文章编号： 1671-7449(2024)04-0413-07

基于Swin Transformer和双层路由注意力的多标签图像分类算法

张　震，王　贺*，宋宏旭

（山西大学 物理电子工程学院， 山西 太原 030006）

摘　要： 图像分类是图像处理中一项基础而又重要的工作。单一标签的图像分类已经无法满足人们的需求，研究者们开始关注于多标签图像分类。本文提出了一种Swin Transformer进行特征提取，由双层路由注意力模块进行特征处理的多标签图像分类框架。Swin Transformer通过分层结构提取多尺度信息，在多目标和更细粒度的图像识别方面优于Vision Transformer；双层路由注意力模块能够实现更灵活的计算分配和内容感知，可根据输入图像的特征自适应地调整注意力权重，灵活地控制注意力的强度和范围。模型在COCO数据集上平均精度均值为87.3，在VOC2007数据集上平均精度均值为96.7，一定程度上提高了多标签图像分类的精度。

关键词： 深度学习；多标签分类；Swin Transformer；双层路由注意力模块

中图分类号： TP391　　**文献标识码：** A　　**doi:** 10.3969/j.issn.1671-7449.2024053

引用格式： 张震，王贺，宋宏旭．基于Swin Transformer和双层路由注意力的多标签图像分类算法[J]．测试技术学报，2024，38(4)：413-419.

ZHANG Zhen，WANG He，SONG Hongxu. Multi-label image classification algorithm based on Transformer [J]. Journal of Test and Measurement Technology，2024，38(4)：413-419.

Multi-Label Image Classification Algorithm Based on Transformer

ZHANG Zhen，WANG He*，SONG Hongxu

（College of Physics and Electronic Engineering，Shanxi University，Taiyuan 030006，China）

Abstract： Image classification is a basic and important direction in image processing. Since there is not only a single label value on an image，the current image classification can no longer meet people′s needs，and multi-label image classification came into being. This paper proposes a multi-label image classification framework using Swin Transformer for feature extraction and a two-layer routing attention module for feature processing. Swin Transformer extracts multi-scale information through a hierarchical structure，and is superior to Vision Transformer in terms of multi-target and finer-grained image recognition. The dual-layer routing attention module enables more flexible computation allocation and content awareness. The dynamic attention mechanism adaptively adjusts the attention weight according to the characteristics of the input image，so that different positions or features can be given different levels of attention，and the intensity and range of attention can be flexibly controlled by adjusting the dynamic attention. The average precision of the model on the COCO dataset is 87.3，and the average precision on the VOC2007 dataset is

收稿日期： 2023-08-23

作者简介： 张　震(1995—)，男，硕士生，主要从事深度学习和图像分类的研究。E-mail：zz13754860102@163.com。

***通信作者：** 王　贺(1983—)，女，博士，讲师，主要从事机器学习和信息处理方向的研究。E-mail：wanghe@sxu.edu.cn。

96.7, which improves the accuracy of multi-label image classification to a certain extent.
Key words: deep learning; multi-label image classification; swin transformer; bi-level routing attention

随着互联网和数码设备的普及，人们每天都需要处理大量图像，因此图像分类的重要性逐渐凸显。传统的图像分类方法为每张图片分配一个标签，但这种方法不够实际，因为它忽略了图像中的其他对象及其属性。因此，人们逐渐开始研究多标签图像分类。与传统的图像分类相比，多标签图像分类为每张图片分配多个标签，能够更全面和准确地描述一张图像的内容和属性。

目前，研究标签之间的相关性已成为多标签图像分类的主要趋势。Wang 等[1]提出著名的CNN-RNN框架，利用递归神经网络(RNN)结合卷积神经网络(CNN)，学习联合图像标签嵌入，以表征语义标签依赖性和图像标签相关性；Chen等[2]提出的ML-GCN，利用卷积神经网络提取特征，并将特征输入到图卷积网络(GCN)[3]中，学习标签相关性，取得了成功；Li 等[4]提出A-GCN，引入了一个即插即用的自适应标签图(LG)模块，结合GCN网络，自动学习标签相关性来提高分类结果。

