驾驶员状态检测 报告

1. 定义

1.1 项目概述

“计算机视觉”是指用摄像机和机器人代替人眼对目标进行识别、跟踪和测量等机器视觉，并作进一步的图像处理，用计算机处理成为更适合人眼观察或传递给仪器检测的图像。其用于创建计算机视觉系统，这类系统的组成部分包括：①过程监控；②事件监控；③信息组织；④物体与环境建模；⑤交感互动。

在人工智能研究的主要问题——让系统具有计划能力和决策能力中，计算机视觉系统，作为一个感知器，为决策提供信息。计算机视觉研究方向包括模式识别和机器学习。因此，计算机视觉时常被看作人工智能与计算机科学的一个分支。计算机视觉相关领域还包括有物理（如红外线）、神经生物学（模拟生物视觉），以及信号处理等。

计算机视觉现今的主要课题是：如何在取得合理的更快执行速度的情况下又不损失足够的精度。

在汽车行驶过程中，驾驶员的状态很大程度上决定了驾驶过程的安全程度。中国的道路交通事故的原因中，由于人的主观故意和过失而引发的交通道路事故，占所有道路交通 95%。造成事故的机动车驾驶员不安全因素包括三个方面：①违反规定、②驾驶疏忽、③操作不当。

在三种不安全因素造成的事故中，占比最大的是违反规定（占比 63%），具体为驾驶员违反交通法规和其他有关交通安全规定引发的事故。主要包括跟车过近、抢道行驶、超速行驶、超载行驶、不按车道行驶、违章停车、无证驾驶等等。

占比第二的是驾驶疏忽（占比 33%），具体是指由于驾驶员心理或生理方面的原因，没有准确的判断与交通行为有关的动态和静态事故而造成的事故。

占比最少的是操作不当（占比 4%），具体是指由于驾驶员技术生疏，经验不足，情绪不安、对车辆道路不熟，遇突发情况时惊慌失措，操作不当而发生的事故[1]。

当驾驶员处于非安全驾驶状态的情况下，比如接打电话、聊天、吃东西、化妆等，极容易造成驾驶疏忽或者操作不当。驾驶员检测技术对降低交通事故率有着重要的作用。对驾驶员的状态监控目前采用的方式有驾驶员移动幅度和次数检测、驾驶员体温检测、车内气体检测、视频处理检测、图像处理检测等。

1.2 问题陈述

本次研究内容为使用计算机视觉功能，对驾驶员的驾驶图片进行驾驶状态检测。

本次研究内容为解决驾驶员的具体状态。驾驶员的状态包括 9 种：

1. c0-安全驾驶
2. c1-右手打字
3. c2-右手打电话
4. c3-左手打字
5. c4-左手打电话
6. c5-调收音机
7. c6-喝饮料
8. c7-拿后面的东西
9. c8-整理头发和化妆
10. c9-和其他乘客说话

研究目标是识别出图片上驾驶员处于以上 10 种状态的概率，输出每种概率的结果。概率为 100%表示确定为该状态，概率为 0%表示确定不为该状态。以概率最高的状态判定为该图片的预测标签。

1.3 评价指标

研究采用的评估指标是损失函数 LogLoss：

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

N： 测试文件图片的数量

M： 图片标签类别的数量

log： 自然对数

y_{ij}： 如果 i 属于 j 的类别，那么 y_{ij} 为 1，否则为 0

p_{ij}： i 属于 j 类别的可能性（几率）

loss 函数是指模型的预测值与真实值的不一致程度，LogLoss 函数是最小化负的似然函数。LogLoss 的值越小，则代表损失函数越小，也就说明模型的预测性能越好。

2. 分析

2.1 数据探索

2.1.1 输入数据

需要对数据集进行更详细的描述，比如图片都有什么特点，图片是从哪些角度拍摄的，人物都有哪些特点等，可以结合几个样本进行介绍。你还需要介绍你将会如何分割这些数据，因为这些样本都是从视频导出的。

