ISSN 2347 - 3983

Volume 9. No.11, November 2021

International Journal of Emerging Trends in Engineering Research

Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter039112021.pdf

https://doi.org/10.30534/ijeter/2021/039112021

Object Detectors' Convolutional Neural Networks backbones : a review and a comparative study

Sara Bouraya[1], Abdessamad Belangour[2]

[1]Laboratory of Information Technology and Modeling,Hassan II University, Faculty of sciences Ben M'sik, Casablanca, Morocco, sarabouraya95@gmail.com

[2]Laboratory of Information Technology and Modeling, Hassan II University, Faculty of sciences Ben M'sik, Casablanca, Morocco,belangour@gmail.com

Received Date : October 04, 2021 Accepted Date : October 25, 2021 Published Date : November 07, 2021

ABSTRACT

Computer vision is a scientific field that deals with how computers can acquire significant level comprehension from computerized images or videos. One of the keystones of computer vision is object detection that aims to identify relevant features from video or image to detect objects. Backbone is the first stage in object detection algorithms that play a crucial role in object detection. Object detectors are usually provided with backbone networks designed for image classification. Object detection performance is highly based on features extracted by backbones, for instance, by simply replacing a backbone with its extended version, a large accuracy metric grows up. Additionally, the backbone's importance is demonstrated by its efficiency in real-time object detection. In this paper, we aim to accumulate the crucial role of the deep learning era and convolutional neural networks in particular in object detection tasks. We have analyzed and have been concentrating on a wide range of reviews on convolutional neural networks used as the backbone of object detection models. Building, therefore, a review of backbones that help researchers and scientists to use it as a guideline for their works.

Key words : Object Detection, Deep Learning, Computer vision, Backbone.

1. INTRODUCTION

Object detection is a computer vision technique used for locating instances of objects in videos or images.Object detection models typically rely on deep learning or machine learning to produce meaningful results. During the last decades, Deep Leaning techniques of Object Detection have been growing rapidly. Thus, we can find a variety of models based on Deep Learning approaches.Deep Learning approaches could be divided into two categories one stage detectors such as Yolo[1], RetinaNet[2], and SSD[3] and two-stage detectors such as R-CNN[4], Fast-R-CNN[5], Faster R-CNN[6], and Mask R-CNN[7] (see Figure 1).

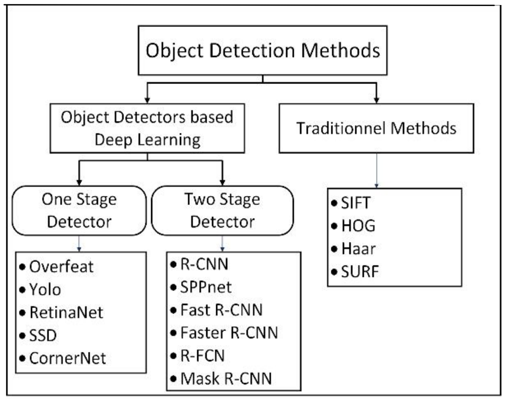

Figure 1: Object Detection Methodologies' Categories

Without ignoring traditional methodologies, these methods are generally based on three different stages. Firstly, informative region selection, when we try to find object location that is appearing in different shapes and different locations. Based on the sliding window this stage could be computationally expensive and capturing irrelevant results Secondly, feature extraction is based on algorithms like HOG or SIFT. Finally, the third stage is relying on some classifiers to classify the target object. These methods' drawbacks are computational costs.

On the other hand, deep learning-based methods are based on different steps that we can summary them up in(see Figure 2)

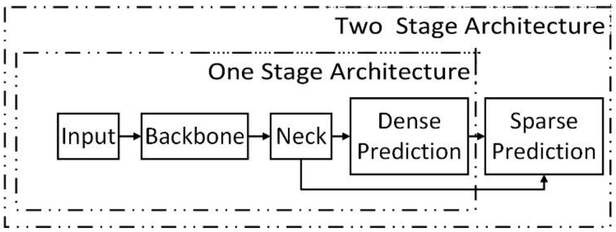

Figure 2: Object Detection Based Deep Learning Architecture

As we can see, there are different steps in reaching object detection based on deep learning starting from an input image or a video frame. Then the next step is feature extraction that can be reached using Backbones that we are going to see in this paper.

