Inference time of YOLOv5

Team C

Sara Mohajerani

October 11, 2022

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.

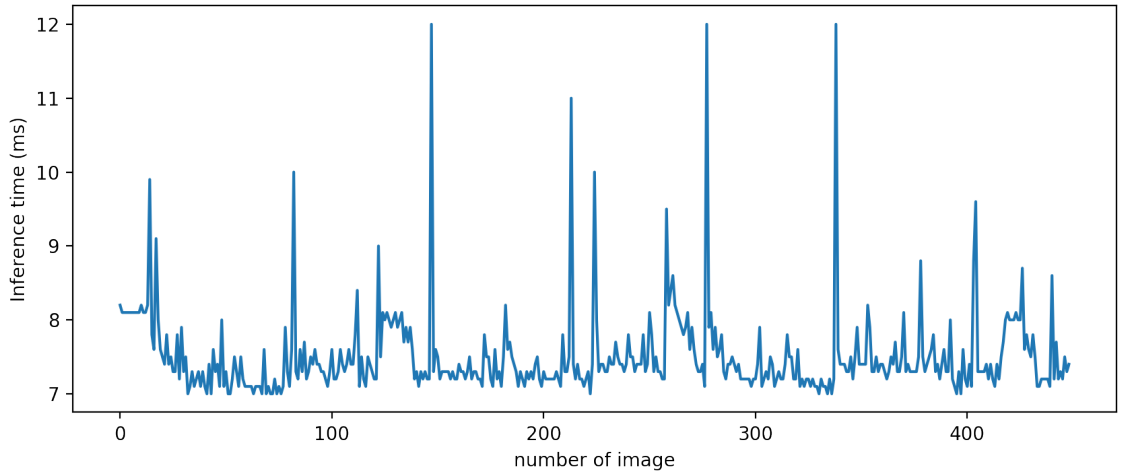

Figure 1: The inference time of YOLOv5s model, when it run on GPU

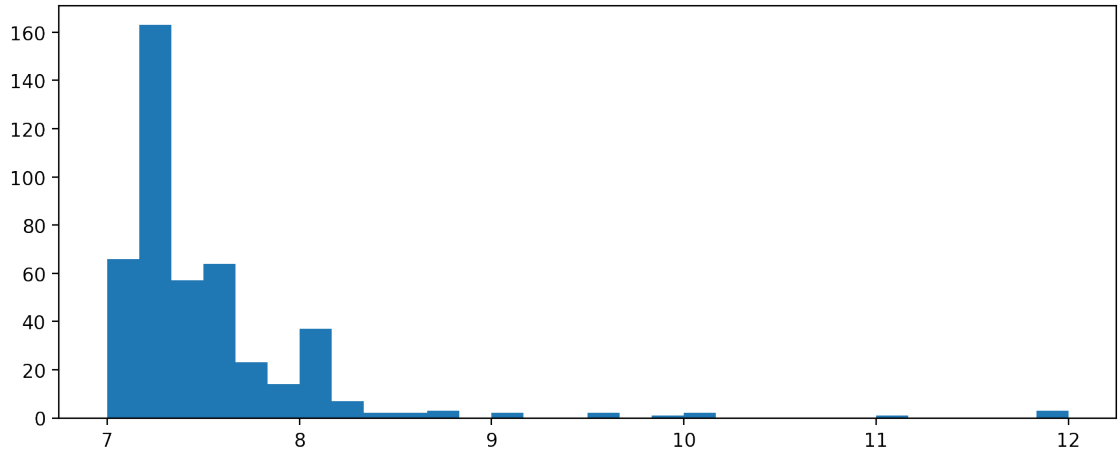

Figure 2: The histogram of inference time of YOLOv5s model, when it run on GPU

GPU pro

Here the GPU model converted to the premium model. The inference time of YOLOv5s model on 470 images in test set of Kaggle data set is recorded and have been shown in Fig. 3 and Fig. 4.