在本问题中涉及的输入为：驾驶员的彩色照片。即 RGB 三通道的图片。数据集获取来自 kaggle- State Farm Distracted Driver Detection。

数据集包含 test 和 train 两个图片文件夹，train 文件夹内的图片作为输入，内有 22,424 张驾驶员照片， train 文件夹内按照司机照片上的状态（安全驾驶、右手打电话、喝饮料等）分为分为 c0-c9 十种状态。

数据集特点：

1. 训练数据集已经按照分类进行排序，训练数据内一共有 26 名不同的司机，每个状态分类中均包含有多个司机图片。
2. 拍摄位置： 输入数据集中的图片均为在汽车内拍摄的驾驶员视频导出图片。拍摄点在副驾驶位置，比驾驶员头部略高的水平位
3. 显示范围： 图片显示的范围主要为驾驶位，包括有方向盘在内（由于汽车品牌不一致方向盘的大小、位置以及样式均有细微差别），其中一些图片显示范围还包括有汽车收音机部件。
4. 人物特点： 驾驶员以白种人居多，包括有黄种人及黑种人。拍摄到的人像为侧面半身像，均可以见到完整头部。驾驶员的手部位置，可以在图片中清晰的观察到，当分类为‘调收音机’和‘拿后面东西’的图片分类中右手手掌部位会有超出图片范围的情况，在其他分类均可见右手掌。

2.1.2 测试数据

测试数据集为 test 文件夹内图片。测试图片数量为 79,726 张图片，拍摄位置与场景与训练数据集内图片一致。

2.2 算法和技术

2.2.1 算法

本次研究是一个多分类问题。用与分类的算法包括有感知机、逻辑回归、支持向量机、决策树以及神经网络等。以上算法的 Input 为图片像素组成的特征向量，output 是司机状态的 10 个类别的概率。

感知机是一种二元线性分类器，感知机主要的本质缺陷是它不能处理线性不可分问题。

逻辑回归（Logistic regression）使用不同的假设类来尝试预测给定的种类概率。

支持向量机（support vector machine,常简称为 SVM）是在分类与回归分析中分析数据的监督式学习模型与相关的学习算法。除了进行线性分类之外，SVM 还可以使用所谓的核技巧有效地进行非线性分类，将其输入隐式映射到高维特征空间中。

决策树[2]（Decision tree）代表的是对象属性与对象值之间的一种映射关系。树中每个节点表示某个对象，而每个分叉路径则代表某个可能的属性值，而每个叶节点则对应从根节点到该叶节点所经历的路径所表示的对象的值。决策树仅有单一输出，若欲有复数输出，可以建立独立的决策树以处理不同输出。 数据挖掘中决策树是一种经常要用到的技术，可以用于分析数据，同样也可以用来作预测。

神经网络（Neural Network，NN）或类神经网络，在机器学习领域，是一种模仿生物神经网络的结构和功能的数学模型或计算模型，用于对函数进行估计或近似。神经网络由大量的人工神经元联结进行计算。大多数情况下人工神经网络能在外界信息的基础上改变内部结构，是一种自适应系统。现代神经网络是一种非线性统计性数据建模工具。

卷积神经网络（Convolutional Neural Network, CNN）是一种前馈神经网络，它的人工神经元可以响应一部分覆盖范围内的周围单元，对于大型图像处理有出色表现。

目前深度学习图像分类 CNN 主要有：VGG16[3]、VGG19、ResNet50、InceptionV3、Xception[4]、MobileNet。这五类 CNN 模型在过去几年 ImageNet[5]竞赛中表现优异。本次研究使用的模型为 ResNet50、MobileNet。

ResNet50 的网络结构依赖于微架构模组（micro-architecturemodules，也被称为 network-in-networkarchitectures）。ResNet 于 2015 年出现在 Heetal 的论文《DeepResidualLearningforImageRecognition》中，它的出现很有开创性意义，证明极深的网络也可以通过标准 SGD 来训练。在 2016 年的著作《IdentityMappingsinDeepResidualNetworks》中，他们证实了可以通过更新残差模组（residualmodule）来使用标志映射（identifymappings），达到提高精度的目的。

