Multi-User Real-time Sign Language Recognition of Words using Transfer Learning

Rashmi Gaikwad (rsgaikwad2020@gmail.com)
DKTE Society's Textile and Engineering Institute
Lalita Admuthe
DKTE Society's Textile and Engineering Institute

Short Report

Keywords: Sign language recognition (SLR), SSD MobileNet V2 FPNLite 320x320, SSD ResNet50 V1 FPN 640x640, OpenCV, Tensorflow, Convolutional Neural Network (CNN), Dataset

Posted Date: April 14th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2777600/v1

Abstract

This paper presents a solution for real-time sign language recognition of words using transfer learning of pre-trained modules available in Tensorflow object detection API. In this research a dataset of 800 images of 8 signs of words (100 images per sign) is generated using OpenCV and labelImg tool. SSD MobileNet V2 FPNLite 320x320 (Single Shot Detector MobileNet) network is implemented first to train and test the images of words using transfer learning technique and the results are compared with that of SSD ResNet50 V1 FPN 640x640 network for same dataset. This research aims to obtain real-time sign language detection of words. Signs obtained from a single user are used for training and testing but real-time detection can be done on signs performed by multiple users. The purpose of this study is to show that one can create his own dataset of signs with simple gestures and communicate with people by converting the sign language to text without any need to wear sensor gloves or any type of hardware.

1. Introduction

Sign Language is a way of communication between the people with hearing impairment and normal people. Normal people don't understand sign language very well. So, an intelligent sign language recognition system which requires minimum efforts from the signer to perform signs will be useful in bridging the communication gap between two communities.

Sign language recognition can be divided into two main categories. First is the glove based approach where the signer is required to wear gloves mounted with sensors and data will be collected and processed from the sensors. Second is the vision based approach which is subdivided into static and dynamic sign language recognition systems. Static signs are processed and detected from images of signs which may be captured in real time or from an available dataset. Dynamic signs are captured in real time by a web camera and processed and detected. The sign language recognition part is carried out by an AI network. The AI network used must be flexible, accurate and precise. The accuracy of the AI network must not be affected by the background conditions and variations.

The deaf and dumb people that are the people with hearing impairment use various sign languages for communication. Some of them are American Sign Language, Chinese Sign Language, Indian Sign Language, Arabic Sign Language etc. Indian sign language varies not only from one state to another state but also from region to region. Comparatively ASL is easier to do research work as it is less complicated and is not varying like ISL. So, this work is focusing first on ASL and then it could be applied on ISL.

2. Related Work

Over the year's sign language recognition systems is done using deep learning neural networks. Convolutional Neural Network based sign language recognition systems are most widely used because of their higher accuracy and precision. One such CNN model for ASL recognition is used in [1]. This system

is designed to recognize signs of letters only and is implemented using Keras, Tensorflow and openCV. The accuracy of recognition is 99.838%. Same CNN model is used for the recognition of Bhutanese sign language of digits in [2]. They have also compared the results of CNN with SVM, KNN, Logistic regression and LeNet5. The performance was evaluated using the parameters like accuracy, precision, recall and F1-score.

Generating text and speech from the input gesture sign language is achieved in [3]. In [4] sign language recognition is done using inflated 3D convolutional networks i.e. I3D ConvNet. Here, the ChaLearn249 dataset is used which consists of images and videos of sign language. In [5] deep learning using CNN is used to locate the signer hands by using 5 layers of CNN and average pooling. The activation function used is ReLu. Long and short term memory neural network (LSTM) is used for encoding and decoding of input frames of variable length by obtaining the temporal structure information. The accuracy of the recognition rate of the network used in this paper is 99%.

A 3D combinational neural network sign language recognition system by extracting spatial-temporal information of signs is proposed in [6]. The region of interest i.e. ROI is the part in an image frame which consists the information about the gestures. All the background is subtracted from the image frame which is obtained from the signer video. The work is focused on large vocabulary Chinese sign language recognition. This system is only implemented for static images and only the signs of words are detected. An accuracy of 94.3% is achieved using this technique. A real time traffic sign recognition system using faster recurrent combinational neural networks and MobileNet is proposed in [7]. The limitations of colors and space which vary from sign to sign have been eliminated by using the RGBN color space and detection of contours and centers of traffic signals. An accuracy of 84.5% and recall of 94% is achieved using this method. A fully convolutional neural network for the detection and recognition of traffic signs is used in [8]. This work is divided into two stages. In the first stage the size and orientation of the traffic sign is detected and in the second stage the text of the traffic sign is detected. A precision of 93.5% and recall of 94% is achieved using this technique.

A method to detect signs of English letters from A to Z in American Sign Language by using Neural Networks is proposed in [9]. The results in this paper show that the speed of processing is improved by the use of neural network and the dataset can be increased to any number of variables. A comparison of the efficiency of different types of Neural Networks in Arabic sign Language Gesture recognition for static as well as dynamic signs is done in [10] and proved that the fully recurrent neural networks have the highest accuracy and minimum error rate. A database of signs for 28 letters of Arabic Sign Language is generated and after training the Neural Network and the signs are detected. The accuracy of recognition using this technique is found to be 95.11%.

Some researchers have used hardware like leap motion controllers[11, 16], Kinect sensors[13, 14], accelerometer and gyroscopic sensors [25, 26] and other such hardware devices [26] to capture hand movements while performing signs and after feature extraction the dataset is fed for training of the AI network which may be a CNN, Support vector machine[16, 22], Self-organizing maps[20], Dynamic time

warping algorithm [13], fuzzy networks[17, 18, 19], Transition movement models[21] and Principal Component Analysis[30, 31]. A method to recognize Indian sign language gestures with the use of flex sensors, gyroscope, accelerometer and microcontroller is proposed in [12]. A platform for hardware and software i.e. microcontroller programming and interfacing is achieved by making the use of Arduino.

Signs of words in a video are recognized using Reinforcement learning and spatial-temporal CNN and bidirectional LSTM in [23].The performance of this system is mapped using the parameter word error rate which comes 28.6% in this research. Hidden Markov Model and K-nearest neighbor are used for recognition of signs in [24]. Real time Indian sign language recognition by using image processing techniques is done in [27]. A system for dynamic sign language recognition system for smart home interaction is proposed in [28]. Meaningful sentences are formed by the use of stochastic linear formal grammar (SLFG) module. This system is not applied to real-time videos and it can be expanded for sign language communication in real time. The accuracy with this system is 98.65%.

Convolutional neural network in combination with long short-term memory (LSTM) in [29] is the most recent development in sign language recognition where the data is collected and signs are recognized using OpenCV and computer vision. This method achieved an accuracy of 95.5%. A human posture recognition system for video surveillance using different classifiers like K Means, Fuzzy C Means, and Multilayer Perceptron, Self-organizing Maps and Neural Networks is proposed in [32]. In [33], a combination of ANN, k-nearest neighbor and support vector machine is used for Italian sign language recognition. This system achieves an accuracy of 93.1%. In [34], dynamic hand gesture recognition system is implemented by using a fusion of 1-D and 2-D CNN with temporal convolutional network. 16 features of each sample are generated using data gloves. An attention-based encoder-decoder model with a multi-channel convolutional neural network is proposed in [35] with a Word Error Recognition (WER) = 10.8%. It translates Chinese sign language into voices.

3. Methodology

Deep learning neural network consists of convolutional layer, MaxPooling layer and fully connected layer.

Tensorflow Object detection API is the framework for creating a deep learning network that solves object detection problems. This module consists of a variety of detection models which are pre-trained on the COCO 2017 dataset. SSD MobileNet V2 FPNLite 320x320 and SSD ResNet50 V1 FPN 640x640 are used for the detection of static and real-time signs in this research. Transfer learning is used to train the two models on the dataset generated in this research. The SSD (Single Shot Detector) architecture is a single convolution network that learns to predict bounding box locations and classify these locations in one pass. The step by step procedure is divided into 5 steps which are- data acquisition, training the network, testing the network, sign recognition in real-time and results.

3.1. Data acquisition

A dataset of signs of five words ‘hello’, ‘yes’, ‘no’, ‘thank you’ ‘thumbs up’, ‘thumbs down’ and ‘one’ was created using openCV. Images were captured using web camera. 100 images of each word were captured so the dataset consisted of 100x8 = 800 images for training from a single user and testing dataset consisted of 15x8 = 120 images. Labeling of images was done using the graphical image annotation tool labelImg where only the hand part was annotated.

