

机器学习纳米学位毕业设计

基于深度学习的走神司机识别

Magicyang 2018 年 5 月 8 日

1 定义

1.1 项目概述

根据 CDC 部门的数据，五分之一的车祸是由一名走神司机引起的。可悲的是，这意味着每年有 425,000 人受伤，3000 人因驾驶走神而死亡。如果能有效的区分驾驶员是否在专心驾驶，并制定相关的一系列规范，可以从一定程度上降低驾驶事故。

这是 KAGGLE 上两年前的竞赛题目。对应网页地址：

<https://www.kaggle.com/c/state-farm-distracted-driver-detection>

训练数据一共有 22424 张图片，26 个司机的数据供训练。一共有 10 个分类：

- 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

希望能尽可能的识别图片中司机的驾驶状态。

1.2 问题陈诉

该项目的数据集来自真实的驾驶司机图片。这个项目是一个监督学习图像多分类的问题，需要从驾驶司机的图片中分析出当前司机的驾驶行为。

基于图像分类的主要算法模型有：

- [VGGNet](#) 14.09 ¹
- [ResNet](#) 15.12 ²
- [Inception v3](#) 15.12 ³
- [InceptionResNetV2](#) 16.02 ⁴
- [Xception](#) 16.10 ⁵
- [NASNet](#) 17.07 ⁶

从中选择一个算法模型作为基准算法模型进行相关的改造和训练，使得图片根据算法分类得出的结果与司机实际的状态一致。

1. 评价指标

KAGGLE 不同的项目评价标准也会有不同。本项目的评判是使用了 Multi class log loss 多分类的对数损失函数。这也是多分类问题通常使用的评价标准。

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

其中 n 是测试集中图像的数目，M 是图像类标签的数目。如果观测对象 i 归属于 j 分类， y_{ij} 为 1，否则为 0。 p_{ij} 表示预测观测对象到 j 分类的概率。为了避免对数函数的极值，用 $\max(\min(p, 1-10^{-15}), 10^{-15})$ 代替。logloss 越小，预测错误率越低，模型也就越好。

KERAS 有对应的损失参数：

categorical_crossentropy：亦称作多类的对数损失，注意使用该目标函数时，需要将标签转化为形如 (nb_samples, nb_classes) 的二值序列。

目标是将损失函数尽量的降低。

具体可以参看：<https://www.kaggle.com/c/state-farm-distracted-driver-detection#evaluation>

本项目需要保证 LOGLOSS 在 Private Leaderboard 排名 10%以内，即：LOGLOSS 小于 0.25634。

2 分析

2.1 数据的探索

KAGGLE 项目提供了对应的数据集。数据集的下载地址：

<https://www.kaggle.com/c/state-farm-distracted-driver-detection/data>

先从训练集开始分析。driver_imgs_list.csv 是训练集的基础数据。

subject	classname	img
p002	c0	img_44733.jpg
p002	c0	img_72999.jpg
p002	c0	img_25094.jpg
p002	c0	img_69092.jpg

