International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 16 (2024) pp. 278–292

Article

Analyzing Visual Content across Frames to Identify Objects and Generating Textual Descriptions for Images

Soumya Upadhyay [1], Kamal Kumar Gola [2,*], Ravindra Kothiyal [3] and Mridula [2]

[1] Department of Cyber Security, College of Smart Computing, COER University, Roorkee 247667, Uttarakhand, India; ersoumya26@gmail.com
[2] Department of Computer Science and Engineering, College of Smart Computing, COER University, Roorkee 247667, Uttarakhand, India; mridulaiitr@gmail.com
[3] Department of Computer Science and Engineering, National Institute of Technology, Jote 791112, Arunachal Pradesh, India; ravindrakothiyal@outlook.com
* Correspondence: kkgolaa1503@gmail.com

Received date: 1 March 2024; Accepted date: 21 March 2024; Published online: 10 July 2024

Abstract: Generating an image description is called a caption. Caption generators have been growing popularity since 2012. Every day, viewers come across many images from different sources that they want to interpret themselves, such as the Internet, TV, news articles, social network sites and more. Though generating caption for an image is a challenging task but recent advancement in Computer Vision has made it possible that a machine can be a storyteller of an image. Most of the time, we do not have images with descriptions. But normal human can easily understand those images. If humans want an automatic image caption from a machine, then machine needs to interpret some image captions. Image Caption Generation model is mostly based on Encoder-Decoder approach like a language translation model. Language translation model uses RNN (Recurrent Neural Network) for Encoding as well as Decoding but Image Caption Generation model uses CNN (Convolutional Neural Network) for Encoding and RNN for Decoding. In this paper, the machine takes an image as input and process a sequence of words as output (caption).

Keywords: CNN-convolutional neural network; RNN-recurrent neural network; RCNN-Region-based convolution neural network; LSTM-Long-short Term Memory; YOLO; SVM-support vector machine

1. Introduction

Object identification is a daunting and demanding topic in computer vision that has gotten a lot of attention [1]. Deep learning improves object detection by training more abstract high-level features and hierarchical data representation with low-level features [2]. For multi-classification tasks, the deep learning-based object identification method outperforms standard detection algorithms in terms of robustness, accuracy and speed. Object recognition approaches based on deep learning are primarily those based on regional recommendations and those that is based on a unified pipeline structure. The first technique creates a number of area recommendations from a target picture, then extracts features from the defined regions and creates an object classifier using a sophisticated neural network.

An older technique of introducing CNN algorithm into the field of object detection was the region-based convolution neural network (RCNN) method [3]. It employs a convolution neural network to fetch the characteristics from the generated region proposals after using a selection search strategy to develop region ideas from the input photographs. The support vector machine is trained using the characteristics gathered. Fast RCNN [4] and Faster RCNN [5] were proposed based on the RCNN approach to shorten learning time and enhance mean average precision. Although region proposal-based approaches have

given increased detecting precision, but the method's layout is really complicated and object identification takes much time than other algorithms [6]. The latter method (which is based on a unified pipeline framework) uses a single feed forward convolutional neural network to predict item placement and class probabilities directly from the entire image, eliminating the need for feature vector construction[7,8]. As a result, the unified pipeline framework technique has a basic structure and can recognize items rapidly, but for regions. It is, however, less precise than that of the proposal-based approach. Both systems have their own set of benefits and are best suited to distinct applications. The unified pipeline framework-based approach is the focus of this paper. In recent years, This YOLO version 2 method [9] is one of the unified pipeline framework-based methodologies presented by researchers. To improve recall, YOLO version 2 uses batch normalization to improve connections and reduce overfitting, as well as anchor boxes to anticipate bounding boxes.

Other improvements include a classifier with high sensitivity, live position forecasting, dimensional clustering, and training on multiple scales, all of which boost object's identification effectiveness. Pedoeem and Huang previously suggested a simplistic actual- time identification alternative for non-GPU programs rely on the YOLO version 2 method [10]. By half the size of the target image and eliminating batch normalization of shallow layers, their method minimizes the number of model parameters. The Fast YOLO approach, introduced by Shafie et al., enables YOLOv2 to be used on embedded devices [11]: This method creates an efficient network architecture optimization using a scalable deep intelligence framework. The motion adaptive inference framework can employ the updated network architecture to boost-up the identification procedure and thereby lower the embedded device's power consumption. Simon et al. [12] devised a tough YOLO approach for RGB picture recognition that uses a complicated regression algorithm to predict multi-glass 3D boxes in Cartesian space. The Single Shot Multi Box Detector (SSD) approach was invented by Liu et al. [13] to detect items of various sizes, it builds multi-scale object maps. This strategy maintains an equilibrium between speed and efficacy in terms of identification. Their approach builds a multi-scale detecting structure to enhance the identification rate of minor objects. Redmon and Farhadi propose utilizing the YOLOv3 approach to use binary cross-entropy losses for class forecasting [14] and scale prediction to predict boxes at different scales to improve detection rate for small items.

The Figure 1 shows that the new caption-based captioning method primarily uses visual space [15] and deep learning methods, but rarely uses multimodal space for captions. Visual Space - In this approach, image features and associated Subtitles are passed independently to the audio decoder. Multimodal Space-Contrastingly, the shared multimodality is picked up from the query image and its corresponding label, and the multimodal space is transferred to the language decoder. Kiros et al. [16] has in advance proposed his paintings on Multimodal area method. Where CNN is used to extract features from an image and the usage of multimodal area represents each concatenation of photograph and textual content for the multimodal illustration learning and photograph captioning. In the equal paper, this approach further introduces multimodal neurolinguistic models, including the Modality-Biased Log-Bilinear model (MLBL-B) and the Factored 3-way Log Bilinear data analysis model (MLBL-F) included with Alex Net. [17]. this version of multimodal neurolinguistic models suffers from huge datasets and it has a longstanding flashback problem. Kiros et al. [18] then expanded their picture of this project [18] in which LSTM was used for encoding and an entirely new version of the language model was launched, namely the LSTM version i.e., Structure Content neural language model. The SC-NLM has a good point that it can transform a sentence structure to its content, which is generated by the encoder. Deep machine learning-based image description technique is primarily based on complete image subtitle strategies and uses specific acquisitions of knowledge from the learning strategies such as supervised learning, unsupervised learning, and reinforcement learning. Encoder-decoder or compositional approaches are used in picture annotation tools. Chi Wang et al. [19] has proposed a Geometry Attention Transformer based on Geometry and position relations.

