融入选择性卷积核的胶囊网络图像分类方法

陈泽轩，于莲芝

（上海理工大学 光电信息与计算机工程学院，上海 200093）

摘　要：传统卷积神经网络对空间信息不敏感，无法学习到不同特征间相对位置的关系，且每一层神经元的感受野被设计为相同大小，导致提取的图像特征信息不够精确。针对这些问题，提出一种选择性卷积核胶囊网络用于图像分类任务。在经典胶囊网络的卷积层融入具有两个分支的选择性卷积核网络，以提取更为丰富、准确的图像特征信息，提高图像分类准确率。采用CIFAR-10、Fashion-MNIST、SVHN经典图像分类数据集进行实验，结果表明，相比于基线胶囊网络模型，新模型的识别精度更高，尤其在CIFAR-10数据集上识别精度提高了1.73%，从而有效提升了图像分类准确率，具有良好的图像识别能力。

关键词：胶囊网络；动态路由；特征提取；选择性卷积核；动态选择机制；图像分类

DOI：10.11907/rjdk.211580　　开放科学（资源服务）标识码（OSID）：

中图分类号：TP317.4　　**文献标识码**：A　　**文章编号**：1672-7800（2022）001-0248-05

Image Classification Method Using Capsule Network Integrated into Selective Kernel Networks

CHEN Ze-xuan, YU Lian-zhi

(*School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai* 200093, *China*)

Abstract: The traditional convolution neural network is insensitive to spatial information and cannot learn the relative position relationship between different features, and the sensory field of each layer of neurons is designed to be the same size, which leads to the inaccuracy of the extracted image feature information. To address these problems, a selective convolutional kernel capsule network is proposed for image classification tasks. The convolution layer of the classical capsule network is integrated into the selective convolution kernel network with two branches, which can extract more abundant and accurate data image feature information and improve the accuracy of image classification. The experimental use CIFAR-10, Fashion-MNIST, SVHN these classical image classification data sets. The results show that the recognition accuracy of the new model is higher than that of the baseline capsule network model, especially the recognition accuracy on the CIFAR-10 data set is improved by 1.73%. The new model effectively improves the accuracy of image classification and has good image recognition ability.

Key Words: capsule network; dynamic routing; feature extraction; selective kernel networks; dynamic selection mechanism; image classification

0　引言

图像分类是计算机视觉领域的一个重要研究方向[1]，旨在从图像或图像序列中提取出目标的判别性特征，以精确判别目标所属类别。随着图像分类技术的快速发展，在许多应用领域取得了较为显著的成绩，并逐步推动人类进入智能时代。目前，图像分类技术在医疗图像处理[2]、智能交通[3]、人脸识别[4]等领域已得到大规模应用。早期使用机器学习算法进行图像分类的流程为：首先提取特征，然后筛选特征，最后将特征向量输入合适的分类器完成分类任务。但传统图像分类模型结构较浅，不能提取较复杂的图像特征，泛化性差，导致图像识别精度不高，只能处理一些简单任务。

卷积神经网络（Convolutional Neural Network，CNN）最早形态是Lecun等[5]提出的LeNet-5，该网络处理数字识别

收稿日期：2021-04-18

作者简介：陈泽轩（1997-），男，上海理工大学光电信息与计算机工程学院硕士研究生，研究方向为图像处理；于莲芝（1966-），女，博士，上海理工大学光电信息与计算机工程学院副教授、硕士生导师，研究方向为图像处理、大数据。本文通讯作者：于莲芝。

任务具有良好效果，但处理其它实际任务时效果不如Boosting等传统算法。2012年，Krizhevsky等[6]提出AlexNet网络结构，并以15.4%创纪录的低失误率夺得2012年ILSVRC年度冠军，表明使用深度学习方法处理图像分类任务出现了较大突破。随后图像分类领域涌现出许多性能优越的深度学习模型，如ZFNet[7]、VGGNet[8]、GoogleNet[9]、ResNet[10]、DenseNet[11]、SENet[12]等。尽管卷积神经网络经过长时间发展，产生了大量优秀模型，但因其本身结构存在一些缺陷，需要研究一种新模型来处理图像分类任务。

