Recent Research in Science and Technology 2023, 15: 1-7
doi: 10.25081/rrst.2023.15.8022
https://updatepublishing.com/journal/index.php/rrst

ISSN: 2076-5061

Perception of multimodal objects in NLP through computer vision

Sakib Hosen Himel[1]*, Mahidul Islam Rana[2]

Department of Computer Science and Engineering, Daffodil International University, Dhaka-1207, Bangladesh

ABSTRACT

This project is based on voice interaction and object detecting properties. It will allow the users to do voice interaction with the artificial intelligence and it will reply with the system voice. That is how users will use their voice to command as a trigger to find out the category of any object by showing it using the camera module. At first, the user will show an object with the help of a camera and ask for identifying it in the system. The object detection system then captures a frame from the camera and predicts through the structure to identify which class the object belongs to by extracting the feature from there. The process of this application is to search the database to match the structural data to find out the exact category of the object. When this system approximately matches with the information of a category then the application will suggest the category for the object by mentioning the category name through voice. This application can also give some basic information by asking for it. Our general-purpose approach can be effective in interpreting the structure and properties of objects in different networks through natural language processing.

Received: October 07, 2022
Revised: December 22, 2022
Accepted: December 26, 2022
Published: January 10, 2023

***Corresponding Author:**
Sakib Hosen Himel
E-mail: sakib15-1598@diu.edu.bd

KEYWORDS: MobileNet, SSD-V3, Object detection, NLP, Computer vision, COCO dataset

INTRODUCTION

Two of the key components of neural networks in artificial intelligence are the human language training of systems as well as the detection of objects. Artificial intelligence is the science or engineering that gives a computer device or a system, the power of intelligence like a human being. A person can use amazing power to communicate and see and these are two of the main characteristics of a human being. In the last few years, there have been many developments in artificial intelligence. If this development continues, as it is, the world of robotics will soon acquire human intelligence.

Different shapes of neurons are formed in the human body, but the main function is to send different information to the human body through the mind. At the same time, neurons are connected to each other and make decisions based on different events. Neurons are at the root of all the things that people can think of or whatever information they get from the outside environment. Everything a person sees or hears is stored in neurons, and based on this information stored at the same time, people can determine their activity through neurons. For example, when two people are talking, neurons store the data of listening to what a person is saying and it is up to the neuron to decide what to say in response (Azevedo *et al.*, 2009).

The neural network is a cornerstone of deep learning where algorithms are inspired by the structure of neurons in the human brain. This part of Artificial Intelligence creates a pattern in which the system is trained through different data at certain stages, and different decisions can be made based on them, for example, recognizing objects based on different features. Let us create a neural network where the system can distinguish between triangles, circles, and quadrilaterals in Figure 1. Here different layers are created that act as neurons in the human body i.e., will take different data and at the same time can make predicted decisions based on different events.

This model "Perception of Multimodal Objects in NLP through Computer Vision" will basically be able to identify different objects and in this case, the whole event will be conducted through human language. We have divided this whole thing into three parts, one of which is to identify objects and the other two is the part of natural language processing that acts as inputs and simultaneously provides outputs based on different inputs.

In this project, the SSD MobileNet-V3 library can create a system that can detect different objects. Simultaneously PyAudio, speech recognition, text to speech libraries can enable a python system to give voice input or output.

RELATED WORK

As children learn to mature, learning (or practice depends on someone's point of view) is an important part of the neural

