**中国食物图像处理**

1. **引言**

食物图像的处理和分析，是计算机视觉领域的一个经典任务。食物识别是视觉食物分析中的主要目标之一，旨在对构成食物图像的食物或成分进行分类。分析食物图像的困难之处，在于食物外观的特性——与一般物体不同，大多数食物并不是刚性物体，因此没有明确的空间排列和形状，这对于算法的准确识别和分类是一个挑战。

对于中国食物的分析则更为困难。因为相比于西方菜肴，中国菜肴种类繁多，传统的就有八大菜系。而且与所谓的“白人饭”不同，中国菜讲究色香味俱全，制作配料和工序错综复杂，并且常常呈现出食物彼此交融在一起的情况，例如炖菜、汤羹等，导致有时肉眼都很难分辨菜品，所以更需要精度极高的算法。

随着深度学习算法，特别是卷积神经网络CNN的兴起，其在各种识别和分类情景中都有广泛的运用。基于深度学习算法的视觉食物分析相比于传统的基于手工特征提取的方法，表现出了巨大的性能提升。同时，因为标注食物类型和烹饪等日常任务可以在几乎没有专业知识的情况下完成，所以其相对于其他领域的标注成本较低，可以有足够的数据来进行训练。

本报告研究了用于构建中国食品识别系统的数据集、机器学习算法等，特别关注了深度学习领域，包括卷积神经网络、迁移学习等技术，并且在个人电脑上完成了训练与验证实验。

**二、食物图像数据集**

1. 数据集分类

一般来说，现有的食物图像数据集根据其特征可分为单类数据集和多类数据集。单一类别数据集通常只包含来自一种烹饪风格的食物，这有助于搭建专有模型，提高其在特定领域的识别精度，例如专注于中国食物的数据集可以提高模型在中餐领域的认知准确性。

然而，数据集的单一类别对于分辨来自不同地域和文化的不同食物形式，存在着局限性。相比之下，多类别数据集覆盖多种地域文化，使用该种数据集可以有效减少模型对特定类别的过拟合，进而提高模型识别的准确性和迁移性能。此外，多类别数据集也更加关注食品营养信息和餐厅地理信息，同时可收集和分析相关的消费者行为、餐厅服务质量等信息，从而更好地解释影响食品分类的各种因素。

2. 中国食物数据集

在广泛阅读论文后，发现目前为止有5个较常见的中国食物图片数据集，如下表所示：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 名称 | 标注类型 | 类别数量 | 图片数量 | 时间 |
| Vireo-Food 172 | 分类&原料 | 172 | 110241 | 2016 |
| ChinFood1000 | 分类 | 1000 | \ | 2017 |
| ChineseFoodNet | 分类 | 208 | 185628 | 2017 |
| ChinaMartFood-109 | 分类&营养元素 | 109 | 10921 | 2022 |
| CNFOOD-241 | 分类 | 241 | 190000 | 2023 |

* Vireo-Food 172从两家中国食物网址中获取了172组食物图片，每一组都收集了超过100张的256×256食物图片。该数据集中食物图片的原料成分由十个家庭主妇标注，产生了共353个原料成分标签，平均每张图像包含3个原料成分。
* ChinFood1000在当时所有公开的食品数据集中拥有最多的食品类别。其类别包括最受欢迎的852类中国菜肴、91类饮料和零食、26种水果和31种其他食物。
* ChineseFoodNet从互联网上收集了19.4 GB的中国食物图片，其中每个图像都被手动标记。该数据集包含208个食品类别的185,628张图像。数据集中的图像不经过任何处理，保持了图像的原始外观。
* ChinaMartFood-109从中国市场收集了109种食品的10921张图片，该数据集不仅包括食物的分类，还包括23种营养成分的分类，涵盖18个主要的食物类群。
* CNFOOD-241是由Bokun Fan等人在论文Automatic Chinese Food recognition based on a stacking fusion model中构建的迄今为止最大的中国菜肴数据集，包含了超过19万张图片，241个类别，涵盖主食、肉类、素食、混合肉和蔬菜、汤、甜点类别。

**三、食物图像识别技术的发展历程**

1. 传统机器学习算法

在2010年之前，在深度学习和大图像数据集还未兴起时，传统的机器学习算法在食品分类中占据主导地位，包括支持向量机、逻辑回归、K近邻、K-means、决策树等方法。这些传统算法利用手工提取特征，通常在小型数据集上拥有良好的性能。

