Inference time of YOLOv5

Team C

Sara Mohajerani

October 24, 2022

The Yolov5 object detection algorithm has been trained on Kaggle data set in three versions of small, medium and large. The inference time for object detection is calculated and reported as the following.

YOLOv5 small

Inference time of GPU

Here you can find the inference time of Yolov5s model on 470 images of kaggle data set, which selected as a test set in Fig. 1. Fig. 2 shows the histogram of these time.

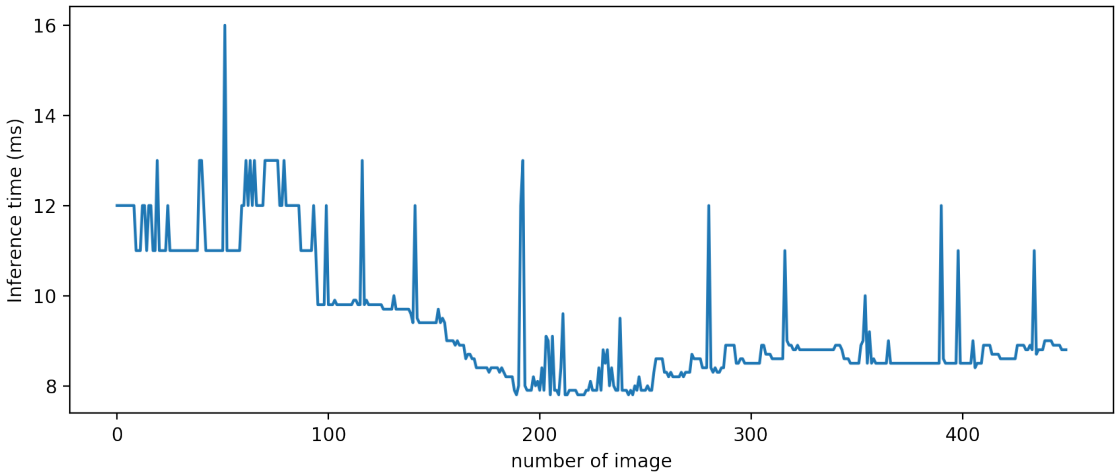

Figure 1: The inference time of YOLOv5s model, when it run on GPU

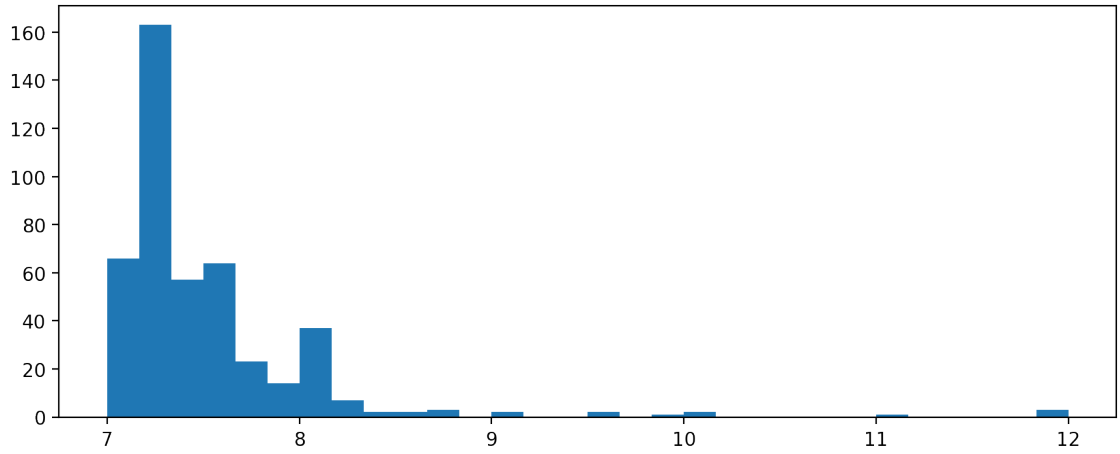

Figure 2: The histogram of inference time of YOLOv5s model, when it run on GPU

The maximum inference time is equal to 16.0(ms). The minimum inference time is 7.8 (ms) and the standard deviation is 1.42. The total time is 4.232 (s).

Half precision

The inference time of YOLOv5s model with half precision on 470 images in test set of Kaggle data set is recorded and have been shown in Fig. 3 and Fig. 4.