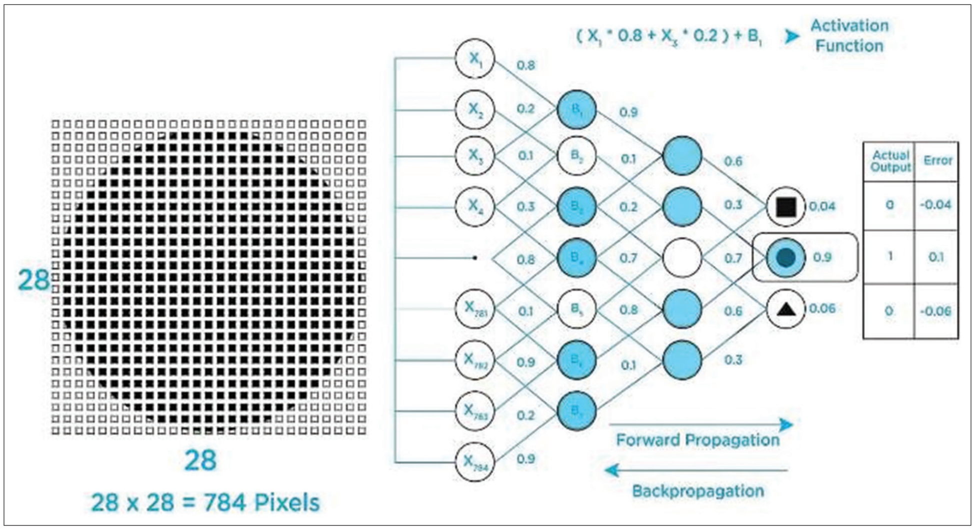

Figure 1: A system that can differentiate triangles, circles, and quadrilaterals

network. The significance of the learning method discussed here is to keep an eye on the learning that the desired results are given and the predictions of the neural network need to be as close as possible. This goal has been achieved in such a way that the weight values are so tuned that the action of error is reduced (Wang, 2003; Li *et al.*, 2015).

Neural models match or outperform the functionality of other sophisticated systems in different NLP functions. Yet the behaviors of the deep learning model can be easily explained by weighing individual human explanatory features (part-speech, nominal entity, word size, syntactic purse features, etc.) and classifying individuals with apt theoretical features. (Part-speech, nominal entity, word size, syntactic purse features, etc.). Deep learning methods focus on sound embedding (low-dimensional, continuous, real-value vectors) using multi-layered neural networks with each layer indicated as an array of hidden neuron units (Li *et al.*, 2015).

Item identification and perception are important from a computer standpoint. Recent advances in target identification include the popularity of the field proposal methodology and region-based conservative neural networks (R-CNN). A suggested framework for object recognition based on prior work employing the SIFT (Scale Invariant Features Transform) keypoint detector and FLNAN (Fast Library for Nearby Nearby), a quick stereo vision technique. Unfortunately, it is not based on the idea of extremely slow and deep learning, as it is used to know the position (estimate of the estimate) (Budiharto, 2014; Yeremia *et al.*, 2013).

Backpropagation networks are noted for their precision, and they help themselves learn and improve so that they may achieve better accuracy. To achieve both precision and training speed for recognizing alphabets, the backpropagation network approach was combined with a genetic algorithm. A genetic algorithm is used to discover the optimum initial settings for the network's architecture and synapses' weight so that the network may reach the best accuracy in the shortest amount of time (Budiharto *et al.*, 2018).

Scope of the Problem

The crops of those who make a living by farming come in contact with numerous insects in remote rural areas. In reality, farmers do not realize it or make any wrong decisions; it is a big loss in the future. In this case, such a device can overcome the issue of "perception of multimodal objects in NLP by computer vision". The system provides features such as object recognition and natural language processing, which allow farmers to speak in simple language about their problems with the system. At the same time, in different situations, one can know about the condition of the crop through this method. If you can provide the necessary programs in the system, this system will help the farmer to solve various problems.

Challenges

To address this dilemma, a major challenge is to find an accurate algorithm for object detection and simultaneously create datasets. Another key feature of our project is the creation of the voice assistant, here too it was a good challenge to know and create the action behind the various commands. The biggest challenge was capturing the camera frame through voice commands, identifying objects from the dataset through feature extraction, and then using voice in the system.

MATERIALS AND PROPOSED METHODOLOGY

The biggest challenges in achieving the objectives of this project can be divided into three parts. These are Object Detection, a voice assistant created by Natural Language Processing and Object Detection as a command of that voice assistant.

At this stage the complete workings of this model will be discussed in detail. Human language has been used to manage this voice assistant. This model extracts human voice input into text. It then matches the data previously trained in the system. Then classify the document and perform the action according to the category. In the case of this model, the output is given as voice. In this case, when the command is related to object detection, it captures a frame from the camera. The object detection architecture detects the object from that frame or still picture and returns it to the system in a detailed text format. In this system, text output is converted to voice output in both object detection and simple commands. Simple commands here refer to commands that do not require object detection (Figure 2).

