Improved Mosaic: Algorithms for more Complex Images

Improved Mosaic: Algorithms for more Complex Images

Wang Hao[1,*], Song Zhili[2]

[1] College of Computer Science and Information Engineering, Shanghai Institute of Technology, ShangHai, 200000, China

[2] College of Computer Science and Information Engineering, Shanghai Institute of Technology, ShangHai, 200000, China

* Corresponding author's e-mail: 1003818796@qq.com

Abstract: Data augmentation plays a vital role in deep learning, and image augmentation, as an important part of target detection and image classification, significantly improves the performance of the algorithm. The Mosaic data augmentation algorithm in YOLOv4 randomly selects 4 pictures from the train set and puts the contents of the 4 pictures into a synthetic picture that is directly used for training. This data augmentation method can improve the model's recognition ability in complex backgrounds. In this paper, we improve the Mosaic data augmentation algorithm. After analyzing the synthesized picture area, it is divided into irregular grids, and a certain number of training set pictures are filled randomly, which further improves the synthesis ability and achieves a synthesis picture that can accommodate 6 and 9 training set pictures. After basic image processing methods such as zooming, flipping, and color gamut transformation, the model's recognition ability under complex backgrounds is improved, and the accuracy of identifying small targets is improved. During the batch normalization operation, the data of 6 or 9 pictures can be calculated at the same time, which makes the model hyperparameter mini-batch need not be set very large, which reduces the GPU memory requirements.

1. Introduction

Data augmentation[6] has always been an integral part of all fields of intelligent algorithms[7]. It plays an important role in object detection[2], image classification[8], and natural language processing[10]. Before entering the era of big data, the data collective amount that can be used by the intelligent model algorithm is still not up to the scale required by the algorithm. In this case, the trained model is prone to overfitting[4]. Therefore, how to get more image data in the only data set is the original intention of designing the data augmentation algorithm.

With the development of data augmentation algorithms, some data augmentation algorithms for single image processing have appeared successively.For example,image rotation[16], image flipping, image zooming[11], image clipping[12], image color transformation[13], image Gaussian noise and so on. However, these methods only change an image, only increase the number of trainable images, and do not provide a more complex and effective training background for the model. As a result, the efficiency of the model's recognition of complex background targets is far lower than that of a simple background. At the same time, in the process of the training model, the single image would lead to the large setting of mini-batch to meet the requirements of efficient training. However, the large mini-batch is a challenge to GPU memory and even cannot be trained under a single GPU, which is difficult for researchers with insufficient computing power to accept.

Now, under the tide of the era of big data, the volume of data set has reached an unprecedented scale, and researchers will not stop for lack of data. Current data augmentation algorithms tend to provide more complex images to cope with multi-level, multi-target scenarios. It is on this basis that the algorithm proposed in this paper further improves the image complexity and fills in the data image by randomly dividing the irregular grid of the blank image, then the composite image can be obtained in the Mosaic of the simple image, giving the model more abundant training background. In the batch normalization[9] operation, the data of 6 or 9 images can be calculated at the same time, which means that the super-parameter mini-batch does not need to be set too large to enable the model to be trained efficiently. In this way, the target detection algorithm can be trained under a single GPU and the calculation cost can be saved.

2. Related work

2.1. Analysis of simple data augmentation algorithm

From the analysis of the existing data augmentation algorithms, it can be seen that although most of the data augmentation algorithms only operate on a single image, there are many methods for image transformation, which can also temporarily solve the dilemma of insufficient data collective. For example, image rotation, gamut transformation, scaling and cropping. Theoretically speaking, the range of rotation angles can be set in an infinite number, and the selection of rotation centers can also be an infinite number. Gamut transformation can also reach a large scale according to different gamut combinations. Although there are only two directions of scaling operation, the scale ratio can be adjusted freely. The scaled part will be adjusted to the original size. There are also a number of options for the direction of the cropping operation, and the rest is restored to the original size after the cropping. The training sets derived from these data augmentation algorithms play an important role in the early training of intelligent models[15] because these operations can meet the requirements of intelligent models on the data collective. The following figure shows an example of both operations.