Subject、classname、img 分别代表：司机编号、司机状态分类、图片名称。

从该表中可以获取到训练集共有 22424 张图片，26 个司机，10 个分类。

下图 1 是训练集各个分类的数目：

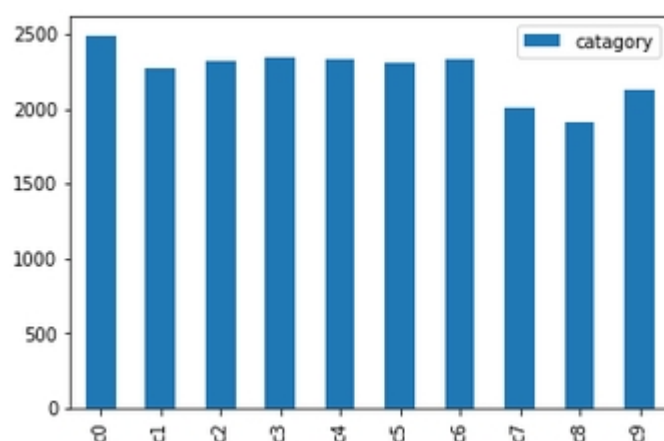


图 1

从图 1 可以看出各个分类的训练图数目基本一致。

图 2 是采样后同一个司机的 10 种状态原始数据展示。

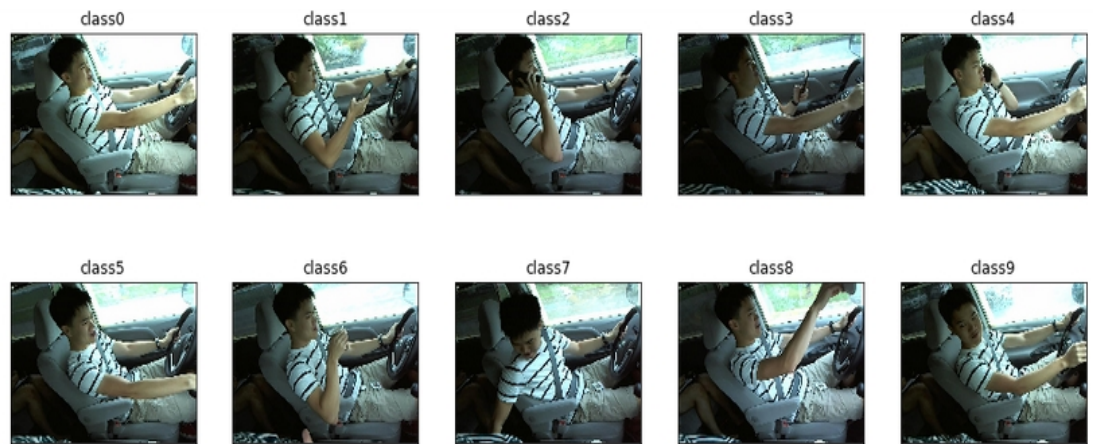


图 2

- 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

从图二上我们大概可以看出，司机状态不同主要是手部和脸部动作决定的。也希望在最后的验证过程中，可以看到对应的结果。

竞赛提交到 KAGGLE 的测试集共有 79726 张图片。最终需要提交测试集的结果到 KAGGLE 上，并获得验证分数。

2.2 算法和技术

自从 Alex 在 2012 年提出的 alexnet 网络结构模型引爆了神经网络的应用热潮，并赢得了 2012 届图像识别大赛的冠军之后，深度学习成为图像识别，尤其是图像分类的主要算法。

最近今年陆续发表了如下的图像分类算法：

- [VGGNet](#) 14.09 ¹
- [ResNet](#) 15.12 ²
- [Inception v3](#) 15.12 ³
- [InceptionResNetV2](#) 16.02 ⁴
- [Xception](#) 16.10 ⁵

- [NASNet](#) 17.07⁶

	Top-1 accuracy	Top-5 accuracy
VGG-16	0.715	0.901
ResNet-152	0.770	0.933
Inception V3	0.782	0.941
Xception	0.790	0.945

图 3

从上图 3 可以看到目前的模型分类准确率越来越高。

本项目主要使用了 VGG16 和 Xception 的模型做迁移学习。

2.2.1 迁移学习

迁移学习(Transfer learning) 顾名思义就是就是把已学训练好的模型参数迁移到新的模型来帮助新模型训练。考虑到大部分数据或任务是存在相关性的, 所以通过迁移学习我们可以将已经学到的模型参数(也可理解为模型学到的知识)通过某种方式来分享给新模型从而加快并优化模型的学习效率不用像大多数网络那样从零学习(starting from scratch, tabula rasa)。

迁移学习的本质就是知识的再利用, 数学上, 迁移学习包含“域”和“任务”两个因素。下面具体说明一下, 迁移学习在域和任务上可能有哪些差异:

给定源域和目标域 D_s 和 D_t , 其中, $D=\{X,P(X)\}$, 并且给定源任务和目标任务 T_s 和 T_t , 其中 $T=\{Y,P(Y|X)\}$ 。源和目标的情况可以以四种方式变化, 我们仍然以我们的文档分类的例子在下面描述:

$X_s \neq X_t$ 。源域和目标域的特征空间不同, 例如, 文档是用两种不同的语言写的。在自然语言处理的背景下, 这通常被称为跨语言适应(cross-lingual adaptation)。

$P(X_s) \neq P(X_t)$ 。源域和目标域的边际概率分布不同, 例如, 两个文档有着不同的主题。这个情景通常被称为域适应(domain adaptation)。

$Y_s \neq Y_t$ 。两个任务的标签空间不同, 例如, 在目标任务中, 文档需要被分配不同的标签。

$P(Y_s|X_s) \neq P(Y_t|X_t)$ 。源任务和目标任务的条件概率分布不同, 例如, 源和目标文档在类别上是不均衡的。

这里再说明一下另外一个问题, 为什么深度神经网络模型具有强大的可迁移能力? 回答这一点需要了解 DNN 的原理, DNN 是一个通过 pre-train 获得数据的分层特征表达, 然后用高层语义分类的模型, 模型的底层是低级语义特征(比如说, 边缘信息, 颜色信息等), 这样的特征实际上在不同的分类任务中都是不变的, 而真正区别的是高层特征。

本项目是一个图片多分类的问题，IMAGENET 是图片分类数据集，如果把 IMAGENET 作为源域，本项目作为目标域，两者是相接近的。因此可以用 IMAGENET 的分类算法来实现迁移学习。

同时也必须看到：源域和目标域的边际概率分布是不同，同时两个任务的标签空间也是不同的。因此我们只能保证最底层的语义特征是相似的，中层和高层特征是不一致的，这也是为什么本项目需要整体训练模型，而不是仅仅训练全连接层的原因。