Backbones are convolutional neural networks based on different layers also, moreover, the Neck stage refers to a collection of layers that collect feature maps and they are composed of several top-down paths and several bottom-up paths. Next, the head of the model that can predict bounding boxes of objects and their classes, can be either a one-stage detector or a two-stage detector. Two-stage detectors are more complicated than one-stage detectors which are elegant and straightforward.

Let us see some of the architectures of two-stage detectors that are complicated and let us observe their improvements. Ranging from object detection to object segmentation. In other words, starting from R-CNN[4] to Mask R-CNN[7].

R-CNN[4]stands for “Region-based Convolutional Neural Networks”. It is one of the famous models that gave a lot of performance to object detection. The idea behind its architecture is composed of two steps. Firstly, relying on selective search to identify several bounding boxes object region candidates that are named region of interest or Roi. Its next step based on CNN can extracts features from each region separately for classification(see Figure 3).

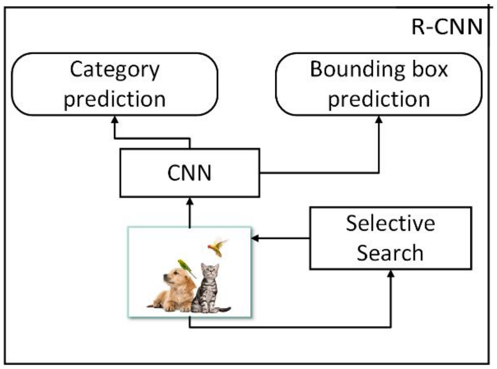

Figure 3: R-CNN Architecture

Fast R-CNN[4] stands for “Fast Region-based Convolutional Neural Networks”. To make R-CNN faster the authors proposed another training ticks to gain more accuracy(see Figure 4). They improved the training process by unifying three models into a jointly trained framework and growing the shared computation result. The model aggregates the feature vectors into one CNN one forward pass over the input and sharing the feature matrix without treating them separately. Next, this matrix was collected to be used as an input for classification tasks and bounding boxes regression.

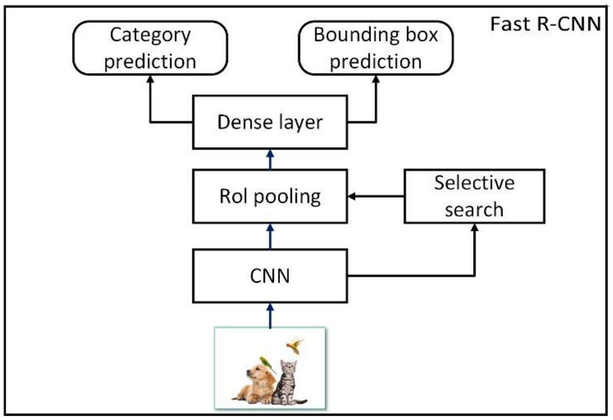

Figure 4: Fast R-CNN Architecture

Faster R-CNN[8] stands for “Faster Region-based Convolutional Neural Networks”. The main idea behind Faster R-CNN [8]is to integrate the region proposal model into CNN which going to make the R-CNN [4]family train rapidly. This model is proposed in 2016 its architecture is based on constructing a unified model composed of region proposal network and Fast R-CNN[5] meanwhile a shared convolutional feature layer.

