IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.11726

IMAGE OBJECT DETECTION andVIDEO CAPTION GENERATION

Akshatha Ravi[1], Mohan Kumar H P[2]

Research Scholar, Department of Master of Computer Applications, P.E.S College of Engineering, Mandya, India[1]

Professor, Department of Master of Computer Applications, P.E.S College of Engineering, Mandya, India[2]

Abstract: Deep Learning methodologies offer great potential for applications that automatically attempt to generate captions or descriptions for images and video frames. With the recent advancements in neural networks, there has been progress in implementing object detection or generating description for images and captions for videos. Our work aims to automatically generate list of objects in the image when image is given and generate caption to video when video is given by reading their content. At present images and videos are annotated with Human intervention and it is difficult or almost an impossible task for a large commercial database to manually caption every photo and video. Image Object Detection and Video Captioning is basically very much useful in many applications like for generating captions or description during real-time and theyare also being used in advance machine and deep learning applications.

Keywords: Deep learning, Object detection, Neural networks, Video captioning.

I. INTRODUCTION

Among all living things on earth, the human brain is the most complex. It can recognize what we see and assess the scenario to determine what's happening. Therefore, it is simple for human beings to glance at an image and describe it using the right words, but the most challenging issue is getting a machine to think like a human. Object recognition, classification, image captioning, video captioning, and many more advanced forthcoming applications that allow for a lot more flexibility and have all benefited from the advancement of image processing, which has played and will continue to play a significant part in these areas. Our work primarily attempts to automatically generate list of objects present in the image and generate captions for videos. This work is widely used in a variety of applications, including captioning live news video and building software for the blind. They are also employed in advanced machine learning applications for human-robot interaction, self- driving cars, video surveillance, and software for self- guiding people with vision impairments. The Inceptionmodel, the VGG Model, and the conventional CNN-RNN Model are some examples of the several models used in deep learning. In our work for image object detection, we have used the SSD (Single Shot Detection) approach and MobileNet V3 as a backbone model which is pre-trained with the COCO dataset, while for video captioning, we have used VGG-16 pre-trained with ImageNet dataset and LSTM to generate captions.

II. RELATED WORK

The technique put out by Upulie H.D.I et al. [1] uses the preset model YOLOv2 for object identification where it includes bounding box limitations based on spatial parameters. Because each cell can only anticipate two boxes and one class, these spatial restrictions are maintained. There are fewer predictable items that are close to one another in clusters as a result. Rachita Byahatti et al. [2] have suggested a technique for categorizing and detecting objects that makes use of a predetermined model called YOLOv3. The classes used in their work were trained using the COCO dataset.

Fast Region-based Convolutional Network approach (Fast R-CNN) is utilized in Ross Girshick's method for object detection [3], which produces results more quickly but with lower accuracy.

In the paper published by Wei Liu et al [4], technique uses the SSD model (Single Shot Detector),which is the quickest and most accurate object detection model for tiny objects.

For video captioning, Krizhevsky A et al. [5] adopted the CNN-CNN approach, where CNN is usedfor both encoding and decoding purposes. Because the captions created here won't be correct andare unrelated to the test image, the CNN-CNN model has a large loss, which is unacceptable.

They have utilized CNN-RNN in the approach suggested by A. Hani et al [6], Contrary to CNN-CNN based models, this model has a lower loss but requires a longer training period.

A Multistream Hierarchical Boundary network was utilized for video captioning in the technique suggested by Nguyen et al [7]. The Multistream Hierarchical Boundary (MHB) model, which uses intrinsic feature boundary cuts to construct clips utilizing a video's fixed hierarchical recurrent architecture, is introduced in this study. To depict a video, border

encoding is employed. Films of different durations receive Parametric Guassian attention.
In the method proposed by Bin Zhao et al [8], technique makes use of a co-attention model (CAM) based recurrent neural network (RNN), where the CAM is used to encode the visual and text data and the RNN serves as the decoder to produce the video caption.
To extract characteristics from picture, audio, and semantic information, Chien Yao Wang et al [9] proposes technique using several neural networks. Before being input into an LSTM for initialization, image and audio information are combined. The integrated audio-image elements aid in the formation of a more effective semantic network over the full semantics.
In the method proposed by Soichiro Oura et al [10], technique makes use of MDNNISF (Multimodal Deep Neural Network with Image Sequence Features) to provide a sentence-length description of a given video clip. By combining S2VT with NeuralTalk2, which is used for captioning images and is known to produce accurate descriptions due to its capacity to learn alignments between text fragments and picture fragments, they have attempted to get around the problem.

