LETTER

Learning Deep Relationship for Object Detection*

Nuo XU[†], ***Nonmember and*** **Chunlei HUO**[†a)], ***Member***

SUMMARY Object detection has been a hot topic of image processing, computer vision and pattern recognition. In recent years, training a model from labeled images using machine learning technique becomes popular. However, the relationship between training samples is usually ignored by existing approaches. To address this problem, a novel approach is proposed, which trains Siamese convolutional neural network on feature pairs and finely tunes the network driven by a small amount of training samples. Since the proposed method considers not only the discriminative information between objects and background, but also the relationship between intraclass features, it outperforms the state-of-arts on real images.

key words: *object detection, Siamese convolutional neural network, remote sensing images, relationship learning*

1. Introduction

Object detection has historically been one of the most important topics in various domains, such as image processing, pattern recognition, remote sensing, and so on [1]. In recent years, satellite remote sensing has entered an unprecedented new stage, and the improved spatial resolution images taken by very high resolution satellites such as QuickBird 2 make it possible to detect objects from satellite images.

Generally, object detection can be modeled as a classification problem, and it consists of two key steps: feature representation and feature classification [2]. Due to the low separation between objects and background, detecting objects from high resolution images is more difficult. With the development of deep learning techniques, learning features from data is promising for discriminating objects from the clutter background. For instance, Cheng proposed learning rotation-invariant HOG [3] and rotation-invariant CNN(convolutional neural networks) [4] to describe objects. Cao [5] tried to use the region-based CNN to detect aircrafts under complex environments. Diao [6] suggested combining unsupervised feature learning and visual salience, where feature learning is performed by deep belief networks.

Despite the novelties of existing approaches, most of them rely mainly on the semantically-labeled data and ignore the relationship between training samples, where the relationship means the similarity and difference between coherent training samples with respect to the labels and features. To simultaneously improve the separability between objects and background and uncover the subtle relationship hidden between features, a novel object detection approach is proposed in this paper. Compared with related works, the novelty of the proposed approach lies in relationship representation and deep relationship learning.

Manuscript received June 15, 2017.
Manuscript revised August 31, 2017.
Manuscript publicized September 28, 2017.

[†]The authors are with National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

*This work was supported by the National Natural Science Foundation of China under Grants 91438105 and 61375024.

a) E-mail: clhuo@nlpr.ia.ac.cn

DOI: 10.1587/transinf.2017EDL8131

2. The Proposed Approach

The rationale of the proposed approach is to take advantage of relationship representation to capture the difference between interclass features (i.e., objects and background) and the similarities within intraclass features. Below, we will elaborate the proposed approach step by step.

2.1 Relationship Representation

As stated before, one of key difficulties of object detection is the low separability between objects and background. In other words, learning differences between interclass features and learning similarities between intraclass features are helpful for feature classification. However, traditional object framework such as SVM focuses purely on maximizing differences between objects and background, and it neglects the similarities within objects or background. For this reason, we learn the relationship between feature pairs instead of individual features.

Motivated by Siamese network [7], we use the label of feature pair to describe the relationship between individual features, and the relationship is learned by minimizing the following contrastive loss function:

$$E = \frac{1}{2N}\sum_{n=1}^{N}(z_n d_n + (1 - z_n)max(m - d_n, 0)) \quad (1)$$

Where $d_n = \|a_n - b_n\|_2^2$. N is the total number of training feature pairs. z_n is the label of the paired data, $z_n = 1$ means that the original features a_n and b_n share the same label, $z_n = 0$ means a_n has different label with b_n. m is a margin, which is set to be 1.

Training feature pairs are generated by randomly combining original training samples, i.e., $p_{ij} = (\boldsymbol{x}_i, \boldsymbol{x}_j, z_{ij})$, where $\{(\boldsymbol{x}_i, l_i)\}$ is the original training set, $\boldsymbol{x}_i$ and l_i denote the feature and label of the ith training sample, respectively. z_{ij} is determined by the consistency between l_i and l_j, i.e.,

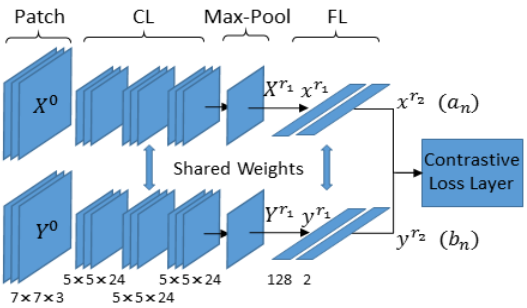

Fig. 1 Siamese CNN for pre-training.

