**测试和评价指标说明**

放假之前让蒋正明同学做了一个baseline的ppt放在了群里。

他的ppt中是将本次比赛的内容作为一个分类任务进行演示的。这种说法虽然也对但还是存在一定的局限性。因为在实际应用中，厂商会不断的推出新品，那么分类模型的扩展性就不够了。为了应对实际场景中存在的新品不断出现的问题，我们可以把赛题作为一个类似ReID的任务 [1]。

* **真实场景**

商家要上架某个产品时，会拍一些相关产品的图像，提取特征之后存入数据库。当前端设备采集了一个**未知品类的商品**的图片后，通过模型提取这个图片中**未知品类的商品**的特征，与数据库中的**商品特征**进行查询和匹配，找到“距离”最近的品类，作为这个**未知品类的商品**的类。

* **数据集**

已经发布的数据集包括100个类别，每个类别有10~15张图片（valid中有相应类别的是15张）。现按照reid任务重新组织为：

Train集合（公布）：92个类目，每个类目10个图片

Valid集合（公布）：分为两个文件夹，一个是key，另一个是query。

另外用于测试排名的非公开数据集，文件夹格式与valid集合相同，包含了train集合中出现的类别，也包括其中未出现过的类别（新品类）。

* **测试方法**

1. 拿query目录下面的每一张图的特征和key目录下图像的特征做比对，要求给query里面的每张图一个预测label，和其自己真实的label比较，得到正确率。
2. 在非公开的测试集里做验证，给出排行。

* **指标 [2]**

**（1）Rank-N准确度**

Rank-N准确度是一种在信息检索和机器学习中常用的评估指标。它主要用于评估一个模型在预测时，正确答案是否出现在模型预测的前N个结果中。

例如，假设我们正在评估一个训练在CIFAR-10数据集上的神经网络模型，CIFAR-10数据集包括10个类:飞机，汽车，鸟、猫、鹿、狗、青蛙、马、船和卡车。给定一张输入图像，模型返回的结果是一个类标签概率列表。我们先看看rank-1的计算，对于每一张图片，取模型预测的类概率列表中最大的概率对应的标签作为该图片的预测结果。如果预测的标签与真实的标签相同，那么我们就说这个预测是正确的。这样，rank-1准确度就是预测正确的图片数量占总图片数量的比例。

现在，我们扩展到rank-5准确度，我们关注的不是top1的预测，而是top5的预测。对于每一张图片，我们看看真实的标签是否出现在模型预测的前5个标签中。如果出现了，那么我们就说这个预测是正确的。这样，rank-5准确度就是预测正确的图片数量占总图片数量的比例。

这种方法也可以很容易地推广到计算rank-N准确度。一般而言，我们只计算rank-1和rank-5准确度——计算rank-1的准确度可以理解，为什么还需要计算rank-5准确度呢？对于大型的、具有挑战性的数据集来说，特别是细粒度的分类，我们往往会增加一个rank-5准确度，也就是说我们不止关心rank-1准确度，也关心rank-5准确度。结合两个准确度来以衡量神经网络的性能。

**（2）mAP：**全称为mean Average Precision，是一种在信息检索和机器学习中常用的评估指标。它主要用于评估一个模型在预测时，对于每个类别的平均精度（Average Precision，AP）的平均值。

AP是对precision-recall曲线进行总结的一种方法，它计算了在每个阈值处实现的精度的加权平均值，其中增加的召回率用作权重。mAP则是对所有类别的AP进行平均，得到的结果。在一些任务中，如目标检测和图像分割，mAP是一个非常重要的评估指标。

* **参考文献**

[1]https://www.bing.com/ck/a?!&&p=33665f5c1fd76b30JmltdHM9MTcwODM4NzIwMCZpZ3VpZD0xYjE2MTMzMS1lNGJlLTY3MWUtMjQ3Yy0wMDdkZTU5MDY2Y2EmaW5zaWQ9NTIyMQ&ptn=3&ver=2&hsh=3&fclid=1b161331-e4be-671e-247c-007de59066ca&psq=%e8%a1%8c%e4%ba%ba%e9%87%8d%e8%af%86%e5%88%ab%e4%bb%bb%e5%8a%a1&u=a1aHR0cHM6Ly96aHVhbmxhbi56aGlodS5jb20vcC80NTYwNjAyMjE&ntb=1

[2] https://www.cnblogs.com/tay007/p/11341701.html