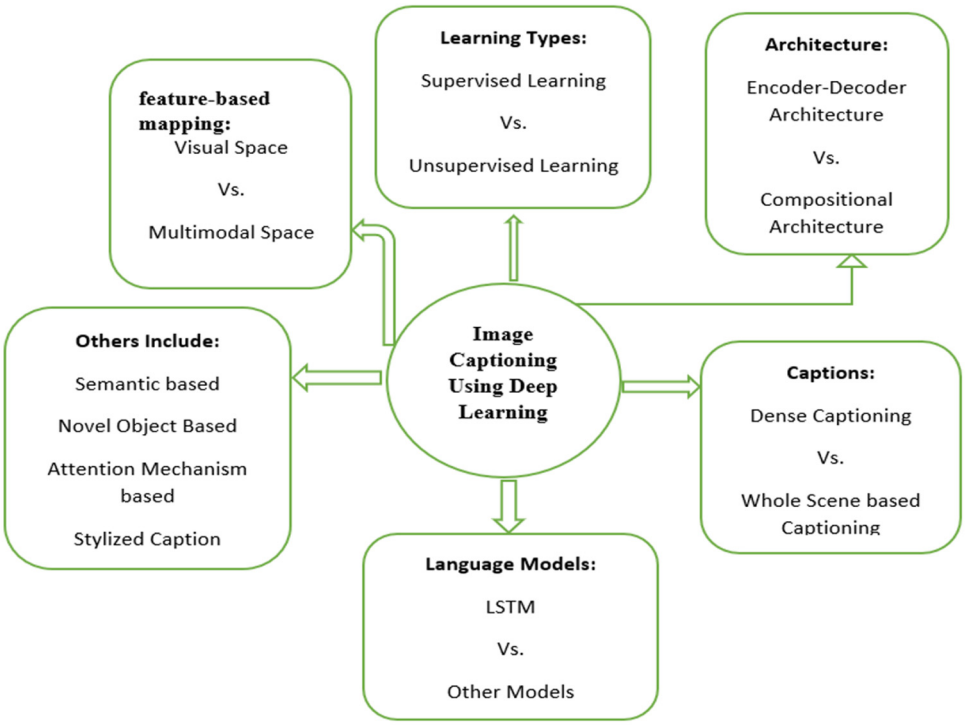

Figure 1. Deep learning-based captioning methods.

They have used Geometry self-attention Refiner to the transformer for the accurate object positioning. It nevertheless keeps a conventional encoder-decoder approach. The encoder uses Geometry self-attention refinement function for the optimization of image representation, with decoder side, a group of position-LSTMs are designed for accuracy of the word sequencing. The authors in [20] has proposed a very effective method in the field of captioning. Although most of the time we see that captioning is based on a single language model but with CaMEL approach, we have image captions based on interplay between two distinct language models. This work has taken the advantage of Mean teacher learning approach, where two different language models learn together during training period. While either language model is used as a teacher, others work as students in the distillation of knowledge. Generating image captions can be a daunting task, but recent advances in computer vision have made it possible for machines to become image storytellers. Deep learning is known for solving the intricacy and grueling of subtitles. To generate caption is similar to generating description like a human brain of an image by the machine. After a thorough study we found that CNN is used for understanding Image Contents and RNN is used for generating descriptions. The combination of CNN+LSTM has proven the better results than the combination of CNN+RNN. The best two mechanisms we found during studies that have been giving a better accuracy are Encoder-Decoder and Attention Mechanism. Nowadays, the use of Transformers has gained a popularity in Image Captioning as it has improved the quality of captions and giving an impressive performance. It has seen that for the better subtitles/captioning, the Geometry and Position relations of different visual objects is one of the critical information.

Here, we are going to explain two different techniques for Image Captioning:

- Machine learning methods that have been used in the past.
- Techniques for deep machine learning.

In traditional techniques based on machine learning, feature extraction is done manually such as directed gradient histogram [21], local binary model [22], etc. Then the manual features are transmitted to a classifier like SVM [23] for object classification. Problem with Traditional Technique- It was only a task specific and feature extraction of large datasets was not feasible. In Deep machine learning approaches, on the other hand, the features are automatically learned from the training datasets, and this technique works well for diverse and tedious data. SoftMax is utilized as a classifier, whereas CNN is mostly employed for feature extraction.

2. Literature Survey

2.1. "Object Detection and Recognition Using a One-Step Improved Model, YOLO v3-Tiny, 2020"

The algorithm for object detection has undergone numerous improvements in order to increase performance in terms of both speed and accuracy. Deep learning algorithms are continuously evolving, with increased object detection performance, thanks to the tireless work of many researchers. Pedestrian detection, medical imaging, robotics, self-driving automobiles, face recognition, and other prominent applications decrease human effort in a variety of industries. It is extremely difficult to include everything at once due to the broad domain and many new techniques. This article covers two types of object detectors to provide a fundamental understanding of object detection methods. RCNN, fast RCNN, and faster RCNN are among the algorithms covered by the two-stage detector, whereas single-stage identifiers are very much concerned with speed, while two-stage detectors are more concerned with accuracy. We'll show you how to use YOLO v3Tiny, an upgraded version of YOLO, and then compare it to other object detection and recognition algorithms.

2.2. "Automated Image Capturing System for Deep Learning-Based Tomato Plant Leaf Disease Detection and Recognition, 2018"

Smart farming, which employs the appropriate infrastructure, is a novel method for increasing the quantity and quality of agrarian production in the state, including tomato production. Diseases cannot be prevented because tomato plant farming takes into account a variety of factors such as the atmosphere, soil, and amount of sunlight. Deep learning-enabled breakthroughs in computer vision have paved the door for camera-assisted disease diagnosis in tomatoes. This research resulted in the creation of a unique approach for disease detection in tomato plants. To identify and distinguish leaf diseases, a motor-controlled picture capturing box was built to capture four sides of each tomato plant. Diamante Max, a special tomato variety, was employed as a test subject. Powdery mildew, leaf borer, and target spot disease are among the ailments that the system can detect. We make a deep complicated NN to recognize three diseases or deficiencies using a dataset of 4,923 leaf pictures from healthy and diseased tomato plants collected under controlled settings. To assess whether tomato illnesses were present in the observed tomato plants, the system used a sophisticated neural network. The trained anomaly detection model from F- RCNN had an accuracy of 95.75 percent, whereas the transfer pathology identification model had an efficiency of 80 percent. The computer-generated image capture model was tested in the field and found to be 91.67 percent accurate in detecting tomato plant leaf diseases.

2.3. "Design and Implementation of High-Speed Background Subtraction Algorithm for Moving Object Detection, 2018"

In the applications of intelligence system such as vehicle navigation, surveillance, and people tracking, object detection is a critical and difficult problem. CCTV is a critical tool in the fight against terrorism and the administration of public safety. Detecting running objects from video is critical for object identification and behavioral analysis in video surveillance. Identification of the running objects in a video stream is a crucial rendering step, and foreground segmentation using background subtraction is a frequent method. A fast background subtraction technique for motion object identification is proposed in this research. To obtain a smooth image, first, we divide the video into a stream, then subjected to a convolution filter, that eliminates components of high-altitude noise. The smoothed pictures are then used to create the detected object in the background image using an adaptive thresholding background subtraction algorithm. The recognized item is subsequently passed through a convolution filter, which improves image quality by removing incorrectly deformed pixels. The suggested method that was created using the VHDL programming language and the Spartan 6 FPGA kit (XC6SLX452csg324). In terms of features and operation speed, the proposed strategy helps performance the previous method. Table 1 shows the gist of few existing work.

Table 1. *G*ist of some other Literature Survey [8–14].

Reference	Image Encoder	Language Model
Kiros et al., 2014 [24]	Alex Net	LBL
Kiros et al., 2014 [25]	Alex Net, VGG Net	LSTM SC-NLM
Karpathy et al., 2014 [2]	Alex Net	DTR
Mao et al., 2015 [3]	Alex Net, VGGNet	RNN
Lu et al., 2017 [4]	ResNet	LSTM
Chen et al., 2017 [5]	VGGNet, ResNet	LSTM
Aneja et al., 2018 [9]	VGG Network	Language Convolutional NN

There are numerous ways for identifying and recognizing things in the proposed system, each with a trade-off between speed and accuracy. However, we cannot assert that one algorithm is superior to another. It is always possible to choose the method that is most appropriate for one's needs. Object detection applications have gained a lot of traction in a short period of time, and due to its broad span of research, there is still a lot to learn about this topic [10,13,26]. Varying approaches for recognizing and localizing objects of various sizes in the input image are compared in this study in terms of accuracy, time, and parameter values. We discovered a new single-stage model methodology that improves speed without losing accuracy. The results of the comparison reveal that YOLO v3-Tiny improves object recognition speed while maintaining accuracy. Object recognition and localization can also be extended from static images to a movie including a dynamic stream of images.