早在20世纪70年代，K-NN算法首次应用于图像识别中的图像分类任务，人们根据图像特征向量之间的距离进行分类；在80年代科学家们开始广泛使用K-means算法进行图像分割，将图像中的像素点聚类成不同的区域，从而实现图像分割和物体识别；同时，逻辑回归、贝叶斯网络、决策树算法在受到广泛研究后开始应用于图像识别和分类任务中；到了90年代，随机森林等集成学习方法的引入，进一步增强了食物图像识别系统的性能，食物图像识别的精度和效率也得到了提升。

传统机器学习在中餐领域的一个经典应用，是2012年来自台湾大学的Mei-Yun Chen等人使用了SVM多分类器，并且结合Adaboost算法，对包含50个主要类别的中餐图片数据完成了种类和数量的识别。

2. 基于卷积神经网络的深度学习算法

然而，手工提取特征意味着需拥有一些专业领域内的先验知识；随着2012年Alexnet的横空出世，人们终于意识到了神经网络的强大。通过基于链式法则的梯度计算和反向传播，可以有效地训练神经网络参数。拥有海量神经元和卷积层等面向视觉的架构，使得神经网络能够排除人工干扰，有效提取潜在特征。

神经网络的一个经典架构是CNN卷积神经网络，专门用于处理图像和视频数据。CNN通过卷积层、池化层和全连接层来提取图像中的特征，并进行分类或其他任务。基于CNN的一些经典模型有AlexNet、VGG16、DenseNet等。在中餐图像识别算法中，基于深度卷积神经网络的研究已经取得了一定进展。下面将列举一些基于CNN的中餐图像识别分类算法。

Hao等（2018）针对单目标中餐图像识别问题，利用迁移学习的思想，将分类性能优秀的CNN预训练模型用于中餐图像数据集，并提出一种基于注意力机制的关键语义特征提取和学习网络，针对图像关键区域进行语义特征提取，减少了噪音干扰，提高了中餐图像的可区分性语义特征的提取能力和对细粒度特征学习的能力。

Teng等（2018）提出了一种用于中餐识别的小型且高效的CNN架构，该架构适用于资源有限的平台。该网络架构设计的处理流水线的思想来源于Bag of Features，即一种基于视觉词袋模型的特征提取方法。

 Liao等(2019)针对传统食品图像识别方法提取特征能力差、准确率差、运行效率差和CNN识别相似食品图像难度大等问题, 提出了一种新的食品图像识别模型China Food-CNN, 以实现对食物的精准分类。在多分类损失函数Softmax With Loss的基础上，提出了最大类间距损失函数(MCSWithLoss) ，以解决食品图像类间相似性大的问题，增大相似类间距。

 Gao等(2021) 为解决多目标中餐图像识别，首先对菜品进行目标检测和分割，然后再进行识别。为了准确定位图像中多个菜品的具体区域，减少背景噪声的影响，并且减小模型复杂度和计算成本，提出了一种基于深度分类卷积的改进的Mask R-CNN。

3. 最近的一些创新算法

2023年的研究首次在中国食品图像识别领域应用了多模型融合技术，以提高模型的鲁棒性和准确性。多模型融合通过堆叠方法整合多个深度学习模型，即在基本模型的预测结果上训练二级学习器（meta-learner），从而得出最终的预测结果。Bokun Fan等在CNFOOD-241数据集上，将ResNet101, DenseNet169和ResNeXt101\_32x32d模型通过堆叠融合，提高了食品图像的识别准确率。其中ResNeXt101\_32x32d模型在单模型测试中表现最优，但通过融合模型进一步提高了准确率至82.88%​​。

到2024年，中国食品图像识别进入了细粒度图像分类的新阶段，其中ResVmamba模型成为了技术发展的新高点。该模型是由Chi-Sheng Chen等开发，基于深度残差学习和今年大火的Mamba架构，即选择性状态空间模型的细粒度食品类别视觉分类系统。Yue Liu等人提出的Vmanba模型，率先将Mamba机制融入到图像分类中，其在细粒度和食品分类方面超过了当前的SOTA模型；而Chen在此基础上提出的ResVmamba，则在没有预训练权重的情况下，将在CNFOOD-241数据集上的分类准确率进一步提高到了79.54%。

1. **应用场景**
2. 营养识别和热量摄入监测

现代社会中的个体时常因摄入过多热量的食物而增加健康风险，导致各种疾病如肥胖、糖尿病、心脏问题的发生。对此，我们可以可以通过智能手机等设备的摄像头捕捉餐盘中的食物图片，应用算法自动识别食物种类和估计食物量，进而计算摄入的热量和营养成分。如Bokun Fan在论文中提到的，堆叠融合模型能够有效帮助用户监控日常饮食，特别是对于糖尿病患者和肥胖者进行饮食控制。

