Comparative Analysis of YOLOv3, YOLOv4 and YOLOv5 for Sign Language Detection

Sahla Muhammed Ali

Graduate Student, Department of Information Technology, Rajagiri School of Engineering and Technology, Kerala, India

ABSTRACT

Sign language is a visual means of communication using hand signals, gestures and body language. It's the main form of communication for deaf people. Hearing-impaired people often find it quite challenging to communicate with others because most of the people do not know sign language. This requires a translator in most cases. Deep learning-based detection methods can solve this issue. Deep learning makes use of different methodologies that can extract features from images, videos and give the conclusion about the different objects. These methods find the meaning of a particular sign from the visual representation of the sign. This paper will discuss the YOLO algorithm used for object detection. YOLO algorithm uses Convolutional Neural Network (CNN) to detect objects in real-time. The algorithm requires only a single forward propagation through a neural network to detect objects. This means that the prediction of the entire image is done in a single algorithm run. YOLO v3 uses Darknet53 as the backbone feature extractor. The YOLOv4 architecture is composed of CSPDarknet53 as a backbone, spatial pyramid pooling additional module, PANet path-aggregation neck and YOLOv3 head [1]. YOLOv5 model can be summarized as Focus structure and CSP network Backbone, SPP block and PANet Neck and YOLOv3 head using GIoU-loss. This paper discusses the performance comparison of YOLOv3, YOLOv4 and YOLOv5 for sign language detection. The comparative analysis is done by using the same data set of sign languages for all three methodologies. The recall, precision and accuracy of each of the algorithms are discussed in the paper.

Keyword: *Deep learning, YOLOv3, YOLOv4, YOLOv5, Convolutional Neural Network, Sign language.*

1. INTRODUCTION

Sign language is used as the first language by the people with hearing impairment. The main components of sign language are postures and gestures [2]. The hearing impaired people find it challenging to express their ideas with others due to the lack of sign language knowledge. To solve this issue Deep learning models can be used to act as a translator. The deep learning techniques make use of different object detection models to carry out this task.

Object detection is the prediction of location of an object along with the class of the object. Here the model needs to predict the location using the rectangular or bounding box. It takes 4 variables to identify the exact location of the rectangle. So, for each object in an image the following variables are to be predicted – class name, bounding box top left x –coordinate, bounding box top left y-coordinate, bounding box width, bounding box height.

The object detector takes the input image and the features of the image are compressed by the convolutional neural network backbone. In object detection multiple bounding boxes are drawn along with the classification of the

image. The object detectors are of two types: one-stage detector and two-stage detector. The one-stage detector does the object localization and the classification at the same time whereas the two-stage detector decouples the task for each bounding box. The YOLO model is a one-stage detector and hence the name You Only Look Once [1].

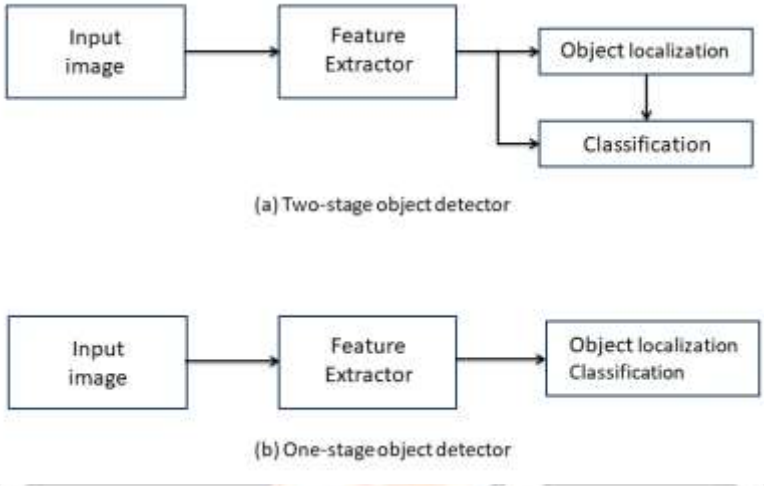

Fig -1: Two-stage and one-stage object detector.