Figure 1. From left to right, the original image, zooming out 10%, and zooming out 20%.

Figure 2. Original image, upper left clipping, and lower right clipping from left to right.

Due to the continuous development of the target detection model, the application scenarios it faces gradually change from simple background to complex background. However, the performance of the object detection model in complex background is not good, and the performance improvement of the

model algorithm by only increasing the data collective quantity is already very small. In order to improve the performance of the model, it is necessary to collect data sets with a complex background in reality. However, it is not easy to collect data sets and standardize the annotation. Therefore, we hope to get more complex training data in the existing simple data sets, so as to replace the collection work.

2.2. Mosaic data augmentation

The Mosaic data augmentation algorithm in YOLOv4[1] refers to the CutMix[3] data augmentation algorithm, which is a further extension of the CutMix data augmentation algorithm. General methods of data augmentation are image flipping, gamut transformation, scaling, and other operations on an image, while CutMix data augmentation is to splicing two images and transfer the spliced images directly to the neural network for training. Mosaic data augmentation[5] algorithm made use of four images for mosaics to form a composite image containing four original images, which improved the efficiency of model training. At the same time, the Mosaic data augmentation algorithm has different choices compared to the number of objects contained in one original image. It can train many different objects in the same composite image, which cannot be achieved in the original data enhancement algorithm. Here's the composite image from running the Mosaic data augmentation algorithm on the VOC2007 dataset.

Figure 3. Unimproved Mosaic data augmentation algorithm.

The Mosaic data augmentation algorithm locates the upper-left corner, upper-right corner, lower-left corner, and lower-right corner of an image, and then place the basic data augmentation processing on the four corners respectively. The learning harvest of one composite image is four times that of the original one. In the real task, the target object is not always presented in the image completely, and sometimes it is blocked or incomplete. You can see that the Mosaic data augmentation algorithm generated a lot of images, but it didn't show the full outline of the object. The model trained with these images can systematically learn the object with unknown full contour. The object detection model can identify the location and type of object by part of the object contour.

2.3. Improved Mosaic data augmentation algorithm

In this paper, based on the Mosaic data augmentation algorithm in YOLOv4, the following improvements were made:

(1) Considering that the goal of the object detection model is to accurately identify more objects if the number of targets in the composite image can be continuously improved, then the number of targets in a model training will increase. Through analysis, it can be known that too much image synthesis in an image will not give more features to the model, but will reduce the overall performance of the target.

(2) Grid division This algorithm is designed in the form of 3×2, 2×3, and 3×3. Combined with the discussion in (1), we can know that the number of grids is consistent with the number of original images. Therefore, when considering how to divide the grids, we should consider that the number of grids should not be too large, and the number of grids should not be too small. Otherwise, the algorithm will degenerate into a Mosaic data augmentation algorithm in YOLOv4 or CutMix algorithm. Finally, this algorithm divides the grid into 6 grids and 9 grids. This gives you the three grid layouts shown in the figure below.

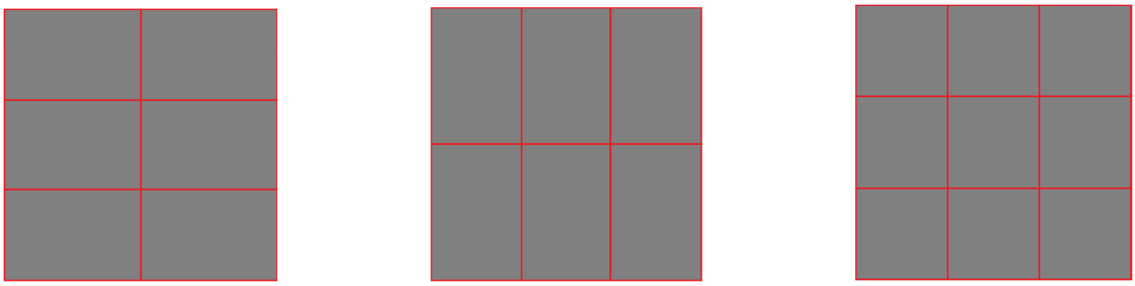


Figure 4. Improved Mosaic data augmentation algorithm three grid layouts.

