Optik 156 (2018) 772–777

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo

Original research article

Optimized CNN Based Image Recognition Through Target Region Selection

Hao Wu [a], Rongfang Bie [a], Junqi Guo [a,*], Xin Meng [b], Shenling Wang [a]

[a] College of Information Science and Technology, Beijing Normal University, China
[b] Electric Power Planning & Engineering Institute, China

ARTICLE INFO

Article history:
Received 25 October 2017
Accepted 22 November 2017

Keywords:
Image recognition
CNN
Target region
Bottom-up region
Enhancement weight

ABSTRACT

Image recognition has plateaued in the last few years. According to this research field, some complicated models typically combined feature extraction and classification models effectively. Moreover, many classic models have already achieved realistic recognition. However, there are still some drawbacks of traditional methods. On the one hand, some unrelated regions of learning instances are often used leading to ignorance of effective features. On the other hand, traditional CNN model don't consider the weights of learning instances which reduces the accuracy of image recognition.

Aiming at the problems above, we proposed one optimized CNN based image recognition model. Firstly, target region selected by bottom-up region proposals contributes to retrieve the target region of each learning instance. Secondly, enhancement weight based model is used to optimize the CNN model contributing to make full use of different learning instances. At last, adequate experiments show our method's superiority, especially compared to some other traditional methods.

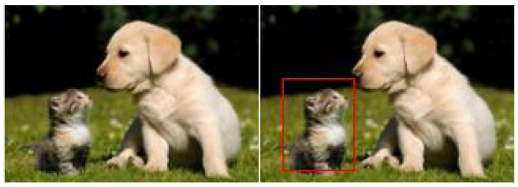
1. Introduction

As one technique which is used to recognize the target object of image, image recognition [1,2] has been widely used in the last few years. In the traditional recognition process, some classic feature descriptors, such as SIFT [3], GIST [4] and HOG [5], could extract the features of images effectively. Moreover, some optimized feature descriptors [6–11] are also presented in the subsequent work. More importantly, compared to traditional feature descriptor, such as color histogram [12], texture histogram [13], they could extract the high-level semantic information from the target image. Within a long duration, SVM [14–16] combined with some high-level feature descriptors is considered as mainstream method of image recognition. Furthermore, image classification [17,18], image retrieval [19], and image annotation [20,21] have taken the same way basically. Even though some innovative methods have been proposed, they are also on the foundation of feature extraction improvement and classification model improvement. However, deep-seated relationship between different pixels couldn't be extracted by previous feature extraction model.

Aiming at the problems above, deep learning model [22] attracts nearly all people's attention using overwhelming experimental results. Deep information could be extracted effectively depending on a large number of learning instances and complicated layered structure. From all of them, CNN-based methods [23–25], RBM-based methods [26–28], Autoencoder-

* Corresponding author.
E-mail address: junqiguobnu@163.com (J. Guo).

https://doi.org/10.1016/j.ijleo.2017.11.153
0030-4026/



(a) (b)

Fig. 1. Image (a) is one learning instance that the target region (cat) only occupies a small portion of the whole image, image (b) shows the target region using one red sliding window. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

based methods [29–31] and Sparse coding-based Methods [32–34] are often considered as four classic categories. In most cases, CNN-based methods contain convolutional layers, pooling layers, and fully connected layers which effectively extract the deep features. Restricted Boltzmann Machine (RBM) could be considered as a generative stochastic neural network which often concludes visible units and hidden units. Encodings are the essential process of learning the features of target images. Sparse coding is a class of unsupervised methods for learning sets of over-complete bases to represent data efficiently. The aim of sparse coding is to find a set of basis vectors that can represent an input vector as a linear combination of these basis vectors .When the theory is used to image, it could represent the image effectively.

The accuracy of image recognition is indeed improved significantly along with the development of deep learning. However, some existing problems also reduce the quality of image recognition. Firstly, traditional CNN model don't consider the weights of different learning instances. In other words, different learning instances are treated equally which reduces the accuracy of recognition obviously. Moreover, CNN is not optimized which increases the computing burden. Secondly, in most cases, if one image is selected as learning instance, the whole regions of image are used for training the classification model. Actually, for most images, the majority of regions are not suitable for feature extraction. For instance, in Fig. 1, the cat just occupies a small part of the target image, but the other unrelated regions are also used for learning the recognition model which directly reduces the accuracy of recognition. So in this paper, we proposed one optimized CNN based image recognition model. The main contributions of this paper are in the following:

(1) Target region could be selected by bottom-up region proposals which avoids the negative influence of unrelated regions.
(2) Enhancement weight based model is used to optimize the CNN model which makes the recognition model simulate the distribution of learning instances more accurately.

