**2020机器学习**

**印花疵点分类**

**组员：**

须子逸 20210240350

李帅 20210240207

李锦程

朱柏霖

郭露

# 1 任务描述

基于机器学习模型对印花疵点数据集中标签为1、2、14的数据以及标签为1、2、4、14、20的数据分别做3分类及5分类任务；

利用ModelArts平台进行模型训练任务。

# 2 数据集介绍

# 3 数据预处理

数据预处理的目标是去除数据中的冗余信息，突出数据中的重要信息，即瑕疵。在印花疵点数据集中，已经注明了瑕疵所在的位置，并且可以作为已知信息用于分类任务，因此本文首先对图片进行裁剪，去除瑕疵位置以外的像素点，从而使分类模型专注于瑕疵所在的位置。

由于Bbox的尺寸各不相同，因此在裁剪之后需要对图片进行resize使得数据的尺寸一致，本文分别尝试了缩放和padding两种resize方式，其中缩放可以减少数据为0的像素点，使得瑕疵填满整张图片，但是会因此失去瑕疵区域的尺寸信息，而padding保留了尺寸信息。根据实验结果发现用padding的进行resize后模型的准确率较高，

由于数据集中某些类别的数量过少，因此我们通过旋转、翻转、改变亮度、增加噪点等方式生成新的图片，进行数据增强，扩充了数据集。实验结果表明数据增强可以有效提高准确率。

在印花疵点数据集中，对每个数据提供了瑕疵图和原图一对图片，为了去除背景冗余信息，本文使用了图片差分的方式，通过将瑕疵图和原图进行差分，去除瑕疵图中属于背景部分的像素，只保留瑕疵信息。

最后，本文也尝试了在差分前对瑕疵图和原图进行池化从而提取特征，本文分别对最大池化和平均池化后的数据集进行训练，通过实验得到了效果最好的预处理方式。

## 3.1 裁剪

在裁剪操作中，根据标注中的Bbox信息对图片对进行裁剪，去除瑕疵区域以外的像素，如图3-1所示：

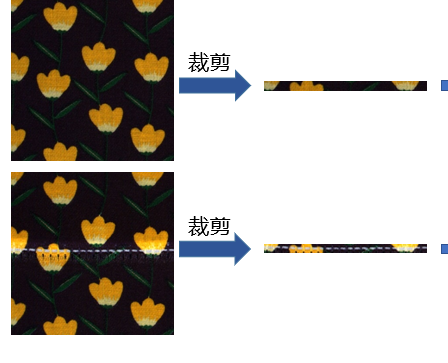


图3-1 裁剪

## 3.2 数据增强

在对图片进行裁剪之后，我们通过旋转随机角度，翻转，改变亮度等操作进行数据扩充。将每一个训练集中类别的图片扩充为600张，从而平衡样本数量。

## 3.3 池化

由于数据集中的图片对没有完全对其，我们尝试对图片进行池化后再进行差分，实验结果表明。。。

## 3.4 差分

对图片对进行差分后，去除了部分背景信息，只保留存在瑕疵的像素点，实验结果表明差分对分类准确率有明显提升。如图3-2所示：

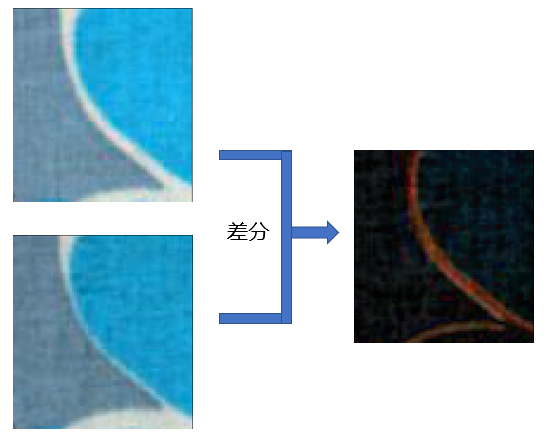


图3-2 差分

## 3.5 resize

将差分后的图片resize成统一大小有两种方法：缩放和padding，缩放通过插值将原图变形成为新的大小，该方法可以使得有效像素填充整个输入向量，但同时会失去图片原来的尺寸信息，如图3-3所示：