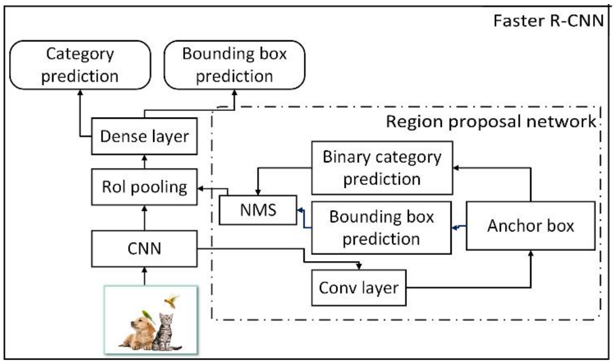

Figure 5: Faster R-CNN Architecture

Mask R-CNN[7] stands for “Fast Region-based Convolutional Neural Networks”. This model was proposed in 2017 to make improvements of Faster R-CNN[6] to deal with image segmentation (see Figure 6). This model's main idea is to predict pixel-level masks. Relying on Faster R-CNN[6], Mask R-CNN [7]adds to its architecture the third branch which is used to predict the mask at the same time as to classification task and bounding box prediction. The mask is also a fully connected network that reaches a segmentation task applied to each region.

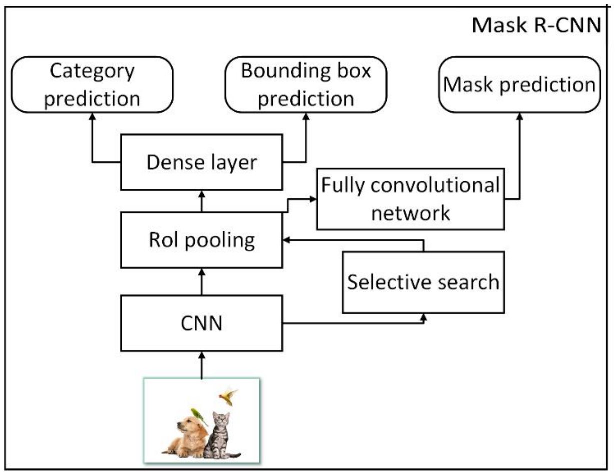

Figure 6: Mask R-CNN Architecture

2. BACKGROUND

All the discussed architectures in the previous section as I said, are relying on the backbone of their architecture. In this section, we are going to discuss some of the useful backbones in Object Detection such as VGG[9], ResNet[10], and so on.

Convolutional Neural Networks have been used in several visual tasks. One of these tasks is image classification. Their main role is feature extraction, which referred us to Backbones. Many scientists implement the successful model in the ImageNet classification contest, to their models to gain better performance. These convolutional neural networks have different architectures and characteristics.

AlexNet[11]is repeatedly considered the pioneer of convolutional neural networks and the beginning point of the deep learning boom. AlexNet[11]competed in the famous ImageNet Large Scale Visual Recognition Challenge in 2012. The proposed network achieved high accuracy. AnAlexNet[11]architectural model is depicted in Figure 7.AlexNet [11]Architecture is composed of 8 layers. It contains eight learned layers, i.e., five convolutional and three fully connected in which three softmax pooling.

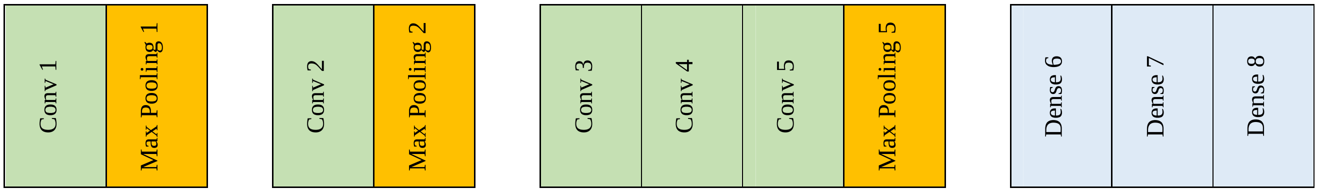

Figure 7: An illustration of the AlexNet architecture

VGG16[9] is convolutional neural network that won ImageNet Large Scale Visual Recognition Challenge competition in 2014. VGG16 [9]has been regarded as the best model at that time. 16 in VGG16[9] refers to its 16 layers. Indeed VGG16 [9]is a large model with 138 parameters approximately. As shown in Figure 8, VGG16[9] have 5 Convolution block and 1 fully connected block. Each convolution block contains a set of convolutional layers with a pooling. Finally, three fully connected layers are referred to as Dense in Figure 8.

