the notions of left/right, up/down, and center and has the same effective receptive field as one 7 x 7). It's more complex, with more non-linearities and fewer parameters.

Convolution filters, which may be thought of as a linear modification of the input channels (followed by non-linearity), are also used in one of the configurations. For 3 x 3 convolutional layers, the convolution stride and spatial padding of the conv. layer input are both set to 1 pixel, ensuring that the spatial resolution is kept after convolution. Spatial pooling is aided by five max-pooling layers that follow parts of the convolutional layers. Max-pooling is done with stride 2 across a 22 pixel frame.

The ConvNet training technique is often carried out by employing mini-batch gradient descent (based on back-propagation) with momentum to optimize the multinomial logistic regression goal. The batch size and momentum were both set to 256 and 0.9, respectively. For the first two fully-connected layers, the training was regularized via weight decay (the L2 penalty multiplier was set to 5 104) and dropout regularization (dropout ratio set to 0.5).

The initialization of the network weights is critical, as poor initialization might cause learning to halt owing to the gradient's instability in deep nets. To get around this, we started training configuration A, which is shallow enough to be learned with random initialization. The first four convolutional layers and the last three fully-connected layers with the layers of net A (the intermediate layers were randomly started) are then initialized for training further architectures. The pre-initialized layers' learning rate is not reduced, enabling them to vary throughout learning. The weights from a normal distribution with a zero mean and 102 variance are sampled for random initialization (where appropriate). The biases were set to zero at the start.

Following a stack of convolutional layers (which have varying depths in different designs), there are three Fully-Connected (FC) layers: the first two have 4096 channels apiece, while the third performs 1000-way ILSVRC classification and so comprises 1000 channels (one for each class). The soft-max layer is the last layer. In all networks, the completely linked levels are configured in the same way. AlexNet (ILSVRC — 2012 winner) replaced huge kernel-sized filters with numerous 3×3 kernel-sized filters one after the

other in the VGG16 architecture.

To obtain the fixed-size 224 x 224 ConvNet pictures, the ConvNet input images were randomly cropped from rescaled training images (one crop per image every SGD iteration). The crops were randomly horizontally flipped and randomly RGB color shifted to add to the training set (Krizhevsky et al., 2012). The process of rescaling training images is described below.

VGG is a CNN which is a light weight deep neural network that made records in ImageNet classification, reaching more than 90% accuracy and it was considered as a state of the art at that time. It is also known as an extended version of AlexNet. VGG became famous for its approach by deepening its layers by using smaller convolutional blocks of 3X3 filters, 2X2 pooling layers etc. for feature learning.
Figure 3 depicts the architecture of VGG16:

F. ResNet 50

A residual neural network (ResNet) is a type of artificial neural network (ANN) that is based on pyramidal cell constructions in the cerebral cortex. Skip connections, or shortcuts, are used by residual neural networks to hop past some layers. The majority of ResNet models use double- or triple-layer skips with nonlinearities (ReLU) and batch normalization in between. To learn the skip weights, an extra weight matrix can be utilized; these models are known as HighwayNets. DenseNets are models that have several parallel skips. A non-residual network is referred to as a plain network in the setting of residual neural networks.

There are two key reasons to add skip connections: to avoid vanishing gradients and to minimize the Degradation (accuracy saturation) problem, in which adding additional layers to a sufficiently deep model results in increased training error. [1] During training, the weights adjust to muffle the upstream layer and magnify the previously skipped layer. Only the weights for the link between neighboring layers are changed in the simplest instance, with no explicit weights for the upstream layer. When only one nonlinear layer is stepped over, or when all intermediate levels are linear, this method works well. Otherwise, an explicit weight matrix for the missed connection should be learnt (a HighwayNet should be used).

In the early phases of training, skipping layers effectively simplifies the network[clarification needed]. Because there are fewer layers to propagate through, the influence of disappearing gradients is reduced, which speeds up learning. As the network learns the feature space, it gradually recovers the skipped levels. When all layers are enlarged near the conclusion of training, it stays closer to the manifold[clarification needed] and so learns quicker. The feature space is explored further by a neural network with no remnant pieces. This makes it more susceptible to disturbances that lead it to depart off the manifold, and thus takes additional training data to recover.