(3) After determining the grid layout, the next step is to fill in the original image within the grid. Considering how to fill to facilitate the implementation of the algorithm, this algorithm decides to take counterclockwise as the order of filling the original image and fills the upper left corner of the original image with basic data augmentation after it is aligned with the upper left corner of the grid. Divide the gray background plate equally, mark the coordinate points equally, and unify the upper left corner of the grid as the image synthesis point. Paste the original image into the grid, and the part of the image beyond the grid will not be displayed. Figure 5 shows the filling order as an example. The yellow dots represent the coordinates in the upper left corner of the grid, the numbers in each grid represent the filling order, starting with the number 1 as the original image sequence number of the first fill, and so on.

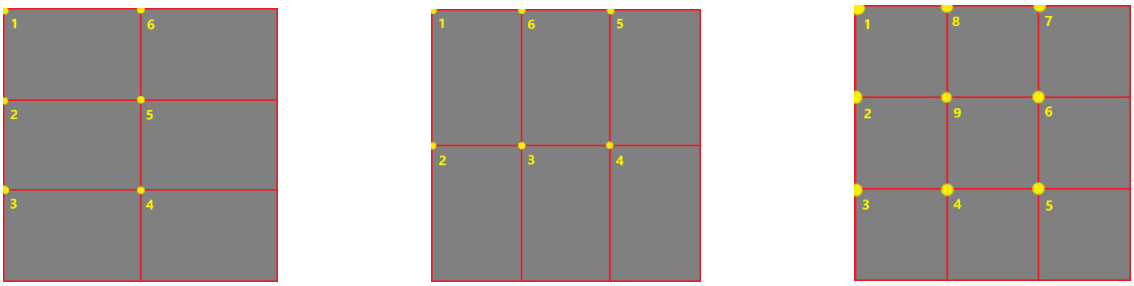


Figure 5. Filling order of the three grid layouts.

(4) Considering that the change of image size will affect the effect of image synthesis, this algorithm sets upper and lower limits for the change amplitude, and the image amplification amplitude should not be too large. The data appearing in the grid should be effective because it is meaningless for only a few pixels to participate in the model training. The image cannot be reduced too much, and the data in the grid needs to be a feature that the model can learn, that is to say, try to meet the features that can appear in each grid.

(5) When the size of the original image changes, such as zooming in or out, the Ground Truth marking the target position will also change. Therefore, when the object changes, the real box will also change. The adjustment of this algorithm on the real box will also change with the target. Also, when the composite image is finally generated, the Ground Truth of those objects that can hardly be seen in the figure should be considered and specified that when the Ground Truth meets in the figure:

$$\begin{cases} xmax - xmin < m \\ ymax - ymin < n \end{cases} \tag{1}$$

When the inequality is present, the border is no longer considered, and the object is not learned as a target. (1) In the formula, xmax represents the maximum abscissa of Ground Truth, xmin represents the minimum abscissa of Ground Truth, ymax represents the maximum ordinate of Ground Truth, ymin represents the minimum ordinate of Ground Truth, m and n represent the threshold displayed in the box, both of which are 5 in this algorithm. As shown in the figure below, taking the VOC2007 dataset as an example, the 2×3 grid layout shows the role of box display threshold in the algorithm.

(a) (b) (c) (d)

Figure 6. The role of thresholds inbox display.

It can be seen from (a)(b) and (c)(d) in Figure 6 that the plane on the left appears a bit of tail in the composite image, but it is not framed by Ground Truth, neither is the computer on the right.