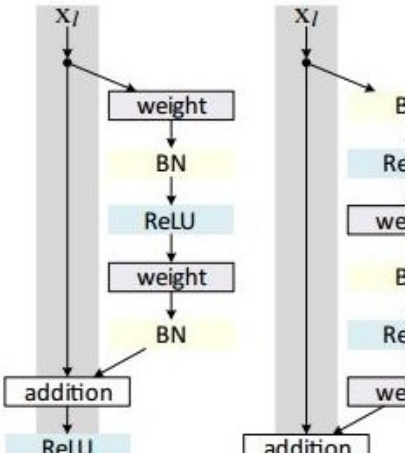
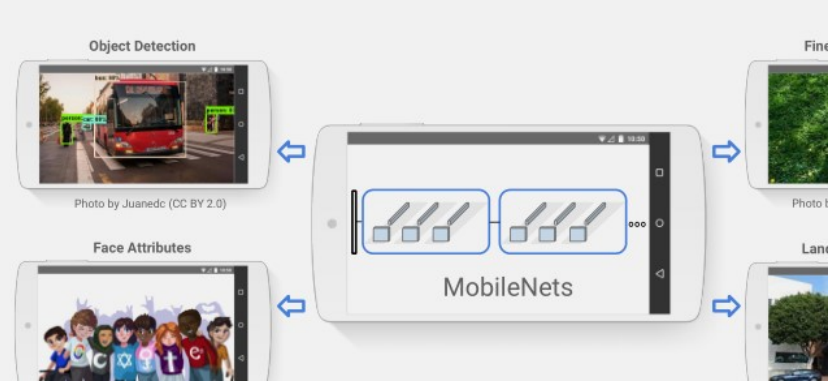


图 4:（左）原始残差模组（右）使用预激活（pre-activation）更新的残差模组

MobileNet[5]是 Google 针对手机等嵌入式设备提出的一种轻量级的深层神经网络，取名为 MobileNets。核心思想就是卷积核的巧妙分解，可以有效减少网络参数。

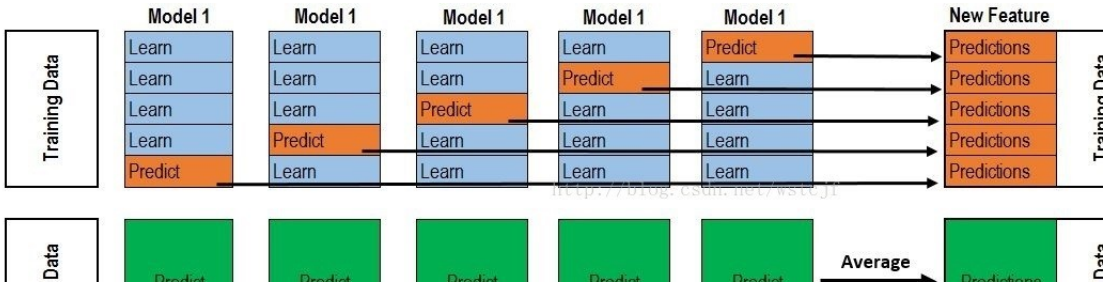


在 ImageNet 竞赛中，各个模型表现如下：

模型	大小	Top1准确率	Top5准确率	参数数
Xception	88MB	0.790	0.945	22,910,
VGG16	528MB	0.715	0.901	138,357,
VGG19	549MB	0.727	0.910	143,667,
ResNet50	99MB	0.759	0.929	25,636,
InceptionV3	92MB	0.788	0.944	23,851,
InceptionResNetV2	215MB	0.804	0.953	55,873,

(数据来源：Keras 文档[6])

在算法的组合上，采用两个算法组合并且使用模型融合技巧 [5] ,将数据集分为多个部分，分别进行预测之后再次采用预测结果作为新的特征值进行模型训练。以下为学习器的第一个模型学习流程。



