Real-Time Object Detection Method Based on Improved Attention Transfer (Postprint)

Authors: Zhang Chi, Liu Hongzhe

Date: 2020-09-28T00:00:00+00:00

Abstract

Deep neural network models currently need to be deployed in resource-constrained environments, thus necessitating the design of efficient and compact network architectures. This paper proposes a model compression method based on improved attention transfer (KE) for designing compact neural networks, which primarily utilizes a wide residual teacher network (WRN) to guide a compact student network (KENet), transferring spatial and channel attention to the student network to enhance performance, and applies this approach to real-time object detection. Image classification experiments on CIFAR verify that the knowledge distillation method with improved attention transfer can improve compact model performance, while object detection experiments on VOC validate that the KEDet model achieves excellent accuracy (72.7 mAP) and speed (86 FPS). The experimental results fully demonstrate that the object detection model based on improved attention transfer exhibits high accuracy and real-time performance.

Full Text

Preamble

Volume 38, Issue 3
Application Research of Computers
ChinaXiv Partner Journal

Real-Time Object Detection Method Based on Improved Attention Transfer

Zhang Chi a,b, Liu Hongzhe a,b†
(a. Beijing Key Laboratory of Information Service Engineering; b. College of Robotics, Beijing Union University, Beijing 100101, China)

Abstract: Deep neural network models currently need to be deployed in resource-constrained environments, necessitating the design of efficient and compact network architectures. This paper proposes a model compression method (KE) based on improved attention transfer for designing compact neural networks. The method primarily uses a wide residual teacher network (WRN) to guide a compact student network (KENet), transferring both spatial and channel-wise attention to the student network to enhance performance, and applies this approach to real-time object detection. Image classification experiments on CIFAR verify that the knowledge distillation method with improved attention transfer can improve the performance of compact models. Object detection experiments on VOC confirm that the KEDet model achieves good accuracy (72.7 mAP) and speed (86 FPS). The experimental results fully demonstrate that the object detection model based on improved attention transfer exhibits excellent accuracy and real-time performance.

Keywords: neural network; deep learning; object detection; knowledge distillation; attention transfer

0 Introduction

Object detection constitutes a critical component of autonomous and assisted driving systems, encompassing tasks such as vehicle detection, pedestrian detection, traffic sign detection, and road marking detection. Convolutional Neural Networks (CNNs) have achieved remarkable success in object detection tasks, largely dependent on substantial computational power and memory resources [**?**]. However, domains like autonomous driving often operate under resource constraints, making neural network deployment challenging. Therefore, effectively reducing the computational and storage costs of neural networks while maintaining performance represents a key urgent problem.

CNN-based object detection primarily employs region proposal-based methods represented by the R-CNN [**?**] series, also known as two-stage approaches. These methods first generate candidate regions (regions of interest) that may contain objects, then predict object categories and locations within these regions. Such approaches achieve high detection accuracy and perform well on large-scale datasets. However, they typically involve substantial computational costs and slow operation speeds. To reduce computational load while preserving accuracy and improving operational efficiency and real-time performance, single-stage object detection frameworks such as YOLO [**?**] and SSD [**?**] have emerged. These methods divide images into grids of equal size and predict object categories and locations based on these grids. Subsequent novel methods have further improved detection speed and accuracy by fusing multi-scale features, incorporating contextual information, and simplifying network structures [**?**].

Model compression represents a general approach for neural network deployment under resource constraints [**?**], enabling compression of object detection models

to enhance real-time detection capabilities. Methods for network parameter compression primarily include pruning [**?**, **?**], quantization [**?**, **?**], and low-rank decomposition [**?**, **?**]. Additionally, more efficient convolutions [**?**, **?**, **?**] can be employed to design more compact structures, or knowledge transfer [**?**, **?**, **?**] (also known as knowledge distillation) can extract knowledge from a large "teacher" model to assist in training a small "student" model, thereby improving the "student" model' s performance. Attention transfer is an improved knowledge distillation method that introduces attention mechanisms into the knowledge distillation model, making the attention activation distributions of the student and teacher networks as close as possible.

Attention transfer is primarily used to improve convolutional neural networks. This paper enhances lightweight convolutional models based on attention transfer, extracting knowledge from both spatial and channel dimensions to compensate for the deficiencies of lightweight models, and proposes a knowledge enhancement distillation method. Building upon this knowledge enhancement approach, the paper further proposes a real-time object detection model based on improved SSD. Through experiments on multiple datasets, the paper verifies that the object detection model based on improved attention transfer achieves excellent accuracy and real-time performance.

