Comparative Review of YOLO & MobileNet Versions: A Case Study

Naethan Jacob[1], Dr. Vishalakshi Prabhu H[2]

[1]Dept. of Computer Science and Engineering R.V College of Engineering, naethanjacob@gmail.com
[2]Dept. of Computer Science and Engineering R.V College of Engineering, vishalaprabhu@rvce.edu.in

---***---

Abstract - *In the current era, computer vision has developed significantly due to advances in deep learning algorithms and architectures. This has allowed for computer vision tasks such as object detection and image classification to achieve state of the performance. This has led to increased demand to perform these tasks on edge devices. This paper outlines the comparative review of leading architectures in the fields of object detection and image classification in real time for edge devices, namely You Only Look Once (YOLO) and MobileNet. Both the models have variants V1,V2 and V3 and the papers highlights the differences among these versions.*

Key Words: convolutional neural networks (CNN), deep learning, image classification, object detection

1. INTRODUCTION

Object detection is a computer technology related to computer vision and image processing that deals with detecting instances of semantic objects of a certain class (such as humans, buildings, or cars) in digital images and video Well-researched domains of object detection include face detection and pedestrian detection. Object detection has applications in many areas of computer vision, including image retrieval and video surveillance. Image classification is the process of taking an input (like a picture) and outputting a class (like “cat”) or a probability that the input is a particular class (“there’s a 90% probability that this input is a cat”). A human can look at a picture and identify the object present, similarly computers are trained to do the same. With the proliferation of Internet of Things (IoT) devices and rapid advances in the computer vision tasks of object detection and image classification. There arose a requirement to perform these tasks in real time for edge devices. In response to these two novel architectures have been proposed YOLO, for object detection and MobileNet for Image classification. This paper is a case study discussing the various versions of these models.

2. YOLO V1-V3

YOLO presented in [1], [2] and [3] is a completely novel neural network based approach to object detection. In previous approaches and works, image classifiers were repurposed using the sliding window approach to perform object detection. Spatially separated bounding boxes with the class possibilities is used to perform object detection via a regression based approach. Category and bounding boxes possibilities directly from full pictures in one analysis is predicted by one neural network, shown in Figure 1. It can be highly optimized to perform end-to-end directly on object detection performance, since the complete architecture is one convolutional neural network. The unified design is extraordinarily quick. The YOLO model can process pictures in a time period at forty five frames per second. Fast YOLO, a smaller version of the neural network, processes an astounding a hundred and fifty five frames per second whereas still achieving double the mAP of different real-time detectors. Compared to progressive detection systems, YOLO makes additional localization errors however is far less seemingly to predict false detections wherever nothing exists. Finally, the proposed model is able to learn very general representations of objects. It performs far better than all other object detection strategies, such as Deformable Parts Model (DPM) and Region CNN (R-CNN), by a very good margin when generalizing from natural pictures to human artwork on both the Picasso Dataset as well as the People-Art Dataset.

