DEEP LEARNING FOR VEHICLE DETECTION IN AERIAL IMAGES

Michael Ying Yang

Scene Understanding Group
University of Twente

Wentong Liao, Xinbo Li, Bodo Rosenhahn

Institute for Information Processing
Leibniz University Hannover

ABSTRACT

The detection of vehicles in aerial images is widely applied in many domains. In this paper, we propose a novel double focal loss convolutional neural network framework (DFL-CNN). In the proposed framework, the skip connection is used in the CNN structure to enhance the feature learning. Also, the focal loss function is used to substitute for conventional cross entropy loss function in both of the region proposed network and the final classifier. We further introduce the first large-scale vehicle detection dataset ITCVD with ground truth annotations for all the vehicles in the scene. The experimental results show that our DFL-CNN outperforms the baselines on vehicle detection.

Index Terms— Vehicle detection, convolutional neural network, focal loss, ITCVD dataset

1. INTRODUCTION

The detection of vehicles in aerial images is widely applied in many domains, *e.g.*traffic monitoring, vehicle tracking for security purpose, parking lot analysis and planning, *etc*. Therefore, this topic has caught increasing attention in both academic and industrial fields [1, 2, 3]. However, compared with object detection in ground view images, vehicle detection in aerial images has a lot of different challenges, such as much smaller scale, complex backgrounds and the monotonic appearance. See Figure 1 for an illustration.

Before the emergence of deep learning, hand-crafted features combined with a classifier are the mostly adopted ideas to detect vehicles in aerial images [4, 1, 2]. However, the hand-crafted features lack generalization ability, and the adopted classifiers need to be modified to adapt the of the features. Some previous works also attempted to use shallow neural network [5] to learn the features specifically for vehicle detection in aerial images [6, 7]. However, the representational power of the extracted features are insufficient and the performance meets the bottleneck. Furthermore, all of these methods localize vehicle candidates by sliding window search. It's low efficient and leads to costly and redundant computation. The window's sizes and sliding steps must be carefully chosen to adapt the varieties of objects of interest in dataset.

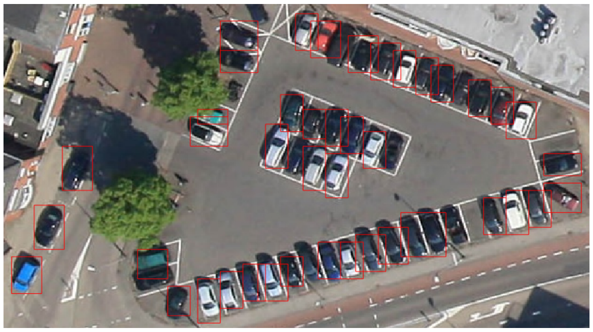

Fig. 1. Vehicles detection results on the proposed dataset.

In recent years, deep convolutional neural network (DCNN) has achieved great successes in different tasks, especially for object detection and classification [8, 9]. In particular, the series of methods based on region convolutional neural network (R-CNN) [10, 11, 12] push forward the progress of object detection significantly. Especially, Faster-RCNN [12] proposes the region proposal network (RPN) to localize possible object instead of traditional sliding window search methods and achieves the state-of-the-art performance in different datasets in terms of accuracy. However, these existing state-of-the-art detectors cannot be directly applied to detect vehicles in aerial images, due to the different characteristics of ground view images and aerial view images [13]. The appearance of the vehicles are monotone, as shown in Figure 1. It's difficult to learn and extract representative features to distinguish them from other objects. Particularly, in the dense park lot, it is hard to separate individual vehicles. Moreover, the background in the aerial images are much more complex than the nature scene images. For examples, the windows on the facades or the special structures on the roof, these background objects confuse the detectors and classifiers.Furthermore, compared to the vehicle sizes in ground view images, the vehicles in the aerial images are much smaller (ca. 50×50 pixels) while the images have very high resolution (normally larger than 5000×2000 pixels). Lastly, large-scale and well

The work is funded by DFG (German Research Foundation) YA 351/2-1 and RO 4804/2-1. The authors gratefully acknowledge NVIDIA Corporation for the donated GPU used in this research. We thank Slagboom en Peeters for providing the aerial images. Contact email: michael.yang@utwente.nl

978-1-4799-7061-2/18/$31.00 ©2018 IEEE

annotated dataset is required to train a well performed DCNN methods. However, there is no public large-scale dataset, such as ImageNet [14], for vehicle detection in aerial images. The two exceptions are VEDAI dataset [15] and DLR 3K dataset [2]. However, the objects in the VEDAI dataset are relative easy to detect because of the small number of vehicles which sparsely distribute in the images, and the background is simple. The more challenging and realistic DLR 3K dataset contains totally 20 aerial images with resolution of 5616×3744. 10 images (3505 vehicles) are used for training. Such number of training samples seems too small for training a CNN model.

