激光与光电子学进展 先进成像

融合张量合成注意力的改进ResNet图像分类模型

邱云飞[1]，张家欣[1*]，兰海[2]，宗佳旭[3]

[1]辽宁工程技术大学软件学院，辽宁 葫芦岛 125105；

[2]中国科学院海西研究院泉州装备制造研究所，福建 泉州 362216；

[3]元启工业技术有限公司，山东 青岛 266000

摘要 针对卷积神经网络处理图像分类任务时提取特征不充分以及提取到的特征不区分贡献度的问题，提出了一种融合张量合成注意力的改进ResNet-101（RTSA Net-101）网络模型。首先，利用ResNet-101骨干网络提取图像特征，并在残差网络卷积结构后嵌入张量合成注意力模块，对获取的特征进行三张量积计算，得到注意力特征矩阵；然后，使用Softmax函数对注意力特征矩阵进行归一化，从而为特征分配权重，以区分特征的贡献度；最后，将得到的权重和对应的键值加权求和，获取最终图像完整特征，以提升模型的图像分类精度。在自然图像数据集CIFAR-10、CIFAR-100和街牌号数据集SVHN上进行了对比实验，模型分类准确率分别为96.12%、81.60%、96.67%，图像平均测试运行时间分别为0.0258 s、0.0260 s、0.0262 s。实验结果表明：相比于其他7种先进图像分类模型，RTSA Net-101模型可以获得更高的分类准确率和更短的测试运行时间，且能够有效地增强网络的特征学习能力，具有一定的创新性、高效性。

关键词 张量合成注意力；残差网络；自注意力；特征提取；图像分类

中图分类号 TP391 **文献标志码** A **DOI**：10.3788/LOP212836

Improved ResNet Image Classification Model Based on Tensor Synthesis Attention

Qiu Yunfei[1], Zhang Jiaxin[1*], Lan Hai[2], Zong Jiaxu[3]

[1]*College of Software, Liaoning Technical University, Huludao* 125105, *Liaoning, China*;

[2]*Quanzhou Institute of Equipment Manufacturing Haixi Institutes, Chinese Academy of Sciences, Quanzhou* 362216, *Fujian, China*;

[3]*Yuanqi Industrial Technology Company, Qingdao* 266000, *Shandong, China*

Abstract An improved ResNet-101 network model that fuses tensor synthesis attention (RTSA Net-101) is proposed to solve insufficient feature extraction and the indiscriminate contribution of the extracted features when processing image classification tasks using a convolutional neural network. First, the image features are extracted using a Resnet-101 backbone network and the tensor synthesis attention module is embedded after the convolution structure of the residual network. The features are calculated using a three-tensor product to obtain the attention feature matrix. Next, the Softmax function is used to normalize the attention feature matrix to assign weights to features and distinguish the contribution of features. Finally, the weighted sum of the weights and critical values are calculated as the final features in our proposed method to improve the image classification performance. Comparative experiments are conducted on natural image datasets, CIFAR-10 and CIFAR-100, and street brand dataset, SVHN. The classification accuracy values of the models are 96.12%, 81.60%, and 96.67%, respectively, and the average test running time of the images are 0.0258 s, 0.0260 s, and 0.0262 s, respectively. The experimental results show that compared with the other seven advanced image classification models, the RTSA Net-101 model can achieve higher classification accuracy and shorter test run time, and it can effectively enhance the feature learning ability of the network, thereby render the proposed model innovative and efficient.

Key words tensor synthesis attention; residual network; self-attention; feature extraction; image classification

1 引 言

近年来，卷积神经网络（Convolutional neural network，CNN）由于其强大的自动提取特征能力逐步成为图像分类领域的研究热点[1]。然而卷积神经网络处理图像分类问题时存在提取特征不充分以及提取到

收稿日期：2021-10-29；**修回日期**：2022-01-01；**录用日期**：2022-01-17；**网络首发日期**：2022-01-27

通信作者：*1595775491@qq.com

的特征不区分贡献度的问题，针对该问题Wang等[2]提出了非局部块(Non-local block，NL Block)，融合了自注意力机制与全局信息以获取更充分的特征信息，但是该网络不能精确地区分特征的贡献度，且计算开销很大[3]。对此，Hu等[4]提出了压缩-激励网络(Squeeze-and-excitation network，SENet)，该网络通过“挤压”“激励”操作建模了通道特征的相关性，在加强有效特征的同时抑制了无效特征，然而却不能够充分利用全局上下文信息[4]。针对SENet的缺点，Roy等[5]提出了空间通道压缩-激励模块(Concurrent spatial and channel squeeze and channel excitation，scSE)，该模块在充分获取全局特征信息的同时，利用注意力机制增强了更重要的特征，多见于图像分割领域。黄盛等[6]提出了基于改进深度残差网络的计算断层扫描图像分类算法[4]，使用该策略可以较好地区分特征的贡献度，并在达到较高分类准确率的同时显著降低了计算量。受SENet启发，Woo等[2,7]提出了卷积注意力模块(Convolutional block attention module，CBAM)，先后使用了通道注意力模块和空间注意力模块，可以为图像提取到更加精确的特征信息。

然而，上述方法中的自注意力机制高度依赖点积操作，这会导致穷举和冗余计算，引起的额外参数和内存消耗使训练过程面临巨大的挑战，特别是在处理大规模数据方面。因此，Tay等[8-9]提出了一种新的模型Synthesizer，该模型提出了新的注意力矩阵学习方法，只通过简单的前馈神经网络便可以得到注意力分数，省去了token之间的点积交互，在自然语言处理任务上取得了很好的效果。然而与一维自然语言向量不同的是，视觉特征通常以多维张量的形式呈现，很难直接学习Synthesizer的线性维数对齐，也很难学习用自注意力机制处理图像特征时，重塑张量特征的尺寸以保持$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$之间的维数对齐，这可能会破坏张量特征的内部结构。

针对以上图像分类问题，提出了一种融合张量合成注意力(Tensor-synthetic attention，TSA)的改进ResNet图像分类模型，主要贡献如下：

1) 提出了一个TSA模块。与以往自注意力模块不同，TSA模块直接对图像张量特征进行三张量积计算，使具有精确图像类别信息的特征得到加强的同时对特征进行权重分配，进而区分特征的贡献度。该模块可以很容易地适应输入数据的结构，不需要重塑原始输入，更好地保存了特征映射的底层结构。

2) 改进了ResNet-101残差网络模型。在ResNet-101残差网络卷积结构后加入了TSA模块，并将改进后的模型命名为RTSA Net-101(Residual Net-101 with tensor-synthetic attention)，其解决了原模型提取的特征分辨性弱的问题，进一步提高了图像分类精度。

2　相关工作

2.1　ResNet

何凯明等[10]提出了残差网络ResNet，其结构如图1所示。一般网络训练多以学习输入输出之间的映射关系为目标，而ResNet主要学习输入x、输出y之间的残差关系$f(x)=y-x$，并计算$f(x)+x$，以获得相应的输出。原始输入x通过跳跃连接与输出直接连接，在很大程度上避免了信息的损失，降低了学习难度[11-12]。

