Traffic Sign Recognition using Spatial Transformer Network with Multi-Structure Convolutional Neural Network

Myo Taeg Lim^{*1} Dong Won Kim^{*2}, Jun Ho Chung[#], Woo Jin Ahn^{\$}, Sang Kyoo Park[@], Tae Koo Kang[&]
^{*1\$@}School of Electrical Engineering, Korea University, Anam-ro, Seongbuk-gu, Seoul, Korea

^{*2}Department of Digital Electronics, Inha Technical College, Inha-ro, Nam-gu, Incheon, Korea

[&]Department of Human Intelligence and Robot Engineering, Sang Myung University, Sangmyungdae-gil, Dongnam-gu, Cheonan-si, Chungcheongnam-do, Korea

(Received 26 September, 2019: accepted 28 December, 2019)

<https://doi.org/10.36224/ijes.120401>

Abstract

Convolutional Neural Network is one of state-of-the-arts and demonstrated its superior performance in various computer visions systems recently. The conventional convolutional neural network has one-way structure to train image information with fixed sizes of filters in general. However, this structure only learns image information followed by one fixed sizes of filters and this is not the best to achieve high performance of the network. In order to achieve high performance of the network, this paper suggests a novel convolutional neural network which consists of spatial transformer network and multi-structure convolutional neural network. Spatial transformer network is robust against distorted images. Multi-structure convolutional neural network uses different sizes of filters for global and local information from the given images. The proposed algorithm, spatial transformer with multi-structure convolutional neural network (SPMCNN) demonstrates its classification performance on German traffic sign recognition benchmark.

Keywords: Spatial transformer network, convolutional neural network, traffic sign recognition, multi-structure, classification

1. Introduction

Many researchers had studied to improve image classification performance through Bayesian classifiers, tree classification, and boosting [1]. When “AlexNet” [2] won classification competition, convolutional neural network (CNN) has become one of state-of-the arts and demonstrated its significant performance in various computer vision systems [3 - 5].

Due to significant performance of CNN, some researchers have studied traffic sign recognition because it has become one of main prior research areas for self-driving automobile. The studies of traffic sign recognition can be divided into two parts which are traffic sign detection and traffic sign classification. The main purpose of traffic sign detection is to find the location of traffic signs from the given images or videos while traffic sign classification is to determine the sub-classes [6]. Human drivers classify traffic signs on the road easily, but it becomes challenging tasks for computer vision because of viewpoint-dependent object variability, and the high in-class variability of many object types [7].

In this paper, a novel CNN, spatial transformer with multi-structure convolutional neural network (STMCNN), is proposed for improving classification performance. Many conventional CNNs create multiple feature maps with one-way structure using fixed sizes of filters and this

^{*}Corresponding authors

Email address: mlim@korea.ac.kr and dwnkim@inhac.ac.kr (Myo Taeg Lim and Dong Won Kim)

ISSN 0976 – 6693. ©2020 SCMR All rights reserved.

forces the network to learn limited information. However, the proposed STMCNN has multi ways to extract global and local feature information with different sizes of filters for learning various feature information. Assuming that input images contain external noises, spatial transformer network (STN) [8] is placed in each beginning part of the network. The main properties of STN is to modify input images by performing spatial transformation which provides clear images to the network.

The reminder of this paper is organized as follows. In Section 2, the STMCNN is proposed. In Section 3, simulation results are presented to demonstrate the performance of the STMCNN. Finally, conclusions are drawn in Section 4.

2. Spatial transformer with multi-structure convolutional neural network

In this section, the architecture of STMCNN is introduced. Figure 1 represents the proposed model in this paper. The STMCNN contains five parts which are input image, image modification using STN, filter adaptation, feature extraction, and fully connected with softmax. Figure 1 (a) shows the input image is provided. Figure 1 (b) represents how STN is operated. STN performs geometric transformation on the given image that provides the network to be spatially invariant to the input image [8]. Thus, data augmentation such as rotation, scale, or cropping is not needed. In order to obtain global and local feature information, small sizes of filters and relatively large sizes of filters are applied to each obtained image as shown in Figure 1 (c). Figure 1 (d) shows how the obtained feature maps are reduced following the process of convolution layer and max-pooling layer. Finally, entire feature information from blue and red filter is reshaped as long vector and the network classifies the class of the input image using softmax.

Figure 1: (a) Input traffic sign, (b) Modification of input image using STN, (c) Applying different sizes of filters, (d) Feature extraction (Grey and Orange: convolution and maxpooling), (e) Concatenating feature vectors (blue, red), fully connected (yellow), and softmax (green)

2.1. Architecture selection for STMCNN

For the implementation of STMCNN, localization network for STN and multi-structure convolutional neural network are constructed. The parameters for localization network are shown in Table 1. For multi-structure convolutional neural network, the number of filters for each convolution layer is decided as 128, 256, and 512. The dataset for demonstration is German Traffic Sign Recognition Benchmark (GTSRB) [9] which contains 43 classes of traffic signs.

