**深度学习系统第六次作业**

ZY2206117 黄海浪

## 实验介绍

### 实验内容

本次实验体验常用的模型轻量化方法（量化、剪枝、蒸馏），感受各方法之间的区别与联系。

### 实验步骤

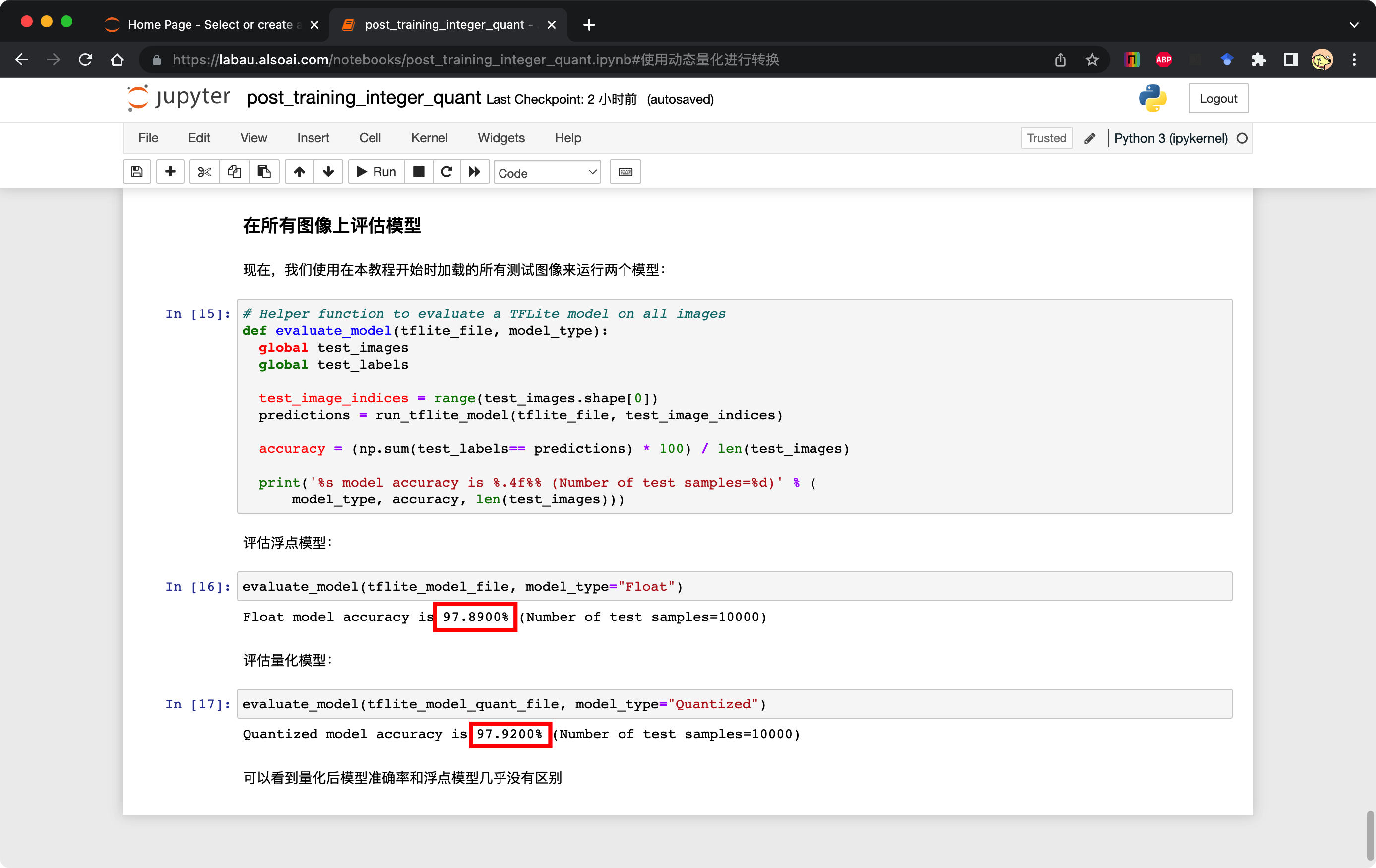
1. 量化部分使用MNIST 数据集从头开始训练一个模型、将其转换为TensorFlow Lite 文件，并使用训练后量化的方法对其进行量化。最后检查转换后模型的准确率并将其与原始浮点模型进行比较。