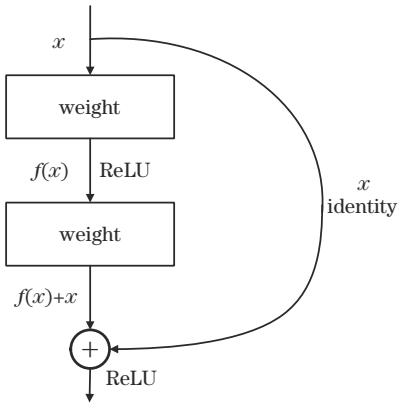

图1　残差块结构

Fig. 1　Structure of residual block

2.2　自注意力机制

图像分类领域中自注意力机制[13]的基本思想是让网络能够在众多特征中忽略无关特征并关注到对当前分类任务更重要的特征[2]。其本质为查询$\boldsymbol{Q}$(Query)到一系列键值对$\boldsymbol{K}$-$\boldsymbol{V}$(Key-Value)的映射，具体做法为采用Encoder-Decoder框架[14]。假设$\boldsymbol{\chi}\in\mathbb{R}^{H\times W\times C}$为一个高度为$H$、宽度为$W$、通道数为$C$的张量输入，为将其送入自注意力机制，首先要将3D张量$\boldsymbol{x}\in\mathbb{R}^{H\times W\times C}$转变为一个2D矩阵$\boldsymbol{X}\in\mathbb{R}^{HW\times C}$。然后将$\boldsymbol{X}$通过映射得到相应的表示$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$，映射分别用$\boldsymbol{Q}=\varnothing_{\mathrm{Q}}(\boldsymbol{X})$，$\boldsymbol{K}=\varnothing_{\mathrm{K}}(\boldsymbol{X})$，$\boldsymbol{V}=\varnothing_{\mathrm{V}}(\boldsymbol{X})$表示[15]，其中，$\varnothing_{\mathrm{Q}}$、$\varnothing_{\mathrm{K}}$、$\varnothing_{\mathrm{V}}$为可训练的变换，如果变换为线性映射，则有对应的标识：$\boldsymbol{Q}=\boldsymbol{X}\boldsymbol{W}_{\mathrm{Q}}$，$\boldsymbol{K}=\boldsymbol{X}\boldsymbol{W}_{\mathrm{K}}$，$\boldsymbol{V}=\boldsymbol{X}\boldsymbol{W}_{\mathrm{V}}$，其中，$\boldsymbol{W}_{\mathrm{Q}}$、$\boldsymbol{W}_{\mathrm{K}}$、$\boldsymbol{W}_{\mathrm{V}}$为线性映射[16]。将$\boldsymbol{Q}$和每个$\boldsymbol{K}$进行相似度计算得到权重$\boldsymbol{S}$，再利用Softmax函数对权重进行归一化，最后将权重$\boldsymbol{S}$和相应的键值$\boldsymbol{V}$进行加权求和得到最终的$\boldsymbol{Y}$(Attention)[17]。其基本结构如图2所示。

自注意力机制计算公式为

$$\boldsymbol{S}=\mathrm{Softmax}\left(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}/\sqrt{d}\right),\tag{1}$$

$$\boldsymbol{Y}=\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\boldsymbol{S}\boldsymbol{V},\tag{2}$$

式中：d为键值的维数；当$\boldsymbol{Q}=\boldsymbol{K}=\boldsymbol{V}$时，则称其为自注意力机制。

2.3　Synthesizer结构

Tay等[9]重新思考了Transformer中注意力机制的必要性，并提出了Synthesizer结构，如图3所示。

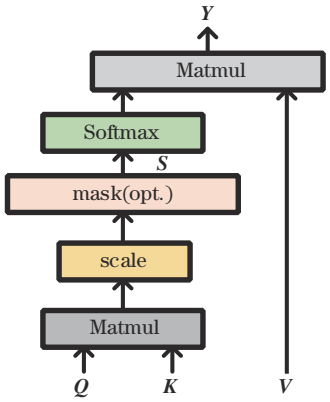

图 2　自注意力机制结构

Fig. 2　Structure of self-attention

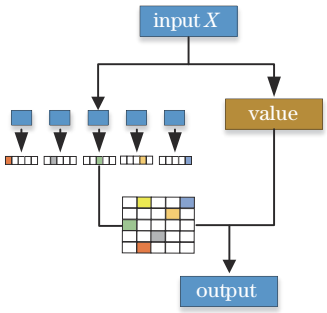

图 3　Synthesizer结构

Fig. 3　Structure of Synthesizer

Synthesizer 去除了自注意力机制中查询键值的概念，直接采用了合成注意力（Synthetic attention）来代替。

这种方式将输入 $\boldsymbol{X}\in\mathbb{R}^{L\times D}$ 映射到输出 $\boldsymbol{Y}\in\mathbb{R}^{L\times D}$，$L$ 为指序列长，D 为模型的维数。采用 $F(\cdot)$ 作为一个参数化函数，用于将输入 $\boldsymbol{X}_i$ 从 D 维投影到 L 维以得到 $\boldsymbol{B}_i\in\mathbb{R}^{L\times L}$，$\boldsymbol{B}_i$ 的计算公式为

$$\boldsymbol{B}_i=F(\boldsymbol{X}_i),\tag{3}$$

式中：i 为 $\boldsymbol{X}$ 的第 i 个 token，可以解释为通过此方法，每个 token 为输入序列中其他的 token 预测权重；映射 F 代替了 $\boldsymbol{QK}^{\mathrm{T}}$ 的计算，整体可以表示为

$$F(\boldsymbol{X})=W_2\left(\delta\left(W_1(\boldsymbol{X})+b_1\right)+b_2\right),\tag{4}$$

式中：δ 为 ReLU 的激活函数，b_1、b_2 为偏置，$W_1\in\mathbb{R}^{D\times D}$ 和 $W_2\in\mathbb{R}^{D\times L}$，因此 $\boldsymbol{B}\in\mathbb{R}^{L\times L}$。此时，输出 $\boldsymbol{Y}$ 为

$$\boldsymbol{Y}=\mathrm{Softmax}(\boldsymbol{B})G(\boldsymbol{X}),\tag{5}$$

式中：$G(\boldsymbol{X})$ 为另一个 $\boldsymbol{X}$ 的参数化函数，类似于注意力机制中的 $\boldsymbol{V}$。

该方法用合成函数 $F(\cdot)$ 代替了标准自注意力机制中的 $\boldsymbol{QK}^{\mathrm{T}}$ 计算，从而完全消除了 $\boldsymbol{QK}$ 间的点积操作。

3　融合 TSA 的改进 ResNet 图像分类模型

针对图像分类问题，提出了一种融合 TSA 的改进 ResNet 图像分类模型 RTSA Net-101，本节将对该模型的设计进行细致阐述。

3.1　RTSA Net-101

当前，卷积神经网络俨然已经成为提取图像特征的重要手段。但随着模型不断加深，梯度消失问题也涌现出来[18]。残差网络因其独特的短路连接解决了上述问题，并在简化学习目标的同时保护了信息的完整性，被广泛应用于各类图像分类任务。本文提出了 RTSA Net-101 网络模型，模型结构如图 4 所示。模型将 TSA 模块嵌入到残差网络模型结构中，在避免了信息在传递过程中丢失与损耗的同时，使模型聚焦分辨性高的特征，提高了模型的分类性能。