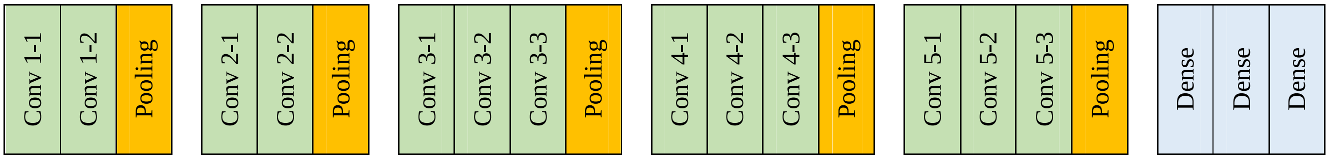

Figure 8: An illustration of the VGG16 architecture

ResNet18[12]is a convolutional neural network that won ImageNet Large Scale Visual Recognition Challenge Classification competition in 2015.Residual Network trained networks with 100 and 1000 layers also. 18 refers to the number of convolutions that are 18 and two pooling.

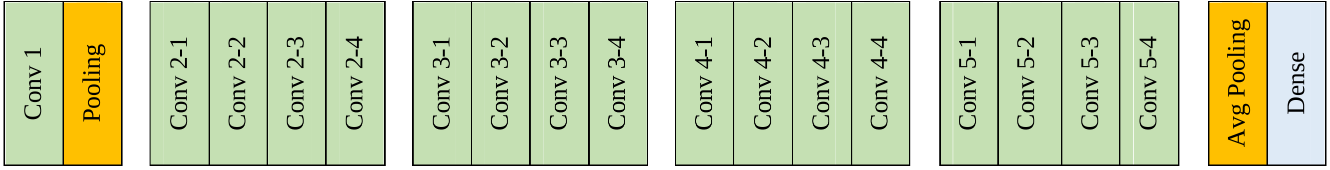

Figure 9: An illustration of ResNet18 architecture

GoogleNet[13]is based on inceptions as shown in figure 10. Each inception is composed of several convolutional layers and max pooling. The inception module contains four parallel operations.

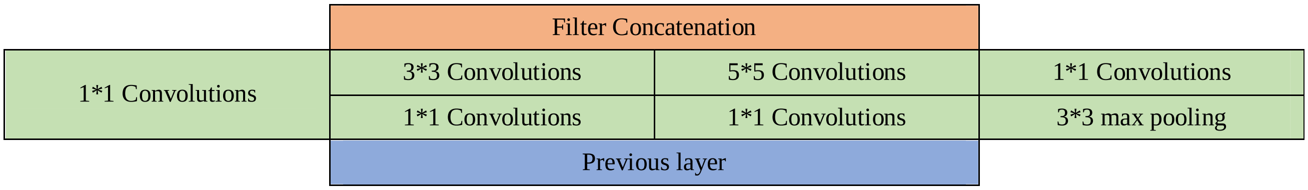

Figure 10: An illustration of Inception architecture

GoogleNet[13] architecture contains 22 layers with 27 pooling layers. In total there are 9 inception modules. After the inception modules, there is the global average pooling as illustrated in Figure 11.

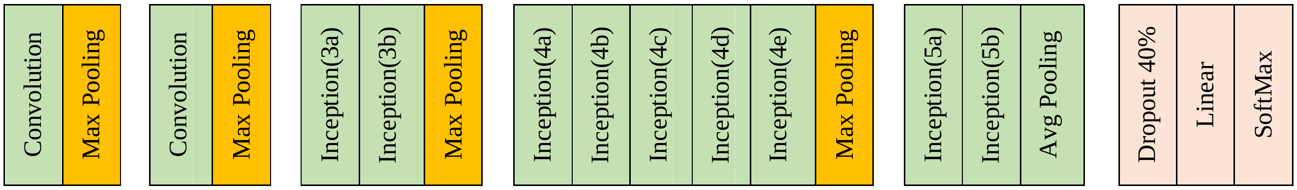

Figure 11: An illustration of GoogleNet architecture

In **DenseNet** [14]architecture, each layer is connected to every other layer, thus the name **Densely Connected Convolutional Network.** This is the main idea of DenseNet that is extremely powerful. Hence, The input of each layer inside **DenseNet** [14]is the concatenation of feature maps from previous existent layers (see Figure 12).