3.2. Training the network

First, the SSD MobileNet V2 network is trained on 100 images of each sign by transfer learning of the pre-trained model and tested on 120 images of each sign. The number of steps for training (epoch) is set to 2000 to achieve best results. The learning rate goes on increasing till 0.08 and then it becomes constant as shown in Fig. 1. The loss metric while training the network starts at 0.7 and goes on reducing till approximately 0.4 at the last step of training as shown in Fig. 2.

The same dataset is used for training and testing of the second network i.e. SSD ResNet50 V1. The learning rate of this network goes on increasing till 0.04 as shown in Fig. 3. The loss metric while training the network goes on reducing till 0.5 as shown in Fig. 4. This training of networks was done on the computer without GPU.

3.3. Testing the network

The trained network is tested on 120 different images of the same signs. Average Precision (AP) and Average Recall (AR) are the evaluation parameters and accuracy in terms of confidence level in percentage is used to indicate the detection of signs in real-time. Figure 5 and Fig. 6 show the evaluation results of SSD MobileNet V2 and SSD ResNet50 V1 respectively.

The maximum precision and recall values of SSD MobileNet V2 are shown in Fig. 7. and Fig. 8. respectively and that of SSD ResNet50 V1 are shown in Fig. 9. and Fig. 10. respectively. The values are also tabulated alongwith the time required to train the two networks in Table 1. The time taken for MobileNet V2 is 1–2 hours whereas that for SSD ResNet50 V1 is nearly 24 hours.

Table 1
Comparison of training and evaluation parameters of the two networks.

Network	Learning Rate at epoch 2000	Total Loss at epoch 2000	Highest Precision Value Achieved	Highest Recall Value Achieved	Time Required for training of network on computer without GPU
SSD MobileNet V2 FPNLite 320x320	0.08	0.41	0.919	0.716	1–2 hours
SSD ResNet50 V1 FPN 640x640	0.04	0.5	0.876	0.743	24hours

3.4. Sign recognition in real-time

Once the network is trained and tested on the dataset of the signs the model can be used to detect real-time signs performed by a person in front of the web camera. OpenCV is used again to capture the images from the video at real-time. SSD uses bounding boxes for every image category while detecting signs in real time. Figure 9 and Fig. 10 show that real- time detection of the sign “Hello” is 100% using both the networks respectively.

3.5. Results

After training and testing of the SSD MobileNet V2 network real-time signs of the same words are detected successfully through the webcam. The same process of training and testing is done for another network SSD ResNet50 V1. It is observed that the time required to train SSD MobileNet V2 is much less (nearly 1 hour) than that required for SSD ResNet50 V1 (nearly 24hours). This training is without any graphics card available on the machine. It is also observed that the time for training goes on increasing if the number of input images is increased. Also, the confidence level of the signs detected in real time goes on increasing as the number of input images is increased from 50 to 100 images per sign. The precision and recall of SSD MobileNet V2 came to be 91% and 71% while that of SSD ResNet50 V1 came to be 87% and 74%. So, there is a difference of around 10% in the detection or evaluation parameters. The time required for training of the network is very much less in case of SSD MobileNet V2 compared to SSD ResNet50 V1. Also, this research proves that signs from different users can be detected in real-time even though the network is trained on signs obtained from a single user as shown in Fig. 11 and Fig. 12.

3.6. Conclusion

Thus in this research OpenCV and computer vision is used to generate a dataset of signs of 8 words. Two Tensorflow object detection modules SSD MobileNet V2 FPNLite 320x320 and SSD ResNet50 V1 FPN 640x640 are implemented to detect real-time signs of words. The networks were trained on the dataset generated using transfer learning of the pre-trained models The performance of both modules is compared using evaluation parameters like accuracy, precision, recall and time required for training of the network. Also, the modules are checked for detecting signs from different signers in real time. The future work will be inclusion of more signs of words and generate sentences from the detected words using grammar rules or sentence prediction modules.

Declarations

Conflict of Interest

The authors declare that they have no conflict of interest.

Funding

No funding was received from any organization for conducting this study.

Authors' Contributions

Rashmi Gaikwad and Lalita Admuthe conceived of the presented idea. Rashmi Gaikwad developed the theory and implemented the methodology. Lalita Admuthe verified the methodology and results. Lalita Admuthe encouraged Rashmi Gaikwad to investigate and supervised the findings of this work. All authors discussed the results and contributed to the final manuscript.

References

1. Kasapbasi A., Eltayeb A., Elbushra A., Al-Hardanee O., Yilmaz A. (2022): DeepASLR: A CNN based human computer interface for American Sign Language recognition for hearing impaired individuals. Elsevier, ISSN: 2666-9900.
2. Wangchuk K., Riyamongkol P., Waranusast R. (2020): Real-time Bhutanese Sign Language digits recognition system using convolutional neural network. ISSN:2405-9595, The Korean Institute of Communications and Information Sciences(KCIS), Elsevier B.V.
3. Thakur A., Budhathoki P., Upreti S., Shrestha S., Shakya S. (2020): Real-time Sign Language Recognition and Speech Generation. Journal of Innovative Image Processing Vol.02/ no. 02 pp. 65-76, ISSN: 2582-4252.
4. Sarhan N. and Frintrop S. (2020): Transfer Learning for Videos: From Action Recognition to Sign Language Recognition. IEEE ICIP.
5. Siming He (2019): Research of a Sign Language Translation System Based on Deep Learning. IEEE International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM).
6. Huang J., Zhou W., Li H. and Li W. (2018): Attention based 3D CNNs for Large Vocabulary Sign Language Recognition. IEEE Transactions on Circuits and Systems on Video Technology.
7. Li J. and Wang Z. (2018): Real-Time Traffic Sign Recognition Based on Efficient CNNs in the wild. IEEE Transactions on Intelligent Transportation Systems.
8. Zhu Y., Liao M., Yang M. and Liu W. (2018): Cascaded Segmentation-Detection Networks for Text-Based Traffic Sign Detection. IEEE Transactions on Intelligent Transportation Systems, Vol. 19, No.1,.
9. Mekala P., Gao Y., Fan J. and Davari A. (2011); Real-time Sign Language Recognition based on Neural Network Architecture. 978-1-4244-9592-4/11/©2011 IEEE.
10. Maraqa M. and Abu-Zaiter R. (2008): Recognition of Arabic sign language (ArSL) using recurrent neural networks. Applications of Digital Information and Web Technologies, pp: 478 – 481.
11. Avola D., Bernardi M., Cinque L., Foresti G. L. and Massaroni C. (2018): Exploiting Recurrent Neural Networks and Leap Motion Controller for the recognition of Sign Language and Semaphoric Hand Gestures. IEEE Transactions on Multimedia.
12. Plouffe G. and Cretu A. (2015): Static and Dynamic Hand Gesture Recognition in Depth Data Using Dynamic Time Warping. IEEE transactions on instrumentation and measurement.

