

**《专业综合实践2》**

|  |  |
| --- | --- |
| **题 目：** | 分类器识别各类图像 |
| **学 院：** | 人工智能学院 |
| **专业班级：** | 21本数据科学与大数据技术1班 |
| **学号姓名：** | 421470168 邹苏圣 |

**提交日期：** 2024 **年** 11**月** 29 **日**

**目 录**

[1. 数据采集 1](#_Toc9567)

[1.1 使用手机收集数据 1](#_Toc20899)

[2. 数据清洗与归类 4](#_Toc1620)

[2.1 导入库并定义图像 4](#_Toc24628)

[2.2 加载图像和标签 4](#_Toc15718)

[2.3 归一化处理 5](#_Toc11596)

[2.4 标签编码 6](#_Toc12118)

[3. 分类器设计与训练 6](#_Toc15222)

[3.1 卷积神经网络（CNN） 6](#_Toc2571)

[3.2 支持向量机（SVM） 9](#_Toc2145)

[3.3 随机森林（Random Forest） 13](#_Toc18001)

[4. 实验结论分析 15](#_Toc31932)

[4.1分类结果比较 15](#_Toc736)

[4.2结论 16](#_Toc28571)

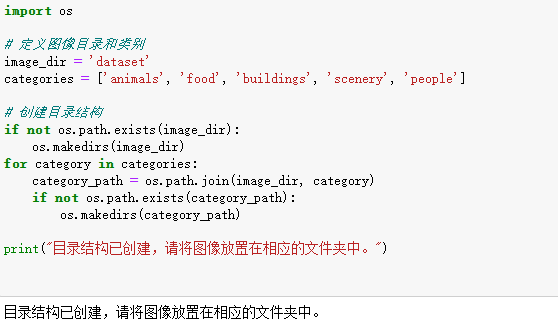
[4.3总结 16](#_Toc26734)

