第 58 卷 第 6 期/2021 年 3 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于改进残差网络的中式菜品识别模型

邓志良，李磊*

南京信息工程大学自动化学院，江苏 南京 210044

摘要 针对传统神经网络无法对相似度较高的中式菜品进行有效分类的问题，提出了一种基于改进残差网络的中式菜品识别 RNA-TL（ResNet with Attention and Triplet Loss）模型。该算法先融合多尺度特征以提取深层次图像的语义信息，然后增加一层注意力机制层，给予图像重要部分更多的关注，最后利用三元组损失（Triplet Loss，TL）计算类间相似度并将结果输入到支持向量机（Support Vector Machine，SVM）中进行分类。实验表明，相较于其他主流算法模型，RNA-TL 模型在中式菜品公共数据集上以及课题组采集的数据集上的识别准确率表现出更优越的性能。

关键词 图像处理；中式菜品识别；三元组损失；卷积神经网络；注意力机制

中图分类号 TP391.4；TP183；TN957.52 **文献标志码** A **doi**：10.3788/LOP202158.0610019

Chinese Food Recognition Model Based on Improved Residual Network

Deng Zhiliang, Li Lei*

School of Automation, Nanjing University of Information Science & Technology, Nanjing, Jiangsu 210044, *China*

Abstract In view of the fact that traditional neural networks cannot effectively classify Chinese food with high similarity, a Chinese food recognition model of RNA-TL (ResNet with attention and triplet loss) based on an improved residual network is proposed. The algorithm first fuses the multi-scale features to extract the semantic information of deep-level images, and then adds an attention mechanism layer to give more attention to the important parts of the images. Finally, the similarity among classes is calculated by using triplet-loss, whose result is input into support vector machine (SVM) for classification. The experimental results indicate that the proposed RNA-TL model possesses more superior performances in recognition accuracy on the public dataset of Chinese food and the dataset collected by our project team, compared with the other mainstream algorithm models.

Key words image processing; Chinese food recognition; triplet loss; convolutional neural network; attention mechanism

OCIS codes 100.2960; 100.3008

1 引 言

随着人们对物质生活要求的不断提高，饮食健康问题越来越突出。为了精准记录每餐摄入的营养信息，可记录每餐的菜品图像并进行菜品名称标注以便后续的营养成分分析，从而为健康饮食提供有效的数据支撑[1]。同时，菜品名称标注如果仅依靠人为标注，不仅成本高而且效率低下。因此，如何实现有效的中式菜品图像识别成为研究热点。

近年来，机器学习与深度学习在菜品识别上的应用越来越广泛。对于菜品图像的识别，目前已经出现了一些比较优秀的算法模型。Yang 等[2]使用传统的机器学习方法，将图像的软像素级分割成八种成分类型以计算局部特征之间的成对统计量，并

收稿日期：2020-07-20；**修回日期**：2020-08-29；**录用日期**：2020-09-09

* **E-mail**：466743943@qq.com

且将结果积累在一个多维的直方图中，最后将这个多维直方图作为一个特征向量输入到分类器中。Zheng 等[3]提出了一种基于超像素的 LDC(Linear Distance Coding)底层特征的方法，通过提取基于超像素分割的判别食物信息，对最大类间方差法进行了改进，该方法在 PFID 食物图像数据库上表现出很好的性能并对噪声和遮挡具有良好的鲁棒性。Mezgec 等[4]提出了一种基于 AlexNet 的改进的 NutriNet 深度卷积神经架构，该方法以高像素的图像作为输入，在 AlexNet 的第一个卷积层之前增加了一个额外的卷积层，最终在食物识别中获取了良好的识别准确率。Martinel 等[5]在残差网络(ResNet)的基础上添加一层以捕捉菜品的垂直结构，并且优化了残差模块，然后将新的结构与残差块结合起来以生成分类的得分，相比于传统的机器学习算法，此方法表现出更好的性能，在食物数据集 Food-101 上达到了 90.27%的准确率。Pan 等[6]提出了一种基于图像增强神经网络的食物识别方法，该方法结合了仿射变换图像增强技术和高级特征向量，利用 SMO 分类器进行分类，在中小规模数据集上表现出优异的性能。Ng 等[7]在现有的卷积神经网络上对不同的数据集进行了大量的对比实验，发现深层网络 Xception、Nasnet-Large 的识别准确率较高，同时还发现，为了得到最佳的训练效果，训练集上每个类至少包含 300 幅图像且不改变形状的图像增强技术更有益。

以上方法针对的都是西餐菜品，由于西餐菜品的相似度相比于中式菜品较低，因此识别起来相对简单。中式菜品比较复杂，种类繁多，高相似度给识别带来了一定的困难[8]。为了解决这一问题，本文提出了一种基于改进残差网络的卷积结构 RNA-TL (ResNet with Attention and Triplet-Loss) 模型。先在 ResNet 的基础上，利用不同结构与大小的卷积核(RNA-A、RNA-B 和 RNA-C)替换尺寸为 3 pixel×3 pixel 的卷积核，从而提取了多尺度的深层次特征，并增加了注意力机制分支，有效特征信息的权重得到增大。然后，利用人脸识别中的三元组损失(Triplet Loss，TL)的原理[9]对损失函数进行了改进，加大或缩小了不同类或相同类之间的相似度。最后，基于得到的训练参数，利用支持向量机(Support Vector Machine，SVM)进行菜品的分类。