13. Heera S. Y., Murthy M. K., Sravanti V. S. and Salvi S. (2017): Talking Hands – An Indian Sign Language to Speech Translating Gloves. International Conference on Innovative Mechanisms for Industry Applications (ICIMIA 2017) 978-1-5090-5960-7/17 IEEE.
14. Yao Y. and Fu Y. (2014): Contour model-based hand-gesture recognition using the Kinect sensor. IEEE Trans. Circuits Syst. Video Technol., vol. 24, no. 11, pp. 1935–1944.
15. Ren Z., Yuan J., Meng J., and Zhang Z. (2013): Robust Part-Based Hand Gesture Recognition Using Kinect Sensor. IEEE transactions on multimedia, vol. 15, no. 5.
16. Saini P., Saini R., Santosh Kumar Behera S., Debi Prosad Dogra D. and Roy P. P. (2017): Real-Time Recognition of Sign Language Gestures and Air-Writing using Leap Motion. Fifteenth IAPR International Conference on Machine Vision Applications (MVA) Nagoya University, Nagoya, Japan, May 8-12.
17. Jia G., Lam H., Ma S., Yang Z., Xu Y., Xiao B. (2020): Classification of Electromyographic Hand Gesture Signals Using Modified Fuzzy C-Means Clustering and Two-Step Machine Learning Approach. IEEE Transactions on Neural Systems and Rehabilitation Engineering, Volume28, Issue6.
18. Várkonyi-Kóczy A. R. and Tusor B. (2011): Human–computer interaction for smart environment applications using fuzzy hand posture and gesture models. IEEE Transactions on Instrumentation and Measurement, vol. 60, no. 5, pp. 1505–1514.
19. Futane P. R. and Dharaskar R. V. (2012): Video Gestures Identification and Recognition Using Fourier Descriptor and General Fuzzy Minmax Neural Network for Subset of Indian Sign Language**.** 978-1-4673-5116-4/12, IEEE.
20. Infantino I, Rizzo R., and Gagli S. (2007): A Framework for Sign Language Sentence Recognition by Common-sense Context**.** IEEE transactions on systems, man, and cybernetics—part c: applications and reviews, vol. 37, no. 5, September.
21. Fang G., Gao W. and Zhao D. (2007): Large-vocabulary continuous sign language recognition based on transition-movement models. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 37, no. 1, pp. 1–9.
22. Quan Y. (2010): Chinese Sign Language Recognition Based on Video Sequence Appearance Modelling. Industrial Electronics and Applications (ICIEA), the 5th IEEE Conference on, pp: 1537 – 1542.
23. Wei C., Zhao J., Zhou W., Li H. (2021): Semantic Boundary Detection with Reinforcement Learning for Continuous Sign Language Recognition. IEEE Transactions on Circuits and Systems for Video Technology Volume31, Issue:3.
24. Mohamed Hassan M., Khaled Assaleh K. and Tamer Shanableh T. (2016): User-dependent Sign Language Recognition Using Motion Detection. International Conference on Computational Science and Computational Intelligence IEEE.
25. Yang X., Chen X., Cao X., Wei S. and Zhang X. (2016).: Chinese Sign Language Recognition Based on An Optimized Tree. IEEE

26. Kosmidou V.E and Hadjileontiadis L. J. (2009): Sign Language Recognition Using Intrinsic-Mode Sample Entropy on sEMG and Accelerometer Data. IEEE transactions on biomedical engineering, vol. 56, no. 12.
27. Subha Rajam P. and Balakrishnan G. (2011): Real Time Indian Sign Language Recognition System to aid Deaf-dumb People. 978-1-61284-307-0/11/ ©2011 IEEE pp 737-742.
28. Abid M.R., Petriu E.M. and Amjadian E. (2015): Dynamic sign language recognition for smart home interactive application using stochastic linear formal grammar. IEEE IEEE Transactions on Instrumentation and Measurement, vol. 64, no. 3, pp. 596–605.
29. Li W., Pu H.,Wang R.(2021): Sign Language Recognition Based on Computer Vision. IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA).
30. Ahuja M. and Singh A.: Static Vision Based Hand Gesture Recognition Using Principal Component Analysis. IEEE (2015).
31. Mohandes M., Deriche M., and Liu J. (2014): Image-Based and Sensor-Based Approaches to Arabic Sign Language Recognition**.** IEEE transactions on human-machine systems, 2168-2291 IEEE.
32. Htike K. K., Khalifa O. O., Ramli H. A. M.and Abushariah A. M. (2014): Human Activity Recognition for Video Surveillance using Sequences of Postures", ISBN: 978-1-4799-3166-8 IEEE.
33. Calado A., Errico V., Saggio G. (2021): Toward the Minimum Number of Wearables to Recognize Signer-Independent Italian Sign Language with Machine-Learning Algorithms. IEEE Transactions on Instrumentation and Measurement (Volume:70) 10.1109/TIM.2021.3109732 .
34. Liu J.; Yan W., Dong Y. (2021): Dynamic Hand Gesture Recognition Based on Signals from Specialized Data Glove and Deep Learning Algorithms**.** IEEE Transactions on Instrumentation and Measurement (Volume:70).
35. Wang Z., Zhao T., Ma J., Chen H., Liu K., Shao H., Wang Q., Ren J. (2020): Hear Sign Language: A Real-time End-to-End Sign Language Recognition System. IEEE Transactions on Mobile Computing.