关于迁移学习的内容可以参考：

<https://www.zhihu.com/question/41979241>

2.2.2 VGG16

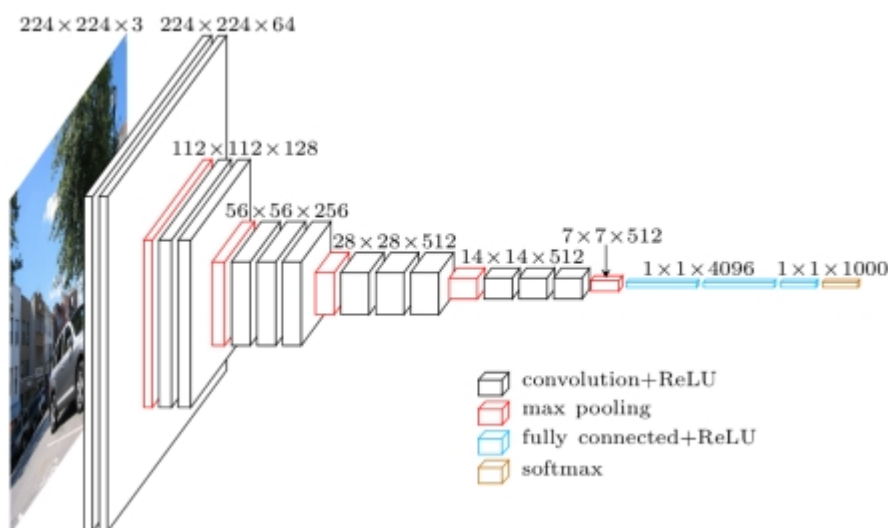


图 4

VGG16 的模型见上图 4。

VGG16 是 2014 年 9 月提出的图像分类模型，从上图 4 可以看到他的模型相对来说比较简单，只有 4 组卷积层，大量的参数运算都集中在全连接层。

下面介绍一下 ReLU 激活函数，目前 CNN 的卷积层都是使用 RELU 的激活函数，SIGMOID 的激活函数会造成梯度消散。ReLU 的公式如下：

$$\text{Relu}(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

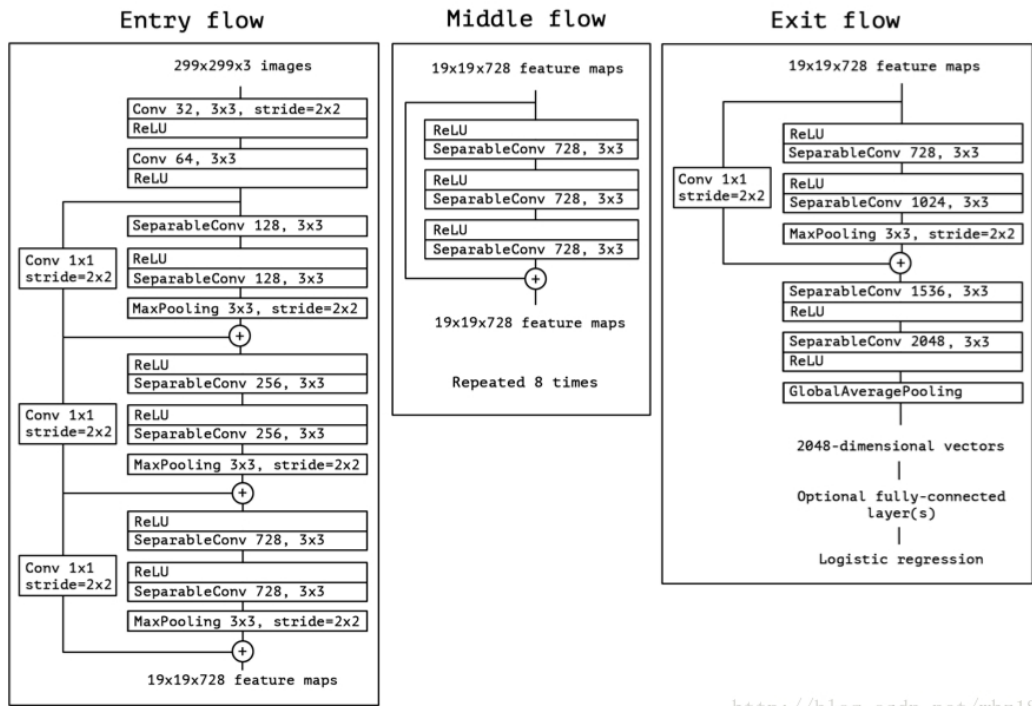
ReLU 有以下优势：对于线性函数而言，ReLU 的表达能力更强，尤其体现在深度网络中；而对于非线性函数而言，ReLU 由于非负区间的梯度为常数，因此不存在梯度消失问题，使得模型的收敛速度维持在一个稳定状态。

ReLU 的另一个优势是在生物学上的合理性，它是单边的，更符合生物学上神经元的特征。

ReLU 也有缺点，大的梯度流经过 ReLU 单元时可能导致神经不会在以后任何数据节点再被激活。当这发生时，经过此单元的梯度将永远为零。所以神经网络训练时需要选择合适的学习率，尽量避免这种情况的出现。

2.2.3 XCEPTION

Figure 5. The Xception architecture: the data first goes through the entry flow, then through the middle flow which is repeated eight times, and finally through the exit flow. Note that all Convolution and SeparableConvolution layers are followed by batch normalization [7] (not included in the diagram). All SeparableConvolution layers use a depth multiplier of 1 (no depth expansion).