2　模型结构

本文提出了一种基于改进残差网络的中式菜品识别模型。该模型分为三个部分。第一部分即 RNA((ResNet with Attention))模块，通过融合多尺度特征以提取深层语义。第二部分即注意力机制层。通过添加注意力机制，动态强化或弱化权重，从而获取模型所关注的重要信息。第三部分，使用优化的 TL 函数计算类与类之间相似度，并将训练得到的特征信息输入到 SVM [10]中进行菜品识别，根据准确率确定目标的最终分类。模型结构如图 1 所示。

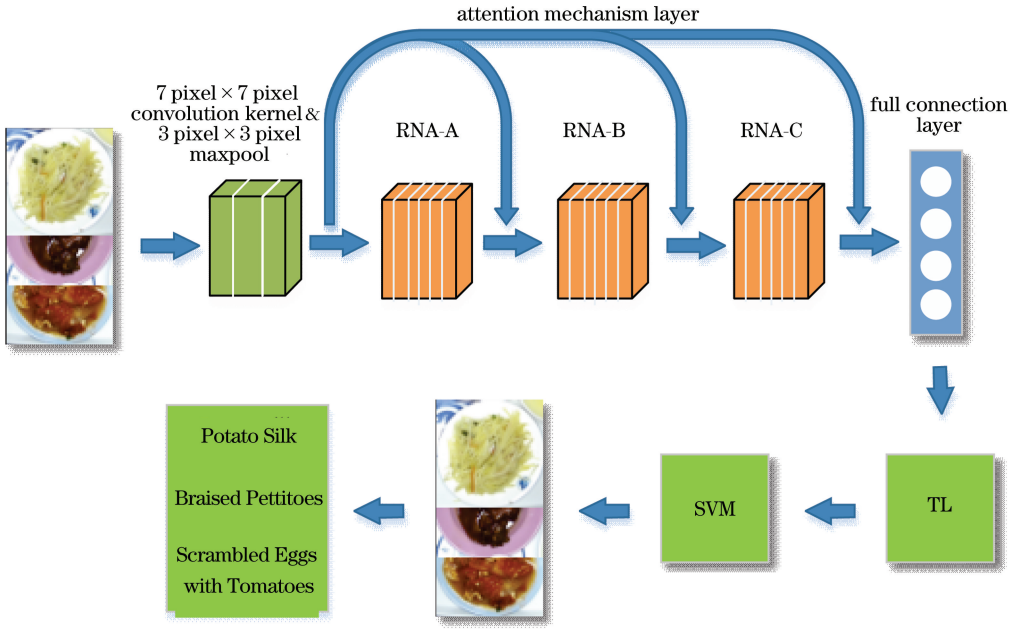

图 1　模型结构

Fig. 1 Model structure

2.1 RNA 模块

由于中式菜品的种类繁多、相似度高，而传统 ResNet 网络在进行菜品识别时会忽略不同尺度下的图像语义信息，且经过多层卷积操作后，语义会丢失。本文将改进后的残差网络 RNA 运用到中式菜品识别中，RNA 模块的结构如图 2 所示，其中 Add 分支为注意力机制层，Relu 为激活函数，并使用 Dropout 策略减少过拟合。

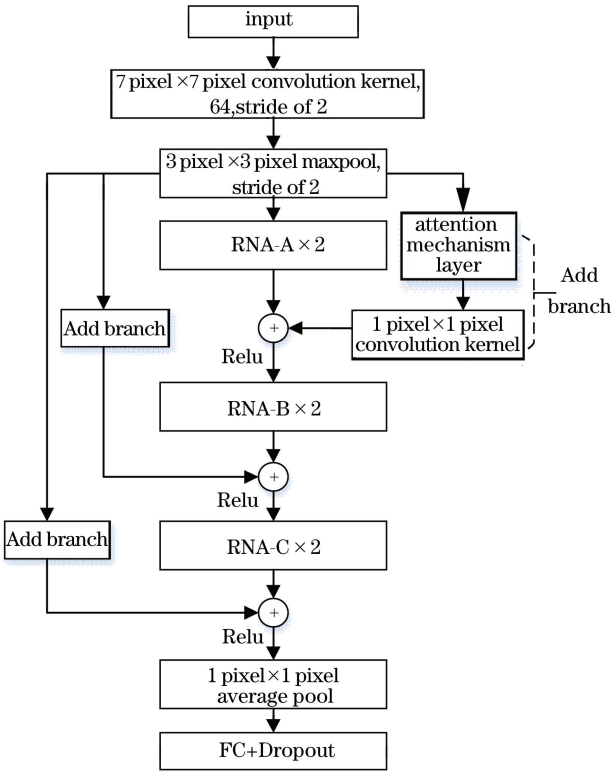

图 2　RNA 网络结构

Fig. 2 RNA network structure

如图 2 所示，将预处理过的菜品图像作为 RNA 的输入，假设输入图像 $\boldsymbol{X}$ 的尺寸为 224 pixel×224 pixel×3 pixel，经过一个尺寸为 7 pixel×7 pixel、步长（stride）为 2 的卷积核，并进行了尺寸为 3 pixel×3 pixel、stride 为 2 的最大池化操作。经过上述操作，原始输入就会变成尺寸为 56 pixel×56 pixel×64 pixel 的特征图，极大地减小了存储空间。