Figures

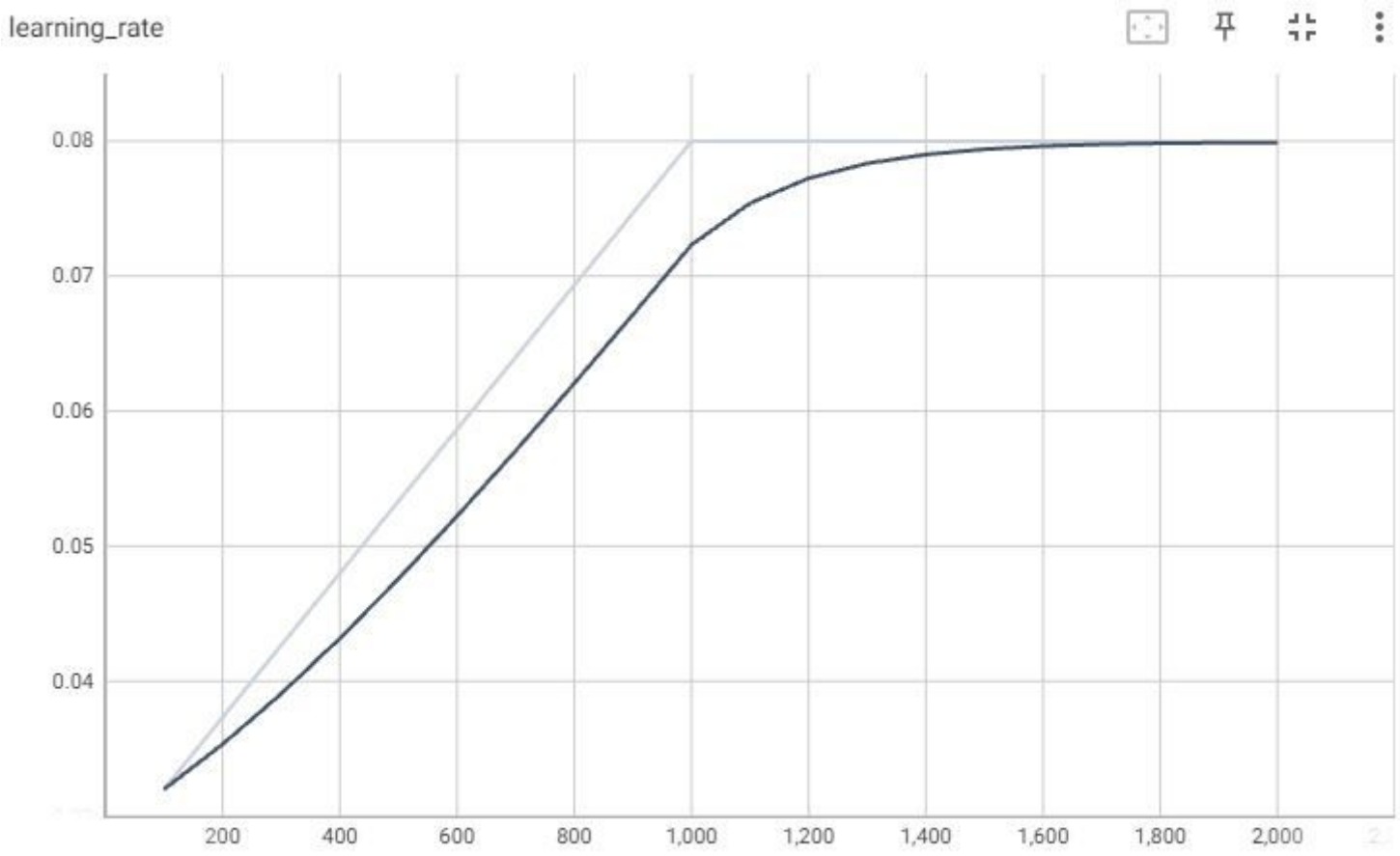


Figure 1

learning rate of SSD MobileNet V2

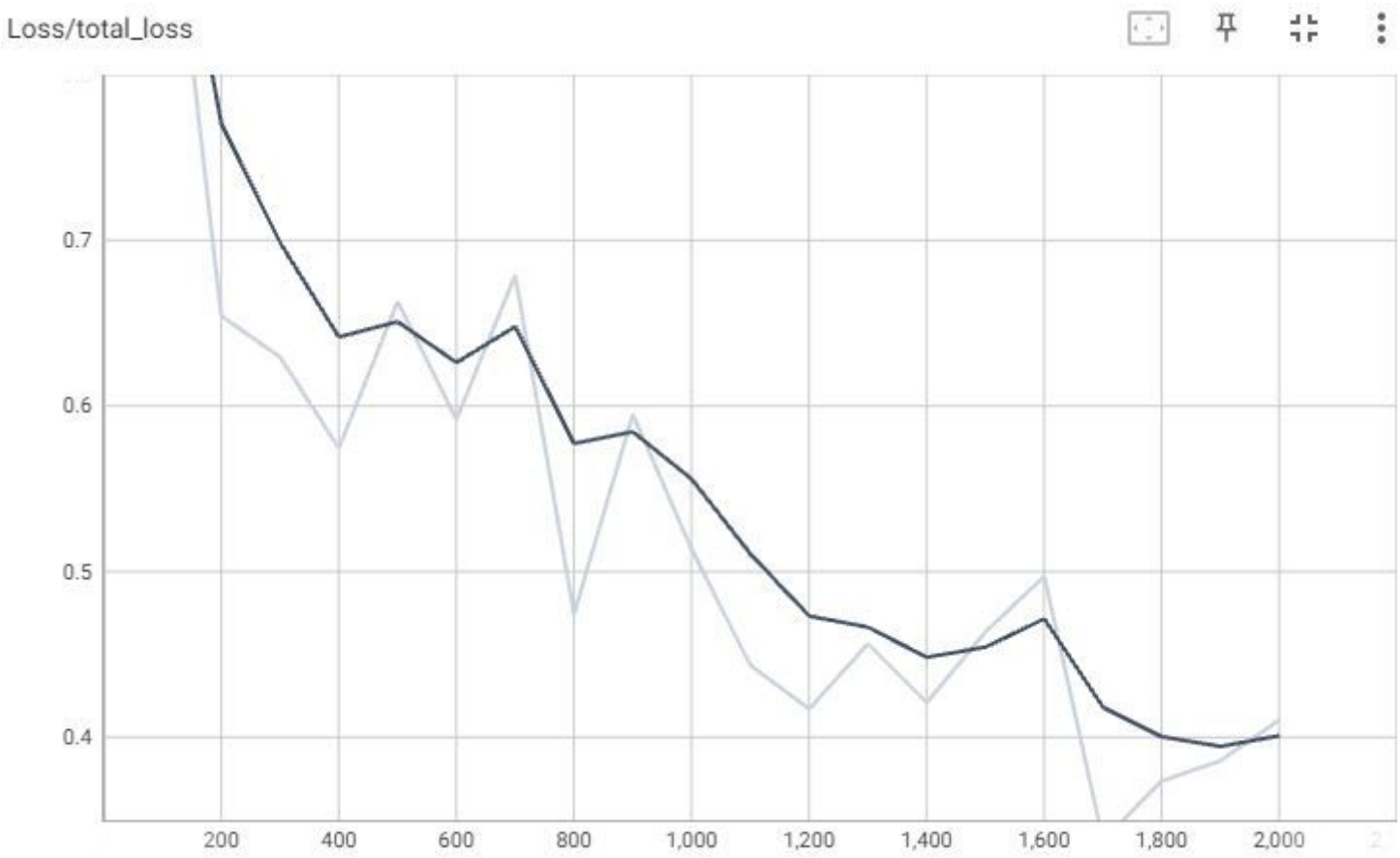


Figure 2

Total loss of SSD MobileNet V2

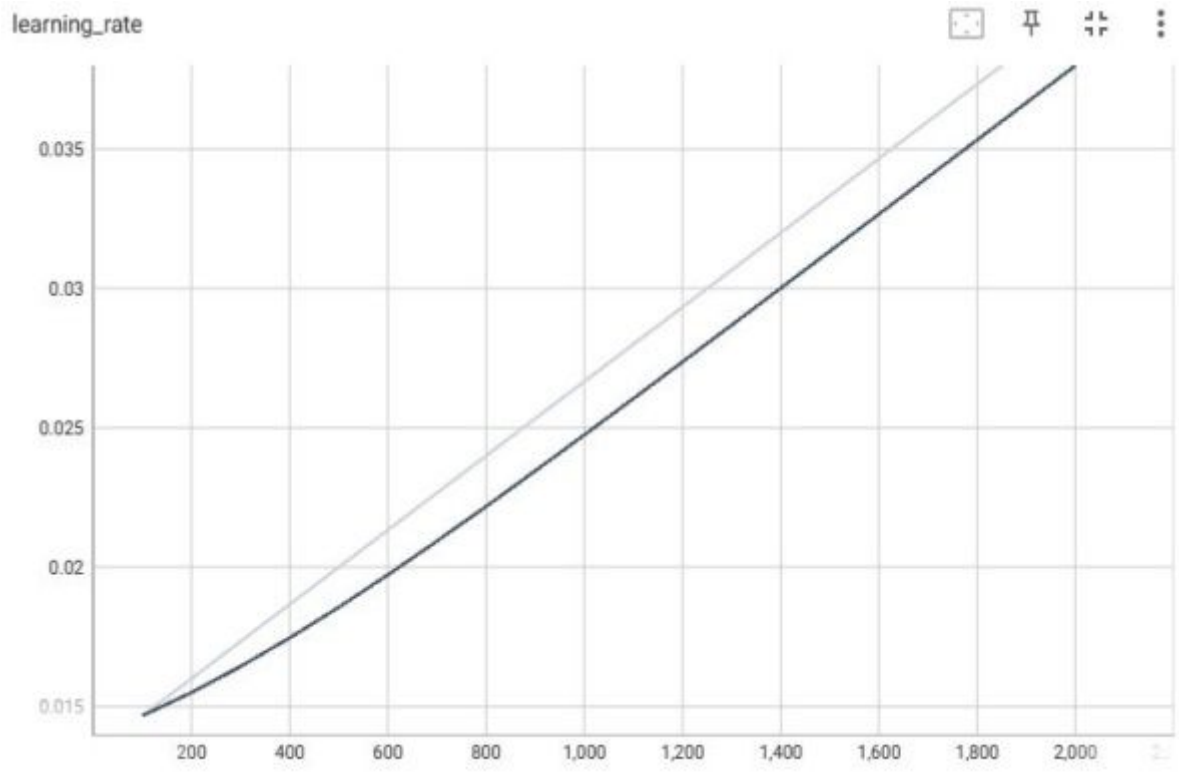


Figure 3

learning rate of SSD ResNet50 V2