胶囊网络（Capsule Network，CapsNet）是在2017年由Sabour等[13]首次提出的，胶囊网络首次提出使用矢量神经元代替传统神经网络中的标量神经元，去掉了池化层，创新性地提出动态路由（Dynamic Routing，DR）算法计算初始胶囊层与高级胶囊层之间的连接权重，并采用新的挤压函数取代Relu函数，从而学习到图像特征之间的空间关系。之后出现了大量关于胶囊网络的研究，如MS-CapsNet[14]、Dcaps[15]、VideoCapsuleNet[16]等。

静态卷积每层使用同一个卷积核，无法提取更丰富的特征信息。2020年，Chen等[17]提出动态卷积网络，动态地聚合多个并行卷积核提取的特征，增强了网络表达能力；华为提出DyNet[18]，核心思想与谷歌提出的CondConv[19]类似；旷视提出DRConv[20]，在动态卷积上引入空间分组；Tian等[21]提出用于实例分割的条件卷积网络。2021年，Li等[22]通过矩阵分解回顾动态卷积。

本文提出一种基于胶囊网络并融入选择性卷积核网络的图像分类模型。胶囊网络对空间信息比较敏感，可学习到不同特征间的位置关系，克服了卷积神经网络识别图像时整体平移的缺点。传统卷积每一层使用单一卷积核，选择性卷积核区别于传统卷积的静态结构，其使用两个分支结构，动态聚合两个并行卷积核提取的特征，比静态卷积具有更强的表示能力。在经典胶囊网络的Conv1基础上增加具有两个分支的选择性卷积核网络，可融合不同卷积核提取到的信息，丰富的图像分类特征有效提高了图像识别准确率。

1　胶囊网络

2011年，Hinton等[23]首次提出胶囊的概念，胶囊由多个向量化神经元构成，可有效保存图像中的特征角度及姿态等更为丰富的信息。胶囊网络区别于传统卷积神经网络，胶囊网络的输入与输出神经元都由标量提升为矢量，池化操作由协议路由取代，并提出新的Squash函数代替原来的ReLU函数。

1.1　总体结构

胶囊网络总体架构如图1所示。网络共有3层，两个卷积层和一个全连接层，是一个比较浅的神经网络。由于常规的卷积操作可得到精确的低级图像特征，第一层采用传统卷积操作，使用256个9×9的卷积核对图像像素作一次局部特征检测。为最大限度地保留图像特征信息，卷积之后没有使用池化层。第二层是含有胶囊的卷积层，将Conv1得到的低级特征送入可进行路由运算的Primary Capsule层，经过再次卷积提取特征后，将特征向量展开成一维，对应特征向量组合得到向量胶囊。数字胶囊层是全连接层，根据向量胶囊的模长大小判断并输出图片类别。

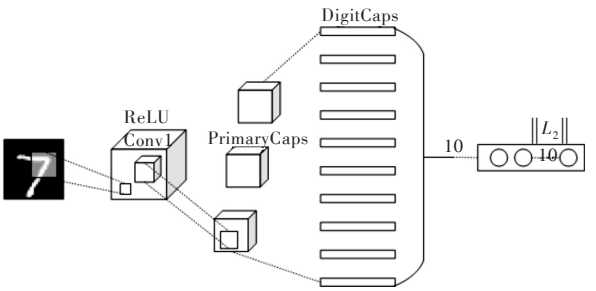

Fig. 1　Overall structure of capsule network

图1　胶囊网络总体结构

1.2　标量神经元提升为向量神经元

卷积神经网络将神经元输入的标量乘以各自权重，加权求和后输入非线性函数进行激活，得到输出标量。胶囊网络在加权求和时比卷积神经网络多一个步骤，需要将输入向量先乘以一个姿态矩阵，生成新的输入向量后再乘以权重矩阵进行加权求和。胶囊网络工作原理如图2所示。

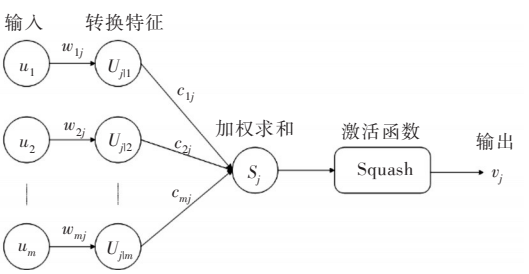

Fig. 2　Working principle of capsule network

图2　胶囊网络工作原理

1.3　动态路由算法

动态路由是PrimaryCap与DigitCaps两者间的一种信息传递机制。其两者间的连接并不是传统神经网络之间简单的标量连接，而是向量与向量之间的连接。低层胶囊与高层胶囊之间的连接权重会在网络训练过程中不断随着学习而改变，并使用动态路由算法学习该权重。动态路由算法如下：

