Improvements in Object Detection

Atik Garg[1], Eric Liu[2], Simeon Gaumart[3]
EECS, National Chiao Tung University, Taiwan[1]
Electrical Engineering , National Chiao Tung University, Taiwan[2]
EECS, Insa Toulouse, France[3]

***Abstract*—This paper analyzes several augmentation methods to propose the optimized strategy for the improvement of classification accuracy in object detection. Data augmentation is one of the most important aspects of the training of neural network models. Even though it is accepted to improve object detection, yet data augmentation is still not properly understood and investigated. Given the large cost of annotation and labeling, the proper understanding of data augmentation has much greater importance for computer vision applications. Therefore, we propose an optimized way to implement several data augment strategies. Our experiments show that various data augmentation strategies affect the performance of object detection differently. Thus by investigating an optimized solution, we can further improve the impact of data augmentation on object detection performance. We will also investigate the Darknet-53 architecture and propose a better feature representation to reduce the information loss during downsampling.**

Keywords—component, formatting, style, styling, insert (*keywords****)***

I. INTRODUCTION

Neural networks are one of the most effective tools of a machine learning system to train a large amount of data for computer vision. And, object detection is one of the important aspects of computer vision as it covers various applications such as robot vision, human-machine interaction, dace recognition, driving, etc. In recent years the role of these applications is fueled further due to rapid improvement in various object detection strategies[1-5]. However, neural models that supported these rapid developments require large amounts of the well-annotated dataset for object detection. And it very arduous task due to the high labor cost and time required for annotation. This problem can be easily solved by efficiently generating more training images and their labels with the help of existing datasets in an effective way. To artificially increase the number of training samples from the existing dataset, several works in the past proposed to create better data augmentation strategies[6-8]. In the image domain, flipping, translating, rotating, etc. the training images are common practices, while many recent works also propose some handcrafted data augmentation methods[9-11].

Some of the latest works shown that manually design augmentation methods performed poorly as compared to the strategies which used an optimal learning policy[12-14]. However, most of these methods do not use a controlled strategy to apply different augmentation strategies. This inspired us to develop an optimized policy which can help in significant improvement of object detection performance. The proposed method generates learning data by selecting effective data augmentation methods and implement them in an optimized manner to get maximum benefit on the custom dataset. The method uses You Only Look Once v3 (YOLOv3)[15] for the detection and classification of objects in the image. Also, this paper modifies the Darknet-53 architecture which is used in YOLOv3. The modification improves the information loss suffered in the down-sampling layer. It replaces each down-sampling layer with down-sampling residual blocks to make the network easier to be optimized and avoid information loss along with an additional residual block in the root stage to enhance image information extraction.

II. RELATED WORK

This paper mainly focused on Data augmentation for object detection. Many the several past works used state of the art models such as MNIST, CIFAR 10, etc. to perform elastic distortions which affect the scale, translation, and rotation[8, 16, 17]. Apart from that image mirroring and random cropping are also commonly used augmentation strategies. Recently object-centric cropping, erasing, or adding noise to patches of images is used in many works to improve the accuracy and robustness of the model[18-21].

Although many of the above methods have worked on classification problems, we proposed an optimization strategy that will generate a specific number of training examples for the different augmentation strategies. This will help the model to efficiently train on the optimized amount of samples with high accuracy for object detection. As a contrast to classification, labeled data for object detection is more difficult because it is costlier to annotate detection data. Our objective is to use the validation set accuracy to help search for an optimized selection of augmentation procedures using custom operations that generalize across datasets, dataset sizes, backbone architectures, and detection algorithms.

III. METHODOLOGY

This paper considers the data augmentation search as a discrete optimization problem and optimizes for generalization performance.

Outline of the proposed method is as follows:

Step 1. Augment data by several data augmentation methods

Step 2. Confirm the effect of each data augmentation method

Step 3. Create an optimal dataset for object detection

First, the proposed method augments object detection data by various methods of data augmentation. Next, the data augmented by each data augmentation method are evaluated. The accuracy result of each model is obtained, whose model is trained with data added by the data augmentation by each augmentation method to the original data. Finally, the optimal dataset is obtained using

augmented data by some methods selected based on the results of Step 2.

The following subsections show the original dataset, augmented data, and the Steps in detail.