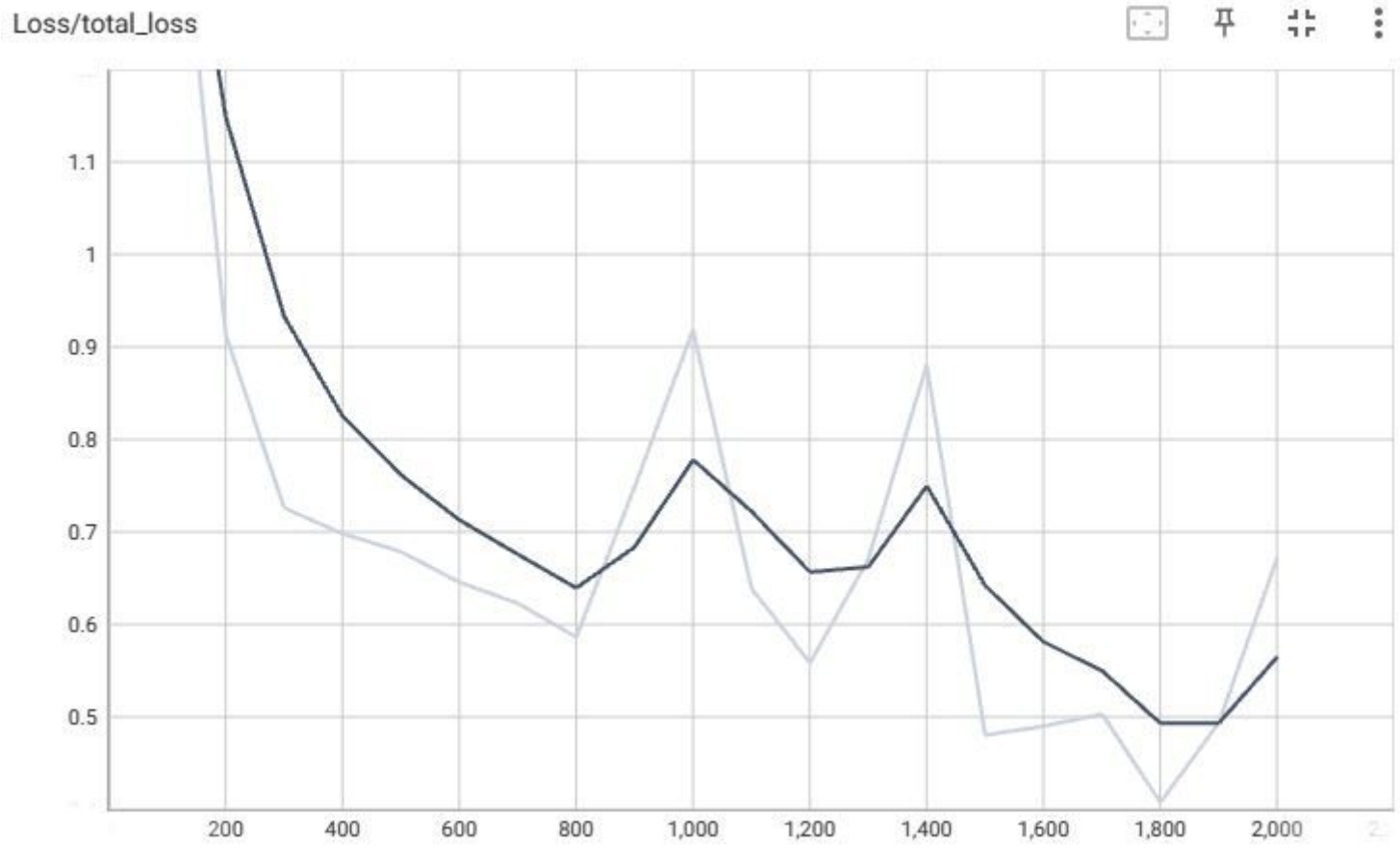


Figure 4

Total loss of SSD ResNet50 V2

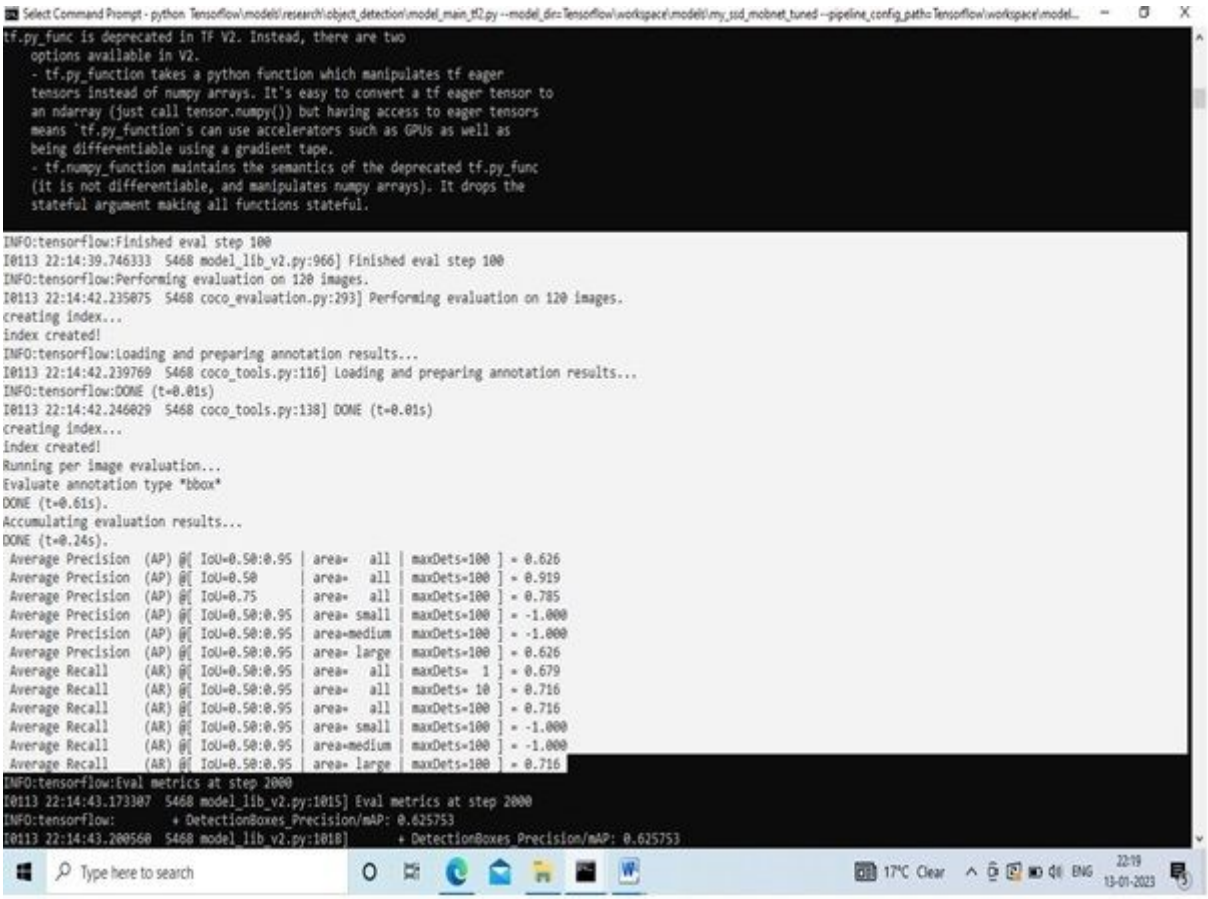


Figure 5

SSD MobileNet V2 FPNLite 320x320 testing

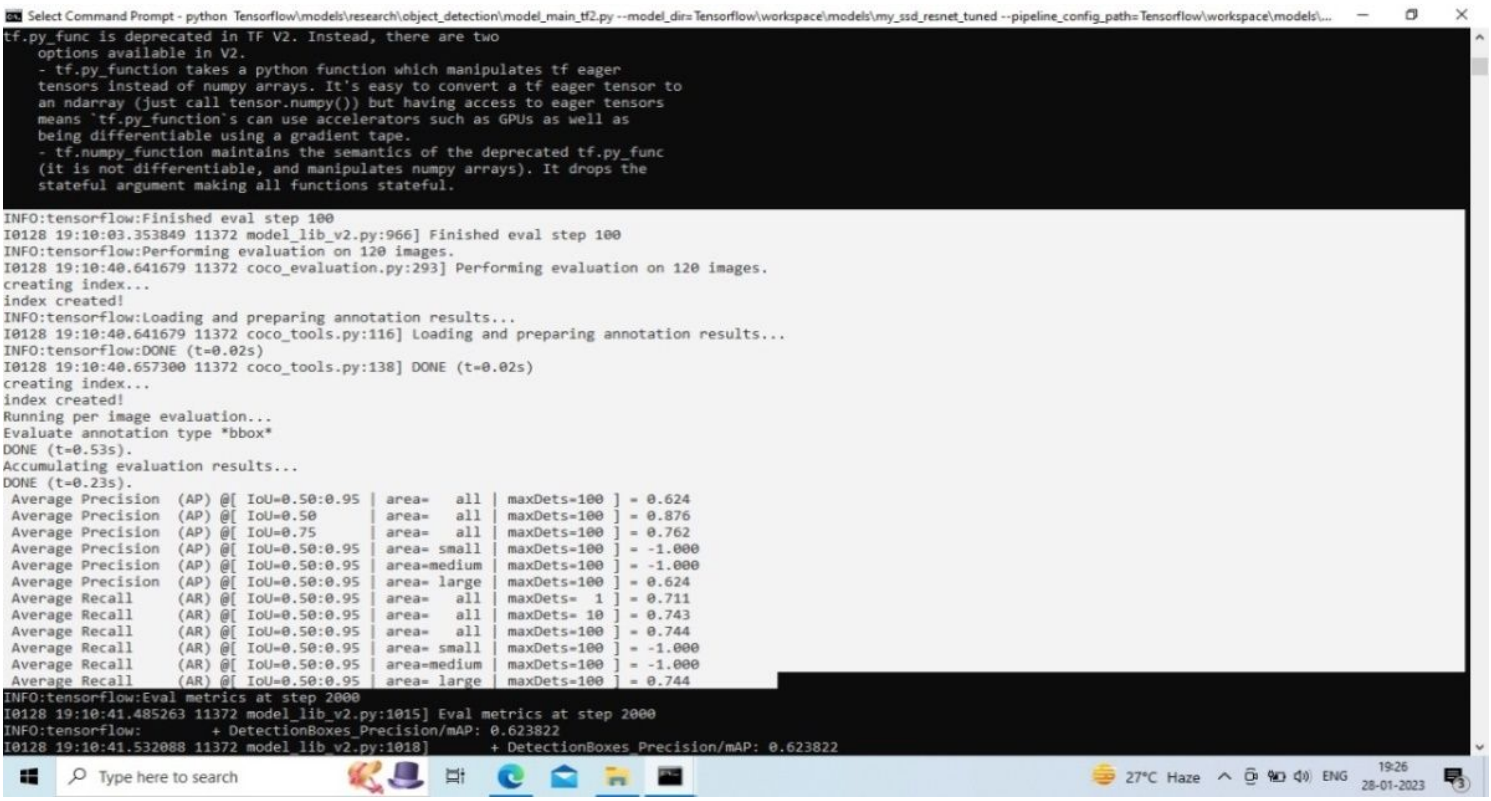


Figure 6

