

课程设计报告

（ 2021秋季学期 ）

|  |  |
| --- | --- |
| 课程名称 | 机器学习课程设计 |
| **项目名称** | **Statoil/C-CORE Iceberg Classifier Challenge** |

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 肖正皇 | 学号 | 20190441125 |

|  |  |  |  |
| --- | --- | --- | --- |
| 专业 | 软件工程 | 班级 | 19软卓01班 |

|  |  |  |  |
| --- | --- | --- | --- |
| 地点 | 雨母楼 | 教师 | 刘永彬 |

1. **项目概述及计划**

赛题的背景：在加拿大的东海岸经常会有漂流的冰山，这对航行在该海域的船舶造成了很大的威胁。挪威国家石油公司(Statoil)是一家在全球运营的国际能源公司，该公司曾与C-CORE等公司合作，C-CORE基于其卫星数据和计算机视觉技术建立了一个监控系统。Statoil发布该赛题的目的是希望利用机器学习的技术，更准确的及早发现和识别出威胁船舶航行的冰山。

Kaggle 冰山分类挑战赛的任务目标是建立一个图像分类器，将输入卫星图像划分为冰山或者船，这项工作在能源勘探领域尤为重要，它让我们能够识别并且避开类似浮冰这样的威胁。对于人类仅凭肉眼准确地区分图片类别都是十分困难。我认为这是一个很好的契机——测试深度学习和计算机视觉能够做到人类做不到的事情。

1. **问题描述**

漂流冰山对加拿大东海岸近海等地区的航行和活动构成威胁。

目前，许多机构和公司使用空中侦察和岸基支持来监测环境条件和评估冰山的风险。然而，在天气特别恶劣的偏远地区，这些方法是不可行的，唯一可行的监测选择是通过卫星。

挪威国家石油公司是一家在全球运营的国际能源公司，与C-CORE等公司密切合作。C-CORE已经使用卫星数据超过30年，并且已经建立了一个基于计算机视觉的监视系统。为了确保操作安全和高效，挪威国家石油公司有兴趣获得一个新的视角，了解如何使用机器学习尽早更准确地检测和识别威胁冰山。

在这场比赛中，你面临的挑战是建立一种算法，自动识别遥感目标是船只还是冰山。所做的改进将有助于降低维护安全工作条件的成本。

1. **模型建模**

这里我们使用TenSorFlow的高级API来搭建卷积神经网络。

我们使用“tf.keras.Sequential()”创建一个序贯模型，序贯模型是多个网络层的线性堆叠，我们使用“tf.keras.Sequential().add()”方法逐层添加网络结构。这里的卷积层我们使用了128个大小为3x3的卷积核，使用了relu激活函数。