由图 4 可知，RTSA Net-101 网络模型将原始图像作为输入，首先，对图像进行裁剪、归一化得到预处理后的图像。这些预处理操作可以使训练集更丰富，从而让模型更具泛化能力。然后，将图像输入到模型卷

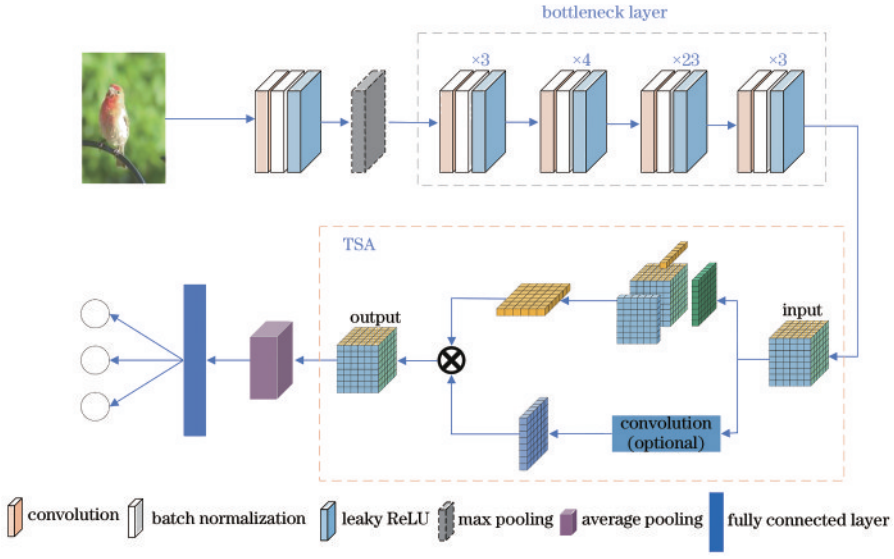

图 4　RTSA Net-101 模型结构概况

Fig. 4　Overview of RTSA Net-101 model structure

积层提取图像特征[18]，卷积层采用1次过滤器大小为3×3的卷积对特征映射进行降维；在每次卷积后，均采用batch normalization[19]来缓解过拟合，并采用leaky ReLU激活函数充分利用梯度信息，保证模型不断收敛；再采用最大池化操作减少特征映射的参数，将图像降维[20]。随即是4组瓶颈残差块，每组残差块卷积操作采用的过滤器大小分别为1×1、3×3、1×1，这种方式可以达到减少计算量和参数量的效果。接着将获取的特征输入TSA模块中，TSA模块嵌入到残差网络卷积结构后，使卷积提取到的特征在被送入平均池化层前，对特征进行权重分配，进而区分特征的贡献度[2]。再利用平均池化层提取TSA模块输出特征映射的空间信息，最后经过全连接层输出图像分类的结果。

3.2　TSA模块

在残差网络中加入自注意力机制能够让网络模型更加关注感兴趣区域，使得模型聚焦所要提取的重要特征，忽略不相关特征；然而自注意力机制的点积操作会导致张量内部结构受到破坏。因此，为了更准确地分析具有复杂背景的图像数据，充分利用图像信息中的显著特征，并直接处理张量输入$\boldsymbol{\chi}\in\mathbb{R}^{T\times H\times W\times C}$，提出了TSA模块。其中包括基础TSA模块和随机张量合成注意力(Tensor synthetic attention random，TSAR)模块两种形式。

给定了张量输入$\boldsymbol{\chi}\in\mathbb{R}^{H\times W\times C}$和其二维重塑矩阵$\boldsymbol{X}\in\mathbb{R}^{HW\times C}$，自注意力机制会产生一个从$\boldsymbol{X}\in\mathbb{R}^{HW\times C}\rightarrow\boldsymbol{X}\in\mathbb{R}^{HW\times HW}$：$\boldsymbol{Q}\rightarrow\boldsymbol{S}$的维度对齐，其中$\boldsymbol{S}$为权重。形式上，设定输入张量$\mathcal{Q}\in\mathbb{R}^{H\times W\times d}$通过映射$Q=\varnothing_{\mathrm{Q}}(\boldsymbol{X})$转换为$\boldsymbol{\chi}$的投影特征，其中$\varnothing_{\mathrm{Q}}(\boldsymbol{X})$被定义为三张量积：

$$\varnothing_{\mathrm{Q}}(\boldsymbol{X})=\mathcal{Q}\times\boldsymbol{A}_{(H)}^{(l)}\times\boldsymbol{A}_{(W)}^{(l)}\times\boldsymbol{A}_{(C)}^{(l)},\tag{6}$$

式中：3个张量$\boldsymbol{A}_{(H)}^{(l)}\in\mathbb{R}^{HW\times H}$，$\boldsymbol{A}_{(W)}^{(l)}\in\mathbb{R}^{HW\times W}$，$\boldsymbol{A}_{(C)}^{(l)}\in\mathbb{R}^{1\times d}$，$l$为层数，三张量积的示例[21]如图5所示。

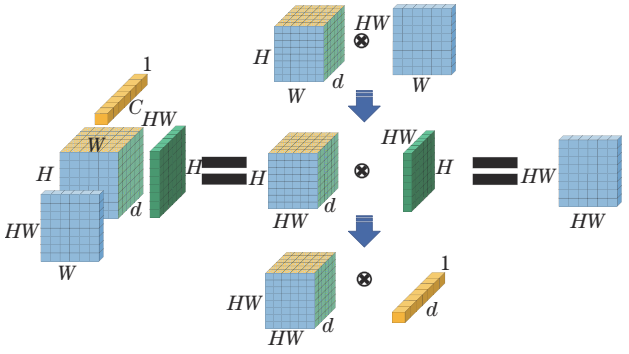

图5　三张量积

Fig. 5　Three-tensor product

为了在没有点积操作情况下在原始张量$\mathcal{Q}$上实现张量的维数对齐，提出了一种TSA模块，其结构如图6所示。该模块首先对输入$\mathcal{Q}$进行三张量积变换以得到注意力矩阵$\boldsymbol{Z}$，并使用Softmax函数对$\boldsymbol{Z}$进行归一化得到权重$\boldsymbol{S}$，从而对特征进行权重分配，以区分特征的贡献度。具体计算方式为