YOLO was proposed by Joseph Redmond in 2015. The algorithm divided each image into a S x S grids and each grid predicts N bounding boxes and confidence. The confidence value indicates the accuracy of the particular bounding box and the presence of object in the box. In short, S x S x N boxes are predicted. The boxes with low confidence values are rejected [4]. The different variances of YOLO algorithm are YOLOv1, YOLOv2, YOLOv3, YOLOv4 and YOLOv5. This paper will compare the YOLOv3, YOLOv4 and YOLOv5 algorithms for the same dataset.

2. METHODOLOGY

YOLOv3 uses Darknet Architecture and has 53 layers trained with ImageNet dataset [6]. YOLOv3 uses residual connections and upsampling. The detection is performed at three different scales. It is more efficient in detecting smaller objects however; it has a longer processing time as compared to the previous versions [5].

The YOLOv4 architecture is composed of CSPDarknet53 as a backbone, spatial pyramid pooling additional module, PANet path-aggregation neck and YOLOv3 head [1].

The YOLOv5 model consists of Focus structure and CSP network as the backbone. The neck is composed of SPP block and PANet. It has a YOLOv3 head using GIoU-loss. YOLOv5 is written in Python programming language however the previous versions are written in C. This makes the installation and integration on IoT devices easier.

The different stages involved in the analysis of the 3 versions of YOLO are shown in the figure Fig-2.

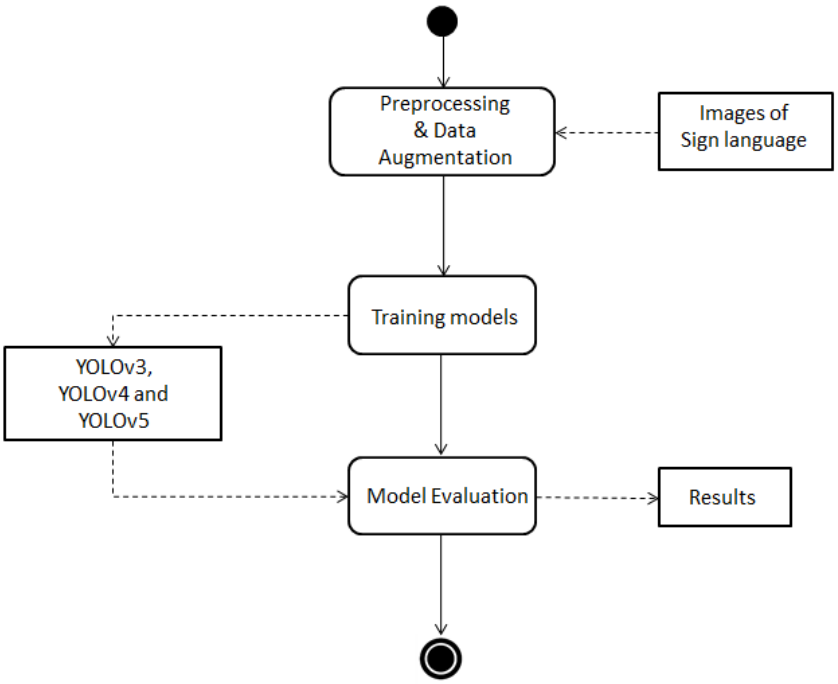

Fig -2: Stages of analysis

First the dataset was preprocessed and augmented. The dataset was then trained and tested using each version. Then the evaluation metrics were compared to know about the accuracy of each of the version of YOLO.

3. RESULTS AND DISCUSSIONS

The training was done using 100 epochs for 2000 images on YOLOv3, YOLOv4 and YOLOv5 respectively. The tensor board module was used to get the metrics for each of the methods for comparison. Table-1 show the precision, recall and mAP_0.5 for each of the algorithm.

Table -1: Precision, Recall and mAP_0.5 of YOLOv3, YOLOv4 and YOLOv5

YOLO Version	Precision	Recall	mAP_0.5
YOLOv5	0.72	0.81	0.87
YOLOv4	0.70	0.80	0.85
YOLOv3	0.52	0.69	0.71

The graphs for precision, recall, mAP_0.5 and mAP_0.5:0.95 for each of the three versions generated using the tensor board module.

