Computer Version Symposium

Zhongqing Chen
ID: 21821324
bryceqing@zju.edu.cn

Abstract

The symposium is focus on a classic computer version topic — detection. The report will unscramble several papers about detection from cvpr2019 and cvpr2018. Undoulely, these paper mostly use CNN to achieve their goals.

1. Introduction

The first is about Single shot aspect detection , which offers an algorithm to detect the object from CMU.[1] The second is about high-precision target detector named Cascade R-CNN that is submited in CVPR2018.[2] In this essay, there are some new methods that you can use in feature work if you want to improve the accuracy of detection. The third is focus on finding tiny faces in the wild with generative adversarial network.[3] The fourth paper mainly talk about domain adaptive faster R-CNN for object detection in the wild.[4] The last paper is about Person Re-identification by Transfer Learning of Spatial-Temporal Patterns with unsupervised Cross-dataset.[5]

2. Background

Nowadays, there are so many methods that can be used to detect objects. With the development of deep–learning , the accuracy of detection has improved greately. In recent years, almost all the papers at CVPR about detection use deep learning to arrivate the perfect effect. CMU submit a new method that can be used for Single-Shot Object Detection.

3. Topic

3.1. Single-Shot

3.1.1 Background

Object scale problem in target detection has always been a difficult problem to solve. So far, it has mainly relied on network structure design, loss function, training method and other aspects to alleviate the troubles caused by scale. Especially for small object detection, there is no good solution now. Among these methods, the most common non-Feature Pyramid Network (FPN) is a multi-level feature map to predict objects of different sizes, with high-level features with high-level semantic information and large receptive fields. Detecting large objects, shallow features with low-level detail semantic information and small receptive fields, suitable for detecting small objects. The FPN gradually integrates deep and shallow features, which gradually increases the high-level semantic information of the shallow features to improve the feature expression and enhance the detection effect. Thanks to the performance enhancements brought by its powerful feature representation capabilities, the FPN architecture has become a standard component of the inspection framework. This article focuses on how to ”feature” the feature to detect objects, and is aimed at a model of single-stage. But in regular ,single-stage does not have the operation of roi pooling. So it's a problem to how to choose the feature.

3.1.2 Method FASF

The article proposes that the FSAF(Feature Selective Anchor-Free Module) module allows each instance to automatically select the most appropriate feature. In this module, the size of the anchor box no longer determines which features to select for prediction, that is, the anchor (instance) size becomes an irrelevant variable. That is the origin of anchor-free. Therefore, the basis of the feature selection is that the original instance size becomes the instance content, and the model auto-learning selection feature is implemented. The structure of FASF is Figure1 . The paper submit the FASF based on RetinaNet with adding branch which can parallel with classification subnet and regression subnet. In particular, FSAF can be integrated into other single-stage models, such as SSDs, DSSDs, and more. The FSAF is designed to automatically select the best feature that is determined by each feature level. The author analyzes the necessity of anchor-free in the abration study section, the importance of the online feature selection, and whether the selected feature level is optimal. It also pointed out that FSAF is very robust and efficient, and has stable rise points under various backbone conditions.

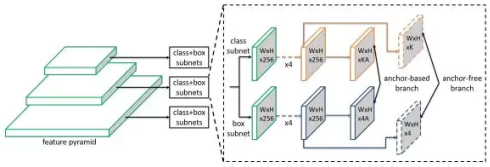

Figure 1: Network architecture of RetinaNet with FASF module.

3.1.3 Conculsion

The article designed a new FSAF module from the perspective of feature selection to improve performance. From the perspective of loss, the efficiency of gradient back-transmission is improved. It is similar to SNIP, only updating the gradient corresponding to objects in a specific scale, but SNIP is different and the efficiency is higher than SNIP.

3.2. Cascade R-CNN

3.2.1 Background

In object detection, an intersection over union (IoU) threshold is required to define positives and negatives. An object detector, trained with low IoU threshold, e.g. 0.5, usually produces noisy detections. However, detection performance tends to degrade with increasing the IoU thresholds. Two main factors are responsible for this:

- Overfitting during training .
- Inference-time mismatch between the IoUs for which the detector is optimal and those of the input hypotheses.