To address these problems, we propose a specific framework for vehicle detection in aerial images, as shown in Figure 2. The novel framework is called double focal loss convolutional neural network (DFL-CNN), which consists of three main parts: 1) A skip-connection from the low layer to the high layer is added to learn features which contains rich detail information. 2) Focal loss function [16] is adopted in the RPN instead of traditional cross entropy. This modification aims at the class imbalance problem when RPN determine whether a proposal is likely an object of interest. 3) Focal loss function replaces the cross entropy in the classifier. It's used to handle the problem of easy positive examples and hard negative examples during training. Furthermore, we introduce a novel large-scale and well annotated dataset for quantitative vehicle detection evaluation - ITCVD. Towards this goal, we collected 173 images with 29088 vehicles, where each vehicle in the ITCVD dataset is manually annotated using a bounding box. The performance of the proposed method is demonstrated with respect to the state-of-the-art baseline. We make our code and dataset online available.

2. PROPOSED FRAMEWORK

An overview of the proposed framework is illustrated in Figure 2. It's modified based on the standard Faster R-CNN [12]. We refer readers to [12] for the general procedure of object detection. In this work, we choose ResNet [17] as the backbone structure for feature learning, because of its high efficiency, robustness and effectiveness during training [18].

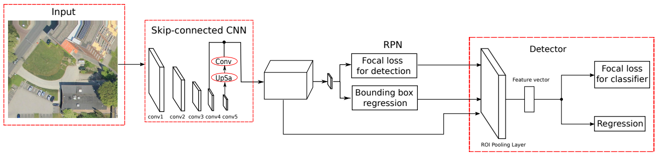

Fig. 2. The overview of the proposed framework DFL-CNN. It consists of three main parts: 1) A skip-connection from the low layer to the high layer is added to learn features which contains rich detail information. 2) Focal loss function [16] is adopted in the RPN instead of traditional cross entropy. 3) Focal loss function replaces the cross entropy in the classifier.

2.1. Skip Connection

It has been proven in the task of semantic segmentation that, features from the shallower layers retain more detail information [19]. In the task of object detection, the sizes of vehicles in aerial images are ca. 30×50 pixels, assuming 10cm GSD. The size of the output feature maps of the ResNet from the $5th$ pooling layers is only one 32nd of the input size [17]. The shorter edges of most vehicles are very small when they are projected on the feature maps after the $5th$ pooling layer. So, they will be ignored because their sizes are rounded up. Furthermore, pooling operation leads to significant loss of detailed information. For densely parked area, it is difficult to separate individual vehicles. For example, the extracted features from the shallow layer have richer detailed information than the features from the deeper layer. In the case of densely parked area, the detail information play an important to separate the individual vehicles from each other. Therefore, we fuse the features from the shallow layers, which contain more detail information, with the features learned by deeper layers, which have more representative abilities, to precisely localize detected individual vehicle. This skip-connected CNN architecture is illustrated in Figure 3. The image fed to the network is 752×674 pixels. The size of the feature maps from the $4th$ and $5th$ pooling layers are $42 \times 47 \times 1024$ and $21 \times 24 \times 2048$ respectively. To fuse them together, the smaller feature maps are upsampled to the size of $42 \times 47 \times 2048$, and then reduced the feature channels into 1024 by a 1×1 convolution layer. Then the two feature maps are concatenated as the skip-connected feature maps.

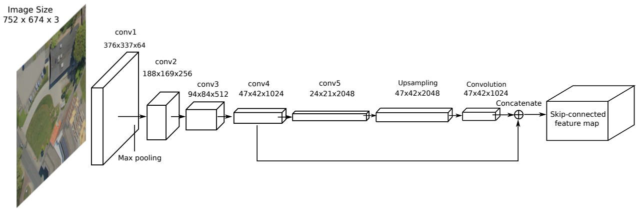

Fig. 3. Structure of skip-connected CNN. The feature maps from the conv5 are upsampled to the same size as the feature maps from conv4. Then, the number of the feature channels are reduced by 1×1 convolution layer into 1024. Finally, the feature maps from conv4 and conv5 are concatenated.