Because deep neural networks take a long time to train and are prone to overfitting, a Microsoft team developed a residual learning framework to speed up the training of networks that are far deeper than those previously utilized. In

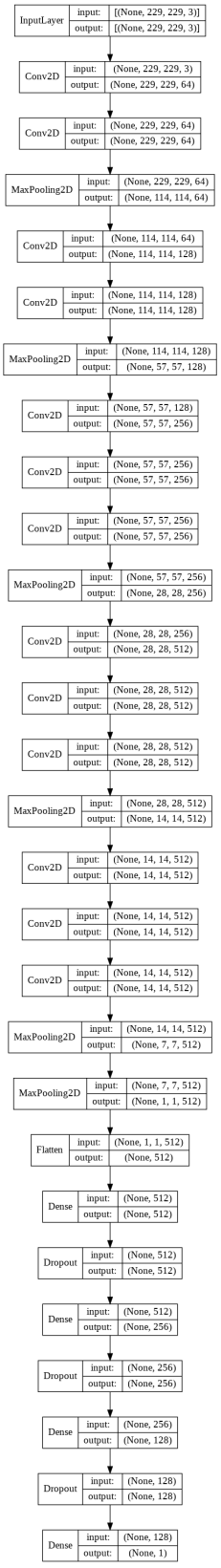

Fig. 3. VGG16 Architecture used.

2015, the results of this study were reported in the publication Deep Residual Learning for Image Recognition. As a result, the well-known ResNet (short for ”Residual Network”) was created.

When training deep networks, there comes a moment where the accuracy reaches a saturation point and then rapidly degrades. The ”degradation problem” is the term for this. This demonstrates that not all neural network topologies are created equal.

To address this problem, ResNet employs a method known as ”residual mapping.” Rather of assuming that every few stacked layers would suit a desired underlying mapping, the Residual Network allows these layers to fit a residual mapping explicitly. The building block of a Residual network is shown below.

ResNets may be used to solve a variety of problems. They’re simple to tune and attain increasing accuracy as the network’s depth grows, resulting in better outcomes than earlier networks. ResNet was trained and tested using ImageNet’s approximately 1.2 million training photos belonging to 1000 distinct classes, much like its predecessors.

ResNets are quite simple to grasp when compared to traditional neural network topologies. A VGG network, a basic 34-layer neural network, and a 34-layer residual neural network are shown below. The layers in the simple network have the same amount of filters for the same output feature map. The number of filters doubles when the quantity of output features is halved, making the training process more difficult.

Meanwhile, as we can see, the Residual Neural Network has considerably fewer filters and lesser complexity during training than the VGG. A shortcut link is added, which transforms the network into its residual counterpart. This shortcut connection conducts identity mapping, padding the dimensions with extra zero entries. There is no extra parameter introduced by this option.

ResNet-50 model (Residual Network) is a Convolutional Neural Network (CNN) that has 50 layers. It is widely used in the area of Computer vision to solve problems related to image recognition, image classification, face recognition, object detection etc. The use of 3-layer bottleneck blocks as well as ‘Skip Connections’ approach improves accuracy of model and reduce training time of data [21].

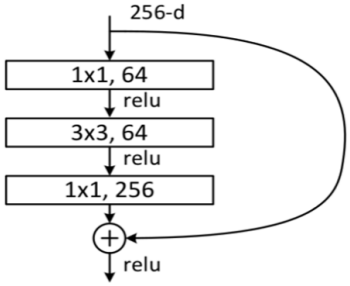

Fig. 4. Residual Block of ResNet-50

Here, we have considered the input size as 224 X 224 X 3. This model performs the initial convolution and max-pooling using 7 x 7 and 3 x 3 kernel size respectively [22].