3. Results & Discussion

3.1. Changes in the grid layout

Compared with the previous CutMix algorithm based Mosaic algorithm, the improved Mosaic algorithm has a big change in the number of grids. The CutMix algorithm has only 2 grids, while the basic Mosaic algorithm has 2×2 grids. The improved Mosaic algorithm has three grid layouts, which are in the forms of 3×2, 2×3, and 3×3 respectively. Compared with the former two algorithms, the improvement in the number of grids in 3×2 and 2×3 grid layout is 200% and 50% respectively, and the improvement in the number of grids in the 3×3 grid layout is 350% and 125% respectively.

Table 1. Comparison between improved Mosaic algorithm and CutMix algorithm.

	The grid number	Increase rate of grid number
CutMix	2	reference
Improved Mosaic 2×3 grid layout	6	200%
Improved Mosaic 3×2 grid layout	6	200%
Improved Mosaic 3×3 grid layout	9	350%

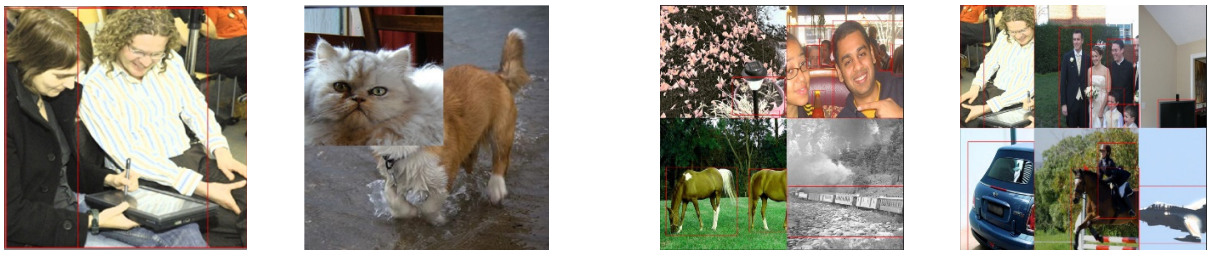
Table 2. Comparison between improved Mosaic algorithm and Mosaic algorithm.

	The grid number	Increase rate of grid number
YOLOv4 Mosaic	4	reference
Improved Mosaic 2×3 grid layout	6	50%
Improved Mosaic 3×2 grid layout	6	50%
Improved Mosaic 3×3 grid layout	9	125%

Through the form data can see, the Mosaic data augmentation algorithm of the improved compared with the Mosaic algorithm and CutMix algorithm, the former can expand training image vision, let a mini-batch[17][18] image features that are included in the amount of data, it can not only increase the trainiing speed, can also reduce the training cost, but higher requirements for calculating the object detec-tion algorithm can also be trained on the lower hardware. For example, on the GPU platform, RTX 2060S video memory was 6G, and the VOC2007 data set was used as a training set to avoid the overflow of video memory into mini-batch. However, the same training data could be obtained in smaller mini-batch using the improved Mosaic data augmentation algorithm.

3.2. Improvement in the number of targets and higher background complexity

The purpose of the object detection algorithm is to accurately detect more objects. This algorithm is compared with the effect of the ordinary data augmentation algorithm. In terms of the image provision of the training set, the latter is still that each image only represents the content of one original image, so it is low in the complexity of the image background. Secondly, the CutMix algorithm is compared with the general data augmentation algorithm. Although the CutMix algorithm is improved compared with the ordinary data augmentation algorithm, the increase is only one more grid data. Finally, the algorithm and the Mosaic YOLOv4 data augmentation algorithm based on the comparison, as a result of the 2×2 hold four original image grid can at most, and the improved Mosaic data augmentation algorithm can accommodate up to 6 and 9 pieces of the original image, the features of the object detection model learn richer, in dealing with complex background would have been better recognition task.