1.1 Lightweight Convolutional Structures

High-performance object detection models often employ lightweight convolutional structures as backbone networks. Let us analyze these structures. In Figure 1: see original paper, N represents the number of input channels, K×K is the size of each convolution kernel, and M is the number of output channels, with total computational cost NK^2M. Traditional convolution has spatial dimension K^2 (kernel size) and channel dimension N×M.

The first method for parameter reduction divides each convolution into G groups, as shown in Figure 1: see original paper. Compared to standard convolution cost NK^2M, this operation reduces computation by 1/G. A $1\$\times 1 pointwise convolution follows the grouped convolution to provide cross - channel in formation, with computational cost N \times M. Total computation becomes NK\hat{}\{2\}\$M/G + NM. This approach has been used in AlexNet [**?**], Xception [**?**], and ShuffleNet [**?**, **?**].

Another method uses narrow and deep residual networks (such as ResNet, MobileNet) to replace wide and shallow networks (such as VGG16), essentially introducing a bottleneck structure. As shown in Figure 1: see original paper, the first $1\$\times 1 pointwise convolution reduces input channel dimension N by a factor of B, followed by K \times K convolution + NK/B + MK/B.

Based on these two methods, this paper proposes a new lightweight convolutional structure: KENet. First, $1\$\times 1 grouped convolution is used for dimensionality reduction, followed by tradi

$NK^{\{2\}M/GB}\{2\}\$ + \text{MK/B}$.

Grouped convolution and bottleneck structures are common methods for parameter reduction, but they also cause information loss. As shown in Figure 2: see original paper, from the spatial dimension perspective, narrow and deep residual networks use smaller convolution kernels, which reduce the receptive field and consequently lose some spatial context information. As shown in Figure 2: see original paper, from the channel dimension perspective, using grouped convolution isolates channels, preventing information flow between different groups. Typical lightweight networks like MobileNet use depthwise separable convolution, which divides each channel into a separate group. Excessive grouping significantly reduces operational speed. To address these issues, a powerful teacher model can be used during training to provide additional supervision signals through knowledge distillation.

1.2 Improved Attention Transfer Algorithm

Attention transfer is an improved knowledge distillation method. It adopts transfer learning principles, treating the activation distribution of teacher network intermediate layers as the source domain and the corresponding distribution of student network as the target domain, achieving attention transfer by minimizing the distance between source and target domains.

Attention transfer measures the distance between student and teacher through attention points at intermediate layers using activation maps. The loss function is defined as:

$$\mathcal{L}_{AT} = \frac{1}{2}\sum_{i=1}^{N}\left\|\frac{F(A_s^i)}{\|F(A_s^i)\|_p} - \frac{F(A_t^i)}{\|F(A_t^i)\|_p}\right\|_2^2 + \beta\mathcal{L}_{CE}$$

where s and t denote student and teacher respectively, σ is the softmax function, and $\mathcal{L}_{CE}$ represents standard cross-entropy loss. i indexes the selected activation layers from teacher and student. A_t^i and A_s^i represent teacher and student activations respectively. F denotes the attention mapping function that maps three-dimensional attention activation tensors to two-dimensional attention activation maps. We use $F(A) = \sum_{j=1}^{C}|a_{ij}|^p$ as the mapping function, where a_{ij} represents the activation feature vector of channel j in layer i. β is a hyperparameter, p is the norm type, set to $p = 2$ here.

To address information loss in lightweight structures, this paper improves upon the attention transfer model by extracting knowledge from both spatial and channel dimensions, redefining attention activation maps, and proposing a model called Knowledge Enhancement. For activation feature map $A \in \mathbb{R}^{H\times W\times C}$, we first use inter-channel relationships to generate channel-wise knowledge. To aggregate spatial information, we employ average pooling to

generate spatial context descriptors, which are then fed into a fully connected network to produce channel-based knowledge $K_c(A)$. Simultaneously, we utilize spatial relationships of feature maps to generate spatial knowledge. We first apply average pooling along the channel axis to generate effective feature descriptors, then input them into a convolutional layer to produce spatial knowledge $K_s(A)$. Based on these two types of knowledge, we enhance the knowledge of attention activation tensors and generate enhanced activation features $E(A)$. The complete computation process is:

$$E(A) = K_c(A) \otimes K_s(A) \otimes A$$

where $\otimes$ denotes element-wise multiplication. Finally, knowledge-enhanced activation maps are generated for both teacher and student: $F_E^i = F(E(A^i))$.