3. Proposed Work

3.1. YOLO V3 Model for Real Time Object Detection

Object detection is regarded as the most difficult task in computer vision. There are several object detection algorithms, but none of them generate the buzz generated by YOLOv3. YOLO, or You Only Look Once, is considered the most popular object detection algorithm. YOLO versions 1–3 were created by Joseph Redmon and Ali Farhadi. One of the fastest object detection techniques is YOLO (You Only Look Once). It's a great option if you need real-time detection without sacrificing too much precision. If you need real-time detection without compromising too much precision, this is a wonderful alternative. To begin, YOLO v3 uses a Darknet variant with a 53-layer ImageNet-trained network. For the detection task, a total of 53 more levels are added, so that we get 106-layers of underlying architecture for YOLO v3. Residual skip connections and up-sampling connections are included in the new architecture. The fact that v3 detects at three different scales is its most striking feature. YOLO is a fully integrated network with a 1 × 1 kernel applied to the feature map as the final output. Detection in YOLO v3 was accomplished by applying 1 × 1 detection kernels to feature maps of three different sizes at three separate network sites. The identified kernel has the structure 1 × 1 × (B × (5 + C)). B stands for the number of bounding boxes that a feature map cell may predict, "5" stands for a feature's four bounding box attributes and confidence, and C stands for the number of classes. YOLO v3 generates predictions at three different sizes, which are accurately determined by down sampling the input image size to 32, 16, and 8 pixels, respectively. The initial detection is handled by the 82nd layer. Because the network down samples the image for the first 81 layers, the

81st layer has a stride of 32. The generated feature map will be 13 × 13 if the image is 416 × 416. A 1 × 1 detection kernel is used here, resulting in a detection feature map of 13 × 13 × 255. The layer 79 feature map is then sampled from 2 × to a size of 26 × 26 before being processed through multiple sophisticated layers of convolution. The layer 61 feature map is concatenated with the layer 61 feature map in terms of depth. To a fused feature map, the few 1 × 1 convolution layer are applied (61). The second detection is then performed by the 94th layer, likely to result in an identification feature map with aspects of 26 × 26 × 255. Layer 91's feature map is subjected to multiple layers of complexity before being coupled at depth to layer 36's feature map in the same way. As before, there are several strata present. After that, a 1 × 1 compound layer is added to integrate information from the previous class (36). On the 106th layer, we complete step 3 by constructing a 52 × 52 × 255 feature map.

YOLO is a Convolutional neural network for real-time object detection (CNN). CNNs are pattern-recognition frameworks that use classifiers to interact with incoming images as structured data arrays. Other object detection models have the benefit of being significantly faster and more accurate than

YOLO.

3.2. Anchor Boxes

There are a total of 9 anchor boxes in YOLO v3 as shown in Figure 2. Each scale has three levels. The K-Means subgroup should be used to produce 9 anchors, if someone using their own data to train YOLO, they will need anchors. Then, in one dimension, arrange the anchors in a way from highest to the lowest. The three largest anchors should be assigned to the first ladder, the next three to the second, and the final three to the third.

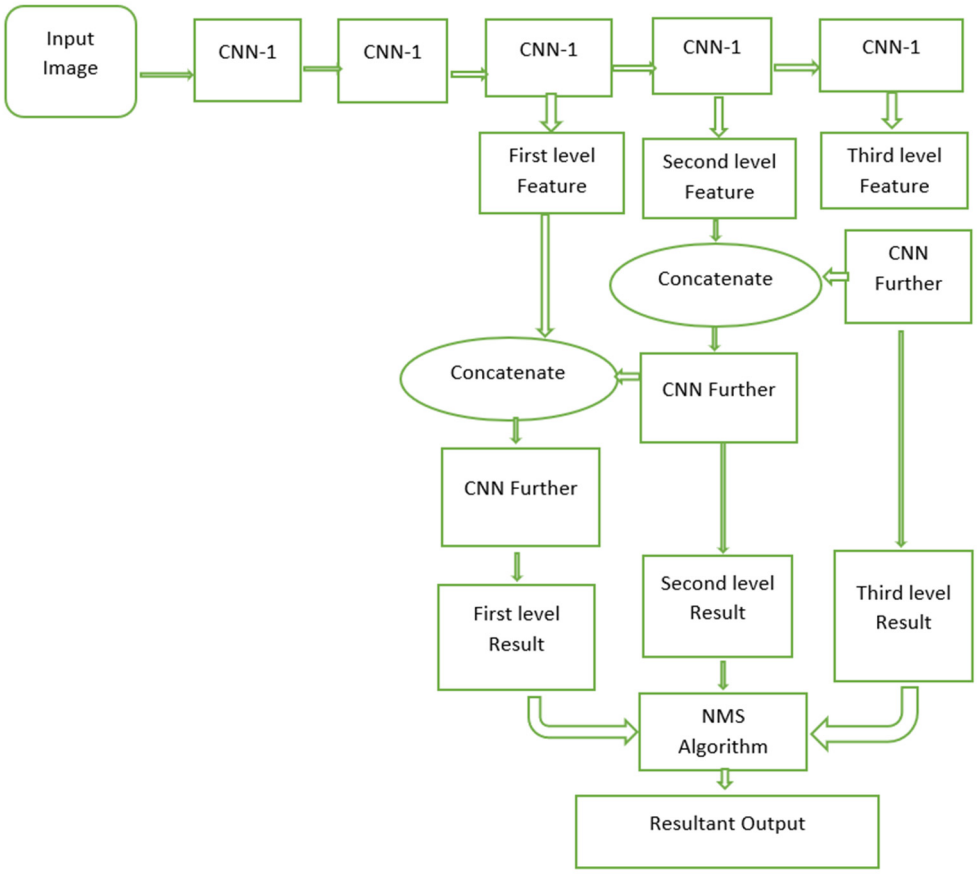

Figure 2. Working Model of YOLO.

4. Implementation

YOLO can be implemented using the Keras or OpenCV deep learning libraries as shown in Figures 3 and 4.

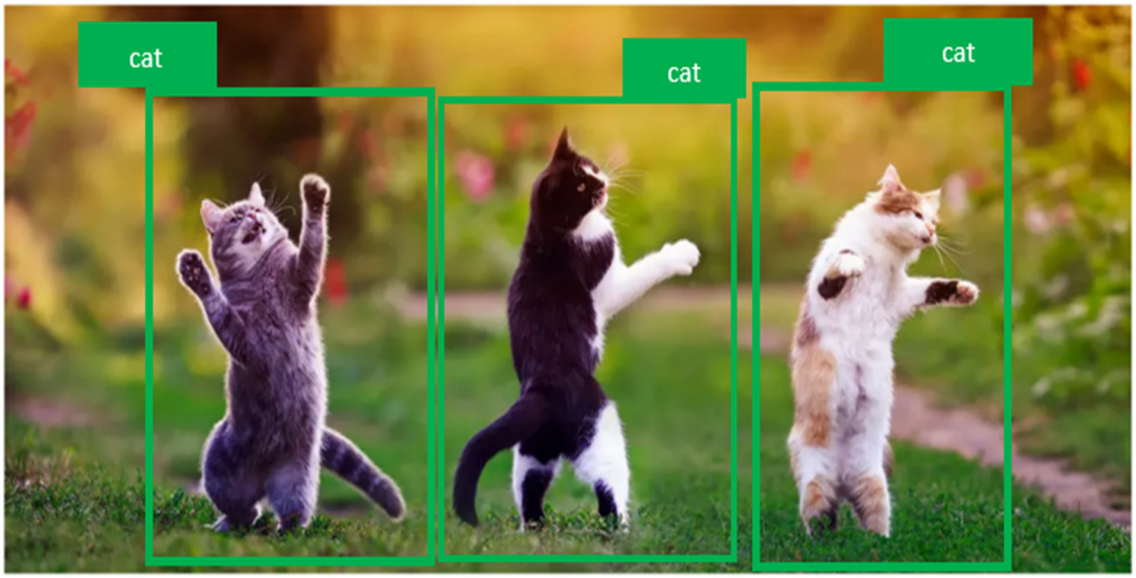

Figure 3. YOLO bounding boxes for object identification.