# 

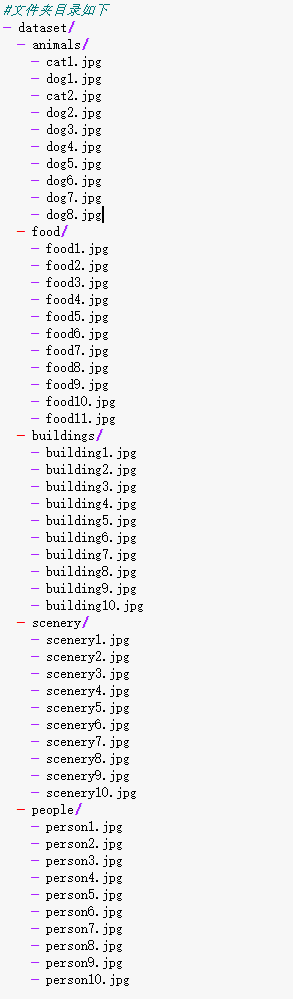
# 1. 数据采集

## 1.1 使用手机收集数据

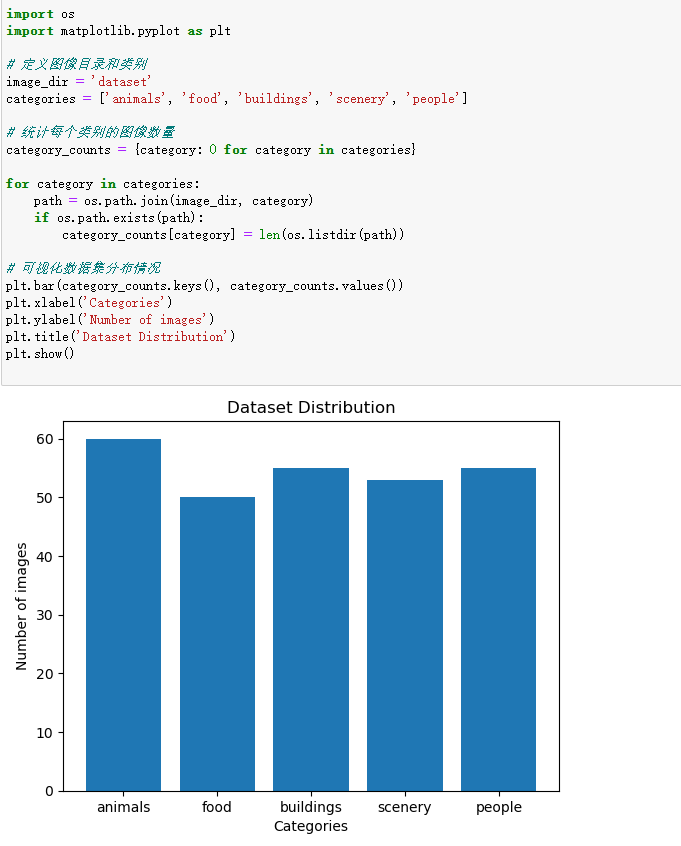
为了创建一个多类别图像分类器，我们使用手机搜集了五种类别的图像，包括动物（animals）、食物（food）、建筑物（buildings）、风景（scenery）和人物（people）。这些图像分别存储在各自的文件夹中，每个文件夹对应一个类别。