The improved attention transfer model is shown in [Figure 3: see original paper]. Since this model extracts spatial and channel knowledge for attention transfer, compensating for information lost in lightweight convolutional models, it is named the Knowledge Enhancement model. Deploying the Knowledge Enhancement model in lightweight models like KENet increases parameters by only approximately 3%, with additional parameters concentrated in the FC layer that generates spatial knowledge, which is negligible for the overall model parameters.

[Figure 4: see original paper] visualizes the activation maps at the last convolutional layer of ResNet50, comparing effects before and after knowledge enhancement. Red regions indicate strong activations that contribute significantly to final predictions. Notably, after applying the proposed knowledge enhancement module, the network tends to focus more on useful regions. In other words, by enhancing spatial and channel knowledge, the proposed method makes network attention more concentrated and improves network performance.

---

2 Real-Time Object Detection Model Based on Improved SSD

SSD (Single Shot Multibox Detector) is a lightweight single-stage object detection algorithm. Given an input image, SSD directly generates classification and localization results, enabling end-to-end object detection. The SSD loss function has the following form:

$$\mathcal{L}_{SSD} = \frac{1}{N}\left(\mathcal{L}_{cls} + \alpha\mathcal{L}_{loc}\right)$$

where $\mathcal{L}_{cls}$ is classification loss, $\mathcal{L}_{loc}$ is localization loss, N is the number of positive samples, and α is a balancing factor. This paper improves SSD based

on knowledge distillation principles, with the overall framework structure shown in [Figure 5: see original paper].

The classification loss is expressed as $\mathcal{L}_{cls} = \mathcal{L}_{hard} + \lambda\mathcal{L}_{soft}$, where $\mathcal{L}_{hard}$ is the classification loss predicted using the student' s ground-truth labels, and $\mathcal{L}_{soft}$ is the distillation loss composed of soft labels from teacher and student. Localization loss can be expressed as $\mathcal{L}_{loc} = \mathcal{L}_{smooth} + \mathcal{L}^b_{loc}$, where $\mathcal{L}_{smooth}$ is the smooth L1 loss between ground-truth and student location predictions, while $\mathcal{L}^b_{loc}$ represents a penalty term that activates when the gap between student regression error and teacher regression error exceeds a threshold m:

$$\mathcal{L}^b_{loc} = \begin{cases} (R^s_y - R^t_y)^2 - m & \text{if } |R^s_y - R^t_y| > m \\ 0 & \text{otherwise} \end{cases}$$

The wide residual network has two main parameters: depth d and width k, where depth d relates to the number of convolution modules n as $d = 6(n + 4)$, and width k determines the channel size of filters in these modules. The convolutional part of the wide residual network consists of an initial convolution layer and three main convolution blocks.

Experiments use WRN-40-2 (wide residual network with depth 40 and width multiplier 2) with standard residual modules. Each standard module comprises two 3$×$3 convolution kernels. This paper conducts comparative experiments using the following modules:

a) Grouped convolution modules, named Gconv(G), where G is the number of groups ranging in {2,4,8,16}.

b) Bottleneck modules with 2× channel reduction, called Bottleneck(B), where B=2 is the channel reduction factor.

c) The typical lightweight CNN MobileNet, using Depthwise Separable Convolution (DSC).

d) The proposed Knowledge Enhancement Network KENet(G,B) combining bottleneck structure with grouped convolution, using B=2 and group numbers G in {2,4,8,16}.

This paper compares the Knowledge Enhancement (KE) method with models trained without knowledge distillation. Training uses 4 Titan V GPUs with minibatch size 128, stochastic gradient descent for 200 epochs, momentum 0.9, initial learning rate 0.1, reduced by 0.2× every 60 iterations. Hyperparameter β is set to 1000.

presents the performance of the above architectures for image classification on CIFAR10. First comparing computational costs of different convolution modules, we observe that effective parameter compression is achieved through

grouped convolution and bottleneck structures, with KENet combining both to compress parameters by 10-20×. Experimental results from knowledge distillation based on different structures show that student models trained with the proposed knowledge enhancement method significantly outperform direct training results.

When KENet serves as the student and uses knowledge enhancement as the distillation model, more effective model compression is achieved with minimal accuracy loss. Although KENet' s parameters are far fewer than the teacher and other student models, inevitably causing accuracy degradation, the knowledge enhancement model provides spatial and channel information, significantly improving the accuracy of KENet trained through knowledge enhancement.