图 2 中的 RNA-A、RNA-B、RNA-C 三个模块的卷积结构图分别如图 3、4、5 所示。可以看出，除了卷积层略有不同外，其整体结构是差不多的，都是卷积层加上一层恒等映射层（Identity）。RNA-A 模块中的卷积层可分为四个分支，第一个分支是由尺寸分别为 1 pixel×1 pixel、3 pixel×3 pixel、3 pixel×3 pixel 的三个卷积核串联组成的，这里使用两个 3 pixel×3 pixel 的卷积核而不使用大一点的卷积核是为了减少计算量。第二个分支是由尺寸分别为 1 pixel×1 pixel、3 pixel×3 pixel 的两个卷积核串联组成的，第三个分支是由一个池化层以及一个尺寸为 1 pixel×1 pixel 的卷积核串联组成的，第四个分支只有尺寸为 1 pixel×1 pixel 的卷积核。各个分支中均使用了尺寸为 1 pixel×1 pixel 的卷积核，其目的是减少计算参数的维度并保证输入到下一层的各分支的参数维度是一致的，以便于后续不同尺度的特征进行特征融合（Filter Concat）[11]。

除了卷积层外，还有一个分支为恒等映射层，其作用是将上一层的输出直接与这一层卷积输出相加以作为这一层的输出，这样可以防止梯度消失或梯度爆炸，同时也能加快网络的收敛，具体结构如图 3 所示。RNA-B 模块和 RNA-C 模块的结构相较于 RNA-A 模块在卷积层上有一些差别。RNA-B 模块引用了深层卷积网络 inception-v3 中一个尺寸为 1 pixel×n pixel 和一个尺寸为 n pixel×1 pixel 的卷积核，将其串联以代替尺寸为 n pixel×n pixel 的卷积核，这样计算量降低为之前的 $1/n$，且当 $n=7$ 时得到的特征效果较好[12]，同时也加深了网络的深度，如图 4 所示。RNA-C 模块只是把 RNA-B 模块中串联的尺寸分别为 1 pixel×n pixel 和 n pixel×1 pixel 的卷积层改成了并联，如图 5 所示，这样在减少计算量的同时可获取不同尺度下的特征向量，使得高维表示更加容易处理，最后将提取到的图像特征信息传送到出输出层。输出表达式为