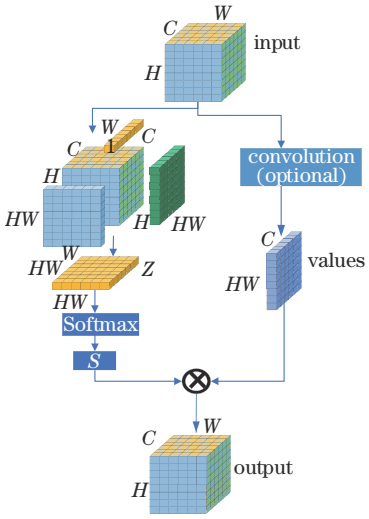

图6　张量合成注意力模块

Fig. 6　Tensor synthesis attention module

$$\boldsymbol{Z}=\mathcal{Q}\times\boldsymbol{A}_{(H)}^{(l+1)}\times\boldsymbol{A}_{(W)}^{(l+1)}\times\boldsymbol{A}_{(C)}^{(l+1)},\tag{7}$$

$$\boldsymbol{S}=\mathrm{Softmax}(\boldsymbol{Z}),\tag{8}$$

式中：3个张量$\boldsymbol{A}_{(H)}^{(l+1)}\in\mathbb{R}^{HW\times H}$，$\boldsymbol{A}_{(W)}^{(l+1)}\in\mathbb{R}^{HW\times W}$，$\boldsymbol{A}_{(C)}^{(l+1)}\in\mathbb{R}^{1\times d}$，$\boldsymbol{Z}\in\mathbb{R}^{HW\times HW\times 1}$，$\boldsymbol{S}\in\mathbb{R}^{HW\times HW\times 1}$。接着，设定$\boldsymbol{V}=\varnothing_{\mathrm{V}}(\boldsymbol{X})\in\mathbb{R}^{HW\times m}$为张量合成注意力模块中的value值，将权重$\boldsymbol{S}$和相应的键值$\boldsymbol{V}$进行加权求和来获取最终图像完整特征输出$\boldsymbol{Y}$，以提升模型的图像分类精度。$\boldsymbol{Y}$的计算方式为

$$\boldsymbol{Y}=\mathrm{Tensor\ Synthetic\ Attention}(\mathcal{Q},\boldsymbol{V})=\boldsymbol{SV}。\tag{9}$$

该模块主要关注特征矩阵中有价值的信息，提取特定目标中分辨性较强的特征。通过合成函数替换掉了自注意力机制中的$\boldsymbol{QK}^{\mathrm{T}}$计算，不需要重塑原始输入，减少了对输入张量的依赖，更容易保存特征映射的底层结构。

此外，考虑到TSA模块很难确定3个张量$\boldsymbol{A}_{(H)}^{(l+1)}$、$\boldsymbol{A}_{(W)}^{(l+1)}$、$\boldsymbol{A}_{(C)}^{(l+1)}$的维数，即如何将维数$HW$、$HW$、1分配给3个张量乘法，尝试采用$\boldsymbol{Z}_{\mathrm{H}}$、$\boldsymbol{Z}_{\mathrm{W}}$、$\boldsymbol{Z}_{\mathrm{C}}$分别作为通过高、宽、通道数3个方向完成对原始张量$\mathcal{Q}\in R^{H\times W\times d}$的三张量积变换后得到的矩阵，结构如图7所示，具体操作如下：

$$\boldsymbol{Z}_{\mathrm{H}}=\mathcal{Q}\times\boldsymbol{A}_{(H)}^{(l+1)}\times\boldsymbol{A}_{(W)}^{(l+1)}\times\boldsymbol{A}_{(C)}^{(l+1)}。\tag{10}$$

其表示将维数1分配给高方向张量，将维数HW、HW分配给宽、通道2个方向张量后得到的注意力矩阵$\boldsymbol{Z}_{\mathrm{H}}$，即3个张量$\boldsymbol{A}_{(H)}^{(l+1)}\in\mathbb{R}^{1\times H}$、$\boldsymbol{A}_{(W)}^{(l+1)}\in\mathbb{R}^{HW\times W}$、$\boldsymbol{A}_{(C)}^{(l+1)}\in\mathbb{R}^{HW\times d}$。

$$\boldsymbol{Z}_{\mathrm{W}}=\mathcal{Q}\times\boldsymbol{A}_{(H)}^{(l+1)}\times\boldsymbol{A}_{(W)}^{(l+1)}\times\boldsymbol{A}_{(C)}^{(l+1)}。\tag{11}$$

其表示将维数1分配给宽方向张量，将维数HW、HW分配给高、通道2个方向张量后得到的注意力矩阵$\boldsymbol{Z}_{\mathrm{W}}$，即3个张量$\boldsymbol{A}_{(H)}^{(l+1)}\in\mathbb{R}^{HW\times H}$、$\boldsymbol{A}_{(W)}^{(l+1)}\in\mathbb{R}^{1\times W}$、$\boldsymbol{A}_{(C)}^{(l+1)}\in\mathbb{R}^{HW\times d}$。

$$\boldsymbol{Z}_{\mathrm{C}}=\mathcal{Q}\times\boldsymbol{A}_{(H)}^{(l+1)}\times\boldsymbol{A}_{(W)}^{(l+1)}\times\boldsymbol{A}_{(C)}^{(l+1)}。\tag{12}$$

其表示将维数 1 分配给通道方向张量，将维数 HW、HW 分配给高、宽 2 个方向张量后得到的注意力矩阵 $\boldsymbol{Z}_{\mathrm{C}}$，即 3 个张量 $\boldsymbol{A}_{(H)}^{(l+1)}\in\mathbb{R}^{HW\times H}$、$\boldsymbol{A}_{(W)}^{(l+1)}\in\mathbb{R}^{HW\times H}$、$\boldsymbol{A}_{(C)}^{(l+1)}\in\mathbb{R}^{1\times d}$。

上述 TSA 模块是将每个输入 $\boldsymbol{\chi}\in\mathbb{R}^{H\times W\times C}$ 投影到维数 HW，其作用就是张量维数的自对齐。基于此，提出了另一种形式的 TSA 模块，即 TSAR 模块，结构如图 7 所示。

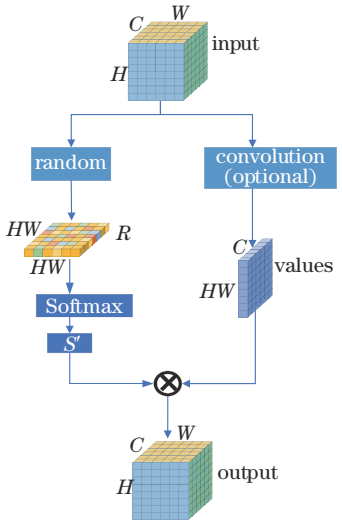

图 7　TSAR 模块

Fig. 7　TSAR module

不同于基础 TSA 模块对原始张量 $\mathcal{Q}$ 进行的三张量积变换以得到注意力矩阵 $\boldsymbol{Z}$，TSAR 模块尝试将注意力矩阵直接设置为随机矩阵 $\boldsymbol{R}$，并使用 Softmax 函数对 $\boldsymbol{R}$ 进行归一化得到权重 $\boldsymbol{S}'$，如此，注意力权重的初始化可以不受任何输入 token 的影响，完全随机初始化，且这些随机初始化的值可以被训练，或保持固定。接着，同基础 TSA 模块一样，设定 $\boldsymbol{V}=\varnothing_{\mathrm{V}}(\boldsymbol{X})\in\mathbb{R}^{HW\times m}$ 为 TSAR 中的 value，将权重 $\boldsymbol{S}'$ 和相应的键值 $\boldsymbol{V}$(Value)进行加权求和来获取最终图像完整特征输出 $\boldsymbol{Y}$，以提升模型的图像分类精度。将 $\boldsymbol{R}$、$\boldsymbol{Y}$ 定义为

$$\boldsymbol{S}'=\mathrm{Softmax}(\boldsymbol{R}),\tag{13}$$

$$\boldsymbol{Y}=\mathrm{Tensor\ Synthesis\ Attention\ Random}(\mathcal{Q},\boldsymbol{V})=\boldsymbol{S}'\boldsymbol{V},\tag{14}$$

