Cross-Level: A Practical Strategy for Convolutional Neural Networks Based Image Classification

Yu Liu[1], Baocai Yin[1], Jun Yu[1], and Zeng-Fu Wang[1,2(✉)]

[1] Department of Automation, University of Science and Technology of China, Hefei 230026, China
{liuyu1,yinbc}@mail.ustc.edu.cn, {harryjun,zfwang}@ustc.edu.cn
[2] Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China

Abstract. Convolutional neural networks (CNNs) have exhibited great potential in the field of image classification in the past few years. In this paper, we present a novel strategy named cross-level to improve the existing CNNs' architecture in which different levels of feature representation in a network are merely connected in series. The basic idea of cross-level is to establish a convolutional layer between two nonadjacent levels, aiming to learn more sufficient feature representations. The proposed cross-level strategy can be naturally combined into a CNN without any change on its original architecture, which makes this strategy very practical and convenient. Three popular CNNs for image classification are employed to illustrate its implementation in detail. Experimental results on the dataset adopted by the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) verify the effectiveness of the proposed cross-level strategy on image classification. Furthermore, a new CNN with cross-level architecture is introduced in this paper to demonstrate the value of the proposed strategy in the future CNN design.

Keywords: Convolutional Neural Networks (CNNs) · Image classification · Network architecture · Feature representation · Deep learning

1 Introduction

As an important issue in the field of computer vision, image classification has achieved great progress in the past decade, which is primarily driven by the ever-increasing demand of image retrieval technique on the internet. Many worldwide competitions on image classification have been carried out, such as the Pattern Analysis, Statistical Modelling and Computational Learning, Visual Object Classes (PASCAL VOC) Challenge from 2005 to 2012 and ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) since 2010. It is notable that the performance of visual object recognition has obtained a dramatic improvement since convolutional neural networks (CNNs) [1,2] were first introduced into image classification by Krizhevsky et al. [3] in 2012. In the last three years, a variety of

H. Zha et al. (Eds.): CCCV 2015, Part I, CCIS 546, pp. 398–406, 2015.
DOI: 10.1007/978-3-662-48558-3_40

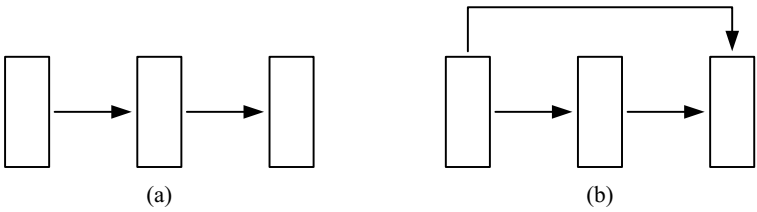

Fig. 1. Comparison of (a) Conventional structure of CNN and (b) the improved structure with cross-level strategy. Note that a block denotes a level of representation and an arrow denotes some operational layers between two levels.

CNN-based classification approaches have been presented [4–7], and the latest reported method [8] can even surpass the human-level performance.

Historically, convolutional network was first applied to visual object recognition by LeCun et al. [1], in which the problem of handwritten digit recognition was well tackled by a network containing two convolutional layers and two fully-connected layers. However, this method did not obtain enough attentions in generalized visual recognition for a long time, until the rise of deep learning theory [9,10] as well as the huge improvement on the computation capacity of hardware. Staring with the AlexNet [3], many representative CNN architectures such as Network in Network (NIN) [5] and GoogLeNet [7] have been proposed in the literature. As a typical category of deep neural networks, CNNs are designed for hierarchical data/feature representation mechanism from lower level to higher level, in which each level consists of a certain number of feature maps. The feature maps in a certain level are obtained from the maps in its previous level through several operations such as linear convolution, non-linear activation and spatial pooling. In this article, to make the following descriptions clearer, we use the term *layer* to specially denote a certain operation between two adjacent levels of feature maps, and the term *level* to indicate the data representation stage which is characterized by a set of feature maps. The existing CNNs share similar architectures, namely, alternate convolutional layers for feature extraction and spatial pooling layers like max-pooling for dimension reduction. Different levels of representation in a network are merely connected in series. In other words, each layer only locates between two adjacent levels, and there is no layer or direct connection between two nonadjacent levels. Fig. 1(a) shows the core structure of existing CNNs. However, the connection mechanism of visual neurons is generally believed to be very complex from the perspective of visual neuroscience [11,12].

