CV703 Assignment 2
Instance Segmentation of Aerial Images on iSAID dataset

Sultan Abu Ghazal
MBZUAI
sultan.abughazal@mbzuai.ac.ae

Aidana Nurakhmetova
MBZUAI
aidana.nurakhmetova@mbzuai.ac.ae

Roba Al Majzoub
MBZUAI
roba.majzoub@mbzuai.ac.ae

Abstract

In the following report, the effect of using different pretraining datasets and number of fine tuned layers in each was studied on the task of instance segmentation. Several experiments were carried out on each of the two pretrained models, where in each experiment a single layer was frozen during training. Generally, it was observed that the model pretrained on the COCO dataset performed better in the instance segmentation task than the model pretrained on ImageNet. This observation falls in line with the expected outcome; since the COCO dataset, compared to the ImageNet dataset, focuses on segmenting individual object instances [7]. Furthermore, it was observed that as layers with higher level features are frozen, the performance deteriorates. Intuitively, high level features play a significant role in defining object boundaries compared to the widely generic, and very common, low-level features. Detailed experiment results are presented in this report.

1. Introduction

Segmentation is a sub-field of the computer vision broad spectrum of applications. It is a classification problem on a finer extent than that of the normal one. Instead of classifying cropped images and computing the probability of each containing a certain object, it performs classification on the pixel level by computing a probabilities vector for each pixel.

Two sub-tasks of segmentation are semantic and instance segmentation. While both perform classification of the image on the pixel level, in instance segmentation pixel level classification is accompanied by object detection to distinguish between different instances of the same category. This allows the distinction of two or more instances of the same class in an image, by representing each instance independently from the other instances, which makes instance segmentation more complex.

2. Datasets

The data used for fine tuning in the experiments are from the iSAID dataset, which is comprised of aerial images that are curated to depict real world challenging scenarios ranging from viewpoint to number of classes and even the distribution of data within those classes[10][12]. There are 15 total classes for objects , where each of the images of the dataset contains multiple instances from the same class and different classes within the same image. Images used are of high resolution where the original images widths range between 800 to 13000, with polygon bounding boxes, segmentation masks, and a total of 2806 high resolution images. Crops from the original images of sizes 800x800 are being used in the following experiments. There are other still unsolved difficulties with over-head images such as high density, arbitrary shapes, orientation, and scale variation, which makes instance segmentation on aerial datasets unique and challenging.

3. Instance segmentation method

The baseline model used is Mask R-CNN [2] which is an extension of Faster R-CNN specifically designed for instance segmentation. Faster R-CNN is a well-known object detection method and has proven to be effective in terms of speed thanks to region proposal networks (RPN) suggested by the authors of the paper [9]. Nowadays RPN is replaced by feature pyramid networks (FPN) [5] to further improve architecture of the model and refine its results. In Mask R-CNN, one more branch is added on each Region of Interest

Table 1. AP results of COCO pre-trained model

frz	AP	AP50	AP75	APs	APm	APl
1	32.72	54.67	34.28	17.34	39.82	50.34
2	32.62	54.93	33.87	17.56	40.01	49.52
3	32.08	54.42	33.04	17.06	39.18	47.57
4	28.59	49.34	29.41	13.87	35.87	45.07
5	26.77	45.32	28.25	15.73	33.34	31.36

(RoI) to predict segmentation masks next to original branch for classification and bounding box regression. Since Faster R-CNN is not an instance segmentation method, it does not classify in a pixel-to-pixel manner. This caused misalignment issue that gave birth to RoIAlign which can preserve precise spatial locations via adopting bi-linear interpolation.

4. Experiments

In the following report we conduct several experiments to study the effect of pretraining on different datasets on transfer learning, as well as how the proportion of fine-tuned layers to frozen layers affects the overall accuracy of the system. This study is conducted on instance segmentation for iSAID dataset. The architecture used to perform the segmentation is the detectron2 which is a large framework published online by FAIR (Facebook Artificial Intelligence Research) for general use by the public which provides a platform for multiple computer vision tasks from object detection, instance and panoptic segmentation as well as person keypoint detection[11].

4.1. Backbone

The main backbone we used is the ResNet101 with FPN. Experiments were conducted on that same backbone with different pretraining settings, one pretrained on ImageNet and another pretrained on COCO which are two of the largest available datasets where ImageNet contains almost 1000 object classes with 1,281,167 training images, 50,000 validation images and 100,000 test images [1] and COCO contains 330,000 images with 80 classes and 1.5 million object instances[6]. For each of the pretrained model we tried several freezing settings for the different layers of the backbone. The model we use here is one of the most famous CNNs which has established baselines for deeper training networks reaching a depth of 152 layers[3]. The model we use, ResNet101, is one of the deeper models and its convolutional blocks are divided into 5 layers, each containing specific number of convolutions. These layers are then frozen and the remaining are fine-tuned with the images from the dataset. Results are reported in Tables 1 and 2.