2.2. Focal loss function

Focal loss function is originally proposed by [16] to dedicate the class imbalance problems for the one-stage object detectors, such as YOLO [20] and SSD [21]. As discussed in the paper, a one-stage detector suffers from the extreme foreground-background class imbalance because of the dense candidates which cover spatial positions, scales, and aspect ratios. A two-stage detector handles this challenge in the first stage: candidates proposal, *e.g.*RPN [12], most of the candidates which are likely to be the background are canceled, and then the second stage: classifier works on much sparser candidates. However, in the scenes with dense objects of interest, *e.g.*, the parking cars in Figure 1, even the state-of-the-art candidates proposal method RPN is not good enough to filter the dense proposals in two aspects: 1) many of the dense proposals cover two vehicles and have high overlap with the ground truth, which makes it hard for the proposal methods to determine whether they are background objects. 2) Too many background objects interfere the training. It is hard to select the negative samples which are very similar as the vehicles to enhance the detector/classifier to distinguish them from the positive samples. Inspired by the idea in [16], we proposed to use the *focal loss function* instead of the conventional CE loss both in the region proposal and the classification stages, dubbed as *double focal loss-CNN* (DFL-CNN).

The Focal loss is derived from the CE loss by adding a modulating factor $(1 - p_t)^\gamma$ with tunable focusing parameter $\gamma \geq 0$:

$$L_{FL}(p_t) = -(1 - p_t)^\gamma \ \log(p_t) \tag{1}$$

The focal loss has two main properties: 1) The loss is unaffected by misclassified examples which have small p_t when the modulating factor is near 1. In contrast, when $p_t \to 1$, the modulating factor is near 0 , which down-weights the loss for well-classified examples. 2)When the focusing parameter γ is increased, the effect of modulating factor is also increased. CE is the special case of $\gamma = 0$. Intuitively, the contribution of the easy examples are reduced while the ones from hard examples are enhanced during the training. For example, with $\gamma = 2$ [1], the focal loss of an example classified with $p_t = 0.9$ is 1% of the CE loss and 0.1% of it when $p_t = 0.968$. If an example is misclassified ($p_t < 0.5$), its importance for training is increased by scaling down its loss 4 times.

2.3. Double Focal Loss CNN

In our DFL-CNN framework, we add a skip connection to fuse the features from the lower (conv4) and higher (conv5) layers, and adopt focal loss function both in the RPN layer and the final classification layer to overcome the class imbalance and the easy/hard examples challenges in our task.

As discussed in Section 2.1, the final feature maps are 1/16 of the original images. Therefore, each pixel in the feature maps corresponds an region of 16×16 pixels. To generate candidates proposal, centered on each pixel in the feature maps, 9 anchors in 3 different areas (30^2, 50^2, 70^2) and 3 different ratios (1:1, 2:1 and 1:2) are generated on the original input image. Every anchor is labeled as either positive or negative sample based on the Intersection-over-Union (IoU) with ground truth. The IoU is formally defined as: IoU $= \frac{area(\text{Proposal}\cap\text{Ground Truth})}{area(\text{Proposal}\cup\text{Ground Truth})}$, where the numerator is the overlapping area of box of candidate and the ground truth box, and the denominator represents the union of them. The proposals which have the IoU more than 0.7, are labeled as positive samples and the ones whose IoU are smaller than 0.1 are labeled as the negative samples. Other proposals are discarded. All the proposals exceeding the boundary of the image are also discarded. During training, each mini-batch consists of 64 positive samples and 64 negative samples.

The loss function for training the RPN using focal loss is defined as:

$$\begin{aligned} L_{RPN}(\{p_i\}, \{t_i\}) =& \frac{1}{N_{cls}} \sum_i L_{cls-FL}(p_i, p_i^*) \\ &+ \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \end{aligned} \tag{2}$$

where L_{cls-FL} is the focal loss for classification, as defined in Eq. (1) and L_{reg} is the loss for bounding box regression. p_i is the predicted probability of proposal i belonging to the foreground and p_i^* is its ground truth label. N_{cls} denotes the total number of samples and N_{reg} is the total number of positive samples. λ is used to weight the loss for bounding box regression [2]. The smooth L_1 loss function is adopted for L_{reg} as in [12]. $t = (t_x, t_y, t_w, t_h)$ is the normalized information of the bounding boxes of the positive sample and t^* is its ground truth.