Figure 4. YOLOv3 is extremely accurate in detecting all the objects.

4.1. Image Captioning—"An Storytelling Approach"

Image captioning is the process of creating captions for an image. Since 2012, Image Caption Generation has grown in popularity.

Every day, we are bombarded with images from a variety of sources, including the Internet, television, and news stories. Social networking sites, and so on, where viewers are encouraged to interpret them for themselves. Most of the time, we do have images with descriptions. But normal human can easily understand those images. If humans want an automatic image caption from a machine, then machine needs to interpret some image captions.

Image Caption Generation model mostly based on Encoder-Decoder Approach like a Language-Translation Model. Language Translation model uses RNN (Recurrent Neural Network) for Encoding as well as Decoding but Image Caption Generation model uses CNN (Convolution Neural Network) for Encoding and RNN for Decoding.

In this project, Machine takes an image as an input, and generates sequence of words as an output

(Captions).

4.2. Proposed Models for Image Captioning

I have used the combination of Residual Network-01 and Long-Short Term Memory for the description of the images.

4.3. Residual Neural Network-101

ResNet (Residual Network) is a kind of NN developed by Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun in their paper "Deep residual learning for image recognition" [27] in 2015. Residual neural network (ResNet) is a well-known deep learning model that was first published in 2015 in the paper "Deep Residual Learning for Image Recognition" [27].

One of the learning models is the ResNet model as shown in Figure 5, it is the most popular and successful deep learning method [28]. This is a highly deep feedback neural network with hundreds of levels, significantly deeper than prior neural networks that just used a few layers to jump over. There are two primary reasons to include a jump coupling: to prevent the degradation problem from disappearing and to alleviate the degradation problem (accuracy saturation), which occurs when adding more layers to a sufficiently deep model results in larger training mistakes.

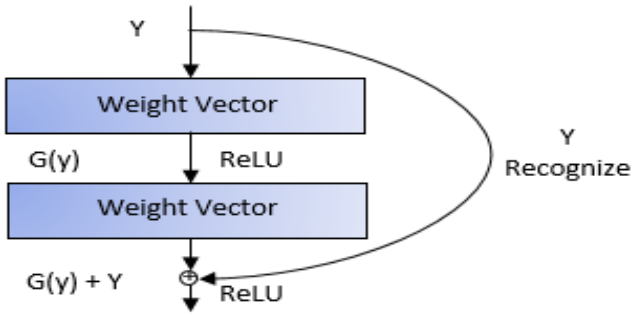

Figure 5. Architecture ResNet 101 is a 101-layer deep convolutional neural network [9].

Figure 6 shows the architecture of ResNet that helps to generate hundreds or even thousands of classes while still achieving convincing performance. Many computer vision applications beyond picture classification, such as object identification and face recognition, have improved as a result of its powerful representation [29,30].

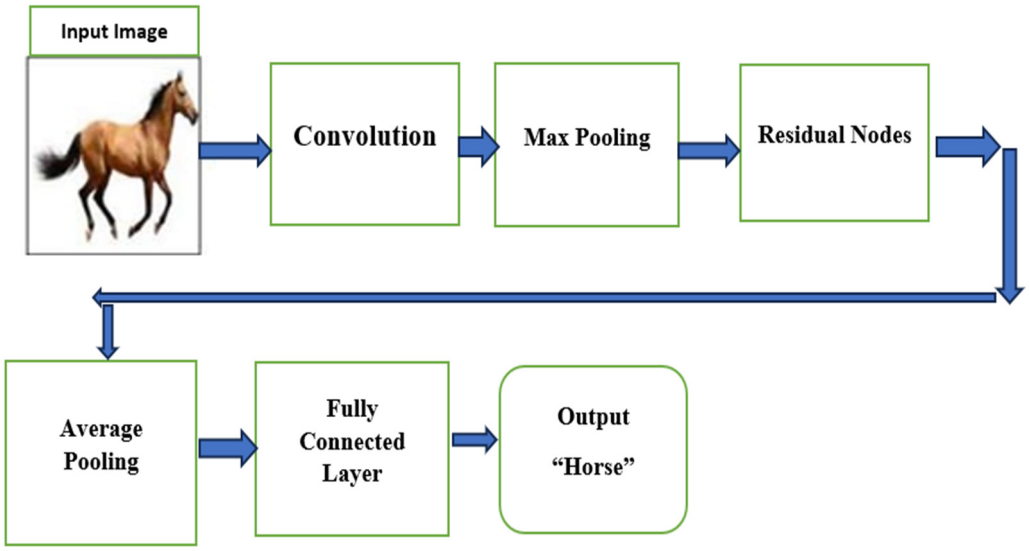

Figure 6. Architecture of ResNet-101.

4.4. An Rnn Model for Description Generation (LSTM)

Long-term dependencies can be overcome using LSTMs, a recurrent neural network that can solve the problem of disappearing gradients that RNNs have. A cyclic neural network, sometimes referred to as RNN, is a type of permanent memory system. As indicated in the Figure 7, The LSTM is divided into three portions, each with its own set of functions. The cell tries to extract updated information from the input during the second phase [31].

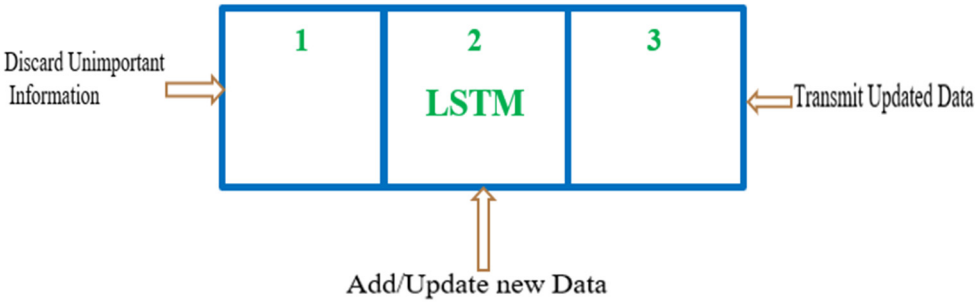

Figure 7. LSTM Model for Description Generation.

Finally, the cell transmits update data from an existing timestamp towards the next timestamp in the third section. An LSTM cell has three parts: gates, switches, and transistors. The forget gate comes first, then the input gate, and then the output gate.

Video Dataset for Video Object Detection shown in Figures 8–10.