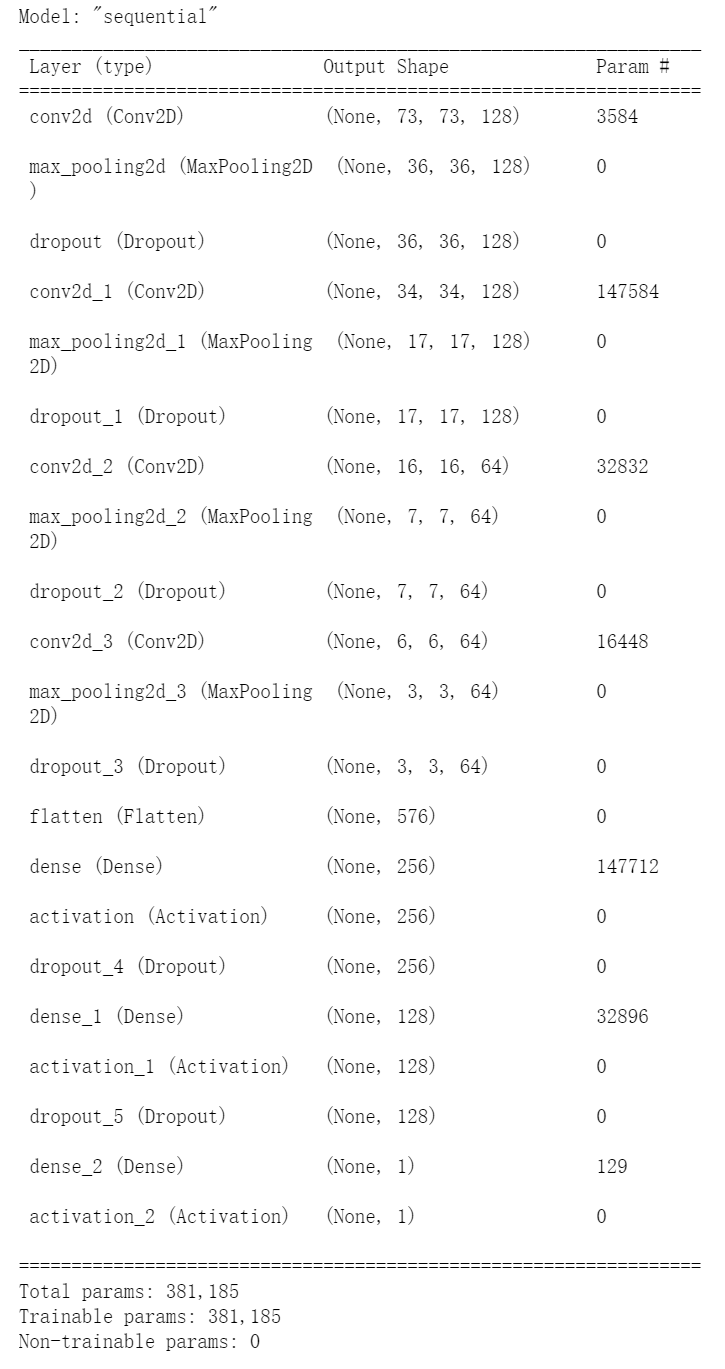
在卷积层后面是一个池化层，采用最大池化，池化窗口的大小为3x3，横向和纵向的步长都为2。在池化层的后面我们使用了Dropout，丢弃了20%的神经元，防止参数过多导致过拟合。

接下来是三个类似的卷积块。使用“Flatten()”将前一层网络的输出转换为了一维的数据，这是为了接下来的全连接操作。第一个全连接层有256个神经元，全连接层后面接relu激活函数，同样使用了Dropout。

由于是二分类问题, 我们使用了一个只有一个神经元的全连接层，并使用了Sigmoid激活函数，得到最终的输出。



训练开始后打印出模型的概况信息如下：



1. **实验分析**

**4.1 数据分析**

赛题提供了两个数据文件“train.json”和“test.json”，其中“test.json”是比赛中用来对模型进行评分的，没有类标。该数据集中有1604个打包过的训练数据，单个样本的数据格式如下