算法1　动态路由算法

Input：Prediction vectors $\hat{\boldsymbol{u}}_{j|i}$，layer $\boldsymbol{l}$，max iterations $\boldsymbol{r}$

Process：

1）　Initialization：$\boldsymbol{b}_{ij} \leftarrow 0$

2）　**For** $\boldsymbol{r}$ iteration **do**

3）　$\boldsymbol{c}_{ij} \leftarrow softmax(\boldsymbol{b}_{ij})$

4）　$\mathbf{s}_j \leftarrow \sum_i \boldsymbol{c}_{ij} \hat{\boldsymbol{u}}_{j|i}$

5）　$\boldsymbol{v}_j \leftarrow squash(\mathbf{s}_j)$

6）　$\boldsymbol{b}_{ij} \leftarrow \boldsymbol{b}_{ij} + \hat{\boldsymbol{u}}_{j|i} \cdot \boldsymbol{v}_j$

7）　**End for**

Output：Layer(l+1)capsules $\boldsymbol{v}_j$

由权重矩阵 $\boldsymbol{W}_{ij}$ 乘以输入 $\boldsymbol{u}_i$ 得到预测向量 $\hat{\boldsymbol{u}}_{j|i}$，对于每个低级别的 $\hat{\boldsymbol{u}}_{j|i}$，$\boldsymbol{c}_{ij}$ 表示对应的底层预测向量输出到高层向量的权重大小，$\boldsymbol{c}_{ij}$ 是由 $\boldsymbol{b}_{ij}$ 使用 softmax 函数得到的，$\boldsymbol{c}_{ij}$ 为非负数，且 $\sum_j \boldsymbol{c}_{ij}=1$；之后使用 $\mathbf{s}_j=\sum_i \boldsymbol{c}_{ij}\hat{\boldsymbol{u}}_{j|i}$ 计算所有预测向量 $\hat{\boldsymbol{u}}_{j|i}$ 的加权和，再使用新提出的 squash 函数得到输出向量 $\boldsymbol{v}_j$。在迭代过程中，权重 $\boldsymbol{b}_{ij}$ 初始赋值为零，采用公式 $\boldsymbol{b}_{ij}=\boldsymbol{b}_{ij}+\hat{\boldsymbol{u}}_{j|i}\cdot\boldsymbol{v}_j$ 更新权重 $\boldsymbol{b}_{ij}$。

2 选择性卷积核胶囊网络

本文提出选择性卷积核胶囊网络图像分类方法，在经典胶囊网络的第一层卷积层中加入选择性卷积核网络，利用网络的动态选择机制自适应地提取输入信息多个尺度的感受野，将不同分支的特征图根据权重进行融合。选择性卷积核网络结构如图 3 所示。

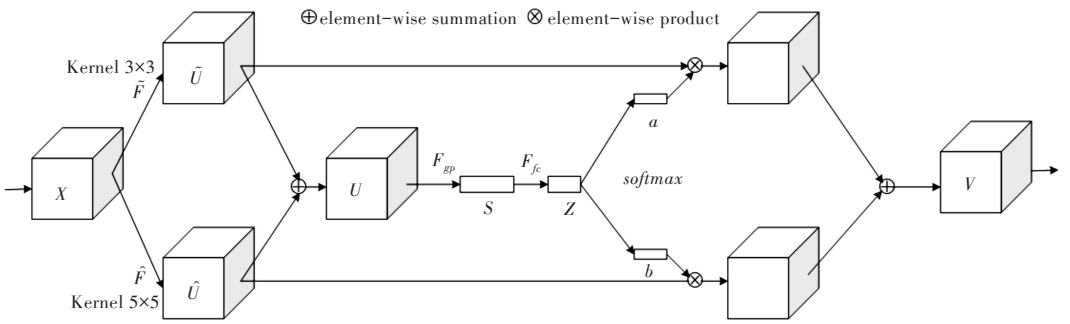

Fig. 3 Network model of selective convolution kernel

图 3 选择性卷积核网络模型

网络通过分离、融合、选择 3 个步骤实现动态选择卷积核。在标准的卷积神经网络中，每一层人工神经元的感受野被设计为相同大小。本网络的不同之处在于提出一种动态选择机制，使用两个分支动态提取输入图像特征。对于输入的特征图 $X\in R^{H'\times W'\times C'}$，并行进行两次变换，如式(1)、式(2)所示。