$z_{ij} = 1$ if $l_i = l_j$. It is worth noting the differences between labels z_{ij} and l_i, l_i means the feature $\boldsymbol{x}_i$ is object or background, while z_{ij} means whether $\boldsymbol{x}_i$ has the same label with $\boldsymbol{x}_j$. In other words, z_{ij} contains the relationship about intra-class similarity and interclass difference.

2.2 Deep Relationship Learning

In this paper, relationship learning is implemented within deep learning architecture. As shown in Fig. 1, deep relationship learning network consists of two identical basic networks, each of which individually learns the features from patches and shares the parameters. For convenience, we describe main operators as follows:

$$\begin{aligned} X^l &= \{X_v^l \mid v = 1, \cdots, s_l\},\ l = 0, \cdots, r_1 \\ X_v^l &= f_1\Big(\sum_{u=1}^{s_{l-1}} (X_u^{l-1} * K_{uv}^l)\Big),\ l = 1, \cdots, r_1 \end{aligned} \tag{2}$$

Where X^l represents the output after the lth convolution layer. $f_1(\cdot)$ is ReLU operator, whose role is to learn sparse features and remove the irrelevant noise. At the lth convolution layer, there are s_l convolution kernel groups, and each group has s_{l-1} convolution kernels K_{uv}^l of size 5×5. X^l contains s_l maps $X_1^l, X_2^l, \cdots, X_{s_l}^l$. In this paper, the deep network consists of three ($r_1 = 3$) convolution layers (CL).

$$x^l = f_2(w^l \cdot x^{l-1} + b^l),\ l = r_1 + 1, \cdots, r_2 \tag{3}$$

After the pooling layer, two fully-connected layers (FL) are stacked, whose neurons are 128 and 2 respectively. The role of fully-connected layers is dimensionality reduction. The matrix X^{r_1} is reformulated to be the vector x^{r_1} by column-wise rearrangement as shown in Fig. 1. In Eq. (3), f_2 is an ordinary linear function, $f_2(x) = x$. The outputs x^{r_2} and y^{r_2} are equivalent to a_n and b_n in Eq. (1).

In this paper, a pair of 7×7 patches X^0 and Y^0 is used as the input, and $s_0 = 3$ represents RGB maps. s_1, s_2 and s_3 denote the number of convolution kernel groups at each layer, and $s_1 = s_2 = s_3 = 24$, which also represent the number of convolution kernel groups at each layer.

The training procedure can be divided into two steps: pre-training and fine-tuning [8]. Pre-training is trained on feature pairs and solved by Adadelta [9] algorithm. As shown in Fig. 2, fine-tuning is to adapt the pre-trained model

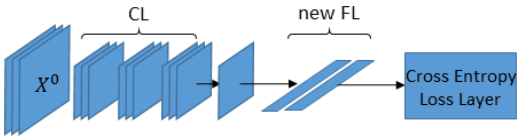

Fig. 2 Fine-tuning. Fine-tuning is performed on the original training samples for object detection task.

(a) Some training images.

(b) Test images .

Fig. 3 Training images and test images. (a) Some training images. (b) Test images.

for original training samples $\{(\boldsymbol{x}_i, l_i)\}$. Cross Entropy loss function is chosen for fine-tuning, and the parameters of the CL in fine-tuning network are initialized by the ones achieved in pre-training, while the parameters of FL are set randomly. A smaller initial learning rate is applied on the CL to adjust parameters progressively.

2.3 Object Detection

In object detection step, the patch feature centered at each pixel in the test image is extracted, and its label is determined by the finely-tuned network. Since Siamese CNN is trained on paired features, it is helpful for considering the relationship between individual features and improving the performance.

3. Experiments

3.1 Experiment Setting

Aircraft is one of the most important objects, and we evaluate the proposed approach in the context of aircraft detection. Training images are shown in Fig. 3 (a), which consist of 17 pan-sharpened multispectral remote sensing images with the size of 600×600 pixels. Test images consist of 5 remote sensing images of the same sizes.