式中：$\boldsymbol{R}\in\mathbb{R}^{HW\times HW}$ 为注意力矩阵随机初始化的值。TSAR 模块不用依赖 token 对之间的交互，较传统自注意力机制更能保证张量特征的内部结构完整性。

4　分析与讨论

4.1　数据集

采用了 3 个图像分类基准数据集，自然图像数据集 CIFAR-10[22]、CIFAR-100[22]、街牌号数据集 SVHN[23]，来验证 RTSA Net-101 模型的有效性 。

CIFAR-10 和 CIFAR-100 数据集均包括 60000 张 32×32 的彩色图像，其中训练集 50000 张，测试集 10000 张[24]。此外，CIFAR-10 包含 10 个类别图像，CIFAR-100 包含 100 个类别图像[24]。

SVHN 是一个来自真实世界的街牌号数据集[23]，该数据集分为 10 类，数据均为 32×32 的彩色街牌号图像，包含 73257 个训练数据和 26032 个测试数据。每张图像上有一个或者多个数字，以正中间的数据作为判断图像类别的依据。数据集详细信息如表 1 所示。

表 1　数据集详细情况

Table 1　Details of datasets

Dataset	Number of data	Category	Training set	Testing set
CIFAR-10	60000	10	50000	10000
CIFAR-100	60000	100	50000	10000
SVHN	99289	10	73257	26032

4.2　实验设置

实验采用的计算机环境为型号 Intel(R) i7-6850k 的处理器，内存为 64 G，实验模型均是使用 PyTorch 深度学习框架实现，并在 NVIDIA 2080Ti GPU 训练得到的。实验训练阶段优化器采用自适应矩估计(Adaptive moment，Adam)算法训练 100 轮，批量样本大小设置为 100。使用准确率(Accuracy)、召回率(Recall)、精准率(Precision)、宏平均 F_1 值(F1)，运行时间(Run time)作为图像分类的评价指标。

4.3　模块鲁棒性实验及结果分析

为了验证 TSA 模块对图像分类的有效性，进行了模块鲁棒性实验，实验搭建了 3 层简单 CNN 网络作为基本骨干网络，将 CIFAR-10 数据集作为实验数据集，分别融合了传统自注意力机制、Synthesizer 各个不同系列的注意力模块，以及所提出的 TSA、TSAR 模块，进行快速实验。为方便表示，各类方法采用缩写方式，方法缩写如表 2 所示。

表 2　各种方法缩写

Table 2　Abbreviations of various methods

Method abbreviation	Concrete model
None	CNN
CA	CNN+Attention
CSD	CNN+Synthesizer Dense
CFSD	CNN+Factorized Synthesizer Dense
CFSR	CNN+Factorized Synthesizer Random
CMS	CNN+Mixture Synthesizers
CTSA(C)	CNN+Tensor Synthetic Attention Channel
CTSA(H)	CNN+Tensor Synthetic Attention Hight
CTSA(W)	CNN+Tensor Synthetic Attention Width
CTSAR	CNN+Tensor Synthetic Attention Random

为了测试各种方法对高斯噪声干扰的鲁棒性，在 CIFAR-10 数据集中加入了不同强度的高斯噪声，利用这种方式破坏图像以进行模块鲁棒性评估，实验结果见表 3、图 8。

表 3　不同方法对高斯噪声图片的分类准确率结果

Table 3　Classification accuracy results of different methods for Gaussian noise images

Method	Gaussian noise									
	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
None	45.96	34.35	27.80	22.83	20.01	17.86	16.23	15.07	14.45	13.76
CA	47.02	35.83	29.21	24.70	21.49	19.33	17.41	16.51	15.42	14.80
CSD	48.71	37.79	31.54	27.30	23.70	22.23	20.18	18.92	17.45	17.03
CFSR	46.51	37.28	31.69	27.56	24.14	22.60	20.62	19.75	18.25	17.76
CFSD	48.90	37.50	30.42	25.89	22.85	20.70	19.33	17.91	16.87	16.07
CMS	48.20	36.80	29.55	25.13	21.45	19.34	17.74	16.61	15.76	15.26
CTSA(*C*)	**49.29**	**36.81**	**31.98**	**28.11**	**25.49**	**23.75**	**22.39**	**21.44**	**20.09**	**19.54**
CTSA(*H*)	44.66	30.92	23.26	19.58	17.08	15.12	14.04	13.35	12.61	12.26
CTSA(*W*)	42.20	30.21	23.65	19.92	17.53	16.41	15.18	14.65	14.33	13.64
CTSAR	45.93	34.02	27.06	23.01	19.91	17.36	16.02	14.93	13.93	13.40

注：加粗字体表示最优结果。

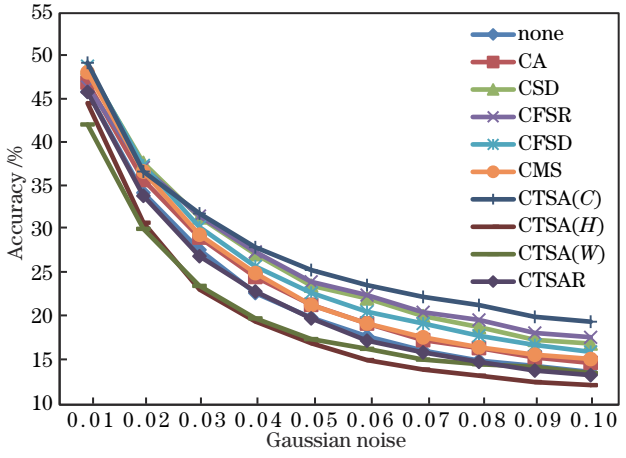

图 8　不同方法在高斯噪声图像上的分类准确率对比

Fig. 8　Comparison of classification accuracy of different methods on Gaussian noise images

由表 3、图 8 可以清晰看出，当噪声比较小，如 N=0.01、0.02 时，CTSA(C)、CFSR、CSD、CFSD、CMS 的准确率都比较高，且相差不大。其中 CTSA(C)的准确率一直处于最优位置；但随着噪声越来越大，会发现这些方法与 CTSA(C)的准确率相差逐渐变大。由此可见，TSA 模块在 CNN 基础上保留了图像丰富结构信息，相比其他方法，抗噪声干扰能力更强，分类效果更佳；而 CTSAR、CTSA(H)、CTSA(W)的分类准确率比原始 CNN 和 CA 要低，鲁棒性相对较差，分析原因可能是张量相乘破坏了原有图像的结构，所以容易受到噪声的干扰。

随机旋转是图像分类中经常面临的场景之一，也是考验分类模型优劣的重要指标。实验将 CIFAR-10 数据集原始图像随机旋转 30°到 330°，且图像增大的角度范围均匀采样，以此来评估这些方法在不同旋转角度下的鲁棒性，实验结果见表 4、图 9。

