Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 2458-9403
Vol. 6 Issue 7, July - 2019

Evaluation of COCO Validation 2017 Dataset with YOLOv3

Dae-Hwan Kim
Department of Electronics,
Suwon Science College, 288 Seja-ro, Jeongnam-myun,
Hwaseong-si, Gyeonggi-do, Rep. of Korea
kimdh@ssc.ac.kr

***Abstract*— Nowadays, deep learning is widely used for various fields such as computer vision, finance, medicine, and agriculture. Object detection and localization in an image is one of the main problems in computer vision, and YOLO is a widely adopted deep learning framework for this analysis. In this paper, the popular COCO dataset is evaluated on YOLOv3. Detection and localization performance is presented in detail for each class and the annotation of the dataset is discussed.**

Keywords—Deep learning; YOLO; COCO dataset; Validation; Evaluation; Detection; Localization

I. INTRODUCTION

Nowadays, deep learning is widely used for various research and commercial fields such as computer vision, finance, natural language processing, medicine, and agriculture. In deep learning, a convolutional neural network (CNN) is commonly adopted for compute vision problems, and SSD [1-2], R-CNN [3-4], and YOLO [5-7] are most widely used algorithms for object detection and localization. Among these, YOLO is far faster than other algorithms.

To evaluate the performance of algorithm, various datasets have been presented [8-10], and among these, COCO dataset [8] is popular because it is large-scale, and contains natural scenes. In this paper, validation set 2017 of COCO dataset is evaluated in detail in the most recent YOLOv3 [7] framework, and COCO annotation is discussed.

The rest of this paper is organized as follows. Section II shows the overview of the COCO dataset and YOLO framework. Section III evaluates COCO dataset with YOLOv3. Conclusions are presented in Section IV.

II. BACKGROUND

A. COCO Dataset

COCO is a large-scale object detection, segmentation, and captioning dataset developed by Microsoft Inc. COCO stands for Common Objects in Context. The images in the dataset are from natural context, which contains common objects in everyday scenes. The dataset contains 80 labeled object categories where the objects are labeled for the precise object localization.

The dataset were initially released in 2014, and then in 2017. The 2017 version includes 164K images including 118K for train, 5K for validation, and 20K for test-dev. To evaluate the detection performance, I use 5K validation data because the testset does not provide label annotation.

Occasionally the number of instances in an image is quite high. Such an example is a dense crowd of people. In these cases, many instances are likely to be tightly grouped together and it is quite difficult to distinguish individual instances. Thus, in the COCO dataset, after 10-15 instances are segmented, the remaining ones are marked as crowed and segmented as one. In this paper, the analysis of crowed instances is not performed due to the lack of the precise label information.

B. YOLO

YOLO stands for You Only Look Once, which is a state-of-the-art, real time object detection framework. It is a unified solution, and uses a single convolutional network, which can simultaneously predict both class probabilities and bounding boxes. The performance of YOLOv3 is 30 FPS and a mAP of 57.9% on COCO test-dev on a Pascal Titan X processor [7].

The algorithm applies the CNN to an entire image. YOLOv3 divides the image into the 19x19 grid cells, and finds the bonding boxes while predicting probabilities for each of these regions.

III. EVALUATION

TABLE I. NUMBER OF IMAGES AND OBJECTS

# of images	# of crowed images	# of non-crowed images	# of objects in non-crowed images
5,000	411	4,589	27,436

The COCO pre-trained weights are used for the YOLO network [11], and evaluation is performed for the COCO validation 2017 dataset. Throughout the paper, AP (Average Precision) is not measured, and total true positive, false positive detection ratios are calculated just for simplicity. Table I shows the

numbers of images, crowed images, non-crowed images, and objects, respectively. The number of images is 5,000 where that of crowed images is 411. There are 27,436 objects in non-crowed images.

Table II describes the detection classification used in this paper. They are true positive, false positive, false negative, and true negative. Considering the object localization together, the detection is true positive when the detection result is correct and IOU (intersection over union) is greater than or equal to the threshold between the ground truth and detection bounding box. When the detection result is wrong or IOU is less than the threshold, it is treated as false positive. False negative indicates when no detection is found for a ground. True negative is not used in this paper.

TABLE II. CLASSIFICATION USED IN THIS PAPER

True positive	The object detection is correct, and IOU is greater than or equal to the threshold.
False positive	The detection result is wrong or IOU is less than the threshold.
False negative	It is the case when no detection is found for a ground truth.
True negative	It is not used in this paper.

Table III shows the detection result by several IOU threshold values where true positive and false positive are denoted by TP and FP, respectively. In the experiment, IOU threshold varies from 0.25, 0.5, and 0.75 while the object confidence threshold is 0.1. As expected, the high IOU threshold decreases TP while increasing FP. When IOU threshold is low, we obtain the contrary result.