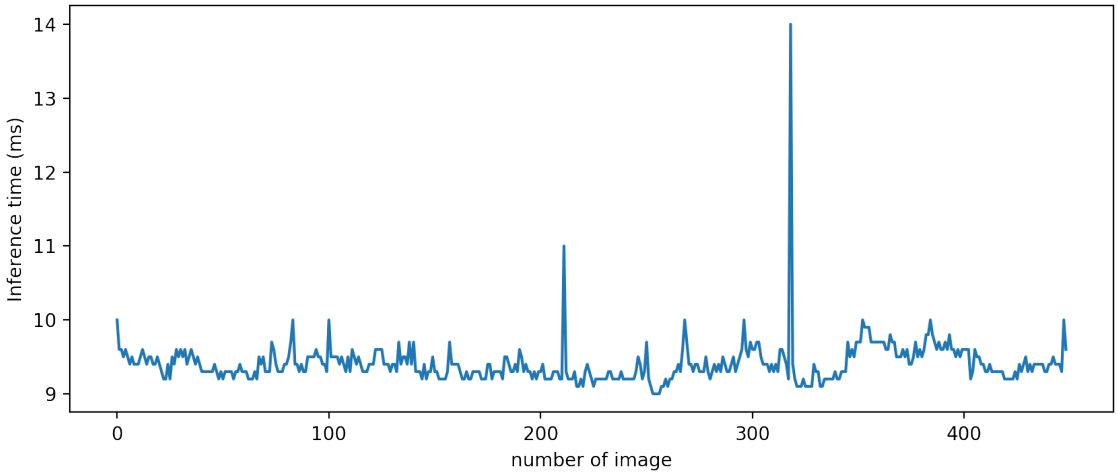

Figure 3: The inference time of YOLOv5s model, when it run on premium GPU

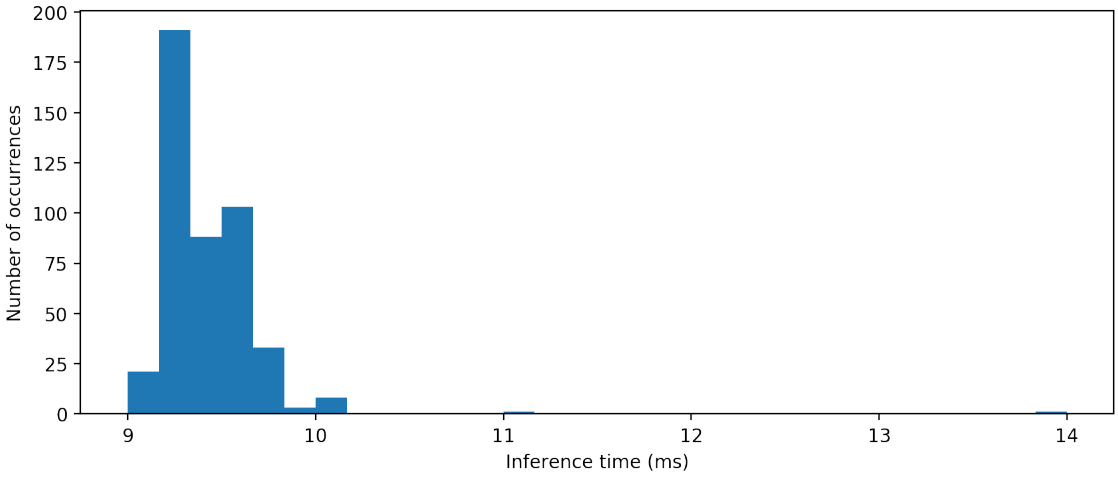

Figure 4: The histogram of inference time of YOLOv5s model, when it run on premium GPU

The maximum inference time is equal to 14.0 ms. The minimum inference time is 9.0 ms and the standard deviation is 0.30. The total time is 4.224 s. In the second run the maximum inference time is equal to 20.0 ms. The minimum inference time is 8.8 ms and the standard deviation is 0.72. The total time is 4.149 s.