The RPN layer output a set of candidates which are likely to be the objects of interest, *i.e.*vehicles in this work, and there predicted bounding boxes. Then, the features covered by these bounding boxes are cropped out from the feature maps and go through the region of interest (ROI) pooling layer to get a fix the size of features.

Finally, the final classifier subnet are fed with these features and classify their labels, and predict their bounding boxes further. The loss function of the classifier subnet for each candidate is formally defined as:

$$L_{classifier}(P, T) = L_{cls-FL}(P, P^*) + \lambda_2 P^* L_{reg}(T, T^*) \tag{3}$$

where T is defined as:

$$\begin{aligned} &T_x = (P_x - A_x)/A_w, \qquad T_y = (P_y - A_y)/A_h, \\ &T_w = \log(P_w/A_w), \qquad T_h = \log(P_h/A_h), \\ &T_x^* = (P_x^* - A_x)/A_w, \qquad T_y^* = (P_y^* - A_y)/A_h, \\ &T_w^* = \log(P_w^*/A_w), \qquad T_h^* = \log(P_h^*/A_h), \end{aligned} \tag{4}$$

The P_x, A_x and P_x^* denote the bounding boxes of prediction results, anchors and ground truth. The other subscripts of y, w and h are the same as x. We set $\lambda_2 = 1$ to equal the influence of classification and bounding box prediction. During training, the classifier subnet is trained using positive and negative samples in ratio of $1 : 3$, same as the conventional training strategy [12].

[1] γ is set to 2 in our experiments.

[2] λ is set to 15 in our experiments.Because the size of final feature maps is 47×42 and totally 128 anchors are chosen, therefore the ratio is ca. 15.

3. ITCVD DATASET

In this section, we introduce the new large-scale, well annotated and challenging ITCVD dataset. The images were taken from an airplane platform which flied over Enschede, The Netherlands, in the height of ca 330m above the ground [3]. The images are taken in both nadir view and oblique view The tilt angle of oblique view is 45 degrees. The Ground Sampling Distance (GSD) of the nadir images is 10cm.

The raw dataset contains 228 aerial images with high resolution of 5616×3744 pixels in JPG format. Because the images are taken consecutively with a small time interval, there is ca. 60% overlap between consecutive images. It is important to make sure that, the images used for training do not have common regions with the images that are used for testing. After careful manual selection and verification, 173 images are remained among which 135 images with 23543 vehicles are used for training and the remaining 38 images with 5545 vehicles for testing. Each vehicle in the dataset is manually annotated using a bounding box which is denoted as (x, y, w, h), where (x, y) is the coordinate of the left-up corner of the box, and (w, h) is the width and height of the box respectively.

4. EXPERIMENTS

4.1. Dataset and experimental settings

We evaluate our method in our ITCVD dataset [4]. To save the GPU memory, each original image in the datasets are cropped into small patches uniformly. The resulting new image patches are in the size of 674×752 pixels. The coordinate information of annotation is also updated in the new cropped patches. The deep learning models are implemented in Keras with TensorFlow backend. The ResNet-50 network [17] is used as the backbone CNN structure for feature learning for Faster R-CNN [12] and our model. We use a learning rate of 0.00001 to train the RPN. The CNN structure are pre-trained on ImageNet dataset [14].

To evaluate the experimental results, the metrics of recall/precision rate and $F1$-score are used, which are formally defined as: Recall Rate (RR) $= \frac{\text{TP}}{\text{TP+FN}}$, Precision Rate (PR) $= \frac{\text{TP}}{\text{TP+FP}}$, F1-score $= \frac{2\times\text{RR}\times\text{PR}}{\text{RR+PR}}$, where TP, FN, FP denote the *true positive*, *false negative* and *false positive* respectively. Furthermore, the relationships between the IoU and RR, PR are also evaluated respectively.