Object Detection Procedure

In this case, whenever the system receives the command of object detection, the system captures a frame from the camera. This captured frame completes the whole process of object detection. This frame is considered a still picture and the object is first identified from this picture. Once the object is found, its features are extracted and matched from the dataset. Since COCO Dataset has been used in this project, the system features the full feature close to this Dataset. From there, when the features match, the object is declared, and the name of the object is sent to the system in string format. In 5 this case, if the object is not found, the whole process is repeated. This full process is also explained in the flowchart in the Figure 3.

Number of Images and Instance in COCO Dataset

COCO stands for Common Object in Context; this dataset was created under the supervision of excellent companies like CVDF (Common Visual Data Foundation), Microsoft, Facebook, and Mighty AI. This dataset has several features such as object segmentation, super pixel staff segmentation, 1.5 million object instances, 91 staff categories, context recognition, and 330K images where more than 200k images are labeled images. The COCO 2020 data set is a large-scale dataset of object recognition, segmentation, and captions, used in this model to classify 91 section objects (COCO, 2021). A significant property of these datasets is that they aim to locate non-iconic images in their natural sense. The sum of specific details contained in an image can be determined by analyzing the types of each item in the diagram and the examples in Figures 4 & 5. For ImageNet, the object recognition verification sets may be plotted since only a single object is labeled in the training results (Lin *et al.*, 2014).

Single Shot Multibook Detector Architecture

Here, it is discussed more SSDv3 for architecture (Figure 6). Both SSD data models have been used for dataset activity/object tracking. Coco datasets were taught by SSDv3 to test object-type subsets. The base deep learning network, which is CNN-based classifiers and applies convolution filters to eventually detect objects by collecting feature maps by SSD object detection (Table 1). This implementation uses MobileNet as the base network (also might be included- VGGNet, ResNet, DenseNet) (others might include- VGGNet, ResNet, DenseNet) (Graetz, 2018). SSD is designed for target tracking in real-time. Faster R-CNN uses an area proposal network to construct boundary boxes and utilizes certain boxes to distinguish objects. Although it is called the start-of-the-art of precision, the entire operation runs at 7 frames per second. SSD speeds up the process by eliminating the 7 need for the area proposal network. To restore the decrease in precision, SSD implements a few enhancements including multi-scale functionality and default boxes. These enhancements allow SSD to equal the Faster R-CNN's precision using lower resolution images, which further drives the speed higher. According to the above contrast, it maintains the real-time processing speed and exceeds the precision of the Quicker RCNN (Accuracy is calculated as the mean average precision map: the precision of the predictions) (Hui, 2018).

SSD uses VGG16 to extract function charts. It then detects artifacts using the Conv4 3 layer. For example, locally here if

Table 1: Extracting feature maps and adding convolution filters are the base component of SSD object detection

System	VOC2007 Test Map	FPS (Titan X)	Number of Boxes	Input Resolution
Faster R-CNN (CGG16)	73.2	7	~6000	~1000 x 600
YOLO (customized)	63.2	45	98	448 x 448
SSD300* (VGG16)	77.2	46	8732	300 x 300
SSD512* (VGG16)	79.8	19	24564	512 x 512

