SUPPLEMENTARY MATERIAL 2

The Interpretation of the Fast Region-Based Convolutional Neural Network and You Only Look Once Models and the Architecture and Implementation of the Convolutional Neural Network Model

Current mainstream depth-based target detection algorithms mainly fall into two categories: two-stage and one-stage detection algorithms. The function of the first stage of a two-stage algorithm is to identify candidate regions in the target image, and the function of the subsequent stage is to classify the candidate regions. Typical two-stage algorithms include the region-based convolutional neural network (R-CNN) (1), Fast R-CNN (2), and Faster R-CNN (3), among others. In contrast, a one-stage detection algorithm does not include a stage in which the candidate region is generated but directly generates the class probability and spatial coordinates of the target. You only look once (YOLO) (4-6) and single-shot multibox detectors (7) are representative one-stage algorithms.

This study used two target detection methods: one based on the Faster R-CNN ResNet-101 object detection Application Programming Interface developed by Google (8), and the other based on the YOLOv3 model using the Darknet-53 network structure (6). The former was a two-stage algorithm. First, feature extraction was performed using a ResNet-101, which was pre-trained using the ImageNet dataset (available at http://image-net.org/) (9). Second, a region proposal network (3) was used to generate a bounding box. In addition, the suggestion box was mapped to the feature map of the last layer of the convolutional neural network (CNN). A fixed-size feature map was generated by the region of interest pooling layer. Finally, joint training of classification probabilities and border regressions was implemented using softmax loss and smooth L1 loss. YOLOv3 is the most advanced one-stage algorithm (6). The feature extraction framework used Darknet-53 with its convolution weight darknet 53. conv.74 pre-trained using the ImageNet dataset (9). The local feature interaction of the network was then implemented in the YOLO portion. Based on this, classification and position regression were performed.

We adjusted the intersection over union (IoU) threshold of the foreground target from a normal value of 0.7 to 0.5, which resulted in additional foreground samples in each minibatch. In addition, an online hard example mining algorithm was used to screen difficult samples and guide the model optimization in the appropriate direction (10). The number of epochs and learning rate were set to eight and 0.0001, respectively, and we applied the stochastic gradient descent optimizer. We used the NVIDIA GTX1080 (NVIDIA, Santa Clara, CA, USA) graphics card and the model training was run using four graphics processing units (GPUs). The GPU version of MXNet (version 1.1.0, available at https://github.com/apache/incubator-mxnet) was used.

REFERENCES

1. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation [updated October 2014]. Cornell University, 2013. Available at: https://arxiv.org/abs/1311.2524. Accessed June 21, 2019
2. Girshick R. Fast R-CNN [updated September 2015]. Cornell University, 2015. Available at: https://arxiv.org/abs/1504.08083. Accessed June 21, 2019
3. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks [updated January 2016]. Cornell University, 2015. Available at: https://arxiv.org/abs/1506.01497. Accessed June 22, 2019
4. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection [updated May 2016]. Cornell University, 2015. Available at: https://arxiv.org/abs/1506.02640. Accessed June 22, 2019
5. Redmon J, Farhadi A. YOLO9000: better, faster, stronger. Cornell University, 2016. Available at: https://arxiv.org/abs/1612.08242. Accessed June 22, 2019
6. Redmon J, Farhadi A. YOLOv3: an incremental improvement. Cornell University, 2018. Available at: https://arxiv.org/abs/1804.02767. Accessed June 22, 2019
7. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: single shot multiBox detector [updated December 2016]. Cornell University, 2015. Available at: https://arxiv.org/abs/1512.02325. Accessed June 23, 2019
8. Huang J, Rathod V, Sun C, Zhu M, Korattikara A, Fathi A, et al. Speed/accuracy trade-offs for modern convolutional object detectors [updated April 2017]. Cornell University, 2016. Available at: https://arxiv.org/abs/1611.10012. Accessed June 21, 2019
9. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. *ImageNet: a large-scale hierarchical image database.* 2009 IEEE conference on computer vision and pattern recognition;2009 June 20-25;Miami, USA
10. Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining. Cornell University, 2016. Available at: https://arxiv.org/abs/1604.03540. Accessed June 23, 2019