YOLOv4 and how one-stage detectors may take over two-stage detectors

Joris LIMONIER

University of Côte d'Azur joris.limonier@hotmail.fr

Abstract. In the last 20 years, the field of object detection has become more and more prevalent. With Viola & Jones (2001), the field started with traditional methods. From 2014 on, more modern Deep Learning techniques were adopted, which lead to two main types of detectors: One-stage detectors and Two-stage detectors. In this paper we focus on the former, more particularly on YOLOv4 (You Only Look Once), and we compare it to the latter. One-stage detectors are meant to focus on speed of execution, while trying to keep a decent accuracy. Today, their accuracy still lacks behind the one of two-stage detectors in some cases, but their fast improvement makes us wonder whether they will take over two-stage detectors. In particular, YOLOv4 represents a great improvement over previous recent methods, showing impressive results both in accuracy and in speed.

Keywords: YOLOv4 · Object detection · One-stage detectors · Two-stage detectors

1 State of the Art

As shown in figure 1, the number of object detection-related publications per year has been greatly increasing over the past 20 years [8]. As a result, many detectors are now available for comparison with the one we are interested in, YOLOv4.

Figure 2 compares YOLOv4 to several of the most recent contenders, namely: YOLOv3 [5], EfficientDet [6], Adaptive Training Sample Selection (ATSS) [7], Adaptively Spatial Feature Fusion (ASFF) [4] and CenterMask [3].

Except for YOLOv3, it appears in figure 2 that for a given speed, YOLOv4 is more accurate than all other state of the art (SOTA) methods, and given an accuracy, YOLOv4 is faster than all other SOTA methods. When compared to its predecessor (YOLOv3), figure 2 doesn't show improvements in speed (not measured), but a great step forward in terms of accuracy. The accuracy improvement is fairly substantial, with an increase of about 30% in prediction correctness (from 32% to 42%). Moreover, the two papers have only been published two years apart from one another.

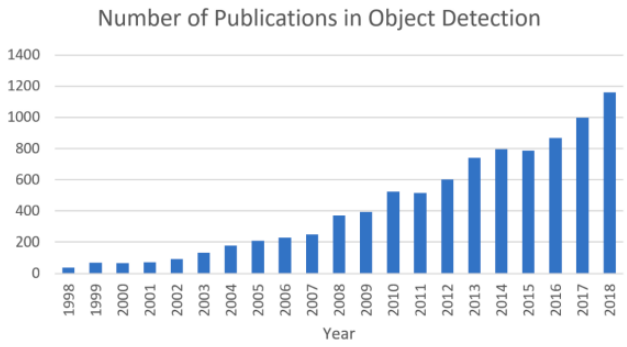

Fig. 1. Number of publications per year between 1998 and 2018 in the field of object detection.

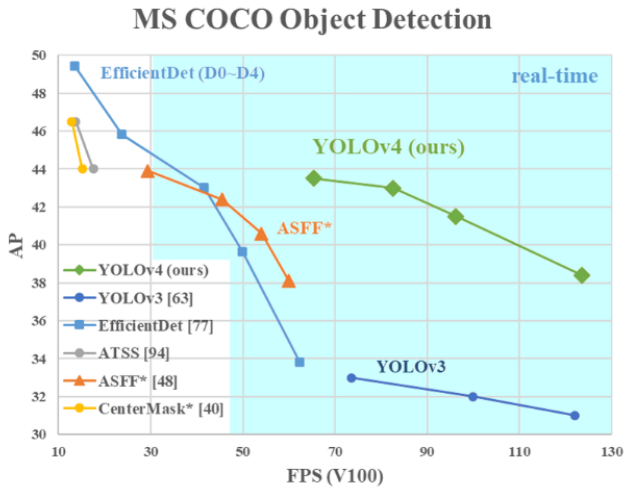

Fig. 2. Comparison between YOLOv4 and several other state of the art object detectors on the COCO dataset. This figure comes from Bochkovskiy et al. [1] and the numbers in the caption refer to their references.

2 Discussion and conclusion

As mentioned previously, there are two types of object detectors: one-stage detectors and two-stage detectors. The former focus on speed and are capable of directly detecting objects without a preliminary steps. The latter are usually more accurate but they need a preliminary step in order to determine regions of interest and before classifying them.
YOLOv4 is a one-stage object detector, but as shown in table 1, it performs similarly, or even higher, than other SOTA two-stage detectors ($38.4\% - 43.0\%$ against 43.0% for YOLOv4). The numbers have been gathered from two sources however: the accuracy for YOLOv4 comes from Bochkovskiy et al. [1], whereas the other accuracies come from Carranza-Garcia et al. [2] so despite our efforts to compare apples to apples, some disparities in the computation of the accuracy may have gone through. As a results, it seems fair to state that YOLOv4 performs "at least as well" as the other two-stage detectors measured. Therefore, based on the observations from 1, it appears that YOLOv4 doesn't perform worse than two-stage detectors, while showing more attractive classification speeds.