Table 2. AP results of ImageNet pre-trained model

frz	AP	AP50	AP75	APs	APm	APl
2	28.65	53.123	27.082	15.143	35.213	40.990
4	22.37	43.97	19.6	10.93	27.83	34.1
5	19.83	39.44	17.34	10.17	26.58	27.53

Table 3. AP per category of COCO pre-trained

category	frz = 2	frz = 4	frz = 5
ship	38.9	36.44	36.119
tennis	74.467	70.176	65.66
bridge	17.823	12.994	9.347
helicopter	4.264	3.659	10.016
soccer	43.703	33.838	21.46
storage tank	35.274	32.548	30.273
basketball	32.695	25.947	11.201
large vehicle	33.056	27.815	28.736
swimming pool	25.446	23.272	20.981
plane	46.319	41.314	51.555
baseball	43.349	41.711	29.833
ground	27.953	22.632	26.124
SV	12.095	10.347	11.655
roundabtt	27.73	27.131	21.364
harbor	26.283	19.16	27.329

Table 4. AP per category of ImageNet pre-trained

category	frz = 2	frz = 4	frz = 5
ship	35.87	29.567	30.513
tennis	71.85	62.793	58.714
bridge	15.796	10.066	8.646
helicopter	3.94	2.727	2.085
soccer	31.62	24.404	13.828
storage tank	31.56	28.917	28.255
basketball	27.69	16.615	12.684
large vehicle	28.3	18.785	19.248
swimming pool	25.64	18.804	16.942
plane	39.48	29.892	29.996
baseball	42.93	34.9536	26.509
ground	19.21	13.889	10.138
SV	11.08	7.75	8.233
roundabtt	25.78	25.243	22.861
n harbor	19.02	11.198	8.86

5. Results and Discussion

The results shown in tables 1 and 2 show that by freezing the same layer (Layer 2), the model that was pretrained on COCO performs much better than the one that was pretrained on ImageNet. Compared to the COCO pretrained model, the mean AP of the ImageNet pretrained model dropped by 3.97%. It is noticeable that for larger objects the AP drops by almost 9%. This is illustrated through the accuracies reported in 3 and 4 where there is a drop between 2% and 3% in AP, however, for larger objects like the soc-

cer playground the AP drops from 43.7% to 31.62% with almost 12% drop.

The observed results are reasonable considering the types of data the backbones had been trained on. The ImageNet pretrained model was pretrained on the classification task, while the COCO pretrained model was pretrained on the instance segmentation task. The COCO dataset is curated for instance segmentation, while the ImageNet dataset was curated for classification [7]. From here it can concluded that the downstream task that the model is being fine-tuned on is affected by the type of pretraining the backbone has gone through and this has been proven through multiple researches [4] [8].

It can be observed in Table 4, an overall trend is that AP decreases with more layers being frozen. Since the later layers contain richer and finer features which determine the decision boundary for segmentation of different objects. For some classes such as ship, large vehicle, plane, and small vehicle, freezing 4 layers yield an AP slightly more than that obtained with freezing all 5. That could be due to the fact that these categories have more number of instances in the dataset.

Table 3 illustrates class-wise AP outputs on COCO fine-tuned model, in which the AP decreases as well with increasing number of freezing layers. Hence, it can be said that the deeper the frozen layers, the lower the accuracy becomes, given that the hyperparameter values are not changed. Although, there is a possibility of a lower drop in accuracy with different hyperparameter settings such as the learning rate, momentum/weight decay etc. Still the learned complex features are very determinantal in the performance of the model on the downstream task, however, that lies outside the scope of the report.

6. Conclusion

In a nutshell, this report summarises the impact of using models pre-trained on two datasets (ImageNet and COCO) on a segmentation downstream task. The effect of freezing various numbers of layers on iSAID, an Aerial dataset for object detection and instance segmentation, was analyzed. From the results obtained, it was concluded that the model pre-trained on COCO yeilds better AP scores for instance segmentation of aerial images than the one pre-trained on ImageNet which refers to the nature of pretraining and how the dataset used for pretraining affects the downstream task. Finally, freezing deeper layers (layers deeper than Layer 2) resulted in a drop of accuracy since those layers contain complex and useful features that are determinantal in the performance of the model.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 2

[4] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty, 2019. 3

[5] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection, 2017. 1

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 2

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Zitnick. Microsoft coco: Common objects in context. volume 8693, 04 2014. 1, 3

[8] Alejandro Newell and Jia Deng. How useful is self-supervised pretraining for visual tasks?, 2020. 3

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. 1

[10] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37, 2019. 1

[11] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 2

[12] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1