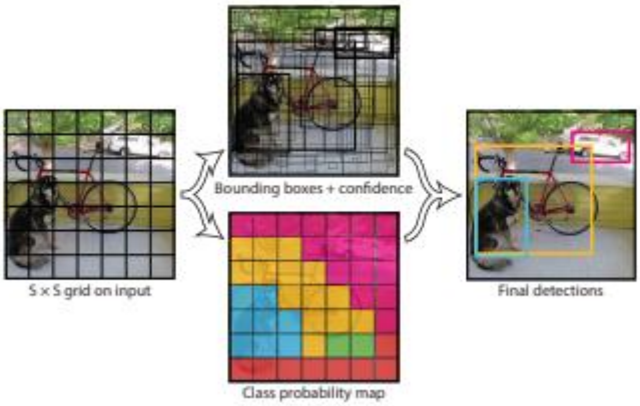
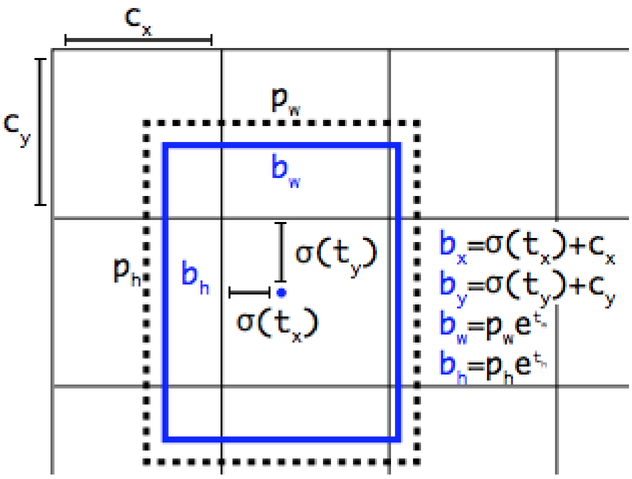
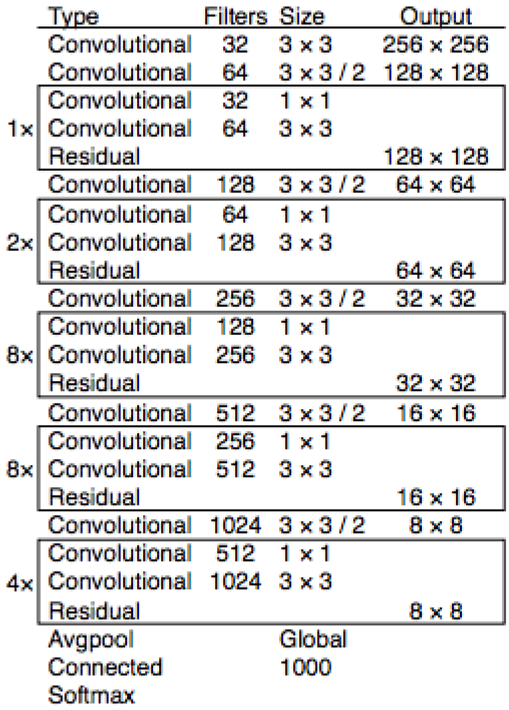


Fig -1: YOLO model dividing the image into a SxS grid and creating the respective bounding boxes [1]

YOLOV2 presented in [2], is the next iteration of [1]. A novel model is proposed "YOLO9000", the model is able to perform object detection over 9000 image categories in real time with state of the art performance . Numerous enhancements to the YOLO detection model, both distinctive and galvanized from previous works, are proposed. The improved model, YOLOv2, achieves state-of-the-art performance on object detection tasks like PASCAL VOC and coco dataset. YOLOv2 gets 76.8 percent mAP on Visual Object Classes (VOC) 2007 running at 67 frames per second. YOLOv2 is able to attain 78.6 mAP, while running at 40 frames per second. Attaining much better performance than faster R-CNN with ResNet and Single Shot Detector (SSD), which are state-of-the-art models while still running considerably quicker. Finally the authors have a tendency to propose how to collectively train on object detection and classification. Mistreatment of this technique the authors have a tendency to train YOLO9000 at the same time on the coco detection dataset and additionally the ImageNet classification dataset. Combined training helps YOLO9000 to predict detections for object categories that are not labeled properly. The model was validated on the ImageNet object detection task. YOLO9000 gets 19.7 mAP on the ImageNet detection and 78.6 on Pascal, validation set despite solely having detection information for forty four of the two hundred categories. On the 156 categories not in coco, YOLO9000 gets 16.0 mAP. However YOLO will detect quite simply two hundred classes; it predicts detections for quite 9000 completely different object classes and it still runs in real-time.