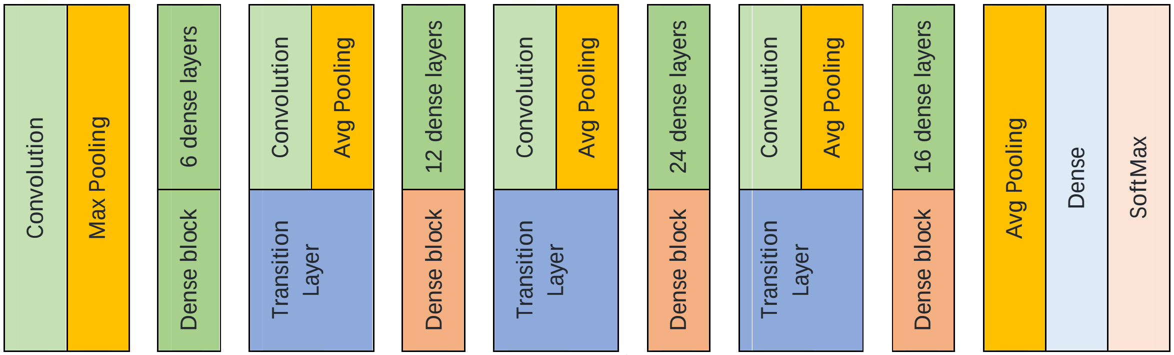

Figure 12: An illustration of DenseNet architecture

MobileNet[15] utilizes depthwise separable convolutions instead of the standard convolutions to reduce computation and model size except for the first layer. Thus, it can be used to construct lightweight deep neural networks for embedded and mobile vision applications. All layers are followed by batch normalization and ReLU non-linearity. However, the final layer is a fully connected layer without any non-linearity and then softmax for classification (see Figure 13).

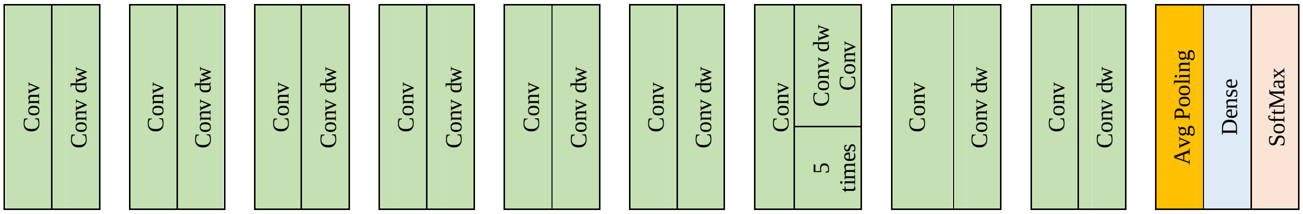

Figure 13: An illustration of MobileNet architecture

3. Comparaison Of Backbones

This table illustrates the deep learning model used for the classification task of the ImageNet Large Scale Visual Recognition Challenge. The number associated with each name referred to the number of layers. This table contains the model name, reference, paper title, accuracy, finally, and time(see Table 1).Our comparison criteria in terms of time and accuracy. Time refers to the training time on the ImageNet dataset. Accuracy is an evaluation metric that describes generally how the model performs across all classes. It is counted based on the ratio of the correct number of predictions to the number total of predictions.