除了对标签相关性的探索，在多标签图像分类方向还出现了一些不同的解决方法。2020年，Wu等[5]对多标签图像分类损失进行改进，提出了Distribution-Balanced Loss，通过重新平衡权重，减轻负标签的抑制，解决标签类别不平衡问题；Ben-Baruch 等[6]提出ASL，动态地降低权重，同时丢弃可能被错误标记的样本，平衡不同样本的概率；Zhu等[7]提出CSRA，这是一种简单而有效的方法，通过最大池化得到每个类别特定的特征，然后与平均池化特征相结合，具有直观的解释性和可视化效果。

近几年，随着Transformer[8]取得巨大成功，人们开始将Transformer引入多标签图像分类领域。Zhao等[9]提出TDRG，基于Transformer探索结构关系和语义关系，将学习到的结构关系合并到语义图中，构建了联合关系图。Liu等[10]提出Query2Label，利用Transformer解码器查询类标签的存在；Ridnik等[11]提出基于Transformer的分类头ML-Decoder，使用Transformer中的decoder部分进行分类，并将自注意力层部分去除，提升了Transformer的检测速度。

基于以上对多标签图像分类的研究，本文设计了一个使用Swin Transformer[12]进行特征提取，使用双层路由注意力模块(BRA)[13]进行特征处理的多标签图像分类网络结构。首先，Swin Transformer引入了一种具有层级的特征表达方式，通过将图像分成多个窗口，并在每个窗口内进行自注意力计算，可以更好地捕获图像中的局部和全局信息；其次，双层路由注意力作为一种动态稀疏注意力，能够实现更灵活的计算分配和内容感知，节省内存并且可以取得很好的效果。与其他多标签图像分类框架相比，本文提出的多标签图像分类框架在两个经典数据集上精度都有明显提升。

1 系统框架

如图1为系统整体网络框图。对于每个图像$I\in R^{H\times W\times 3}$，通过图块分割层(Patch Partition)将其划分为一组互不重叠的图像块，分别通过Stage1，Stage2、Stage3、Stage4，得到特征图；再将特征图送进双层路由注意力模块中进行特征处理，该模块可以自适应地调整注意力权重，获取不同位置的重要程度，灵活控制注意力的强度和范围；最后通过一个自制的分类器得到分类结果。

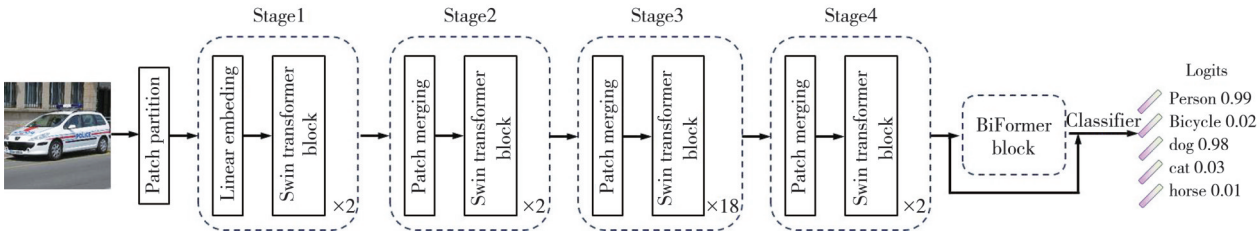

图1 系统整体网络框架

Fig. 1 Structure diagram of network

1.1　Swin Transformer网络

Swin Transformer通过将图像划分为不重叠的窗口来学习图像特征，降低了计算复杂度。另外，Swin Transformer通过分层结构提取多尺度信息，在多目标和更细粒度的图像识别方面优于Vision Transformer[14]。

如图 2 所示，Swim Transformer通过下采样把图像分为多个小尺寸的图像块，在图像块组成的窗口内进行自注意力计算，然后再将相邻补丁进行合并，输出的特征图的高和宽会减半，并且深度翻倍，这与卷积神经网络的结构类似，其中通过Stage1，Stage2，Stage3，Stage4输出的特征图分别为 $\frac{H}{4}\times\frac{W}{4}\times C$，$\frac{H}{8}\times\frac{W}{8}\times 2C$，$\frac{H}{16}\times\frac{W}{16}\times 4C$，$\frac{H}{32}\times\frac{W}{32}\times 8C$，$C$表示通道数。