$$\tilde{F}:X\rightarrow\tilde{U}\in R^{H\times W\times C} \tag{1}$$

$$\hat{F}:X\rightarrow\hat{U}\in R^{H\times W\times C} \tag{2}$$

式中，两个 F 函数依次由高效卷积、批处理规范化与 Relu 函数组成。使用核尺寸大小分别为 3 和 5 的卷积核提取特征信息，使用门控制来自两个分支的信息流，两个分支携带不同尺寸的特征信息进入下一层神经元。首先将两个分支信息通过 element-wise summation 进行融合，$u=\tilde{u}+\hat{u}$，然后通过简单地使用全局平均池化嵌入全局信息生成信道统计信息，如式(3)所示。

$$s_c=F_{gp}(U_c)=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}U_c(i,j) \tag{3}$$

其中，F_{gp} 为全局平均池化函数，S_c 为信道统计信息。此外，为精确地引导自适应选择，创建一个紧凑特征 $z\in\boldsymbol{R}^{d\times 1}$，并通过全连接(fc)层进行实现，如公式(4)所示。

$$z=F_{fc}(s)=\delta(B(Ws)) \tag{4}$$

其中，δ 为 ReLU 激活函数，$\boldsymbol{B}$ 代表批处理，$\boldsymbol{W}\in\boldsymbol{R}^{d\times c}$。在压缩特征描述符 z 的指导下，使用跨通道的软注意机制自适应地选择不同空间的尺度信息，并在通道级数字上应用 softmax 操作，如式(5)所示。

$$a_c=\frac{e^{A_c z}}{e^{A_c z}+e^{B_c z}}\qquad b_c=\frac{e^{B_c z}}{e^{A_c z}+e^{B_c z}} \tag{5}$$

式中，a、b 分别是 $\tilde{u}$ 和 $\hat{u}$ 的软注意力矢量，最后特征由每个卷积核提取的特征乘以各自的注意力权重，再相加求和得到，如公式(6)所示。

$$V_c=a_c\cdot\tilde{U}+b_c\cdot\hat{U} \tag{6}$$

式中，V_c 为加权融合两个分支后的最终特征。

将选择性卷积核提取到的特征送入初始胶囊层中，由动态路由算法学习权重参数，数字胶囊层输出图片类别。胶囊网络中的 Conv1 层原本使用 256 个 9×9 卷积核提取特征，本文在该卷积操作后融入具有两个分支的选择性卷积核网络，分别使用 3×3 和 5×5 的卷积核并行提取图像特征，得到初级胶囊(primary capsule)的输入。初级胶囊是多维实体的最低级别，具有 32 个 8 通道的特征图，每个卷积层具有 8 个步幅为 2、尺寸为 9 的卷积核。PrimaryCapsules 的输出是 32×6×6 的胶囊(每个输出是一个 8 维矢量)，将得到的胶囊展成一维，与对应特征的胶囊进行组合得到向量神经元。其中，6×6 网格中每个胶囊彼此共享权重，权重通过路由算法进行学习更新。最后一层(DigitCaps)中的每个类由一个 16 维的向量胶囊表示，每个胶囊都接收来自上层级中所有胶囊的输出作为输入，依据胶囊向量的模长预测并输出图像类别。采用 Marginloss 作为损失函数，如公式(7)所示。

$$L_k=T_K\max(0,m^+-\|v_k\|)^2+\lambda(1-T_K)\max(0,\|v_k\|-m^-)^2 \tag{7}$$

式中，k 为分类类别。当 k 存在时，$T_k=1$；当 k 不存在时，$T_k=0$。

3 实验结果与分析

3.1 实验数据集

为测试选择性卷积核胶囊网络处理图像分类任务的有效性，采用图像分类任务中的 3 个经典数据集进行实验，分别为：CIFAR-10、Fashion-MNIST、SVHN。CIFAR-10 包

含 10 种类别的图片，共 6 万张；Fashion-MNIST 为涵盖不同商品的图片，共 7 万张；SVHN 为包含街景门牌号码的图片，共 10 万张。CIFAR-10、Fashion-MNIST 的测试集与训练集已划分好，SVHN 可根据不同网络结构自行划分测试集与训练集。数据集详细情况介绍如表 1 所示。