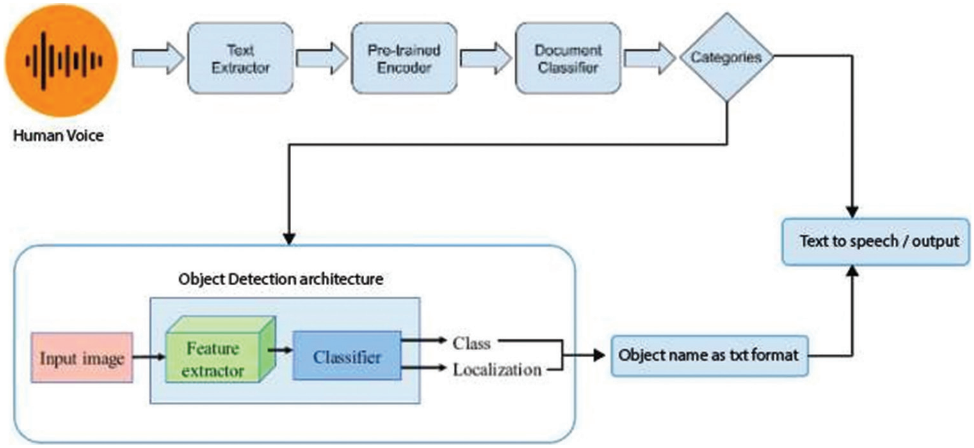

Figure 2: Working procedure flowchart

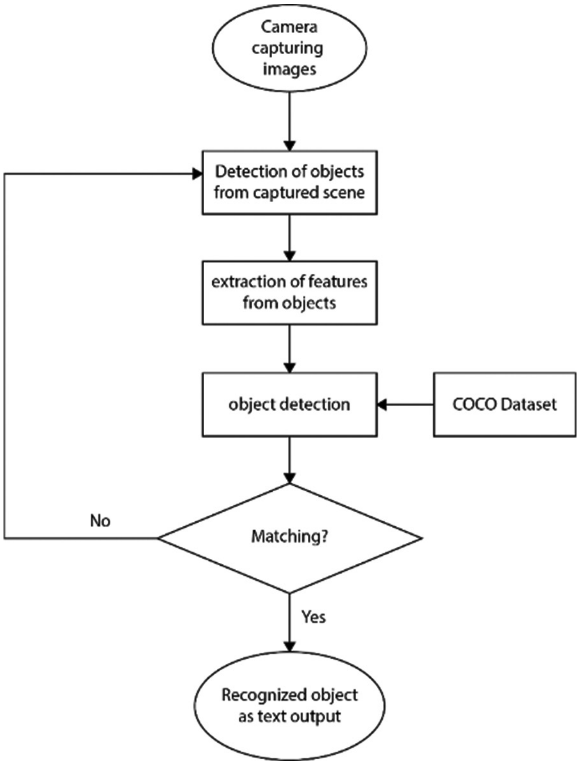
Camera capturing images
Detection of objects from captured scene
extraction of features from objects
object detection
COCO Dataset
Matching?
No
Yes
Recognized object as text output

Figure 3: Object detection procedure flowchart

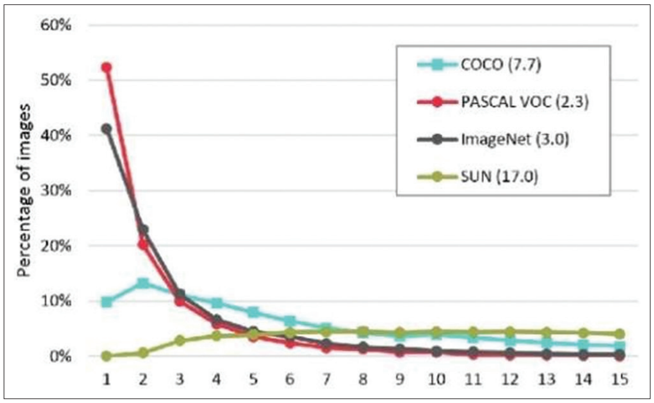

Figure 4: Number of instances in COCO dataset

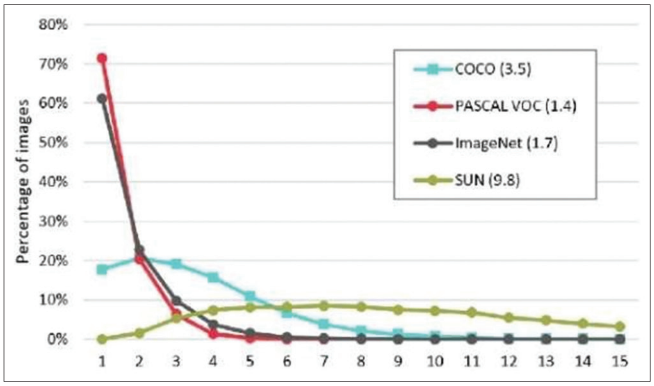

Figure 5: Number of categories in COCO dataset

Conv 4 3 is drawn from 8 × 8 (it should be 38 × 38). It predicts 4 entities for each cell (also called location).

A boundary box and 21 scores for each class (for no object there is one extra class) composed of each prediction, and the class for the bounded object is the selection by the highest score. A total of 38 × 38 × 4 predictions is made by Conv4 3: four projections per cell because of the scope of the function maps. Several forecasts include no object as predicted (Figure 7). SSD reserves a class "0" to mean it has no objects (Hui, 2018).