<http://blog.csdn.net/whz1861>

图 5

XCEPTION 的模型见上图 5。

XCEPTION 的模型相对于前面的 VGG16 来说就复杂了很多，它的核心思想是将 Inception modul 拆分成一系列操作，独立处理 spatial-correlations 和 cross-channel correlations，网络处理起来更加简单有效。

从上图 5 的模型中可以看到在卷积块之间是使用了 1*1 卷积相加的操作，类似于 RESNET 的操作，主要用于空间数据整合。而在卷积块中间而是使用了 SeparableConv 进行 CHANNEL 卷积。

关于 SeparableConv 的结构可以看下面两张图：

Figure 3. A strictly equivalent reformulation of the simplified Inception module.

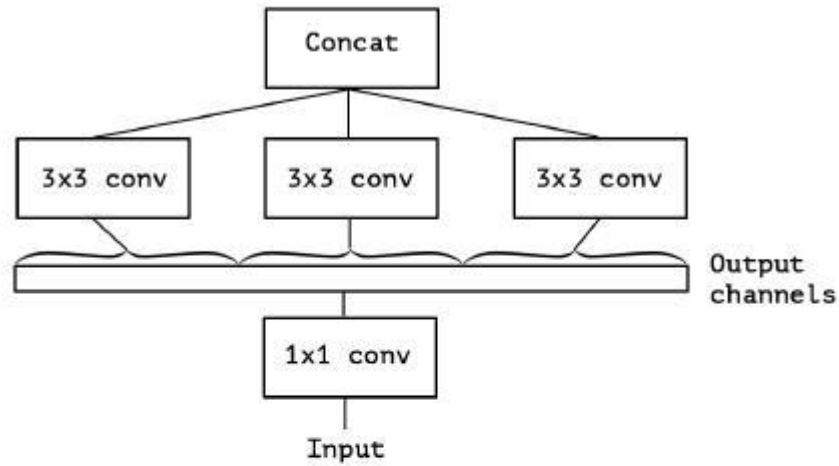


Figure 4. An “extreme” version of our Inception module, with one spatial convolution per output channel of the 1x1 convolution.

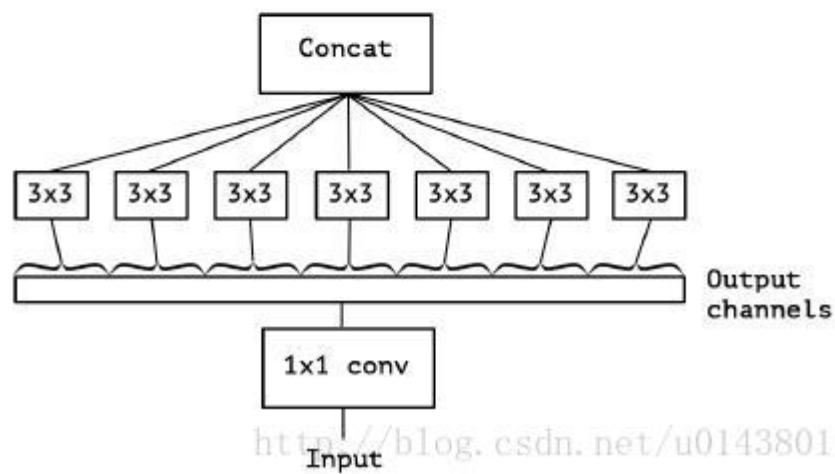
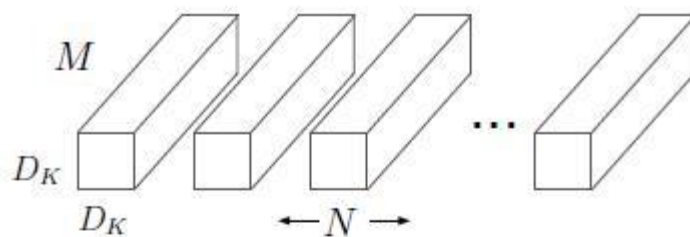
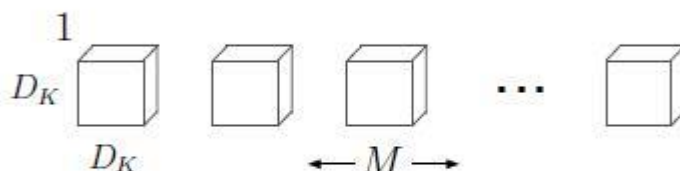