|  |  |
| --- | --- |
| **字段名** | **字段说明** |
| id | 图像的id |
| band\_1,band\_2 | 卫星图像数据，band\_1,band\_2是以特定入射角下不同极化方式产生的雷达后向散射为特征的信号，分别对应HH（水平发射/水平接收）和HV（水平发射/垂直接收）两种计划方式的数据。 |
| inc\_angle | 获得该数据时的入射角度。该字段部分缺少的数据，标记为”na”。 |
| Is\_iceberg | 类标，0：船只；1：冰山 |

比赛的数据有两个特别有意思的地方：

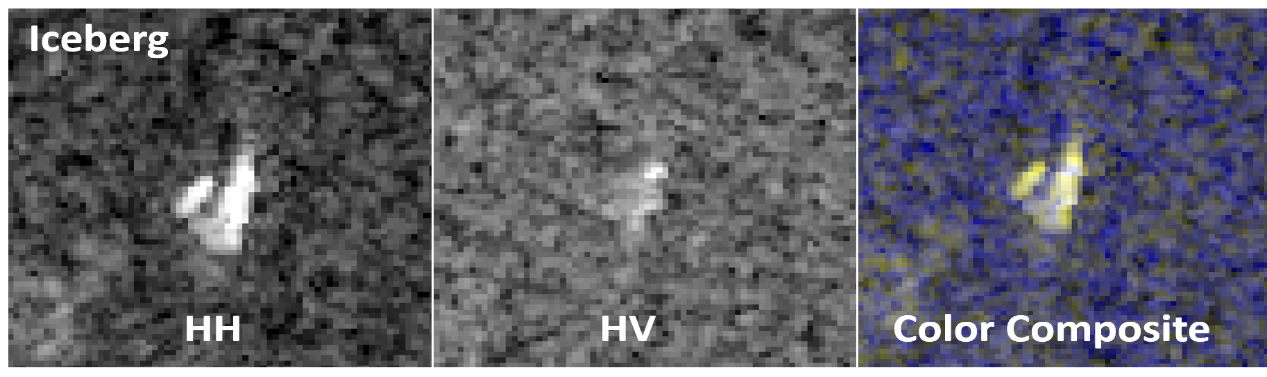
首先，数据集规模相当小，训练集中只有 1604 张图片。这样一来，从硬件的角度来说，参与竞赛的门槛很低。然而，另一方面，使用有限的数据完成图像分类任务又是十分困难的。