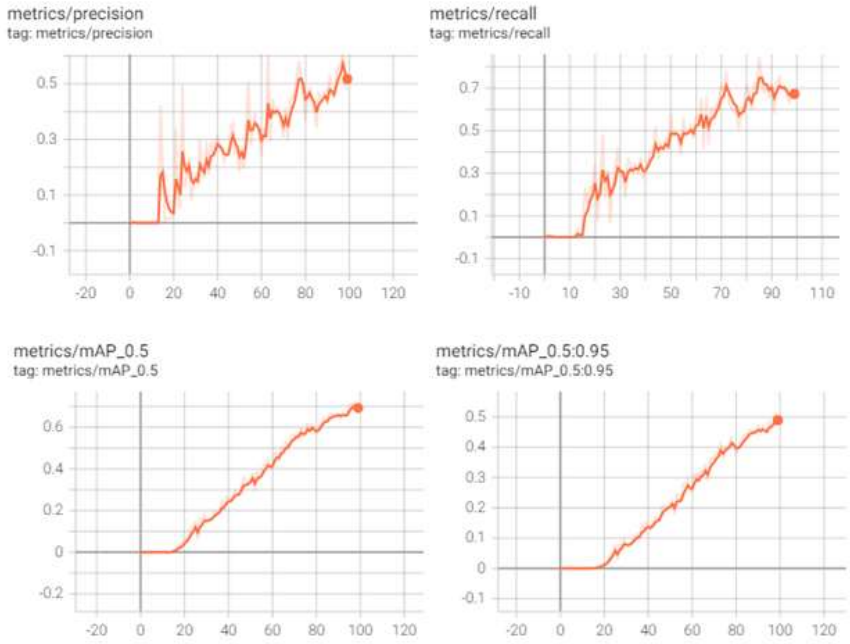

Fig -4: The graph of YOLOv3

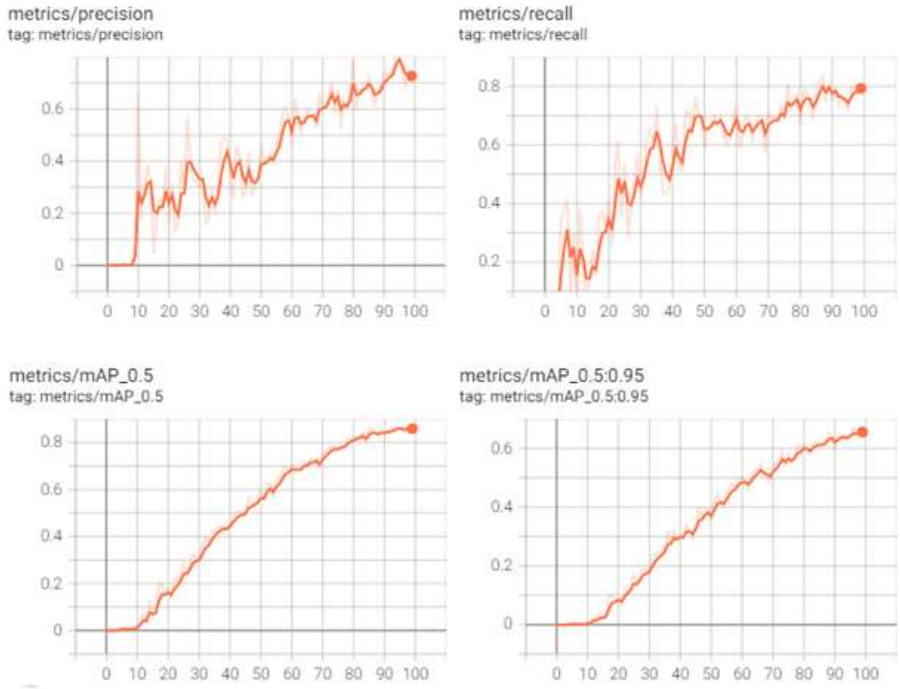

Fig -4: The graph of YOLOv4

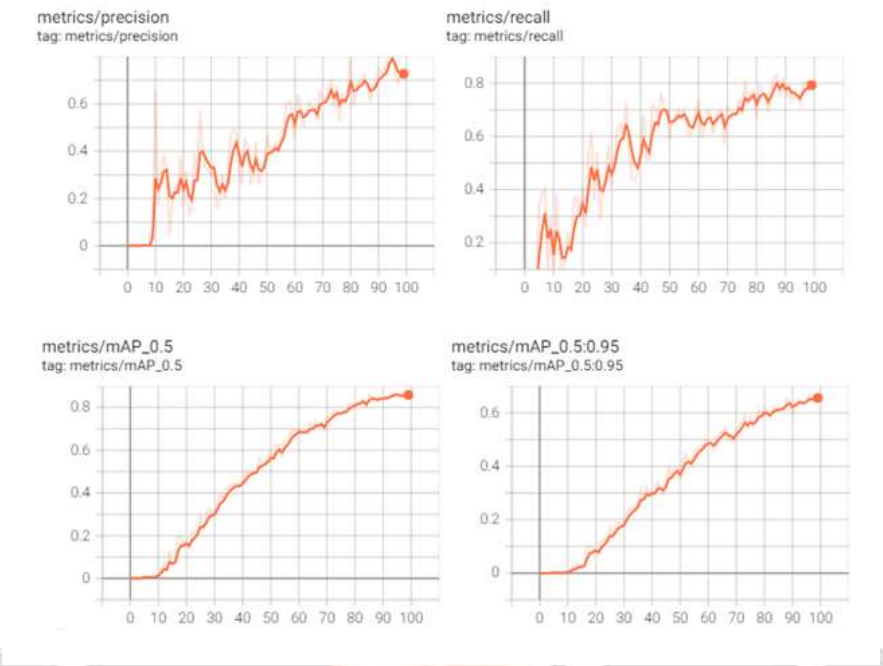

Fig -5: The graph of YOLOv5

4. CONCLUSIONS

The dataset for sign language was trained with YOLOv3, YOLOv4 and YOLOv5 algorithms. The accuracy, recall and the precision of the result from each of the algorithm were found out. By comparing the accuracy values and the results obtained from the three versions the best suitable version is YOLOv5. But the time taken for YOLOv3 is lower as compared to the time taken for YOLOv4 and YOLOv5.

5. REFERENCES

[1]. Alexey Bochkovskiy,Chien-Yao Wang, Hong-Yuan Mark Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection" arXiv:1506.02640, 2020.

[2]. M. A. Jalal, R. Chen, R. K. Moore, L. Mihaylova, "American sign language posture understanding with deep neural networks," in 2018 21st International Conference on Information Fusion (FUSION), 573–579, IEEE, 2018, doi: 10.23919/ICIF.2018.8455725.

[3]. H. Lahamy and D. D. Lichti, "Towards real-time and rotation invariant American sign language alphabet recognition using a range camera," *Sensors (Switzerland)*, vol. 12, no. 11, pp. 14 416–14 441, 2012.