In this paper, we mainly argue that the existing serial connection approach can be improved by adding direct connections between two nonadjacent levels. Specifically, a convolutional layer is established between two nonadjacent levels to realize this idea. This strategy is logically named cross-level, and it can be naturally combined into a CNN without any change on its original architecture. The illustration of cross-level strategy is shown in Fig. 1(b). The primary motivation

of this strategy is to learn more sufficient feature representations to pursue a better performance on image classification. The rest of this paper is organized as follows. In Section 2, three popular CNNs for image classification are reviewed. The implementation details of the cross-level strategy are presented in Section 3. The experimental results for validation are given in Section 4. Finally, Section 5 concludes the paper and puts forward some future work.

2 Related Work

In this section, we briefly review three representative deep convolutional neural networks presented for image classification in the last three years, which are the AlexNet [3], Network-in-Network (NIN) [5] and GoogLeNet [7].

The AlexNet [3] proposed in 2012 can be viewed as a milestone in the field of image classification. It is the first time that CNN was employed for generalized image classification. The classification method based on AlexNet is the winner of ILSVRC 2012 with a significant breakthrough with respect to the previous approaches. The AlexNet reported in [3] contains five convolutional layers and three fully-connected layers, and each of these layers is followed by a point-wise non-linear activation layer called Rectified linear units (ReLUs). In this work, the non-linear activation is also viewed as a layer for the consistency of layer definition in Section 1. There is a local response normalization (LRN) layer that follows the first as well as the second convolutional layer (Actually, it is after the ReLU layer. Since a convolution layer in a CNN is usually followed by a non-linear layer like ReLU, the non-linear layer will not be explicitly mentioned later). There are three max-pooling layers in AlexNet. The first two follow the two LRN layers, respectively. The last max-pooling layer follows the fifth convolutional layer. The core structure of AlexNet locates between the second and third max-pooling layers, which contains three convolutional layers each with 3×3 convolution kernel. Four levels of feature maps of spatial size 13×13 are connected by these three convolutional layers. The authors reported in [3] that the removal of any of these layers leads to a loss of about 2% in terms of top-1 performance. The core structure of AlexNet is shown in Fig. 2(a).

Lin et al. [5] proposed NIN to obtain a better representation of local patches by adding a multi-layer perceptron after a convolution layer. In their method, they use a three-layer perceptron, and it is essentially equivalent to add two 1×1 convolutional layers after a 3×3 or 5×5 convolutional layer. Thus, the core structure or unit of NIN has three convolutional layers in series, as shown in Fig. 2(b). The network applied in [5] has four such units and there is a max-pooling layer between every two units. Furthermore, after the last three-layer convolution unit, instead of employing traditional fully-connected layers, the authors generate one feature map for each class and use the global average pooling scheme to obtain the resulting vector, which can reduce the number of parameters to a great extent and prevent overfitting for neural networks.

GoogLeNet, a 22-layer deep convolutional network proposed by Szegedy et al. [7], is the winner of ILSVRC 2014 classification competition. Since

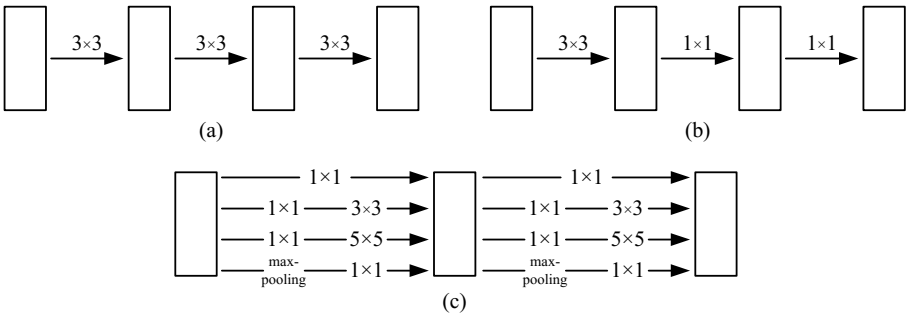

Fig. 2. Core structures of three CNNs. (a)AlexNet, (b)NIN, and (c)GoogLeNet.