III. METHODOLOGY

A. Image object detection

In our work, we have used a deep learning-based approach to solve the problem. The goal of our work is to automatically create captions for videos and description for images.
For image object detection, we have used, MobileNet V3 as a backbone model for the SSD (Single Shot Detector) model for object recognition in the images which is pre-trained using the COCO dataset. The COCO dataset comprises 3,28,000 total images of 80 different items like car, person, dog, boat etc. For the most effective application of detection and tracking, we have used the combination of SSD and MobileNet V3.
Single Shot Detector (SSD) models consist of two parts: the head and the backbone. In our work, MobileNet V3 serves as the backbone model for the SSD. The bounding boxes and classes of objects in the spatial location of the final layer's activations are translated from this backbone model's outputs, which are attached to an SSD head. This model has already been trained using the MS COCO dataset. While training, an image is provided as input using MobileNet V3, object characteristics are mapped, weighted, and the matching label is placed into the list. We employ the SSD (Single Shot Detector) model for object detection, which is built on top of this trained model. Let's now talk about some of the important SSD parameters.

1) Grid cell:

Instead of employing sliding windows, which are used in image classification models, SSD splits the picture into grid cells, with each grid cell being in charge of object detection in that area of the image. Therefore, location and background class are disregarded when no objects are found or identified in that area. Predicting the kind and placement of an object inside that area is all that is involved in object detection.

2) Anchor box:

We employ an anchor box when there are two or more different objects in a single grid cell to be identified. In SSD, multiple anchor boxes can be allocated to each grid cell. Each of these anchor boxes, which are pre-defined, is in charge of the size and form of a grid cell. In order to match the proper anchor box with the bounding boxes of each ground truth item within an image, SSD performs a matching phase during training. Essentially, the position and class of an item are predicted by the anchor box that has the greatest degree of overlap with it. After network training, this attribute is utilized to forecast the identified objects' positions. Each anchor box is specified by an aspect ratio and a zoom level.

3) Zoom level:

The size of the anchor box and the grid cell might not always match. In a grid cell, the zoom level is utilized to find small or large objects. Zoom parameter is used to specify how much the anchor boxes need to be zoom in or zoom out with respect to each grid cell.

4) Receptive field:

The fundamental principle of SSD architecture is the receptive field because it allows to detect objects at different scales and output a tighter bounding box. SSD further compared to other model i.e., MobileNet V3 by applying multiple convolutional layers to the backbone feature map and each of these convolutional layers provides object detection results. Due to the backbone model, i.e., MobileNet V3 layers with smaller receptive field can represent smaller size objects, predictions from these layers help when working with smaller size objects.

DOI: 10.17148/IJARCCE.2022.11726

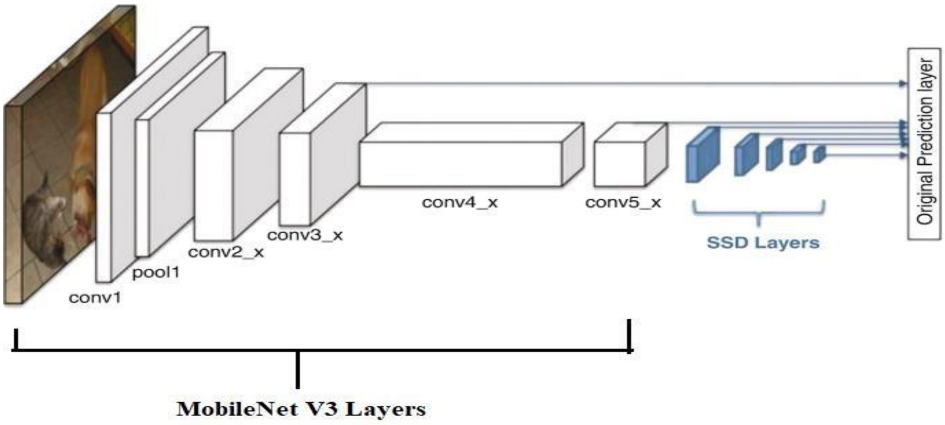

Fig. 1 Object detection architecture

B. Video captioning

In our work for video captioning, we have used a sequence learning end-to-end encoder-decoder CNN-RNN model, i.e., VGG-16 a CNN model which is pre-trained on ImageNet dataset and acts as an encoder where the feature extraction and training of images happens and we have used LSTM a RNN model as a decoder where it uses greedy search for generating captions. Firstly, the pre-trained model VGG-16. This model is trained using ImageNet dataset where an image dataset is organized according to the WordNet hierarchy. In WordNet each meaningful concept isdescribed by multiple words or word phrases which iscalled synset. ImageNet aims to provide on average 1000 images to illustrate each synset. Now let’s discuss about the working of VGG-16 model, video is a collection of many image that is arranged one after the other and each image is called the frames. When a video is given as input, 80 images are extracted and 4096 features are extracted from a single frame. So 80x4096 numpy array for each video is extracted using pretrained CNN model VGG-16. Input data is in sequence of features of frames. After analyzing and extracting the feature, to process the sequences of frames for each video sequence models are used. So, LSTM is used as the decoder to perform the task of estimating the probability of occurrence of consecutive video caption terms from the vocabulary. In this decoder component, we have implemented a greedy search algorithm for efficient expression probability estimation. The main goal of this approach is to calculate the term with the maximum possible conditional probability from a given list of known terms. The algorithm terminates when the sequence of terms selected in the search does not reach a limit. This satisfies the condition of the algorithm and helps to terminate, which leads to the output sequence and is assigned as a video caption for the given input coded features and vocabulary.

