Object Detection in Live Streaming Video Using Deep Learning Approach

Jeberson Retna Raj, Senduru Srinivasulu
Department of Information Technology, Sathyabama Institute of Science and Technology, Chennai, India

Abstract. Object detection in a complex scene is of challenge as the image contains numerous features. Object detection is important in surveillance, healthcare, recognition, manufacturing, military and agriculture etc. The traditional approaches are limited in accuracy and mainly focussing on image datasets. With the application of deep learning algorithm, the accuracy can be increased. In this paper, the Convolutional Neural Network (CNN) is introduced to detect the objects in the image frames. The input of video is decomposed into image frame and the particular image frame is considered for processing. The image is preprocessed and the noises are removed for further processing. In the next step, feature extraction process contains edge detection and extracting the meaningful shape and texture in the image. The shape which contains the higher level feature is then feed into convolutional layer. The convolutional layer consist of four parts such as number of filter, filter size, pooling layer and max pooling layer. With the application of activation function Relu, the objects can be identified. Furthermore, the Gradient descent algorithm is applied for optimization to improve the accuracy. The system has been tested with ImageNet and the results are very promising.

1. Introduction
Data Science is a broad area which employs Artificial Intelligence, Machine Learning and Deep learning. Artificial intelligence means that to make intelligent externally for any kind of the system. Furthermore, it taking decision based on insight of the available data. Machine learning allows machine to crunch data and construct knowledge and intelligence. It combines applied statistics and optimization theory. Deep learning is the sub topic of machine learning which focuses on building, training and deploying large and deep neural networks. The neural network works based on the data in parallel. The journey of the problem can be defined as data science is a path, Machine Learning is a tool and AI is the Goal. The term analytics can be defined as the combination of discern and analysis made from a given data source. There are three types of analytics defined in literature. First descriptive analytics focus on what happened to the system or change of state after applying the algorithm. The fundamental is database and what we seen as a dashboard. The second is diagnosis analysis deals about root cause analysis of the problem. Data mining is the fundamental source to discover the facts in the data. The data mining techniques aims to discovering facts from wealth of information. Diagnosis analysis is a level 2 which comprising of data mining and ETL. The third is predictive analytics deals with historical things. The prediction is based on the historical data. The mathematical model comprising AI and ML allows for solving the problem. The final one is prescriptive analytics deals with decision making process. It focuses on the term ‘what I do next’ phrase. It is a antology or knowledge representation techniques. For performing this task, a toolkit is required with computational form. Data science employs with the techniques data handling and harnessing of data.
with the system. It focuses on creation, manipulation, transformation and diffusion of row archived data into information turned knowledge with support of decision.

The need for deep learning is to address the N-gram feature representation of data. The existing techniques are limited in accuracy and performance. In essence, the deep learning focuses on reusing the existing trained models. The convolutional Neural Network is one such algorithm to solve the n-gram feature vector of the image or data. The word convolution is derived from the image processing applications. The canny filter is used to identifying the edges and corners from n layers and detects the edges and boundaries of the given image. The first layer is feed with the raw data and all intermediate each subsequent layer are used to extract the data. The n-th layer is for displaying the final output. The classification is divided into binary classification and multiclassification. The binary classification techniques are used to classify the output based on one neuron at the output layer. Whereas, the multiclass classification used to classify the data with multiple class labels. In essence, the neural network passes the input value along with the weight value. The activation function used to process the data with the weight values. Many activation functions are available in literature which includes sigmoid,Tannet and Relu functions for processing hidden layer neurons. The softmax is the activation function mainly used for multiclass classification problems.

For multilayer perceptron, the following key functions are required. First, we need to define the cost function, back propagation, learning rate alpha, optimizer and regularizer parameters. The optimizer may be gradient descent or any other similar algorithm can be used. The epoch used to perform one forward and backward pass of all training examples. In essence, there are two problems can occur in these kind of processing. The over-fitting and under-fitting problems are exist in classification. The gap between the training accuracy and validation accuracy is defined as over-fitting. The under-fitting is categorized as the training itself the model not provide accuracy. To solve this over-fitting problem the drop out mechanism has been used. It randomly removes connection in hidden layer. The second important solution to the over-fitting problem is based on data augmentation technique. It is based on increase the amount of data based on available data. The parameter epoch is used to fine tune the model for early stopping of iteration.