increasing the depth of a network directly needs a sharp increase use of computational resources and tends to cause severe overfitting, the GoogLeNet is designed to make a balance between the network size and computational budget. The core structure adopted in GoogLeNet is called Inception. Fig. 2(c) shows two serial Inceptions. In each Inception, the feature maps in the output level are obtained from four branches, namely, a 1×1 convolution layer, a 3×3 convolution layer with a 1×1 layer for parameter reduction, a 5×5 convolution layer with a 1×1 layer for parameter reduction, and a max-pooling layer followed by a 1×1 layer to limit the number of output feature maps for parameter reduction in the next level. It is worthwhile to note that the intermediate feature maps generated in the last three branches do not construct a level of representation since those three 1×1 layers are essentially designed for parameter reduction. Therefore, there are only three levels of representation in Fig. 2(c). In GoogLeNet, there are totally nine Inceptions which are separated into three parts. The first part and last part both have two Inceptions just like the illustration given in Fig. 2(c). The middle part has five Inceptions in series. Moreover, there is no max-pooling layer within each of the three parts, so all the feature maps within each part have the same spatial size. In GoogLeNet, there exists a max-pooling layer between every two parts for dimension reduction of feature maps.

3 Cross-Level

In this section, we mainly describe the implementation details of the cross-level strategy via the above three convolutional networks, namely, the AlexNet [3], Network-in-Network (NIN) [5] and GoogLeNet [7]. Fig. 3 shows the improved structure of each network after applying the cross-level strategy. The basic idea of cross-level is to establish a convolutional layer between two nonadjacent levels. Naturally, the added convolution layer can be called cross layer. Thus, the feature maps in the output level come from two aspects: the layers in the original structure and the cross layer. In our approach, considering the cost of computational resource, the size of convolution kernel in each cross layer is fixed to 1×1,

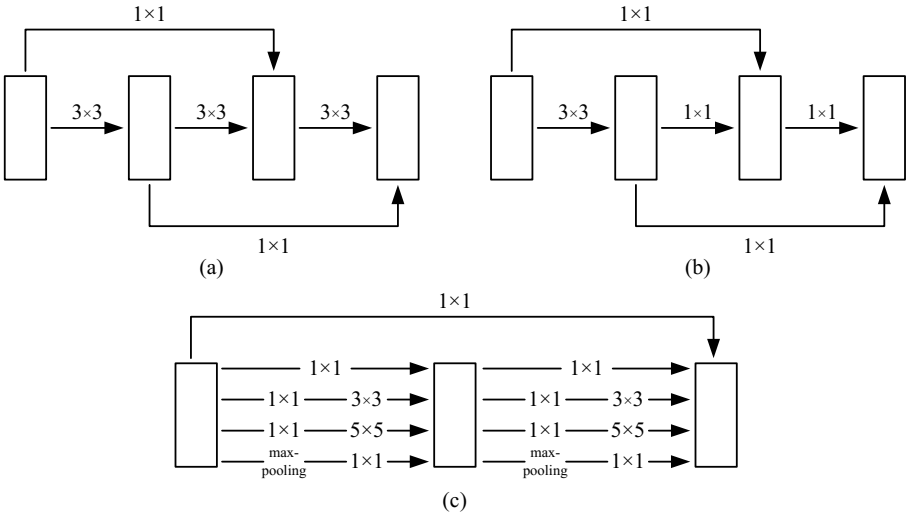

Fig. 3. The improved structure of three networks after applying the cross-level strategy.(a)AlexNet, (b)NIN, and (c)GoogLeNet.

and the number of feature maps generated by a cross layer is universally set as half number of the original maps in that level.

As shown in Fig. 3(a), for the AlexNet, two 1×1 convolutional layers are established from the first and second levels to the third and fourth levels, respectively. Notice that the core structure shown in Fig. 2(a) appears only once in the AlexNet, all the other parts of the network are not changed. The situation of NIN is similar to that of AlexNet, as shown in Fig. 3(b). The only difference is there are several core structures/units (see Fig. 2(b)) in the NIN architecture. For each unit except the first and last one, two 1×1 convolutional layers are added on the original structure. Thus, when there are four units [5], only four 1×1 layers are created on the second and third units, while the other parts in NIN remain unchanged. Finally, Fig. 3(c) shows the modified structure of GoogLeNet with cross-level strategy, which connects the input level of the former Inception and the output level of the latter one with a 1×1 convolutional layer. As mentioned before, the GoogLeNet also contains a structure of five consecutive Inceptions. The cross-level strategy deals with this situation just using the same approach in AlexNet (see Fig. 3(a)) and NIN (see Fig. 3(b)). Accordingly, there are totally six 1×1 convolutional layers added on the original GoogLeNet after applying the cross-level strategy.