SSD ResNet50 V1 FPN 640x640 Testing

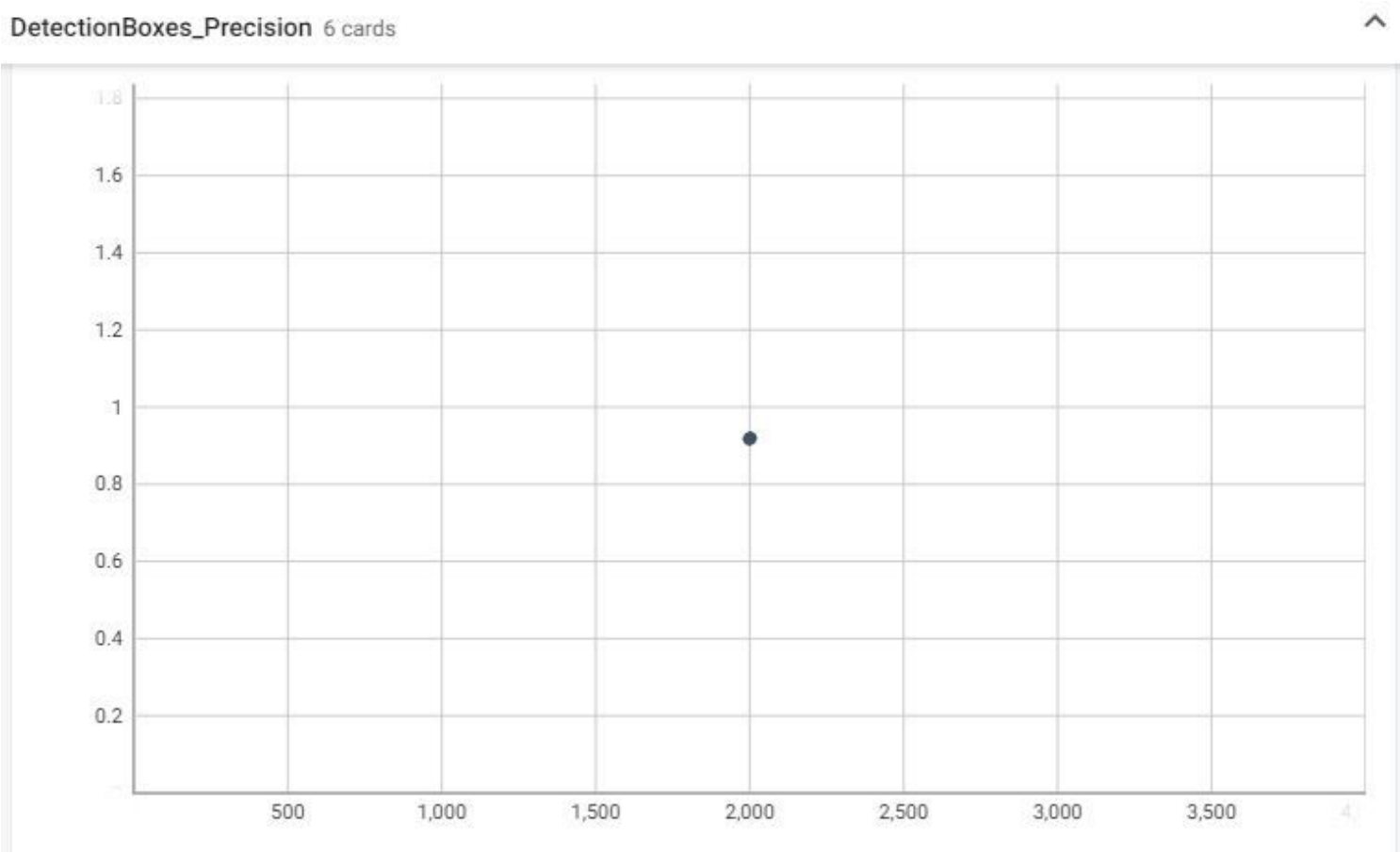


Figure 7

SSD MobileNet V2 precision

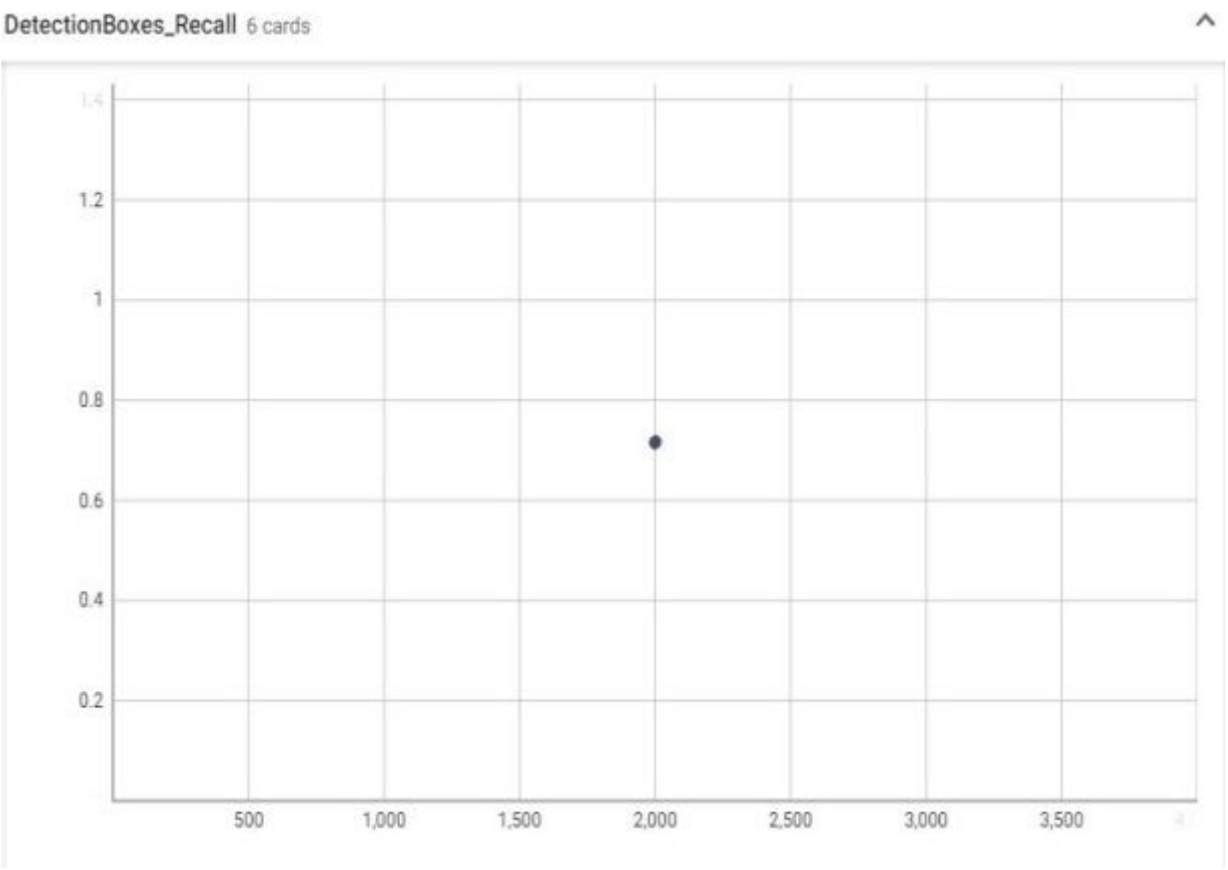


Figure 8

SSD MobileNet V2 recall

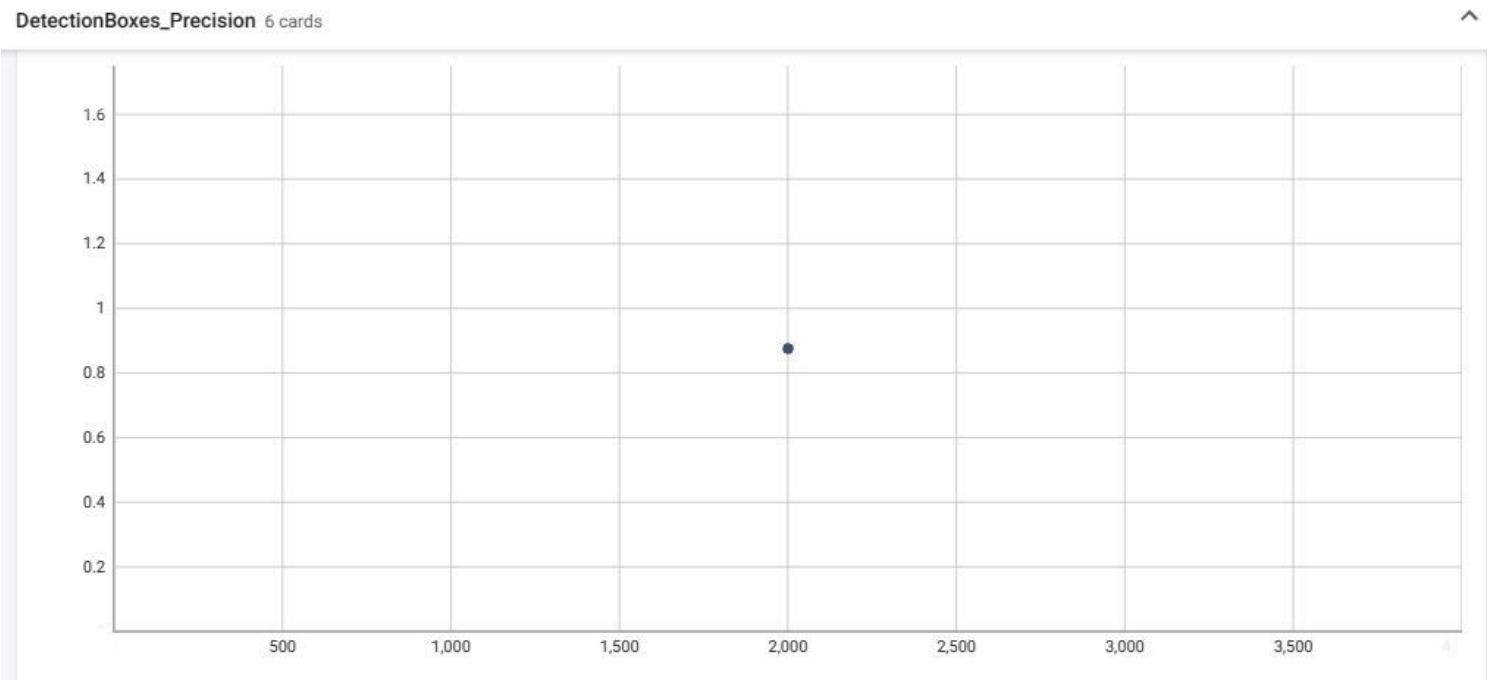


Figure 9