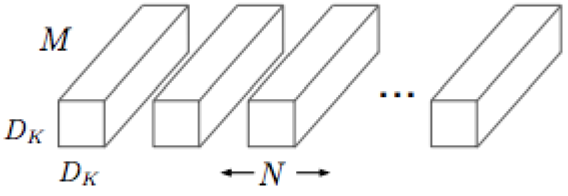
	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Fig -2: Bounding box calculations [3] **Fig -3**: Network Architecture [3]

YOLOV3 presented in [3], is the next iteration of [1] and [2]. In this paper, authors present updates to the YOLO model. They have proposed a bunch of little design changes to YoloV2 to make it better, as shown in Figure 2 and 3. The authors also trained this new network and have provided the weights under an open source license. The updated model is a little bigger than the previous works but more accurate. The model is still fast despite it's increased size. With an input image size of 320x320 pixels YoloV3 is able to perform predictions in 22 ms with 28.2 mAP, this is as accurate as the SSD model but much faster. When compared with the old .5 IOU mAP detection metric YoloV3 achieves good performance. The model is able to perform predictions in 51 ms on a Titan X GPU with 57.9 mAP@50 , compared to 57.5 mAP@50 in 198 ms by RetinaNet, as shown in Table 3, similar performance but 3.8x faster

3. MobileNet V1-V3

MobileNet V1, proposed in [4] presents a category of extremely economical neural network models referred to as MobileNets for mobile in addition to embedded computer vision applications, shown in Figure 5. MobileNets use an efficient design that involves depth-wise separable convolutions operations, shown in Figure 4, to form light- weight deep convolutional neural networks. The authors have introduced two easy global hyper-parameters that allow for a trade off between accuracy and latency in an exceedingly very economical manner. These hyper-parameters permit the top user to choose the acceptable sized model for his or her application supporting the software needs of the task. The authors present intensive experiments on resource and accuracy tradeoffs and show robust performance compared to alternative standard models on ImageNet classification. The effectiveness of the MobileNet models is shown by the authors across a decent vary of applications and use cases as well as object detection, fine grain classification, face attributes and large scale geo-localization



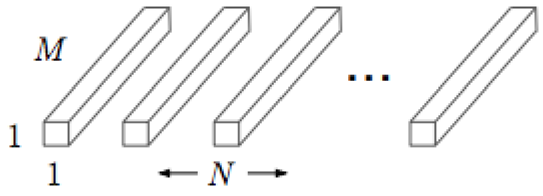


Fig -4.1: Standard convolutional filters [4] **Fig -4.2**: Depthwise convolution filters [4]

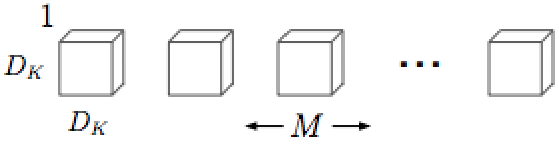


Fig - 4.3: 1×1 Convolutional Filters called Pointwise Convolution in the context of Depthwise Separable Convolution [4]

MobileNet V2, presented in [5] is the next iteration of [4]. The authors describe a completely unique mobile neural network, MobileNetV2, that improves considerably on the previous state of the art performance of mobile neural networks on many tasks and benchmarks together with a spectrum of various neural network sizes. The authors even have delineated economical ways in which of applying these mobile neural networks to object detection during a distinctive framework that's remarked as SSDLite. Additionally to the current, the authors demonstrate the way to produce mobile linguistics segmentation neural networks through a reduced sort of DeepLabv3 that is remarked as Mobile DeepLabv3. The MobileNetV2 model relies on associate degree inverted residual structure, as depicted in Figure 7 , wherever the input and the output of the residual block are skinny bottleneck layers opposite to ancient residual models that use distended representations within the input an MobileNetV2 uses light-weight depthwise convolutions to filter features within the intermediate expansion layer, shown in Figure 6. Additionally, the authors realize that it's necessary to get rid of nonlinearities within the narrow layers in order to take care of representational power. The authors demonstrate that this improves performance and supplies an intuition that led to the current style. Finally, the author's approach permits decoupling of the input/output domains from the quality of the transformation, which provides a convenient framework for any analysis. The authors measure the model's performance on Imagenet classification, coco object detection, VOC image segmentation. The authors evaluate the newly proposed neural network on trade offs between accuracy, and sort of operations measured by multiply-adds (MAdd), also as the variability of parameters