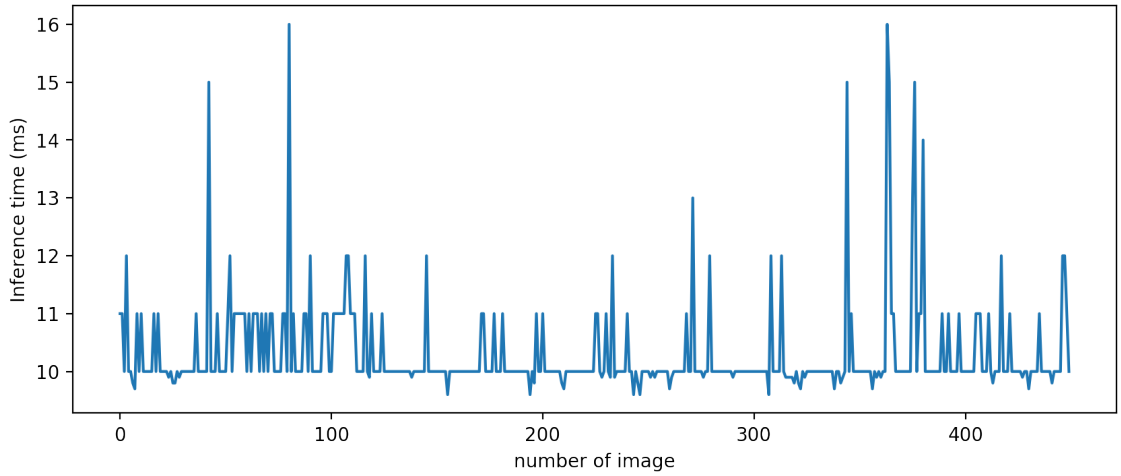

Figure 3: The inference time of YOLOv5s model, when it run on GPU with half precision

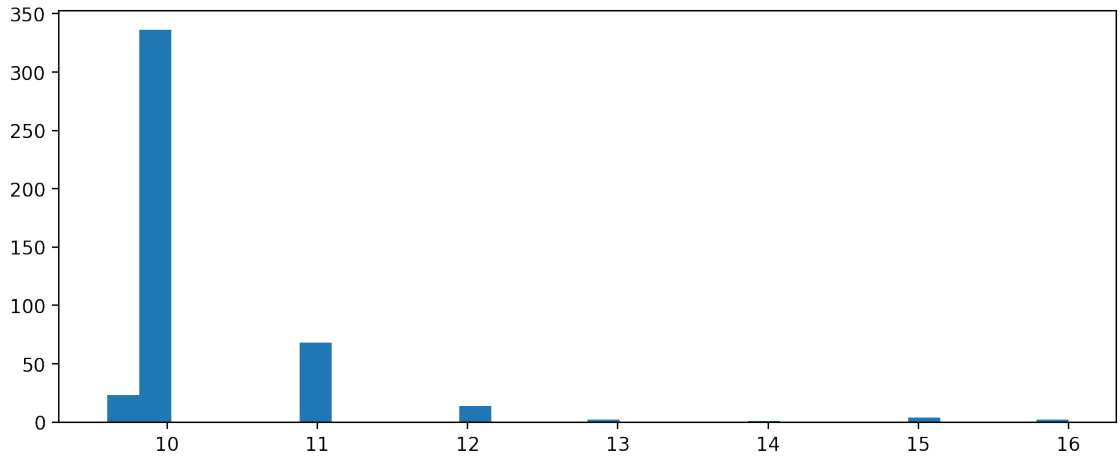

Figure 4: The histogram of inference time of YOLOv5s model, when it run on GPU with half precision

The maximum inference time is equal to 16.0 (ms). The minimum inference time is 9.6 (ms) and the standard deviation is 0.81. The total time is 4.629 (s).

Inference time of CPU

Here you can find the inference time of Yolov5s model on 470 images of kaggle data set, which selected as a test set in Fig. 5. Fig. 6 shows the histogram of these time.