2. Algorithm

In the process of image recognition proposed by our paper, learning instance optimization, target region selection and enhancement weight consideration are the essential parts which could contribute to the improvement of recognition significantly.

2.1. Learning instance optimization

Although many learning instances are effective for learning the final recognition model, some double learning instances are often used which obviously increase the computing burden. In response to the challenge of double learning instance existence, GIST-based model [35] is used to delete the double learning instance. More concretely, we could represent the learning instance using the GIST descriptor. Then we could calculate the similarity between different learning instances. If one learning instance is similar to each other, we just keep one learning instance through deleting the redundant learning instances.

Compared to some other optimization models [36–38], our model is not complicated and comprehensive. But it is still very effective which could contribute to save the computing resource and time obviously. Especially for the images of scene category, this method shows its effectiveness for reducing the computing burden obviously.

2.2. Target region selection

Sliding window based model [39] could orientate the target region in a certain extent and some previous methods based on it have already played an important role in the field of computer vision. However, for some special images concluding more objects, it is difficult to orientate the target object exactly. In response to this challenge, we study thoughts of R-CNN model [40]. As the essential algorithm of R-CNN model, bottom-up region proposal is used to assume many regions selection.

In this paper, we generate around 2000 categories-independent region proposals for the input images regardless of the region's shape. After selection of 2000 categories-independent region proposals, how to select the most suitable target region has become one challenge. In response to this challenge, the minimum energy function is used to evaluate which region is the most suitable target region in our paper.

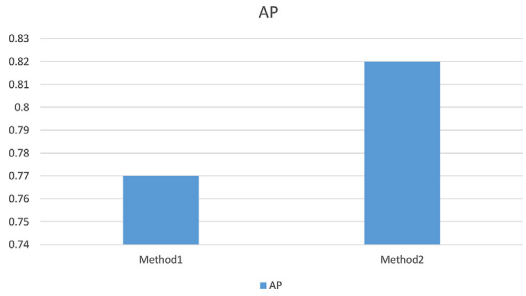


Fig. 2. The mean average precision (mAP) using different methods. Method1: Common method [41]. Method2: Target region selection based model.

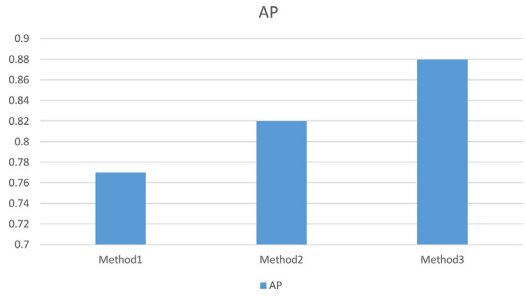


Fig. 3. The mean average precision (mAP) using different methods. Method1: Common method [41]. Method2: Target region selection based model. Method3: Optimized CNN based model in this paper.

Compared to original learning instance, the main content of the image is used effectively through our model. Based on it, features would be extracted more fully which directly supports the final recognition accuracy.

2.3. Enhancement weight consideration

Similar to structure and parameters in the paper [41], we extract a 4096-dimensional feature vector from the target region using Caffe [42] model. The model is introduced by Krizhevsky and features are computed by forward propagating a mean-subtracted 227*227 RGB image through five convolutional layers and two fully connected layers. The network architecture details could be learnt from papers [42,43].

In this step, we refer to the idea of reinforcement learning. Reinforcement learning (RL) is about an agent interacting with the environment, learning an optimal policy, by trail and error, for sequential decision making problems in a wide range of fields in both natural and social sciences, and engineering.

Based on the thoughts of reinforcement learning, we adapted enhancement weight based model to make full use of different learning instances. More concretely, if the learning instances are the target images, we give more weights. Otherwise, they are given as less weights. Compared to some other traditional models, our model could make full use of different learning instances. Through using the learning instances with different weights, the parameters and architecture could be adjusted through different cost functions.