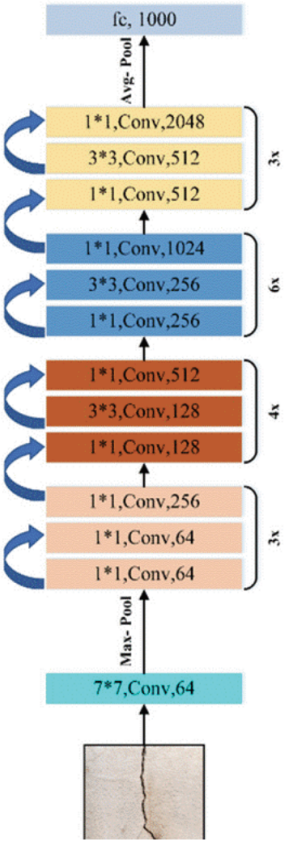

Fig. 5. ResNet-50 Architecture

G. Conditions and Assumptions

These are the assumptions that we considered for the implementation of this study: 1. The datasets used in this project have true information and are not biased. 2. The LIAR and LIAR-PLUS labels have been merged to perform similarly in the real world. 3. Some columns have been filtered, and these are the list of columns that we used in each dataset: The statement and subject columns in the LIAR dataset, the statement and justification columns in the LIAR-PLUS dataset, and the title column in the ISOT dataset are considered.

H. Simulation Analysis

In this study, we first split each dataset into train and validation sets using train_test_split. Then, we converted all inputs and labels into torch tensors, the required datatype for our model. The DataLoader needs to know our batch size for training, a batch size of 16 or 32 is recommended, so we used 32 in this experiment. After the completion of each training epoch, we measured our performance on the validation set.

V. Computational Experiments

A. Evaluation metrics

Accuracy: Accuracy is the ratio of correct predictions to total predictions. An accuracy metric is used to measure the algorithm's performance in an interpretable way and is usually determined after the model parameters and is calculated in the form of a percentage. It is the measure of how accurate the model's prediction is compared to the true data.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

Loss: A loss function is used to optimize a machine learning algorithm. The loss is calculated on training and validation, and its interpretation is based on how well the model is doing in these two sets. It is the sum of errors made for each example in training or validation sets. Loss value implies how poorly or well a model behaves after each iteration of optimization. As for the loss function, Cross entropy is used for evaluation in this project.

As the final part of our evaluation, the model was checked against the test set. The data required for this was created, and the model's 'evaluate' method in TensorFlow was used to compute the metric values for the trained model. This method returns the loss value and metrics values for the model in test mode.

B. Implementation Details

In this approach, for shoplifting event identification, we validate our proposed system by randomly taking 20% of the frames generated out of dataset as a testing dataset and 80% as training data. We fine-tune the training options like mini-batch size, learning rate, and number of epochs for the training. We set the mini-batch size to 32 for faster processing and the number of epochs were set to 10. We have performed some data augmentation operations on the training images to improve some accuracy. The operations include image resizing according to network input where the image sizes are varying in the dataset to make distinct image channels, randomly flipping in the y-direction, rotating and rescaling.

Figure 6 represents the architecture and flow of data in our proposed system and following are the actions taken place at every stage:

1) Both Shoplifting and Normal videos are considered categorically as input to the system.
2) Once the videos are received the system will process them and extract the required number of frames and store them in image format.
3) After extracting the Images in previous step we have augmented the images by re-scaling,re sizing etc. to improve accuracy.
4) The augmented data is fed to the CNN Models that is either RESNET50 or VGG16 or InceptionV3 for training.
5) Once the models are trained the system outputs the test accuracy and loss.