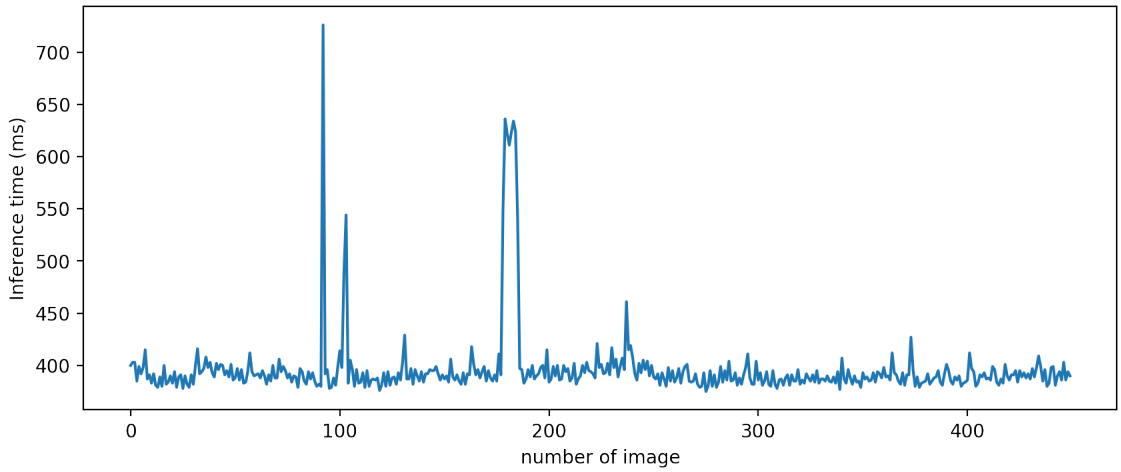

Figure 5: The inference time of YOLOv5s model, when it run on CPU

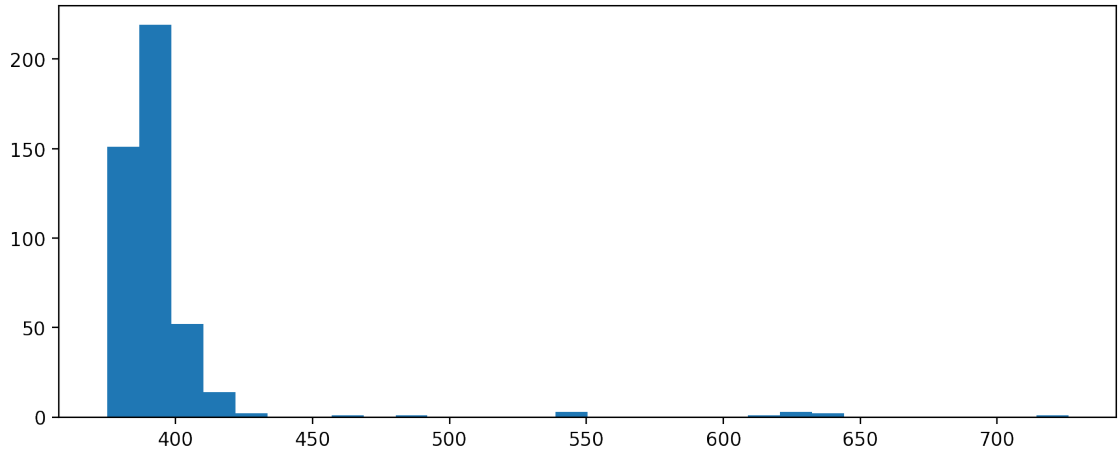

Figure 6: The histogram of inference time of YOLOv5s model, when it run on CPU

The maximum inference time is equal to 726.0 (ms). The minimum inference time is 375.0 (ms) and the standard deviation is 34.82. The total time is 178.255 (s).

YOLOv5 Medium

Inference time of GPU

Here you can find the inference time of Yolov5m model on 470 images of kaggle data set, which selected as a test set in Fig. 7. Fig. 8 shows the histogram of these time.