A. *Dataset used for Object Detection*

This paper customized the original dataset from the link[22]. It originally contains 80,000 images in 80 categories but we use the 1800 images for the data augmentation. Indeed, in our project, we need to train and test the model many times with different dataset (with different augmentation). However, the time needed for training YOLOv3 with the full coco dataset is around 3 days (for 150 epochs with a RTX 2080ti GPU). So, because of a lack of time, we decided to cut the coco dataset and use only 1800 images of coco, so the training time goes to 4 hours for the same number of epochs and same GPU.

However, we couldn't reduce the coco dataset with choosing 1800 random images. We have to choose our images so our cut dataset is balanced as the original coco dataset (for example 30% of label 0, 1% of label 1, etc.). As a matter of fact, the testing accuracy with such a small dataset will be really low, but we don't care of it because what is interesting us is the variation of the accuracy depending on the augmentation performed, with a dataset representing the best the full coco dataset (to be sure to get around the same results with the full coco dataset). The following table list the data augmentation types used in this paper

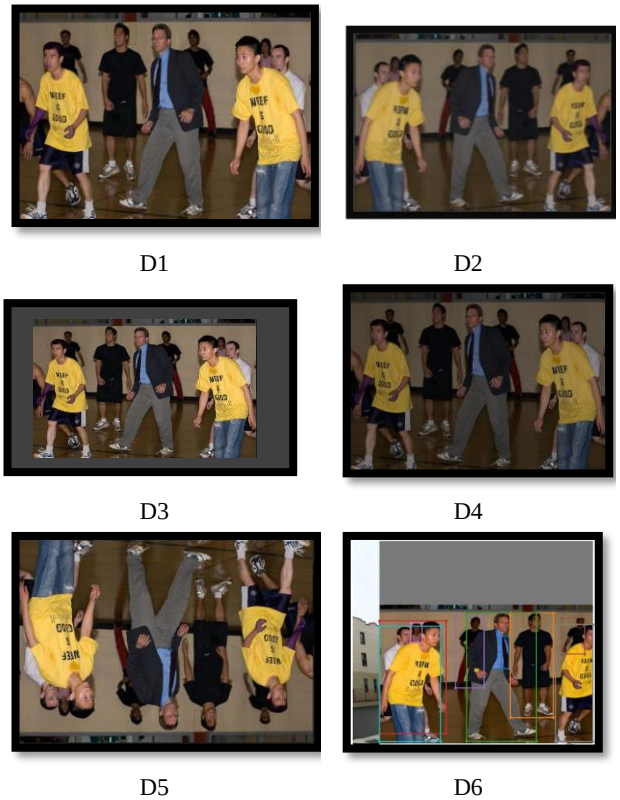

Figure 1Examples of Training Datasets

B. *Data Augmentation by Various Data Augmentation Methods*

In Step 1, this paper augments the training data by various data augmentation methods. The proposed method uses 6 different data augmentation methods individually. The particulars of each 6 augmentation methods are shown in Table 1.

Table 1 Details of Data Augmentation Strategies

Sr. No	Dataset	Amount	Dataset Name
1	COCO	1800	D1
2	COCO + Mirror	3600	D2
3	COCO + Resize	3600	D3
4	COCO + Brightness	3600	D4
5	COCO + Rotate	3600	D5
6	COCO + Mosaic	1800	D6

C. *Details of Implementation*

In step 2, the accuracy of the detection results is evaluated for each dataset with the help of the YOLOv3 model. After obtaining the performance profile of all the data augmentation strategies, we optimized all the augmentation method in the proposed approach to get maximum benefit. All the networks are tuned with the same hyper-parameters. The hyper-parameters of YOLOv3 for all datasets are shown in Table 2.

Table 2 Hyperparameters used for this paper

Batch Size	IOU Threshold	Optimizer	Learning Rate	Iterations
8	0.2	SGD	0.01	150

The proposed method selects the classes of the dataset which are augmented and selects effective data augmentation methods for each class independently. Thus, the selection of classes causes improving classification accuracy efficiently. The selection of effective augmentation methods prevents decreasing accuracy and increases it remarkably. In the starting, the classes of which data are augmented are determined by APs that are lower than the mAP. Then with the iteration of the highest mAP is augmented. Next, AP at the iteration with the highest mAP in the result with each dataset is confirmed in each selected class. Some datasets with higher APs are selected. The augmentation methods by which the datasets are augmented are determined as the selected data augmentation methods. After that, the data are augmented. The images including the chosen classes are selected randomly and are applied to the data augmentation method selected randomly from the selected methods. The number of augmented data for each class is determined to be as close as possible to the number of data of other classes.

To determine the additional factors, which affect the performance for object detection, we separately applied architecture proposed by Fei Gao et al[23].