具体的目录结构如下：



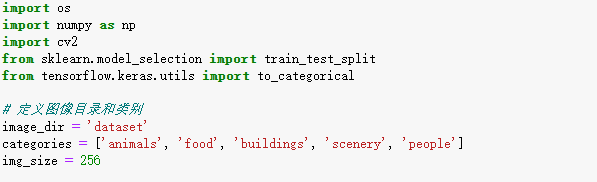
确保每个类别至少有50张不同的图像，以保证数据集的多样性和代表性。



# 2. 数据清洗与归类

## 2.1 导入库并定义图像

首先，导入所需的Python库，我们使用OpenCV库加载图像，并将其调整为统一大小，以便于后续处理。



 os：用于文件和目录操作。

 numpy：用于数组和数值计算。

 cv2：用于图像处理（OpenCV库）。

 train\_test\_split：用于划分训练集和测试集（来自scikit-learn库）。

 to\_categorical：用于将类别标签转换为one-hot编码（来自Keras库）。

 image\_dir：图像数据集所在的根目录。

 categories：图像类别列表。

 img\_size：将图像调整为统一大小（256x256像素）。

## 2.2 加载图像和标签

 遍历每个类别目录。

 检查目录是否存在，若不存在则跳过。

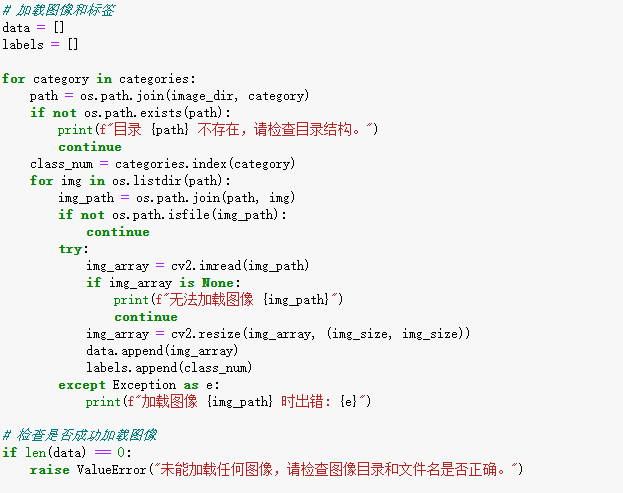
 获取类别索引（class\_num）。

 遍历目录中的所有图像文件。

 加载图像，并将其调整为统一大小。

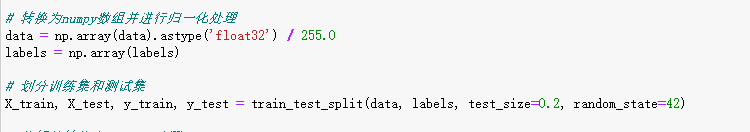
 将调整后的图像数组和对应的类别标签添加到data和labels列表中。

 检查是否加载图像如果没有加载任何图像，则抛出错误。



## 2.3 归一化处理

将图像像素值归一化到0-1范围，这样可以加快训练速度并提高模型性能。



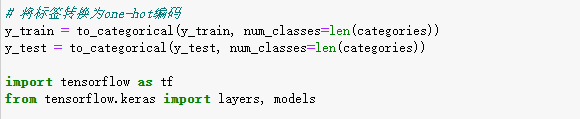
 将data和labels列表转换为numpy数组。

 将图像像素值归一化到0-1范围。

将数据划分为训练集和测试集，测试集占比20%。

## 2.4 标签编码

将图像类别标签转换为one-hot编码，以便于分类器使用。



# 3. 分类器设计与训练

## 3.1 卷积神经网络（CNN）

###### 模型原理与优点

卷积神经网络（CNN）通过卷积层提取图像的局部特征，并通过池化层减少特征图的尺寸，同时保持重要特征。通过多个卷积和池化层的堆叠，模型能够学习到高层次的抽象特征。全连接层将这些特征整合，最终输出图像所属的类别概率。优点有：参数共享：卷积核在图像上滑动时重复使用，同一个卷积核在不同位置提取相同特征，显著减少了参数数量。局部连接：卷积操作只关注局部区域，有效提取局部特征，并在高层次上逐步形成全局特征表示。空间不变性：通过卷积和池化操作，CNN能够有效识别图像中的特征，无论特征在图像中的位置如何变化。