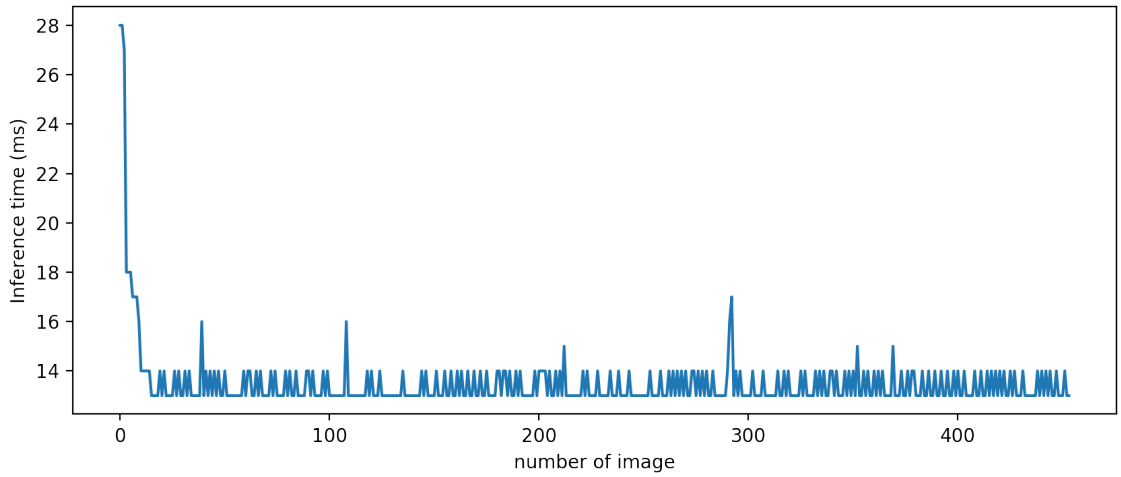

Figure 7: The inference time of YOLOv5m model, when it run on GPU

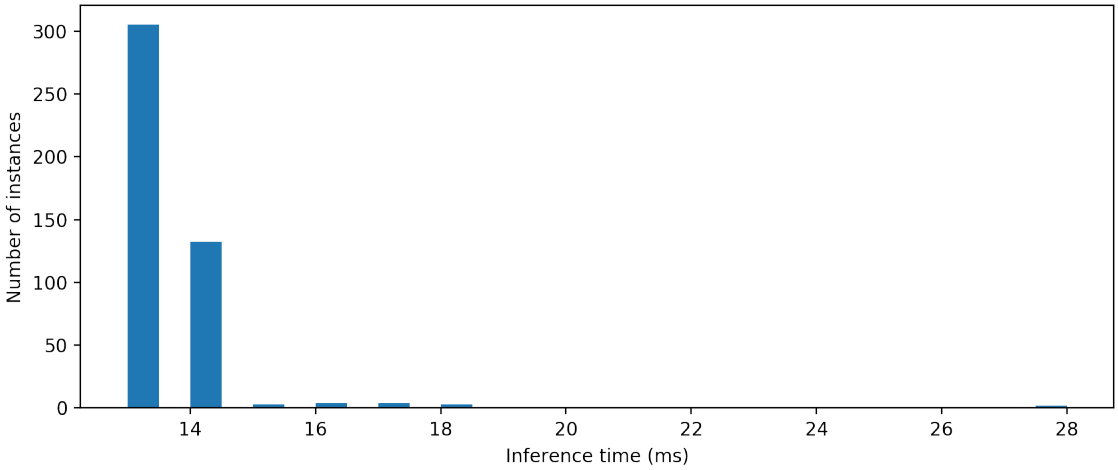

Figure 8: The histogram of inference time of YOLOv5m model, when it run on GPU

The maximum inference time is equal to 28.0 (ms). The minimum inference time is 13.0 (ms) and the standard deviation is 1.37. The total time is 6.127 (s).

Half precision

The inference time of YOLOv5m model with half precision on 470 images in test set of Kaggle data set is recorded and have been shown in Fig. 9 and Fig. 10.

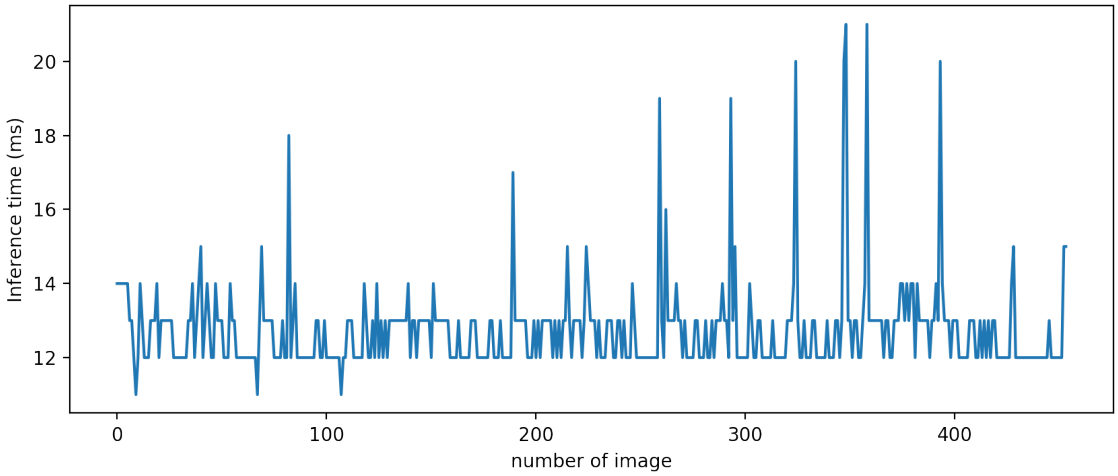

Figure 9: The inference time of YOLOv5m model, when it run on GPU with half precision