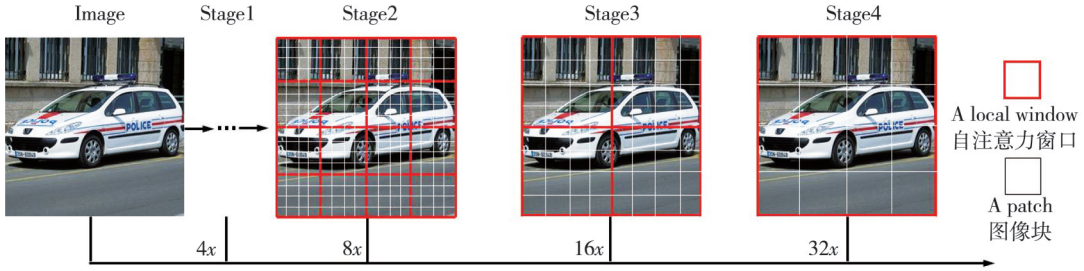

图 2　Swim Transformer 的分层特征图

Fig. 2　Hierarchical feature map of Swing Transformer

Swin Transformer的优点是自注意力的计算仅在7×7的小窗口内进行，从而在计算复杂度上优于Vision Transformer。Swin Transformer基于移位窗口构造并用基于窗口的多头自注意力机制（Window-based Multi-Head Self-Attention，W-MSA）替代了Vision Transformer中的多头自注意力模块。如图 3 所示，Swin Transformer块依次经过归一化层（Layer Norm）、W-MSA 和多层感知机（Multi Layer Perceptron，MLP），又继续通过Layer Norm层、移位窗口的多头自注意力机制（Shifted Window-based Multi-Head Self-Attention，SW-MSA）和MLP层，通过移动窗口实现了原本互不交流窗口的信息交流。

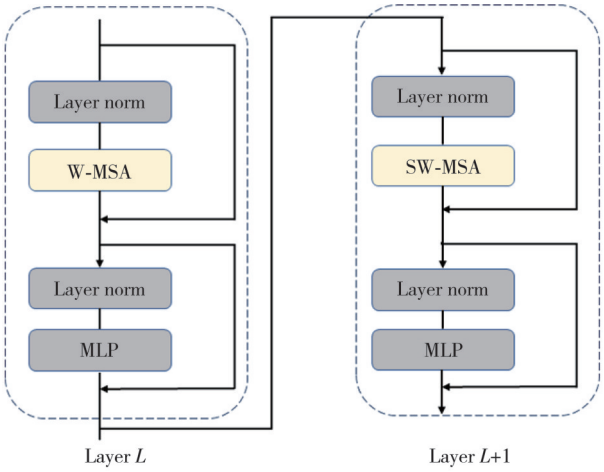

图 3　Swim Transformer块

Fig. 3　Swim Transformer block

计算过程及公式为

$$\hat{z}^{l}=\text{W-MSA}(LN(z^{l-1}))+z^{l-1},\tag{1}$$

$$z^{l}=\text{MLP}(LN(\hat{z}^{l}))+\hat{z}^{l},\tag{2}$$

$$\hat{z}^{l+1}=\text{SW-MSA}(LN(z^{l}))+z^{l},\tag{3}$$

$$z^{l+1}=\text{MLP}(LN(\hat{z}^{l+1}))+\hat{z}^{l+1},\tag{4}$$

式中：z^{l}和z^{l+1}分别为L和$L+1$层的输出特征；$\hat{z}^{l}$和$\hat{z}^{l+1}$分别为特征经过W-MSA和SW-MSA得到的输出特征；W-MSA表示基于窗口的多头注意力机制；SW-MSA表示位移窗口的多头自注意力机制；LN为Layer Norm归一化层；MLP表示多层感知器。

1.2　双层路由注意力模块（BRA）

引入双层路由注意力（BRA）模块，是一种新的动态稀疏注意力，能够实现更灵活的计算分配和内容感知。BRA是Transformer模型的变体，在原始Transformer模型中引入了动态注意力机制。动态注意力机制根据输入图像的特征自适应地调整注意力权重，灵活控制注意力的强度和范围。

如图 4 所示，BRA模块分为3个步骤：当输入一个特征图时，将其划分为几个区域，通过线性映射获得查询（Q）、关键字（K）和值（V）；其次，使用邻接矩阵来构建有向图，以找到不同键值对对应的参与关系，可以理解为每个给定区域应该参与的区域；最后，通过从一个区域到另一

