TRAFFIC SIGN RECOGNITION USING VISUAL ATTRIBUTE LEARNING AND CONVOLUTIONAL NEURAL NETWORK

RONG-QIANG QIAN[1], YONG YUE[1], FRANS COENEN[2], BAI-LING ZHANG[1]

[1]Department of Computer Science and Software Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, P.R.China.
[2]Department of Computer Science, University of Liverpool, Liverpool L69 3BX, United Kingdom
E-MAIL: rongqiang.qian@student.xjtlu.edu.cn

Abstract:

The problem of extracting high level information from digital images and videos is frequently faced in the area of computer vision and machine learning. For the recognition of traffic signs, a lot of outstanding methods have been proposed, and deep models demonstrates their powerful representation capacity, achieved dominant performances. In this paper a method for recognizing traffic signs is proposed founded on a novel visual attribute mechanisms; whereby attributes are generated using Convolutional Neural Networks (CNN). In comparison with previous methods founded on the use of CNN for feature extractor and Multi-Layer Perception (MLP) as classifier, the Max Pooling Positions (MPPs) proposed in this paper predict visual attributes that provide a useful linkage between low-level features and high-level sematic tasks. The results show that outstanding performances can be achieved using MPPs.

Keywords:

Advanced Driver Assistance; Traffic sign recognition; Deep learning; Convolutional neural networks; Max pooling; Visual attributes

1. Introduction

Traffic Sign Recognition (TSR) using computer vision and machine learning technology is a key component with respect to many applications, such as self-driving, driverless vehicles, traffic mapping and traffic surveillance. Recognition of traffic signs allows a driver to be warned of inappropriate actions and potential dangers. The dataset German Traffic Sign Recognition Benchmark (GTSRB) [1] provides a number of difficult challenges due to "open-set problems" such as viewpoint variations, poor illumination conditions, motion-blur, occlusions, colors fading, and low resolutions. As a consequence, the recognition of traffic signs has been significantly developed, and a number of outstanding approaches have been reported in the literatures [2, 3, 4].

Due to the huge success achieved by the large Convolutional Neural Networks (CNN) [5] proposed for image-level object recognition on the ImageNet challenge [6]. Deep learning has attracted much attention in computer vision research as more and more promising results are published on a range of different vision tasks. It has been demonstrated that CNN are able to automatically learn and extract the optimized features that richly describe image content as well as allowing for discrimination between object categories. Although excellent performance has been achieved by CNN, exploring, understanding and interpreting the internal working principles of CNN remains an open problem. Some pioneering works have been conducted; for examples in [7, 8] have shown that features extracted from a pre-trained CNN model using ImageNet have the potential for application to other computer vision tasks, including object detection, fine-grained recognition and action recognition.

Recently activation schemes [9, 10, 11, 12] have present a mechanism for investigating and understanding CNN models. For example, in fine-grained recognition tasks, the required part detectors can be discovered by analyzing the gradient maps of the network outputs and finding activation centers spatially related to annotated semantic parts or bounding boxes. Moreover, the relationship between visual attributes and convolutional networks is discussed in [13], Attribute Centric Nodes (ACNs) are selected and used for predicting visual attributes.

Inspired by previous researches that aims at visualizing and activating CNNs for performing particular tasks, the work presented in this paper focus on exploring the relationships between visual attributes and CNN using Max Pooling Positions (MPPs), as illustrated in Fig. 1. Instead of directly using features extracted from CNN for training classifies, which is com-

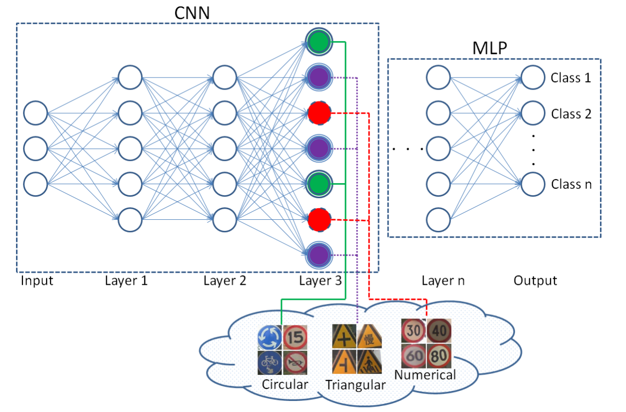

FIGURE 1. Based on a pre-trained CNN network, visual attributes seem to be encoded by the network's activations. This work focuses on finding the exact activation locations for each specific visual attributes.

mon to all of the CNN models, we use MPPs to sample and encode the features; the encoded pooling sequence will then be used to predict visual attributes. The main contributions of our proposed system are:

- A CNN model to learn a compact yet discriminative feature representation.
- A novel method to discover the relationship between visual attributes and CNN based on MPPs.