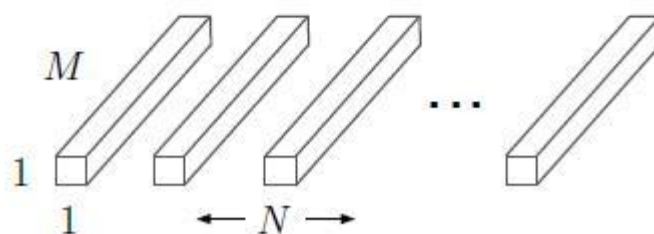
Figure4 是 INCEPTION 的终极模型，先进行 1*1 卷积，接着在 CHANNEL 的维度进行分离，接着对分离结果分别做 3*3 卷积，最后整合成和原 CHANNEL 维度相同的结果。



(a) Standard Convolution Filters



(b) Depthwise Convolutional Filters

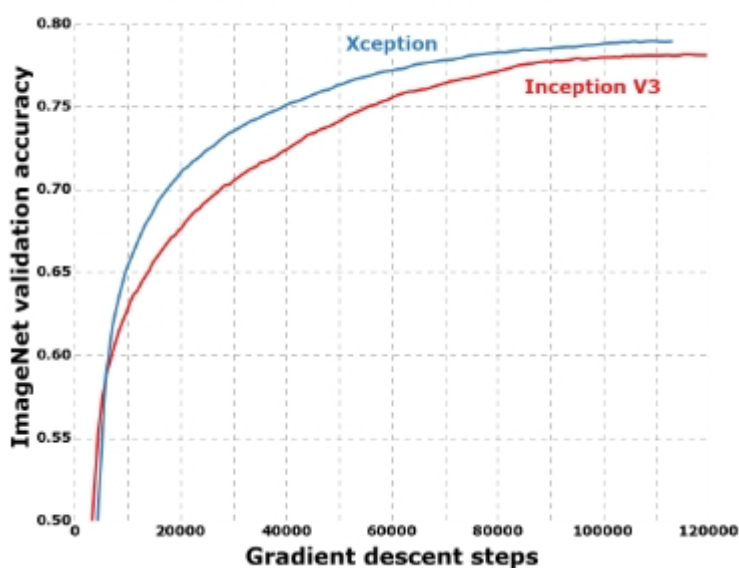


(c) 1×1 Convolutional Filters called Pointwise Convolution in the context of Depthwise Separable Convolution

其实 depthwise separable convolution 和上面的 Figure4 是很像的。差别有两个：1、顺序不一样，在 depthwise separable convolution 中是先进行一个 channel-wise 的 spatial convolution，也就是上图的 (b)，然后是 1×1 的卷积。而在 Figure4 中是先进行 1×1 的卷积，再进行 channel-wise 的 spatial convolution，最后 concat。

从整体模型来看，XCEPTION 一方面同时结合了 INCEPTION 和 RESNET 的优点，另一方面模型同时从 CHANNEL 和空间两个方向去做卷积去提取特征，在论文⁵中也可以看到其测试结果优于其他 CNN 卷积网络。

Figure 6. Training profile on ImageNet



从图 3 可以看到 Xception 在主流算法里 IMAGENET 验证集里的准确率是最高的，同时对比上图可以发现对比 INCEPTIONV3，XCEPTION 有更快的收敛速度。

项目中使用 KERAS 工具构建模型，KERAS 已经集成了 VGG16 和 XCEPTION 的模型，大量降低了我们构造模型错误的可能性。

2.3 基准模型

本项目需要保证 LOGLOSS 在 Private Leaderboard 排名 10%以内，即：LOGLOSS 小于 0.25634。

在训练过程中，希望验证集的 LOGLOSS 尽可能的低。

VGG 的模型相对简单，可以用来先做相应的测试。而 XCEPTION 模型相对性能更好，在需要时候替换模型。

3 方法

3.1 数据预处理

3.1.1 数据分割

该项目中一共有 20000 多张图片，需要正确的分割训练集和验证集。从输入文件可以看到这 20000 张图片分属于 26 个司机。从数据探索可以看到相同司机的图片差异很小，相同状态下不同司机的图片差异更大。所以

不能直接对整个训练集进行数据划分，需要对司机进行训练集和验证集的划分，即训练某些司机的图片，在其他司机上进行验证。

3.1.2 特征提取

对于不同的基准模型，输入的像素是不一样的。如 VGG16 是 $224*224*3$ ，而 XCEPTION 是 $299*299*3$ 。项目统一使用 OPENCV 读取图片特征。

另外需要注意的是我们都是使用的 IMAGENET 的训练集作为迁移学习的基础，对于不同的模型，特征预处理是不一样的。KERAS 已经封装好了预处理方法。可以详见：

https://github.com/yangchengtest/keras/blob/master/keras/applications/imagenet_utils.py

如果需要一次处理特征，需要关注一下内存的使用情况。

3.2 执行过程

3.2.1 读取特征的内存分析

前面提到需要考虑特征读取的内存。例如使用 XCEPTION 模型，使用 UINT8 来记录特征所需要的内存是： $224*224*299*299*3$ 约 6G 内存。在实际处理过程中，我把预处理统一提前在 CPU 一次处理，不用每个 EPOCH 处理一次，使用了 FLOAT16 来记录特征这就是需要 12G 内存。如果开始阶段就把所有训练数据的特征提取出来，再分割成训练集特征和验证集特征，就需要 2 倍的内存开销，对于 64G 以下的用户来说，可能出现 OOM。因此我个人是在切分完验证集和训练集之后再获取图片特征。还有一个办法就是使用 UINT8 记录特征，在 KERAS 训练模型里做预处理，也可以节约内存，各有利弊。