2. 剪枝部分使用MNIST 数据集从头开始训练一个模型，使用剪枝API 微调模型并查看模型准确率，最后结合剪枝和后训练量化，生成一个是原模型10倍小的TFLite 模型。

3. 蒸馏部分包含教师模型和学生模型，首先使用MNIST 数据集训练教师模型，接着使用训练后的教师模型指导学生模型的训练过程，最后比较学生模型有无指导情况下的准确率。

## 模型量化

使用全整数量化



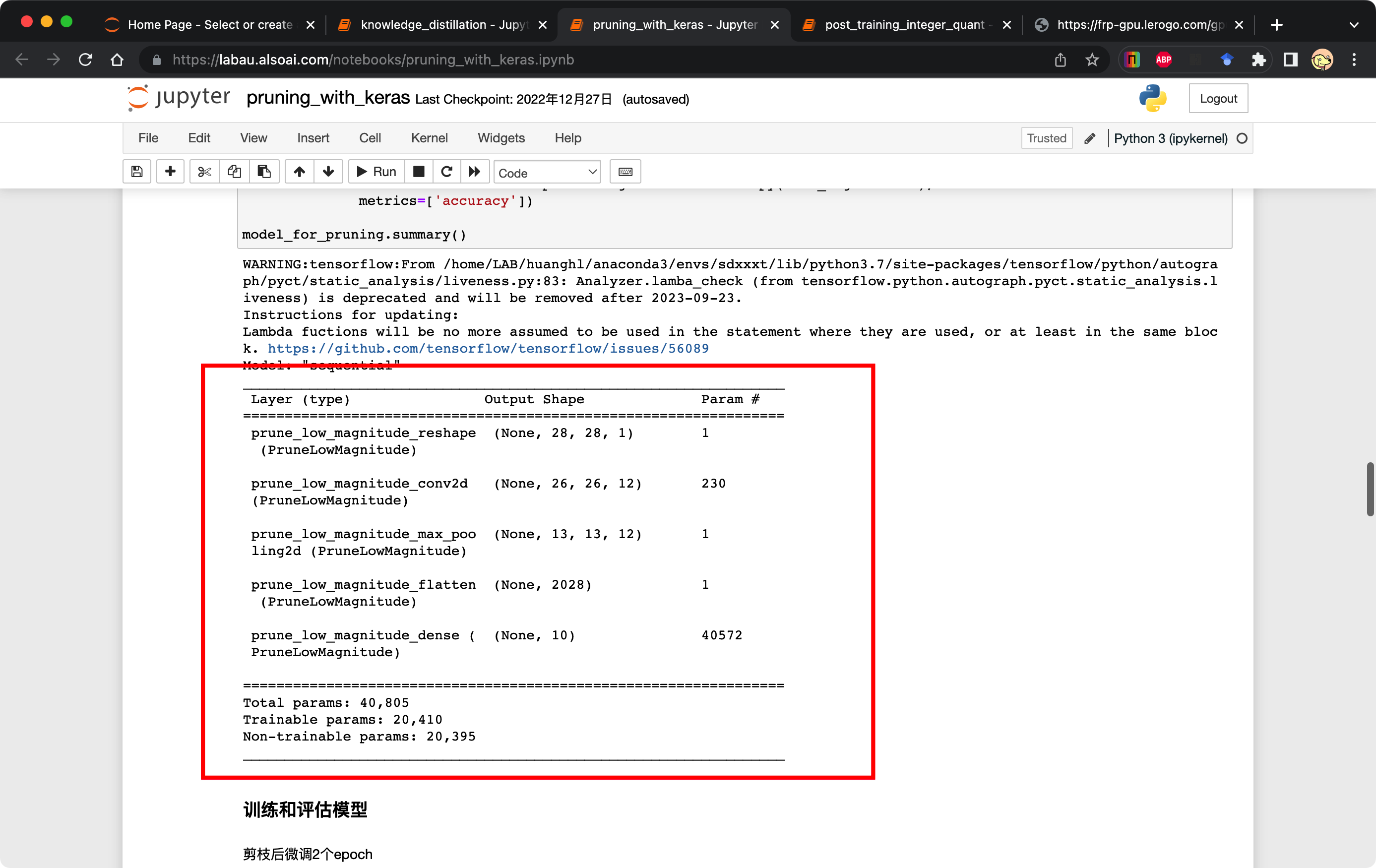
转化前准确率：97.8900%

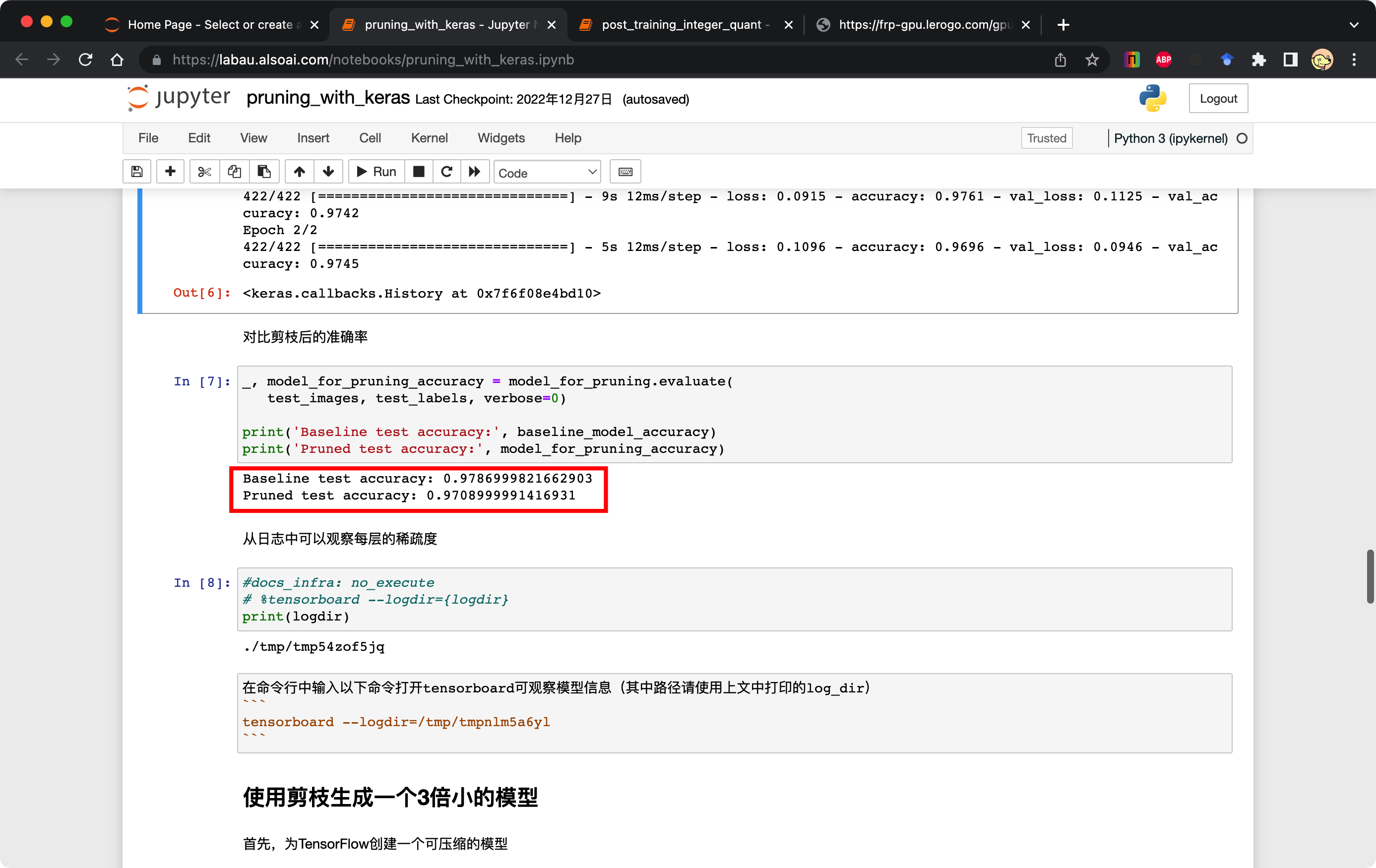
转化后准确率：97.9200%

