机器学习（进阶）毕业项目开题报告

1 选题背景

该项目是 StateFarm 公司在 Kaggle 上发起的一个竞赛项目[1]，目的是得到一个模型，用来检测驾驶员在驾驶过程中是否走神。

根据疾病预防控制中心（CDC）的研究成果[2]，五分之一的汽车事故是由于驾驶过程中分心导致的，这意味着每年有约 425000 人因此受伤，3000 人丧命。典型的驾驶分心行为主要包括视野离开了路面、双手离开方向盘或注意力不在驾驶上。StateFarm 公司希望通过仪表盘上的摄像头自动检测驾驶过程中的分心行为，进而提醒驾驶员集中注意。

2 问题陈述

该项目要求给定一张 2D 图片作为输入数据，输出 10 种驾驶状态的概率。本质上是监督学习中的分类问题，提供的作为训练的数据集已经做好了分类标记。构建合适的模型，利用交叉验证对训练集中的数据进行训练。训练完的模型用来预测测试集，将测试集的结果保存成 csv 文件，每一行记录图片名和 10 种行为的预测概率。

3 数据集和输入

此数据集可以从 kaggle 上下载，包含三个文件：1）imgs.zip，train/test 图片文件的压缩包；2）sample_submission.csv，给出了提交文件的格式模板；3）driver_imgs_list.csv，训练集中驾驶员 ID、图片编号和标签的对应关系。

数据集为彩色图片，已经划分为训练集和测试集，图片内容是驾驶员在驾驶过程中的不同行为。训练集分为 10 类，标签为 c0 ~ c9，分别表示：

c0：安全驾驶

c1：右手打字

c2：右手打电话

c3：左手打字

c4：左手打电话

c5：调收音机

c6：喝饮料

c7：拿后面的东西

c8：整理头发和化妆

c9：和其他乘客说话

同时，训练集提供了驾驶员编号、包含驾驶行为的图片编号和行为类别标签的对应关系；测试集为待预测图片的集合。

对数据中的训练集进行简单的分析可知，训练集中的数据共 22424 个，驾驶员编号共 26 个，c0~c9 每个标签对应的数据量比较均衡，均在 1900~2500 之间。按照驾驶员分类，每个驾驶员类别中都包含 c0~c9 所有标签，且大部分类别中的标签分布与整个数据集中的标签分布相似。为了能更好的泛化，训练集和验证集可以按照驾驶员 ID 进行分割。

4 解决方案

项目目的是通过学习训练集的图片数据，可以对测试集的图片进行分类，是一个基于监督学习的多分类识别问题。对于图片分类问题，卷积神经网络（CNN）[3]可以说是目前公认的最有效的方法。从 2012 年 AlexNet 在 ImageNet 的 LSVRC（大规模视觉识别挑战赛）中以远超第二名的成绩夺得冠军，CNN 开始重新回到大众视野，此后的每年 LSVRC 大赛都是卷积神经网络的天下，并发展出更多的模型，如 VGGNet（2014.09）、GoogLeNet（2014.09）、ResNet（2015.12）、Inception v3(2015.12)、Inception v4(2016.02)、Xception(2016.10)和 ResNeXt(2016.11)。

一般这些模型已经在大数据集上进行了预训练，可以使用这些预训练过的模型作为初始模型或者特征提取器，进行迁移学习。该项目可以使用 opencv 进行图片预处理；scikit-learn 做交叉验证和评价指标的计算；keras 用来创建神经网络模型，是对 tensorflow 的高级封装。选取三个模型 VGG16[5]、ResNet50[6]和 InceptionResNetV2[7]分别进行 finetune。

VGG16 是由 AlexNet 演化而来，分成 5 层（组），每层有 2~3 个 convolution 层，每层之间用 max pooling 层分开，最后再加三层 fully connected 层，filter 的大小为 3x3。默认输入尺寸是 224 x 224。ResNet 即深度残差网络，其出现是为了解决网络深度增加性能下降的问题，这里选用 ResNet50 深度为 50 层，默认输入为 224x224。InceptionResNetV2 由 Inception v3 模型演化而来，同时结合了残差网络的思想，默认输入尺寸为 299x299。

5 基准模型

kaggle 上有两个榜单，Public Leaderboard 使用 31%的数据集计算 log loss，Private Leaderboard 使用剩下的 69%计算 logloss。计算结果越小排名越高，最终排名以 Private Leaderboard 为准。kaggle 上有 1440 只队伍提交了有效结果，本项目的目标是进入 kaggle 排行榜前 10%，即第 144 名，对应的 Private Leaderboard 上的 logloss 为 0.25634。

6 评价指标

评价指标分为两部分，一部分是预测结果的准确性，另一部分是训练和预测的时长。

准确性采用 multi-class logarithmic loss 评价，是对所有测试集预测概率取对数后加权得到的一个值，值越接近于 0 表示模型预测越。准确，具体方程为：

$$\text{logloss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\, log(p_{ij}) \ ,$$

其中，N是测试集中图片的数量；M是分类标签的数量；log是自然对数；当输入i属于j类时y_{ij}为 1，否则为 0；p_{ij}为输入i属于j类时的预测概率。对于单个确定的输入图片，应该输出 10 种类别对应的预测概率，为避免 log 函数取极端值，p_{ij} 取值为 $\max(\min(\mathrm{p}, 1-10^{-15}), 10^{-15})$。

7 项目设计

首先对数据进行预处理，将训练集中的数据按照驾驶员 ID 分类读取，读取图片的同时记录对应的分类，使图片和分类一一对应。对图片中的数据进行转换，存储为 numpy array 结构，对于 VGG16 和 ResNet50 模型，将图片缩放成 224x224 的尺寸；对于 InceptionResNetV2 模型，缩放为 299x299 尺寸。然后对图像数据进行零均值化处理，对标签数据进行独热编码。

然后使用 Keras 自带的模型进行迁移学习，分类器从 softmax-1000 classifier 改为 softmax-10。使用在 ImageNet 上预训练的 weights 作为卷积层的 weights，微调全连接层的 weights，等到全连接层趋近收敛，再微调感兴趣的层。训练过程中采用 K 折交叉验证，根据驾驶员 ID 切分训练集和验证集。

训练完成之后，批量读取测试集中的图片，并根据不同模型的需求对图片进行预处理，然后输入模型中进行预测，并将预测结果按照格式写入 csv 文件中。

参考资料：

[1] https://www.kaggle.com/c/state-farm-distracted-driver-detection.
[2] https://www.cdc.gov/motorvehiclesafety/distracted_driving/.
[3] https://en.wikipedia.org/wiki/Convolutional_neural_network.
[4] https://keras-cn.readthedocs.io/en/latest/.
[5] https://arxiv.org/abs/1409.1556
[6] https://arxiv.org/abs/1512.03385
[7] https://arxiv.org/abs/1602.07261