标题

杭州海康威视数字技术股份有限公司

XXX 字 XXX 号

深度学习性能测试指标规范

密级级别：AA 级商业秘密
生效时间：2022 年 4 月 11 日
保密期：3 年

目录

1. 简介

1.1 编写目的

本文档主要介绍深度学习相关的各算法模块及业务场景所需的性能测试指标，旨在统一不同资源组的性能测试指标，以便于后续进行模型基线版本的更新维护。

本文档的预期读者为算法开发人员、项目经理以及相关的管理人员。

1.2 背景

目前，深度学习技术在公司内部得到越来越广泛的应用，但是各组之间都有自己的评价指标，给算法的交流与更新带来一定的阻碍。即使几个小组使用了相同的测试指标，也可能存在具体的实现差异或者参数设置差异，使得性能测试数据无法完全对齐。因此，必须推出一套具有一定通用性的性能测试指标并开发相应的脚本，作为各组共同使用、维护的性能测试工具，从而实现性能测试指标及数据的统一。

1.3 术语和缩写

术语/缩写	含　义
AP	Average Precision，指 PR 曲线下方的面积
F1	F1_Score，又称平衡 F 分数，同时兼顾了分类模型的精确率和召回率

1.4 参考资料

1.5 测试环境

1.5.1 软件环境

操作系统	Windows7/8/10　ubuntu16.04/18.04/20.04
编译环境	Vscode/anaconda/pycharm　python3.x
依赖库	
静态库名	暂无

1.5.2 硬件环境

CPU	Intel 系列
内存	8G/16G/32G/512G

2. 评价指标

2.1 基础概念

2.1.1 混淆矩阵

混淆矩阵就是统计分类模型的分类结果，即：统计归对类，归错类的样本的个数，然后把结果放在一个表里展示出来，这个表就是混淆矩阵。以猫狗猪三分类的混淆矩阵为例，可见图 2.1.1。需要注意的是混淆矩阵会受到后处理的影响，在实际使用时需要注意。

混淆矩阵		真实值		
		猫	狗	猪
预测值	猫	10	1	2
	狗	3	15	4
	猪	5	6	20

图 2.1.1 混淆矩阵示例

2.1.2 IoU 与 MIoU

在基于检测框的性能评价时，判定某一样本是预测正确还是错误，则是根据其交并比（IoU）来判定的。所谓 IoU，是指是预测框与原标记框的交叠率，即它们的交集与并集的比值需要注意的是这里的预测框和标记框都是指多点模式标注的多边形，而不仅仅是矩形框。在计算时，可以统一按照多边形的模式，计算其 IoU。最理想情况是完全重叠，即比值为 1。具体可参见图 2.1.1。计算公式为：

$$\mathrm{IoU} = \frac{\mathrm{area(C)} \cap \mathrm{area(G)}}{\mathrm{area(C)} \cup \mathrm{area(G)}} \qquad \text{（式 2-1）}$$

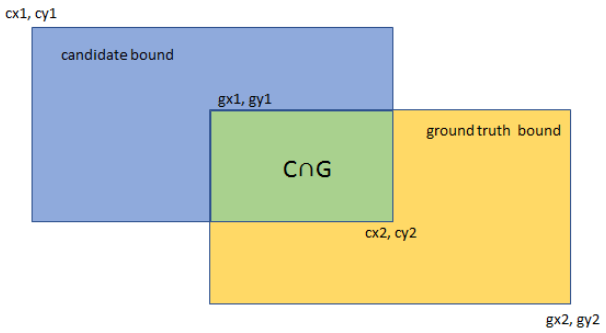

图 2.1.1 IoU 示意图

2.1.3 Precision 与 Recall

在进行分类模型性能评价时，常用的指标包括召回率（Recall），精确率（Precision）。在实际业务中，往往还会将每一个类别具体的精确率和召回率作为评价指标，及 C_Recall 和 C_Precision。

精确率是针对预测结果而言的，它表示的是预测为正的样本中有多少是真正的正样本。预测为正有两种可能，一种就是把正类预测为正类(TP)，另一种就是把负类别预测为正类别(FP)；预测为负也有两种可能，一种是把负类预测为负类(TN)，另一种是把正类预测为负类(FN)。因此有

$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{（式 2-2）}$$

召回率是针对原有样本而言的，它表示的是样本中的正例有多少被预测正确了。也有两种可能，一种是把原来的正类预测成正类(TP)，另一种就是把原来的正类预测为负类(FN)，即

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \qquad \text{（式 2-3）}$$

具体可参见图 2.1.2。

图 2.1.2 Precision 与 Recall

相对应的，C_Recall 和 C_Precision 计算公式如下：

$$\text{C\_Precision} = \frac{TP_c}{TP_c+FP_c} \quad \text{（式 2-4）}$$

$$\text{C\_Recall} = \frac{\text{TP}_c}{\text{TP}_c+\text{FN}_c} \quad \text{（式 2-5）}$$

一般而言，精确率与召回率是此消彼长的关系。对于具体的某一个模型而言，如果想要保证较高的精确率，那么相应的召回率可能会比较低；反之亦然。如果想要同时提高精确率和召回率，则需对现有的网络模型加以改进，提升模型的整体性能。需要注意的是这两个指标与置信度阈值强相关，在实际使用时需要注意。

2.1.4 F1_Score

F1_Score 又称平衡 F 分数，同时兼顾了分类模型的精确率和召回率，其最大值是 1，最小值是 0，计算公式如下：

$$\text{F1} = 2 * \frac{Precision*Recall}{Precision+Recall} \quad \text{（式 2-6）}$$

F1_Score 的公式可以看出 Recall 或者 Precision 较小的那个将会决定 F1_Score 结果，即具有短板效应，而均值的方法不具有这样的效果。例如 Recall=1，Precision≈0 的情况下 F1_Score≈0。F1_Score 比均值的方法更能说明一个模型的好坏，因为很多时候都需要 Precision 和 Recall 的均衡，任意一个指标太差都是无法接受的。