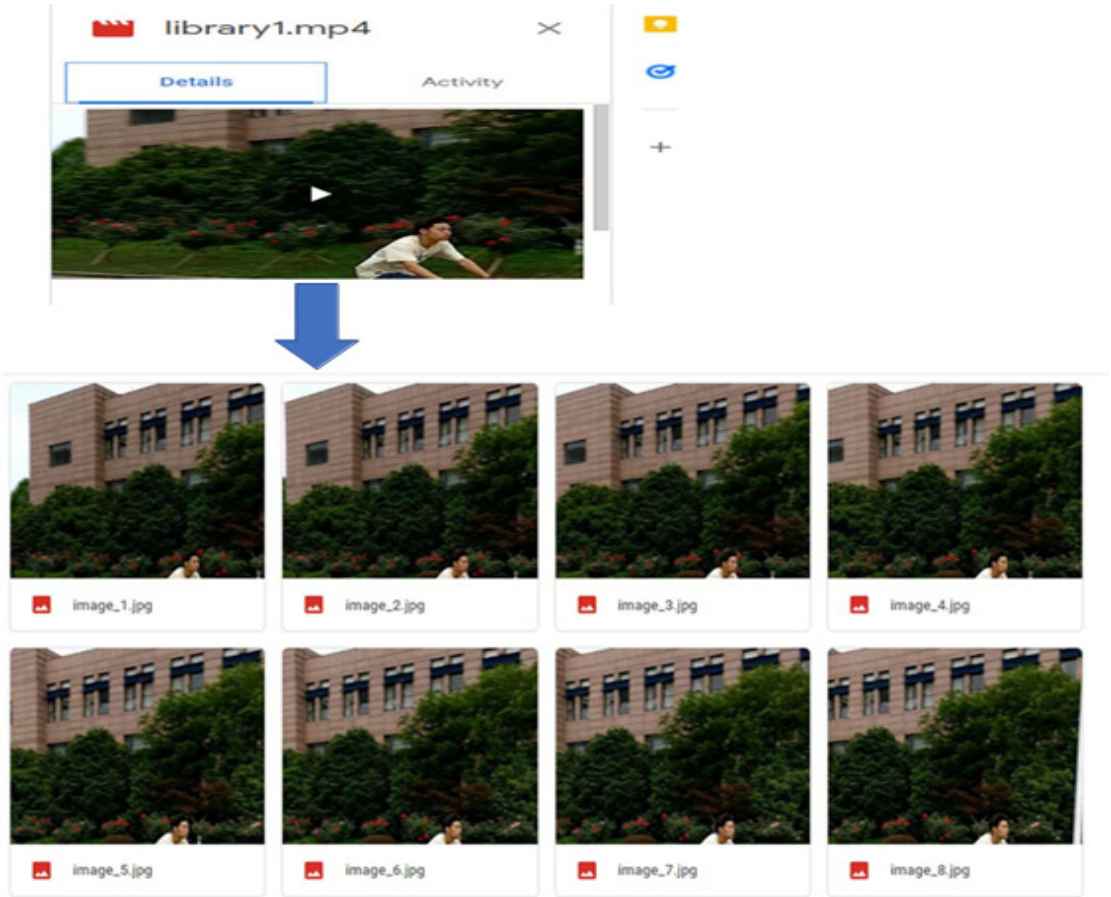

Figure 8. Datasets for YOLO V3.

In dataset, selected a video dataset and then converted that video into frames and saved those frames in a different folder.

Images Dataset for Image Captioning:

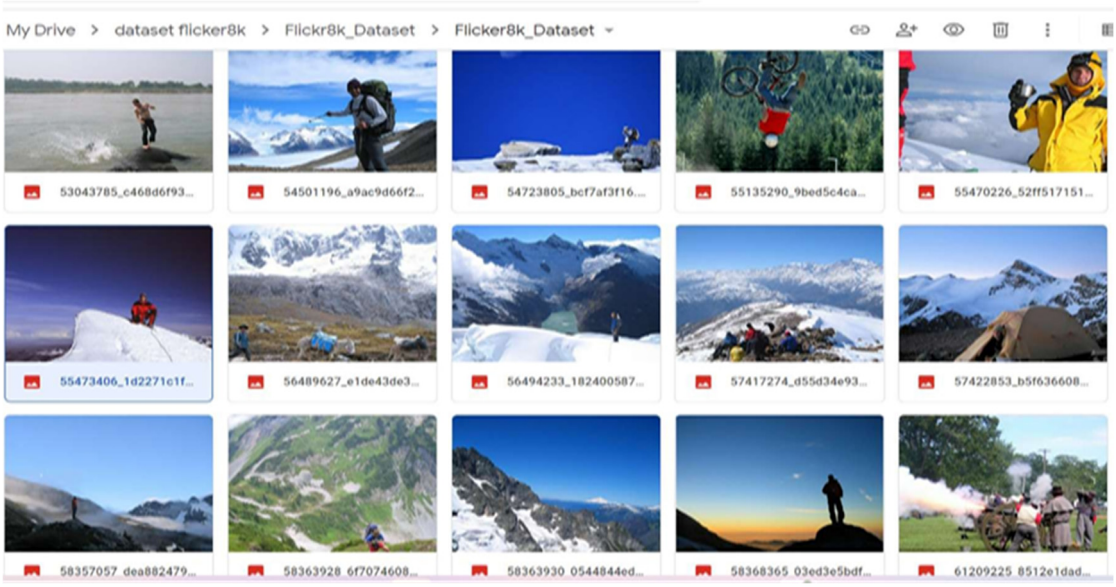

Figure 9. Datasets for YOLO V3.

```
1000268201_693b08cb0e.jpg#0     A child in a pink dress is climbing up a set of stairs in an entry way .
1000268201_693b08cb0e.jpg#1     A girl going into a wooden building .
1000268201_693b08cb0e.jpg#2     A little girl climbing into a wooden playhouse .
1000268201_693b08cb0e.jpg#3     A little girl climbing the stairs to her playhouse .
1000268201_693b08cb0e.jpg#4     A little girl in a pink dress going into a wooden cabin .
1001773457_577c3a7d70.jpg#0     A black dog and a spotted dog are fighting
1001773457_577c3a7d70.jpg#1     A black dog and a tri-colored dog playing with each other on the road .
1001773457_577c3a7d70.jpg#2     A black dog and a white dog with brown spots are staring at each other in the street .
1001773457_577c3a7d70.jpg#3     Two dogs of different breeds looking at each other on the road .
1001773457_577c3a7d70.jpg#4     Two dogs on pavement moving toward each other .
1002674143_1b742ab4b8.jpg#0     A little girl covered in paint sits in front of a painted rainbow with her hands in a bowl .
1002674143_1b742ab4b8.jpg#1     A little girl is sitting in front of a large painted rainbow .
1002674143_1b742ab4b8.jpg#2     A small girl in the grass plays with fingerpaints in front of a white canvas with a rainbow on it .
1002674143_1b742ab4b8.jpg#3     There is a girl with pigtails sitting in front of a rainbow painting .
1002674143_1b742ab4b8.jpg#4     Young girl with pigtails painting outside in the grass .
1003163366_44323f5815.jpg#0     A man lays on a bench while his dog sits by him .
1003163366_44323f5815.jpg#1     A man lays on the bench to which a white dog is also tied .
1003163366_44323f5815.jpg#2     a man sleeping on a bench outside with a white and black dog sitting next to him .
1003163366_44323f5815.jpg#3     A shirtless man lies on a park bench with his dog .
1003163366_44323f5815.jpg#4     man laying on bench holding leash of dog sitting on ground
1007129816_e794419615.jpg#0     A man in an orange hat starring at something .
1007129816_e794419615.jpg#1     A man wears an orange hat and glasses .
1007129816_e794419615.jpg#2     A man with gauges and glasses is wearing a Blitz hat .
1007129816_e794419615.jpg#3     A man with glasses is wearing a beer can crocheted hat .
1007129816_e794419615.jpg#4     The man with pierced ears is wearing glasses and an orange hat .
1007320043_627395c3d8.jpg#0     A child playing on a rope net .
1007320043_627395c3d8.jpg#1     A little girl climbing on red roping .
1007320043_627395c3d8.jpg#2     A little girl in pink climbs a rope bridge at the park .
1007320043_627395c3d8.jpg#3     A small child grips onto the red ropes at the playground .
1007320043_627395c3d8.jpg#4     The small child climbs on a red ropes on a playground .
1009434119_febe49276a.jpg#0     A black and white dog is running in a grassy garden surrounded by a white fence .
1009434119_febe49276a.jpg#1     A black and white dog is running through the grass .
1009434119_febe49276a.jpg#2     A Boston terrier is running in the grass .
1009434119_febe49276a.jpg#3     A Boston Terrier is running on lush green grass in front of a white fence .
1009434119_febe49276a.jpg#4     A dog runs on the green grass near a wooden fence .
1012212859_01547e3f17.jpg#0     A dog shakes its head near the shore , a red ball next to it .
1012212859_01547e3f17.jpg#1     A white dog shakes on the edge of a beach with an orange ball .
1012212859_01547e3f17.jpg#2     Dog with orange ball at feet , stands on shore shaking off water
1012212859_01547e3f17.jpg#3     White dog playing with a red ball on the shore near the water .
1012212859_01547e3f17.jpg#4     White dog with brown ears standing near water with head turned to one side .
1015118661_980735411b.jpg#0     A boy smiles in front of a stony wall in a city .
1015118661_980735411b.jpg#1     A little boy is standing on the street while a man in overalls is working on a stone wall .
```

