2022-06-16
Olof Lennartsson, Elias Sonnsjö
Infotiv

ML Verification Results [Z]

Introduction

The testing of the ML model is based on assessing the object detection accuracy for the sequestered verification data set. These results provide evidence that the ML model satisfies the ML safety requirements on the verification data. A fundamental aspect of the verification argument is that this data set was never used in any way during the development of the ML model.

The ML model test cases provide results for both 1) the entire verification data set and 2) eight slices of the data set that are deemed particularly important. The selection of slices was motivated by either an analysis of the available technology or ethical considerations, especially from the perspective of AI fairness. Consequently, we measure the performance for the following slices of data. Identifiers in parentheses show direct connections to requirements.

S1 The entire verification data set

S2 Pedestrians close to the ego car (longitudinal distance < 50 m)(SYS-PER-REQ1, SYS-PER-REQ2)

S3 Pedestrians far from the ego car (longitudinal distance ≥ 50 m)

S4 Running pedestrians (speed ≥ 3 m/s) (SYS-ROB-REQ2)

S5 Walking pedestrians (speed > 0 m/s but < 3 m/s) (SYS-ROB-REQ2)

S6 Occluded pedestrians (entering or leaving the field of view, defined as bounding box in contact with any edge of image) (DAT-COM-REQ4)

S7 Male pedestrians (DAT-COM-REQ2)

S8 Female pedestrians (DAT-COM-REQ2)

S9 Children (DAT-COM-REQ2)

For the ML model in SMIRK, we evaluate pedestrian detection at IoU = 0.5, which for each image means:

TP True positive: IoU ≥ 0.5

FP False positive: IoU < 0.5

FN False negative: There is a ground truth bounding box in the image, but no predicted bounding box.

TN True negative: All parts of the image with neither a ground truth nor a predicted bounding box. This output carries no meaning in our case.

Results

The total number of images in the verification data is 208,884 (202,712 pedestrians (97.0%) and 6,172 non-pedestrians (3.0%)). Figure 1 depicts four subplots representing IoU = 0.5: A) P vs R, B) F1 vs. Conf, C) P vs. Conf, and D) R vs. Conf. We observe that the appearance of the four subplots closely resembles the corresponding plots for the internal test data (see the Internal ML Model Test Protocol 2022-06.16.pdf).