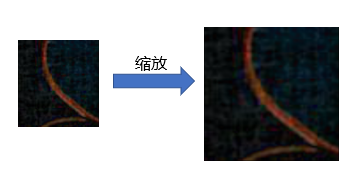


图3-3 缩放

而padding通过在周围填充黑色像素点从而适应形状，该方法保留了图片原来的尺寸信息，并且纹理不会发生变形。如图3-4所示：

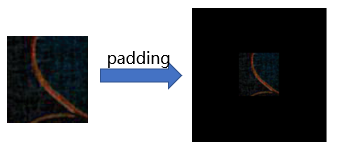


图3-4 padding

# 4 模型介绍及实验

## 4.1 决策树

# 4.2 SVM

# 4.3 softmax回归

**环境**

Tensorflow1.8、python3.6、ModelArts

**数据集**

64\*64\*3

**算法介绍**

当考虑多分类的问题时，，可以使用​个参数，表示每个类别结果的可能性。显然，。经过一系列推导可以得到下面的softmax函数，，其中，。进一步推导，可得



假设函数的输出为



**算法实现**

softmax\_regression.py

**训练结果**

对不同的数据集，上述代码的参数有所不同，最后的结果如下



# 4.4 简单CNN

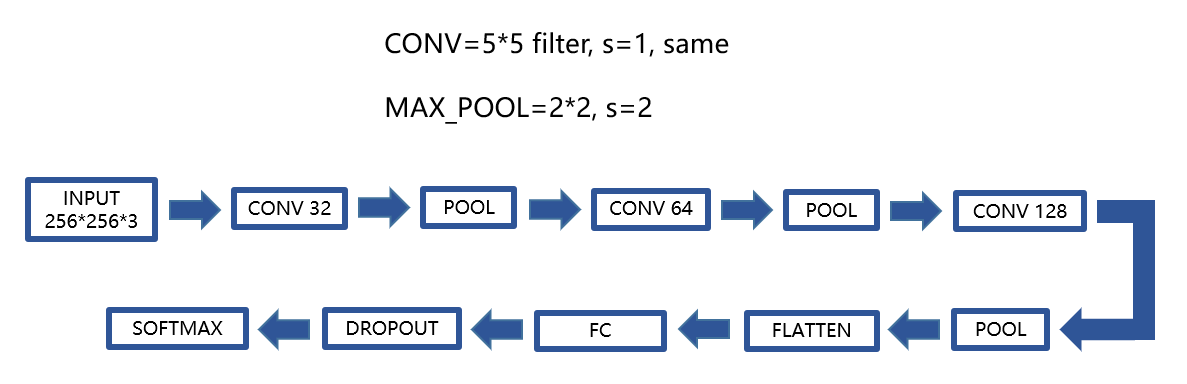
**环境**

Tensorflow1.8、python3.6、ModelArts

**数据集**

256\*256\*3

**算法介绍**



**算法实现**

simple\_cnn.py

**训练结果**

对不同的数据集，上述代码的参数有所不同，最后的结果如下



对task1的4\_diff\_padding\_augment，loss的变化如下



# 4.5 ResNet-18

## 4.6 ResNet-50 on ModelArts

使用ModelArts平台提供的ResNet-50模型进行训练。

## 4.6.1 数据上传

ModelArts的数据导入的文件存放方式如图4-1所示：

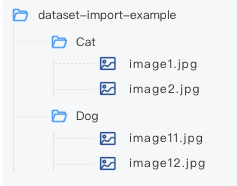


图4-1 ModelArts数据导入文件存放方式示意图

根据指定文件存放方式上船数据，如图4-2所示：



图4-2 上传图片数据到OBS桶

## 4.6.2 创建并发布数据集

# 5 实验结果及分析

## 5.x 结果对比

三分类：

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 预处理方式 | | | | | | 分类模型 | | | | | | |
| 差分 | 最大池化 | 平均池化 | 缩放 | 填充 | 数据增强 | 决策树 | 逻辑回归 | SVM | CNN | ResNet-18 | ResNet-50 |
|  |  |  |  | √ |  |  |  |  |  |  |  |
| √ |  |  | √ |  |  |  |  |  |  |  |  |
| √ |  |  |  | √ |  |  |  |  |  |  |  |
| √ |  |  |  | √ | √ |  |  |  |  |  |  |
| √ | √ |  |  | √ | √ |  |  |  |  |  |  |
| √ |  | √ |  | √ | √ |  |  |  |  |  |  |

五分类：

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 预处理方式 | | | | | | 分类模型 | | | | | | |
| 差分 | 最大池化 | 平均池化 | 缩放 | 填充 | 数据增强 | 决策树 | 逻辑回归 | SVM | CNN | ResNet-18 | ResNet-50 |
|  |  |  |  | √ |  |  |  |  |  |  |  |
| √ |  |  | √ |  |  |  |  |  |  |  |  |
| √ |  |  |  | √ |  |  |  |  |  |  |  |
| √ |  |  |  | √ | √ |  |  |  |  |  |  |
| √ | √ |  |  | √ | √ |  |  |  |  |  |  |
| √ |  | √ |  | √ | √ |  |  |  |  |  |  |