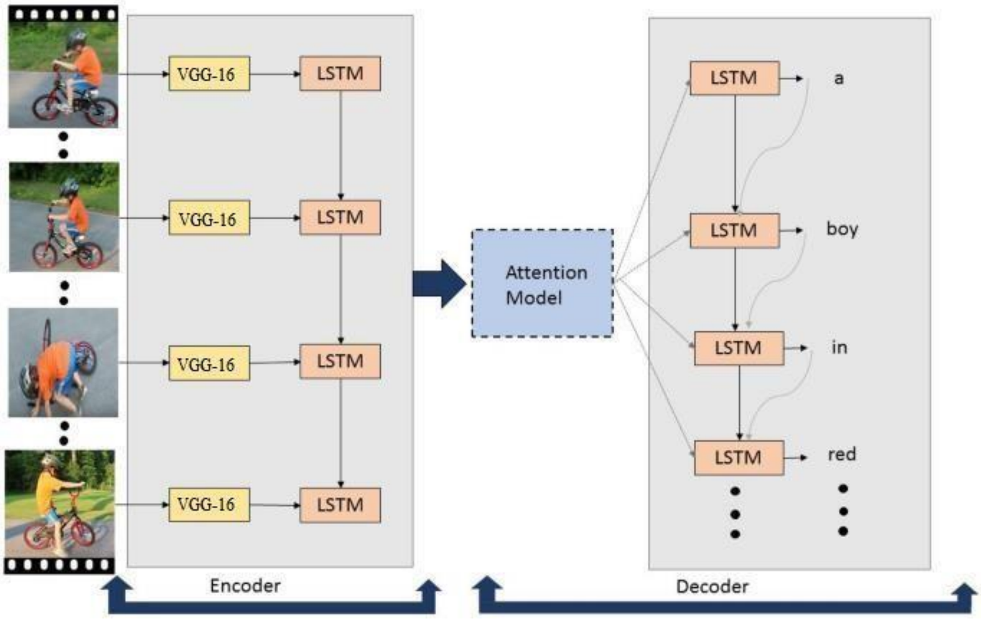

Fig. 2 Video captioning architecture

'

DOI: 10.17148/IJARCCE.2022.11726

IV. RESULT AND ANALYSIS

A. RESULT

In our work we have worked on both image object detection and video caption generation. I have used pre-trained models MobileNet V3 and VGG-16 which are trained using COCO and ImageNet dataset respectively.

For image object detection, we have tested the model using google photos. The test cases of 5 images from the testing data is given in the table 1.

For object detection it is not easy to find the accuracy like any other classification. It includes many steps. First step is to find the IoU (Intersection over Union) where, IoU = area of overlap/ area of union. For COCO dataset IoU = 0.5. If IoU > 0.5 then it is TP (True Positive i.e., predicted positive and are actually positive), if IoU < 0.5 then it is FP (False Positive i.e., predicted positive and are actually negative). Second step is to find the precision and recall where precision = TP/(TP+FP) and recall = TP/Total number of objects in that image. Then last step is plotting the precision vs. recall graph for the test cases given in the table and finding the average precision using 11 Point Interpolation, where Average Precision = 1/11*(interpolated precision at each point)

TABLE I RESULT FOR OBJECT DETECTION

Image No	Detection	TP	FP	Precision	Recall
Image 1	Motorbike	1	0	1	0.083
	Person	1	0	1	0.16
Image 2	Dog	1	0	1	0.25
	Person	1	0	1	0.33
Image 3	Person	1	0	1	0.41
	Person	1	0	1	0.5
	Boat	1	0	1	0.58
	Boat	1	0	1	0.66
	Person	0	1	0.72	0.66
Image 4	Bear	1	0	0.9	0.75
	Bear	1	0	0.9	0.83
	Cow	0	1	0.83	0.83
Image 5	Airplane	1	0	0.84	0.91
	Airplane	0	1	0.78	0.91

For video captioning, we have tested the model using random short 20 videos for 2 categories where the video has 2 objects and more than 2 objects of different category of objects. To calculate accuracy of the model we have used BLEU (Bilingual Evaluation Understudy)

If BLEU score is <10 it means interpretation is almost useless.