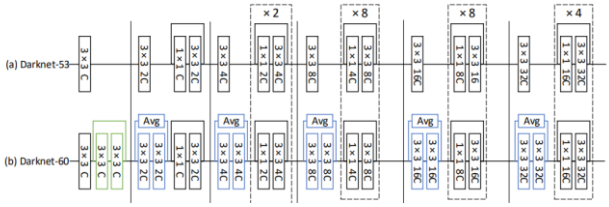

Figure 2 Darknet-53 vs. Darknet-60

As shown in Figure 2, Darknet-60 includes two types of modifications to improve the performance of Darknet-53, as follows. first, the original ResNet adopts a 7×7 convolution layer to extract features from input images. A useful but straightforward scheme is replacing the 7×7 convolutions with three 3×3 convolutions. So, a residual block in the root stage is added and is shown as a green block in Figure 2.

Second, a downsampling block is added to strengthen the gradient propagation in the network, the downsampling layer is replaced with a downsampling residual block, which is shown as blue blocks in Fig. 3(b). In the projection shortcut path of such residual block, a 2×2 average pooling layer with a stride of 2 is added before the 1×1 convolution layer, and the stride of 1×1 convolution is changed to 1. In comparison with the original downsampling block in ResNet, the improved structure proposed in [23] can avoid information loss in projection shortcuts.

IV. Experiment Results

To prove the effectiveness and potential of the proposed method.

Initially, the data of chosen classes are augmented based on the proposed method. The effect of augmentation can also be seen in Table 3, F1 scores of 400 test images are calculated. Scores in red and green indicate the lowest and highest score among all the augmentation as along with a comparison of baseline score.

Table 3 Selected classes and class prediction scores on YOLOv3 model

Score / Class	D1	D2	D3	D4	D5	D6
Person	.655	.613	.673	.656	.611	.649
Motorcycle	.387	.612	.382	.421	.28	.383
Airplane	.758	.76	.600	.494	.679	.495
Bus	.646	.554	.825	.721	.783	.710
Fire hydrant	.418	.195	.372	.264	.482	.433
Bottle	.376	.37	.333	.368	.381	.332
Teddy bear	.505	.433	.602	.349	.514	.601
Snowboard	0	0	0	.184	.231	0
Chair	.309	.338	.319	.349	.246	.33
Dining table	.201	.282	.244	.301	.350	.316
Average	.382	.362	.376	.353	.367	.406

In Table 3, various augmentation strategies are proving effective against individual classes even with a small set of 1800 images for the 80 categories. However, the average score is mostly less in all augmentation as compared to baseline because many times it's harmful to the model training as shown in Table 3. Therefore, in optimizing the augmentation procedure, we performed controlled augmentation for the chosen categories e.g. for the image with bus category we used to resize augmentation and for the fire hydrant category, we used rotate augmentation. In the cases, where we have several categories with conflicting augmentation then we choose the augmentation which classes are appearing the most e.g. if the image has two bottles, one chair, and one teddy bear then we will choose rotate augmentation but in case of a tie, we will don't use that particular image case for the augmentation. In the last, we performed mosaicking for all the augmented data.

In the final experiment, we again utilized the same dataset of 1800 images which is augmented to increase the dataset 'D7' size up to 2446 and performed training again on the YOLOv3 model.

A. *Objective Evaluation*

As shown in Table 4, dataset 'D7' has the highest average F1 score of .424 which is 11 percent as compared to an initial dataset of baseline score and also beating the score of the Mosaicking augmentation.

Table 4 Average scores of model training with and without augmentation

Score / Class	D1	D2	D3	D4	D5	D6	D7
Average	.382	.362	.376	.353	.367	.406	.424

B. *Visual Evaluation*

As shown in Figure 3, Controlled gives better detection results as compared to baseline training on same YOKLOv3 neural net