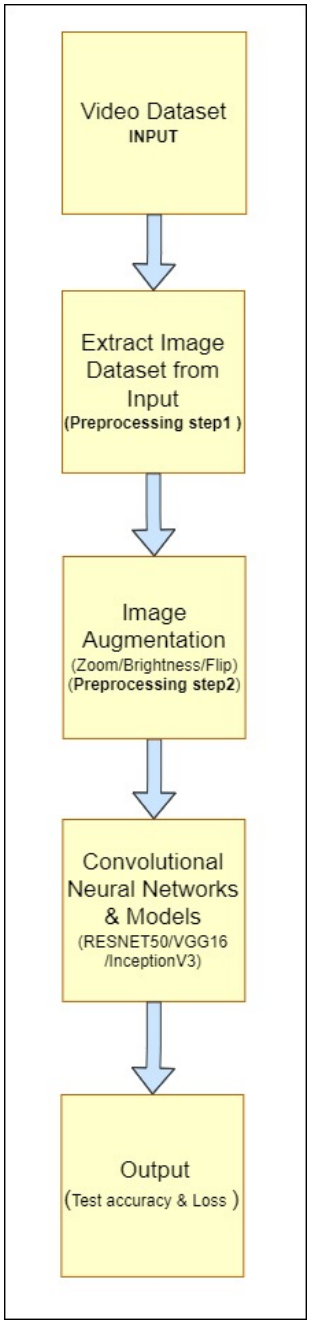

Fig. 6. Proposed system Architecture

C. Results and Analysis

In the experimental setup, the results obtained after running the CNN Machine learning techniques on shoplifting dataset which consists of 100 videos from which we have extracted frames of size 229 x 229 that has images/frames of two categories, such as videos when shoplifting event occurred and the videos without any suspicious activities i.e Normal videos. To construct the algorithms, we use Python3 IDE in Google Compute Engine (Google Collab), Scikit Learn Libraries as Knowledge sources. This experiment is performed on Duo Core with 2.20 GHz CPU and 12GB RAM running

on Linux Platform. Comparison result after running each CNN algorithms given below:

In this approach, for shoplifting event identification, we validate our proposed system by randomly taking 20% of the frames generated out of dataset as a testing dataset and 80% as training data. We fine-tune the training options like mini-batch size, learning rate, and number of epochs for the training. We set the mini-batch size to 32 for faster processing and the number of epochs were set to 10. We have performed some data augmentation operations on the training images to improve some accuracy. The operations include image resizing according to network input where the image sizes are varying in the dataset to make distinct image channels, randomly flipping in the y-direction, rotating and rescaling.

In our experimentation, we tested 3 different models on our dataset. The first was Inception V3 which gave an accuracy of 68% and a loss of 0.6288. The second model which we implemented was VGG 16 which had an accuracy of 68.1% and loss of 0.636. The third model which we implemented was Resnet50 which had an accuracy of 68.1% and loss of 0.630.

Graphs in Figure 7 and Figure 8 depicts the trends of training and validation metrics for loss and for accuracy in InceptionV3 model. Figure 7 represents depicts the training accuracy and validation accuracy Vs Num of epochs and Figure 8 represents

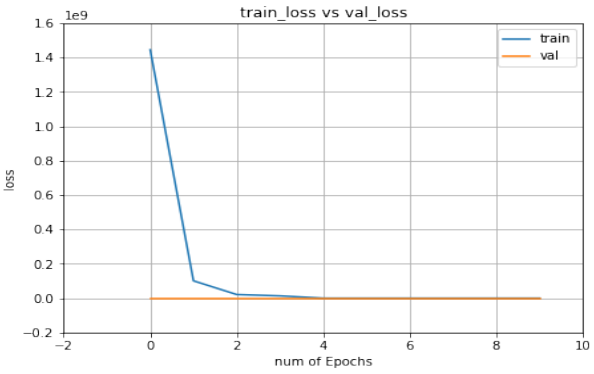

Fig. 7. Traning VS Validation Accuracy Mapping for InceptionV3 for 10 epochs

depicts the training loss and validation loss Vs Num of epochs when model InceptionV3 model is used.

Graphs in Figure 9 and Figure 10 represents the trends of training and validation metrics for loss and for accuracy in VGG16 model. Figure 9 represents the training accuracy and validation accuracy Vs Num of epoch and Figure 10 depicts the training loss and validation loss Vs Num of epochs when model VGG 16 model is used.