个区域的索引矩阵，可以捕获到补丁与补丁之间的关系。

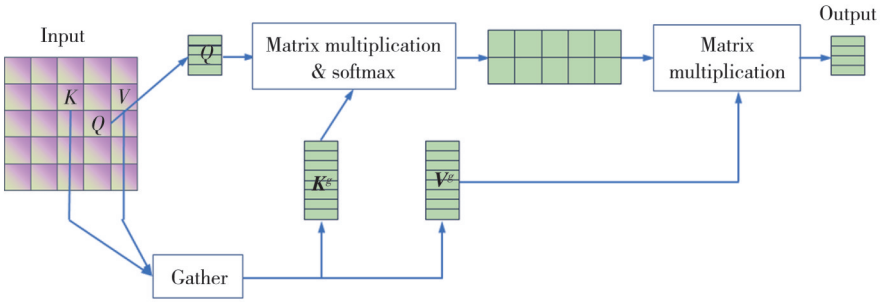

图 4 BRA结构示意图

Fig. 4 Schematic diagram of BRA structure

从图 4 中可以看出，BRA模块利用稀疏性运算跳过最不相关区域的计算，从而节省了参数和计算成本。

1.3 自制分类器（Classifier）

构造了一个简单的分类器，用于多标签分类，并在后续消融实验中通过该分类器对Vision Transformer作为基准进行分类。具体来说，给定一个特征$X\in R^{H\times W\times C}$，首先对其进行降维，使得通道数C降为类别数c，并且使用平均池化和最大池化将$H\times W$大小的特征处理为1×1，最后得到分类结果，如图 5 所示。

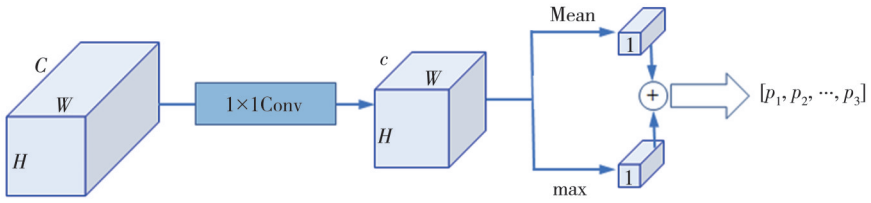

图 5 分类器结构图

Fig. 5 Classifier structure diagram

$$[p_1, p_2, \cdots, p_c]=\mathrm{Max}P(C_{1\times 1}(X))+\mathrm{Avg}P(C_{1\times 1}(X)), \tag{5}$$

式中：$\mathrm{Max}P(\cdot)$为最大池化；$\mathrm{Avg}P(\cdot)$为平均池化；$C_{1\times 1}(\cdot)$为1×1卷积；$[p_1, p_2, \cdots, p_c]$为经过分类器得到各个类别的概率值。

2 实验与分析

2.1 数据集及实验环境介绍

选择了两个公开的数据集进行试验：MS COCO2014和Pascal VOC 2007。前者包含了80个常见的对象类别标签，由82 783个训练图像和40 504个验证图像构成；后者经常被用于图像分类和目标检测任务，由5 011张图像的训练集和4 952张图像的测试集构成，包含了20个常见的类别。

在ubuntu20.04操作系统上使用单内存为24 G的NVIDIA GeForce RTX 3090上完成所有的实验。在训练阶段，将图片分辨率随机裁剪并调整为448×448，共经过15个epoch的训练，使用SGD算法作为优化器，批量大小设置为16，动量设置为0.9，权重衰减为0.000 1，初始学习率为0.01，在第30和40个epoch时将学习率设为1/10。实验在ImageNet上进行预训练，且去掉Swin Transformer中分类部分，只保留主干部分用来提取特征。

2.2 评价指标

以平均精度均值（P_{mA}）作为主要的评价指标，以总体精密度（P_{O}）、召回率（R_{O}）、F1-measure（F_{O1}）和每类精密度（P_{C}）、召回率（R_{C}）、F1-measure（F_{C1}）为辅助的评价指标进行比较，定义如下

$$P_{\mathrm{mA}}=\frac{\sum_{i=1}^{N}P_{A_i}}{N},\quad \left(P=\frac{P_{\mathrm{T}}}{P_{\mathrm{T}}+P_{\mathrm{F}}},\ P_{\mathrm{A}}=\frac{\sum P}{k}\right), \tag{6}$$