其次，如果用肉眼去观察这些图片，它们看上去就像电视屏幕上出现的「雪花」

图——一堆乱七八糟的噪点。人类用肉眼完全不可能看出来哪些图像是船，哪些

图像是冰山。

我们将数据可视化后进行观察，如下图所示。图像上方是冰山图像的可视化效果，三幅图分别对应“HH”计划方式、“HV”计划方式以及两者结合后的数据。图像下方是船只图像的可视化效果。

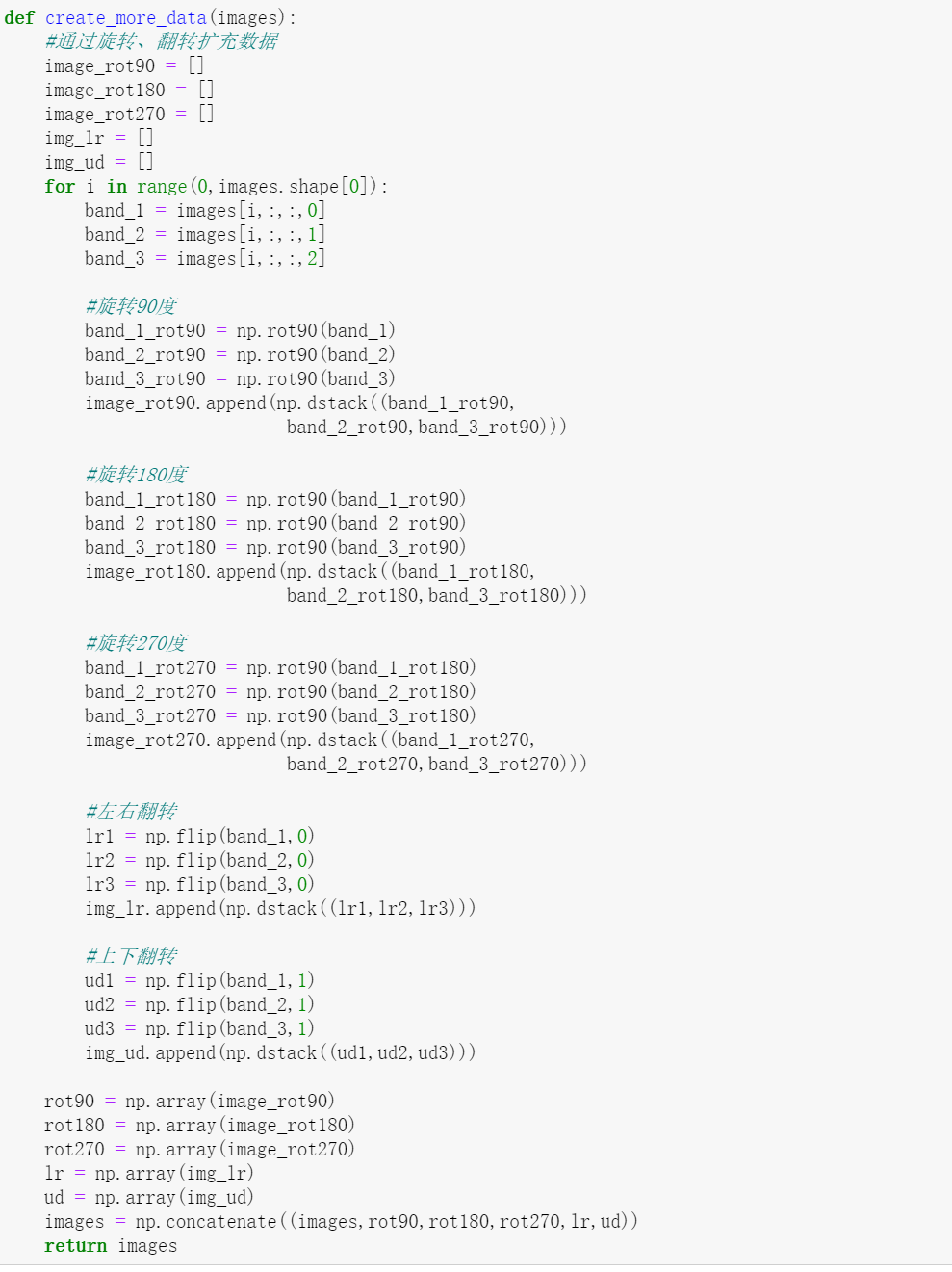
由于数据集规模比较小，所以必须对数据适当增强。

我们对样本数据进行了处理，除了原有的“band\_1”和“band\_2”，我们增加了“band\_3”，band\_3=band\_1+band\_2。最后我们使用numpy的“dstack”将三种数据进行堆叠，因此我们单个样本的数据维度为75x75x3

调用“create\_more\_data”函数对训练数据进行扩充，之后是对训练集的类标数据进行扩充，因为“create\_more\_data”函数将训练数据扩充为了原来的6倍，因此这里对应的也要将类标扩充为原来的6倍。



通过对图像进行旋转和翻转来扩充数据集，虽然旋转前后的图像是同一张，但是由于特征的位置发生了变化，因此对于模型来说就是不同的数据，旋转或翻转操作是扩充图像数据集的一个简单有效的方法。定义了5个列表，用来保存扩充的数据集，对应的操作分别是逆时针旋转90度、逆时针旋转180度、逆时针旋转270度、左右翻转和上下翻转。具体实现如下：



使用numpy的“rot90”和“flip”函数对图像进行旋转和翻转操作。“flip”函数的第二个参数控制翻转的方式，“0”为左右翻转，“1”为上下翻转。使用numpy的“concatenate”函数将扩充的数据预原数据进行拼接。

**4.2 实验环境**

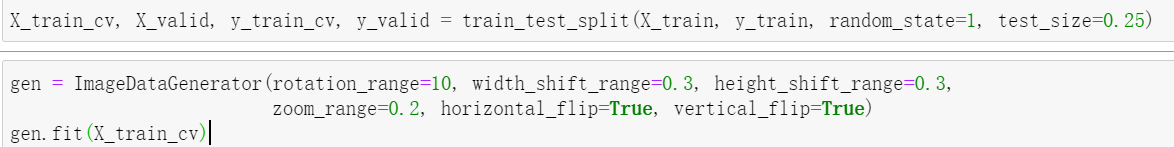
**工具和程序库：**TensorFlow 、 Keras、numpy、pandas等