1.1. Literature Survey

Lex et al introduced a dilated kernel based CNN algorithm is used to classify the earth surfaces of satellite images. The method replaces the traditional kernel with the dilated kernel to increase the classification accuracy. The hybrid dilated CNN model is used to address the information loss problem [1]. Lin Yuan et al proposed a CNN model for extracting the image feature of MRI brain images to classify tumor cells. The authors use multiple convolutional layers to extract the gradient information in images and combine the spatial features [2]. In [3], an ensemble based CNN method is applied with the medical images for classification. The method is applied with various dataset like Googlenet and Alexnet for measuring the performance of the method. Hou et al proposed an object detection method in panchromatic images using spatial template matching method. The approach making use of single shot multibox detector (SSD) method to extract the multi scale features used to classify the image and identify location [4]. Tang et al proposed an object detection mechanism to linking the objects in the same frame. Furthermore, cuboid network proposal method used to extract the spatio temporal features in the image [5]. Fang et al proposed a faster R-CNN algorithm for object detection. The R-CNN experiencing low object experiences in bounding boxes, to increase this Faster R-CNN algorithm is introduced [6]. Haijun et al compared the performance of different state-of the art algorithms for object detection in UAV satellite images. The algorithms successfully detect buildings, trucks, car, building collapse etc. The algorithm detects the bigger objects and unable to detect smaller objects [7]. Vipul et al presented a detailed survey on object detection using deep learning algorithms. They are focusing on image dataset and not applicable for video data [8]. Yong Tian et al introduced a metal object detection system using SVM algorithm for wireless charging technology for electric vehicles [9]. In [10], object detection in UAV images is presented. Several deep learning algorithms
includes Faster RCNN, cascade RCNN, R-FCNN, YOLO and its variants are discussed. These algorithms are applied for the static images and not applied for video data.

2. Proposed Methodology

The proposed method using video as the source of input and the object of interest is detected as output. The video is decomposed into image frames and the selected frame is then taken for detecting objects. In the first step, the features of the images are extracted in the image. Not all the pixel is feed into the neural network, the shape and structure is enough to identify the meaningful features in a given image. Using feature extraction process, the edges of the image can be detected. With the help of the edges, the shapes can be determined. Then the shape used to detect the higher level features such as vehicle, people, sign board, etc. The first layer is convolution layer which consist of four parts such as number of filter, filter size, pooling layer and max pooling layer. The sequence is important in CNN method. First starts with convolutional layer by perform the feature extraction process. The low pass filter is used to smooth the edges and high pass filter used to identify the edges in the image. Basically, the filter used to train the image. In our approach, sobel operator is used to train the image. For corner pixel feature, we use the 0 padding concept.

The kernel value is defined using the formula. If H*W and F*F are defined as 5*5 and 3 * 3, then K can be computed as

\[ K = \frac{H - F + 2P}{S} + 1 \]

Where s is a stride value. For example H=5, F=3 and P=0, the filter size is odd. This means that the filter size is odd which used to fitting center pixel. Padding can be defined using the formula

![Figure 1. Flow of object detection using CNN Algorithm](image)
\[ p = \frac{F - 1}{2} \]

where \( p = 1 \)

Next we have to define the max pooling layer. The maxpooling layer is used to reduce the spatial pattern. There are two pooling is available, one is max pooling and another is average pooling. Max pooling layer is used to reduce the pixel window size. The maximum pixel value is considered for reducing the window size. For example 4*4 window size is reduce into 2*2 window. In our system, the max pool is reduced into 2*2 and the stride value which makes the filter is go with a non-overlapping fashion. Furthermore, maxpooling provides better results.

The convolutional neural network is an important algorithm in deep learning algorithms. The applications of CNN which includes object identification, face recognition, classification, etc. in essence, the computer process the image as array of pixels of height, width and dimension. For example 6*6*3 as RGB color image and 6*6*1 as grayscale image. Generally, CNN algorithm works with train and test data. The process involves of convolutional filter, that is kernel, pooling, fully connected layer and finally apply softmax function to classify the objects with the probabilistic values between 0 and 1.

2.1. Convolutional layer

Convolution is the first layer used to train the image by extracting the image features from an input image. It requires two input, one is sample window \( h \times w \times d \) and kernel input \( fh \times fw \times d \) and the output is convolutional featured 2D image.

2.2. Stride

Stride used to shift the number of pixels over the input image. stride equals 1 means one pixel shift and stride value 2 means shifting with two pixel positions in the image. in our work, the stride value is set as 2.

2.3. Padding

Padding is used to fit the kernel in the image window. There are two paddings used to fit the kernel in the image region. To cover the boundary region of the image zero padding mechanism used. otherwise, the part of the image can be apply with the kernel filter which doesn’t cover corner pixels.

2.4. Relu

The softmax function Relu stands for Rectified Linear Unit.

\[ f(x) = \max(0, x) \]

The Relu function used to apply the input data for getting non linearity in the convolutional network. This provides the CNN with non-negative linear values. The other functions Tanh and Sigmoid can be used instead of Relu. However, the Relu provides higher performance comparing with others.