The accuracy metric is between 0% and 100%. There are also other performances such as Recall, Precision,

Table 1: Accuracy and Time of Classification models based on deep learning

Model	Ref	Paper title	Accuracy %	Time
vgg16	[9]	Very Deep Convolutional Networks For Large Scale Image Recognition	70.79	24.95
vgg19			70.89	24.95
resnet18	[10]	Deep Residual Learning for Image Recognition	68.24	16.07
resnet50			74.81	22.62
resnet101			76.58	33.03
resnet152			76.66	42.37
resnet50v2			69.73	19.56
resnet101v2			71.93	28.80
resnet152v2			72.29	41.09
resnext50	[16]	Aggregated residual transformations for deep neural networks	77.36	37.57
resnext101			78.48	60.07
densenet121	[14]	Densely connected convolutional networks	74.67	27.66
densenet169			75.85	33.71
densenet201			77.13	42.40
inceptionv3	[17]	Rethinking the Inception Architecture for Computer Vision	77.55	38.94
xception	[18]	Xception: Deep learning with depthwise separable convolutions	78.87	42.18
inceptionresnetv2	[19]	Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning	80.03	54.77
seresnet18	[20]	Squeeze and Excitation Networks	69.41	20.19
seresnet34			72.60	22.20
seresnet50			76.44	23.64
seresnet101			77.92	32.55
seresnet152			78.34	47.88
seresnext50			78.74	38.29
seresnext101			79.88	62.80
senet154			81.06	137.36
nasnetlarge	[21]	Learning Transferable Architectures for Scalable Image Recognition	**82.12**	116.53
nasnetmobile			74.04	27.73
mobilenet	[15]	MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications	70.36	15.50
mobilenetv2	[22]	MobileNetV2: Inverted Residuals and Linear Bottlenecks	71.63	18.31

After gathering the main methods to compare them (see Table 1).One on the one hand, with a view to detect the best model in term of time.

Here is a bar plot shows the best method.

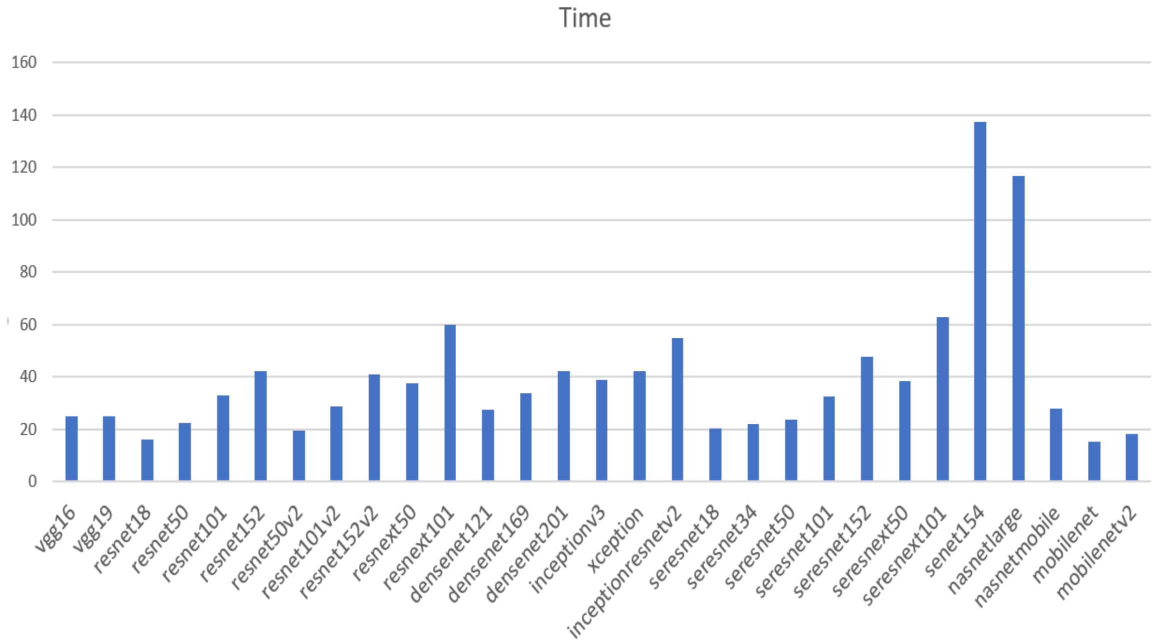

Figure 12: Time Training Comparisonof Classification Models Based Deep Learning

On the other hand, in term of accuracy this is the bar plot(see Figure 13).