Half precision

Here the GPU model converted to the premium model. 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. 5 and Fig. 6.

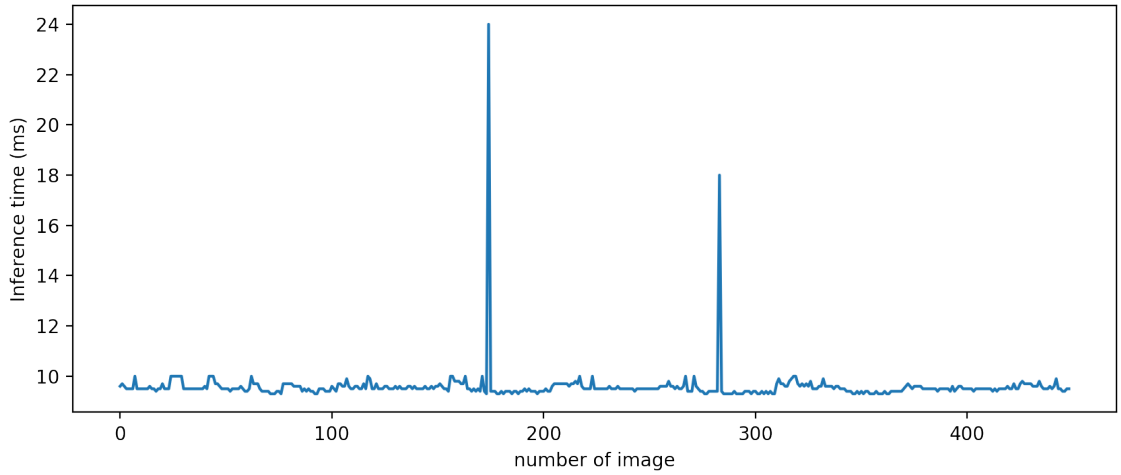

Figure 5: The inference time of YOLOv5s model, when it run on premium GPU with half precision

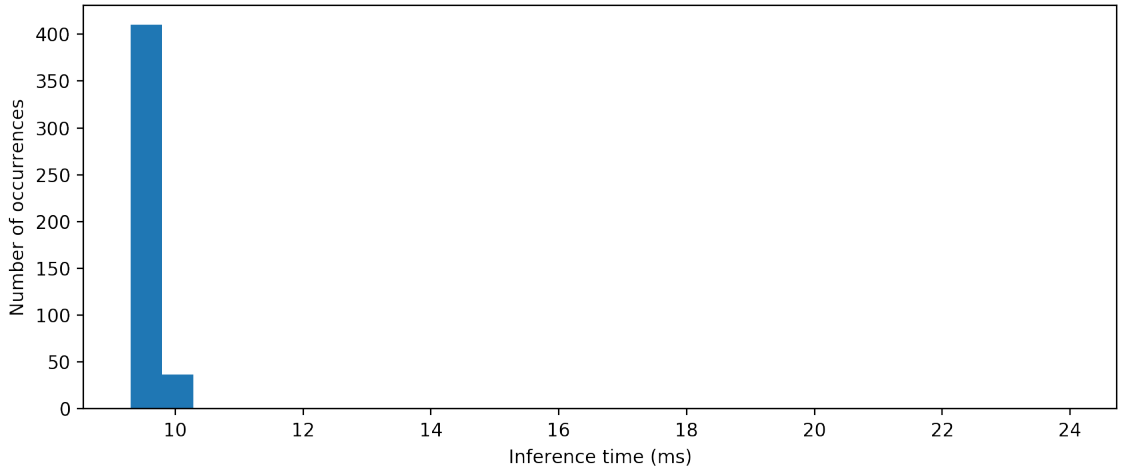

Figure 6: The histogram of inference time of YOLOv5s model, when it run on premium GPU with half precision

Here the inference time of Yolov5s model on 470 images of kaggle data set, with half precision is shown in Fig. 7. Fig. 8 shows the histogram of these recorded time.