[4] Joseph Redmon , Santosh Divvala, Ross Girshick , Ali Farhadi "You Only Look Once: Unified, Real-Time Object Detection" Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 2016, doi: 10.1109/CVPR.2016.91.

[5] Redmon J., Farhadi A. YOLOv3: An Incremental Improvement. *arXiv*:1804.02767, 2018.

[6] Krizhevsky A., Sutskever I., Hinton G.E. "ImageNet classification with deep convolutional neural networks" Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS '12); Lake Tahoe, NV, USA. 3–6 December 2012; pp. 1097–1105.

[7] L. Y. Bin, G. Y. Huann, L. K. Yun, "Study of Convolutional Neural Network in Recognizing Static American Sign Language," in 2019 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 41–45, IEEE, 2019, doi:10.1109/ICSIPA45851.2019.8977767.

[8] W. Tao, M. C. Leu, Z. Yin, "American Sign Language alphabet recognition using Convolutional Neural Networks with multiview augmentation and infer- ence fusion," Engineering Applications of Artificial Intelligence, 76, 202–213, 2018, doi:10.1016/j.engappai.2018.09.006.

[9] J. Wu, L. Sun, R. Jafari, "A wearable system for recognizing American sign language in real-time using IMU and surface EMG sensors," IEEE journal of biomedical and health informatics, 20(5), 1281–1290, 2016, doi: 10.1109/JBHI.2016.2598302.

[10] V. Bheda, D. Radpour, "Using deep convolutional networks for gesture recog- nition in American sign language," arXiv preprint arXiv:1710.06836, 2017.