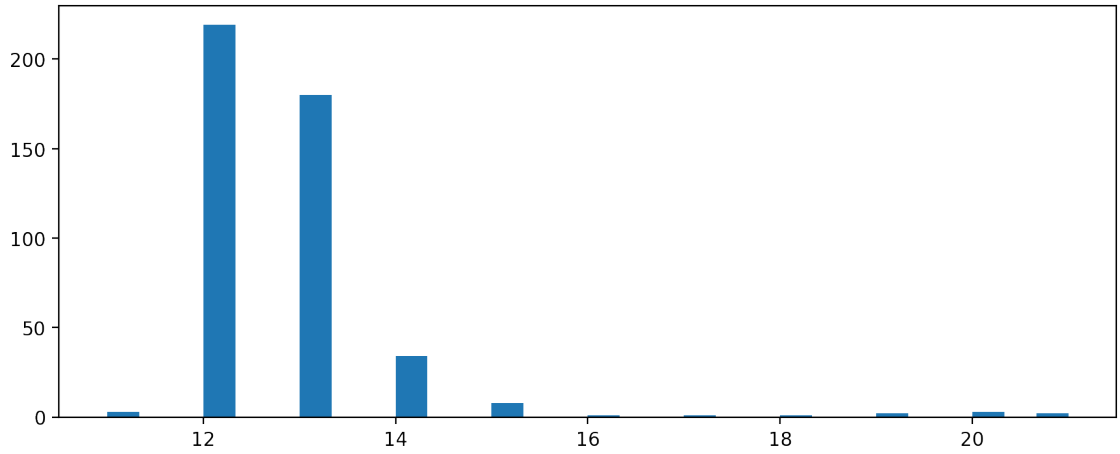

Figure 10: The histogram of inference time of YOLOv5m model, when it run on GPU with half precision

The maximum inference time is equal to 21.0 (ms). The minimum inference time is 11.0 (ms) and the standard deviation is 1.21. The total time is 5.788 (s).

Inference time of CPU

Here you can find the inference time of Yolov5s model on 470 images of kaggle data set, which selected as a test set in Fig. 11. Fig. 12 shows the histogram of these time.