The rest of this paper is organized as follows: Section 2 outlines some related research on traffic sign recognition, visual attributes and CNN applications; Section 3 provides a detailed description of the proposed system; Implementation details and experimental results will be provided in Section 4, followed by conclusion in Section 5.

2. Related works

2.1. Traffic sign recognition

TSR has been an active area of research in the computer vision community for many years. The common approaches for TSR using classification can be divided into two groups: (i) multi-class classification whereby traffic signs are classified according to a set of class labels (more than two) and (ii) binary-classification where traffic signs are classified in terms of a two class problem. Among the many classification (prediction) models, Support Vector Machines (SVMs) have provided good performance [14, 15]. A robust sign similarity measurement using SimBoost and fuzzy regression tree was proposed in [16]. An ensemble of classifiers, based on the Error-Correcting Output Code (ECOC) framework, was introduced in [17], where the ECOC framework was designed using a forest of optimal tree structures that are embedded in the ECOC matrix. Due to the powerful representational learning capabilities of CNN, CNN models have become the dominant approaches for traffic sign recognition in recent years. For example, committee CNN [3], multi scale CNN [4], multi column CNN [18] and hinge-loss CNN [19].

2.2. Visual attributes

The works presented in [20, 21] expounded the advantages of using visual attributes in various vision tasks, and visual recognition in particular. Visual attributes are human-nameable sematic properties shared by objects, which provide a useful linkage between low-level features and high-level categorical labels. The attributes of interest can be categorised into two types: (i) visual attributes and (ii) data driven attributes. The first are defined by humans and tend to be represented by words which have a clear sematic meaning. The second are discovered from the data. In recent years, studies directed at attribute-based representations have drawn a lot of attention, with representational works including Zero-Shot Learning (ZSL) [22], action recognition [23], and scenes representation [24, 25].

2.3. Convolutional neural network

A CNN is a special type of multi-layer neural network that extracts features by combining convolution, pooling and activation layers. The most successful CNN architecture [5] is trained usin back-propagation; excellent performance has been achieved using benchmark data sets. Although CNN models have been proven to have powerful description capabilities, it remains unclear how features are learned inside the network. This lack of understanding of the internal operation of CNNs has attracted significant research interest directed at a desired to seek deeper insight into its working principle. For example, Part Detector Discovery (PDD) has been proposed based on analyzing the gradient maps of the network outputs and finding activation centers [9]. An unsupervised fine-grained recognition scheme was introduced in [10] whereby part models were generated by finding constellations of neural activation patterns. In [11], the visualisation of image classification models were displayed based on computing the gradient of the class score with respect to the input image. Motivated by the wish to visualize and understand CNNs, the work presented in [12] proposed a novel scheme to visualize activations based on a multi-layered Deconvolutional Network (deconvnet). The relationship between visual attributes and convolutional networks is discussed in [13], where Attribute Centric Nodes (ACNs) are used for predicting visual attributes.

3. Approach

An overview of the proposed recognition system is shown in Fig. 2. In the pre-processing stage, the input image will be normalized by using Contrast-Limited Adaptive Histogram Equalization (CLAHE) [26]. Then, the normalized image will be passed to a CNN model for extracting discriminative features. Finally, MPPS is adopted to predict each visual attributes of the input image.

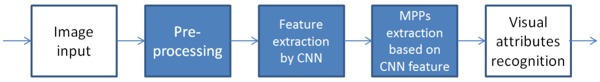
Image input
Pre-processing
Feature extraction by CNN
MPPs extraction based on CNN feature
Visual attributes recognition

FIGURE 2. System overview

3.1. Network architecture

The basic network structure of our proposed model is similar to the work in [3], which can be illustrated as shown in Fig. 3. In the training stage, the network consists of three convolution stages followed by fully connection layers and softmax layer. Each convolution stage includes convolutional layer, non-linear activation layer and max pooling layer. ReLU [5] is employed as the activation function for the convolutional layers and the full connection layers. The final softmax layer has 26 outputs, corresponding to each category in our traffic sign dataset. On the other hand, in the testing stage, the original full connection layers and softmax layer are replaced with several visual attribute classifiers. Each of the classifiers implements a two-class classification. Instead of training all the classifiers using standard back-propagation, all the parameters are simply selected and fine-tuned by our MPPs method. The details will be elaborated on in the next section.