If BLEU score is 10-19 it means interpretation is, hard to get the gist.

If BLEU score is 20-29 it means interpretation is, the gist is clear, but has significant grammatical errors. If BLEU score is 30-39 it means interpretation is, understandable to good translation.

If BLEU score is 40-49 it means interpretation is, high quality translations.

If BLEU score is 50-59 it means interpretation is, very high-quality translations, adequate, and fluent translation. If BLEU score is >60 it means interpretation is, quality better than human.

TABLE II RESULT FOR VIDEO CAPTIONING

Type of test video used	BLEU score
Video which has 2 different objects	42
Video which has more than 2 objects	35

B. ANALYSIS

- In image object detection we have used SSD and MobileNet V3 models as SSD gets rid of the feature resampling stage and combines all of the computed results into one piece. For areas with limited computing capacity,

DOI: 10.17148/IJARCCE.2022.11726

such as mobile devices, MobileNet V3 is a compact and light weight compared to other network model that employs depth-wise separable convolution. So, processing time is less while not sacrificing accuracy or performance, these models work together to recognize objects quickly and effectively.

- MobileNet V3 is used in which accuracy is more as it detects even the smallest objects in the image and also SSD is used which gives better accuracy thanthe sliding window approach models. The combination of proposed models gives better accuracy that is 90% and gives the output fast.
- The dataset used in image object detection is COCO which has 80 classes so it is better than Pascal VOC which has only 20 classes in the dataset.
- In video captioning we have used VGG-16 and LSTM a CNN-RNN model which are better than CNN-CNN models as it experiences fewer losses and gives better accuracy.

V. CONCLUSION

Image object detection and video captioning are considered to be intellectually challenging problems in image science. Our work utilizes Deep Learning techniques to offer automatically generating list of objects detected in images and captions for videos. In our work we have used MobileNet V3 which is pre- trained and also light weight model, takes less memoryspace so it can be used in mobile phones also which uses COCO dataset and acts as a backbone model for SSD (Single Shot Detector) model for object detectionin the image and for video captioning we have used VGG-16 as an encoder and LSTM as decoder. Models used for our work are more accurate than the existing system. In future this model can be extended to build complete Image-Speech conversion by converting captions of images to speech. This is very much helpful for blind people. It can also be extended to build higher machine learning applications like robot interactions. Even though deep learning is advanced up to now, sometimes model cannot generate the exact captions because machines cannot think or make decisions as accurately as human do. So, in future with the advancement of hardware and deep learning models we hope to generate captions with higher accuracy.

REFERENCES

[1]. Upulie H.D.I, Lakshini Kuganandamurthy, “Real- Time Object Detection using YOLO: A review”, May2021Varsha Kesavan, Vaidehi Muley, Megha Kolhekar, “Deep Learning based Automatic Image Caption Generation”, 03 February 2020.

[2]. Rachitha Byahatti, Dr. S. V. Viraktamath, Madhuri Yavagal, “Object Detection and Classification usingYOLOv3”, February 2021.

[3]. Ross Girshick “Fast R-CNN”, In International Conference on Computer Vision (ICCV), 07-13 December 2015.

[4]. Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, andAlexander C. Berg “SSD: Single shot multibox detector”, In ECCV, 2016.

[5]. Krizhevsky, A., Sutskever, I. and Hinton, G.E. “ImageNet Classification with Deep Convolutional Neural Networks”, Communications of the ACM, 2017.

[6]. A. Hani, Nailba Tegougui M. Kherallah, “ImageCaption Generration using a Deep Architecture”, International Arab conference on Information Technology (ACIT), 1 December 2019.

[7]. Thang Nguyen, Shagan Sah, Raymond Ptucha, "Multistream Hierarchical Boundary Network forVideo Captioning", IEEE Western New YorkImage and Signal Processing Workshop (WNYISPW), 2017.

[8]. Bin Zhao, Xuelong Li, Xiaoqiang Lu, "CAM-RNN: Co-Attention Model Based RNN for Video Captioning",IEEE Transactions on Image Processing, vol. 28, no. 11, November 2019.

[9]. Chein-Yao Wang, Pei-Sin Liaw, Kai-Wen Liang, Jai-Ching Wang, Pao-Chi Chang. “Captioning Based on Image–Audio Deep Learning Techniques", IEEE 9th International Conference on Consumer Electronics (ICCE-Berlin), 2019.

[10]. Soichiro Oura, Tetsu Matsukawa, Einoshin Suzuki, "Multimodal Deep Neural Network with Image Sequence Features for Video Captioning”, International Joint Conference onNeural Networks(IJCNN), 2018.