式中：P_{T}全称为True Positive，指的是被判定为正样本，事实上也是正样本；P_{F}全称为False Positive，指的是被判定为正样本，但事实上是负样本；P为准确率；k表示数据集中含有该类别的图片数量；N代表总类别个数；P_{A}为单标签精度值，

P_{mA}代表所有目标类别精度的平均值。

$$P_O=\frac{\sum_i N_i^c}{\sum_i N_i^p},\ P_C=\frac{1}{C}\sum_i\frac{N_i^c}{N_i^p},\tag{7}$$

$$R_O=\frac{\sum_i N_i^c}{\sum_i N_i^g},\ R_C=\frac{1}{C}\sum_i\frac{N_i^c}{N_i^g},\tag{8}$$

$$F_{O1}=\frac{2\times P_O\times R_O}{P_O+R_O},\ F_{C1}=\frac{2\times P_C\times R_C}{P_C+R_C},\tag{9}$$

式中：N_i^c是第i个类别中被正确预测的图像数量，N_i^p是第i个类别中被预测的图像数量，N_i^g是第i个类别中真实图像的数量。在这些指标中，P_{mA}，F_{C1}，F_{O1}是最重要的指标，可以提供更全面的评价。

2.3　实验结果分析

2.3.1　在MS-COCO14上的表现

在MS-COCO14上进行比较的结果如表 1 所示。可以看出，本文提出的框架在平均精度均值上优于其他多标签图像分类框架。另外两个重要指标CF1和OF1也均高于其他多标签分类框架，整体结果在所有模型中最好，说明了本文框架的有效性。

表 1　各算法在COCO 14数据集上的指标

Tab. 1　Indicators of each algorithm on the COCO 14 dataset　　单位：%

模型	All							TOP3					
	P_{mA}	P_C	R_C	F_{C1}	P_O	R_O	F_{O1}	P_C	R_C	F_{C1}	P_O	R_O	F_{O1}
SRN	77.1	81.6	65.4	71.2	82.7	69.9	75.8	85.2	58.8	67.4	87.4	62.5	72.9
ResNet-101	78.3	80.2	66.7	72.8	83.9	70.8	76.8	84.1	59.4	69.7	89.1	62.8	73.9
ML-GCN	83.0	85.1	72.0	78.0	85.8	75.4	80.3	87.2	64.6	74.2	89.1	66.7	76.3
MCAR	83.8	85.0	72.1	78.0	88.0	73.9	80.3	88.1	65.5	75.1	91.0	66.3	76.7
SSGRL	83.8	89.9	68.5	76.8	91.3	70.8	79.7	91.9	62.5	72.7	93.8	64.1	76.2
STMG	84.3	85.8	72.7	78.7	86.7	76.8	81.5	91.6	51.9	66.2	92.7	61.8	74.1
C-Trans	85.1	86.3	74.3	79.9	87.7	76.5	81.7	90.1	65.7	76.0	92.1	71.4	77.6
ADD-GCN	85.2	84.7	75.9	80.1	84.9	79.4	82.0	88.8	66.2	75.8	90.3	68.5	77.9
CSRA	86.9	89.1	74.2	81.0	89.6	77.1	82.9	92.5	65.8	76.9	93.4	68.1	78.8
本文改进模型	87.3	87.2	77.0	81.8	88.0	79.0	83.2	91.0	67.5	77.5	92.5	68.8	78.9