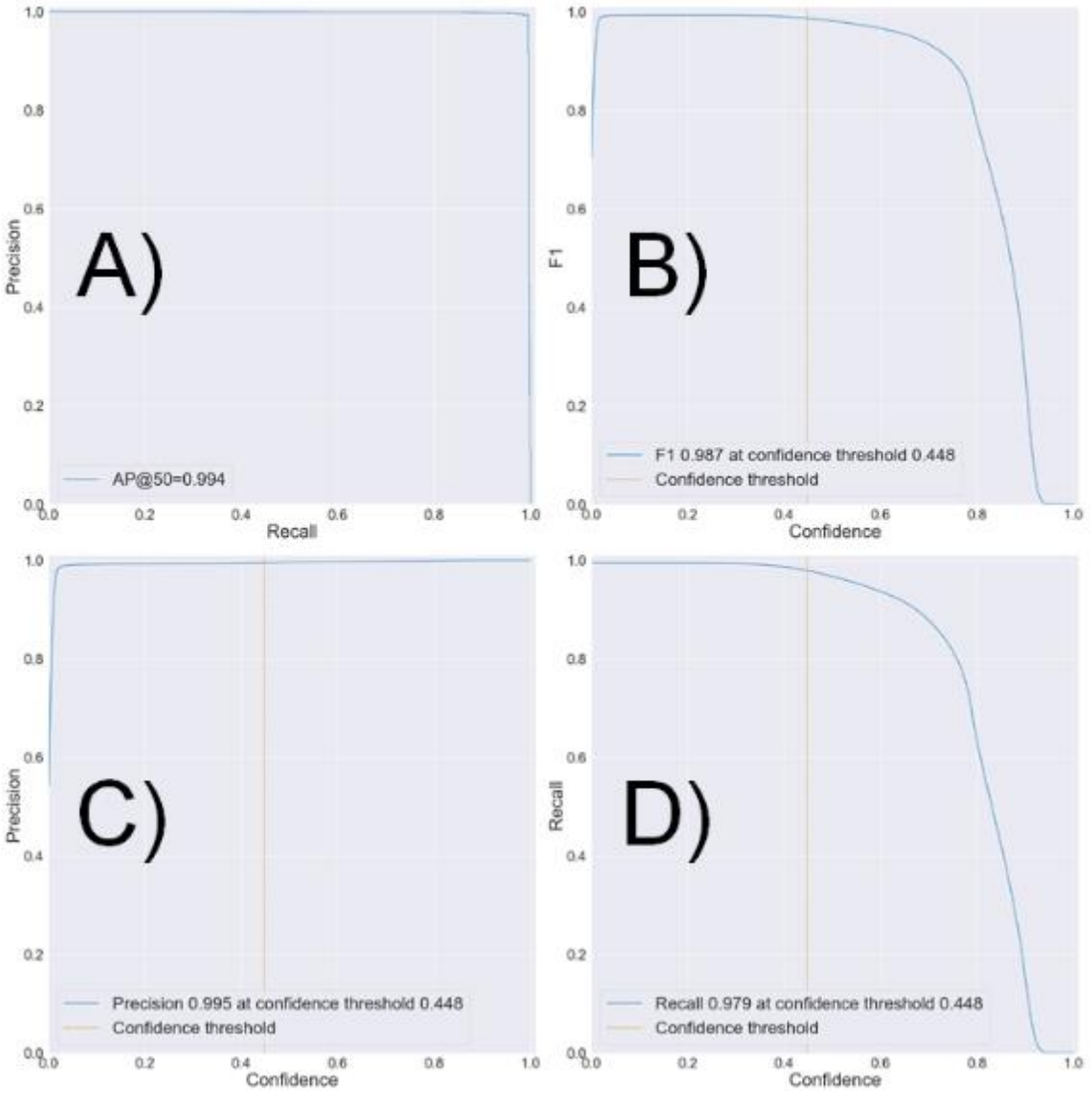

Figure 1: Evaluation of the ML model on the verification data at IoU=0.5.

Table 1 shows the output from the ML model using the Conf threshold 0.448 on the verification data. The table is organized into 1) all distances from the ego car, 2) within 80 m, and 3) within 50 m, respectively. The table also shows the effect of adding OOD detection using the autoencoder, i.e., the number of FPs is decreased just as for the internal test data.

Table 1: ML model accuracy on the verification data at the Conf threshold 0.448.

Distance	Total	TP	FP	FN	P	R	F1	AP@0.5
All	208,884	198,457	990	4,255	0.9950	0.9790	0.9870	0.9942
+OOD		195,695	533	7,017	0.9973	0.9654	0.9811	0.9878
$\leq$ 80 m	158,066	151,976	330	210	0.9978	0.9986	0.9982	0.9945
+OOD		149,214	23	2,972	0.9998	0.9905	0.9901	0.988
$\leq$ 50 m	92,592	86,847	178	165	0.9980	0.9981	0.9980	0.9949
+OOD		84,085	21	2,972	0.9998	0.9805	0.9901	0.988

Table 2 demonstrates how the ML model satisfies the performance requirements on the verification data. The FPPI (0.21%) is too high to satisfy SYS-PER-REQ3 without additional OOD detection, i.e., we observed 330 FPs (roughly an equal share of pyramids and children). The rightmost column in Table 2 shows how the FPPI decreased to 0.015% with the autoencoder. All basic shapes were rejected, instead children at a long distance with too low IoU scores dominate the FPs. We acknowledge that it is hard for the YOLOv5 to achieve a high IoU for the few pixels representing a child almost 80 m away. However, commencing emergency braking in such cases is an appropriate action – a child detected with a low IoU is not an example of the ghost braking hazard. SYS-PER-REQ4 is satisfied as the fraction of rolling windows with more than a single FN is 2.3%. Figure 2 shows the distribution of position errors. The median error is 1.0 cm, the 99% percentile is 5.4 cm, and the largest observed error is 12.8 cm. Consequently, we show that SYS-PER-REQ5 is satisfied for the verification data.

Table 2: ML model satisfaction of the performance requirements on the verification data at the Conf threshold 0.448. R1–R4 = SYS-PER-REQ1–4.

Req.	Expected	Observed (Model)	Observed (Model+OOD)
R1	TP rate $\geq$ 93% for $\leq$ 80 m	$\frac{151,976}{158,066} = 96.1\%$	$\frac{149,214}{158,066} = 94.4\%$
R2	FN rate $\leq$ 7% for $\leq$ 50 m	$\frac{165}{92,592} = 0.18\%$	$\frac{2,927}{92,592} = 3.2\%$
R3	FPPI $\leq$ 0.1% for $\leq$ 80 m	$\frac{330}{158,066} = 0.21\%$	$\frac{23}{158,066} = 0.015\%$
R4	$\leq$ 3% of rolling windows contain $\geq$ 2 misses in 5 frames for $\leq$ 80 m	$\frac{201}{152,395} = 0.13\%$	$\frac{3,499}{152,090} = 2.3\%$