MobileNet SSD V3 Loss Function

To create such a single object identification network that produces 4 + n values, here we will quantify the loss function. Predicting the location of an object (n target class) is a classification problem. Trying to predict the four positions for both, the bounding box is indeed a regression task. A loss function will be needed that incorporates these two problems. Allow beginning with the bounding boxes: The localization loss may be the L1 loss of the projected bounding box coordinates y pred as well as the ground truth bounding moving targets x label:

$$l_{loc} = \sum_{i=1}^{4} \left| x_{pred,i} - x_{label,i} \right|$$

This is the sum of all the absolute discrepancies between the bounding box coordinates of the four predicted and four ground truths. The larger the gap between the expected and ground truth bounding boxes, the greater the loss. It is also be used as L2 loss, but at the cost of typical examples that have a much smaller error, the model would be more susceptible to outliers and will adapt to eliminate single outlier events. Then, adding the softmax activation function to these n values in order to transform them into probabilities. The cross-entropy loss function is then used to equate them to the mark. This is referred to as "trust loss." The combined loss attribute is simply the weighted sum of the loss of localization (bounding boxes) and loss of trust (classes):

$$L = L_{loc} + a.L_{conf} \, (2)$$

Networks can only learn boundary box predictions and neglect the classification function. As a result, the values of all the losses must be looked at and one of them has to be multiplied by a formula that makes them almost equal in size (Graetz, 2018).

SYSTEM ARCHITECTURE IN NLP

The most important thing for this model is natural language processing. In this system, the whole process is controlled through human language. Therefore, NLP's approach is very important for creating this kind of system. To create this type of system, voice input is first taken through the microphone and then the inputted voice is converted into a string. Then the information given in the string is matched with the dataset. In this case, if the data match, the system can perform the action according to

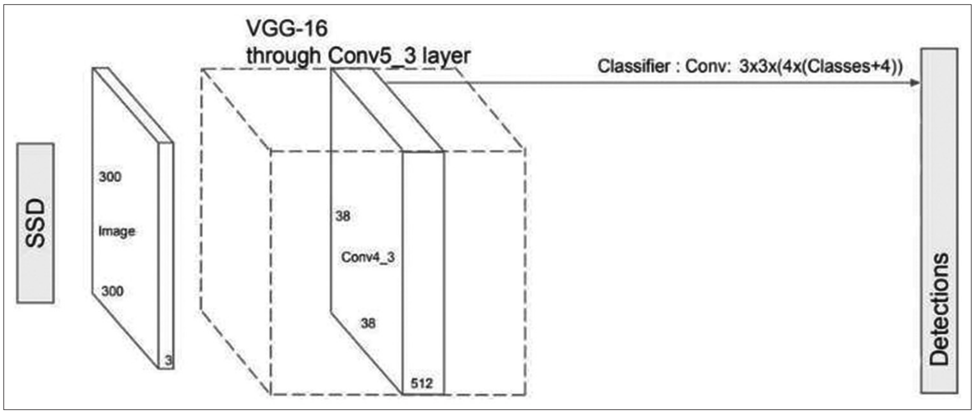

Figure 6: Mobile net SSD architecture

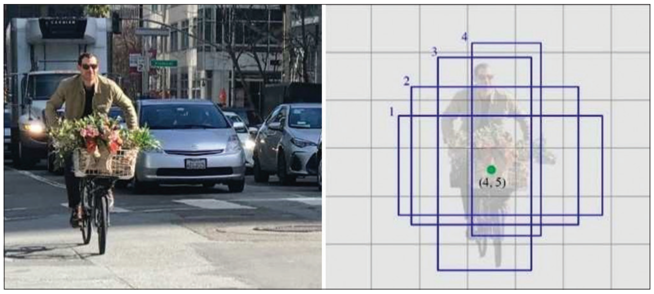

Figure 7: The original picture vs 4 predictions at each cell

the information. In the case of this model, string-based answers are given for action performance. The string is then converted to voice again and this voice is selected as the output. The complete procedure is shown in the Figure 8 as a flowchart.