From the above three examples, we can see that the cross-level strategy can be easily applied to an existing CNN without changing it original architecture, and the depth of the network also remains the same. The only requirement is that all the feature maps within the two cross-connected levels must have the same spatial size. That is to say, there must be no inside spatial pooling layers with stride larger than one.

It is worthwhile to notice that some existing CNN architectures have partly applied the cross-level strategy in some specific applications. Fan et al. [13] introduced a CNN with multiple paths for human tracking. In their method, the network between the first convolutional and the output layer is split into two branches, namely, global branch and local branch. The global branch is the same as traditional CNN architecture, which consists of several convolutional layers and pooling layers. The purpose of global branch is to enlarge the receptive field to address global structures. The local branch only has a convolutional layer, which aims to extract more details about local structures. Sermanet and LeCun [14] employed a similar multi-scale CNN architecture for traffic sign recognition. In [15], Sun et al. proposed a face verification method based on CNN, in which the last hidden layer is connected with both the third and fourth convolutional layers. The main purpose of this design is to avoid the loss of useful information, since the fourth layer contains too few neurons. The networks used in the publications referred above are generally known as multi-scale CNNs. Although these CNNs have bypassing connections, there exist clear difference between them and the CNNs applying the proposed cross-level strategy. In the above multi-scale CNNs, bypassing connections only connected with the output layer. Moreover, the main motivation using multi-scale CNNs is for specific object recognition such as human and face, in which features with different scales are all required in the output layer. However, the target of the proposed cross-level strategy is generalized object classification [3,5,7], and the basic motivation of this strategy is to extract more features with different scales at each feature representation level, not just the output one. Thus, the design of CNNs using the cross-level strategy is more flexible.

4 Experiments

The AlexNet [3], Network-in-Network (NIN) [5] and GoogLeNet [7] are first employed to verify the effectiveness of the proposed cross-level strategy for image classification. In this work, we use the dataset adopted by ILSVRC, which is a subset of ImageNet. It contains 1000 categories and each category has about 1300 images. Totally, there are about 1.28 million training images and 50000 validation images. The experimental setup is exactly similar to the approach reported in [3]. All the images are first down-sampled to a fixed spatial resolution of 256×256 and the mean intensity over the training set from each pixel is subtracted. All the models are learned using stochastic gradient decent algorithm. All the experiments are conducted on Caffe [16], which is a popular deep learning framework created by Jia et al. The implementation files of all the above three networks are available on Caffes website [17], and the parameters in our experiments are set as default values. The cross-level strategy is applied to these three networks by modifying the corresponding network definition files. For simplicity, the modified versions of these three networks are named AlexNet-Cross, NIN-Cross and GoogLeNet-Cross, respectively. For a fair comparison, all the parameters with respect to model training remain the same with the original networks.

The top-1 and top-5 accuracy rates are tested for each learned CNN model using the validation image sets. For each test image, only the central patch of appropriate size is extracted for prediction, i.e., single-view prediction is applied. It is worthwhile to note that we do not apply some widely used strategies such as multi-view prediction and model fusion [3] to pursue a high accuracy rate, which are always required in ILSVRC competition. The main purpose here is to make a pure comparison between a network and its improved version with cross-level strategy. Thus, we just test the accuracy rate based on single model as well as single view in this paper. Table 1 lists the top-1 and top-5 accuracy rates of six learned CNN models. For all of these three networks, it can be seen from Table 1 that the cross-level strategy results in a rise of about 1% in terms of both top-1 and top-5 accuracy rates. In particular, the performance improvement of GoogLeNet is the most significant. From our perspective, this is mainly because the proportion of levels which are influenced by the cross-level strategy in GoogLeNet is the highest among these three networks.

Table 1. The top-1 and top-5 accuracy rates of six learned CNN models.

Model	Top-1	Top-5
AlexNet	56.48%	79.56%
AlexNet-Cross	57.37%	80.52%
NIN	59.42%	81.60%
NIN-Cross	60.56%	82.61%
GoogLeNet	68.93%	88.90%
GoogLeNet-Cross	70.28%	90.08%

In addition to the existing networks, the cross-level strategy can be also used for the design of new networks. To verify this point, as well as to further demonstrate the effectiveness of the cross-level strategy from the other point of view, we design a new CNN architecture by referring to GoogLeNet. Specifically, we just remove two branches in the Inception of GoogLeNet, while all the other structures remain the same, mainly including the depth of network and the number of feature maps each branch generates. We apply the cross-level strategy to this new network just as the way to GoogLeNet. The core structure of the designed network is shown in Fig. 4, in which only the 1×1 and 3×3 branches are preserved. The same training and testing approaches are used to this network. The top-1 and top-5 accuracy rates obtained are 68.74% and 88.78%, respectively. We can see from Table 1 that the performance of this new network is very close to that of GoogleNet, but the number of feature maps as well as training parameters is significantly decreased.