3.2.2 Method Cascaded RCNN

Cascaded RCNN achieves the goal of continuously optimizing prediction results by cascading several detection networks. Unlike common cascading, several detection networks of cascade R-CNN are trained on positive and negative samples determined by different IOU thresholds. The output of a detection model is used as the input of the latter detection model, and therefore is the training method of the stage by stage, the IOU threshold for defining the positive and negative samples is constantly rising like Figure2

3.2.3 Conclusion

Most of the cascade R-CNN experiments were done in the COCO dataset, and the results were very good.

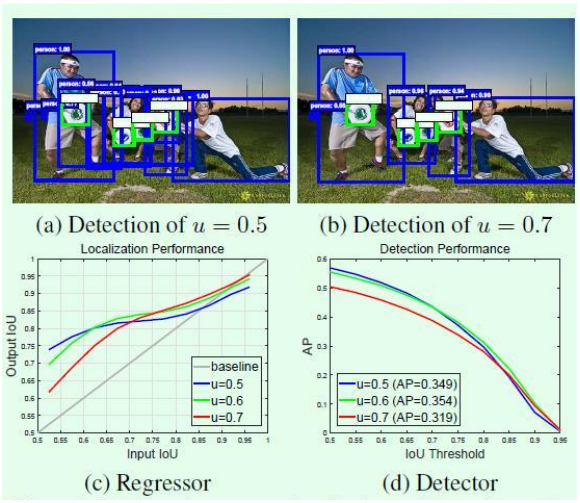

Figure 2: The detection outputs,localization and detection performance of object detectors of increasing IoU threshold u.

3.3. Tiny faces find

3.3.1 Background

Minimal target detection is a very challenging problem in the current target detection field and a key factor that constrains the current detector ranking on various lists. This problem is not limited to face detection, but it is particularly prominent in the field of face detection, especially in the real scene, the scale of face targets is more extensive. As we all know, the difficulty of detecting very small targets is that very small targets appear as small size and low resolution on the image, so the lack of highly discriminative information makes it difficult for the detector to distinguish them from the background.

3.3.2 Method GAN

Generative Adversarial Nets (GAN) has achieved many successful applications in image synthesis (such as DCGAN). One of the successful applications is super-resolution (SR), published in CVPR2017. Using GAN has made a major breakthrough in the performance of super-segment reconstruction.The progress of detection is Figure 3

3.3.3 Conclusion

This paper successfully use GAN to generate very small face images into high-resolution face images, and the effect of generation is more realistic than the existing methods.In addition, it successfully applied the idea of GAN to

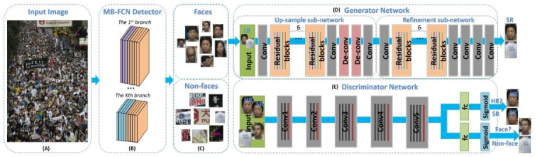

Figure 3: The pipeline of the proposed tiny face detector system.

super-reconstruction to the target detection field, and improved the performance of minimal face detection.

3.4. Domain Adaptive

3.4.1 Background

The Domain Adaptation (DA) problem has been extensively studied in image classification tasks and has made amazing progress. There are also many related work on CVPR in 2018. The essence is a kind of migration learning. The problem is how to make the trained classifier on the Source Domain migrate well to the target domain without label data. Two of the representative jobs are: DSN and ADDA .

3.4.2 Method

This paper first demonstrates the necessity of domain adaptation from the perspective of probability distribution, and draws two contributions from this paper: image-level adaptation (Image-Level Adaptation) and target-level adaptation (Instance-Level Adaptation)), which is used to solve the domain adaptation problem of using different data training for the target detection task in the automatic driving scenario.The structure of model is Figure4. The author tackle the domain shift on two levels, the image level and the instance level. A domain classifier is built on each level, trained in an adversarial training manner. A consistency regularizer is incorporated within these two classifiers to learn a domain-invariant RPN for the Faster R-CNN model.

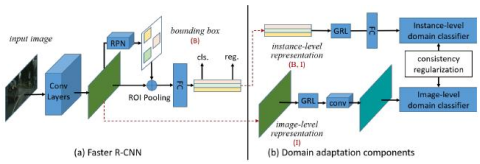

Figure 4: An overview of paper Domain Adaptive Faster R-CNN model.