Figure 10. Datasets for YOLO V3.

5. Results and Discussion

Resultant Image of YOLO V3 Model for Object Detection in a running video are shown in Figures 11–13.

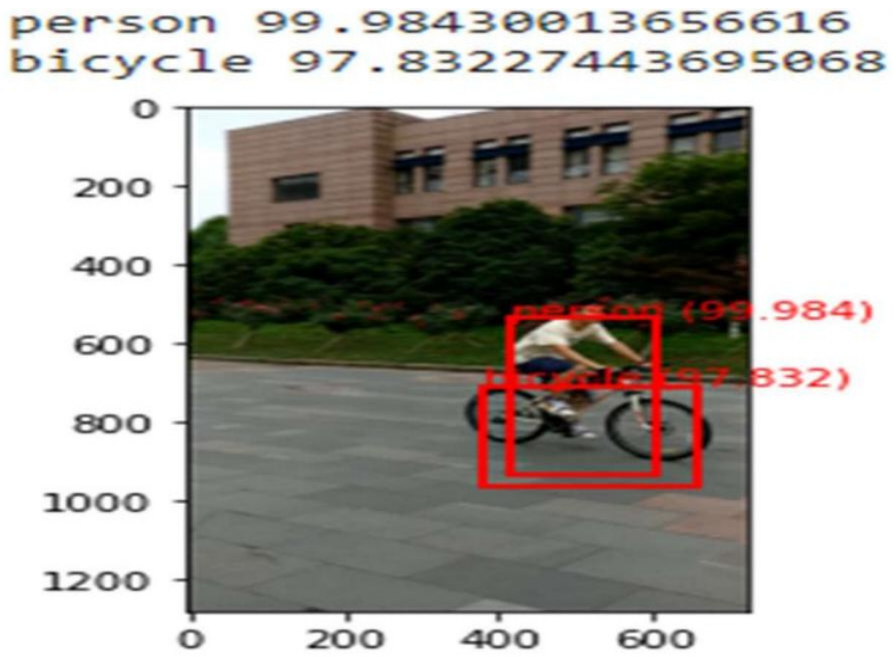

Figure 11. Accuracy of YOLO-V3 Model.

Results of Image Captioning Using ResNet-101 Model

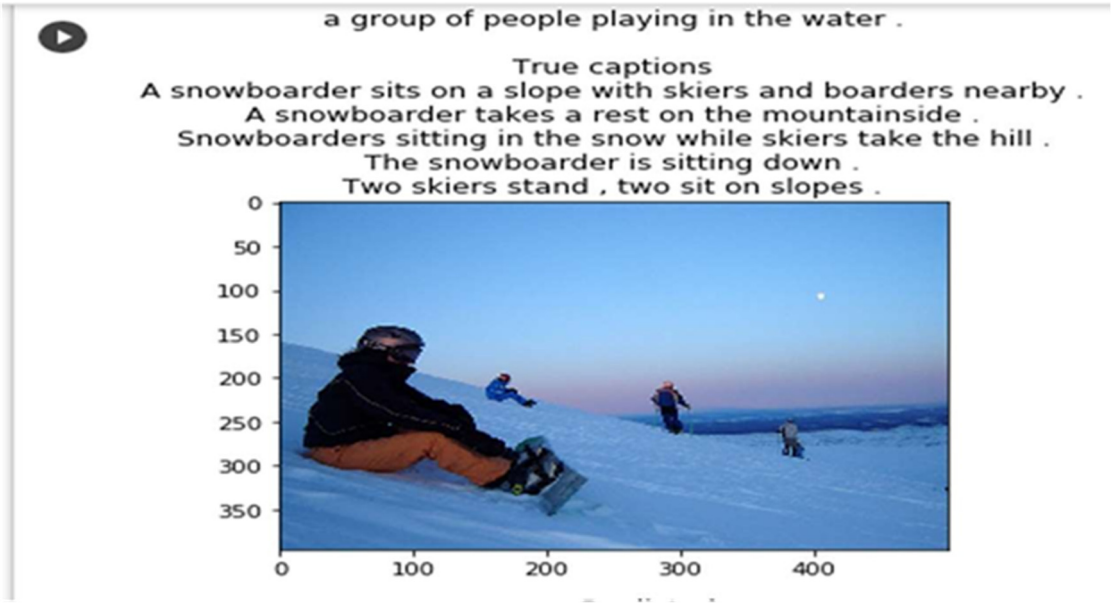

Figure 12. Results of Image Captioning.