SYS-PER-REQ4 is met as the fraction of rolling windows with more than a single FN is 0.24%, i.e., ≤ 3%. Figure 2 shows the distribution of position errors in the object detection for pedestrians within 80 m of ego car, i.e., the difference between the object detection position and ESI Pro-SiVIC ground truth. The median error is 1.0 cm, the 99% percentile is 5.6 cm, and the largest observed error is 12.7 cm. Thus, we show that SYS-PER- REQ5 is satisfied for the internal test data, i.e., ≤ 50 cm position error for pedestrian detection within 80 m. Note that satisfaction of SYS-PER-REQ6, i.e., sufficient inference speed, is demonstrated as part of the system testing.

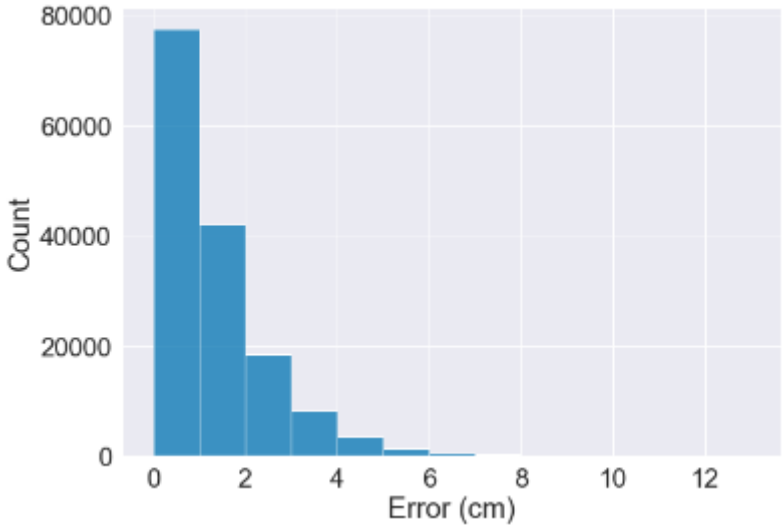

Figure 2: Distribution of position errors for the verification data.

Table 3 presents the output of the ML model on the nine slices of the verification data defined in Section 9.1. In relation to the robustness requirements, we notice that there the accuracy is slightly lower for S9 (children). This finding is related to the size requirement SYS-ROB-REQ3. Table 10 contains an in-depth analysis of children at different distances with OOD detection. We confirm that most FPs occur outside of the ODD, i.e., 507 out of 512 FPs occur for children more than 80 m from go car. In extension, we find that the performance requirements are still satisfied for the most troublesome slice of data as follows:

- TP rate children ≤ 80m: $_{50,402/\,50,696}$ = 99.4%
- FN rate children ≤ 50m: $_{249/30,731}$ = 0.81%
- FPPI children ≤ 80m: $_{5/52,463}$ = 0.0099%

Table 3: ML model accuracy on eight slices of the verification data. Every second rows show results for the ML model followed by OOD detection using the autoencoder.

Slice	Total	TP	FP	FN	P	R	F1	AP@0.5
S1	208,884	198,457	990	4,255	0.995	0.979	0.987	0.9942
	+OOD	195,695	533	7,017	0.9998	0.9616	0.9803	0.9878
S2	92,028	86,691	22	165	0.9997	0.9981	0.9989	0.995
	+OOD	83,929	21	2,927	0.9997	0.9663	0.9827	0.9761
S3	65,330	65,285	2	45	1	0.9993	0.9996	0.995
	+OOD	65,285	2	45	1	0.9993	0.9996	0.995
S4	58,267	56,130	110	716	0.998	0.9874	0.9927	0.9949
	+OOD	54,964	110	1,882	0.998	0.9669	0.9822	0.9818
S5	149,617	142,328	424	3,538	0.997	0.9757	0.9863	0.9949
	+OOD	140,732	423	5,134	0.997	0.9648	0.9806	0.9882
S6	1,031	866	22	165	0.9752	0.84	0.9026	0.9289
	+OOD	805	21	226	0.9746	0.7808	0.867	0.8783
S7	69,292	67,555	15	54	0.9998	0.9992	0.9995	0.995
	+OOD	65,009	14	2,600	0.9998	0.9616	0.9803	0.9741
S8	69,291	67,495	7	74	0.9999	0.9989	0.9994	0.995
	+OOD	67,482	7	87	0.9999	0.9987	0.9993	0.995
S9	69,301	63,408	512	4,126	0.992	0.9389	0.9647	0.9879
	+OOD	63,205	512	4,329	0.992	0.9359	0.9631	0.9871