1. 食品安全和质量检测

可利用图像识别技术检测食品的新鲜度、是否变质，以及是否含有非法添加剂等。如Chi-Sheng Chen在论文中提到的，可使用细粒度的图像分类模型如ResVmamba，对食品细节进行深入分析，从而判别食品质量和安全性

1. 智能厨房管理

在智能厨房环境中，图像识别技术可以帮助自动识别存储的食材，记录食材的使用情况，提醒食材的保质期限。将图像识别与物联网设备结合，实现食材的智能管理和自动补给提醒。

**五、当前技术的优劣势和改进方案**

优势：现代的深度学习模型，如卷积神经网络（CNN）和视觉转换器（ViT），在图像分类和识别方面表现出高效的性能​​，而一些创新的方法正在不断涌现。同时，自动化处理食品图像，减少了人工操作的需求，提高了处理速度和准确性。

劣势：高精度的模型通常需要大量的算力资源，而这是GPU紧缺的时代。而且，目前大多数食品图像数据集主要还是集中在西餐上，中餐数据集较少。并且由于中餐图像的特性，其处理相较于西餐而言，语义特征提取能力差、识别准确率低。

改进方向：

1. 利用多模态学习，结合图像数据和其他形式的数据（如文本描述、用户偏好等），提高食品识别的准确性和系统的适应性。
2. 开发更多的轻量化模型，以适应移动设备和边缘计算设备，减少对计算资源的依赖。
3. 建立更加全面和多样化的中餐数据集。

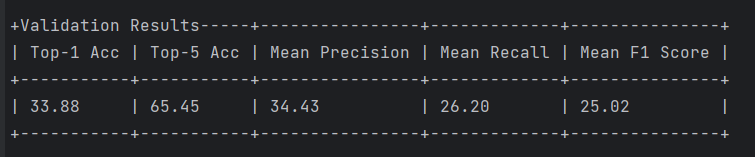
**六、实验与分析结果**

实验设置：

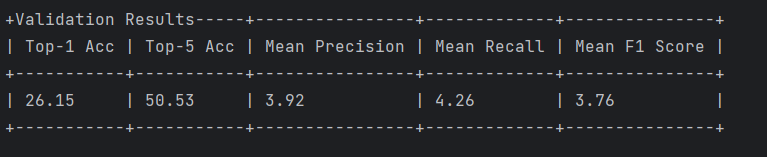
为探究不同模型在中国食品图片上的性能，我们选用ResNet、DenseNet、ViT这三种基准模型，并分别在CNFood241数据集上进行微调训练，划分相同比例的训练集和验证集。本次实验使用显卡为4070ti，显存12g。由于设备原因，训练时选择减少batch\_size和训练周期数，学习率参照原库无改动，预训练模型使用原库提供模型，同时由于每个类别的数据量较少，选择关闭图像增强手段来避免污染数据。训练完后打印验证结果，即分类报告。

结果分析：

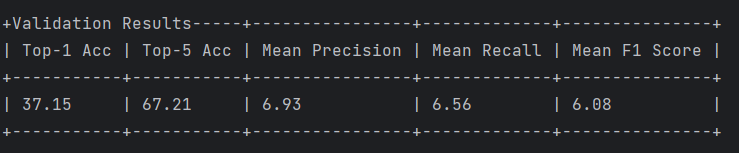
使用Resnet152，设置三个周期，每个训练周期时长约为2min，batch\_size为32。



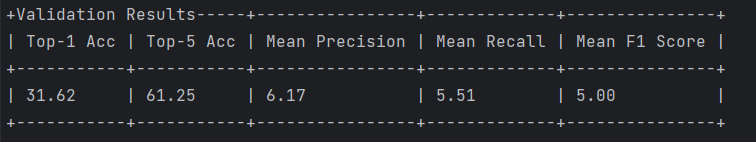
使用Densenet201，设置三个周期，每个训练周期时长约为1.5min，batch\_size为32。



使用Vision\_transformer\_base\_p32\_384，设置三个周期，每个训练周期时长约为1min，batch\_size为16。由于batch\_size的调整，所以该模型得到了较好的结果。



若batch\_size与前面相同，则每个训练周期时长约为0.5min。结果变为：



由于项目时长和设备原因，我们未能完全微调出针对中餐图片的最优参数，但是在每个模型参数调节差别不大的条件下，本实验能够得出一定结论。对于中餐图片分类任务，ViT的速度最快，但由于参数调整原因没有得到较好的结果。Densenet速度次之，但结果较差，可能原因为学习率设置不当。Resnet最慢，但有着不错的准确率。初次之外，我们还在数据集上实验了传统的SVM算法，但因运行速度过慢，且效果十分欠佳而舍弃。