如果有多个模型学习器，那么在最终分别预测测试集后，将测试结果取平均。

2.2.2 技术

本次研究使用深度学习框架 TensorFlow [7]。TensorFlow 可以利用多核 CPU 和 GPU 的优势。

2.3 基准指标

本次研究采用基准模型/基准阈值为 kaggle 上'Private Leaderboard'的排名第 144 名（总参加人数为 1440 人）。准确的 Multiclass Loss 分数为 0.25634。

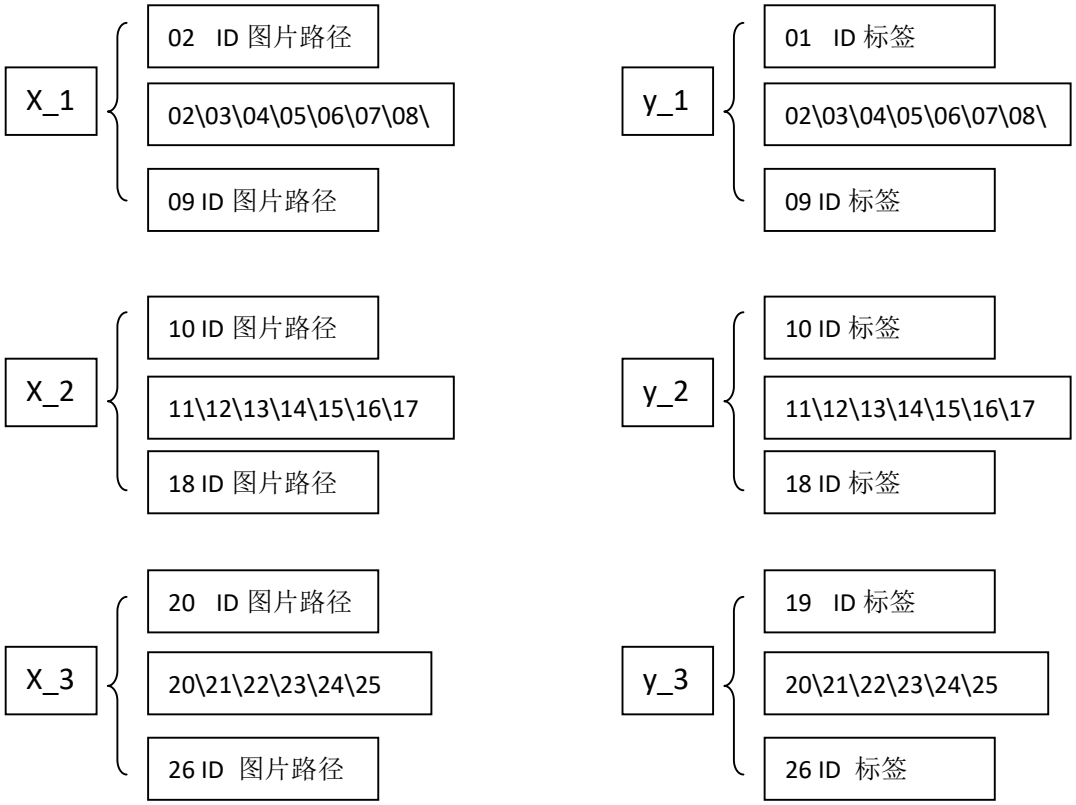
3. 具体方法

3.1 数据分割

本次数据为视频截图的司机状态，因此连续的多张图片相似性高，如果同时将该组图片放置于训练集和验证集中，容易产生过拟合。正因如此，在数据集处理的第一步是将图片按照司机 ID 进行分类，以使得在训练时某一个 ID 下的图片只处于训练集或者验证集中。

1) 读取已知 CSV 文件'driver_imgs_list.csv'，在其中分为 3 列：subject 表示司机 ID；classname 表示图片中司机状态（c0 – c9）；img 为图片名称。
2) 根据 CSV 文件提供信息， 将图片路径按照列表中的顺序重新排列，形成对应的两个列表 X_path, y_。再按照 26 个司机 ID 建立包含 26 个子列表的对应两个数据列表：X_data, y_data。
3) 将数据集分为 3 个部分，分别为：前 9 个 ID 数据集、9-18 个 ID 数据集以及后 10 个 ID 数据集，定义为 X_1, X_2, X_3, y_1, y_2, y_3。