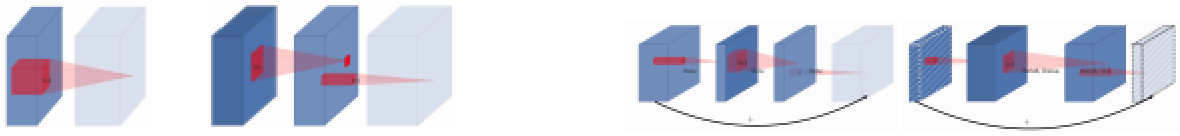


Fig - 6.1: Regular [5] **6.2:** Separable [5] **6.3:** Separable with linear bottleneck [5] **6.4:**Bottleneck with expansion layer [5]

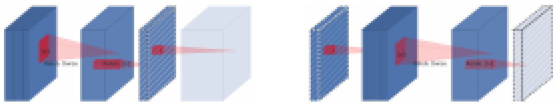


Fig - 7.1: Residual Block [5] **Fig - 7.2**: Inverted Residual Block [5]

MobileNet V3, presented in [6] is the next iteration of [4] & [5]. The authors present the latest version of MobileNets architectures based on a mixture of complementary search techniques also as a unique neural network design. MobileNetV3 is geared towards mobile phone CPUs through a mixture of hardware-aware network architecture search (NAS) supported by the NetAdapt architecture then is further improved through unique neural network architecture advances. This paper explores the idea of how automated neural network architecture search algorithms and neural network design can work in conjunction to leverage complementary approaches increasing the overall state of the art performance. Through this method the authors propose two novel MobileNet models for public use: MobileNetV3-Large and MobileNetV3-Small which are intended for top and low resource software requirements. This is depicted in chart 1. These neural networks are then repurposed and applied to the pc vision tasks of object detection and semantic segmentation. For the matter of semantic segmentation (or any dense pixel prediction), the authors have proposed a completely unique efficient segmentation decoder Lite Reduced Atrous Spatial Pyramid Pooling (LR-ASPP). The authors have reported new state of the art results for mobile network classification, detection and segmentation.

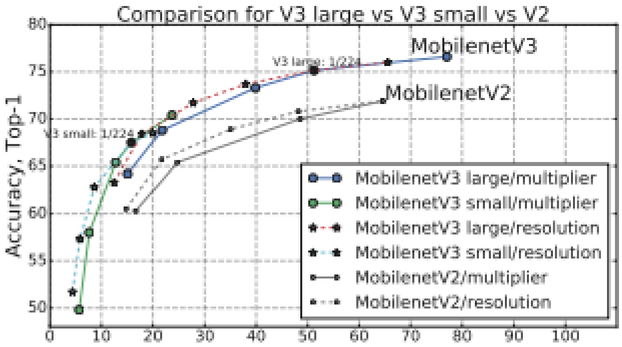


Chart - 1: Latency on pixel 2 [6]

4. CONCLUSION

This paper has provided state of the art advances in object detection model YOLO from version one to version three, as well as the advances made in image classification model MobileNet from version one to version three. Both the models, namely YOLO and MobileNet are compared with other detection frameworks and network architectures. Progress in these architectures will enable the AI dream to come true.

REFERENCES

[1] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, "You Only Look Once: Unified, Real-Time Object Detection", Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR 2016), pp:779-788, June 2016, Las Vegas, USA.

[2] Joseph Redmon, Ali Farhadi, "YOLO9000: Better, Faster, Stronger", arxiv:1612.08242.

[3] Joseph Redmon, Ali Farhadi (2018). YOLOv3: An Incremental Improvement, arxiv:1804.02767.

[4] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, arxiv:1704.04861.

[5] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen (2019). MobileNetV2: Inverted Residuals and Linear Bottlenecks, arxiv:1801.04381.

[6] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, Hartwig Adam (2019). Searching for MobileNetV3, arxiv:1905.02244.