Furthermore, introducing the improved attention transfer (knowledge enhancement) algorithm described above adds a knowledge enhancement loss:

$$\mathcal{L}_{KE} = \frac{1}{N}\sum_{i=1}^{N}\left\|F_E^i(A_s) - F_E^i(A_t)\right\|_p$$

The final loss function for the improved real-time object detection model KEDet is:

$$\mathcal{L}_{KEDet} = \mathcal{L}_{cls} + \alpha\mathcal{L}_{loc} + \beta\mathcal{L}_{KE}$$

where F_E^i are knowledge-enhanced activation maps.

3.1 Effectiveness of KENet and Knowledge Enhancement Algorithm

This paper evaluates the performance of knowledge enhancement through image classification experiments, training three types of student networks (Gconv, Bottleneck, and KENet) on the Canadian Institute For Advanced Research (CIFAR) dataset for image classification, using the official metric (top-1 error rate) as evaluation criterion. Wide Residual Networks (WRN) serve as the basic experimental structure.

The wide residual network has two main parameters: depth d and width k, where depth d relates to convolution module count n as $d = 6(n+4)$, and width k determines filter channel sizes in these modules. The convolutional portion comprises an initial convolution layer and three main convolution blocks.

Experiments use WRN-40-2 (wide residual network depth 40, width multiplier 2) with standard residual modules. Each standard module consists of two 3$×$3 convolution kernels. This paper uses the following modules for comparative experiments:

a) Grouped convolution modules, named Gconv(G), where G is the number of groups in {2,4,8,16}.

b) Bottleneck modules with 2× channel reduction, called Bottleneck(B), where B=2 is the channel reduction factor.

c) Typical lightweight CNN MobileNet, using Depthwise Separable Convolution (DSC).

d) The proposed Knowledge Enhancement Network KENet(G,B) combining bottleneck structure with grouped convolution, using B=2 and group numbers G in {2,4,8,16}.

This paper compares the Knowledge Enhancement (KE) method with models trained without knowledge distillation, using 4 Titan V GPUs, minibatch size 128, stochastic gradient descent for 200 epochs, momentum 0.9, initial learning rate 0.1, reduced by 0.2× every 60 iterations. Hyperparameter β is set to 1000.

shows classification error for various architectures on CIFAR10. Comparing computational costs first reveals that grouped convolution and bottleneck structures achieve effective parameter compression, with KENet combining both to compress parameters by 10-20×. Knowledge distillation results based on different structures demonstrate that student models trained with the proposed knowledge enhancement method significantly outperform direct training.

When KENet serves as the student using knowledge enhancement for distillation, more effective model compression is achieved with minimal accuracy loss. Although KENet' s parameters are far fewer than the teacher and other student models, inevitably causing accuracy drops, the knowledge enhancement model provides spatial and channel information, yielding significant accuracy improvements for KENet trained via knowledge enhancement.

3.2 Evaluation of Real-Time Object Detection Model

To verify the effectiveness of the proposed real-time object detection model KEDet, we compare the improved SSD detector with the original SSD model and also include the two-stage detector Faster-RCNN for comparison, evaluating detection performance using different backbone networks. The Pattern Analysis, Statistical Modelling and Computational Learning Visual Object Classes (PASCAL VOC) dataset serves as the benchmark for model performance. Mean Average Precision (mAP) evaluates detection accuracy, while Frames Per Second (FPS) measures real-time capability.

Experiments use VOC2007 and VOC2012 as training datasets with input image size 300, training for 250 epochs on a server with 4 TITAN V GPUs, and evaluating performance on the VOC2007 test set. Testing equipment is a mobile

terminal (laptop) with a single GTX 1080 GPU.

presents test results for Faster-RCNN, SSD, and KEDet models. Faster-RCNN, as a typical two-stage detector, achieves high detection accuracy with both VGG16 and ResNet101 backbones but suffers slow detection speed, failing to achieve real-time performance (above 30 FPS). SSD (based on VGG16), as a representative single-stage detector, achieves real-time speed (45 FPS) while maintaining detection accuracy. However, SSD based on VGG16 does not deploy well in practice, motivating the SSD object detection model based on MobileNet, which sacrifices some detection accuracy (68.1 mAP) to improve speed (83 FPS) and performs well in actual deployment. Nevertheless, as described in this paper, MobileNet uses Depthwise Separable Convolution, where excessive grouping leads to inter-group information isolation and reduced speed. Therefore, this paper proposes KENet to replace MobileNet as a new backbone network, reducing group count and improving performance.