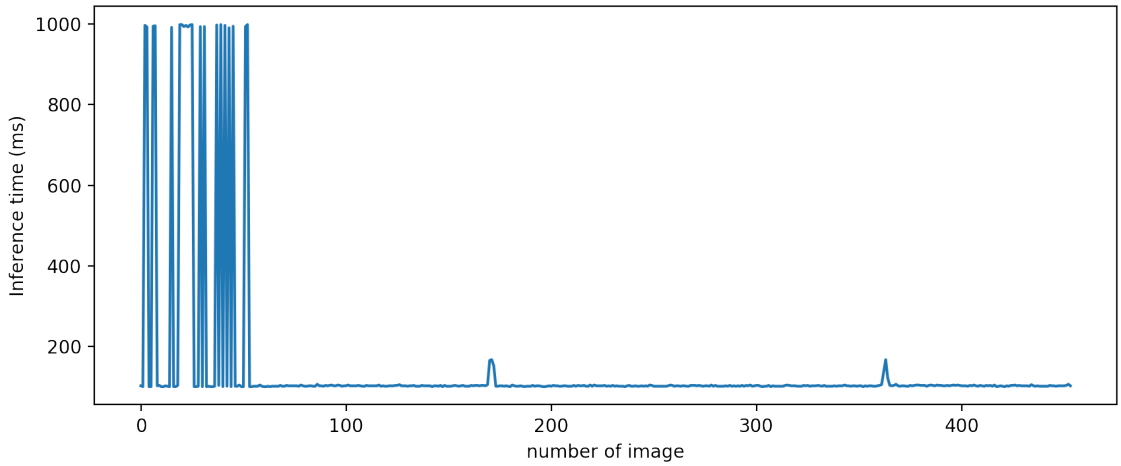

Figure 11: The inference time of YOLOv5s model, when it run on CPU

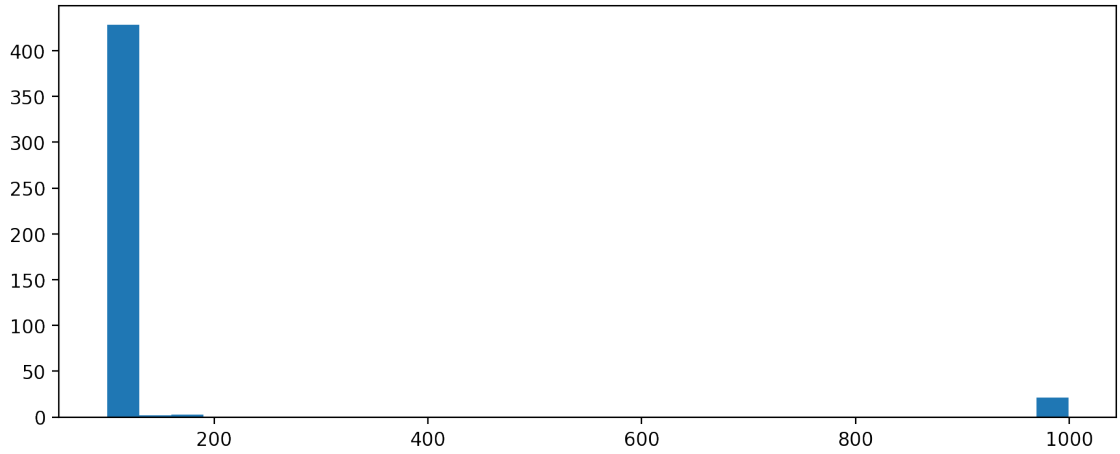

Figure 12: The histogram of inference time of YOLOv5m model, when it run on CPU

The maximum inference time is equal to 999.0 (ms). The minimum inference time is 100.0 (ms) and the standard deviation is 187.73. The total time is 65.311 (s).

YOLOv5 large

Inference time of GPU

Here you can find the inference time of Yolov5l model on 470 images of kaggle data set, which selected as a test set in Fig. 13. Fig. 14 shows the histogram of these time.

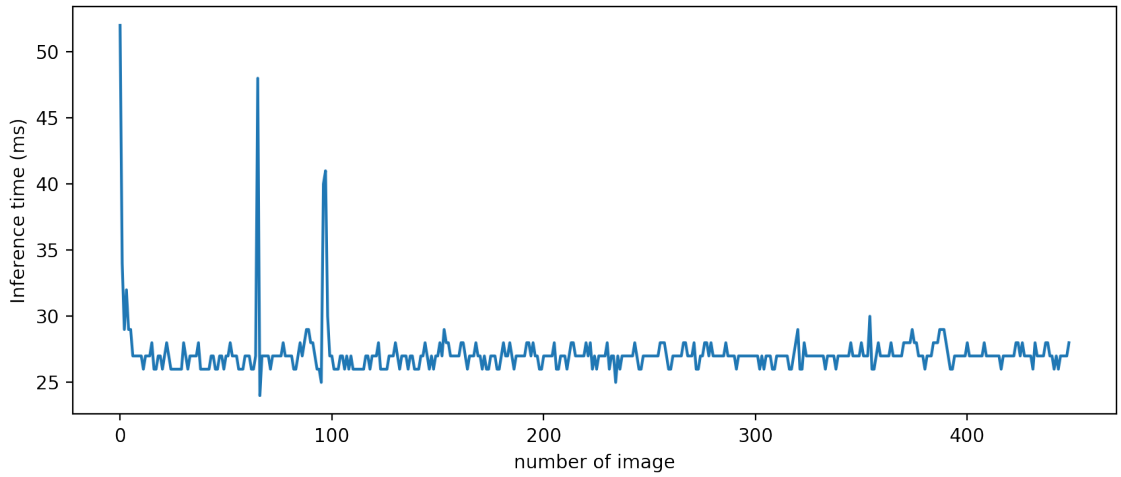

Figure 13: The inference time of YOLOv5l model, when it run on GPU

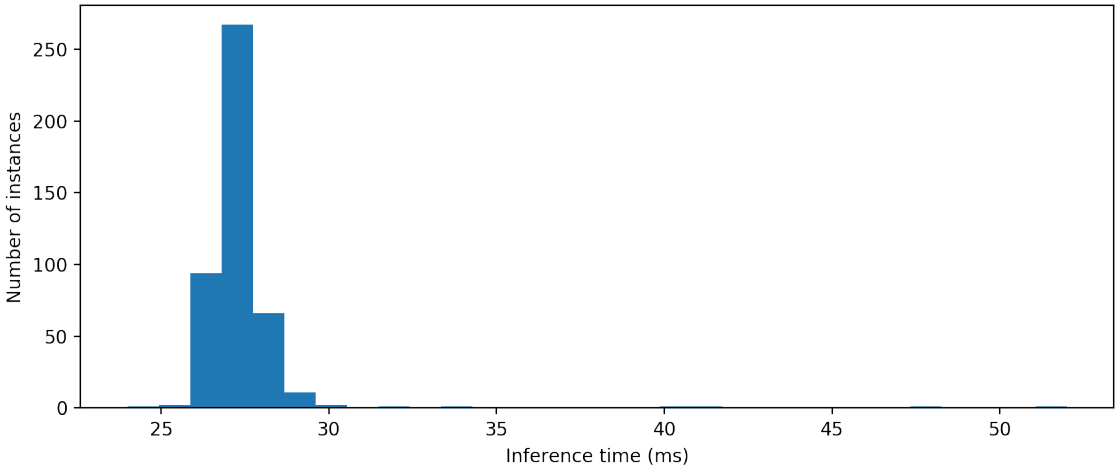

Figure 14: The histogram of inference time of YOLOv5l model, when it run on GPU

he maximum inference time is equal to 52.0 (ms). The minimum inference time is 24.0 (ms) and the standard deviation is 1.96. The total time is 12.201 (s).