Table 1　Presentation of data sets

表 1　数据集详细情况介绍

数据集	类别数量	训练集	测试集
CIFAR-10	10	50 000	10 000
Fashion-MNIST	10	60 000	10 000

3.2　实验设置

本文实验环境如下：CPU 为 intel E5-2678V3，显卡为 GTX1080，内存为 16G，操作系统为 Windows10。使用 pytorch 深度学习框架，设置 batch_size 为 30，epochs 为 50，初始学习率为 0.001。在训练胶囊网络时，采用性能优越的 Adam 优化器。

设置选择性卷积核胶囊网络参数如下：路径数 M 为 2，采用两个分支结构，卷积核尺寸分别为 3×3 和 5×5，决定每个路径的组数 G 为 8，控制融合操作比率的参数 r 为 2，步长默认为 1。胶囊网络使用动态路由，激活函数为 Squash 函数，分类层的胶囊个数依据数据集包含的图片种类数量设定。

3.3　结果分析

采用基线胶囊网络与本文模型分别在 3 个数据集上进行对比实验。评价指标使用 top1 正确率，测试错误率迭代图分别如图 4-图 6 所示。模型分类精度对比如表 2 所示。

图 4-图 6 的测试结果迭代图可直观表现出本文模型的测试误差明显低于基线胶囊网络。表 2 列出了两个网络模型在 3 个公开数据集上的分类精度对比，其中基线胶囊网络是简单的具有两层卷积的网络结构，在 3 个数据集上得到的分类精度分别为 78.78%、92.89%、94.88%。本文模型是融入了选择性卷积核的胶囊网络模型，相比基线网络分类精度分别提升了 1.73%、0.69%、0.53%，表明选择性卷积核在提取特征时具有明显优势，弥补了传统网络提取特征时采用相同大小卷积核的缺点，可提取到更加丰富的图像分类特征。