3.2.2 模型的选择和是否冻结的问题

前面提到图片比较相似，需要全部训练模型，在实际的操作中，个人还是尝试了先通过卷积层先训练出图片卷积层的特征，然后只训练全连接层。实际尝试后发现，验证集的 logloss 无法降低。

接着开始使用 VGG16 的模型来进行训练，在实际训练中，开始忘记在模型中加入 IMAGENET 的权重，导致训练集的 logloss 一直无法降低。我个人是这么理解的：如果不加初始权重实际上就不是迁移学习了，是整个识别算法全部重训练，这个需要大量的训练数据集和大量的训练次数来迭代完成，而我们的训练集只有 20000 多张图片，很多分类特征根本训练不出来。如果看到 logloss 一直很高，看一下是否忘记加模型权重了。

最后还是选择了 XCEPTION 作为基准模型，这部分在后面的优化里去描述。

3.2.3 优化器的参数选择

CNN 的反向传播本质上都是梯度下降，随着深度学习的发展，优化器也在不断的发展。

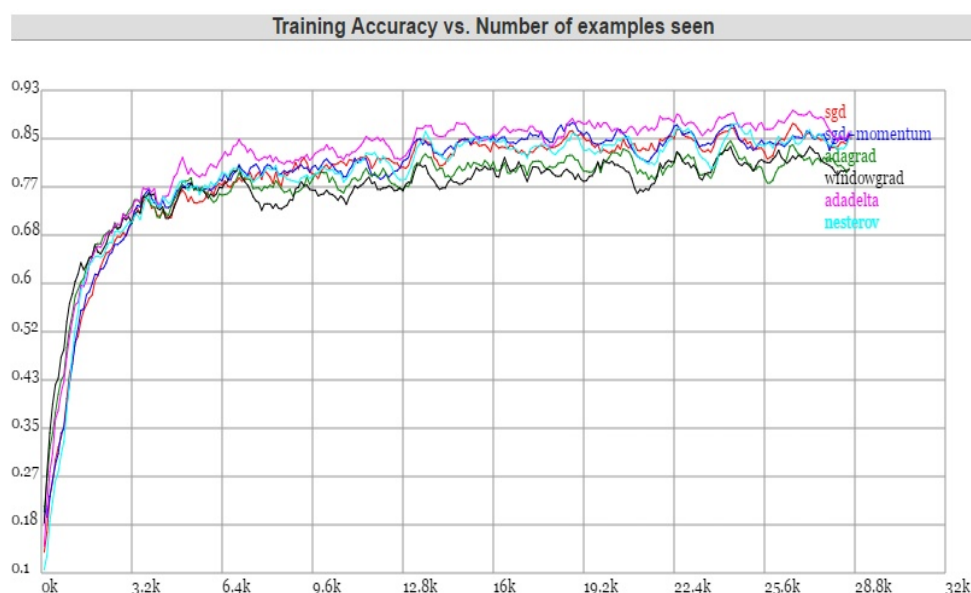


图 6

从图 6 可以看到各个算法的准确率标识。

各个优化器的优劣可以参考知乎的文章：

<https://zhuanlan.zhihu.com/p/22252270>

目前主流的优化器选择是 ADAM⁷ 和 SGD-Nesterov。

提到优化器，就不得不提学习率。当学习率过低的时候，可能陷入局部极值，而当学习率过大的时候，又可能导致反向传播过快，出现梯度爆炸。因此在选择优化器后，选择合适的学习率比较重要。例如在本项目中，使用了 ADAM 优化器，ADAM 的学习率推荐是 10^{-3} 到 10^{-5} 。选择 10^{-3} 时就会出现梯度爆炸的情况。



3.3 完善

3.3.1 使用 KFOLD

这里需要提到的是交叉验证法。它的基本思想就是将原始数据（dataset）进行分组，一部分做为训练集来训练模型，另一部分做为测试集来评价模型。交叉验证用于评估模型的预测性能，尤其是训练好的模型