3. Experiments

Some images of previous databases are too idealistic which is not inconsistent with the actual situation. In order to make the experimental results more convincing, we selected 128,910 images from google, Yahoo and some other websites totally. Moreover, the selected images are as complicated as possible, each image contains more than one object in most cases.

Firstly, we evaluate the contribution of target region selection through AP value. Fig. 2 shows that target region selection model improves the recognition accuracy in certain extent. Then, we combined target region selection model and enhancement weight consideration model to do more experiments. As shown in Fig. 3, enhancement weight consideration model further improves the accuracy of recognition on the foundation of target region selection model. Next, in order to show our method's extensive applications, we applied our method to different deep learning models. In the process, GoogLeNet [44], VGG [45]. SPP [46] and AlexNet [47] are used as baselines to evaluate our method's generalization ability. Figs. 4, 5 show that our model could be used to improve other deep learning models' power effectively evaluated by AP value and AUC value respectively. Figs. 6–9 show some groups of recognition results using our model.

After adequate experiments, we could see that the experimental result is consistent with theory expectation. On the one hand, our model could achieve realistic recognition results through target region selection and enhancement weight consideration. On the other hand, our model could improve other recognition models' power significantly.

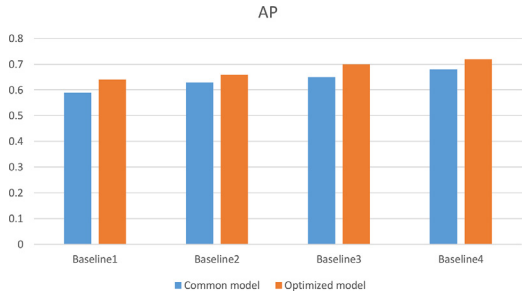


Fig. 4. The mean average precision (mAP) of recognition between different methods. The orange histograms show the mAP of different methods combined with our model. The blue histograms show the mAP of different methods. Baseline 1: GoogLeNet [44]. Baseline 2:VGG [45]. Baseline 3:SPP [46]. Baseline 4:AlexNet [47]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

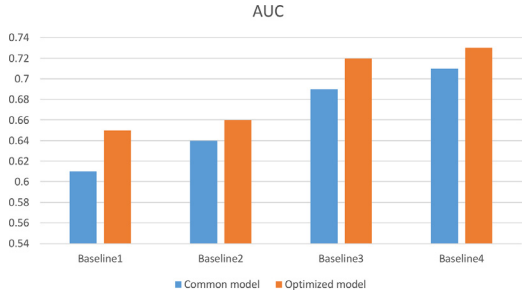


Fig. 5. Receiver Operating Characteristic (AUC) of recognition between different methods. The orange histograms show the AUC of different methods combined with our model. The blue histograms show the AUC of different methods. Baseline 1: GoogLeNet [44]. Baseline 2:VGG [45].Baseline 3:SPP [46]. Baseline 4:AlexNet [47]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

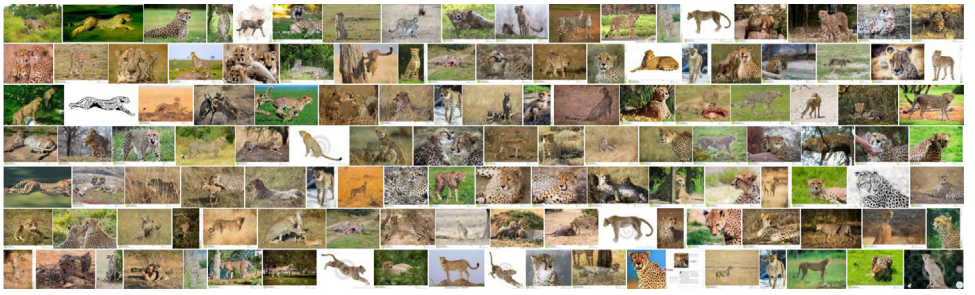


Fig. 6. Recognition results using our model. (cheetah).

Fig. 7. Recognition results using our model. (dragonfly).

Fig. 8. Recognition results using our model. (bridge).