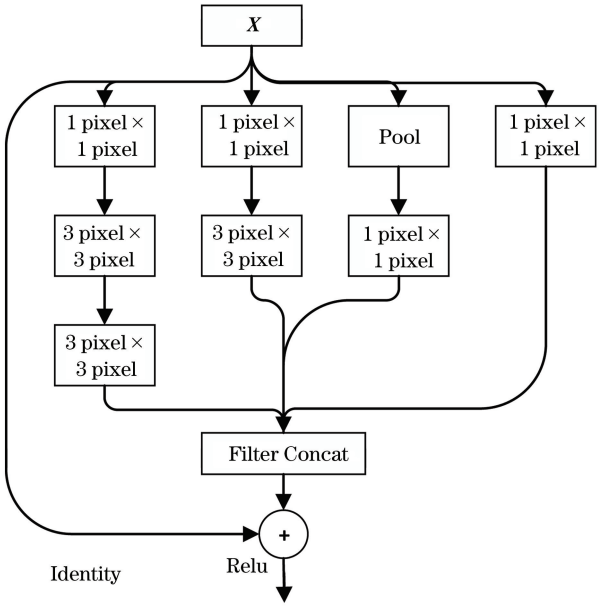

图 3　RNA-A 的卷积结构

Fig. 3 Convolution structure of RNA-A

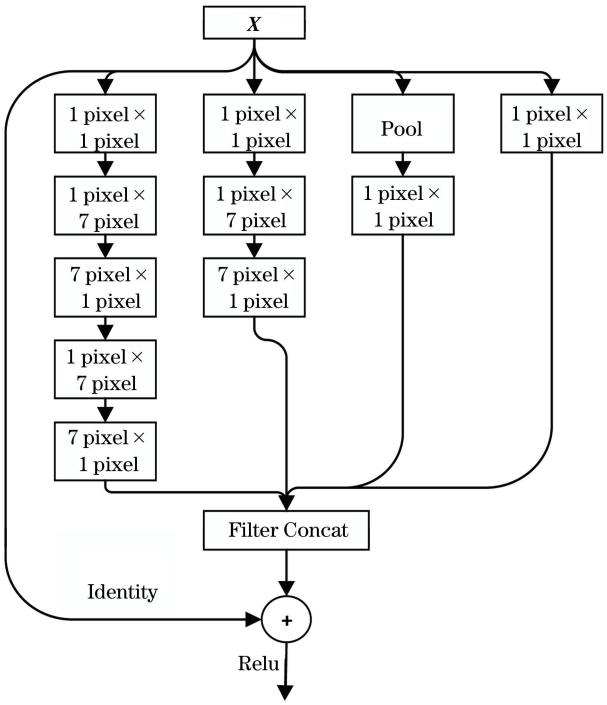

图 4　RNA-B 的卷积结构

Fig. 4 Convolution structure of RNA-B

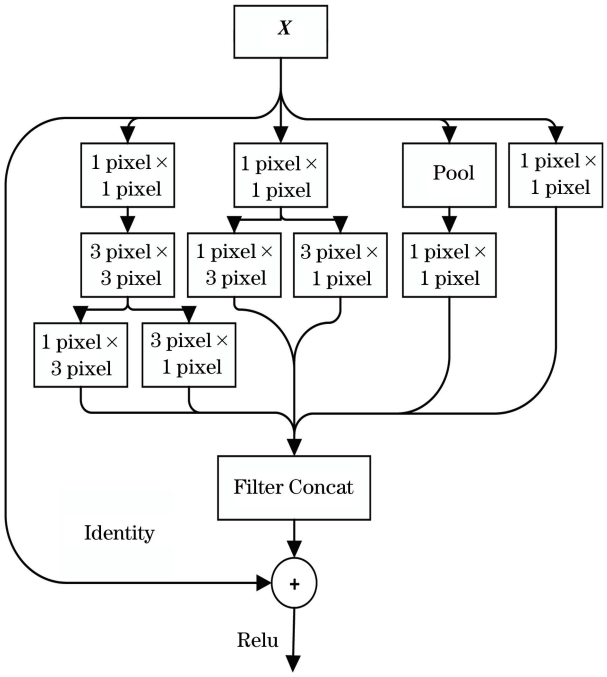

图 5　RNA-C 的卷积结构

Fig. 5 Convolution structure of RNA-C

$$O_{\mathrm{u}}=f_n\{\cdots f_2[f_1(\boldsymbol{X}\boldsymbol{W}_1+b_1)\boldsymbol{W}_2+b_2]\cdots\times\boldsymbol{W}_n+b_n\}\ , \tag{1}$$

式中：$\boldsymbol{W}_i$ 表示第 i 层的权重参数矩阵，$i=1,2,3,\cdots,n$，其中 n 为卷积层中的网络层数；b_i 为第 i 层的偏置项；$f_i(\cdot)$ 为第 i 层的激活函数；O_{u} 为 RNA 网络的输出。

2.2　注意力机制层

图 2 中的 Add 分支是利用注意力机制层将上一层的输出分别加到 RNA-A、RNA-B 和 RNA-C 的输出里。增加 Add 分支可以使有用的特征得到更大的权重并弱化对结果影响比较小的特征，极大提升了神经网络处理信息的能力[13]。图 6 所示为注意力机制模块的基本结构图。其中，$x_t(1\leqslant t\leqslant n)$ 表示原始的图像输入，是 $\boldsymbol{X}$ 的矩阵元；$\boldsymbol{h}_t(1\leqslant t\leqslant n)$ 是 x_t 经过前面卷积层后的隐层特征；$\alpha_t(1\leqslant t\leqslant n)$ 是各个隐层特征在新的隐层特征中的权重系数。最终通过加权求和获取新的图像区域特征表达 $\boldsymbol{s}$：

$$\boldsymbol{s}=\sum_{t=1}^{n}\alpha_t\boldsymbol{h}_t\ , \tag{2}$$

$$\alpha_t=\frac{\exp[v_t\tanh(w_t\boldsymbol{h}_t+b_t)]}{\sum_{t=1}^{n}[v_t\tanh(w_t\boldsymbol{h}_t+b_t)]}, \tag{3}$$

式中：w_t 是 t 时刻隐含层的权重系数；v_t 是用来衡量 t 时刻特征点重要度的权重；b_t 是 t 时刻相应的偏移量。通过(2)式，就可以实现输入的初始特征到新的注意力特征的转换。

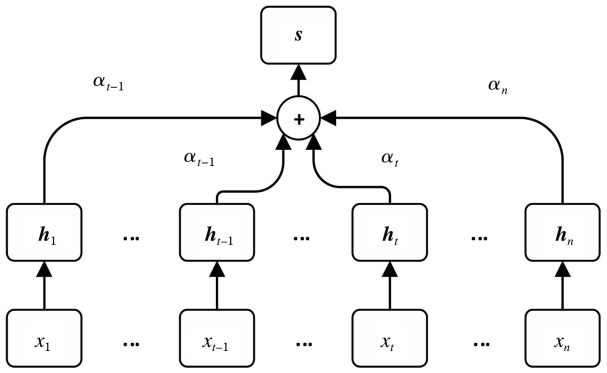

图 6　注意力机制模块的基本结构

Fig. 6 Basic structure of attention mechanism module

2.3　三元组损失

如图 7 所示，三元组损失是一个三元组。首先从训练数据集中随机选取一个样本图像 Anchor(记为 x^{a})，然后在 Anchor 所属类别中随机选取一个样本图像 Positive(记为 x^{p})，并在与 Anchor 所属类别不同的类别中随机选取一个样本图像 Negative(记为 x^{n})，这样三个样本图像(Anchor、Positive、Negative)组成的一组图像样本即成为一个三元组。

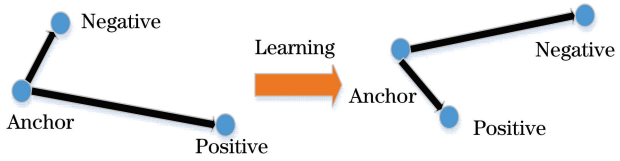

图 7　TL 学习示意图

Fig. 7 Schematic of TL learning

一个三元组经过 RNA-TL 模型的一系列特征提取及空间映射后，这个三元组中三个样本的特征表达分别记为 $f(x_t^{\mathrm{a}})$，$f(x_t^{\mathrm{p}})$，$f(x_t^{\mathrm{n}})$。TL 原理就是通过不断的训练学习，使 x^{a} 和 x^{p} 之间的欧氏距离尽可能小且 x^{a} 和 x^{n} 之间的欧氏距离尽可能大，同时 x^{a} 和 x^{p} 之间的欧氏距离比 x^{a} 和 x^{n} 之间的欧氏距离小一个阈值 α（常量）。转化公式[9]为

$$\| f(x_t^{\mathrm{a}})-f(x_t^{\mathrm{p}}) \|_2^2+\alpha < \| f(x_t^{\mathrm{a}})-f(x_t^{\mathrm{n}}) \|_2^2, \tag{4}$$