4.2. Results on ITCVD dataset

The state-of-the-art object detector Faster R-CNN [12] is implemented to provide a strong baseline. In addition, traditional HOG + SVM method [22] is provided as a weak baseline. Figure 4 depicts the relationship between recall rate and the precision rate of DFL-CNN, Faster R-CNN and HOG+SVM algorithms with different IoU in the ITCVD dataset. It is obvious that the CNN based methods (DFL-CNN in green curve and Faster R-CNN in red curve) are significantly better than the traditional method (HOG+SVM in black curve). In the relation between recall and precision, our DFL-CNN method also perform better than Faster R-CNN. According to these relationship curves, $IoU = 0.3$ is a good balance point for the following experimental settings, which reports high recall rate and precision at the same time. Note that, it is also a conventional setting in the task of object detection. The quantitative results of these three methods are given in Table 1 (the results are given with $IoU = 0.3$). We can see that, our method outperforms the others.

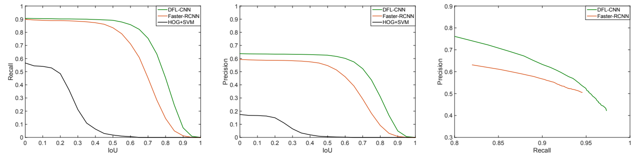

Fig. 4. The relationship between IoU and recall rate (a), IoU and precision rate (b) and recall and precision (c) of DFL-CNN, Faster R-CNN, HOG+SVM in the ITCVD dataset.

	HOG+SVM	Faster R-CNN	DFL-CNN
RR	21.19%	88.38%	**89.44%**
PR	6.52%	58.36%	**64.61%**
F1-score	0.0997	0.7030	**0.7502**

Table 1. Comparison of baselines and the DFL-CNN method in ITCVD dataset.

5. CONCLUSION

In this paper, we have proposed a specific framework DFL-CNN for vehicle detection in the aerial images. We fuse the features properties learned in the lower layer of the network (containing more spatial information) and the ones from higher layer (more representative information) to enhance the network's ability of distinguishing individual vehicles in a crowded scene. To address the challenges of class imbalance and easy/hard examples, we adopt focal loss function instead of the cross entropy in both of the region proposal stage and the classification stage. We have further introduced the first large-scale vehicle detection dataset ITCVD with ground truth annotations for all the vehicles in the scene. Compared to DLR 3K dataset, our benchmark provides much more object instances as well as novel challenges to the community. For future work, we will extend DFL-CNN to recognize the vehicle types and detect the vehicle orientations.

[3] http://www.slagboomenpeeters.com/

[4] Further experiments on DLR 3K dataset [2] can be found in the technical report `https://arxiv.org/abs/1801.07339`.

6. REFERENCES

[1] Joshua Gleason, Ara V Nefian, Xavier Bouyssounousse, Terry Fong, and George Bebis, "Vehicle detection from aerial imagery," in *IEEE International Conference on Robotics and Automation*, 2011, pp. 2065–2070.

[2] Kang Liu and Gellert Mattyus, "Fast multiclass vehicle detection on aerial images," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 9, pp. 1938–1942, 2015.

[3] Ziyi Chen, Cheng Wang, Huan Luo, Hanyun Wang, Yiping Chen, Chenglu Wen, Yongtao Yu, Liujuan Cao, and Jonathan Li, "Vehicle detection in high-resolution aerial images based on fast sparse representation classification and multiorder feature," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 8, pp. 2296–2309, 2016.

[4] Tao Zhao and Ram Nevatia, "Car detection in low resolution aerial images," *Image and Vision Computing*, vol. 21, no. 8, pp. 693–703, 2003.

[5] Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in neural information processing systems*, 1990, pp. 396–404.

[6] Hsu-Yung Cheng, Chih-Chia Weng, and Yi-Ying Chen, "Vehicle detection in aerial surveillance using dynamic bayesian networks," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 2152–2159, 2012.

[7] Xueyun Chen, Shiming Xiang, Cheng-Lin Liu, and Chun-Hong Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *IEEE Geoscience and remote sensing letters*, vol. 11, no. 10, pp. 1797–1801, 2014.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[9] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[11] Ross Girshick, "Fast r-cnn," in *IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[13] Guisong 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," *CoRR*, vol. abs/1711.10398, 2017.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[15] Sebastien Razakarivony and Frederic Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," *Journal of Visual Communication and Image Representation*, vol. 34, pp. 187–203, 2016.

[16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," *arXiv preprint arXiv:1708.02002*, 2017.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] Alfredo Canziani, Adam Paszke, and Eugenio Culurciello, "An analysis of deep neural network models for practical applications," *arXiv preprint arXiv:1605.07678*, 2016.

[19] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[20] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision*. Springer, 2016, pp. 21–37.

[22] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886–893.