SSD ResNet50 V1 precision

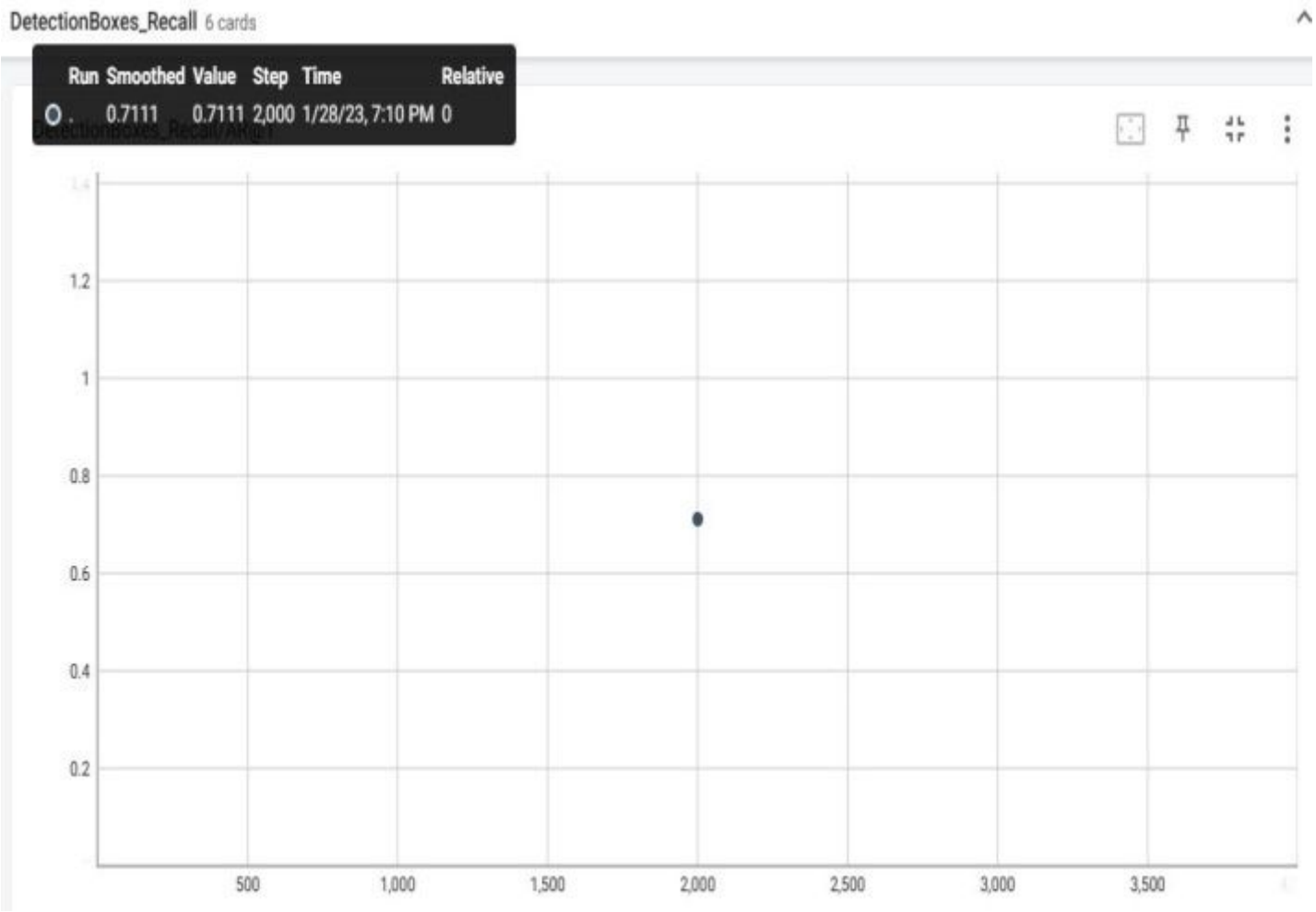


Figure 10

SSD ResNet50 V1 recall

Figure 11

Fig. 9 Hello sign detected in real-time using SSD MobileNet V2

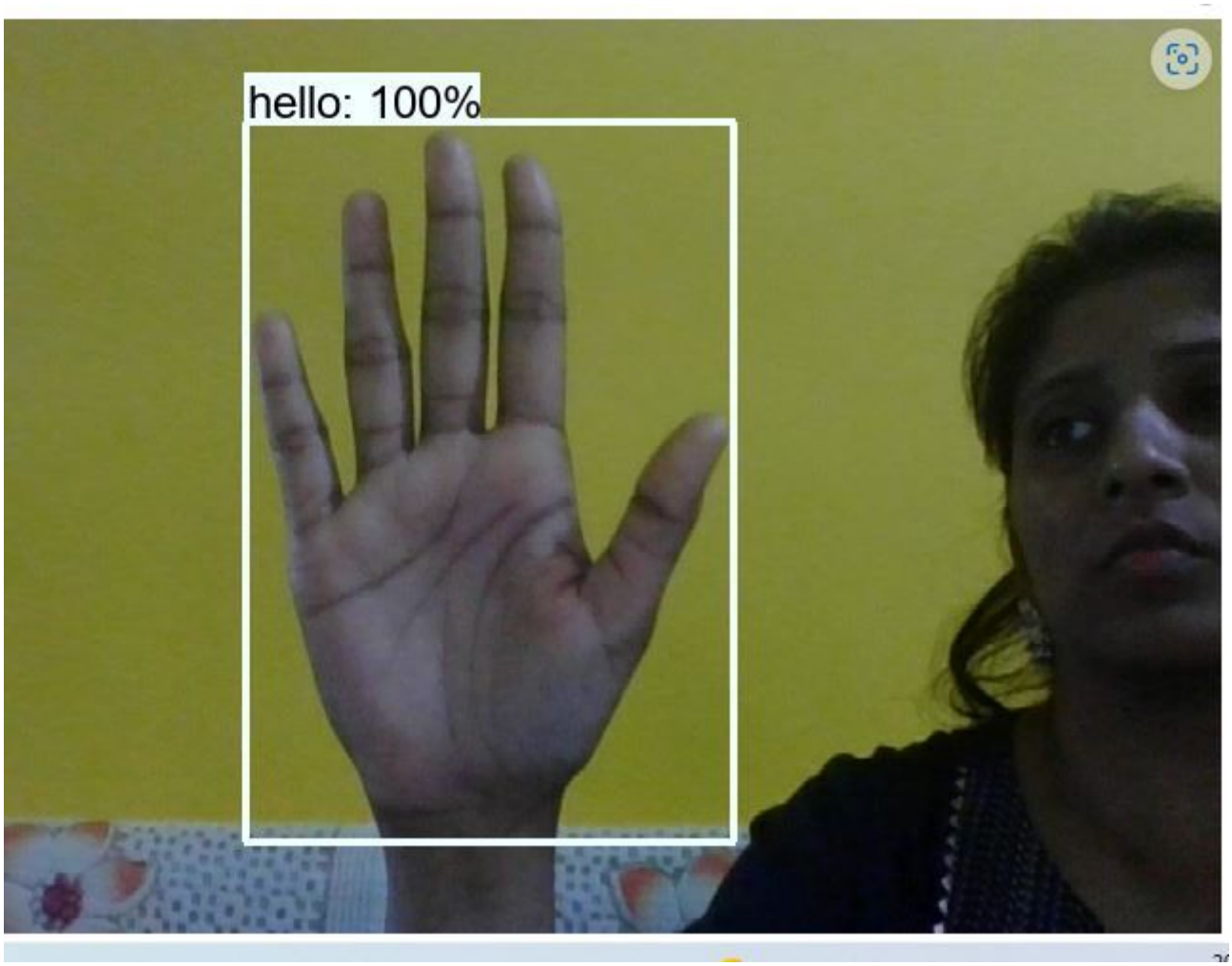


Figure 12

Fig. 10 Hello sign detected in real-time using SSD ResNet50 V1