式中：$\forall(x_t^{\mathrm{a}},x_t^{\mathrm{p}},x_t^{\mathrm{n}})\in T$，$T$ 表示整个数据集。

由此可得损失函数为

$$L=\sum_t^N[\| f(x_t^{\mathrm{a}})-f(x_t^{\mathrm{p}}) \|_2^2-\| f(x_t^{\mathrm{a}})-f(x_t^{\mathrm{n}}) \|_2^2+\alpha]_+, \tag{5}$$

式中：N 为样本的总个数；$(\cdot)_+$ 表示当括号内的值小于 0 时损失取为 0，大于 0 时取该值为损失值。

由(5)式可以看出，最小化损失 $L=d_{\mathrm{ap}}-d_{\mathrm{an}}+\alpha$，其中 d_{ap} 为类内距离，d_{an} 为类间距离。

在实际情况中，当数据量很大时，我们很难穷举所有数据。因此，在每次训练过程中，随机选取 A 个训练数据类别数并在每个类别中随机选取 B 个样本图像进行训练。(4)式只比较了 x^{a} 和 x^{p} 之间的欧氏距离及 x^{a} 和 x^{n} 之间的欧氏距离，为了加大类内距离与类间距离的差距以明显提升分类效果，对 x^{n} 和 x^{p} 之间的欧氏距离进行比较。针对(5)式进行如下优化：

$$L'=\sum_t^N\{\| f(x_t^{\mathrm{a}})-f(x_t^{\mathrm{p}}) \|_2^2-\min[\| f(x_t^{\mathrm{a}})-f(x_t^{\mathrm{n}}) \|_2^2,\| f(x_t^{\mathrm{p}})-f(x_t^{\mathrm{n}}) \|_2^2]+\alpha\}_+。 \tag{6}$$

经过大量实验，发现将损失函数中的欧式距离的平方去掉，得到的分类效果更好，查阅资料，在相关文献中也有类似结论[14]。所以最终的损失函数可以优化为

$$L''=\sum_t^N\{\| f(x_t^{\mathrm{a}})-f(x_t^{\mathrm{p}}) \|_2-\min[\| f(x_t^{\mathrm{a}})-f(x_t^{\mathrm{n}}) \|_2,\| f(x_t^{\mathrm{p}})-f(x_t^{\mathrm{n}}) \|_2]+\alpha\}_+。 \tag{7}$$

最终利用 SVM 进行分类。对于包含 k 个类别的分类问题，训练出 $\frac{k(k-1)}{2}$ 个分类器，即每两个类别训练出一个分类器，并依据 $\frac{k(k-1)}{2}$ 个分类器的结果，通过“投票”方式预测结果。

3　模型描述

本文提出的模型描述如下。

输入：中式菜品数据集(Food208、Food292)。

输出：训练后的菜品分类模型。

Step 1：对准备好的数据集进行图片预处理，包括对图像进行线性变换增强，将图像的亮区域变暗，暗区域变亮；对图像的像素尺寸进行统一裁剪。

Step 2：从处理好的数据集中随机抽取 A 个类别并从对应的每个类别中随机抽取 B 个样本，这些样本组合成一个批次(Batch)。然后在抽取的样本中进行三元组的组合，并将所有的三元组输入到 RNA 网络中以进行特征提取。

Step 3：基于提取的特征计算 TL 损失函数，进行同类别以及类别间的相似度计算比较，根据结果不断迭代更新权重系数。

Step 4：将上述训练好的参数模型输入到 SVM 中以进行菜品的分类。

4　实验结果与分析

4.1　实验数据准备与处理

为了验证 RNA-TL 模型的有效性，实验选用了一个公开的中国菜品数据集[15]和课题组采集的中式快餐菜品数据集。公开的中国菜品数据集一共有 208 个类，平均每个类有 800 张左右的图片，一共有 16 万张，这里简称 Food208，对应的样例示图如图 8 所示。从比较流行的中式快餐餐厅中收集了 292 类菜品图像，每个类约有 300 张图片，一共 87603 张，简称 Food292，其对应的样图示例如图 9 所示。