在新数据上的表现，可以在一定程度上减小过拟合。还可以从有限的数据中获取尽可能多的有效信息。

项目的模型并不大，这里使用的 KFOLD(K 折交叉验证)，通过 K 折后的预测平均可以有效的提高准确率，降低单训练的过拟合。

submission_10.csv a minute ago by magicyang xception-10	0.36330	0.31445	<input type="checkbox"/>
submission_9.csv 33 minutes ago by magicyang xception-9	0.41243	0.48640	<input type="checkbox"/>
submission_8.csv 34 minutes ago by magicyang xception-8 	0.33843	0.30654	<input type="checkbox"/>
submission_7.csv an hour ago by magicyang xception-7	0.43228	0.39570	<input type="checkbox"/>
submission_6.csv an hour ago by magicyang xception-6	0.34553	0.39912	<input type="checkbox"/>
submission_5.csv an hour ago by magicyang xception-5	0.32300	0.35430	<input type="checkbox"/>
submission_4.csv an hour ago by magicyang xception-4	0.38370	0.35981	<input type="checkbox"/>
submission_3.csv 3 hours ago by magicyang xception-3 	0.39483	0.33210	<input type="checkbox"/>
submission_2.csv 3 hours ago by magicyang xception-2	0.35226	0.37786	<input type="checkbox"/>
submission_1.csv 3 hours ago by magicyang xception-1	0.37598	0.37278	<input type="checkbox"/>
submission.csv 6 days ago by magicyang XCEPTION KFOLD	0.22823	0.23925	<input type="checkbox"/>

上图记录了 10 个 XCEPTION 提交 KAGGLE 的分数和最终使用 KFOLD 模型的分。单模型的 PRIVATE SCORE 都在 0.3-0.5 之间，当使用 KFOLD 后 PRIVATE SCORE 下降到了 0.2223。使用交叉验证法可以显著提升模型效果。

3.3.2 模型的改进

Submission and Description	Private Score	Public Score	Use for Final Score
submission.csv 3 days ago by magicyang XCEPTION KFOLD	0.22823	0.23925	<input type="checkbox"/>
submission.csv 5 days ago by magicyang add submission details	0.29202	0.30417	<input type="checkbox"/>

图 7

先上图，图 7 是 kaggle 提交的最终结果，在使用 VGG16 的模型情况下，使用 KFOLD 运算，logloss 并没有进入 10%，而使用 XCEPTION 则相对来说比较容易的进入了 10%。个人分析原因是相对来说，XCEPTION 相对于 VGG16 来说，VGG16 的参数主要集中在全连接层，卷积层的参数并不多，越多的卷积层参数表示底层可描述的细节越多。这就导致在细节特征提取上，XCEPTION 相对于 VGG16 有天然的优势。

下面就开始选择优化器，ADAM 有更快的收敛速度，本项目最终还是选择 ADAM 作为优化器。主要还是调整学习率，合适的学习率可以看到 VAL_LOSS 先是慢慢下降，后续过拟合再升高的变化。

最后为了降低训练时间和防止过拟合，使用 EASY STOPPING 来结束模型训练。

4 结果

4.1 模型的评价

最终模型简述为如下的步骤：

- (1) 将 XCEPTION 作为基础模型，在卷积层后加入一层 AveragePooling 和一层激活函数为 SOTFMAX 10 的全连接层。使用 IMAGENET 数据集的卷积层权重作为新模型的卷积层权重。
- (2) 对数据进行预训练，使用 KFOLD 方法，折线数 10，将司机 ID 分解。从 KFOLD 分解出的训练集司机 ID 和验证集司机 ID 中分别读取图片特征。由于训练数据读取是有顺序的，在提交训练之前需要打乱。
- (3) 选择学习率为 10^{-4} 的 ADAM 优化器作为优化器进行训练。主要参数还有 batch_size=32,epoch=10,early_stop 的 patience=2。

下面说一下训练参数的选择。

学习率前面已经提过，过高过低都会有问题，目前的学习率是试验出来的。

BATCH_SIZE 由于我们都是使用 GPU 进行训练，2 进制的 **BATCH_SIZE** 更方便机器学习，在 GPU 内存允许的条件下可以提升一下 **BATCH_SIZE** 数目，**BATCH_SIZE** 过低，**VAL_LOSS** 也会无法收敛。

early_stop 的选择则相对无奈，这个跟 GPU 的性能是强相关的，我自己的 GTX1070 的 8G 显卡，**BATCH_SIZE** 只能支持 16，一个代次需要 18 分钟。后续使用 AWS 的 P3.2XLARGE，一个代次只需要 4 分钟。由于使用了 **KFOLD** 需要训练 10 个模型，时间规模上是巨大的。因此使用了较小的 **early_stop**。其实这跟优化算法也有关系，如果使用 SGD-Nesterov，训练代次会增加，收敛速度会变慢一点，也可能会得到更加低的 **VAL_LOSS**。时间和金钱成本需要考虑。

在模型训练中使用了司机 ID 分离训练集和验证集，以及使用 **KFOLD** 降低单模型的过拟合。使用 **XCEPTION** 替代 **VGG16** 增加卷积层的描述性，我认为项目的最终模型是优于初始模型的。

4.2 模型的验证

本项目是 KAGGLE 的竞赛项目，最终是需要提交到 KAGGLE 上进行验证的。测试集中一共有 79726 张图片。基于内存的考虑，分解成了 2 个测试集进行模型预测。

前面提到训练过程中使用了 **KFOLD**，在验证的时候，需要对同一个输入图片，需要分别使用 10 个训练后的模型对输入进行预测，然后对 10 个预测进行平均，最后得到最终的训练结果。