EXPERIMENT RESULT AND DISCUSSION

This part will talk about the experiments and the results based on the test data to sort out and compare the accuracy of the model. In this model, Computer Vision has merged with NLP where NLP is represented as the project controller. This model runs smoothly when the microphone system and internet speed are efficient. It is able to respond more efficiently model gets a clear voice and correct pronunciation as input. For the sight of the model, this project uses version 3 of MobileNet SSD, which has an incredible volume of data to make an average accuracy of 85% for a variety of items. In this case, accuracy can work better if the camera control is tightly controlled. It is also possible to make this phenomenon more subtle through advanced cameras.

TEST RESULTS FOR OBJECT DETECTIONS

For visual extraction in this model, 91 categories of COCO Dataset objects are used with version 3 of MobileNet SSD. In this case for object identification, Mobile Net SSD V3 provides reasonable accuracy. There are some object identification precautions (Figure 9).

The most used methods of advanced deep learning models for performing a wide range of tasks on embedded systems are mobile networks and binary neural networks. A Single-Shot multibox detection (SSD) network configured to perform target detection is the mobilenet-ssd model. Using the Caffe* paradigm, this model is applied. The detection algorithm of this item has been tested up to 14 fps, so any low quality cameras will achieve decent results for any fps. Here SSD algorithms display indoor and outdoor video frames using webcams in experiments but can distinguish the positions of objects between two consecutive frames. The video and algorithm obtained by the webcam transform the size of a single frame that is known to be 300 x 300m. By creating a boundary box around the detected object, the SSD can detect the object frame by frame with the accuracy of a class name (Lin *et al*., 2014). The results obtained from this technology in the handmade video frame are shown in Table 2.

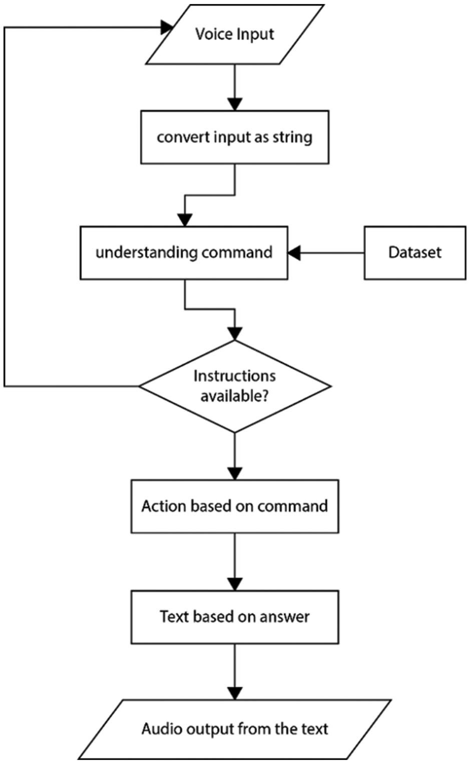

Figure 8: NLP working Procedure flowchart

A monitor in Figure 10 with a confidence level of 97.79 percent and a person with a 92.57 percent (i.e., probability) input frame of a video series identification, while the entire face of the person is not seen, CNN has a highly reliable detection algorithm for human characteristics. For various groups with different trust levels, the SSD will generate several bounding boxes as like in Figure 11 using a higher proportion of default boxes that will have a stronger effect, where different boxes are used for each position.

Table 2: Validation set accuracy for still image

Object Categories	Average accuracy	Mean average accuracy
Person	91.67%	
Chair	87.35%	87.2177%
Cat	79.5%	
Mouse	89.79%	
Monitor	94.83%	
Laptop	95.13%	
Keyboard	82.14%	
Suitcase	76.41%	
Umbrella	88.14%	

Patterns can be identified based on frame variations using this proposed single-shot multi-box detection method. In addition, the frame analyzes the effectiveness of the proposed method and is able to test the accuracy and consistency of the proposed method with the effects of foggy weather sensations.

It is very difficult to achieve a clear contrast between different object detectors. This model is best is not an easy one to clear. We like to combine precision and speed with real-life applications. The main purpose here is to illustrate the computer vision object recognition process as an operation of the processing of natural language. For this whole concept, this model is a successful and beautiful enough process.