3.2. Visual attributes prediction by max pooling positions

Since the final layer of the proposed CNN has 250 feature maps with 4 × 4 neurons, the extracted features have a dimension of 4000. As Fig. 4 demonstrates, for each 4 × 4 feature map, max pooling with a kernel size 2 × 2 and stride of 2 is performed, with the position of each max value being recorded. Then, each recorded position is encoded into a four bits binary value. Finally, the whole MPPs sequence can be obtained by concatenating all the binary values. The dimension of MPPs sequence is also 4000.

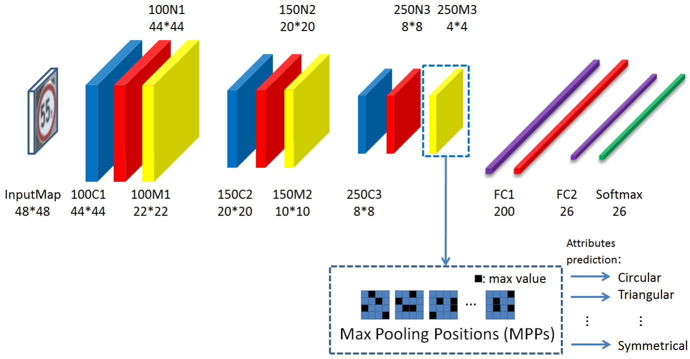

FIGURE 3. Network architecture

For each visual attribute classifier, the training stage can be described as follows.

Step 1. Data collection. All the available data is divided into two class based on existence of current visual attributes.

Step 2. Data processing. In order to measure the similarities of MPPs belonging to the same class, all of the MPPs sequences from same class are accumulated together and normalized by dividing by the number of samples. In this manner we get a series of sequences that indicate the probability of appearance for each of the max value positions. Therefore, two probability sequences are achieved, namely, $p_{positive}$ and $p_{negative}$.

Step 3. Activation selection. The purpose of this step is to select a decision matrix for each visual attribute classifier with a dimension of 2 × 4000. The corresponding decision matrix $d(x, y)$ can be computed by comparing the magnitude of the probability of each channel in $p_{positive}$ and $p_{negative}$; a channel is activated according to the larger probability, the activation value for the selected channel is set to be $|p_{positive}(y) - p_{negative}(y)|$ and that for the other channel to 0.

Step 4. Fine-tuning. The obtained decision matrix $d(x, y)$ needs further tuning. For each iteration, if the current predicted label is wrong, the corresponding decision matrix is fine-tuned by adding the product of the current MPPs sequence and a learning rate. The decision matrix is normalized after each iteration. Normally, the tuning requires only a few iterations.

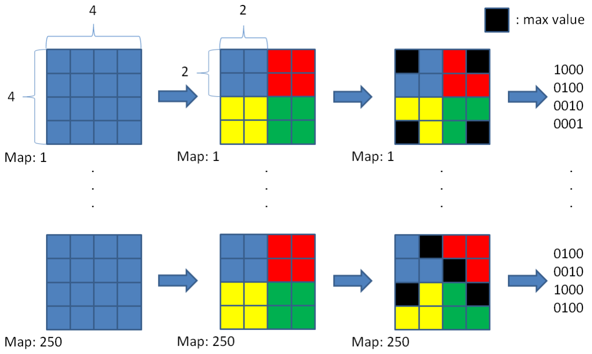

FIGURE 4. Classification by max pooling positions (MPPs)

4. Experiment

In this section, we will introduce the implementation details of our CNN model, including architecture selection and training. The proposed MPPs based visual attribute recognition method will then be discussed. Finally, performance comparison will be provided. To evaluate the performance of our traffic sign recognition system, a computer with Xeon 3.3GHz CPU and 24GB memory was employed. To accelerate the training, a very efficient Titan GPU based implementation was built using NVIDIA CUDA and CUDNNv3. The details will be introduced with the corresponding experiments in the following.