Half precision

The inference time of YOLOv5l model with half precision on 470 images in test set of Kaggle data set is recorded and have been shown in Fig. 15 and Fig. 16.

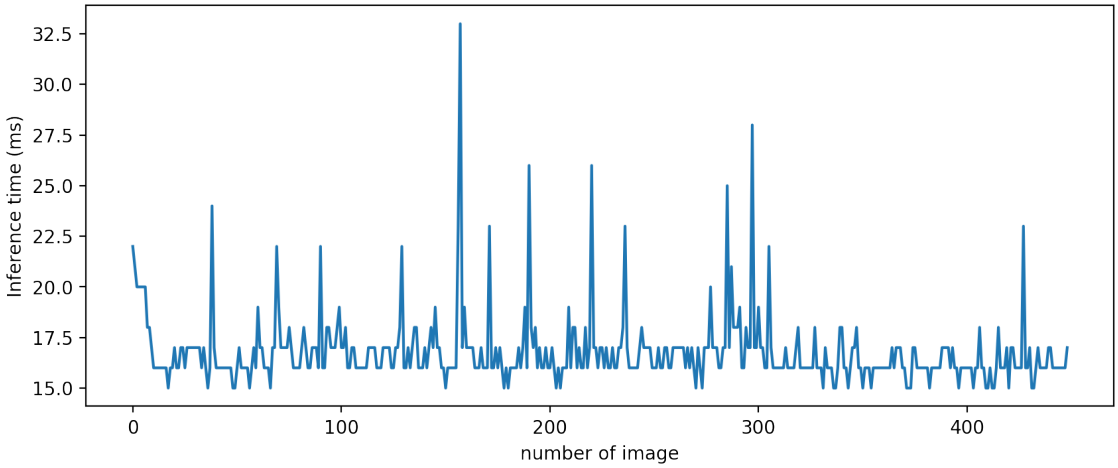

Figure 15: The inference time of YOLOv5l model, when it run on GPU with half precision

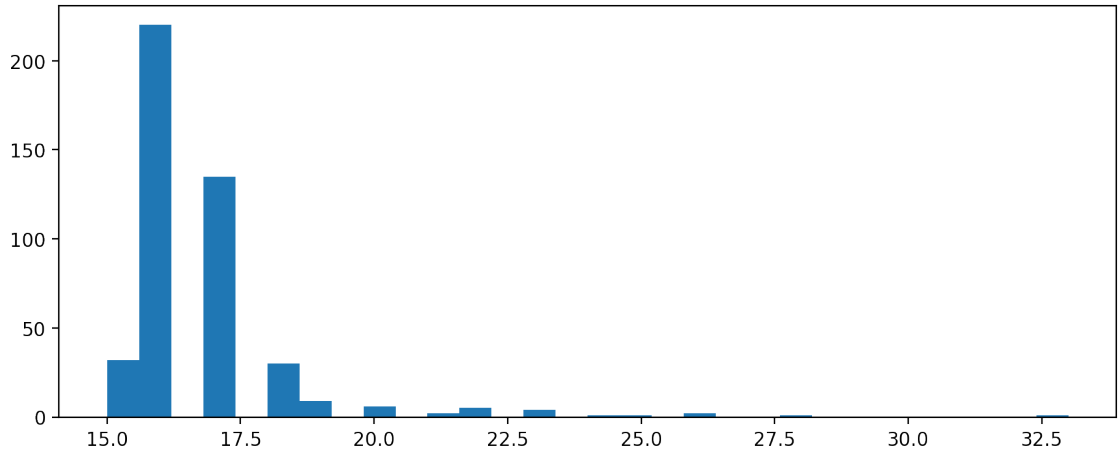

Figure 16: The histogram of inference time of YOLOv5l model, when it run on GPU with half precision

The maximum inference time is equal to 33.0 (ms). The minimum inference time is 15.0 (ms) and the standard deviation is 1.76. The total time is 7.532 (s).

Inference time of CPU

Here you can find the inference time of Yolov5l model on 470 images of kaggle data set, which selected as a test set in Fig. 17. Fig. 18 shows the histogram of these time.