**Windows10操作系统**

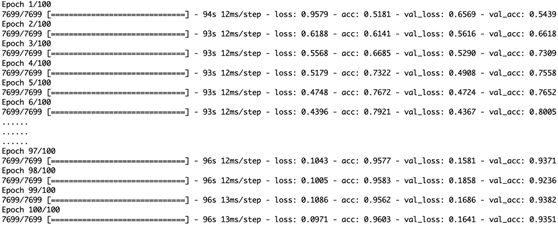
**Anaconda** Python发行版本

**4.3 实验结果**

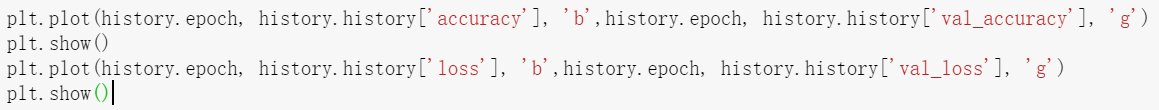
**4.3.1、读取数据，并训练模型**

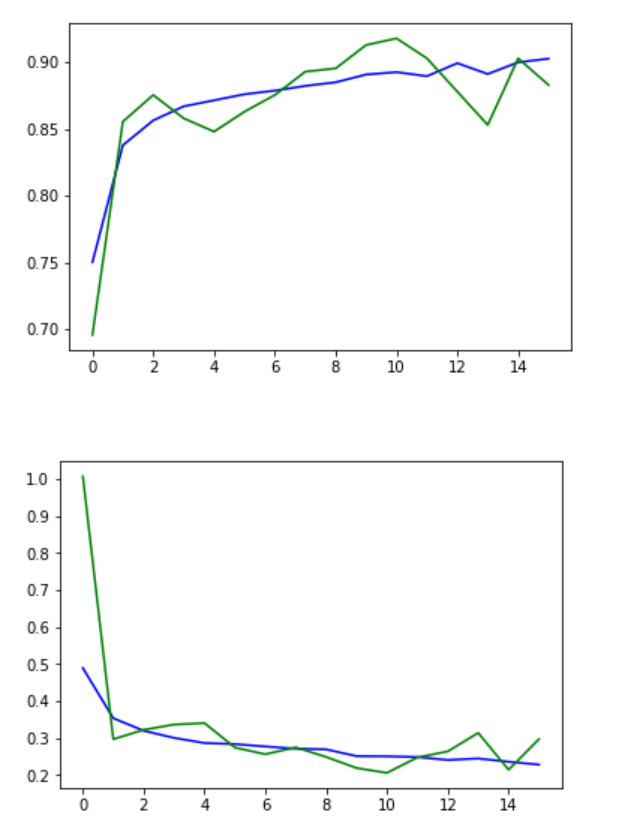


模型的训练过程和结果如下：



**4.3.2、loss和accuracy曲线**

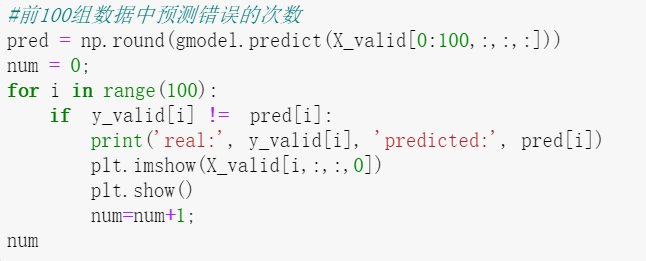




损失值（loss）值不断减小，证明它正在趋于收敛，准确率（accuracy）值不断增加，证明它的准确率在不断提升

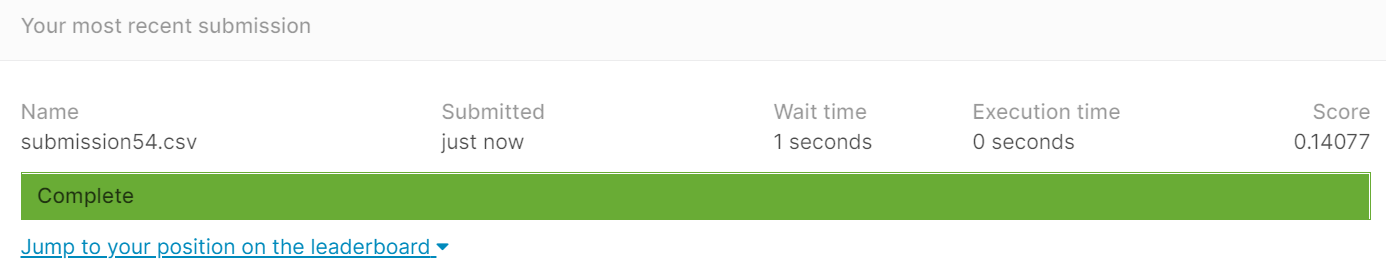
**4.3.3、预测结果**

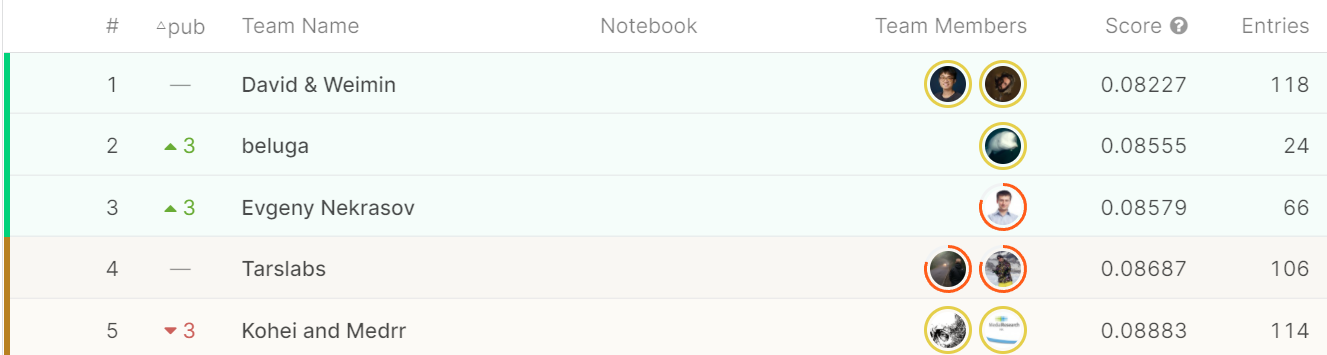
**打印前100组数据中预测错误的图片并统计数量：**

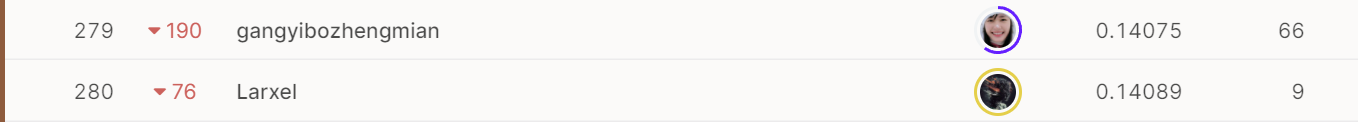


发现最终输出为4，即对于这个二分类任务，我们的模型准确率已经达到了大概96%。

**4.4 实验对比（竞赛排名）**







****

**由于比赛已经结束，故提交不记入排名，只能对成绩进行相对比较**

**最终排名为280/3330**

当然这个排名对我来说已经是很高了，毕竟，我还只是一个小白。然后我肯定是想向更优秀的人学习的。之后我仔细研究了这次比赛第一名的代码，发现真的是自愧不如。他们使用了一个十分庞大的卷积神经网络结构，包含超过 100 个改造过的卷积神经网络和类似于 VGG 的结构，然后他们还使用 greedy blending 策略和两层结合了其他的图像特征的集成学习算法将模型结果融合起来，最后为了防止过拟合将所有的模型都要经过4折交叉验证。