To demonstrate the effectiveness of the proposed approach, three SVM-based approaches are used for comparison: **SVM** [10], **D-SVM** (Doublet-SVM) [11], **Fk-SVM** (FkNN-SVM) [12]. For these three approaches, DAISY [13] feature is used since it's a fast dense local image feature descriptor. D-SVM is used to learn the shallow relationship by a single-layer architecture, while Fk-SVM trains SVM to learn the local spatial relationship. For convenience, the proposed approach is abbreviated as **SCNN**.

To investigate how relationship learning help improve

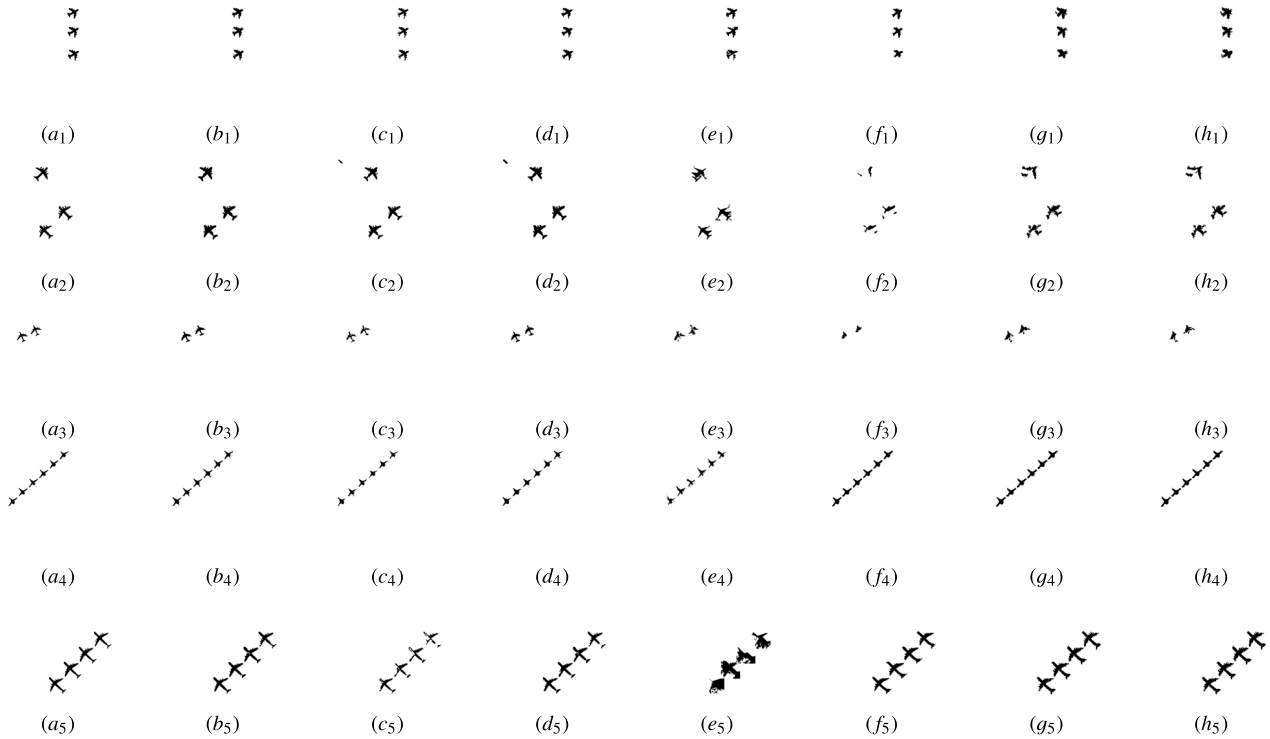

Fig. 4 Results comparison of different approaches. Results on the ith test image are shown in the ith row. Column 1: Ground Truth, Column 2: SCNN, Column 3: CNN, Column 4: DeepDesc, Column 5: GraphSeg, Column 6: SVM, Column 7: D-SVM, Column 8: Fk-SVM.

the performance, three other deep learning approaches are also added for comparison:

1) **CNN**. Traditional CNN is used for feature learning and feature classification. This approach is aimed at understanding the necessity of relationship learning.