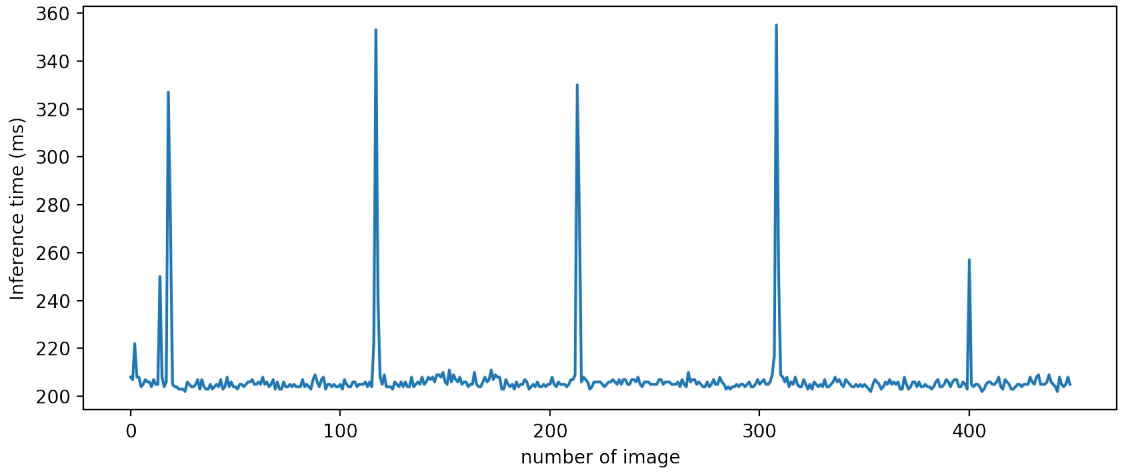

Figure 17: The inference time of YOLOv5l model, when it run on CPU

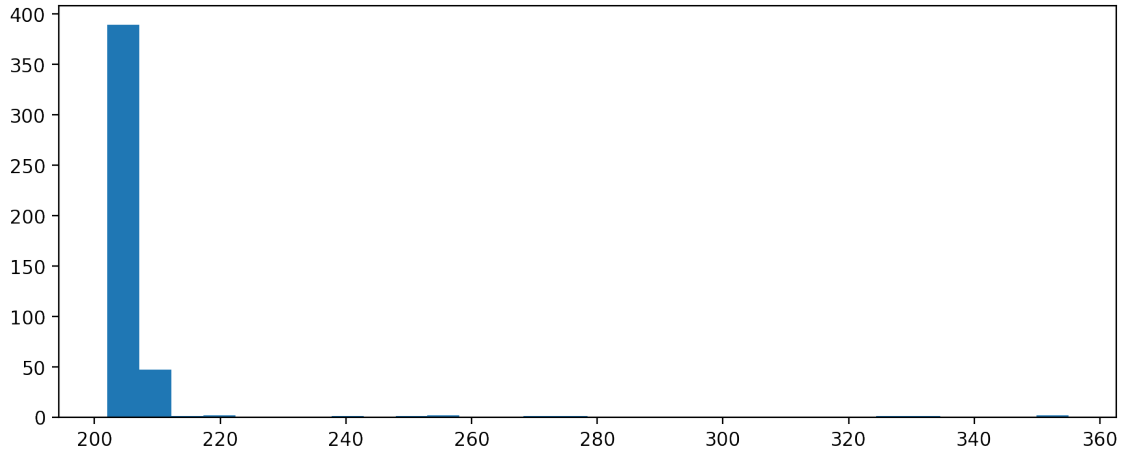

Figure 18: The histogram of inference time of YOLOv5l model, when it run on CPU

The maximum inference time is equal to 355.0 (ms). The minimum inference time is 202.0 (ms) and the standard deviation is 14.36. The total time is 93.110 (s).

Compare the results

Table 1: The inference time with different scenarios for YOLOv5s model

Name	Min (ms)	Max (ms)	Average (ms)	Std	Total time (s)
GPU	7.8	16	9.4	1.42	4.23
GPU-half	9.6	16	10.29	0.81	4.63
CPU	375	726	396.12	34.82	178.255

Table 2: The inference time with different scenarios for YOLOv5m model

Name	Min (ms)	Max (ms)	Average (ms)	Std	Total time (s)
GPU	13	28	13.49	1.37	6.127
GPU-half	11	21	12.75	1.21	5.788
CPU	100	999	143.85	187.73	65.311

Table 3: The inference time with different scenarios for YOLOv5l model

Name	Min (ms)	Max (ms)	Average (ms)	Std	Total time (s)
GPU	24	52	27.17	1.96	12.201
GPU-half	15	33	16.77	1.76	7.532
CPU	202	355	207.37	14.36	93.11