2.3.2　在Pascal VOC 2007上的表现

在Pascal VOC 2007数据集上进行实验，并列出了20个类别各自的精度以及最后的平均精度均值，如表 2 所示。

表 2　各算法在VOC 2007数据集上的指标

Tab. 2　Indicators of each algorithm on the VOC 2007 dataset

类别	精度/%						
	VGG+SVM	FeV+LV	HCP	RDAL	MCAR	CCD-R101	本文改进模型
aero	98.9	97.9	98.6	98.6	99.7	99.9	99.9
bike	95.0	97.0	97.1	97.4	99.0	98.2	99.5
bird	96.8	96.6	98.0	96.3	98.5	98.4	99.4
boat	95.4	94.6	95.6	96.2	98.2	98.9	99.0
bottle	69.7	73.6	75.3	75.2	85.4	84.9	87.5
bus	90.4	93.9	94.7	92.4	96.9	97.7	97.9
car	93.5	96.5	95.8	96.5	97.4	97.8	98.7
cat	96.0	95.5	97.3	97.1	98.9	99.0	99.1
chair	74.2	73.7	73.1	76.5	83.7	86.4	85.4
cow	86.6	90.3	90.2	92.0	95.5	98.8	98.9
table	87.8	82.8	80.0	87.7	88.8	90.2	92.9
dog	96.0	95.4	97.3	96.8	99.1	99.2	99.3
horse	96.3	97.7	96.1	97.5	98.2	98.9	99.1
mbike	93.1	95.9	94.9	93.8	95.1	97.8	98.8
person	97.2	98.6	96.3	98.5	99.1	98.8	99.3
plant	70.0	77.6	78.3	81.6	84.8	87.3	89.3
sheep	92.1	88.7	94.7	93.7	97.1	99.4	99.6
sofa	80.3	78.0	76.2	82.8	87.8	88.8	89.5
train	98.1	98.3	97.9	98.6	98.3	99.7	99.7
tv	87.0	89.0	91.5	89.3	94.8	96.6	97.1
	P_{mA}/%						
	VGG+SVM	FeV+LV	HCP	RDAL	MCAR	CCD-R101	本文改进模型
	89.7	90.6	90.9	91.9	94.8	95.8	96.7

从表 2 可以看出，所提出的框架在 19 个类别上都有提升，而且平均精度均值也要优于这些方法，进一步验证了提出方法的有效性。但是由于椅子的样本太少，而且它经常出现在被遮挡的情况下，样本质量较差，所以导致椅子的精度低于 CCD-R101 模型中的椅子的精度。

2.4 消融实验

为了验证本文算法中各部分的重要性，对每个模块进行消融实验。分别为：1）使用 Vision Transformer 进行多标签图像分类。2）使用 Swin Transformer 进行多标签图像分类。3）使用 Swin Transformer 和双层路由注意力模块进行多标签图像分类。在 COCO 14 和 VOC 2007 数据集上分别进行实验。为了避免其他因素的影响，消融实验中其他部分框架均相同。表 3 给出了上述 3 种方法的分类精度。

从表 3 可以看出，Swin Transformer 作为骨干网络，与 Vision Transformer 相比，在两个数据集上评价指标都有了明显的提升，这主要是因为 Swin Transformer 通过分层结构提取特征，可以更好地捕获图像中的局部和全局信息。在时间上 Swin Transformer 也比 Vision Transformer 快了很多，在加入 BRA 后，评价指标又有了继续的提升，这是因为 BRA 能够实现更灵活的计算分配和内容感知，而且通过调整动态注意力来灵活地控制注意力的强度和范围。综合两点可以让多标签图像获取到更高的分类精度。

表 3 在 COCO 14 和 VOC 2007 数据集上进行消融实验的结果

Tab. 3 Results of ablation experiments on COCO 14 and VOC 2007 datasets 单位：%

方法	MS COCO14			VOC2007		
	P_{mA}	F_{C1}	F_{O1}	P_{mA}	F_{C1}	F_{O1}
Vision Transformer	80.4	74.5	78.6	92.5	83.9	86.4
Swin Transformer	85.5	79.8	82.1	95.4	88.2	89.7
Swin Transformer+BRA	87.3	81.8	83.2	96.7	90.6	91.4

在表 4 中对比了各类方法在 COCO14 数据集上的参数量和运行时间。从表 4 中可以看出在分辨率提高的情况，Swin Transformer 作为骨干网络，与 Vision Transformer 相比，参数量只提升了 2 M，但在运行时间上却减少了 3.5 h，这说明了 Swin Transformer 在计算效率上要优于 Vision Transformer。在 Swin Transformer 的基础上加入 BRA 模块，参数量和运行时间变化较小，但是可以对平均精度均值起到很好的提高作用，说明了其有效性。

表 4 各类方法在在 COCO 14 数据集上的参数量和运行时间

Tab. 4 The number of parameters and running time of various methods on the COCO 14