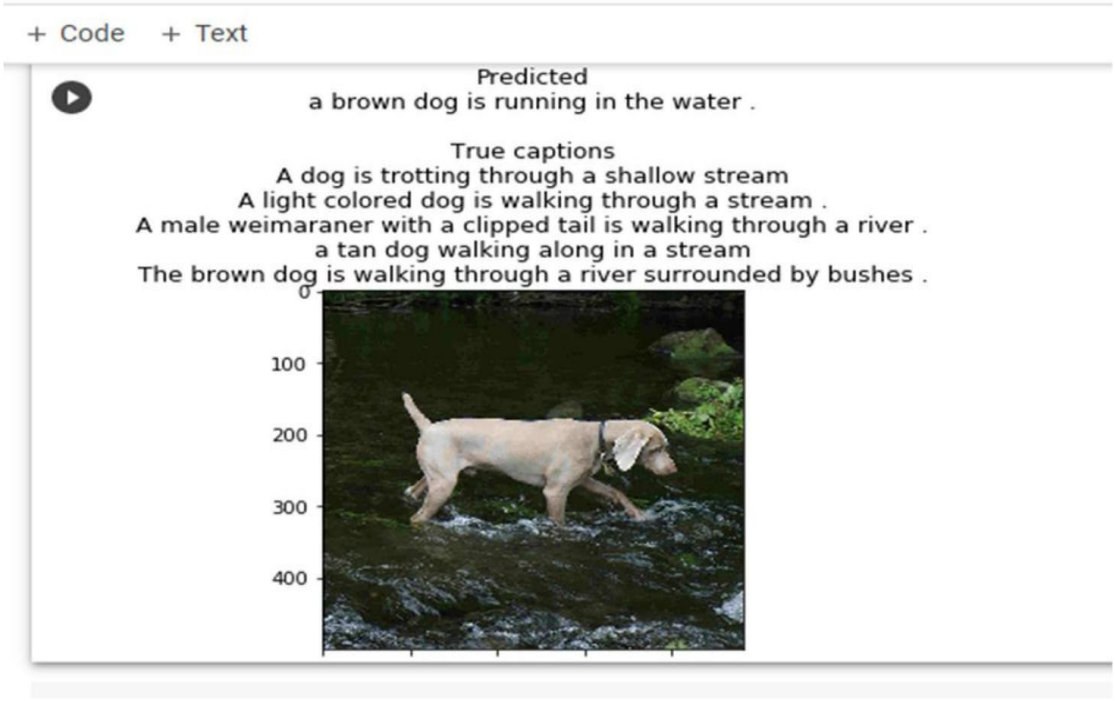

Figure 13. Results **o**f Image Captioning.

Accuracy of Image Captioning

In Figure 14 the graph shows Loss History is having Safron line that shows validation loss and blue line shows Training loss.

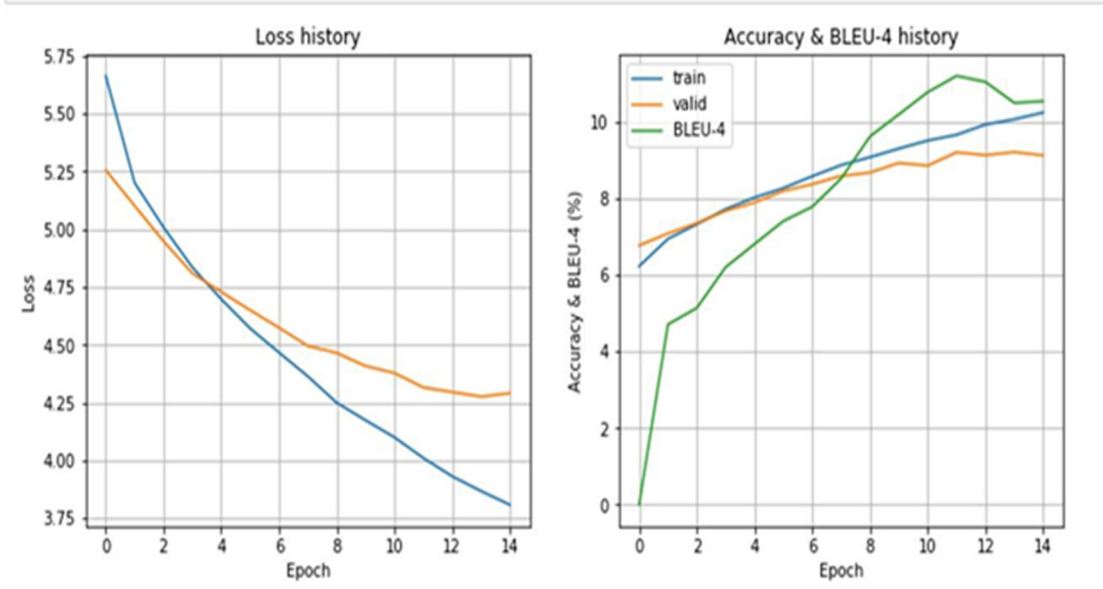

Figure 14. BLEU-4 Accuracy metrics shows the results of Testing.

This graph represents a comparison between Training loss and Validation loss.

6. Conclusions

To identify and localize objects, there exist many methods with a trade-off in speed performance and accuracy of result. But yet we can't say any single algorithm is best over others. One can always select the method that suits the requirement at best. In a short period, object detection applications got much popularity and still a lot to cover in this area because of its vast scope of research [9–11]. This project

presents the comparison of various algorithms to identify and localize objects based on accuracy, time, and parameter values with varying sizes of the input image. We have identified a new methodology of single stage model for improving speed without sacrificing much accuracy. The comparison results show that YOLO v3-Tiny increases the speed of object detection while ensures the accuracy of the result. We can also extend object localization and recognition from static pictures to a video containing the dynamic sequence of images. In this project, we have known about many methods of generating image captions based on Traditional techniques on machine learning and based on Deep machine learning techniques. With the traditional methods, it is not feasible to extract features of large-datasets but with deep machine learning approach it is possible to work with diverse and large datasets. These methods, using the datasets available, have established a standard for future researchers and have given a powerful image annotation method capable of generating high-quality descriptions of query images. Attention mechanism is also playing a vital role in storytelling approach. It is an effective improvement to the image description approaches.

In this paper, we have known about many methods of generating image captions based on Traditional techniques on machine learning and based on Deep machine learning techniques. With the traditional methods, it is not feasible to extract features of large-datasets but with deep machine learning approach it is possible to work with diverse and large datasets [32]. These methods, using the datasets explained above, have established a standard for future researchers and have given a powerful image annotation method capable of generating high-quality descriptions of query images. The CaMEL model (a novel transformer model) outperforms conventional approaches in terms of quality of captions and less requirement of parameters. Nowadays, it is very easy to figure out that the use of Transformers with encoder- decoder mechanism has left basic model structures far behind in terms of better outputs. Attention mechanism is also playing a vital role in storytelling approach. It is an effective improvement to the conventional approaches.

7. Future Enhancement

The future of object detection technology is in the process of proving itself, and much like the original Industrial Revolution, it has the potential to free people from menial jobs that can be done more efficiently and effectively by machines.

Stupendous work has also been done on this research topic. If we talk about future perspective, we can say that we can use a large dataset for training and testing purposes in order to getting more accurate results and caption quality. In the future, more resource-rich network designs, more complex object properties, more functional geometric relationships, and positional encoding strategies still generate enormous research expectations for further advancing the image captioning domain. We can train our model for real time webcam-based project that would be more useful in the coming future.

Author Contributions
Conceptualization, S.U. and R.K.; methodology, S.U.; software, R.K.; validation, S.U., R.K. and M.; writing—original draft preparation, S.U.; writing—review and editing, K.K.G. and M. All authors have read and agreed to the published version of the manuscript.

Funding
This research received no external funding.

Conflict of Interest Statement
The authors declare no conflicts of interest.

Data Availability Statement
Dataset is used in this paper is Flickr 8K. **Source:** https://www.kaggle.com/datasets/ming666/flicker8k-dataset.

References

1. A. Krizhevsky, I. Sutskever and G.E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *NIPS*, 2012, doi: 10.1201/9781420010749.
2. J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," 30th IEEE Conf. Comput. Vis. Pattern Recognition, *CVPR*, vol. 2017-Janua, pp. 6517–6525, 2017, doi: 10.1109/CVPR.2017.690.
3. J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," arXiv Prepr., 2018.
4. N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *IEEE CVPR*, vol. 1, pp. 886–893, 2005, doi: 10.1109/CVPR.2005.177.
5. C. Szegedy, W. Liu, Y. Jia, and P. Sermanet, "Going Deeper with Convolutions," *CVPR*, 2015, doi: 10.1108/978-1-78973-723- 320191012.