Graphs in Figure 11 and Figure 12 represents the trends of training and validation metrics for loss and for accuracy in Resnet50 model. Figure 11 depicts the training accuracy and validation accuracy Vs Num of epochs and Figure 12 depicts the training loss and validation loss Vs Num of epochs when

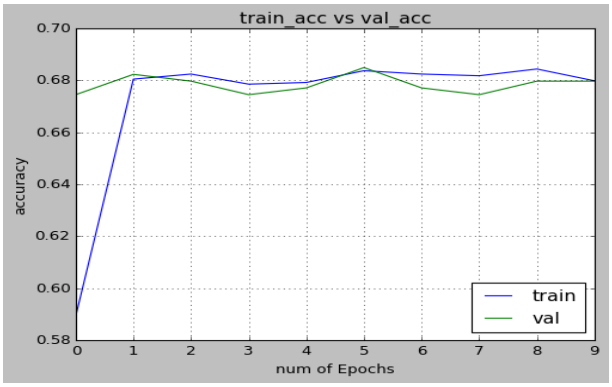

Fig. 8. Traning VS Validation Loss Mapping for InceptionV3 for 10 epochs

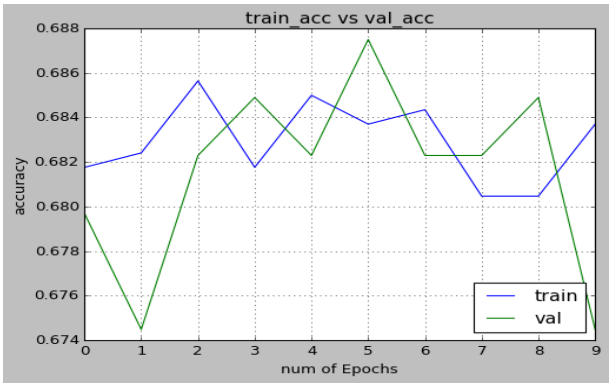

Fig. 9. Traning VS Validation Accuracy Mapping for VGG16 for 10 epochs

model Resenet50 model is used.

TABLE I
STATISTICS OF METRICS USED IN DIFFERENT ALGORITHMS

CNN Algorithm	Validation accuracy	Validation Loss	Test Accuracy	Test Loss
Inception V3	0.6849	0.6245	0.6810	0.6270
VGG16	0.6745	0.6312	0.6810	0.6267
Resenet50	0.6771	0.6305	0.6810	0.6280

As we can see in the table I, the accuracy and loss of all the 3 models are almost the same. We can infer that for less epochs and limited data, all the models perform the same. To see differences in the model we need to have more data. This could be achieved by extrapolating new frames from the video and stacking it. However, due to computational resources we refrain from it.

VI. CONCLUSION

A. Summary

Our goal in this project is to explore ways to mitigate shoplifting using computer vision. We use UCF Crime dataset to achieve this. We notice that the UCF Crime dataset is weakly labelled, meaning if duration of shoplifting happens for 30 seconds, the length of the video labelled shoplifting

Fig. 10. Traning VS Validation Loss Mapping for VGG16 for 10 epochs

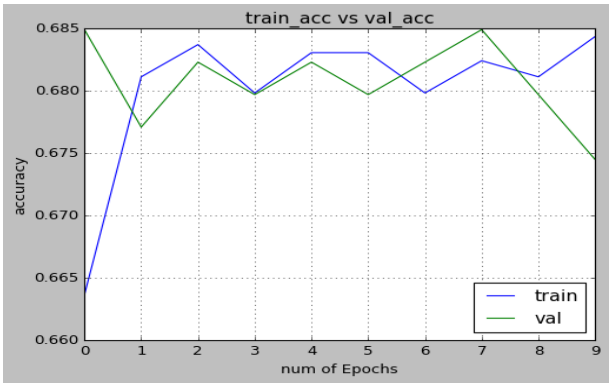

Fig. 11. Traning VS Validation Accuracy Mapping for Resnet50 for 10 epochs

is over 3 minutes. Therefore, we manually edit the videos to increase the accuracy and reliability of our models. We then convert the videos to images. We then preprocess the images. We then implement 3 differnet models namely VGG 16, Resnet50 and Inception V3.

On our dataset, we experimented with three distinct models. The first was Inception V3, which had a 68 percent accuracy and a 0.6288 loss. The second model we used was VGG 16, which had a 68.1 percent accuracy and a loss of 0.636 percent. Resnet50 was the third model we used, and it had an accuracy of 68.1 percent and a loss of 0.630.