具体见下图：



3.2 训练过程

单个模型训练过程：

1） 将 X_1 + X_2 作为训练集，X_3 作为预测集，使用 K 折验证法训练模型。

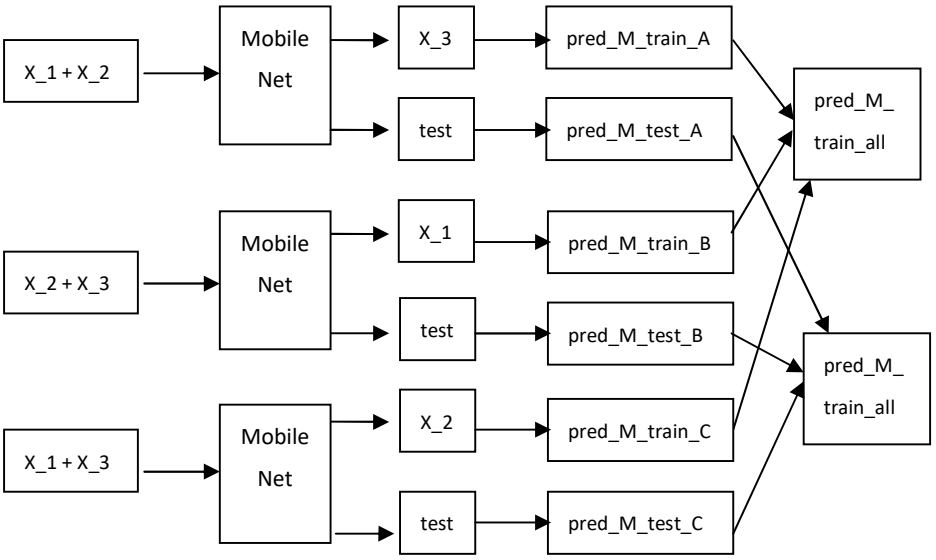
a) 按照本轮训练集中文件个数定义空矩阵 train_feature 和 train_label、verify_feature 和 verify_label，矩阵格式为 np.uint8。
b) 读取将 X_1 和 X_2 中的图片路径所对应的图片逐个写入矩阵，同时将特征值转换为独热编码。
c) 特征值进行模型训练之前，使用 Lambda 层进行数据归一化处理。
d) 将预处理之后的数据以及标签使用迁移学习 MobileNet 网络进行训练。
e) 用本次训练之后的模型对 X_3 预测，得到预测文件 pred_M_train_A，对测试集进行预测，得到预测文件 pred_M_test_A。

2） 将 X_2 + X_3 作为训练集，X_1 作为预测集，使用 K 折验证法训练模型。得到两个预测文件：
pred_M_train_B 以及 pred_M_test_B

3） 同第二步，更换训练集和预测集得到两个预测文件：
pred_M_train_C 以及 pred_M_test_C

4） 将三次训练得到的训练集三个部分预测值合并在一起，可以得到一个完整的训练集预测：
pred_M_train_all

5） 三次预测验证机的预测值采用取平均的方法得到完整的验证集预测：pred_M_test_all



将上述 1-6 步骤在使用 ResNet50 模型替换 MobileNet 之后重复一次，得到预测结果 pred_V_train_all，以及 pred_V_test_all。

以 pred_M_train_all + pred_V_train_all 作为新的特征，将之与对应的标签进行第二个阶段模型的 K 折训练。最后用第二阶段的模型在 pred_M_test_all/pred_V_test_all 上进行预测。得到测试集预测结果 A/B。