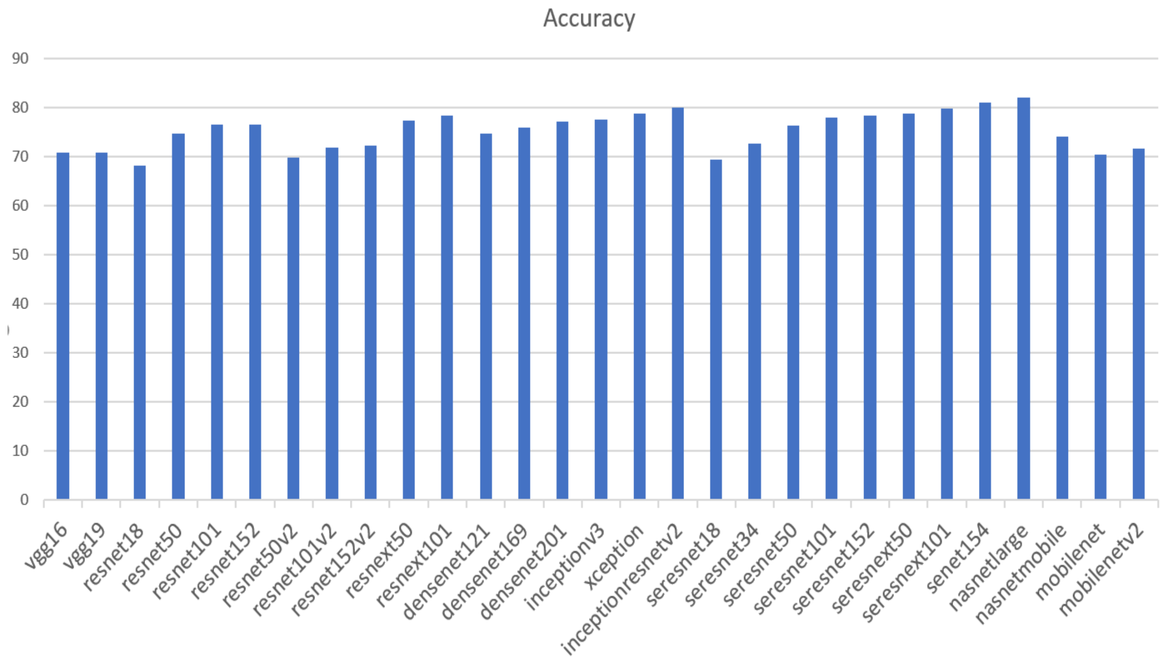

Figure 13: Accuracy Metric comparison Of Classification Models Based Deep Learning

4. DISCUSSION

This paper covers lot of models, starting from gathering the relevant methods of object detection that are divided into two categories traditional approaches and those based deep learning. We are interested in deep learning based that are divided into two techniques one stage detectors and two stage detectors.

In our second stage, we said that deep learning-based models are list of small models of deep learning, their architecture contains backbone, neck, and at the end sparse prediction or dense prediction it depends of each category. So, we had interested in one backbone part. We gathered the famous methodologies in this part.

After gathering backbone methodologies based deep learning saying deep learning saying a sit of layers. We discussed each method separately. Without ignoring the architecture of each backbone model.

At the end, and after discussing and analyzing the architectures, we defined a benchmark table that contains the performance in terms of time and accuracy. Our methods are implemented on ImageNet Dataset.

After all of these steps, plots have handed based on bar plot that visualize the best and the worst methods in terms of time and accuracy.