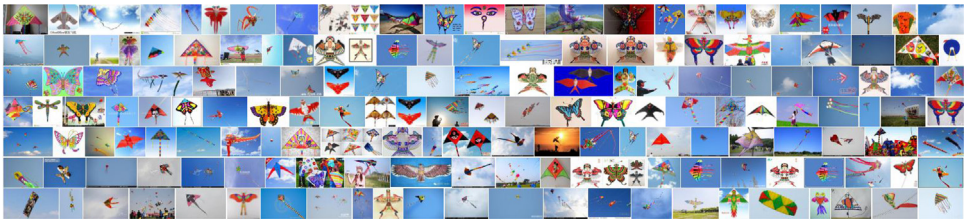


Fig. 9. Recognition results using our model. (Kite).

4. Conclusion

In this paper, we presented one optimized CNN based image recognition model combined with bottom-up region proposals. Target region selection and enhancement weight consideration are the main contributions of our paper. Based on these improvements above, adequate experiments using a larger number of images are also used to show our method's validness and robustness.

However, further improvements are still needed in the future. For instance, if the target region is not selected accurately, the error could be enhanced as a result of ineffective selection. Moreover, the enhancement weight is based on subjective evaluation but subjective evaluation would have certain tendency. In summary, target region selection model and enhancement weight consideration model need further improvements to make our proposals more convincing.

Acknowledgement

This research is sponsored by National Natural Science Foundation of China (Nos. 61571049, 61371185, Nos. 61601033, 61401029), Fundamental Research Funds for the Central Universities (No. 2016NT14), China Postdoctoral Science Foundation Funded Project (No. 2016M591109) and Beijing Advanced Innovation Center for Future Education (BJAICFE2016IR-004).

References

[1] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).
[2] K. He, X. Zhang, S. Ren, et al., Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[3] D.G. Lowe, Distinctive image features from scale-invariant keypoints, J. Comput. Vis. 60 (2) (2004) 91–110.
[4] Aude Oliva, Antonio Torralba, Building the gist of a scene: the role of global image features in recognition, Prog. Brain Res. 155 (2006) 23–36.
[5] A. Bosch, A. Zisserman, X. Munoz, Representing shape with a spatial pyramid kernel, in: Pro. ACM International Conference on Image and Video Retrieval, ACM, New York, NY, 2007, pp. 672–679.
[6] Yan-Tao Zheng, et al., Toward a higher-level visual representation for object-based image retrieval, Vis. Comput. 25 (1) (2009) 13–23.
[7] E. Bart, S. Ullman, Cross-generalization: learning novel classes from a single example by feature replacement, in: Pro. IEEE Conference on Computer Vision and Pattern Recognition, IEEE, San Diego, CA, 2005, pp. 672–679.
[8] A. Torralba, K.P. Murphy, Sharing visual features for multiclass and multiview object detection, IEEE Trans. Pattern Anal. Mach. Intell. 29 (5) (2007) 854–869.
[9] H. Wu, Y. Li, Z. Miao, et al., Creative and high-quality image composition based on a new criterion, J. Vis. Commun. Image Represent. 38 (2016) 100–114.
[10] H. Wu, Y. Li, Z. Miao, et al., A new sampling algorithm for high-quality image matting, J. Vis. Commun. Image Represent. 38 (2016) 573–581.
[11] Kai Kunze, et al., The Wordometer–Estimating the Number of Words Read Using Document Image Retrieval and Mobile Eye Tracking, in: Document Analysis and Recognition (ICDAR), 2013 12th International Conference on. IEEE, 2013.
[12] Michael J. Swain, Dana H. Ballard, Indexing via color histograms, in: Active Perception and Robot Vision, Springer, Berlin, Heidelberg, 1992, pp. 261–273.
[13] Chad Carson, et al., Blobworld: a system for region-based image indexing and retrieval, in: Visual Information and Information Systems, Springer, Berlin, Heidelberg, 1999.
[14] Cho-Jui Hsieh, et al., A dual coordinate descent method for large-scale linear SVM, in: Proceedings of the 25th International Conference on Machine Learning, ACM, 2008.
[15] Mahesh Pal, Giles M. Foody, Feature selection for classification of hyperspectral data by SVM, IEEE Trans. Geosci. Remote Sens. 48 (5) (2010) 2297–2307.
[16] Serafeim Moustakidis, et al., SVM-based fuzzy decision trees for classification of high spatial resolution remote sensing images, IEEE Trans. Geosci. Remote Sens. 50 (1) (2012) 149–169.
[17] S. Maji, A.C. Berg, Max-margin additive classifiers for detection, in: Proc. IEEE International Conference on Computer Vision, IEEE, Kyoto, 2009, pp. 40–47.
[18] N. Kumar, et al., Attribute and simile classifiers for face verification, in: Proc. IEEE International Conference on Computer Vision, IEEE, Kyoto, 2009, pp. 365–372.
[19] Z. Zha, et al., Joint multi-label multi-instance learning for image classification, in: Proc. IEEE International Conference on Computer Vision, Anchorage, AK, 2008, pp. 1–8.
[20] B.C. Russell, et al., LabelMe:a database and web-based tool for image annotation, Int. J. Comput. Vis. 77 (1-3) (2008) 157–173.
[21] A. Ulges, et al., Indentifying relevant frames in weakly labeled videos for training concept detectors, in: Proceedings of International Conference on Content-Based Image and Video Retrieval, Niagara Falls, Canada, 2008, pp. 9–16.
[22] Y. Guo, Y. Liu, A. Oerlemans, et al., Deep learning for visual understanding: a review, Neurocomputing 187 (2016) 27–48.
[23] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. (2012) 1097–1105.
[24] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, Springer International Publishing, 2014, pp. 818–833.