In addition, the qfgaohao model with pre-trained weights from the merged VOC2007 and VOC2012 datasets was used in both experiments for both forms of networks. To prevent data development, the original code was changed to the same data and the base networks were retrained only with pre-trained

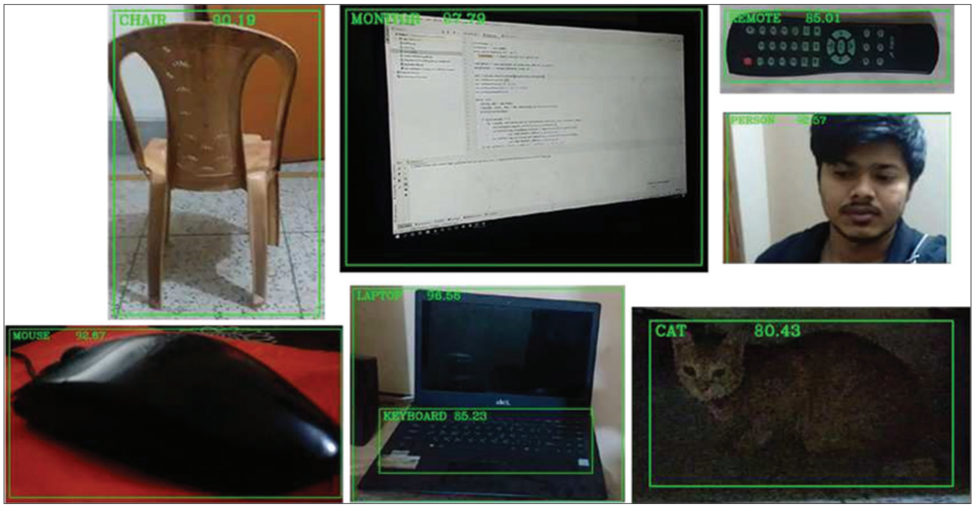

Figure 9: Object Identification via SSD V3

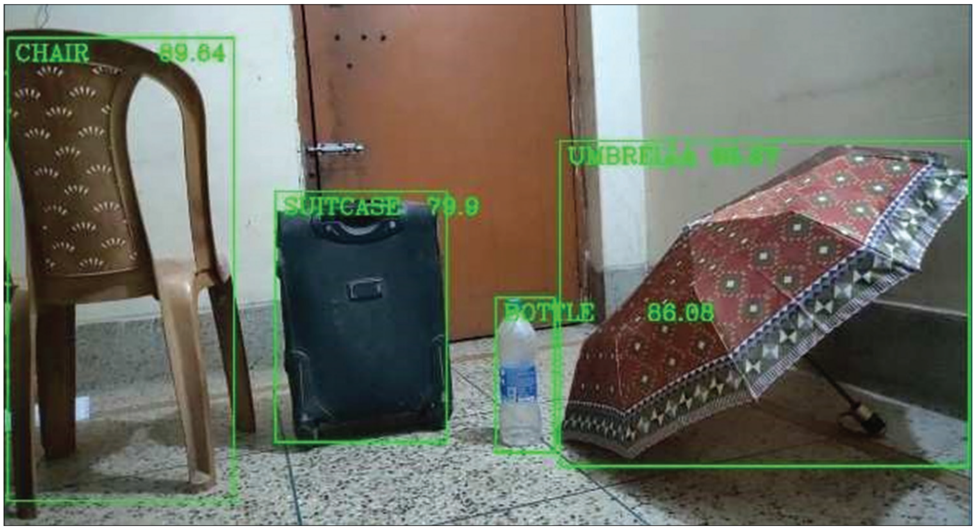

Figure 10: SSD multiple bounding boxes with varying trust levels for different groups

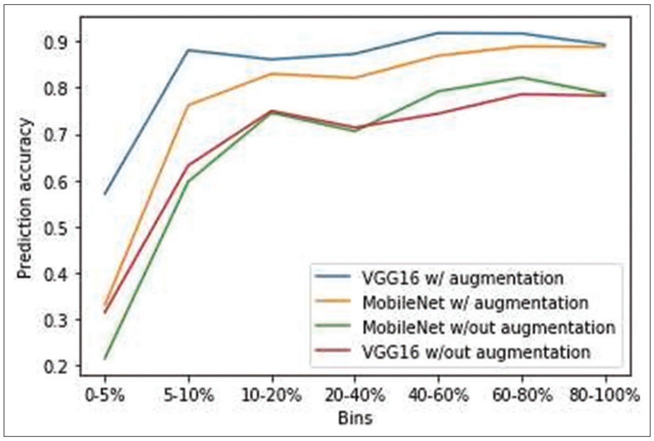

Figure 11: Experiment results via SSD V3

weights (e.g., VGG16 and MobileNet). It's also worth noting that after training networks for 100 epochs in Google Collab with K-80 GPU, the models that are conditioned (without data augmentation) use weights that decrease the optimum validity during training (Medium, 2021).