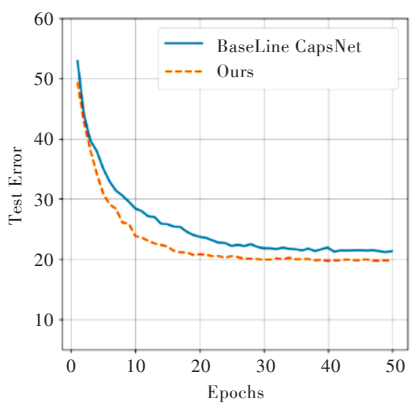

Fig. 4　CIFAR-10 dataset test results

图 4　CIFAR-10 数据集测试结果

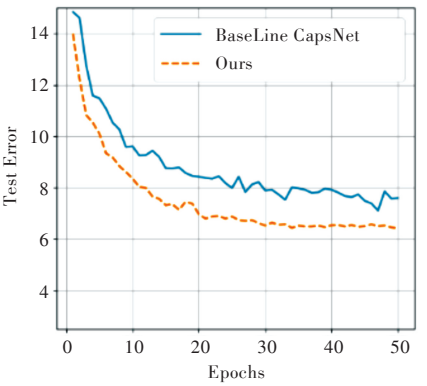

Fig. 5　Fashion-MNIST dataset test results

图 5　Fashion-MNIST 数据集测试结果

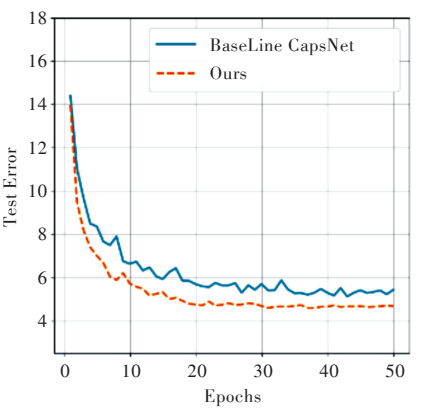

Fig. 6　SVHN dataset test results

图 6　SVHN 数据集测试结果

Table 2　Comparison of model classification precision

表 2　模型分类精度对比　　单位：%

模型	CIFAR-10	Fashion-MNIST	SVHN
Bas	78.78	92.89	94.88
Ours	80.51	93.58	95.41

4　结语

本文提出应用于图像分类任务的选择性卷积核胶囊网络，并在多个图像分类数据集中对模型进行实验测试与评估。针对卷积神经网络不能灵敏识别各特征之间相对位置关系的问题，采用矢量胶囊表征实体不同特征间的空间位置关系。最大池化导致网络不能精确反映实体特征与实体间的关系，采用胶囊网络的动态路由机制学习初级胶囊层与高级胶囊层之间的权重系数，以准确表征特征与实体的层级结构。针对传统卷积神经网络中每一层人工神经元的感受野被设计为相同大小这一问题，采用选择性卷积核网络融合不同尺寸卷积核提取的特征信息。在 CIFAR-10、Fashion-MNIST、SVHN 数据集上的测试结果表明，相比于基线胶囊网络，本文提出的网络模型具有更高的识别精度，分别为 80.51%、93.58%、95.41%。下一步研究将注意力机制与网络相结合以提高模型识别精度，注意力机制可抑制一些无关信息，着重关注所需的重要信息，以

提高模型分类准确率。

参考文献：

[1] ZHANG P P, LI Q S, YANG C H. Image classification algorithm based on lightweight group attention module[J]. Computer Applications, 2020, 40(3): 645-650.
张盼盼, 李其申, 杨词慧. 基于轻量级分组注意力模块的图像分类算法[J]. 计算机应用, 2020, 40(3): 645-650.

[2] BEI C Y, YU H B, PAN M, et al. Gland cell image segmentation algorithm based on improved U-Net network[J]. Electronic Technology, 2019, 32(11): 18-22.
贝琛圆, 于海滨, 潘勉, 等. 基于改进U-Net网络的腺体细胞图像分割算法[J]. 电子科技, 2019, 32(11): 18-22.

[3] LI K J, CHEN S B, LI W Q. Simulation of traffic sign detection system based on deep learning [J]. Software Guide, 2020, 19(9): 31-34.
李克俭, 陈少波, 李万琦. 基于深度学习的交通标志检测系统仿真[J]. 软件导刊, 2020, 19(9): 31-34.

[4] LIU T B. Research on face detection algorithm based on convolutional neural network[J]. Software Guide, 2020, 19(10): 66-70.
刘天保. 基于卷积神经网络的人脸检测算法研究[J]. 软件导刊, 2020, 19(10): 66-70.

[5] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[6] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25: 1097-1105.

[7] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks [C]//European Conference on Computer Vision, 2014: 818-833.

[8] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [DB/OL]. https://arxiv. org/abs/1409.1556.

[9] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1-9.

[10] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[11] HUANG G, LIU Z, VAN D M L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4700-4708.

[12] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.

[13] SABOUR S, FROSST N, HINTON G E. Dynamic routing between capsules [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 3859-3869.

[14] XIANG C, ZHANG L, TANG Y, et al. MS-CapsNet: a novel multi-scale capsule network[J]. IEEE Signal Processing Letters, 2018, 25(12): 1850-1854.

[15] ZHANG X, SUN Y, WANG Y, et al. A novel effective and efficient capsule network via bottleneck residual block and automated gradual pruning[J]. Computers & Electrical Engineering, 2019, 80: 106481.

[16] DUARTE K, RAWAT Y S, SHAH M. VideoCapsuleNet: a simplified network for action detection[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018: 7621-7630.

[17] CHEN Y, DAI X, LIU M, et al. Dynamic convolution: attention over convolution kernels [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11030-11039.

[18] SU Z, FANG L, KANG W, et al. Dynamic group convolution for accelerating convolutional neural networks [C]//European Conference on Computer Vision, 2020: 138-155.

[19] YANG B, BENDER G, LE Q V, et al. CondConv: conditionally parameterized convolutions for efficient inference[DB/OL]. https://arxiv.org/abs/1904.04971.

[20] CHEN J, WANG X, GUO Z, et al. Dynamic region-aware convolution[DB/OL]. https://arxiv.org/abs/2003.12243.

[21] TIAN Z, SHEN C, CHEN H. Conditional convolutions for instance segmentation[DB/OL]. https://arxiv.org/abs/2003.05664v3.

[22] LI Y, CHEN Y, DAI X, et al. Revisiting dynamic convolution via matrix decomposition[DB/OL]. https://arxiv.org/abs/2103.08756v1.

[23] HINTON G E, KRIZHEVSKY A, WANG S D. Transforming auto-encoders[C]//International Conference on Artificial Neural Networks & Machine Learning, 2011: 44-51.

（责任编辑：黄 健）