使用 **DataFrame** 记录训练后的结果，最终提交到 KAGGLE 上，获得分数，评价是否满足项目的要求。

Submission and Description	Private Score	Public Score	Use for Final Score
submission.csv 3 days ago by magicyang XCEPTION KFOLD	0.22823	0.23925	<input type="checkbox"/>
submission.csv 5 days ago by magicyang add submission details	0.29202	0.30417	<input type="checkbox"/>

最终 Private Score 为 0.22823 低于目标 0.25634，项目达成。

4.3 CAM 可视化

最后来看一下 CAM⁸ 可视化，可以参考附录的论文。对于实现来说就是用卷积层的结果点乘全连接层分类，对于 **XCEPTION** 来说， $10 \times 10 \times 2048$ 点乘 2048×1 会得到一个 10×10 的图，放大到输入特征图大小(299×299)，用热度图的方式展现出来。

可以用下图 8 对比图 2，在 CAM 图例，司机的特征区域都比较明显的。模型训练结果满足需求。

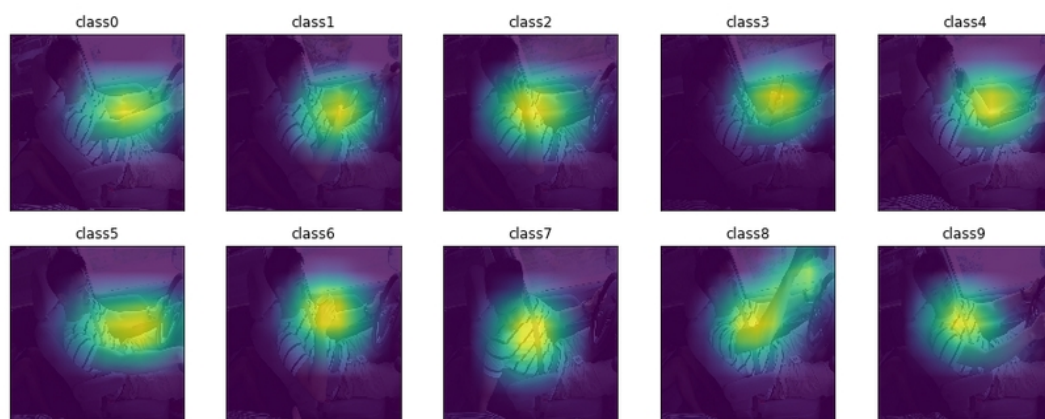


图 8

5 项目结论

5.1 对项目的思考

总体来看这是一个非常好的项目，项目中需要掌握以下的知识：

- (1) 数据预处理，司机项目需要考虑内存，考虑特征规模。
- (2) 图像处理，OPENCV 的基础函数。
- (3) 评价标准，司机项目使用 **logloss** 评价体系，对该评价体系有了深入了解。
- (4) 模型使用，对于简单模型，可能无法达到标准，这就需要我们更多的了解新模型，理解新模型，使用新模型
- (5) 如果不使用 **KFOLD** 结果也不会优秀，教会我们使用交叉验证法。
- (6) 走神司机的项目模型不算小，从该项目中大概了解了模型的计算规模，学会了使用 **AWS** 的 **P3.2XLARGE** 进行编程。以后遇到相应的问题，可以更快的评价计算量和对项目进行整体规划。
- (7) 学会了 **CAM** 可视化，看懂论文也花了很长的时间，对可视化有了一定的探索。
- (8) 调参很重要，大概试错了一些，还需继续加强。

从开始做毕业项目到完成这个毕业项目，总共花了 **12** 个工作日的时间，如果需要做到更好需要更多的时间，很遗憾给自己定了一些小目标，后面接着要去做新的项目，很感谢这个项目给我个人的锻炼。

5.2 后续改进

这里还有一些做的不够完美的内容，个人觉得项目可以在以下方面继续加强：

- (1) 调参不是特别充分，**ADAM** 收敛太快，换 **SGD** 可能结果更好。
- (2) 可以尝试融合模型，相对单模型，融合模型结果应该会更好。

(3) 没有针对训练集做数据增强，做数据增强的化结果应该会更好。

参考文献：

[1] Karen Simonyan, Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition.(arXiv:1409.1556,2014)

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun.Deep Residual Learning for Image Recognition.(arxiv:1512.03385,2015)

[3] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Zbigniew Wojna .Rethinking the Inception Architecture for Computer Vision. (arXiv:1512:00567,2015)

[4] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, Alex Alemi. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. (arXiv:1602.07261,2016)

[5] François Chollet .Xception: Deep Learning with Depthwise Separable Convolutions. (arXiv:1610.02357,2016)

[6] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, Quoc V. Le.Learning Transferable Architectures for Scalable Image Recognition. (arXiv:1707.07012,2017)

[7] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. CVPR'16 (arXiv:1512.04150, 2015).

[8] Diederik P. Kingma, Jimmy Ba.Adam: A Method for Stochastic Optimization (arXiv:1412.6980,2014).