最后将 A 和 B 取平均值作为本次项目的最终测试结果提交。

迁移学习模型	训练步骤	训练集		验证集	
		损失函数	准确率	损失函数	准确率
MobileNet	训练 X_1 + X_2	0.0563	0.9857	0.3582	0.9065
	训练 X_2 + X_3	0.0310	0.9924	0.1301	0.9427
	训练 X_1 + X_3	0.0272	0.9927	0.0209	0.996
VGG16	训练 X_1 + X_2	0.0143	0.9952	0.0481	0.985
	训练 X_2 + X_3	0.0124	0.9962	0.002	0.999
	训练 X_1 + X_3	0.0034	0.9994	0.0024	0.998

3.3 改进

本次研究的过拟合问题比较严重。因此在模型训练时候的分数无法完全代表模型的性能。在改进过程中，以提交 kaggle 分数作为评价模型的标准。采用的方法包括：加入噪声层、加入正则、采用 ModelCheckpoint、更改优化器以及调整学习率和衰减率。具体见下表

方案	训练集		验证集		kaggle 分数
	损失函数	准确率	损失函数	准确率	
初始	0.0210	0.9784	0.0520	0.9210	0.54015
正则 权重正则 0.01	0.2751	0.9440	0.1851	0.9581	0.31182
高斯噪声层（标准差 0.8）	0.0268	0.9487	0.0268	0.9980	0.26081
高斯噪声层（标准差 2.0）	0.2751	0.9440	0.1851	0.9581	0.24997
高斯噪声层（标准差 3.0）	0.2320	0.9487	0.0260	0.9980	0.28330
更换优化器 SGD	0.1761	0.9575	0.1807	0.9556	0.24693

在训练过程中，由于训练集的损失函数衰减过快，有回弹的现象。为此在改进的过程中，加入 checkpointer，储存每次最优的权重。判断最优的标准是验证集损失函数 val_loss，储存最小损失函数对应的权重。在每次训练完成后，读取最佳函数再进行预测。

4. 结果

4.1 模型评价

在使用模型融合之后，模型能在使用把第一部的预测当作特征的问题空间中得到更多的信息。比单一模型进行训练能够更好的获得目标特征。

在使用 SGD 作为优化器并加入噪声层之后，取得 kaggle 上相当于’Private Leaderboard’的排名 134 名成绩，进入前 10%排名。

5. 结论

5.1 可能的应用

模型适用检测的场景为汽车内部副驾驶座位观察驾驶员状态。检测场景固定，但是在用于安全驾驶检查方面可以有所建树。

5.2 总结

研究成功达成了设定的目标，即达到 kaggle 排名前 10%分数。在车内的司机状态有多种，并且衣着和发饰有所不同。为了解决这个问题，由两个不同的模型做第一步的识别并形成预测结果，再由最终模型在此预测结果上进行分类，保证了模型的泛化性能。

5.3 改进

在模型方面，可以采用更多的神经网络作为第一阶段的模型来训练。由于每个网络有不同的结构，能够学习到不同的信息。在第一步训练时，还可以加强防止过拟合的技巧，加入噪声或者正则。为了增加模型能够学习到的信息，提高的技巧还可以对训练集进行图像增强，增强方法包括放大、裁剪、镜像、旋转等。为了让网络学习到更多的特征，也可以对图片进行局部遮盖。

参考文献

[1] 数据来源：百度词条-道路交通事故损伤

[2] Rokach, Lior; Maimon, O. (2008). *Data mining with decision trees: theory and applications*. World Scientific Pub Co Inc. ISBN 978-9812771711.

[3] Karen Simonyan. *Very Deep Convolutional Networks for Large-Scale Image Recognition*[C] arXiv:1409.1556v6,2015

[4] François *Chollet. Xception: Deep Learning with Depthwise Separable Convolutions*[C] arXiv:1610.02357,2017

[5] Andrew G. Howard .*MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications* .[C] arXiv:1704.04861v1

[6] Keras 文档 http://keras-cn.readthedocs.io/en/latest/

[7] TensorFlow 是谷歌基于 DistBelief 进行研发的第二代人工智能学习系统，其命名来源于本身的运行原理。Tensor（张量）意味着 N 维数组，Flow（流）意味着基于数据流图的计算，TensorFlow 为张量从流图的一端流动到另一端计算过程。