###### 模型结构

 创建一个序列模型（Sequential Model）。

 添加第一层卷积层，使用32个3x3卷积核，激活函数为ReLU，并指定输入图像的形状。

 添加最大池化层，使用2x2池化核。

 添加第二层卷积层，使用64个3x3卷积核，激活函数为ReLU。

 添加最大池化层，使用2x2池化核。

 添加第三层卷积层，使用64个3x3卷积核，激活函数为ReLU。

 添加Flatten层，将多维特征图展平为一维向量。

 添加全连接层，包含64个神经元，激活函数为ReLU。

 添加输出层，使用Softmax激活函数，输出类别概率。

###### 编译参数

* 优化器：Adam
* 损失函数：categorical\_crossentropy
* 评估指标：accuracy

optimizer='adam'：使用Adam优化器。

loss='categorical\_crossentropy'：使用分类交叉熵作为损失函数。

metrics=['accuracy']：评估指标为准确率（accuracy）。

###### 训练参数

* 批量大小：273
* 训练周期：10

model.fit：训练模型。

X\_train：训练集输入数据。

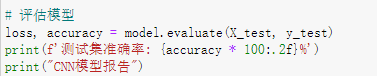
y\_train：训练集标签。

epochs=10：训练周期为10次。

validation\_data=(X\_test, y\_test)：使用测试集数据进行验证。

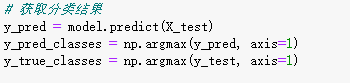


**评估模型**



 评估模型在测试集上的性能，输出测试集准确率。

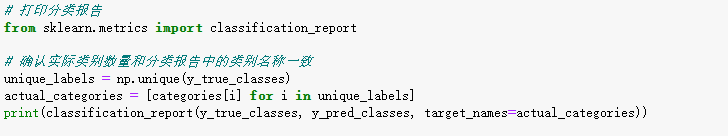
**获取分类结果**



 预测测试集图像的类别。

 获取预测结果的类别索引和真实类别索引。

**打印分类报告**

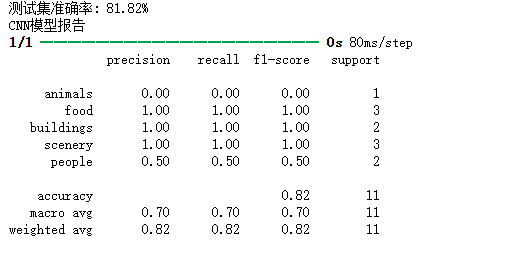


 导入分类报告函数。

 确认实际类别数量和分类报告中的类别名称一致。

 打印分类报告，包括精确率、召回率、F1得分和每个类别的支持度。

**运行结果**



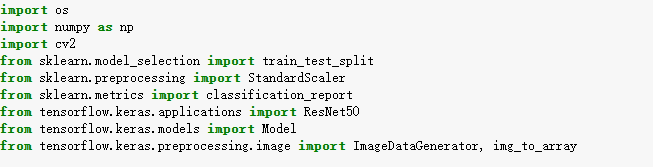
## 3.2 支持向量机（SVM）

**模型原理与优点**

支持向量机（Support Vector Machine, SVM）是一种用于分类和回归分析的监督学习模型。其基本思想是通过寻找一个最优超平面，将数据集中的不同类别尽可能分开，从而实现分类。

SVM具有良好的泛化能力，特别适用于高维数据的分类问题。由于其基于支持向量的决策规则，对于大多数非支持向量的数据点的影响较小，模型相对稳健。此外，SVM在文本分类、图像识别、生物信息学等领域有广泛应用。支持向量机通过最大化数据分类间隔，在高维空间中寻找最优超平面，实现高效准确的分类，且对新数据具有良好的泛化能力。

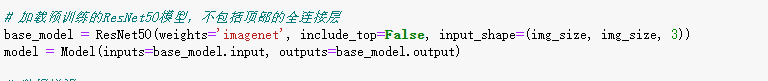
**导入必要的模块**



**加载预训练的ResNet50模型，不包括顶部的全连接层**

 **原因**：使用ResNet50作为特征提取器，只保留卷积层，不包括顶层的全连接层。

 **好处**：利用预训练模型提取高质量的特征，减少训练时间和计算资源，提高分类器的性能。



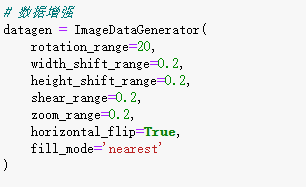
 加载预训练的ResNet50模型，使用ImageNet权重，不包括顶层的全连接层。