4.1. Implementation details

The basic network structure of our proposed model is similar to that described in [3]. However, instead of using hyperbolic tangent as the activation function, ReLU [5] is employed.

4.2. Pre-training by GTSRB dataset

The pre-training scheme for our system is as follows: (i) 43 categories of training data were acquired and pre-processed from the GTSRB [1]; (ii) initial weights of the convolution layers and full connection layers were allocated using a uniform random distribution in the range [-0.05, 0.05]; (iii) the learning rate was set to 0.001 and the training will be terminated after 200 epochs ;(iv) cross-entropy loss and stochastic gradient descent [5] were used during training.

4.3. Fine-tuning by self-captured dataset

The fine-tuning was continued with our traffic sign dataset, which contains 13340 traffic signs that incorporated 26 classes, 70% of the images were used for training and the rest for testing. The adopted fine-tuning scheme was as follows: (i) all of the CNN parameters were kept unchanged except that the final 43-way classification layer was replaced with a randomly initialized 26-way classification layer; (ii) the learning rate was set to 0.001; (iii) training was conducted using cross-entropy loss and stochastic gradient descent.

4.4. Traffic sign recognition by visual attributes

The visual attributes adopted with resect to the reported experimentation are listed in Table 1. From the table it can be seem that 11 visual attributes were identified based on the shapes, colors and content of the traffic signs. The corresponding prediction accuracy of each attribute is given in the rightmost column of Table 1. Since visual-level attributes are not sufficient for classifying all the traffic signs, we also use category-level attributes. The recognition process was as follows: (i) giving an input image, the visual-level attributes are labeled; (ii) compare the predicted labels with the attributes table, and choose the nearest category; (iii) checking whether category-level attributes are needed. The recognition rate is about 94.63% based on the training network. By introducing the proposed recognition scheme, the overall accuracy rate is about 95.53%, an encouraging result. Some of the unsuccessful examples are given in Fig. 5. The main possible reasons for the failed recognition include: (i) motion-blur; (ii) poor illumination; (iii) disadvantageous viewpoint variations.

TABLE 1. Results of attribute prediction

														5	15	30	40	60	80		让	停	慢				Accuracy
Numbers:	385	410	875	755	475	910	765	825	600	785	310	370	275	270	250	500	530	715	390	525	585	260	395	820	360	865	
Circular	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0.988
Triangular	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1	0.987
Symmetrical	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0.975
Numerical	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0.989
Blue	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.994
Red	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0.990
Yellow	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0.994
People	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0.979
Chinese	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0.984
Arrow	0	1	1	1	1	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.973
Vehicle	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0.981
Category-level attribute:	5 15 30 40 60 80																										

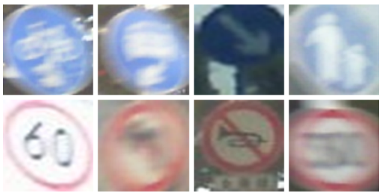

FIGURE 5. Example images where the recognition failed.

5. Conclusion

In this paper, a novel visual attribute based traffic sign recognition system is proposed. The main contributions of the work include: (i) a CNN model to learn a compact yet discriminative feature representation; (ii) a novel method to discover the relationship between visual attributes and convolutional networks based on MPPs. By introducing visual attributes for recognition, excellent prediction performance was achieved.

References

[1] J. Stallkamp, M. Schlipsing, J. Salmen, C. Igel, Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition, Neural Networks 32 (2012) 323 – 332, selected Papers from {IJCNN} 2011. doi:http://dx.doi.org/10.1016/j.neunet.2012.02.016.

[2] M. Mathias, R. Timofte, R. Benenson, L. Van Gool, Traffic sign recognition - how far are we from the solution?, in: Neural Networks (IJCNN), The 2013 International Joint Conference on, 2013, pp. 1–8. doi:10.1109/IJCNN.2013.6707049.

[3] D. Ciresan, U. Meier, J. Masci, J. Schmidhuber, A committee of neural networks for traffic sign classification, in: Neural Networks (IJCNN), The 2011 International Joint Conference on, 2011, pp. 1918–1921. doi:10.1109/IJCNN.2011.6033458.

[4] P. Sermanet, Y. LeCun, Traffic sign recognition with multi-scale convolutional networks, in: Neural Networks (IJCNN), The 2011 International Joint Conference on, 2011, pp. 2809–2813. doi:10.1109/IJCNN.2011.6033589.