Meanwhile, this paper improves the detection model through the knowledge enhancement method, proposing the KEDet detection model. Through improved attention transfer for knowledge distillation, the model compensates for information lost in lightweight models. Compared to VGG16, KENet uses bottleneck structures to deepen the network and significantly compress parameters. Compared to MobileNet, KENet replaces depthwise separable convolution with minimal grouping, improving detection accuracy and speed. Experimental results demonstrate that KEDet improves detection accuracy (72.7 mAP) while maintaining detection speed (86 FPS), achieving excellent accuracy and real-time performance.

---

4 Conclusion

This paper proposes a knowledge distillation method based on attention transfer and applies it to improve the SSD model, presenting a real-time object detection model called KEDet. By analyzing lightweight convolutional structure characteristics, we first propose KENet, a lightweight convolutional model combining grouped convolution and bottleneck structures. We then propose a knowledge distillation method based on attention transfer to enhance KENet with knowledge enhancement, validating its effectiveness through image classification experiments on CIFAR datasets. Building upon this, we propose KEDet, an improved real-time object detection model based on SSD, and verify it on the VOC dataset. Experimental results show that the proposed KEDet model achieves high detection accuracy and speed, simultaneously possessing both accuracy and real-time capability. Future research could combine pruning algorithms and neural architecture search to further explore more efficient network structures and compression algorithms to improve detection efficiency.

References

[1] Zhang Junyang, Wang Huili, Guo Yang, et al. Review of deep learning [J]. Application Research of Computers, 2018, 35(7): 1921-1928.

[2] Ren Shaoqing, He Kaiming, Girshick R, et al. Faster r-cnn: Towards Real-time Object Detection With Region Proposal Networks [C]// Advances in Neural Information Processing Systems. 2015: 91-99.

[3] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2016: 779-788.

[4] Liu Wei, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector [C]// Proc of European Conference on Computer Vision. Cham: Springer, 2016: 21-37.

[5] Zhang Linna, Chen Jianqiang, Chen Xiaoling, et al. Lightweight SSD network for real-time object detection in automotive videos [J]. Computer Science, 2019, 46(7): 233-237.

[6] Cao Wenlong, Rui Jianwu, Li Min. Survey of neural network model compression methods [J]. Application Research of Computers, 2019(3): 649-656.

[7] Yoon J, Hwang S J. Combined group and exclusive sparsity for deep neural networks [C]// Proc of the 34th International Conference on Machine Learning. 2017: 3958-3966.

[8] Liu Zhuang, Li Jianguo, Shen Zhiqiang, et al. Learning efficient convolutional networks through network slimming [C]// Proc of IEEE International Conference on Computer Vision. 2017: 2736-2744.

[9] Courbariaux M, Bengio Y, David J P. Binaryconnect: Training deep neural networks with binary weights during propagations [C]// Advances in Neural Information Processing Systems. 2015: 3123-3131.

[10] Hubara I, Courbariaux M, Soudry D, et al. Binarized neural networks [C]// Advances in Neural Information Processing Systems. 2016: 4107-4115.

[11] Wang Weiqi, Sun Yifan, Eriksson B, et al. Wide compression: Tensor ring nets [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2018: 9329-9338.

[12] Zhang Xiangyu, Zou Jianhua, He Kaiming, et al. Accelerating very deep convolutional networks for classification and detection [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2015, 38(10): 1943-1955.

[13] Sandler M, Howard A, Zhu Menglong, et al. Mobilenetv2: Inverted residuals and linear bottlenecks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4510-4520.

[14] Zhang Xiangyu, Zhou Xinyu, Lin Mengxiao, et al. Shufflenet: An extremely

efficient convolutional neural network for mobile devices [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6848-6856.

[15] Ma Ningning, Zhang Xiangyu, Zheng Haitao, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design [C]// Proc of European Conference on Computer Vision. 2018: 116-131.

[16] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network [EB/OL]. (2015-03-05). https://arxiv.org/abs/1503.02531.

[17] Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer [EB/OL]. (2016-12-12). https://arxiv.org/abs/1612.03928.

[18] Yim J, Joo D, Bae J, et al. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4133-4141.

[19] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [C]// Advances in Neural Information Processing Systems. 2012: 1097-1105.

[20] Chollet F. Xception: Deep learning with depthwise separable convolutions [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1251-1258.

Note: Figure translations are in progress. See original paper for figures.