6. T. Sato, H. Kishi, S. Murakata, Y. Hayashi, T. Hattori, S. Nakazawa, Y. Mori, M. Hidaka, Kasahara, Y. A. Kusuhara, K. Hosoya, H. Hayashi, A. Okamoto. A new deep-learning model using YOLOv3 to support sperm selection during intracytoplasmic sperm injection procedure. *Reprod. Med. Biol.* 2022 Apr 4;21(1):e12454. doi: 10.1002/rmb2.12454. PMID: 35414764; PMCID: PMC8979154.
7. Zhao, Liquan & Li, Shuaiyang. (2020). Object Detection Algorithm Based on Improved YOLOv3. *Electronics* 9. 537. 10.3390/electronics9030537.
8. Van Houdt, Greg & Mosquera, Carlos & Nápoles, Gonzalo. (2020). A Review on the Long Short-Term Memory Model. *Artificial Intelligence Review* 53. 10.1007/s10462-020-09838-1.
9. J. R. R. Uijlings, K. E. A. Van De Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," Int. J. Comput. Vis., vol. 104, no. 2, pp. 154– 171, 2013, doi: 10.1007/s11263-013-0620- 5.
10. Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, "Object Detection with Deep Learning: A Review," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, 2019, doi: 10.1109/TNNLS.2018.2876865K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional.
11. K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *ECCV*, pp. 346–361, 2014, doi: 10.1023/B: KICA.0000038074.96200.69.
12. R. Nabati and H. Qi, "RRPN: RADAR REGION PROPOSAL NETWORK FOR OBJECT DETECTION IN AUTONOMOUS VEHICLES," *IEEE Int. Conf. Image ----Process.*, pp. 3093–3097, 2019.
13. D. Wang, C. Li, S. Wen, X. Chang, S. Nepal, and Y. Xiang, "Daedalus: Breaking Non- Maximum Suppression in Object Detection via Adversarial Examples," *arXiv* arXiv Prepr., 2019.
14. Q. C. Mao, H. M. Sun, Y. B. Liu, and R. S. Jia, "Mini-YOLOv3: RealTime Object Detector for Embedded Applications," *IEEE Access*, vol. 7, pp. 133529–133538, 2019, doi: 10.1109/ACCESS.2019.2941547.
15. Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. ACM, 144–152.
16. Kiros et al. (2014a). Multimodal Neural Language Models. Proceedings of the 31st International Conference on Machine Learning, PMLR 32(2):595–603.
17. Kiros, R., Salakhutdinov, R. & Zemel, R.(2014). Multimodal Neural Language Models. In Proceedings *of the 31st International Conference on Machine Learning, Proceedings of Machine Learning Research* 32(2):595-603 Available from https://proceedings.mlr.press/v32/kiros14.html.
18. A Comprehensive Survey of Deep Learning for Image Captioning MD. ZAKIR HOSSAIN, Murdoch University, Australia FERDOUS SOHEL, Murdoch University, ustralia MOHD FAIRUZ SHIRATUDDIN, Murdoch University, Australia HAMID LAGA, Murdoch University, Australia.
19. Qingzhong Wang and Antoni B Chan. 2018. CNN+ CNN: Convolutional Decoders for Image Captioning. *arXiv* arXiv:1805.0901.
20. Geometry Attention Transformer with Position-aware LSTMs for Image Captioning2 Chi Wang1, Yulin Shen1, Luping Ji∗ School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, P.R. China (2021).
21. Timo Ojala, Matti PietikAd'inen, and Topi MAd'enpAd'Ad'. 2000. Gray scale and rotation invariant texture classification with local binary patterns. In European Conference on Computer Vision. Springer, 404–420.
22. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9,2015.
23. Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition, 2005. In *Proceedings of the IEEE Computer Society Conference on* CVPR 2005, Vol. 1. IEEE, 886–893.
24. W. Liu, D. Anguelov, D. Erhan, and C. Szegedy, "SSD: Single Shot MultiBox Detector," *ECCV*, vol. 1, pp. 21–37, 2016, doi: 10.1007/978- 3-319-46448-0.
25. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real- time object detection," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 779–788, 2016, doi: 10.1109/CVPR.2016.91.
26. J. Huang et al., "Speed/accuracy trade-offs for modern convolutional object detectors," In *Proceedings of the 30th IEEE Conf. Comput. Vis. Pattern Recognition*, CVPR, vol. 2017-Janua, pp. 3296– 3305, 2017, doi: 10.1109/CVPR.2017.351.
27. CaMEL: Mean Teacher Learning for Image Captioning Manuele Barraco, Matteo Stefanini, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, Rita Cucchiara University of Modena and Reggio Emilia (2022).
28. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv* arXiv:1512.03385,2015.z.
29. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105,2012.
30. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv* arXiv:1409.1556,2014.
31. Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, and Tat-Seng Chua. 2017. SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6298– 6306.
32. Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. 2018. Convolutional image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5561–5570.

Author Biographies

Soumya Upadhyay currently holds the position of Assistant Professor at COER University in Roorkee, India. She earned her Engineering Degree in Computer Science and Engineering from Rajiv Gandhi Proudyogiki Vishwavidyalaya in Bhopal. Additionally, she holds an MBA in Human Resource and Marketing Management from Indore University, and an M.Tech. Degree in Computer Science and Engineering from the National Institute of Technology in Arunachal Pradesh. Currently, she is pursuing her Ph.D. in Computer Science and Engineering from the Indian Institute of Technology in Mandi. With a combined experience of six years in industry and academia, her research focuses on Artificial Intelligence and Computer Vision.

Kamal Kumar Gola is an Assistant Professor at COER University, Roorkee, India. He completed his B.Tech. Degree in Computer Science and Engineering from Moradabad Institute of Technology and his M.Tech. Degree in Computer Science and Engineering from Uttarakhand Technical University. He is pursuing a Ph.D. in Computer Science and Engineering from Indian Institute of Technology, Jodhpur. With over 14 years of university teaching experience and six months of industrial experience, Kamal Kumar Gola is a seasoned educator with a wealth of knowledge in his field. He has published more than fifty papers in international journals and has participated in various international and national conferences where he has presented his research. He kept himself updated by attending various professional development programs, workshops, and training courses organized by esteemed institutions such as IIT, NIT, IIIT, and others. His main research interests lie in Underwater Acoustic Sensor Networks, Algorithms and Security.

Ravindra Kothiyal serves as a Cyber Security Analyst within the Ministry of Communication in New Delhi, India. He earned his B.Tech. in Computer Science and Engineering from Noida Institute of Engineering and Technology, followed by an M.Tech. in Computer Science and Engineering from the National Institute of Technology in Arunachal Pradesh. His expertise lies in the realms of Cyber Security and Artificial Intelligence, and he has gained extensive experience through collaborations with various organizations specializing in these fields.

Dr. Mridula, holding B Tech, M Tech, and PhD degrees from IIT Roorkee, specializes in Pattern Recognition and Image Processing. Awarded the "Best Teacher of the Year" by Uttarakhand's Chief Minister, she boasts over a decade of teaching experience, guiding 3 PhD scholars. With a significant research portfolio, she has authored numerous papers in international journals and conferences, including SCI indexed papers, and contributed to a book chapter. Dr. Mridula has secured research grants and holds patents. Known for mentoring students, she currently serves as Professor at Haridwar University's Computer Science and Engineering Department.