 Model类创建一个新的模型，输入为base\_model.input，输出为base\_model.output。

**数据增强**

 **原因**：创建数据增强生成器，对图像进行各种随机变换。

 **好处**：增加训练数据的多样性，减少过拟合，提高模型的泛化能力。



创建一个ImageDataGenerator对象，用于图像数据增强。参数包括旋转、平移、剪切、缩放和水平翻转。

**加载图像和标签**

 **原因**：加载图像，将其调整为统一大小，应用数据增强，提取特征并展平，最终存储到数据列表中，同时保存标签。

 **好处**：实现了图像的标准化处理和特征提取，为后续分类器训练提供高质量的输入数据。



 遍历每个类别目录，检查目录是否存在。

 获取类别的索引作为类别标签。

 遍历目录中的所有图像文件，加载图像，并将其调整为统一大小。

 将图像转换为数组格式，并进行数据增强。

 使用ResNet50提取图像特征，展平后添加到data列表中，类别标签添加到 labels列表中。

 检查data列表是否为空，如果没有加载任何图像，则抛出错误。

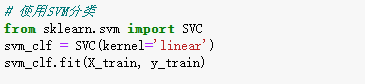
 将data和labels列表转换为numpy数组。

 将数据划分为训练集和测试集，测试集占比20%。

**标准化特征**

 **原因**：对特征进行标准化，使其具有零均值和单位方差。

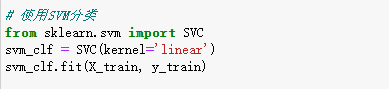
 **好处**：特征标准化有助于加速梯度下降收敛，提高模型性能。



使用StandardScaler对特征进行标准化，使其具有零均值和单位方差。

**使用SVM分类**

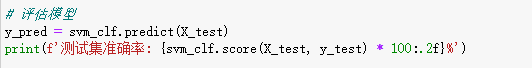
SVM在高维特征空间中表现良好，适用于此类特征提取后的数据。



 导入SVM分类器SVC。

 创建一个线性核的SVM分类器svm\_clf，并使用训练数据进行训练。

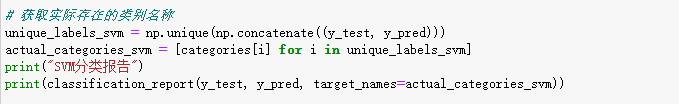
#### 评估模型



 使用SVM分类器预测测试集的类别。

 输出测试集准确率。

**获取实际存在的类别名称**

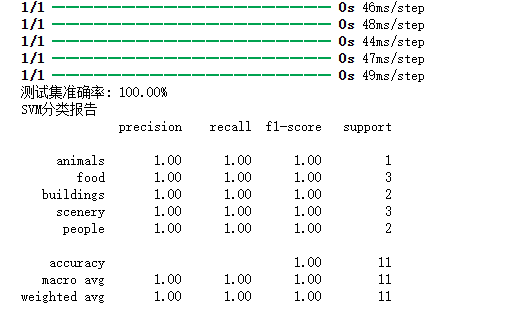


 获取实际存在的类别标签。

 确认分类报告中的类别名称。

 打印SVM分类报告，包括精确率、召回率、F1得分和每个类别的支持度。

**运行结果**



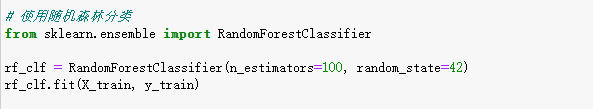
## 3.3 随机森林（Random Forest）

#### 模型原理与优点

随机森林（Random Forest）是一种基于决策树的集成学习算法，由Leo Breiman在2001年提出。其基本思想是通过构建多个决策树，并结合它们的预测结果来提高模型的准确性和稳定性。具体来说，随机森林利用了以下关键技术：Bootstrap采样 随机特征选择 决策树训练 集成预测。 优点有**高准确性**：通过结合多个决策树的预测结果，随机森林通常比单个决策树有更高的准确性和稳定性。**抗过拟合**： 由于引入了Bootstrap采样和随机特征选择，随机森林能够有效地降低过拟合风险，具有较强的泛化能力。**处理高维数据**： 随机森林能够处理高维数据，不需要特征选择。它能够自动评估各个特征的重要性。 **处理缺失值**： 随机森林具有处理缺失值的能力，通过替代缺失值或构建多个模型来处理缺失数据。