从结果可以看到，量化后模型的准确率并没有下降，而是和float模型保持相对稳定的准确率。

## 模型剪枝

剪枝API微调模型



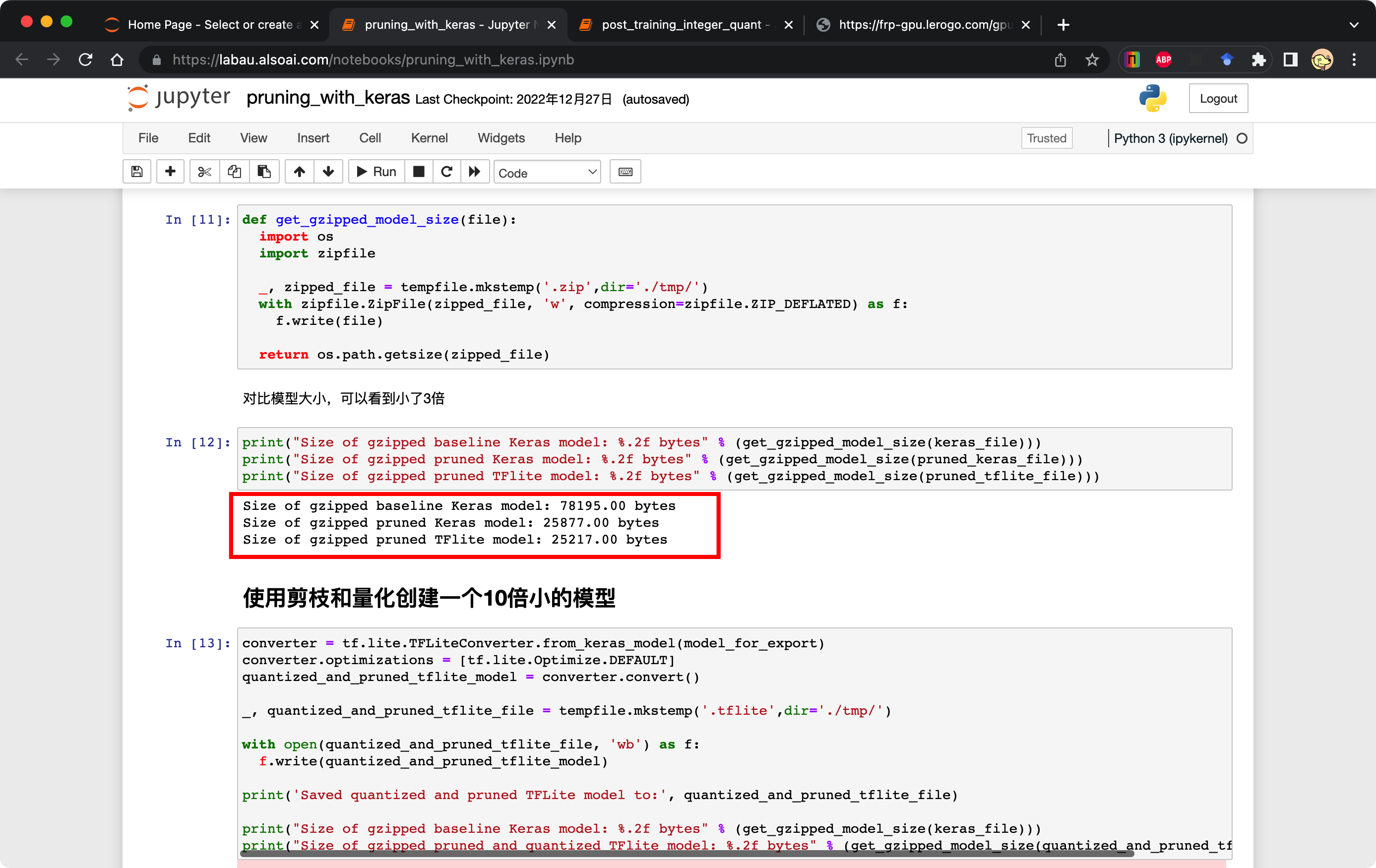


baseline: 0.9786999821662903