表 4　不同方法对旋转图片的分类准确率结果

Table 4　Classification accuracy results of different methods for rotating pictures

Method	Rotation angle/(°)										
	30	60	90	120	150	180	210	240	270	300	330
None	37.58	26.51	23.07	21.12	19.40	19.39	18.68	18.89	18.27	17.74	17.40
CA	36.54	36.57	23.27	21.25	19.75	19.44	19.70	18.61	18.19	18.46	17.75
CSD	37.90	38.17	23.24	21.79	19.26	19.45	19.36	18.51	17.55	17.60	17.98
CFSR	37.04	26.86	23.10	20.69	19.72	18.69	18.70	18.27	17.69	17.24	17.08
CFSD	39.27	28.85	23.55	22.85	20.37	19.85	19.95	18.71	18.86	18.34	17.88
CMS	37.91	26.93	22.96	21.74	19.61	18.89	18.76	18.01	18.06	17.62	17.65
CTSA(*C*)	38.71	38.36	24.51	21.80	20.97	20.36	20.89	19.93	19.86	19.12	20.00
CTSA(*H*)	42.32	30.44	26.28	24.53	22.17	22.18	**22.37**	**21.83**	**20.85**	**20.07**	**20.16**
CTSA(*W*)	**45.53**	**38.64**	**27.21**	**25.20**	**22.39**	**22.23**	22.03	21.27	20.61	20.01	19.89
CTSAR	33.72	24.70	22.06	19.88	19.00	18.17	18.03	17.22	17.03	17.23	16.79

注：加粗字体表示最优结果。

由表 4、图 9 可知，随着图像旋转角度越来越大，所有方法的分类性能整体呈下降趋势。在整个旋转过程中，旋转角度为 30°至 180°时，CTSA(W)方法的分类准确率最优；旋转角度为 210°至 330°时，CTSA(H)方

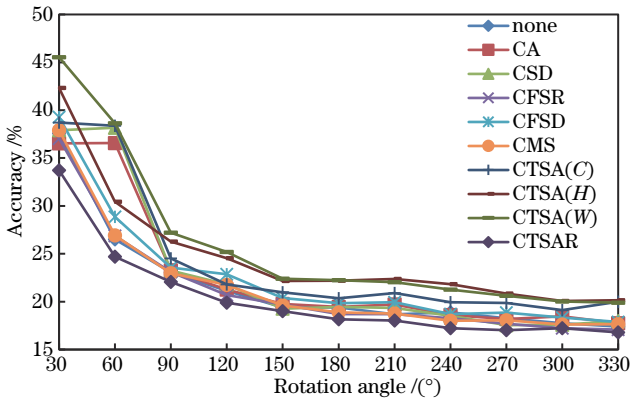

图 9　不同方法对旋转图像的分类准确率对比

Fig. 9　Comparison of the classification accuracy of different methods for rotating images

法的分类准确率一直处于最优，表明了 TSA 模块对图片旋转的抗干扰能力较强。而 CTSAR 方法的分类精度一直处于最低，由此可见 TSAR 模块并不适用于随机旋转图像分类。

中心裁剪作为计算机视觉的图像增强方法之一，其目的是从图片的中心开始，裁剪出其四周指定长度和宽度的图片，即获取原图的中心部分。经过中心裁剪的图像特征变得更少，将更不利于图像分类。因此，将图像中心剪裁成不同的尺寸来测试图像的分类准确率，以评估各类方法的鲁棒性。实验结果见表 5、图 10。

表 5　不同方法对中心裁剪图像的分类准确率结果

Table 5　Classification accuracy results of different methods for center cropped images

Method	Center crop size/pixel				
	10	15	20	25	30
None	14.89	20.22	32.68	51.15	61.20
CA	13.20	17.98	28.00	47.41	60.58
CSD	13.79	19.29	30.15	48.85	61.38
CFSR	14.03	19.19	30.23	48.89	61.23
CFSD	13.53	19.03	31.28	51.87	63.26
CMS	15.12	20.12	32.58	51.31	62.16
CTSA(*C*)	13.74	18.93	32.60	**53.56**	**63.62**
CTSA(*H*)	14.70	20.15	31.51	50.56	61.84
CTSA(*W*)	**15.22**	**20.53**	**33.23**	52.12	62.21
CTSAR	14.03	19.19	30.23	48.89	61.23

注：加粗字体表示各列最优结果。

由表 5 和图 10 可知，当中心裁剪图片尺寸较大时，所有的方法都有较好的分类性能。整体来看，CTSA(*W*)方法的分类精度性能始终优于 CNN 方法，而 CA 方法与 CNN 本身相比分类精度较低，表明了直接添加 Attention 模块会影响 CNN 自身的分类精度，不利于经过中心裁剪图像的分类。当裁剪尺寸为 10×10、15×15、20×20 时，CTSA(*W*)的分类精度在所有方法中最高；当裁剪尺寸为 25×25、30×30 时，CTSA(*C*)的分类准确率最高。由此可见，TSA 模块有助于 CNN 在图像进行中心裁剪时获更高的分类精度。

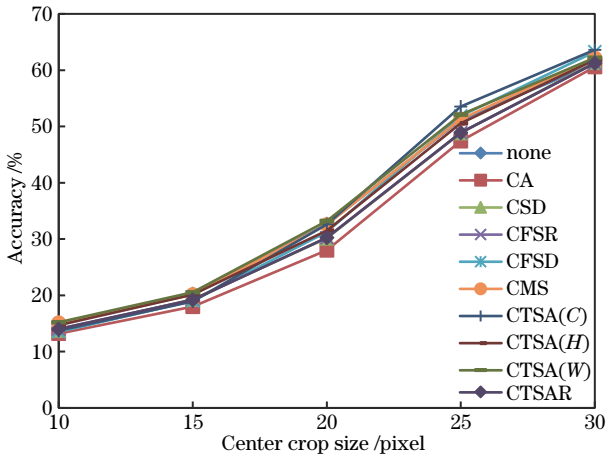

图 10　不同方法对中心裁剪图像的分类准确率对比

Fig. 10　Comparison of the classification accuracy of different methods for the center cropped image

采用上述 10 种方法对 CIFAR-10 数据集原始图像进行分类，其分类对比结果如表 6 所示。可以看出，CTSA(*C*)方法对图像各个类别的准确率、召回率、精确率、宏平均 F_1 值，都优于其他方法，且 CTSA(*C*)、CTSA(*H*)、CTSA(*W*)、CTSAR 整体分类效果要比其他方法好，从而进一步验证了模块的有效性。

表 6　不同方法分类结果

Table 6　Classification results of different methods

Method	Accuracy /%	Recall /%	Precision /%	F_1 /%
None	83.52	62.31	52.30	71.59
CA	82.30	71.28	68.46	82.37
CSD	84.23	80.56	85.77	78.38
CFSR	72.94	73.59	83.21	80.97
CFSD	87.71	81.32	77.32	67.98
CMS	65.23	88.57	60.58	56.35
CTSA(*C*)	**91.65**	**90.20**	**89.91**	**86.72**
CTSA(*H*)	90.50	89.30	86.35	84.23
CTSA(*W*)	89.96	87.23	87.25	83.37
CTSAR	88.67	88.59	89.34	82.56