**注：**定义图像目录和类别，加载预训练的ResNet50模型，数据增强，加载图像和标签，转换为numpy数组，划分训练集和测试集，标准化特征，以上步骤与支持向量机（SVM）一样，这里不做赘述。

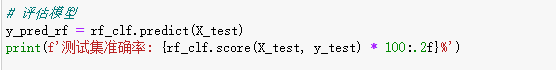
**使用随机森林分类**



 导入随机森林分类器RandomForestClassifier。

 创建一个随机森林分类器rf\_clf，设置树的数量为100，并使用训练数据进行训练。

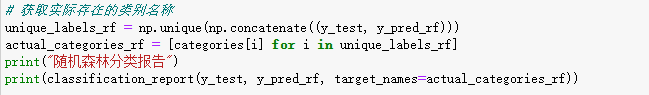
**评估模型**



 使用随机森林分类器预测测试集的类别。

 输出测试集准确率。

**获取实际存在的类别名称**

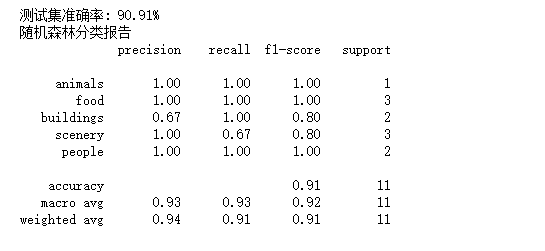


 获取实际存在的类别标签。

 确认分类报告中的类别名称。

 打印随机森林分类报告，包括精确率、召回率、F1得分和每个类别的支持度。

**运行结果**



# 4. 实验结论分析

## 4.1分类结果比较

通过比较三种分类器的分类报告和测试集准确率，我们可以得出以下结论：

**准确率**：

* 1. SVM：100%
  2. 随机森林：90.91%
  3. CNN：81.82%

SVM的准确率最高，达到了100%，表明在测试集上分类没有错误。随机森林的准确率略低，为90.91%，而CNN的准确率最低，为81.82%。

**精确率、召回率和F1得分**：

* 1. SVM的精确率、召回率和F1得分均为1.00，表现最佳。
  2. 随机森林在大多数类别上的表现也非常好，精确率和召回率较高，但在某些类别（如风景）上表现不佳。
  3. CNN的表现相对较差，在某些类别上（如动物）精确率和召回率较差，说明模型在这些类别上的识别能力较弱。

## 4.2结论

实验结果表明，在当前的数据集上，支持向量机（SVM）的表现最好，其测试集准确率达到了100%，而随机森林（Random Forest）和卷积神经网络（CNN）的测试集准确率分别为90.91%和81.82%。这可能是由于以下几个原因：

1. **数据增强**：使用ImageDataGenerator进行了多种数据增强，增加了模型的泛化能力。
2. **预训练模型**：使用ResNet50作为特征提取器，提高了分类器的性能。
3. **数据集大小**：虽然数据集较小，但通过数据增强和特征提取，提高了模型的泛化能力。

## 4.3总结

通过本次实验，我们成功地使用三种不同的分类算法（卷积神经网络、支持向量机和随机森林）对五种类别的图像进行了分类。实验的主要步骤包括数据收集、数据清洗、特征提取、模型训练和评估。以下是实验的详细总结：

#### 数据收集与清洗

**1.数据收集**：

* 1. 使用手机搜集了五种类别的图像，包括动物、食物、建筑物、风景和人物。
  2. 确保每个类别至少有50张不同的图像，以保证数据集的多样性和代表性。

**2.数据清洗**：

* 1. 使用OpenCV库加载图像，并将其调整为统一大小。
  2. 应用数据增强技术（旋转、平移、剪切、缩放和水平翻转）以增加训练数据的多样性。

**3.特征提取**：

* 1. 使用预训练的ResNet50模型提取图像特征。ResNet50模型不包括顶部的全连接层，仅保留卷积层，以提取高质量的特征。

#### 模型训练与评估

**1.卷积神经网络（CNN）**：

* 1. 构建了一个包含卷积层、最大池化层和全连接层的CNN模型，并使用数据增强后的图像数据进行训练。
  2. 评估结果：测试集准确率为81.82%。

**2.支持向量机（SVM）**：

* 1. 使用ResNet50提取的图像特征作为输入，训练了一个线性核的SVM分类器。
  2. 评估结果：测试集准确率为100%。

**3.随机森林（Random Forest）**：

* 1. 使用ResNet50提取的图像特征作为输入，训练了一个包含100棵树的随机森林分类器。
  2. 评估结果：测试集准确率为90.91%。

**分类器比较**

 支持向量机（SVM）在我们的实验中表现最佳，其测试集准确率达到了100%，在所有分类指标上均优于其他分类器。这表明，SVM在处理这种小规模、多类别图像分类问题时具有良好的性能。

 随机森林（Random Forest）的表现也很接近SVM，测试集准确率为90.91%。它在大多数类别上表现较好，但在某些类别上的识别能力稍有不足。

 卷积神经网络（CNN）的表现相对较差，测试集准确率为81.82%。这可能是由于我们的数据集较小，而CNN通常需要大量数据才能充分训练和泛化。

 未来工作可以考虑增加数据集的规模、改进数据增强技术和调整模型结构，以进一步提高分类器的性能。