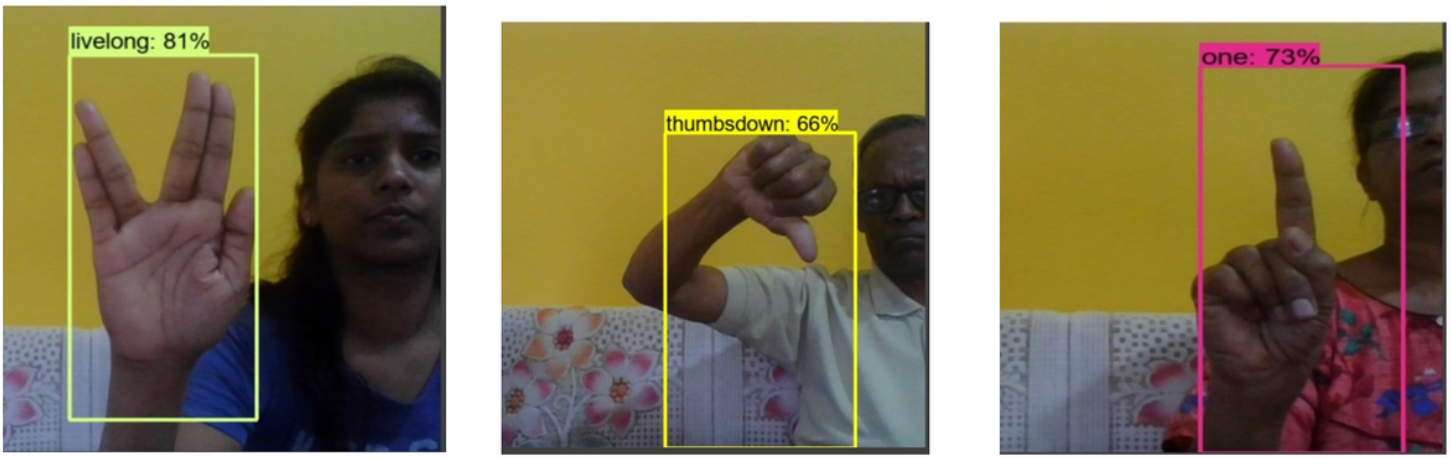
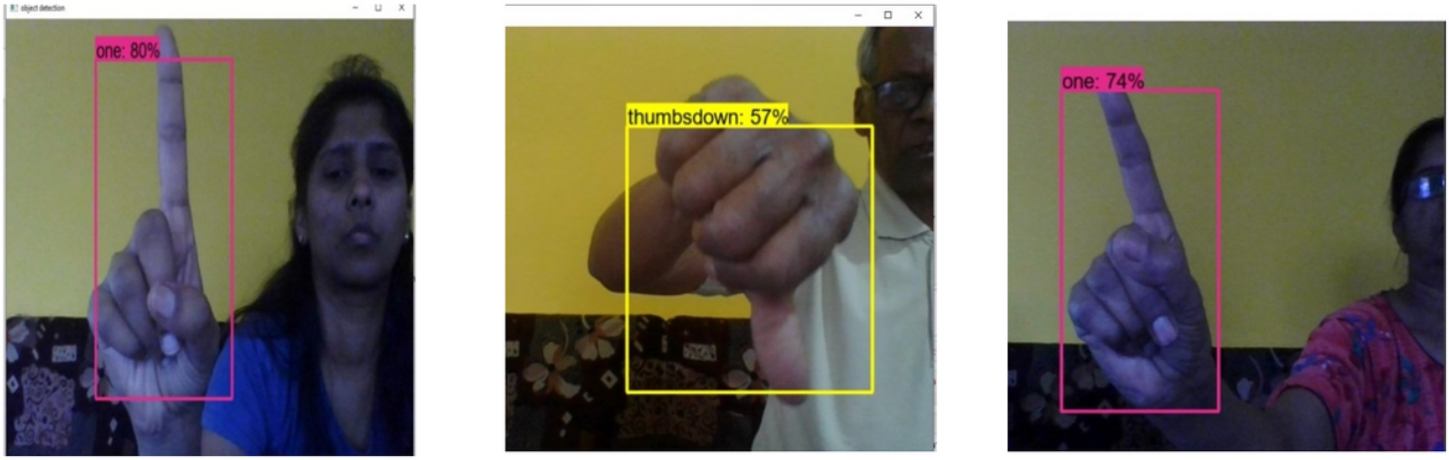


Figure 13

Fig. 11 Real time output of SSD MobileNet V2



Signer1 Signer 2 Signer 3

Figure 14

Fig. 12 Real-time output of SSD ResNet50 V1