B. Future Research

There are a lot of scopes for further extend this project of ours. There are several ways we can go about doing this. For further research, we can add more classes to test our system. Moreover, as we have manually edited the videos by visually judging the criminal activity, research can be conducted to turn this into an automated process. Which will extract the frames by the type of category we select. Furthermore, there is another option to consider weighting the frames to get the optimum number of frames so that our training uses lesser resources and become faster. This will also open the door to the real-time shoplifting detection from just the pre-crime behavior and comprehensive crime moment.

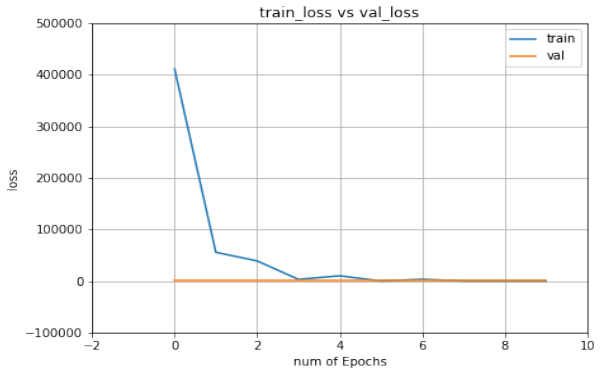

Fig. 12. Traning VS Validation Loss Mapping for Resnet50 for 10 epochs

C. Open Problems

Although our implemented models successfully detected shoplifting, it can be further improved by using applying or taking a different approach than ours. deeper networks, ensemble techniques and transfer learning. In this part, we'll discuss several unsolved difficulties that need to be investigated further: 1. Designing a video classification model with high performance on any shoplifting data 2. Using several CNN Model modifications to compare and evaluate results. 3. Incorporate a deeper network 4. Larger data-set 5. Using transfer learning.

ACKNOWLEDGEMENT

The project code and final results are accessible on the link: shorturl.at/fvAG3

REFERENCES

[1] "Retail shrink totaled $61.7 billion in 2019 amid rising employee theft and shoplifting/ORC", NRF, 2022. [Online]. Available: https://nrf.com/media-center/press-releases/retail-shrink-totaled-617-billion-2019-amid-rising-employee-theft-and. [Accessed: 03- Jan- 2022].

[2] "2018 National Retail Security Survey", Cdn.nrf.com, 2022. [Online]. Available: https://cdn.nrf.com/sites/default/files/2018-10/NRF-NRSS-Industry-Research-Survey-2018.pdf. [Accessed: 03- Jan- 2022].

[3] "The National Association for Shoplifting Prevention", NASP, 2022. [Online]. Available: https://www.shopliftingprevention.org/. [Accessed: 03- Jan- 2022].

[4] G. Martínez-Mascorro, J. Abreu-Pederzini, J. Ortiz-Bayliss, A. Garcia-Collantes and H. Terashima-Marín, "Criminal Intention Detection at Early Stages of Shoplifting Cases by Using 3D Convolutional Neural Networks",Computation, vol. 9, no. 2, p. 24, 2021. Available: 10.3390/computation9020024.

[5] K. Singh, D. Arora and P. Sharma, "Identification of Shoplifting Theft Activity Through Contour Displacement Using OpenCV",Computational Methods and Data Engineering, pp. 441-450, 2020. Available: 10.1007/978-981-15-6876-3_34 [Accessed 6 December 2021].

[6] O. Rajpurkar, S. Kamble, J. Nandagiri and A. Nimkar, "Alert Generation on Detection of Suspicious Activity Using Transfer Learning", 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020. Available: 10.1109/icccnt49239.2020.9225263 [Accessed 6 December 2021].

[7] M. Ansari and D. Singh, "An expert video surveillance system to identify and mitigate shoplifting in megastores", Multimedia Tools and Applications, 2021. Available: 10.1007/s11042-021-11438-2 [Accessed 6 December 2021].