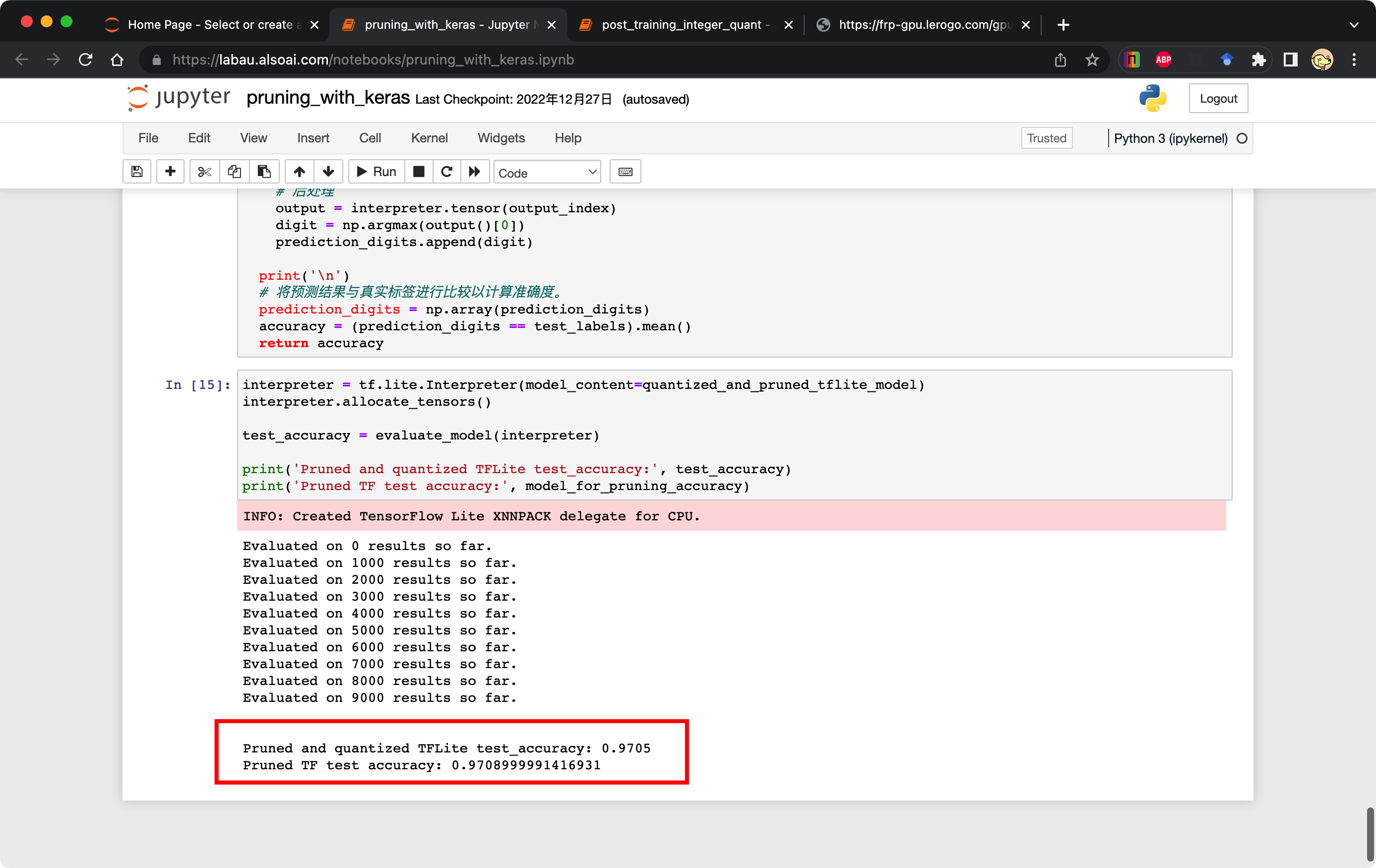
剪枝后：0.9708999991416931

从结果可以看到，模型剪枝后，比起baseline结果并无较大差距，性能几乎一致。

通过剪枝将TF和TFLite模型大小缩减为原来的3倍



结合剪枝和后训练量化，生成一个原模型10倍小的TFLite模型