[25] K. He, X. Zhang, S. Ren, et al., Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1904–1916.
[26] G.E. Hinton, S. Osindero, Y.W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.
[27] R. Salakhutdinov, G.E. Hinton, Deep Boltzmann machines, in: AISTATS, 2009, 1: 3.
[28] J. Ngiam, Z. Chen, P.W. Koh, et al., Learning deep energy models, in: Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 1105–1112.
[29] C. Poultney, S. Chopra, Y.L. Cun, Efficient learning of sparse representations with an energy-based model, in: Advances in Neural Information Processing Systems, 2006, pp. 1137–1144.
[30] P. Vincent, H. Larochelle, Y. Bengio, et al., Extracting and composing robust features with denoising autoencoders, in: Proceedings of the 25th International Conference on Machine Learning, ACM, 2008, pp. 1096–1103.
[31] S. Rifai, P. Vincent, X. Muller, et al., Contractive auto-encoders: explicit invariance during feature extraction, in: Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 833–840.
[32] R. Memisevic, U. Ca, D. Krueger, Zero-bias autoencoders and the benefits of co-adapting features, Stat 1050 (13) (2014).
[33] X. Zhou, K. Yu, T. Zhang, et al., Image classification using super-vector coding of local image descriptors, in: European Conference on Computer Vision, Springer, Berlin, Heidelberg, 2010, pp. 141–154.
[34] S. Gao, I.W.H. Tsang, L.T. Chia, et al., Local features are not lonely–Laplacian sparse coding for image classification, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010, pp. 3555–3561.
[35] J. Hays, A.A. Efros, Scene completion using millions of photographs, in: ACM Transactions on Graphics (TOG), ACM, 2007, p. 4, 26 (3).
[36] H. Wu, Z. Miao, Y. Wang, et al., Optimized recognition with few instances based on semantic distance, Vis. Comput. 31 (4) (2015) 367–375.
[37] Y. Li, R. Bie, C. Zhang, et al., Optimized learning instance-based image retrieval, Multimedia Tools Appl. (2016) 1–18.
[38] H. Wu, Z. Miao, Y. Wang, et al., Image completion with multi-image based on entropy reduction, Neurocomputing 159 (2015) 157–171.
[39] P. Sermanet, K. Kavukcuoglu, S. Chintala, et al., Pedestrian detection with unsupervised multi-stage feature learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3626–3633.
[40] R. Girshick, J. Donahue, T. Darrell, et al., Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.
[41] S. Shankar, D. Robertson, Y. Ioannou, et al., Refining architectures of deep convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2212–2220.
[42] Y. Jia, Caffe: An Open Source Convolutional Architecture for Fast Feature Embedding (2013) http://caffe.berkeleyvision.org/.
[43] A. Krizhevsky, I. Sutskever, G. Hinton, ImageNet classification with deep convolutional neural networks, in: NIPS, 2012.
[44] C. Szegedy, W. Liu, Y. Jia, et al., Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[45] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
[46] K. He, X. Zhang, S. Ren, et al., Spatial pyramid pooling in deep convolutional networks for visual recognition, in: European Conference on Computer Vision, Springer International Publishing, 2014, pp. 346–361.
[47] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.