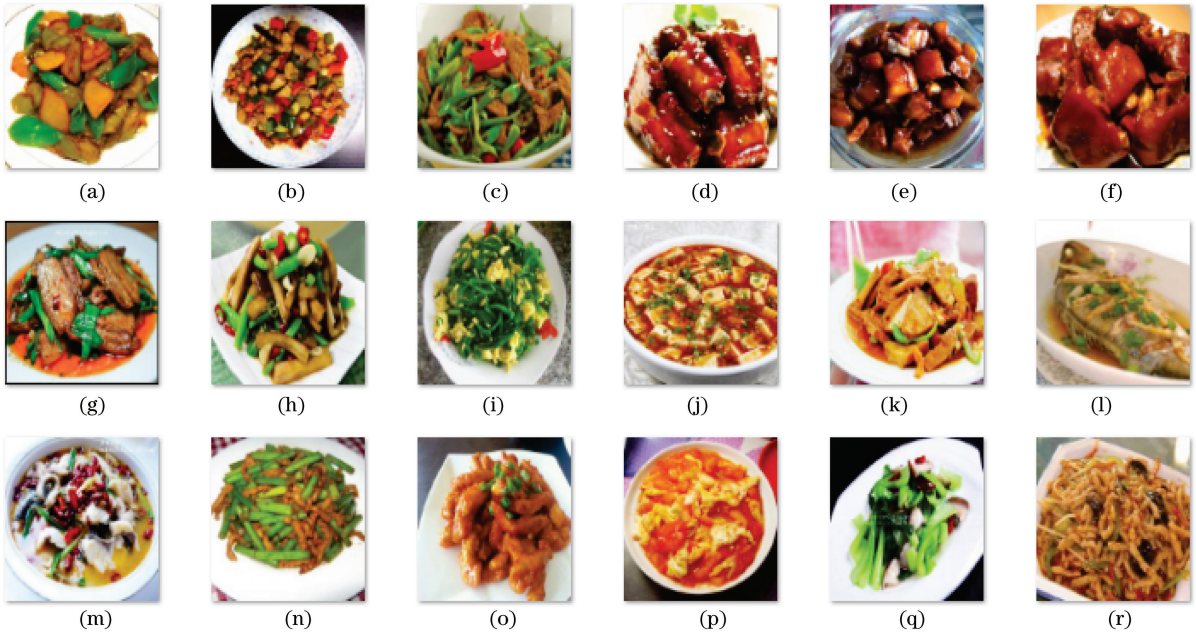

图 8　Food208 数据集样图示例

Fig. 8 Samples in Food208 dataset

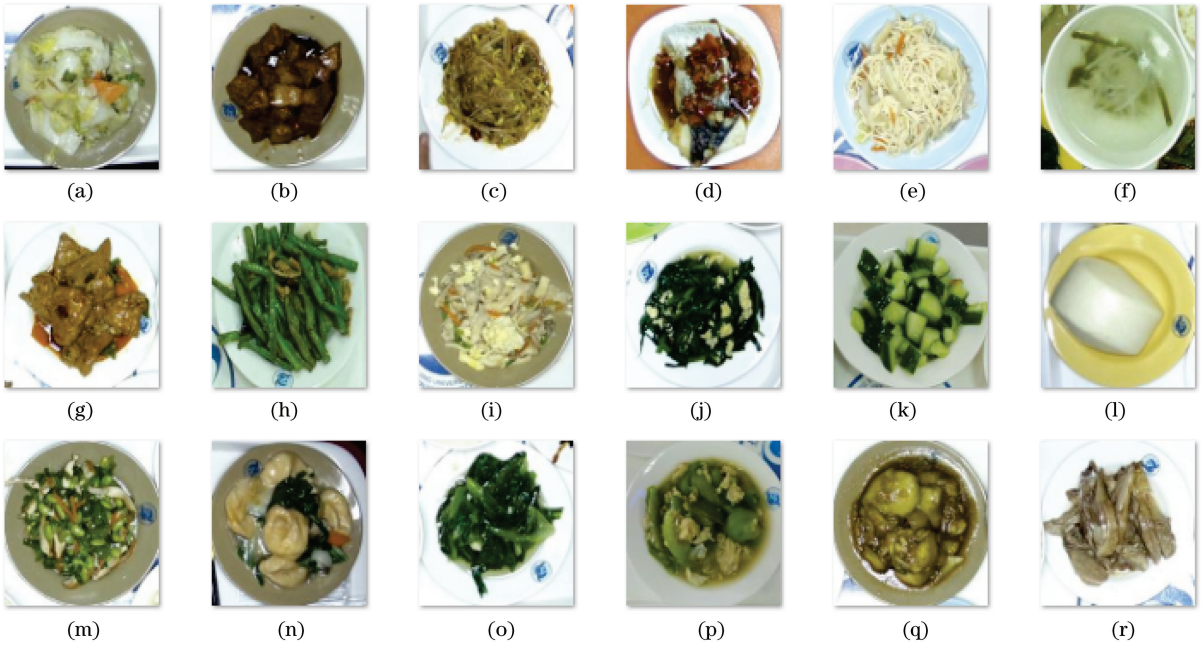

图 9　Food292 数据集样图示例

Fig. 9 Samples in Food292 dataset

对图片进行预处理。先对图像进行线性变换增强，适当调亮图像中比较暗的区域并适当调暗图像中比较亮的区域。然后将所有的图像尺寸统一裁剪成 224 pixel×224 pixel。最后将处理过的数据划分为训练集和测试集，按照 8∶2 的图片张数比例随机分配，具体分配数量如表 1 所示。

表 1　不同数据集中训练集与测试集的图片数量

Table 1 Image numbers of training dataset and test dataset for different datasets

Dataset	Training	Test
Food208	128016	20214
Food292	70082	17521

4.2 实验环境与参数设置

本实验采用 Tensorflow 深度学习框架，用 python3.6 进行编程，在 Win10 系统上进行实验，硬件方面选用了 CPU 9700K，显卡为 GTX 1080 8G。

本次实验使用三元组的方法，因此每次输入必须是 3 的倍数，这里将批大小设置成 90。每次训练随机选取 40 个类别，每个类别随机抽取 150 个样本，因此一个 epoch 需要迭代 6000 次。实验中使用 Adam 优化器对学习率进行优化，为了防止过拟合，增加了 Dropout 层，并将其参数设置为 0.4。模型的具体参数如表 2 所示。

表 2 模型参数设置

Table 2 Model parameter setting

Parameter	Content
Input size	224 pixel×224 pixel
Epoch	200
Batch size	90
Optimizer	Adam
Dropout parameter	0.4