[8] M. Ansari and D. Singh, "An Expert Eye for Identifying Shoplifters in Mega Stores", Advances in Intelligent Systems and Computing, pp. 107-115, 2021. Available: 10.1007/978-981-16-3071-2_10 [Accessed 6 December 2021].
[9] U. Gim, J. Lee, J. Kim, Y. Park and A. Nasridinov, "An Automatic Shoplifting Detection from Surveillance Videos (Student Abstract)",Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 10, pp. 13795-13796, 2020. Available: 10.1609/aaai.v34i10.7169 [Accessed 6 December 2021].
[10] N. Aggrawal, "Detection of Offensive Tweets: A Comparative Study", Computer Reviews Journal, 1(1), pp. 75-89, 2018.
[11] S. Hamidian, M. Diab, "Rumor identification and belief investigation on twitter", In Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 3-8, 2016.
[12] Qian et al., "Fresh Tea Leaves Classification Using Inception- V3," 2019 IEEE 2nd International Conference on Information Com- munication and Signal Processing (ICICSP), 2019, pp. 415-419, doi: 10.1109/ICICSP48821.2019.8958529.
[13] A. Krizhevsky, I. Sutskever, G. E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks," Neural Information Processing Systems, pp.1097–1105, 2012
[14] Hiroaki Tsushita and Thi Thi Zin. A study on detection of abnormal behavior by a surveillance camera image. In Thi Thi Zin and Jerry Chun-Wei Lin, editors, Big Data Analysis and Deep Learning Applications, pages 284–291, Singapore, 2019. Springer Singapore.
[15] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6479–6488, 06 2018.
[16] Tatsuya Ishikawa and Thi Thi Zin. A study on detection of suspicious persons for intelligent monitoring system. In Thi Thi Zin and Jerry Chun-Wei Lin, editors, Big Data Analysis and Deep Learning Applications, pages 292–301, Singapore, 2019. Springer Singapore.
[17] Singh, D. K., Paroothi, S., Rusia, M. K., Ansari, M. A. (2020). Human crowd detection for city wide surveillance. Procedia Computer Science, 171, 350-359.
[18] Javed, O., Shafique, K., Shah, M. (2002, December). A hierarchical approach to robust background subtraction using color and gradient information. In Workshop on Motion and Video Computing, 2002. Proceedings. (pp. 22-27). IEEE.
[19] Goswami, P. P., Paswan, D., Singh, D. K. (2016). Detecting moving objects in traffic surveillance video. International Journal of Control Theory and Applications, 9(17), 8423-8430.
[20] Wren, C. R., Azarbayejani, A., Darrell, T., Pentland, A. P. (1997). Pfinder: Real-time tracking of the human body. IEEE Transactions on pattern analysis and machine intelligence, 19(7), 780-785.
[21] "Deep Residual Networks (ResNet, ResNet50) - Guide in 2021 - viso.ai", viso.ai, 2022. [Online]. Available: https://viso.ai/deep-learning/resnet-residual-neural-network/. [Accessed: 03- Jan- 2022].
[22] "Detailed Guide to Understand and Implement ResNets - CV-Tricks.com", CV-Tricks.com, 2022. [Online]. Available: https://cv-tricks.com/keras/understand-implement-resnets/. [Accessed: 03- Jan- 2022].
[23] J. Chang and J. Sha, "An efficient implementation of 2D convolution in CNN", IEICE Electronics Express, vol. 14, no. 1, pp. 20161134-20161134, 2017. Available: 10.1587/elex.13.20161134.

APPENDICES

In Project :
Roisul Islam Rumi : Data preprocessing.
Naga Jyothirmayee Dodda : VGG16 Implementation.
Sumaiya Deen Muhammad : ResNet50 implementation.
Prem Shanker Mohan : InceptionV3 implementation.

In Report :
Roisul Islam Rumi : Data preprocessing,Data Augmentation,Future work, Conclusion
Naga Jyothirmayee Dodda : Introduction, Abstract, Problem Statement, Methodology -Preprocessing,Computational Experiment - VGG16
Sumaiya Deen Muhammad : Literature Review, Methodology - ResNet 50, Computational Experiment - ResNet50
Prem Shanker Mohan : Problem Statement, Methodology - Inception v3, 3DCNN , Computational Experiment - Inception v3