(a) (b) (c) (d)

Figure 7. 4 Comparison of the four data augmentation algorithms: (a) The generated graph enlarged by the ordinary data augmentation algorithm to 30% outward; (b) the generated graph was generated by CutMix algorithm; (c) the generated graph was generated by Mosaic algorithm in YOLOv4; (d) the generated graph was generated by the improved 2×3 grid layout Mosaic algorithm.

From the four images, it can be seen that the image generated by the common data augmentation algorithm is just a simple background image without more complex features. However, the Mosaic algorithm and CutMix algorithm have improved the background complexity of the composite image to some extent, but they are still not complex enough. It can be seen that the algorithm proposed in this paper has pushed the background complexity to a new height, thus it can be seen that the object detection[14] model trained by the improved Mosaic data augmentation algorithm will perform better in the face of objects with complex background.

3.3 Constraints on Ground Truth are added

When the size of the original image changes, such as zooming in or out, the Ground Truth marking the target position will also change. Therefore, when the object changes, the real box will also change. The adjustment of this algorithm on the real box will also change along with the target. The following inequality group is the coordinate processing constraint conditions of the Mosaic data augmentation algorithm:

$$\begin{cases} ymin > cuty \\ xmin > cutx \\ ymax - ymin < m \\ xmax - xmin < n \end{cases} \tag{2}$$

In equation (2), xmax and ymax represent the maximum abscissa and vertical ordinate of Ground Truth, xmin and ymin represent the minimum abscissa and vertical coordinates of Ground Truth, cutx and cuty represent the equiequinox of the horizontal and vertical axes of the background image, m and n are the box display thresholds, and the unit is pixels. In the Mosaic algorithm, 5 is selected.

In the improved Mosaic data augmentation algorithm, taking the grid layout as 2×3 as an example, constraints on Ground Truth were added in the processing of the third image and the sixth image, so that the Ground Truth of the object on the composite image could be kept intact. In addition, different from the Mosaic data augmentation algorithm, when judging the size of Ground Truth, this algorithm doubled

the box display threshold, which eliminated more useless information and increased the training speed of the object detection model.The coordinate processing of the two images satisfies the following inequalities respectively:

$$\begin{cases} ymax < cuty \\ xmax < cutx1 \\ xmin > cutx2 \\ ymax - ymin < M \\ xmax - xmin < N \\ xmax - xmin < K \end{cases} \tag{3}$$

$$\begin{cases} xmax < cutx1 \\ ymin > cuty \\ xmin > cutx2 \\ xmax - xmin < M \\ ymax - ymin < N \\ xmax - xmin < K \end{cases} \tag{4}$$

In equation (3) and (4), xmax and ymax represent the maximum abscissa and ordinate value of Ground Truth, while xmin and ymin represent the minimum abscissa and ordinate value of Ground Truth. cuty represents the equal point of the background image on the vertical axis. cutx1 and cutx2 represent the two equal points on the horizontal axis of the background image. M, N and K represent the display threshold of the box, and the unit is pixel, this algorithm takes 10.

4. Conclusions

This paper provides a better target detection data augmentation algorithm, which is better than the basic Mosaic algorithm and the CutMix algorithm in providing complex background images. This algorithm can not only improve the ability of the target detection model to recognize objects in complex background, but also save GPU memory and enable people with poor computing power to use it. These characteritics make it possible to widely use the improved Mosaic data augmentation algorithm.

Acknowledgments

This thesis is completed under the kind care and careful guidance of my tutor, Mr. Song Zhili. His serious scientific attitude, rigorous academic spirit, and excelsior working style deeply infected and inspired me. Mr. Song not only gave me careful guidance in my studies but also gave me meticulous care in my thoughts and life. Here I would like to express my sincere gratitude and high respect to Mr. Song. It is better to teach a man to fish than to teach him to fish. I learned to accept new ideas and to think independently. I would also like to thank my girlfriend(Lou Yaqin) who encouraged me silently behind my back. Thanks to your help and support, I was able to overcome difficulties and doubts one by one until the successful completion of this article.