4.3 实验评价指标

本实验采用宏平均查准率（Macro average precision，P）作为实验评价指标，其指标表达式为

$$P=\frac{1}{m}\sum_{k=1}^{m}P_k, \tag{8}$$

式中：P_k 为第 k 个类别的查准率；m 为总类别数。

4.4 实验结果对比与分析

4.4.1 模型对比

为了比较改进卷积神经网络 RNA 的性能优劣，将卷积神经网络更换为当今比较成熟的几种卷积神经网络（inception-v3、inception-v4、resnet-18、inception-resnet-v1 和 inception-resnet-v2）进行实验。其中除 epoch 以外，其他训练参数不变，最后得到的最佳准确率如表 3 所示。

表 3 不同卷积神经网络在 Food208、Food292 数据集上的识别准确率

Table 3 Recognition accuracies of different CNNs on Food208 and Food292 datasets

Convolutional neural network	Food208	Food292
Inception-v3[16]	70.52%	80.65%
ResNet-18[17]	74.64%	82.41%
Inception-v4[18]	79.51%	83.16%
Inception-ResNet-v1[18]	79.93%	85.53%
Inception-ResNet-v2[18]	80.36%	86.10%
RNA	83.66%	90.31%

由表 3 可知，相比于其他网络，RNA 在 Food208、Food292 数据集上的表现更好，RNA 在 Food208 上的准确率比 Inception-ResNet-v2 高出了 3.3 个百分点，在 Food292 上的准确率比 Inception-ResNet-v2 高出了 4.21 个百分点。Inception-ResNet-v1 以及 Inception-ResNet-v2 也采用了 ResNet 中的恒等映射方法，并且增加了网络深度以及卷积核大小，相比 Inception-v3，这两个网络在 Food208、Food292 数据集上的表现都有很大提升。RNA 在改进的残差网络卷积层中加入了注意力机制层，使得有用的特征获得更大的权重，而对结果影响比较小的特征被弱化，从而在两个数据集上的表现更加良好。

其中，RNA 在数据集 Food292 上的表现更好，准确率可以达到 90.31%，相比于其在 Food208 上的准确率，高出了 6.65 个百分点。比较两个数据集可以发现，Food208 数据集中带有很多无用背景信息，而 Food292 数据集中的菜品是盛装在形状类似的餐盘中，除去菜品部分，背景信息的干扰不大。由此可得，RNA 更适合应用在场景变化不大的中式菜品识别中，比如大部分的中式餐厅以及中式食堂中。

对于表 3 中的 6 种卷积神经网络，将 TL 函数改为(2)式，进行了对比实验，其结果如表 4 所示。

表 4 当损失函数为(2)式时不同卷积神经网络的准确率

Table 4 Accuracies of different CNNs when loss function is formula (2)

Convolutional neural network	Food208	Food292
Inception-v3	69.52%	79.53%
ResNet-18	73.15%	81.12%
Inception-v4	78.30%	81.91%
Inception-ResNet-v1	78.21%	84.46%
Inception-ResNet-v2	78.95%	84.95%
RNA	82.50%	89.10%

对表 3 和表 4 进行对比发现，损失函数经过优化后，5 个卷积神经网络在数据集 Food208 以及 Food292 上的识别准确率都有所提升，RNA 在 Food208、Food292 上的准确率分别提高了 1.16 个百分点和 1.21 个百分点。由此可以推断，优化过的 TL 函数在一定程度上提高了模型准确率。

4.4.2 参数对比

基于 RNA-TL 模型，利用上述参数设置，将 epoch 分别设置为 50，100，150，200 和 250，在 Food208 和 Food292 数据集上进行实验，得到的结果如表 5 所示。

表 5 不同 epoch 下 RNA-TL 得到的准确率

Table 5 Accuracies obtained by RNA-TL under different epochs

Epoch	Food208	Food292
50	78.95%	85.63%
100	80.22%	87.02%
150	82.13%	89.23%
200	83.60%	90.31%
250	83.62%	90.32%

由表 5 可知，在数据集 Food208 和 Food292 上，当 epoch 大小为 200 时，得到的准确率趋于稳定。由此可得，在使用 RNA-TL 训练时，epoch 设置为 200 比较合适。

4.4.3 训练时间对比

图 10 列出了不同卷积神经网络在数据集 Food208 和 Food292 上得到最佳识别准确率时所需要的时间。可以看出，RNA 在保证取得比较好的识别准确率的同时，所用时间也相对较短。

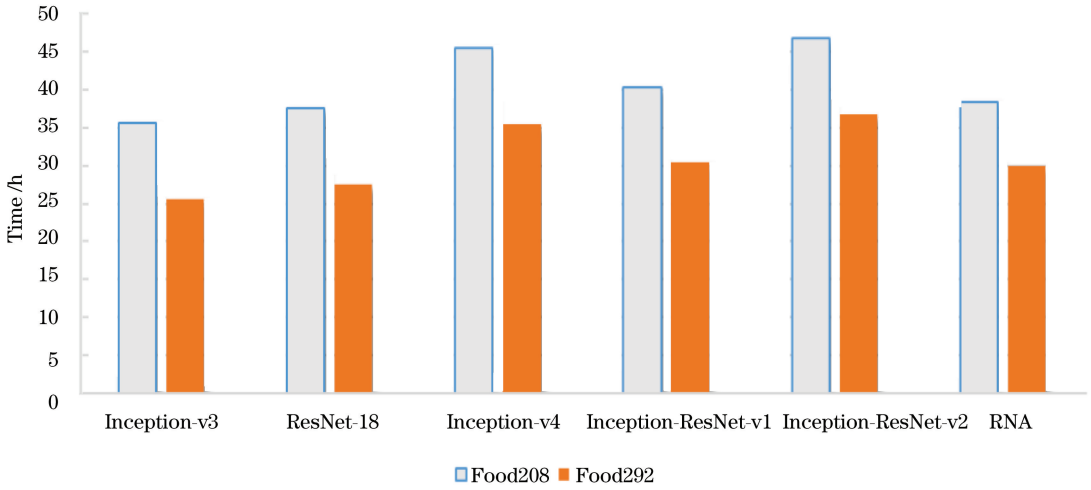

图 10 不同卷积神经网络在数据集 Food208 和 Food292 上训练时所需的时间

Fig. 10 Time required for training different CNNs on Food208 and Food292 datasets