3.4.3 Conclusion

The experimental part demonstrates that the training strategy of this paper can improve the performance of the detector in the target domain. However, will it cause the detector's detection performance in the source domain to degrade is doubt. Intuitively, the network will be optimized in the direction of the indistinguishable of the source and target domains during the training process. However, it's ambigious this optimization is beneficial to the target domain,the problem remains to be further verified by the experiment.

3.5. Person Re-identification

3.5.1 Background

Person Re-identification is an image retrieval problem. Given a set of images, for each picture in the probe, find the picture that most likely belongs to the same pedestrian from the candidate picture set.

The human identification data set is taken by a series of surveillance cameras, and the detection algorithm is used to pull the pedestrians out and make the pedestrian match. In these data sets, the face is very blurry and cannot be used as a matching feature. Because of the different viewing angles of multiple cameras, the same person may be photographed on the front, side, and back, with different visual features, so it is more difficult. Image matching problem.

Before CVPR2018, the unsupervised work officially published in the Person Reid field was only CVPR2016's UMDL: Unsupervised Cross-Dataset Transfer Learning for Person Re-identification, based on the dictionary learning method, learning cross-data set invariance dictionary on multiple source data sets, Migrate to the target dataset. However, the accuracy rate is still very low.

3.5.2 Method Spatial-Temporal Patterns

The so-called space-time model, that is, the distribution of migration time between pedestrians in a camera network. The author looked at all the Reid datasets and found that there are three datasets with occasional null information, Market1501, GRID, DukeMTMC4ReID. Among them, DukeMTMC-ReID came out in the second half of 2017. The time comparison hastily contained no relevant information in the paper. experiment. Market1501 is a relatively large Person Reid dataset. GRID is a relatively small Person Reid dataset and has six cameras (although 8 cameras are introduced in GRID, there are actually only 6 cameras).

The author first pre-train a convolutional neural network on other datasets (so we can say that this is a cross-dataset task), and then use this convolutional neural network to extract features from the target dataset, using cosine distance calculations. Feature similarity ranks the top ten as the same person Using this "same person" information + maximum likelihood estimation to construct a spatiotemporal model. Unsupervised space-time model construction is

Figure5.

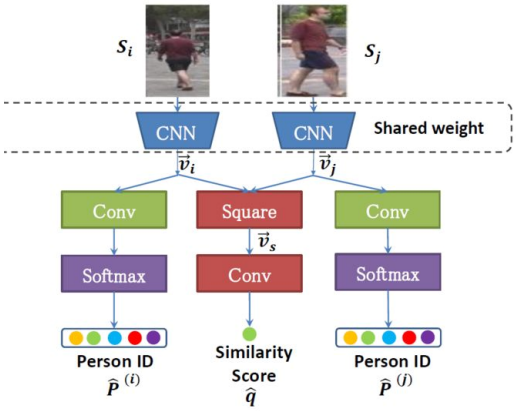

Figure 5: The Siamese Network to classify

3.5.3 Conclusion

From the current experimental results, the first migration learning is relatively large, and the latter promotion is relatively small. This phenomenon is better to say that convergence is fast, but it is broken, although the image classifier has been improved, but there is no phenomenon that the image classifier is promoted larger than the fusion classifier, so there should be something to study further.

4. Summary

From the papers published in 2018 and 2019, we can see that in detection domain ,you must use CNN or other learning algorithm to improve your accurancy. It's significant to construct an excellent model when you want to get satisified result. In a word, it maybe a trend to use CNN or other knowledge futher your computer vision study.

References

[1] Chenchen Zhu, Yihui He, and Marios Savvides. Feature selective anchor-free module for single-shot object detection. *arXiv preprint arXiv:1903.00621*, 2019.

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6154–6162, 2018.

[3] Yancheng Bai, Yongqiang Zhang, Mingli Ding, and Bernard Ghanem. Finding tiny faces in the wild with generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–30, 2018.

[4] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3339–3348, 2018.

[5] Jianming Lv, Weihang Chen, Qing Li, and Can Yang. Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7948–7956, 2018.