方法	分辨率	参数量/M	运行时间/h	P_{mA}/%
Vision Transformer	224×224	88	22	80.4
Swin Transformer	448×448	90	15.5	85.5
Swin Transformer+BRA	448×448	91.9	17	87.3

2.5 可视化

如图 6 所示，在模型的热力图中，颜色强度代表了它们在模型中的相对重要性，颜色越深代表越重要。

第 1 列为原图，从图中可以发现，它们都包含多个标签值，所以我们在第 2、第 3、第 4 列图中，在给定标签值的情况下，得出模型具体关注图像的哪个部分。说明该模型可以为每个类别生成精确的注意力热图，确保了模型关注特定类别最相关的信息，引导模型关注所在的区域进行精确分类。

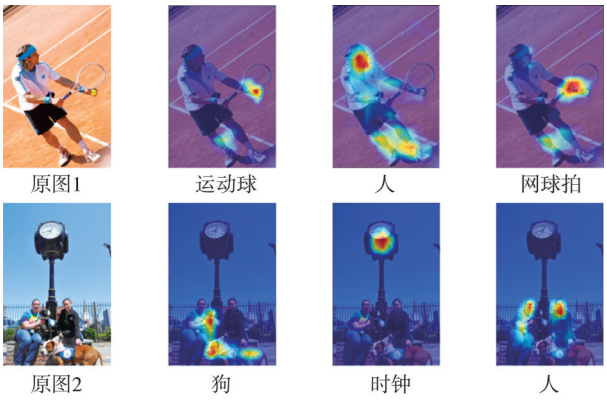

图 6 模型可视化

Fig. 6 Model visualization

3　结　论

提出一种基于Swin Transformer和双层路由注意力的多标签图像分类算法。该模型使用Swin Transformer提取特征，使用双层路由注意力（BRA）模块处理特征。文中对Swin Transformer和Vision Transformer网络进行比较，发现Swin Transformer在处理多标签图像分类中具有显著优势。使用双层路由注意力模块处理特征，既考虑到了使用Transformer带来的显存问题，又用该模块提升了多标签分类的精度。同时还用分类器对模型加以改进，完善了多标签分类整体框架。但是也可以看到，当样本质量较差和样本较少时，该算法的分类精度会有一定的下降。未来我们将继续完善该算法，通过对Swin Transformer网络的改进和对数据进行图像增强，进一步提高该算法在样本较少的类别上的分类精度。

参考文献：

[1] WANG J, YANG Y, MAO J, et al. CNN-RNN: a unified framework for multi-label image classification [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2285-2294.

[2] CHEN Z M, WEI X S, WANG P, et al. Multi-label image recognition with graph convolutional networks [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 5172-5186.

[3] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks [EB/OL]. http://arxiv.org/abs/1609.02907v4.

[4] LI Q, PENG X, QIAO Y, et al. Learning label correlations for multi-label image recognition with graph networks [J]. Pattern Recognition Letters, 2020, 138: 378-384.

[5] WU T, HUANG Q, LIU Z, et al. Distribution-balanced loss for multi-label classification in long-tailed datasets [C]//European Conference on Computer Vision, 2020: 162-178.

[6] RIDNIK T, BEN-BARUCH E, ZAMIR N, et al. Asymmetric loss for multi-label classification [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 82-91.

[7] ZHU K, WU J. Residual attention: a simple but effective method for multi-label recognition [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 184-193.

[8] WASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]// Conference and Workshop on Neural Information Processing Systems, 2017: 6000-6010.

[9] ZHAO J, YAN K, ZHAO Y, et al. Transformer-based dual relation graph for multi-label image recognition [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 163-172.

[10] LIU S, ZHANG L, YANG X, et al. Query2Label: a simple transformer way to multi-label classification [EB/OL]. http://arxiv.org/abs/2107.10834v1.

[11] RIDNIK T, SHARIR G, BEN-COHEN A, et al. ML-decoder: scalable and versatile classification head [C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023: 32-41.

[12] LIU Z, LIN Y, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 9992-10002.

[13] ZHU L, WANG X, KE Z, et al. BiFormer: vision transformer with bi-level routing attention [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 10323-10333.

[14] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale [EB/OL]. http://arxiv.org/abs/2010.11929v1.