2) **DeepDesc**. To illustrate the advantage of fine tuning, a Siamese-like approach DeepDesc [14] is used to learn discriminative features, and kNN is utilized to classify the features.

3) **GraphSeg**. R-CNN [15] is utilized to detect the bounding box of the object, and graph based image segmentation algorithm [16] is applied to refine the object boundary.

3.2 Experiment Analysis

Since we model the object detection task as a classification problem, we use the following four metrics to compare the above methods quantitatively: Precision, Recall, CA (Classification Accuracy) and F-measure. Results of different approaches are shown in Fig. 4, and the performances are listed in Table 1. From Table 1, it can be found that SCNN achieves the best performance with respect to four metrics. For instance, its average classification accuracy and F-measure are 99.7% and 0.893, respectively, while average CA and F-measure of DeepDesc are 99.6% and 0.878, which are inferior to SCNN.

For SVM-based methods, D-SVM and Fk-SVM achieve higher Recalls than SVM, whose Recalls on IMG3 are 65.0%, 59.5% and 43.7%, respectively. D-SVM achieved higher Recalls since it captured the relationship between features by feature couplets, while Fk-SVM improved the performance by a local decision. Despite the advantages of relationship learning in improving Recalls, many false alarms are obtained, which can be observed from Fig. 4. The underlying reason lies in the fact that the relationship captured by a single layer architecture is limited in reflecting the fine structure between objects and background. However, such subtle relationship is caught by deep learning. To understand how deep relationship learning improve the performance, we analyze different approaches in detail by taking IMG2 for an example. By comparing Fig. 4(b_2), Fig. 4(g_2) and Fig. 4(h_2), we find that many pixels are wrongly identified as the background by D-SVM and Fk-SVM, whose Recalls are 56.2% and 58.4% respectively, which are significantly lower than SCNN, 97.2%.

Noting that GraphSeg is worst among deep learning approaches, the key reason is that R-CNN aims at detecting the object's bounding box and the performances are difficult to improve even with the help of segmentation. As can be seen from Fig. 4(e_5), results obtained by GraphSeg are inconsistent with the ground truths especially near the object boundary. DeepDesc is similar to the pre-training step of SCNN with respect to Siamese network, but SCNN outperforms DeepDesc definitely. For instance, their average F-measures are 0.893 and 0.878, respectively. The reason lies in the lack of fine-tuning. In detail, DeepDesc is promising in obtaining discriminative features for patch comparison, but for the pixel-wise feature learning and feature classifi-

Table 1 Object detection performance comparison

Image	Indicator	SCNN	CNN	DeepDesc	GraphSeg	SVM	D-SVM	Fk-SVM
IMG1	CA	**99.8%**	**99.8%**	**99.8%**	99.7%	99.6%	99.5%	99.5%
	Precision	89.5%	**94.1%**	91.5%	85.2%	86.9%	72.5%	71.1%
	Recall	**95.3%**	87.6%	94.4%	89.8%	79.6%	85.3%	89.5%
	F-measure	0.923	0.907	**0.929**	0.875	0.831	0.784	0.792
IMG2	CA	**99.5%**	**99.5%**	99.4%	98.9%	98.5%	98.5%	98.5%
	Precision	79.9%	**83.8%**	76.1%	69.1%	67.8%	56.0%	57.2%
	Recall	97.2%	92.1%	**97.5%**	71.1%	31.4%	56.2%	58.4%
	F-measure	0.877	**0.878**	0.855	0.701	0.429	0.561	0.578
IMG3	CA	**99.9%**	99.8%	99.8%	99.6%	99.7%	99.6%	99.6%
	Precision	79.6%	**85.0%**	78.9%	55.0%	74.4%	59.2%	58.8%
	Recall	**93.8%**	79.9%	89.0%	55.9%	43.7%	65.0%	59.5%
	F-measure	**0.861**	0.823	0.836	0.554	0.550	0.619	0.591
IMG4	CA	**99.8%**	**99.8%**	99.7%	99.4%	99.6%	99.5%	99.5%
	Precision	84.1%	**91.9%**	83.6%	71.5%	72.0%	67.3%	67.3%
	Recall	95.5%	86.4%	91.5%	71.8%	**99.5%**	99.2%	99.1%
	F-measure	**0.894**	0.891	0.873	0.716	0.836	0.802	0.801
IMG5	CA	**99.5%**	99.1%	**99.5%**	96.6%	99.1%	98.8%	98.8%
	Precision	86.8%	**91.9%**	86.3%	40.3%	74.1%	68.8%	67.8%
	Recall	96.0%	70.7%	92.9%	75.2%	96.3%	**98.3%**	98.0%
	F-measure	**0.912**	0.799	0.895	0.525	0.837	0.809	0.802
AVG	CA	**99.7%**	99.6%	99.6%	98.8%	99.3%	99.2%	99.2%
	Precision	84.0%	**89.3%**	83.3%	64.2%	75.0%	64.8%	64.4%
	Recall	**95.5%**	83.3%	93.0%	72.7%	70.1%	80.8%	80.9%
	F-measure	**0.893**	0.860	0.878	0.674	0.697	0.715	0.713