剪枝和量化后结果：0.9705

baseline Keras model: 78195.00 bytes

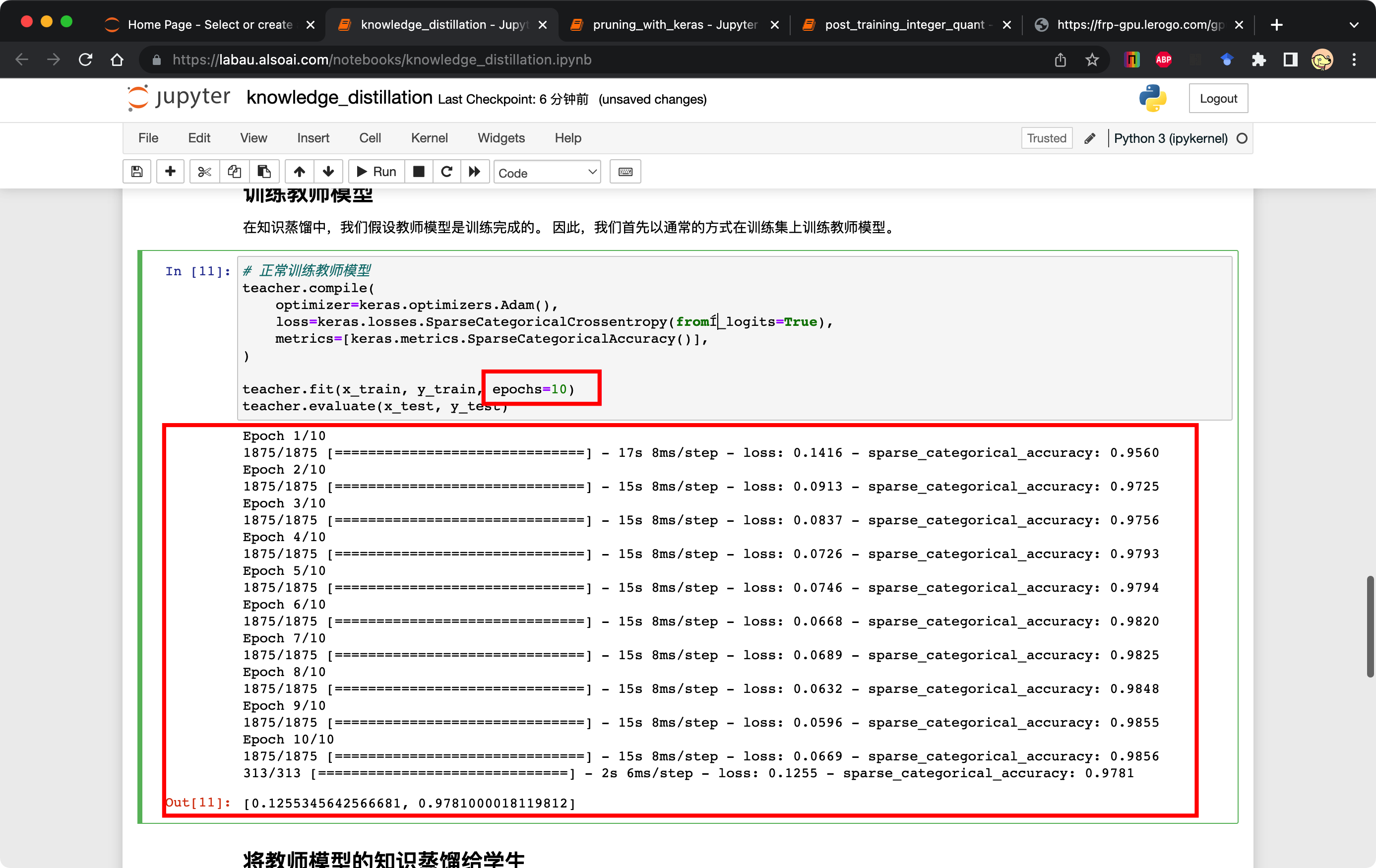
pruned and quantized TFlite model: 8130.00 bytes