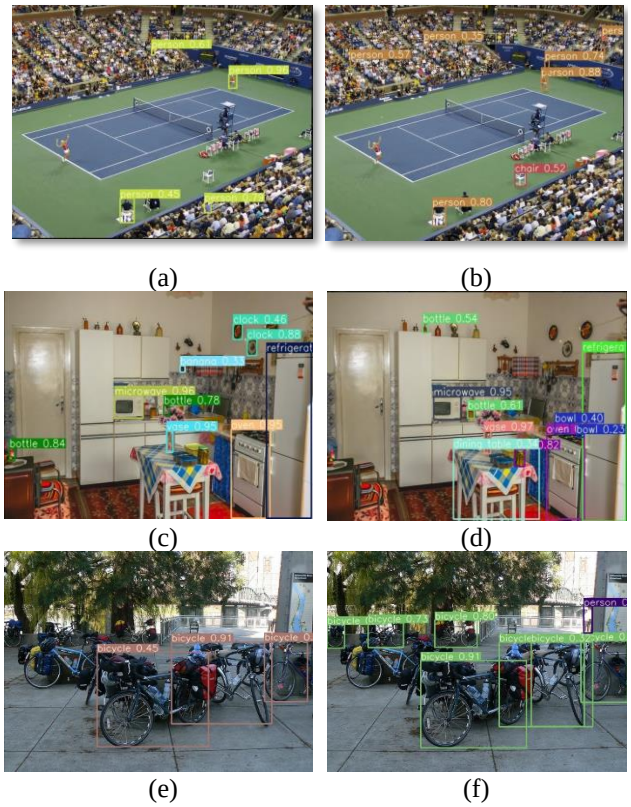

Figure 3 Comparison of baseline detection vs. controlled augmentation

V. Conclusion

This paper proposed a method for optimizing the data augmentation for the custom online dataset of a small set of 1800 images to improve the accuracy of detection and classification. The effective performance of each data augmentation method depends on the class. The five augmentation methods are tested for each class independently. The optimal learning dataset is generated by applying effective data augmentation methods. YOLOv3 trained using the generated dataset can obtain better results than that using the original dataset. The quantitative and visual evaluation of two YOLOv3s shows that YOLOv3 trained using the dataset generated by the proposed method can detect and classifies the objects with high accuracy as compared to individual or random augmentation approaches

VI. Future Work

We believe that we can further optimize the augmentation by including some more augmentation strategies such as superpixel augmentation, pixel removal, etc. Also, we can implement and improve the architecture of Darknet-53 which we currently unable to perform due to more time is required

Acknowledgment *(Heading 5)*

This work is performed as the final project for the Machine Learning course in under the supervision of Prof. Chun Shu Wei, Assistant professor at National Chiao Tung University, Taiwan

References

[1] Y. Amit, P. Felzenszwalb, and R. Girshick, "Object detection," *Computer Vision: A Reference Guide*, pp. 1-9, 2020.

[2] D. R. High, M. D. Atchley, K. Kay, R. C. Taylor, and D. C. Winkle, "Shopping facility assistance object detection systems, devices and methods," ed: Google Patents, 2020.

[3] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12697-12705.

[4] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980-2988.

[5] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3588-3597.

[6] D. M. Montserrat, Q. Lin, J. Allebach, and E. J. Delp, "Training object detection and recognition CNN models using data augmentation," *Electronic Imaging*, vol. 2017, no. 10, pp. 27-36, 2017.

[7] N. McLaughlin, J. M. Del Rincon, and P. Miller, "Data-augmentation for reducing dataset bias in person re-identification," in *2015 12th IEEE International conference on advanced video and signal based surveillance (AVSS)*, 2015: IEEE, pp. 1-6.

[8] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *arXiv preprint arXiv:1708.04896*, 2017.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[11] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9.

[12] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.

[13] A. J. Ratner, H. Ehrenberg, Z. Hussain, J. Dunnmon, and C. Ré, "Learning to compose domain-specific transformations for data augmentation," in *Advances in neural information processing systems*, 2017, pp. 3236-3246.

[14] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 113-123.

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

[16] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Icdar*, 2003, vol. 3, no. 2003.

[17] I. Sato, H. Nishimura, and K. Yokoi, "Apac: Augmented pattern classification with neural networks," *arXiv preprint arXiv:1505.03229*, 2015.

[18] R. G. Lopes, D. Yin, B. Poole, J. Gilmer, and E. D. Cubuk, "Improving robustness without sacrificing accuracy with patch gaussian augmentation," *arXiv preprint arXiv:1906.02611*, 2019.

[19] N. Ford, J. Gilmer, N. Carlini, and D. Cubuk, "Adversarial examples are a natural consequence of test error in noise," *arXiv preprint arXiv:1901.10513*, 2019.

[20] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[21] W. Liu *et al.*, "Ssd: Single shot multibox detector," in *European conference on computer vision*, 2016: Springer, pp. 21-37.

[22] [Online]. Available: http://cocodataset.org/#download.

[23] F. Gao, C. Yang, Y. Ge, S. Lu, and Q. Shao, "Dense Receptive Field Network: A Backbone Network for Object Detection," in *International Conference on Artificial Neural Networks*, 2019: Springer, pp. 105-118.