CONCLUSION

In this study, the accuracy of object detection and natural language processing of neural models was observed. To help us understand how neural models work, the object identification model, and helps to explain that the model shows remarkable features. Various methods have been adopted here to create some aspects of performance in these tasks. This model can provide insights into the behavior of object identification models in language-based work; also, this model identifies the initial step in understanding how to artificially perceive human language in natural language processing. At the same time in this research paper, the step of understanding the computer's vision is also responded as it evaluates the continuous method for object detection from complex visuals, mobile net SSD dependent image formatting. Adding further development to this project in the future will further revolutionize the neural networking system of Artificial Intelligence. However, it can be said that a satisfactory result has been obtained from this model.

Performing action through different languages can be a unique way to develop this model. Understanding human expression, he is able to create an advanced neural model of action performance. This model will also be able to give output that is more extraordinary if a better system for object detection is created. Once a large dataset is created, this model will be able to create systems for performing various activities.

ACKNOWLEDGMENTS

The authors wish to acknowledge themselves for all their efforts and resources inputted to ensure the success of the research manuscript. Also, thanks to their family for the support during this research.

CONFLICT OF INTEREST

There is no conflict of interest.

REFERENCES

Azevedo, F. A. C., Carvalho, L. R. B., Grinberg, L. T., Farfel, J. M., Ferretti, R. E. L., Leite R. E. P., Filho, W. J., Lent, R., & Herculano-Houzel, S. (2009). Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *The Journal of Comparative Neurology, 513*(5), 532-541. https://doi.org/10.1002/cne.21974

Budiharto, W. (2014). Robust vision-based detection and grasping object for manipulator using SIFT keypoint detector. International Conference on Advanced Mechatronic Systems (pp. 448-452). IEEE. https://doi.org/10.1109/ICAMechS.2014.6911587

Budiharto, W., Gunawan, A. A. S., Suroso, J. S., Chowanda, A., Patrik, A., & Utama, G. (2018). Fast object detection for quadcopter drone using deep learning. International Conference on Computer and Communication Systems (pp. 192-195). IEEE. https://doi.org/10.1109/CCOMS.2018.8463284

COCO. (2021). Common Objects in Context. Retrieved from https://cocodataset.org/#home

Graetz, F. M. (2018). RetinaNet: how Focal Loss fixes Single-Shot Detection. Retrieved from https://towardsdatascience.com/retinanet-how-focal-loss-fixes-single-shot-detection-cb320e3bb0de

Hui, J. (2018). SSD object detection: Single Shot MultiBox Detector for real-time processing. Retrieved from https://jonathan-hui.medium.com/ssd-object-detection-single-shot-multibox-detector-for-real-time-processing-9bd8deac0e06

Li, J., Chen, X., Hovy, E., & Jurafsky, D. (2015). Visualizing and understanding neural models in NLP. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 681-691). Association for Computational Linguistics.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. Lecture Notes in Computer Science: European conference on computer vision (vol. 8693, pp. 740-755). Cham: Springer. https://doi.org/10.1007/978-3-319-10602-1_48

Medium. (2021). Object Detection with SSD and MobileNet. Retrieved from https://medium.com/@aditya.kunar_52859/object-detection-with-ssd-and-mobilenetaeedc5917ad0

Wang, S. C. (2003). Artificial neural network. In *Interdisciplinary computing in java programming* (pp. 81-100). Boston, US: Springer.

Yeremia, H., Yuwono, N. A., Raymond, P., & Budiharto, W. (2013). Genetic algorithm and neural network for optical character recognition. *Journal of Computer Science, 9*(11), 1435-1442. https://doi.org/10.3844/jcssp.2013.1435.1442