In terms of time, MobileNet and ResNet 18 have less time in training, not like senNet150 and NesNetLarge.

Based on our comparisonNasNetLarge and SeNet154reached the high performance in terms of accuracy however relying on time they are the worst.

ResNext101, InceptionResNetsV2, SerResNext101 are reaching greater than 76% and in terms of time they take medium place.Some other models are great in terms of accuracy, for example, ResNet18, MobileNet but in terms of accuracy, they reach greater than 68.

In general,more layer increases performance relied on accuracy metric and increases the training time which is not good. The main purpose of researchers in deep learning erea,is looking for higher accuracy metric and less training time.

5. Conclusion

This paper has givena whole globalvision about Object Detection one and two-stage detectors as well as a close up view of their backbone part. Finally,it has given you a comparison of some classification models.

In addition, we have presented the techniques of object detection, the traditional ones and those based on deep learning. We have focused on Two-stage detectors that are based on the backbone or feature extraction stage. Furthermore, we have stated the most relevant techniques based on deep learning and their architecture.

Additionally, a survey has been made on the most relevant image classification techniques for the ImageNet Large Scale Visual Recognition Challenge Classification competition.The architecture of these techniques has been discussed and decorticated.

After gathering some techniques of image classification based on deep learning, we have made a comparison of these models in terms of time and accuracy because of their importance in this field.

The future work is to implement these techniquesusing TensorFlow in object detection.Some of the models are good in terms of accuracy and others in terms of time. Thus, our future work will focus on finding a new model that combines less time and high accuracy. This is our main challenge that is going to be implemented in object detection-based deep learning approaches.

References

1.J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 779–788, 2016.

2. T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal Loss for Dense Object Detection,” IEEE Trans. Pattern Anal. Mach. Intell., pp. 318–327, 2020.

3. W. Liu et al., “SSD: Single shot multibox detector,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), pp. 21–37, 2016.

4. R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 580–587, 2014.

5. R. Girshick, “Fast R-CNN,” Proc. IEEE Int. Conf. Comput. Vis., pp. 1440–1448, 2015,“Fast R-CNN,” Proceedings of the IEEE International Conference on Computer Vision. pp. 1440–1448, 2015.

6. S. Ren, K. He, and R. Girshick, “Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks,” pp. 1–9.

7. K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” IEEE Trans. Pattern Anal. Mach. Intell., pp. 386–397, 2020.

8. S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” IEEE Trans. Pattern Anal. Mach. Intell., pp. 1137–1149, 2017.

9. Karen Simonyan∗& Andrew Zisserman+, “VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION Karen,” Am. J. Heal. Pharm., pp. 398–406, 2018.

10. K. He and J. Sun, “Deep Residual Learning for Image Recognition,” pp. 1–9.“Deep Residual Learning for Image Recognition.” pp. 1–9.

11. H. G. Krizhevsky A., Sutskever I., “ImageNet Classification with Deep Convolutional Neural Networks,” NIPS, pp. 84–90, 2012.

12. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 770–778, 2016.

13. C. Szegedy et al., “Going deeper with convolutions Christian,” Popul. Health Manag., pp. 186–191, 2015.

14. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017.

15. A. G. Howard and W. Wang, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” 2012..

16. S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, pp. 5987–5995, 2017.

17. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 2818–2826, 2016.

18. F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, pp. 1800–1807, 2017.

19. M. Längkvist, L. Karlsson, and A. Loutfi, “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning,” Pattern Recognit. Lett., pp. 11–24, 2014.

20. J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze and Excitation Networks,” IEEE Trans. Pattern Anal. Mach. Intell., pp. 2011–2023, 2020.

21. B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, “Learning Transferable Architectures for Scalable Image Recognition,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 8697–8710, 2018.

22. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 4510–4520, 2018.

23. S. Kumar, J. Tiwari, “A Review: Machine Learning Approach and Deep Learning Approach for Fake News Detection“, International Journal of Emerging Technologies in Engineering Research (IJETER) Volume 9, Issue 8, August (2021).