The sizes of GTSRB is varying from 15×15 to 250×250 , the size of dataset for implementation is set to 32×32 . The sizes of filters for filter adaptation is selected as 3×3 and 5×5 , respectively. Using 3×3 sizes of filters focuses local feature information while 5×5 sizes of filters are for global feature information. Zero-padding is applied to convolution layer from feature extraction. When zero-padding is applied to convolution layer, the network maintains the dimension of the feature maps and this offers the ability to preserve the edge information from the input image. The entire parameters for STMCNN are described in Table 2.

Table 1: Architecture selection for localization network of STN

		Layer information		
		Type	Size	Channel
Operation	1	Input image	32×32	3
	2	Convolution	26×26 (7×7)	16
	3	Max-pooling	13×13	16
	4	Convolution	9×9 (5×5)	32
	5	Max-pooling	4×4	32
	6	Convolution	2×2 (3×3)	64
	7	Max-pooling	1×1	64
	8	Concatenate	1×1	
	9	Fully connected	1×1	128
	10	Fully connected	1×1	64
	11	Fully connected	1×1	6

Table 2: Architecture selection for STMCNN

		Layer information			
		Type	Size (Upper path)	Size (Below path)	Channel
Operation	1	Input image	32×32	32×32	3
	2	STN	32×32	32×32	
	3	Convolution(Zero-padding)	32×32 (3×3)	32×32 (5×5)	128
	4	ReLU	32×32	32×32	128
	5	Max-pooling	16×16	16×16	128
	6	Convolution (Zero-padding)	16×16 (3×3)	16×16 (5×5)	256
	7	ReLU	16×16	16×16	256
	8	Max-pooling	8×8	8×8	256
	9	Convolution(Zero-padding)	8×8 (3×3)	8×8 (5×5)	512
	10	ReLU	8×8	8×8	512
	11	Max-pooling	4×4	4×4	512
	12	Concatenate	1×1	1×1	
	13	Fully connected	1×1	1×1	256
	14	softmax	1×1	1×1	43

3. Experiments

The simulation results are conducted on a Core i7-4790 (3.60 GHz), 16 GB DDR3, and GTX 1080 Ti. The GTSRB dataset consists of 39,202 training dataset with 43 traffic sign classes and 12,630 test dataset. The number of training steps, batch size, and optimizer are selected as 20, 32, and Adam [10].

Table 3 shows the GTSRB classification performance comparisons between the basic CNN and STMCNN. The architecture of basic CNN follows upper path of Table 2. Using the multi-structure convolutional neural network, the network learns global and local features from the input images via different sizes of filters. Moreover, STN performs spatial transformation on the input images and this provides clearer images. Due to the properties of STN, spatial transformation behaves certain tasks such as rescale, rotation, or cropping. Thus, data augmentation for training is not required. Adopting STN and multi-structure, the classification performance of STMCNN reaches to 97.450 while basic CNN is 91.005 which is relatively low performance.

Table 3: Classification comparisons between basic CNN and STMCNN

	Basic CNN	STMCNN
Classification Accuracy (%)	91.005	97.450

4. Conclusion

In this paper, a novel CNN, spatial transformer with multi-structure convolutional neural network is proposed. The STN adjusts input images as intended before training and this provides clearer images containing distortion. Then multi-structure convolutional neural network extracts global and local feature information from the modified images. Multi-structure provides the network more image information from different properties than conventional CNN. Through demonstration, the performance of STMCNN is superior to basic CNN. Since the STMCNN outperforms the basic CNN, it is expected to be widely used in classification research.

Acknowledgement

This work was supported by the Brain Korea 21 Plus Project in 2019.

References

1. Junqi J., Kun F., and Changshui Z., *Traffic sign recognition with hinge loss trained convolutional neural networks*, IEEE Trans. Intell Transp. Syst., 2014, 15(5)
2. Alex K., Ilya S., and Geoffrey E. H., *Imagenet classification with deep convolutional neural networks*, Proc. Adv. Neural Inf. Process. Syst. 2012, 1106- 1114
3. Jia L., Maoguo G., Kai Q., and Puzhao Z., *A deep convolutional coupling network for change detection based on heterogeneous optical and radar images*, IEEE Trans. Neural Netw. Learn. Syst., 2018, 29(3) 545- 559
4. Yanwei P., Manli S., Xiaoheng J., and Xuelong L., *Convolution in Convolution for Network in Network*, IEEE Trans. Neural Netw. Learn. 2018, 29 (5) 1587- 1597
5. Li X., Jimmy S. R., Ce L., and Jiaya J., *Deep convolutional neural network for image deconvolution*, Advances in Neural Information Processing Systems, 2014, 1790- 1798

6. Yang Y., Luo H., Xu H., and Wu F., *Towards real-time traffic sign detection and classification*, IEEE Trans. Intell Transp. Syst., Volume 2016, 17 (7) 2022- 2031
7. Ciresan D., Meier U., and Schmidhuber J., *Multi-column deep neural networks for image classification*, Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, 2015, 3642- 3649
8. Max J., Karen S., Andrew Z., and Koray K., *Spatial Transformer Networks*, Advances in neural information processing systems, 2015, 2017- 2025
9. Johannes S., Marc S., Jan S., and Christian I., *Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition*, Neural Networks, 2012, 32, 323- 332
10. Diederik P. K., and Jimmy L. B., *ADAM: A method for stochastic optimization*, arXiv preprint arXiv:2014, 1412.6980