TABLE III. IOU THRESHOLD AND DETECTION RATIO

IOU Threshold	# of GT objects	# of TP (ratio)	# of FP (ratio)
0.25	27,436	21,478 (78.3%)	12,237 (44.6%)
0.5		19,965 (72.8%)	13,750 (50.1%)
0.75		13,307 (48.5%)	20,408 (74.4%)

Table IV shows the result by the confidence threshold. The threshold varies from 0.1, 0.2 to 0.3. As expected, when the threshold value is high, less detections are found, and thus, both the true positive and the false positive ratios are decreased. When the threshold is low, we obtain the contrary result.

TABLE IV. OBJECT CONFIDENCE THRESHOLD AND DETECTION RATIO

Object confidence threshold	# of GT objects	# of TP (ratio)	# of FP (ratio)
0.1	27,436	19,965 (72.8%)	13,750 (50.1%)
0.2		18,574 (67.7%)	7,347 (26.8%)
0.3		17.475 (63.7%)	4,646 (16.9%)

Table V shows the object detection ratio for each class where the total numbers of GT and TP are summarized for all images, respectively. In the experiment, the IOU threshold and object confidence threshold are set 0.5, and 0.1, respectively. The cat class shows the best detection ratio, and the top five most detected classes are cat, dog, bear, bus, and train. The bottom five classes are knife, backpack, handbag, toaster, and hair drier.

TABLE V. OBJECT CLASS AND TRUE POSITIVE RATIO

Class	*# of TP*	*# of GT*	*ratio*	*Class*	*# of TP*	*# of GT*	*ratio*
cat	184	197	93.4	dog	186	205	90.7
bear	64	71	90.1	bus	193	217	88.9
train	162	183	88.5	horse	205	234	87.6
frisbee	94	109	86.2	airplane	116	135	85.9
fire hydrant	85	99	85.9	microwave	46	54	85.2
mouse	86	101	85.1	tennis racket	153	181	84.5
tv	208	248	83.9	laptop	182	218	83.5
person	6130	7402	82.8	giraffe	190	230	82.6
toilet	146	178	82.0	skateboard	128	157	81.5
zebra	192	236	81.4	clock	194	240	80.8
pizza	195	243	80.2	parking meter	48	60	80.0
bed	124	156	79.5	elephant	162	208	77.9
refrigerator	92	119	77.3	umbrella	207	268	77.2
baseball glove	88	114	77.2	motorcycle	201	261	77.0
sheep	184	239	77.0	Teddy bear	119	156	76.3
surfboard	169	222	76.1	couch	185	245	75.5
stop sign	55	73	75.3	cake	164	218	75.2
donut	140	187	74.9	sink	163	222	73.4
keyboard	106	145	73.1	oven	102	140	72.9
car	1113	1532	72.7	baseball bat	80	111	72.1
Sports ball	134	187	71.7	bowl	367	515	71.3
cup	532	753	70.7	snowboard	36	51	70.6
tie	120	171	70.2	sandwich	115	165	69.7
cellphone	167	243	68.7	wineglass	182	266	68.4

truck	229	335	68.4	cow	166	245	67.8
hotdog	63	93	67.7	potted plant	198	297	66.7
vase	148	222	66.7	bird	144	217	66.4
bottle	521	786	66.3	dining table	378	571	66.2
traffic light	347	536	64.7	remote	162	252	64.3
chair	787	1225	64.2	suitcase	137	214	64.0
bicycle	138	216	63.9	kite	96	154	62.3
fork	119	191	62.3	orange	148	238	62.2
banana	148	246	60.2	skis	105	177	59.3
apple	106	181	58.6	bench	187	327	57.2
boat	168	297	56.6	carrot	153	277	55.2
toothbrush	31	57	54.4	scissors	18	34	52.9
broccoli	121	248	48.8	book	327	680	48.1
spoon	104	219	47.5	knife	131	276	47.5
backpack	116	256	45.3	handbag	171	384	44.5
toaster	3	9	33.3	hair drier	1	11	9.1

Table VI shows the ten highest false positive ratio classes where the ratio is the FP number divided by the number of GT. Book class is the worst where the number of false positives and ground truths is 843, and 680, respectively. Next classes are apple, carrot, dining table, spoon, broccoli, tooth brush, bowl, vase, and knife.

TABLE VI. THE HIGHEST FALSE POSITIVE RATIO CLASSES

Class	# FP	# GT	ratio(%)
book	843	680	124.0
apple	203	181	112.2
carrot	267	277	96.4
dining table	483	571	84.6
spoon	170	219	77.6
broccoli	179	248	72.2
tooth brush	40	57	70.2
bowl	360	515	69.9
vase	154	222	69.4
knife	182	276	65.9

Table VII shows the ten lowest false positive ratio classes. Hair drier class is the best where the no false positive is generated at all though the number of ground truths is also the lowest. Next classes are zebra, airplane, giraffe, kite, toaster, elephant, fire hydrant, train, and frisbee.