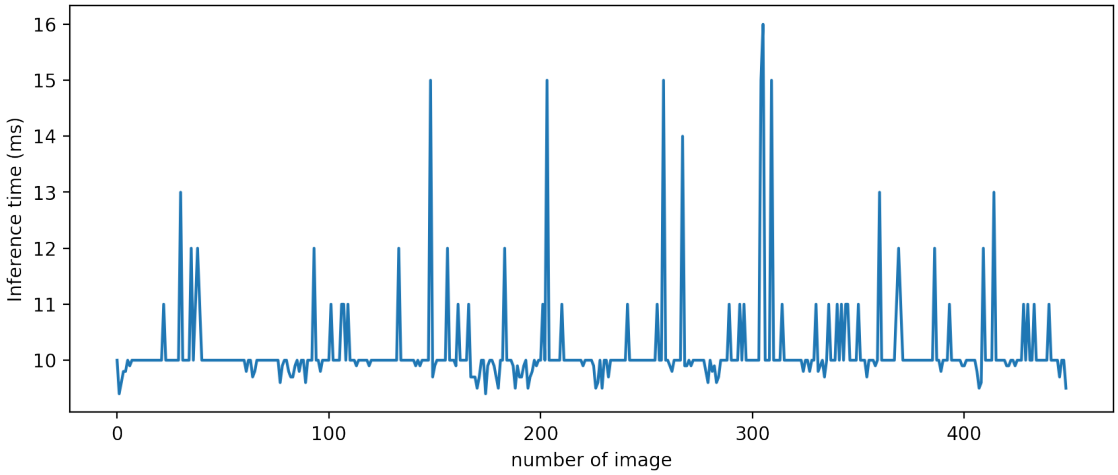

Figure 7: The inference time of YOLOv5s model with half precision, when it run on GPU

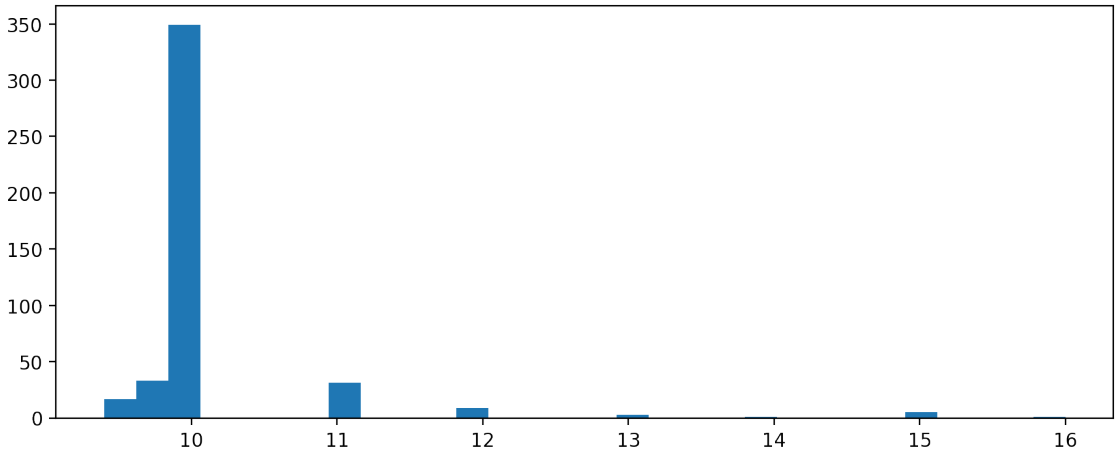

Figure 8: The histogram of inference time of YOLOv5s model with half precision, when it run on GPU

GPU pro

The maximum inference time is equal to 24.0 ms. The minimum inference time is 9.3 ms and the standard deviation is 0.81. The total time is 4.302 s. In the second run, we achieve the maximum inference time equal to 19.0 ms. The minimum inference time is 9.3 ms and the standard deviation is 0.59. The total time is 4.363 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. 9. Fig. 10 shows the histogram of these time.