[5] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: F. Pereira, C. Burges, L. Bottou, K. Weinberger (Eds.), Advances in Neural Information Processing Systems 25, Curran Associates, Inc., 2012, pp. 1097–1105.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.

[7] A. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, Cnn features off-the-shelf: An astounding baseline for recognition, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on, 2014, pp. 512–519. doi:10.1109/CVPRW.2014.131.

[8] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Decaf: A deep convolutional activation feature for generic visual recognition, in: Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014, 2014, pp. 647–655.

[9] M. Simon, , E. Rodner, , J. Denzler, Computer Vision – ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part II, Springer International Publishing, Cham, 2015, Ch. Part Detector Discovery in Deep Convolutional Neural Networks, pp. 162–177.

[10] M. Simon, E. Rodner, Neural activation constellations: Unsupervised part model discovery with convolutional networks, in: International Conference on Computer Vision (ICCV), 2015.

[11] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, CoRR abs/1312.6034.

[12] M. D. Zeiler, R. Fergus, Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I, Springer International Publishing, Cham, 2014, Ch. Visualizing and Understanding Convolutional Networks, pp. 818–833.

[13] V. Escorcia, J. Niebles, B. Ghanem, On the relationship between visual attributes and convolutional networks, in: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, 2015, pp. 1256–1264. doi:10.1109/CVPR.2015.7298730.

[14] S. Maldonado-Bascon, S. Lafuente-Arroyo, P. Gil-Jimenez, H. Gomez-Moreno, F. Lopez-Ferreras, Road-sign detection and recognition based on support

vector machines, Intelligent Transportation Systems, IEEE Transactions on 8 (2) (2007) 264–278. doi:10.1109/TITS.2007.895311.

[15] M. Shi, H. Wu, H. Fleyeh, Support vector machines for traffic signs recognition, in: Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on, 2008, pp. 3820–3827. doi:10.1109/IJCNN.2008.4634347.

[16] A. Ruta, Y. Li, X. Liu, Robust class similarity measure for traffic sign recognition, Intelligent Transportation Systems, IEEE Transactions on 11 (4) (2010) 846–855. doi:10.1109/TITS.2010.2051427.

[17] X. Baro, S. Escalera, J. Vitria, O. Pujol, P. Radeva, Traffic sign recognition using evolutionary adaboost detection and forest-ecoc classification, Intelligent Transportation Systems, IEEE Transactions on 10 (1) (2009) 113–126. doi:10.1109/TITS.2008.2011702.

[18] D. Ciresan, U. Meier, J. Masci, J. Schmidhuber, Multi-column deep neural network for traffic sign classification, Neural Networks 32 (2012) 333 – 338, selected Papers from {IJCNN} 2011. doi:http://dx.doi.org/10.1016/j.neunet.2012.02.023.

[19] J. Jin, K. Fu, C. Zhang, Traffic sign recognition with hinge loss trained convolutional neural networks, Intelligent Transportation Systems, IEEE Transactions on 15 (5) (2014) 1991–2000. doi:10.1109/TITS.2014.2308281.

[20] C. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 2009, pp. 951–958. doi:10.1109/CVPR.2009.5206594.

[21] A. Farhadi, I. Endres, D. Hoiem, D. Forsyth, Describing objects by their attributes, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 2009, pp. 1778–1785. doi:10.1109/CVPR.2009.5206772.

[22] S. Antol, C. L. Zitnick, D. Parikh, Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV, Springer International Publishing, Cham, 2014, Ch. Zero-Shot Learning via Visual Abstraction, pp. 401–416.

[23] Y. Fu, T. M. Hospedales, T. Xiang, S. Gong, Computer Vision – ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, Ch. Attribute Learning for Understanding Unstructured Social Activity, pp. 530–543.

[24] G. Patterson, J. Hays, Sun attribute database: Discovering, annotating, and recognizing scene attributes, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, 2012, pp. 2751–2758. doi:10.1109/CVPR.2012.6247998.

[25] J. Shao, K. Kang, C. C. Loy, X. Wang, Deeply learned attributes for crowded scene understanding, in: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, 2015, pp. 4657–4666. doi:10.1109/CVPR.2015.7299097.

[26] K. Zuiderveld, Contrast limited adaptive histogram equalization, in: Graphics gems IV. Academic Press Professional, 1994, pp. pp. 474–485.