注：加粗字体表示各列最优结果。

综合以上实验结果表明，本文提出的 TSA 模块嵌入 CNN 后在高斯噪声、随机旋转、中心裁剪干扰下均能较准确关注图像贡献度高的特征信息，且可以较好地完成图像分类，具有良好的鲁棒性。

4.4　模型对比实验及结果分析

目前，主流的残差卷积神经网络模型有 ResNet-18、ResNet-34、ResNet-50 及 ResNet-101[20]。为了选择更适应图像数据复杂度的模型，分别采用经过 ImageNet 预训练的上述残差网络模型对自然图像数据集 CIFAR-10 进行训练，得到的分类结果如表 7 所示。

表 7 残差网络模型分类结果对比

Table 7 Comparison of classification results of residual network models

Model	Accuracy /%
ResNet-18	89.33
ResNet-34	88.76
ResNet-50	86.72
ResNet-101	**93.12**

注：加粗字体表示最优结果。

表 8 不同RTSA Net模型分类精度对比

Table 8 Comparison of classification accuracy of different RTSA Net models

Model	Accuracy /%
RTSA Net-18	90.30
RTSA Net-34	91.23
RTSA Net-50	92.98
RTSA Net-101	**96.12**

注：加粗字体表示最优结果。

由表 7 可知，ResNet-101 在 CIFAR-10 数据集上的分类准确率最高，因此选择 ResNet-101 作为本文的骨干网络模型。

综合上文模块鲁棒性实验结果，采用 TSA(*C*)模块分别嵌入上述残差网络模型中，得到 RTSA Net-18、RTSA Net-34、RTSA Net-50 及 RTSA Net-101 等网络模型，并分别对 CIFAR-10 图像进行分类实验。其中，所有实验都在相同的实验环境和实验参数中进行，分类结果如表 8 所示。

由表 8 可知，嵌入了 TSA 模块的 ResNet 模型分类精度整体要比模型本身高，分类准确率分别提高了 0.97、2.47、6.26、3.00 个百分点，其中 RTSA Net-101 网络模型的分类效果最优。由此可见，本文提出的 TSA 模块可以为图像选取并利用更加重要的特征，解决了原模型提取的特征分辨性弱的问题，具有较好的分类能力。

为进一步验证模型的图像分类效果，将本文提出的 RTSA Net-101 网络模型分别在 CIFAR-10、CIFAR-100、SVHN 数据集上与其他 7 种先进图像分类模型 ANODE[25]、Sign-symmetry[26]、Improved GAN[27]、CLS-GAN[28]、NNCLR[29]、CCT-6/3x1[30]、ResNet56 with reSGHMC[31]进行分类准确率和图片平均测试运行时间对比。在实验过程中，均采用原文献提供的模型或源代码对图像进行测试，实验对比结果如表 9 所示。

表 9 与其他先进方法分类准确率和平均测试运行时间对比

Table 9 Comparison with other advanced methods for classification accuracy and average test running time

Model	CIFAR-10		CIFAR-100		SVHN	
	Accuracy /%	running time /s	Accuracy /%	running time /s	Accuracy /%	running time /s
ANODE[25]	60.60	0.0378	—	—	83.50	0.0308
Sign-symmetry[26]	80.98	0.0341	52.25	0.0301	89.84	0.0317
Improved GAN[27]	88.17	0.0291	—	—	91.89	0.0269
CLS-GAN[28]	91.70	0.0431	—	—	94.02	0.0277
NNCLR[29]	93.70	0.0289	79.00	0.0289	—	—
CCT-6/3x1[30]	95.29	0.0321	77.31	0.0271	—	—
ResNet56 with reSGHMC[31]	96.10	0.0278	80.14	0.0278	—	—
RTSA Net-101	**96.12**	**0.0258**	**81.60**	**0.0260**	**96.67**	**0.0262**

注：加粗字体表示各列最优结果，“—”表示未知结果。

通过对比表 9 中所有模型的分类准确度和平均测试运行时间可知，RTSA Net-101 网络模型在不同的分类图像数据集上均优于其他 7 种先进图像分类模型。对于 CIFER-10 数据集而言，RTSA Net-101 模型的分类精度较其他模型均有不同程度的提高，图片平均测试运行时间也最短。对于 CIFER-100 数据集而言，其样本量较小，且图像种类较多，图像分类任务具有一定的难度，因此可以看出分类准确率整体偏低；其中，RTSA Net-101 与 Sign-symmetry、NNCLR、CCT-6/3x1、ResNet56 with reSGHMC 方法相比，准确率分别提高了 29.35、2.60、4.29、1.46 个百分点，平均运行时间分别缩短了 0.0041 s、0.0029 s、0.0011 s、0.0018 s。对于 SVHN 数据集而言，由于该数据集相对于其他两个数据集样本量要大一些，且种类较少。因此，几种模型的整体分类精度都较高，RTSA Net-101 分别比 ANODE、Sign-symmetry、Improved GAN、CLS-GAN 模型准确率提高了 13.17、6.83、4.78、2.65 个百分点，平均运行时间分别缩短了 0.0046 s、0.0055 s、0.0007 s、0.0015 s。实验结果证实了本文所提的 RTSA Net-101 图像分类模型在图像分类方面具有一定的创新性、实用性和高效性。

5 结 论

为了解决卷积神经网络处理图像分类任务时提取特征不充分以及提取到的特征不区分贡献度的问题，提出了 RTSA Net-101 图像分类模型。该模型在原

ResNet-101网络模型的残差结构中嵌入TSA模块，避免了信息在传递过程中的丢失与损耗，使模型在充分提取图像特征信息的同时，能够聚焦贡献度高的特征，解决了原模型提取的特征分辨性弱的问题，提高了模型的分类性能。同时，对TSA模块进行了鲁棒性测试，验证了模块的有效性，在3个自然图像数据集上进行了对比实验，RTSA Net-101图像分类模型均获得了较其他7种先进模型更具有竞争力的分类准确率和图像平均测试运行时间，并能够有效增强网络的特征学习能力，具有一定的创新性、高效性。在未来研究工作中，将优化模块结构，进一步提升模型分类性能，并丰富任务类型。

参 考 文 献

[1] 张祥东，王腾军，朱劭俊，等. 基于扩张卷积注意力神经网络的高光谱图像分类[J]. 光学学报，2021，41(3)：0310001.
Zhang X D，Wang T J，Zhu S J，et al. Hyperspectral image classification based on dilated convolutional attention neural network[J]. Acta Optica Sinica，2021，41(3)：0310001.

[2] 张永鹏，张春梅，白静. 基于DenseNet-Attention模型的高光谱图像分类[J]. 图学学报，2020，41(6)：897-904.
Zhang Y P，Zhang C M，Bai J. DenseNet-attention for hyperspectral remote sensing image classification[J]. Journal of Graphics，2020，41(6)：897-904.

[3] Wang X L，Girshick R，Gupta A，et al. Non-local neural networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition，June 18-23，2018，Salt Lake City，UT，USA. New York：IEEE Press，2018：7794-7803.