References

[1] Alexey Bochkovskiy*, Chien-Yao Wang*, Hong-Yuan Mark Liao, (2020) YOLOv4: Optimal Speed and Accuracy of Object Detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Online virtual hosting.pp.126-138.

[2] Jiwoong Choi, Dayoung Chun, Hyun Kim, and Hyuk-Jae Lee.(2019) Gaussian YOLOv3: An accurate and fast object detector using localization uncertainty for autonomous driving. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Los Angeles, USA.pp. 502–511.

[3] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo, (2019). CutMix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) . Seoul, Korea. pp. 6023–6032.

[4] Wyk, A. B. V , Engelbrecht, A. P,(2010) Overfitting by PSO trained feedforward neural networks[C].In: Evolutionary Computation. IEEE.In China.pp.155-163.

[5] Alexey Bochkovskiy*,Chien-Yao Wang*,Hong-Yuan Mark Liao,(2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Online virtual hosting. pp.163-178.
[6] Wu, L. Y. N, (1999). Parameter expansion for data augmentation. In: Publications of the American Statal Association, 94(448), 1264-1274.
[7] Suk H I , Shen D, (2013). Deep Learning-Based Feature Representation for AD/MCI Classification[C]. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Med Image Comput Comput Assist Interv. pp.263-280.
[8] Zhou X, Yu K, Zhang T, et al, (2010). Image Classification Using Super-Vector Coding of Local Image Descriptors[C]. In: Computer Vision-eccv -european Conference on Computer Vision. DBLP.
[9] Nadeemhashmi S , Gupta H , Mittal D , et al,(2018). A Lip Reading Model Using CNN with Batch Normalization[C]. In: 2018 Eleventh International Conference on Contemporary Computing (IC3). IEEE Computer Society.In Chengdu, China. pp.235-245.
[10] Ibrahim M, Ahmad R, (2010). Class Diagram Extraction from Textual Requirements Using Natural Language Processing (NLP) Techniques[C]. In: 2010 Second International Conference on Computer Research and Development. IEEE. pp.598-610.
[11] N. Shezaf, H. AbramovSegal, I. Sutskover,(2002). Adaptive low complexity algorithm for image zooming at fractional scaling ratio[C]. In: IEEE. pp.2203-2253.
[12] Ciocca G, Cusano C, Gasparini F, et al, (2007). Self-Adaptive Image Cropping for Small Displays[C]. In: International Conference on Consumer Electronics. IEEE.
[13] Lin D, Tang X, (2013). Coupled space learning of image style transformation[C]. In: Tenth IEEE International Conference on Computer Vision. IEEE. pp.235-245.
[14] Jiwoong Choi, Dayoung Chun, Hyun Kim, and Hyuk-Jae Lee,(2019). Gaussian YOLOv3: An accurate and fast object detector using localization uncertainty for autonomous driving. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Seoul, Korea. pp. 502–511.
[15] Redmon J , Farhadi A, (2017). YOLO9000: Better, Faster, Stronger[C]. In: IEEE Conference on Computer Vision & Pattern Recognition. IEEE, pp. 6517-6525.
[16] Islam S M M , Debnath R , Hossain S K A, (2007). DWT Based Digital Watermarking Technique and its Robustness on Image Rotation, Scaling, JPEG compression, Cropping and Multiple Watermarking[C]. In: International Conference on Information & Communication Technology. IEEE Xplore. pp.235-245.
[17] Andrew Cotter, Ohad Shamir, Nathan Srebro, (2011). Better mini-batch algorithms via accelerated gradient methods[C]. In: International Conference on Neural Information Processing Systems. Curran Associates Inc. pp.2333-2356.
[18] Feyzmahdavian, Hamid Reza, Aytekin, Arda, Johansson, Mikael, (2016). An asynchronous mini-batch algorithm for regularized stochastic optimization[C]. In: IEEE Conference on Decision & Control. IEEE. pp.265-285.