cation problem, fine-tuning is important and necessary for achieving accurate region shape.

The above comparisons demonstrate the importance of three factors: relationship representation, deep learning and fine-tuning. In other words, SCNN cannot obtain the best performances if either factor is out of consideration.

4. Conclusion

In this paper, a novel object detection approach based on deep relationship learning is proposed, which considers not only the discriminative information between objects and background, but also the relationship between training couples. Considering the importance of relationship representation and deep relationship learning, in the future work, transductive Siamese CNN will be developed for semi-supervised object detection.

References

[1] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp.680–688, 2016.

[2] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," ISPRS Journal of Photogrammetry and Remote Sensing, vol.117, pp.11–28, 2016.

[3] G. Cheng, P. Zhou, X. Yao, C. Yao, Y. Zhang, and J. Han, "Object detection in VHR optical remote sensing images via learning rotation-invariant HOG feature," IEEE International Workshop on Earth Observation and Remote Sensing Applications, pp.433–436, 2016.

[4] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," IEEE Trans. Geosci. Remote Sens., vol.54, no.12, pp.7405–7415, 2016.

[5] Y. Cao, X. Niu, and Y. Dou, "Region-based convolutional neural networks for object detection in very high resolution remote sensing images," IEEE International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, pp.548–554, 2016.

[6] W. Diao, X. Sun, X. Zheng, F. Dou, H. Wang, and K. Fu, "Efficient saliency-based object detection in remote sensing images using deep belief networks," IEEE Geosci. Remote Sens. Lett., vol.13, no.2, pp.137–141, 2016.

[7] J. Bromley, J.W. Bentz, L. Bottou, I. Guyon, Y. Lecun, C. Moore, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," Advances in Neural Information Processing Systems, vol.6, pp.737–744, 1994.

[8] P. Agrawal, J. Carreira, and J. Malik, "Learning to see by moving," IEEE International Conference on Computer Vision, pp.37–45, 2015.

[9] M.D. Zeiler, "Adadelta: an adaptive learning rate method," arXiv preprint arXiv:1212.5701, 2012.

[10] R.E. Fan, P.H. Chen, and C.J. Lin, "Working set selection using second order information for training support vector machines," Journal of Machine Learning Research, vol.6, no.Dec, pp.1889–1918, 2005.

[11] F. Wang, W. Zuo, L. Zhang, D. Meng, and D. Zhang, "A kernel classification framework for metric learning," IEEE Trans. Neural Netw. Learning Syst. vol.26, no.9, pp.1950–1962, 2015.

[12] N. Segata and E. Blanzieri, "Fast local support vector machines for large datasets," International Workshop on Machine Learning and Data Mining in Pattern Recognition, vol.5632, pp.295–310, 2009.

[13] E. Tola, V. Lepetit, and P. Fua, "Daisy: An efficient dense descriptor applied to wide-baseline stereo," IEEE Trans. Pattern Anal. Mach. Intell., vol.32, no.5, pp.815–830, 2010.

[14] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," IEEE International Conference on Computer Vision, pp.118–126, 2015.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," IEEE conference on computer vision and pattern recognition, pp.580–587, 2014.

[16] P.F. Felzenszwalb and D.P. Huttenlocher, "Efficient graph-based image segmentation," Int. J. Comput. Vision., vol.59, no.2, pp.167–181, 2004.