[4] 郭玉荣，张珂，王新胜，等. 端到端双通道特征重标定DenseNet图像分类[J]. 中国图象图形学报，2020，25(3)：486-497.
Guo Y R，Zhang K，Wang X S，et al. Image classification method based on end-to-end dual feature reweight DenseNet[J]. Journal of Image and Graphics，2020，25(3)：486-497.

[5] Roy A G，Navab N，Wachinger C. Concurrent spatial and channel ‘Squeeze & Excitation’ in fully convolutional networks[M]//Frangi A F，Schnabel J A，Davatzikos C，et al. Medical image computing and computer assisted intervention-MICCAI 2018. Lecture notes in computer science. Cham：Springer，2018，11070：421-429.

[6] 黄盛，李菲菲，陈虬. 基于改进深度残差网络的计算断层扫描图像分类算法[J]. 光学学报，2020，40(3)：0310002.
Huang S，Li F F，Chen Q. Computed tomography image classification algorithm based on improved deep residual network[J]. Acta Optica Sinica，2020，40(3)：0310002.

[7] Woo S，Park J，Lee J Y，et al. CBAM：convolutional block attention module[M]//Ferrari V，Hebert M，Sminchisescu C，et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham：Springer，2018，11211：3-19.

[8] 舒意恒. 告别自注意力，谷歌为Transformer打造新内核Synthesizer[EB/OL]. [2020-06-09]. https://zhuanlan.zhihu.com/p/148900206.
Shu Y H. Saying goodbye to self attention，Google builds a new kernel Synthesizer for transformer[EB/OL]. [2020-06-09]. https://zhuanlan.zhihu.com/p/148900206.

[9] Tay Y，Bahri D，Metzler D，et al. Synthesizer：rethinking self-attention in transformer models[EB/OL]. (2020-05-02)[2021-02-03]. https://arxiv.org/abs/2005.00743.

[10] He K M，Zhang X Y，Ren S Q，et al. Deep residual learning for image recognition. [C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas，Nevada，USA：IEEE，2016. 770-778.

[11] 陈琳琳，朱惠娟，朱俊，等. 基于卷积神经网络的多尺度注意力图像分类模型[J]. 南京理工大学学报，2020，44(6)：669-675.
Chen L L，Zhu H J，Zhu J，et al. Multiscale attention model for image classification based on convolutional neural network[J]. Journal of Nanjing University of Science and Technology，2020，44(6)：669-675.

[12] 于福升，余江，鲁远甫，等. 基于残差网络的虹膜图像性别分类[J]. 激光与光电子学进展，2021，58(16)：1610022.
Yu F S，Yu J，Lu Y F，et al. Gender classification of iris image based on residual network[J]. Laser & Optoelectronics Progress，2021，58(16)：1610022.

[13] Vaswani A，Shazeer N，Parmar N，et al. Attention is all you need[EB/OL]. (2017-06-12) [2021-02-03]. https://arxiv.org/abs/1706.03762.

[14] 张忠林，李林川，朱向其，等. ON-LSTM和自注意力机制的方面情感分析[J]. 小型微型计算机系统，2020，41(9)：1839-1844.
Zhang Z L，Li L C，Zhu X Q，et al. Aspect sentiment analysis combining ON-LSTM and self-attention mechanism[J]. Journal of Chinese Computer Systems，2020，41(9)：1839-1844.

[15] 王茜. 基于注意力机制和复曲波融合网络的多光谱图像融合及地物分类[D]. 西安：西安电子科技大学，2019.
Wang Q. Multispectral image fusion and classification based on attention mechanism and complex curvelet fusion network[D]. Xi’an：Xidian University，2019.

[16] Cordonnier J，Loukas A，Jaggi M. On the relationship between self-attention and convolutional layers[EB/OL]. (2019-11-08)[2021-02-03]. https://arxiv.org/abs/1911.03584.

[17] 朱虹. 基于神经网络和自注意力机制的文本表示与分类研究[D]. 重庆：西南大学，2020.
Zhu H. Research on text representation and classification based on neural networks and self-attention mechanism [D]. Chongqing：Southwest University，2020.

[18] 汪鹏，刘瑞，辛雪静，等. 基于残差网络的光学遥感图像场景分类算法[J]. 激光与光电子学进展，2021，58(2)：0210001.
Wang P，Liu R，Xin X J，et al. Scene classification of optical remote sensing images based on residual networks

[J]. Laser & Optoelectronics Progress, 2021, 58(2): 0210001.

[19] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[20] 乔思波, 庞善臣, 王敏, 等. 基于残差混合注意力机制的脑部CT图像分类卷积神经网络模型[J]. 电子学报, 2021, 49(5): 984-991.
Qiao S B, Pang S C, Wang M, et al. A convolutional neural network for brain CT image classification based on residual hybrid attention mechanism[J]. Acta Electronica Sinica, 2021, 49(5): 984-991.

[21] de Lathauwer L, de Moor B, Vandewalle J. A multilinear singular value decomposition[J]. SIAM Journal on Matrix Analysis and Applications, 2000, 21(4): 1253-1278.

[22] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images[M]//Handbook of systemic autoimmune diseases. Amsterdam: Elsevier, 2009.

[23] 付晓, 沈远彤, 李宏伟, 等. 基于半监督编码生成对抗网络的图像分类模型[J]. 自动化学报, 2020, 46(3): 531-539.
Fu X, Shen Y T, Li H W, et al. A semi-supervised encoder generative adversarial networks model for image classification[J]. Acta Automatica Sinica, 2020, 46(3): 531-539.

[24] 常东良, 尹军辉, 谢吉洋, 等. 面向图像分类的基于注意力引导的Dropout[J]. 图学学报, 2021, 42(1): 32-36.
Chang D L, Yin J H, Xie J Y, et al. Attention-guided Dropout for image classification[J]. Journal of Graphics, 2021, 42(1): 32-36.

[25] Dupont E, Doucet A, Teh Y W. Augmented neural ODEs[EB/OL]. (2019-04-02) [2021-02-03]. https://arxiv.org/abs/1904.01681.

[26] Liao Q L, Leibo J Z, Poggio T. How important is weight symmetry in backpropagation? [EB/OL]. (2015-10-17)[2021-03-02]. https://arxiv.org/abs/1510.05067.

[27] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training GANs[EB/OL]. (2016-06-10) [2021-02-05]. https://arxiv.org/abs/1606.03498.

[28] Qi G J. Loss-sensitive generative adversarial networks on Lipschitz densities[J]. International Journal of Computer Vision, 2020, 128(5): 1118-1140.

[29] Dwibedi D, Aytar Y, Tompson J, et al. With a little help from my friends: nearest-neighbor contrastive learning of visual representations[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2021: 9568-9577.

[30] Hassani A, Walton S, Shah N, et al. Escaping the big data paradigm with compact transformers[EB/OL]. (2021-04-12)[2021-02-03]. https://arxiv.org/abs/2104.05704.

[31] Shen Z Y, He L S, Lin Z C, et al. PDO-eConvs: partial differential operator based equivariant convolutions[EB/OL]. (2020-07-20) [2021-02-05]. https://arxiv. org/abs/2007.10408.