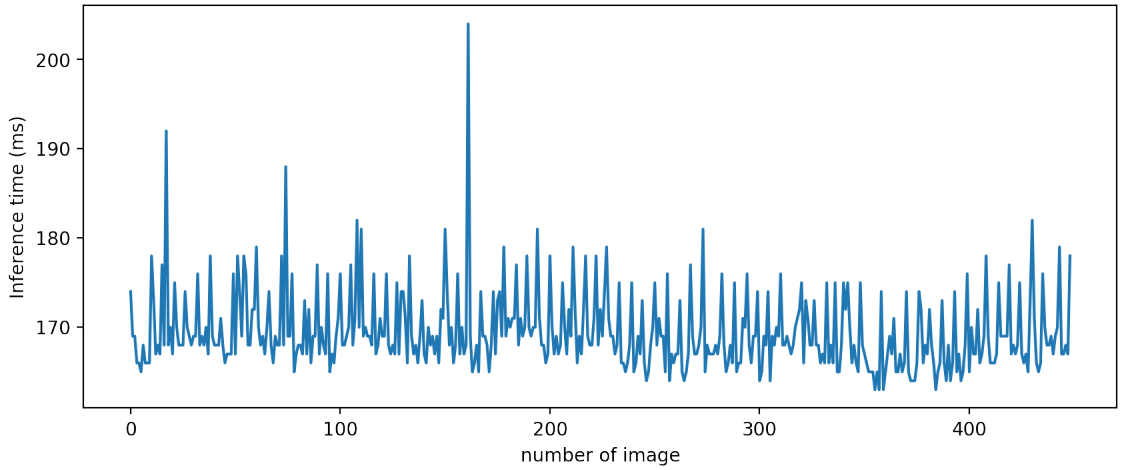

Figure 9: The inference time of YOLOv5s model, when it run on CPU

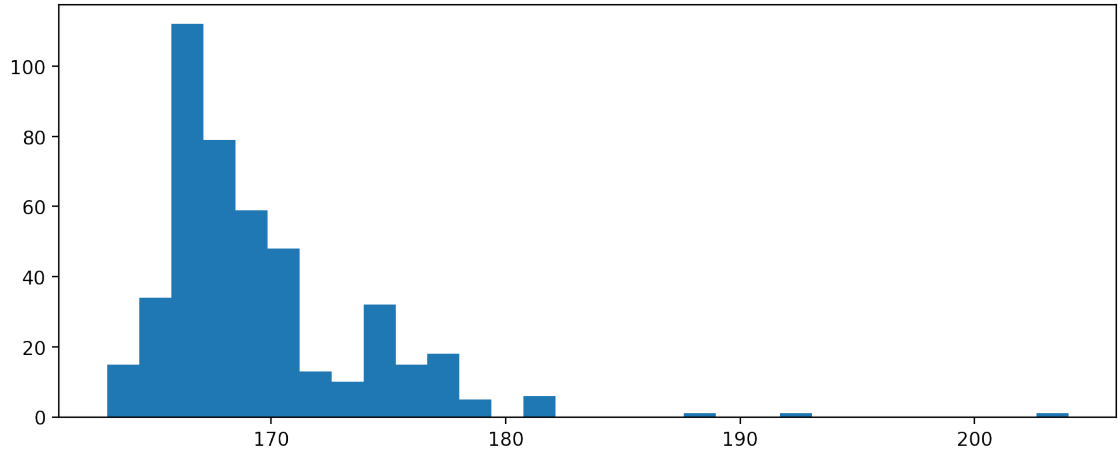

Figure 10: The histogram of inference time of YOLOv5s model, when it run on CPU

Compare the results

Table 1: The inference time with different scenarios

Name	Min (ms)	Max (ms)	Std	Total time (s)
GPU	7	12	0.59	3.372
GPU-half	9.4	16	0.77	4.563
GPU(Pro)	9	14	0.30	4.224
GPU-half (Pro)	9.3	24	0.81	4.302
CPU	163	204	4.38	76.096