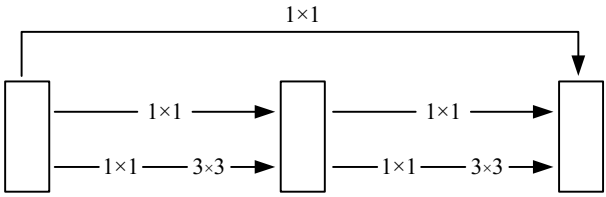

Fig. 4. The core structure of the new designed network.

5 Conclusion

Contribution- This paper presents a novel strategy called cross-level for CNN-based image classification. The basic idea is to establish a convolutional layer between two nonadjacent levels in the network, which aims to learn more sufficient feature representations for a better classification performance. Experimental results on three popular convolutional networks demonstrate the effectiveness of the proposed cross-level strategy. We also exhibit the potential of the cross-level strategy used for the design of new networks.

Limitation- There still exist some limitations in this work. First, the number of feature maps generated by a cross layer is normally set as half number of the original maps in that level in our method. The impact of this proportional factor on the classification performance is not fully studied, which is mainly due to the reason that CNN model training is very time-consuming. Second, only one single model for each network is learned and only the central patch in each test image is extracted for prediction. Thus, this work has not been completed and we have not obtained an ultimate result on classification accuracy.

Future Work- Considering the above limitations, we will conduct more experiments to further study the impact of the above proportional factor. Furthermore, we will design some new networks using the cross-level strategy and attempt to obtain a competitive result via the approaches like model fusion as well as multi-view prediction.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 61472393 and No. 61303150), the National Science and Technology Major Project of the Ministry of Science and Technology of China (No. 2012GB102007), and the Anhui Province Initiative Funds on Intelligent Speech Technology and Industrialization (No. 13Z02008). The authors greatly acknowledge the support of IFLYTEK CO.,LTD.

References

1. LeCun, Y., Boser, B., Denker, J.S., et al.: Backpropagation applied to handwritten zip code recognition. Neural Computation **1**, 541–551 (1989)
2. LeCun, Y., Kavukcuoglu K., Farabet C., et al.: Convolutional networks and applications in vision. In: IEEE International Symposium on Circuits and Systems, pp. 254–256 (2010)
3. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convoluntional neural networks. Advances in Neural Information Processing Systems **25**, 1106–1114 (2012)
4. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part I. LNCS, vol. 8689, pp. 818–833. Springer, Heidelberg (2014)
5. Lin, M., Chen Q., Yan, S.: Network in network (2013). arXiv: 1312.4400 [cs.NE]
6. He, K., Zhang, X., Ren, S., et al.: Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition (2014). arXiv: 1406.4729 [cs.CV]
7. Szegedy, C., Liu, W., Jia Y., et al.: Going deeper with convolutions (2014). arXiv: 1409.4842 [cs.CV]
8. He, K., Zhang, X., Ren, S., et al.: Delving deep into rectifiers: Surpassing human-Level performance on imageNet classification (2015). arXiv: 1502.01852 [cs.CV]
9. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**, 504–507 (2006)
10. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**, 1798–1828 (2013)
11. Desimone, R., Duncan, J.: Neural mechanisms of selective visual attention. Annual review of neuroscience **18**, 193–222 (1995)
12. Spirkovska, L., Reid, M.B.: Robust position, scale, and rotation invariant object recognition using higher-order neural networks. Pattern Recognition **25**, 975–985 (1992)
13. Fan, J., Xu, W., Wu, Y., et al.: Human tracking using convolutional neural networks. IEEE Transactions on Neural Networks **21**, 1610–1623 (2010)
14. Sermanet, P., LeCun, Y.: Traffic sign recognition with multi-scale convolutional networks. In: International Joint Conference on Neural Networks, pp. 2809–2813 (2011)
15. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1891–1898 (2014)
16. Jia, Y., Shelhamer, E., Donahue, J., et al.: Caffe: Convolutional architecture for fast feature embedding. In: ACM International Conference on Multimedia, pp. 675–678 (2014)
17. Caffe website. http://caffe.berkeleyvision.org/