2.5. Pooling layer

Pooling layer used to reduce the dimension in the feature map. There are three pooling are used for dimensionality reduction of the image feature. They are max pooling, average pooling and sum pooling. The max pooling takes the maximum pixel value in the image grid of (2*2) with the stride value of 2 and the average pooling is to sum of all pixel in the image grid and take average. The sum pooling is to sum of all elements in the selected image grid.

2.6. Fully connected layer

The fully connected layer used to flatten the matrix into vector data. The input vector data is feed into neural network. We create a model with the help of these fully connected layers to combine the features. With the help of softmax function, we can classify the objects. Finally, the Gradient descent algorithm is used for optimization to increase the accuracy.
3. Experimental Result
The proposed system is implemented using the CNN algorithm. The video input is processed and decomposed into image frame and the image is used for further processing. The image is preprocessed and the noise is removed. The sobel operator is used to detect the edges and corner pixels. The shape is identified and based on the shape the higher level features are identified. These higher level features are feed into CNN layer. The CNN network consist of sequence of steps containing convolutional filter with the activation function, maxpooling. Figure 1 shows the input image of road view is considered and the CNN algorithm is applied. The image consists of many objects like car, person, bus, auto etc. The system is trained such a way that it can able to detect the objects separately. The image is divided into meaningful segments. The region of proposal is extracted from the image.

![Figure 2. Sample image for detecting the objects](image1)

![Figure 3. Output image with objects detected using the CNN algorithm](image2)

Figure 2 shows that the objects which are detected in a video frame using CNN algorithm. The objects car, man tree are detected based on the training of data. The output clearly shows that the algorithm successfully detecting the objects.

| Algorithm                  | Accuracy (%) |
|----------------------------|--------------|
| Random forest              | 85           |
| CNN with Gradient descent  | 92           |

Table 1 shows that the accuracy of the proposed method of object detection with gradient descent algorithm. With this algorithm, the accuracy is increased from 85% to 92%.

4. Conclusion
Detecting objects in a streaming video is complex due to the depth of information available in each frame. The video is sliced into image frames and the image of interest will be identified for processing. The images are divided into training and testing set for experiment. The system successfully processed with the images. The CNN algorithm with the Gradient descent optimization algorithm is used and the results are compared. The results are very promising when the gradient method is integrated with the CNN. The system can be extended with Fast RCN method for similar data sets.

References
[1] Lei, X., Pan, H., & Huang, X. (2019). A dilated CNN model for image classification. IEEE
[2] Yuan, L., Wei, X., Shen, H., Zeng, L., & Hu, D. (2018). Multi-center Brain Imaging Classification Using A Novel 3D CNN Approach. IEEE Access, PP: 124082-124095, 1–1. doi:10.1109/access.2018.2868813

[3] Kumar, A., Kim, J., Lyndon, D., Fulham, M., & Feng, D. (2017). An Ensemble of Fine-Tuned Convolutional Neural Networks for Medical Image Classification. IEEE Journal of Biomedical and Health Informatics, 21(1), 31–40. doi:10.1109/jbhi.2016.2635663

[4] B. Hou, Z. Ren, W. Zhao, Q. Wu and L. Jiao, "Object Detection in High-Resolution Panchromatic Images Using Deep Models and Spatial Template Matching," in IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 2, pp. 956-970, Feb. 2020, doi: 10.1109/TGRS.2019.2942103.

[5] P. Tang, C. Wang, X. Wang, W. Liu, W. Zeng and J. Wang, "Object Detection in Videos by High Quality Object Linking," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 5, pp. 1272-1278, 1 May 2020, doi: 10.1109/TPAMI.2019.2910529.

[6] F. Fang, L. Li, H. Zhu and J. Lim, "Combining Faster R-CNN and Model-Driven Clustering for Elongated Object Detection," in IEEE Transactions on Image Processing, vol. 29, pp. 2052-2065, 2020, doi: 10.1109/TIP.2019.2947792.

[7] Haijun Zhang, Mingshan Sun, Qin Li, Linlin Liu, Ming Liu, Yuzhu Ji, “An empirical study of multi-scale object detection in high resolution UAV images”, Neuro Computing 421(20201), PP:173-182

[8] Vipul Sharma Roohie Naaz Mir, A comprehensive and systematic look up into deep learning based object detection techniques: A review, Computer Science Review 38 (2020), 100301

[9] YongTian, ZhengLi, Yawen Lin, LijuanXiang, Xiaoyu Li, Yonghong Shao, JindongTian, Metal object detection for electric vehicle inductive power transfer systems based on hyperspectral imaging, Journal of Measurement 168 (2021) 108493

[10] Payal Mittal, Raman Singh, Akashdeep Sharma, Deep learning-based object detection in low-altitude UAV datasets: A survey, Image and Vision Computing, 104 (2020) 104046