从结果可以看到，生成原模型10倍小的TFLite模型性能与baseline相比并没有多大差别，但是模型大小可以大大减小，减少了资源占用和计算量。

## 模型蒸馏

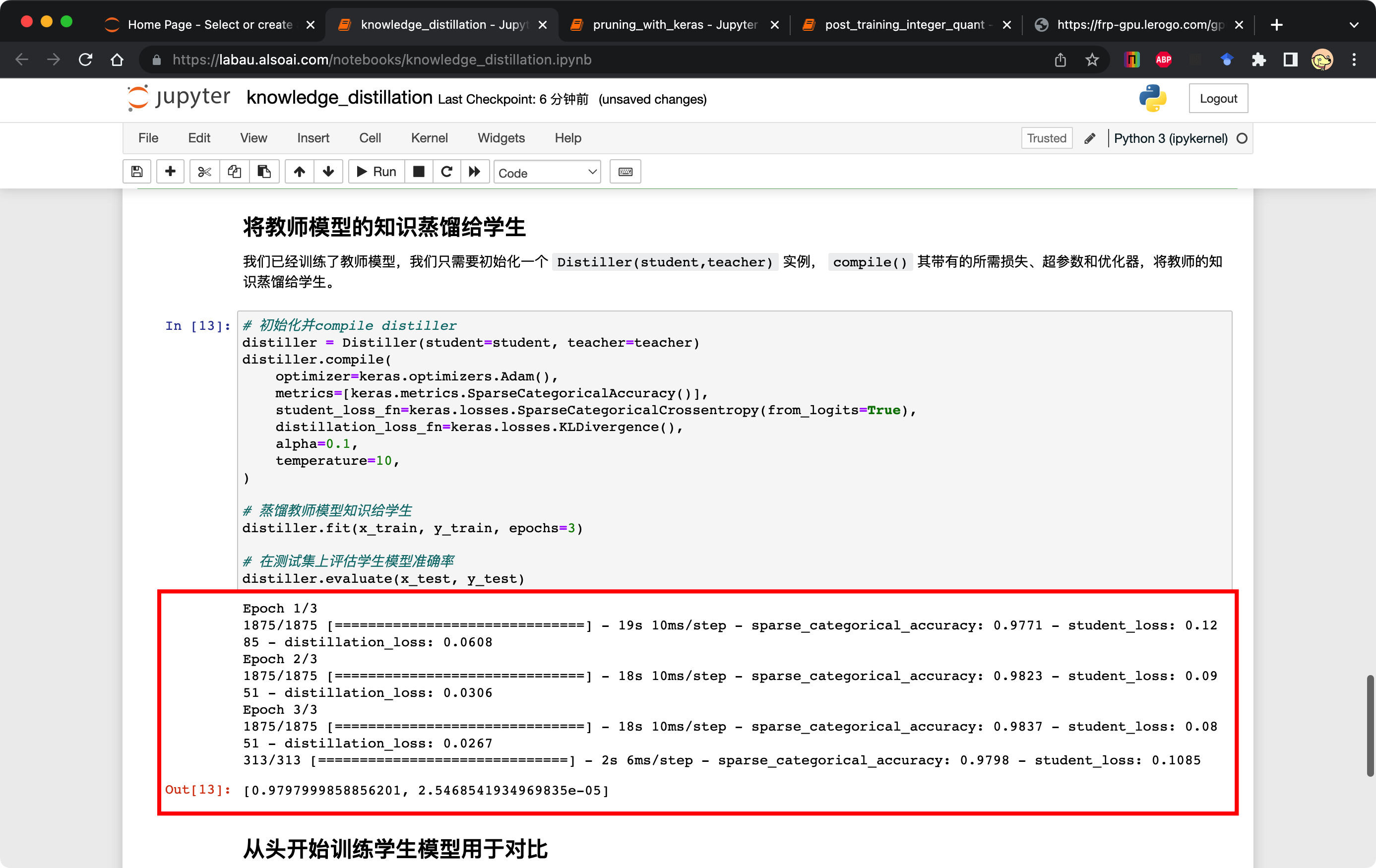
老师模型：

老师模型为了得到更高的准确率，实验将代码的epochs改为10，进行训练，准确率为0.9781



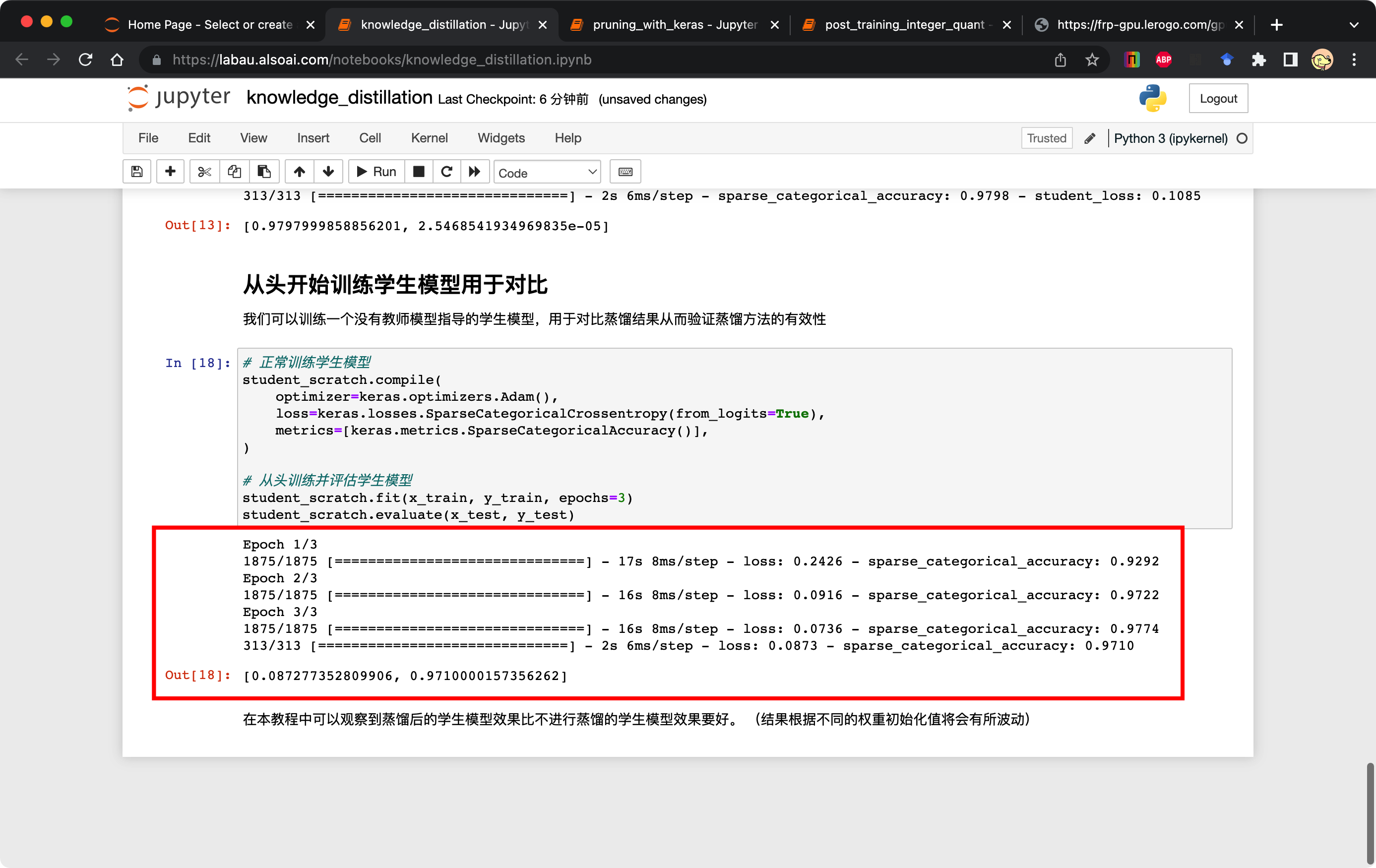
蒸馏模型：

蒸馏模型训练了3个epoch，准确率为0.9798



单独训练学生模型：

单独训练学生也为3个epoch，准确率为0.9710



从结果可以看到，通过教师蒸馏得到的模型性能比单独训练模型效果要好。但是由于数据集比较简单，差距并不是很大。不过该方法是很好且常用的方法。

## 小结

网络剪枝的主要思想就是将权重矩阵中相对“不重要”的权值剔除，然后再重新fine tune 网络进行微调。随机剪枝方法对硬件非常不友好，往往在硬件实现的过程中不一定能够很好地对网络起到加速和压缩的效果，而且一些标准化模块conv2d可能都不能用。后来就使用成块出现的结构化剪枝，包括Filter Pruning，梯度Pruning等方法。量化则可以分为低比特量化、总体训练加速量化和分布式训练梯度量化，本次实验使用的是低比特量化。知识蒸馏将深度网络中所学到的知识转移到另一个相对简单的网络中去，这种方法对于没有标签的数据集也可以拿来训练。另外大模型的输出相较于GTlabel来说包含更多信息，如类间距和类内方差等，所以这也可以作为数据增广的一种手段。

网络量化和剪枝是属于后端压缩的方法，这种方法不可逆，而蒸馏的方法属于前端压缩的方法，该方法通过复杂网络去监督简单网络对模型进行简化，是可逆的。这些方法的目的都是最大程度的减小模型复杂度，减少模型存储需要的空间，也致力于加速模型的训练和推测。