1. **项目总结**

当然这是第一次弄机器学习，通过python打的比赛，说不难能不难吗，本来python就是这学期速成的，很多条条框框都还不太理解，甚至很多语法都还不太记得清楚，然后就赶鸭子上架去打比赛了，然后打开比赛网址一看，喔哦，英文的，翻译一下，然后一脸懵，图像识别什么的，预测分析什么的，甚至还有人脸识别，这些平常感觉高大上的东西怎么就这么个比赛就要用上了，可是完全不会啊，然后看别人公开的代码，就是一堆框架，都不会，甚至很多python包都没见过，然后就开始学习之路，做课设之前再学一门课先，就去网上学习各种会用上的框架，特别是CNN,大概知道是怎么一回事之后，已经是好几天之后的事了，然后就正式开始比赛了。

虽然这次比赛做的确实牵强了一点，但是对于以后如果有机会从事机器学习、深度学习乃至人工智能等方面的研究，绝对是一个很好的铺垫。

在做一个比赛之前，最重要的步骤当属预先进行的探索性分析，从而对数据有更好的了解。探索性分析的结果表明，有一个图像数据之外的特征中非常重要，它能够帮助消除数据中大量的噪声。在我看来，我们在平常比赛或者学习等问题中最容易忽视的步骤之一，就是需要事先理解数据，并且利用这些知识帮助我们做出最佳设计选择。

现成的算法如今更加易于获得和引用，我们往往会不假思索、简单粗暴地将这些算法应用于待解决的问题上。然而我们却没有真正想清楚这些算法是不是这一任务的最佳选择，或者没有想清楚在训练之前或之后是否需要对数据进行一些适当的处理工作。