TABLE VII. THE LOWEST FALSE POSITIVE CLASSES

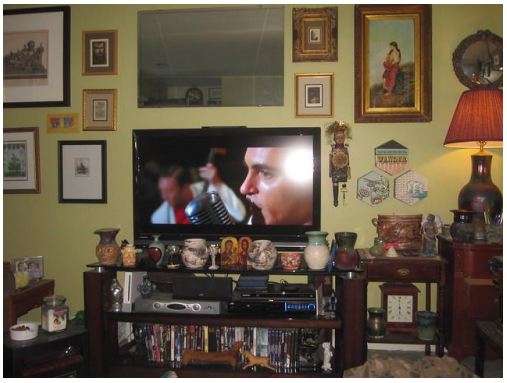
Class	# FP	# GT	ratio(%)
hair drier	0	11	0.0
zebra	20	236	8.5
airplane	12	135	8.9
giraffe	23	230	10.0
kite	16	154	10.4
toaster	1	9	11.1
elephant	25	208	12.0
fire hydrant	12	99	12.1
train	24	183	13.1
frisbee	17	109	15.6

Table VIII shows the true positive ratio by the object size. In the experiment, objects are grouped in their size, and true positive ratio is measured for the objects in the same group. As objects are small, the detection ratio is low. The 10% small objects are only detected 34.3% while the largest objects are of 93.7%. There exists an evident correlation between the object size and the detection ratio.

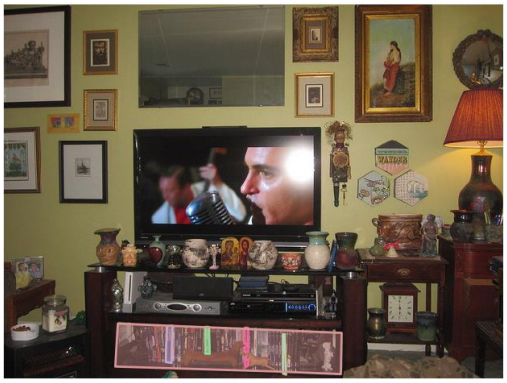
TABLE VIII. TRUE POSITIVE RATIO BY OBJECT SIZE

Area decile	Ratio (%, # TP/ # GT)
0%~10%	34.3%
10%~20%	53.1%
20%~30%	64.3%
30%~40%	69.0%
40%~50%	74.7%
50%~60%	79.0%
60%~70%	83.4%
70%~80%	86.4%
80%~90%	89.7%
90%~100%	93.7%

Some labeling information seems to be not perfect in the COCO dataset. In Fig. 1 (a), there are several ground truth boxes for books in the bottom area. However, the number of GT and the bounding box width are not correct as shown in Fig. 1 (b). Similarly, in Fig 1 (c), the green ground truth box for ski is too wide. The number of ground truth labeling is also somewhat confusing in some images. In Fig. 2, it is not easy to determine how many people, banana, apple, or, orange are in the image where the number of GT is shown below.



(a)
http://cocodataset.org/#explore?id=479248



(b)

(c)

http://cocodataset.org/#explore?id=334767

Fig. 1. Example of incorrect bounding box area

(a)
http://cocodataset.org/#explore?id=885
# of ground truth people: 8

(b)
http://cocodataset.org/#explore?id=4134
# of ground truth people: 13

(c)
http://cocodataset.org/#explore?id=6040
# of ground truth people: 9

(d)
http://cocodataset.org/#explore?id=2149
# of apple: 1

(e)
http://cocodataset.org/#explore?id=45472
# of oranges: 1

Fig. 2. Example of unclear object count

IV. CONCLUSIONS

In this paper, YOLOv3 algorithm is evaluated on the COCO dataset. Various true and false positive ratios are shown for each object in detail. It is quantitatively shown that small objects are not well detected. I discuss the COCO dataset labeling information some of which is not clear in the image or seems to be incorrect. This evaluation may be expected to help other detection studies for the performance analysis. The further evaluation of false positive and false negative remain as future works.

REFERENCES

[1] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In European conference on computer vision, pp. 21–37. Springer, 2016.

[2] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg. Dssd: Deconvolutional single shot detector. arXiv preprint arXiv:1701.06659, 2017.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, pp. 580–587. IEEE, 2014.

[4] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In NIPS, pages 91–99, 2015.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788, 2016.

[6] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7263–7271, 2017.

[7] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv: 1804.02767, 2018.

[8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft COCO: Common objects in context. In ECCV. 2014.

[9] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. IJCV, pp. 303–338, 2010.

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), 115(3), pp. 211–252, 2015.

[11] https://pjreddie.com/darknet/yolo/