Table 1. Comparison in accuracy between SOTA two-stage detectors (using Faster RCNN) and one-stage detectors on the COCO dataset.

	Architecture	**Feature extractor**	**Accuracy (%)**
Two-stage detectors	Faster RCNN	FPN Res2Net101	43.0
	Faster RCNN	FPN ResNeXt101	41.2
	Faster RCNN	FPN ResNet152	40.1
	Faster RCNN	FPN ResNet101	39.8
	Faster RCNN	FPN ResNet50	38.4
One-stage detectors	YOLOv3	DarkNet-53	33.4
	YOLOv4	CSPDarknet-53	43.0

So what are the advantages of two-stage detectors ? Object recognition is making its way into more and more devices around us. These devices often strongly rely on fast identification of what is coming through video input (*e.g.* self-driving cars). Other devices require reliable identification given little computing power (*e.g.* robot vacuum cleaners), which goes alongside faster classification speeds.
Furthermore, the rate of improvement shown in figure 2 and table 1 between YOLOv3 and YOLOv4 in only two years time hints that even more impressive results could be obtained in the upcoming years. A similar trend would allow the elaboration of significantly faster and more accurate detectors in the near future. By continuing on a similar progression, it seems that one-stage detectors could make two-stage detectors obsolete rather sooner than later.
It appears to us that one-stage detectors may be the way to go and will probably take over the majority of cases of object detection. A word of caution however, in

the case of slow processes (not heavily time reliant), it is possible that two-stage detectors may still remain in use. Such cases may value accuracy much more than they value speed, therefore preferring those longer detectors. Additionally, Carranza-Garcia et al. [2] also bring proof of some specific cases where two-stage detectors perform better than their counterpart, such as high resolution. Indeed, although the accuracy is comparable between the two type of detectors in low-resolution images ($\approx 42\%$ against $\approx 45\%$), the gap widens for high-resolution images ($\approx 48\%$ against $\approx 56\%$).
So the only question remaining is whether one-stage detectors will become absolutely more accurate (that is, not relatively to speed but simply beating two-stage detectors in terms of accuracy). Our prediction on that front is that **they will**, however this will probably still take a couple years, or even decades. The dynamic seems in favor of one-stage detectors but they still have some way to go before being completely on part with two-stage ones. They have even more way to go (if it ever happens) until systematically being superior, that is in all edge cases, for all types of objects and under all circumstances (such as the image resolution). The whole process of identifying a region of interest before clasifying objects in it seems too cumbersome to be true. Analogically, it seems like issues faced by physicists with two irreconcilable theories that one day someone manages to merge in a more elegant way. Our opinion is that the field of object detection is waiting for such a kind of event to occur and that this reunification will come eventually.

References

1. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv:2004.10934 [cs, eess] (Apr 2020), http://arxiv.org/abs/2004.10934, arXiv: 2004.10934
2. Carranza-García, M., Torres-Mateo, J., Lara-Benítez, P., García-Gutiérrez, J.: On the Performance of One-Stage and Two-Stage Object Detectors in Autonomous Vehicles Using Camera Data. Remote Sensing **13**(1), 89 (Dec 2020). https://doi.org/10.3390/rs13010089, https://www.mdpi.com/2072-4292/13/1/89
3. Lee, Y., Park, J.: CenterMask: Real-Time Anchor-Free Instance Segmentation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13903–13912. IEEE, Seattle, WA, USA (Jun 2020). https://doi.org/10.1109/CVPR42600.2020.01392, https://ieeexplore.ieee.org/document/9156712/
4. Liu, S., Huang, D., Wang, Y.: Learning Spatial Fusion for Single-Shot Object Detection. arXiv:1911.09516 [cs] (Nov 2019), http://arxiv.org/abs/1911.09516, arXiv: 1911.09516
5. Redmon, J., Farhadi, A.: YOLOv3: An Incremental Improvement. arXiv:1804.02767 [cs] (Apr 2018), http://arxiv.org/abs/1804.02767, arXiv: 1804.02767
6. Tan, M., Pang, R., Le, Q.V.: EfficientDet: Scalable and Efficient Object Detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10778–10787. IEEE, Seattle, WA, USA (Jun 2020). https://doi.org/10.1109/CVPR42600.2020.01079, https://ieeexplore.ieee.org/document/9156454/

7. Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: Bridging the Gap Between Anchor-Based and Anchor-Free Detection via Adaptive Training Sample Selection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9756–9765. IEEE, Seattle, WA, USA (Jun 2020). https://doi.org/10.1109/CVPR42600.2020.00978, https://ieeexplore.ieee.org/document/9156746/
8. Zou, Z., Shi, Z., Guo, Y., Ye, J.: Object Detection in 20 Years: A Survey. arXiv:1905.05055 [cs] (May 2019), http://arxiv.org/abs/1905.05055, arXiv: 1905.05055