5 结 论

针对中式菜品识别，提出了一种基于残差网络的改进 RNA-TL 菜品识别模型。利用残差网络的恒等映射方法，并在残差网络中融合多个不同大小的卷积核，从而提取了多尺度特征向量。通过添加一层注意力机制层，有效特征信息的权重得到增大。为了防止特征信息的丢失，将 Add 分支分别与 RNA-A、RNA-B、RNA-C 的输出进行拼接。接着基于所提取的特征信息，利用改进的 TL 函数比较类间的相似度，并不断迭代更新权重参数。最终将训练得到的特征输入到 SVM 中进行菜品识别分类。相比于现有的卷积神经网络，RNA-TL 在数据集 Food208 以及 Food292 上表现出更加优异的性能。

中式菜品的种类繁多，且受材料以及制作菜品人员水平的影响，训练的数据集样本比较单一，所提算法是否能适应多样化的数据集还有待验证。通过结合先进的目标检测算法，所提算法能更好地应用于实际生活中。

参 考 文 献

[1] Mezgec S, Seljak B K. Using deep learning for food and beverage image recognition[C]// 2019 IEEE International Conference on Big Data (Big Data), December 9-12, 2019, Los Angeles, CA, USA. New York: IEEE, 2019: 19393612.

[2] Yang S, Chen M, Pomerleau D, et al. Food recognition using statistics of pairwise local features [C] // 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 13-18, 2010, San Francisco, CA, USA. New York: IEEE, 2010: 11500651.

[3] Zheng J N, Wang J, Ji X Y. Food image recognition via superpixel based low-level and mid-level distance coding for smart home applications [J]. Sustainability, 2017, 9(5): 1-17.

[4] Mezgec S, Koroušić Seljak B. NutriNet: a deep learning food and drink image recognition system for dietary assessment[J]. Nutrients, 2017, 9(7): E657.

[5] Martinel N, Foresti G L, Micheloni C. Wide-slice residual networks for food recognition [C] // 2018

IEEE Winter Conference on Applications of Computer Vision, March 12-15, 2018, Lake Tahoe, NV, USA. New York: IEEE, 2018: 17751381.

[6] Pan L L, Qin J H, Chen H, et al. Image augmentation-based food recognition with convolutional neural networks [J]. Computers, Materials & Continua, 2019, 59(1): 297-313.

[7] Ng Y S, Xue W Q, Wang W, et al. Convolutional neural networks for food image recognition: an experimental study [C] // Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management, October 15-21, 2019, Nice, France. New York: ACM Press, 2019: 33-41.

[8] Duan X M, Zhu M, Bao T L. Application of bilinear model in Chinese image classification[J]. Journal of Chinese Computer Systems, 2019, 40(5): 1050-1053.
段雪梅, 朱明, 鲍天龙. 双线性模型在中国菜分类中的应用[J]. 小型微型计算机系统, 2019, 40(5): 1050-1053.

[9] Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering [C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 15524223.

[10] Rozi M F, Mukhlash I, Soetrisno, et al. Opinion mining on book review using CNN-L2-SVM algorithm[J]. Journal of Physics: Conference Series, 2018, 974: 012004.

[11] Wu Q, Li Q, Guan X. Optical music recognition method combining multi-scale residual convolutional neural network and bi-directional simple recurrent units[J]. Laser & Optoelectronics Progress, 2020, 57(8): 081006.
吴琼, 李锵, 关欣. 基于多尺度残差式卷积神经网络与双向简单循环单元的光学乐谱识别方法[J]. 激光与光电子学进展, 2020, 57(8): 081006.

[12] Wang C, Chen D L, Hao L, et al. Pulmonary image classification based on inception-v3 transfer learning model[J]. IEEE Access, 2019, 7: 146533-146541.

[13] Chu J H, Tang W H, Zhang S, et al. An attention model-based facial expression recognition algorithm[J]. Laser & Optoelectronics Progress, 2020, 57(12): 121015.
褚晶辉, 汤文豪, 张姗, 等. 一种基于注意力模型的面部表情识别算法[J]. 激光与光电子学进展, 2020, 57(12): 121015.

[14] Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification [EB/OL]. (2017-11-21) [2020-06-15]. https://arxiv.org/abs/1703.07737.

[15] Chen X, Zhu Y, Zhou H, et al. ChineseFoodNet: a large-scale image dataset for Chinese food recognition [EB/OL]. (2017-10-15) [2020-06-15]. https://arxiv.org/abs/1705.02743.

[16] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision [EB/OL]. (2015-12-11) [2020-06-15]. https://arxiv.org/abs/1512.00567.

[17] He K, Zhang X, Ren S, 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, 2016:16541111.

[18] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning [EB/OL]. (2016-08-23) [2020-06-15]. https://arxiv.org/abs/1602.07261.