英伟达-阿里云异构计算TensorRT加速AI推理Hackathon2021挑战赛说明

项目介绍：准备基于

<https://github.com/PaddlePaddle/PaddleDetection>，项目进行TensorRT AI模型加速的引擎

生成，主要考虑几点：

1. 相比较于已经广泛应用的tensorflow和torch框架，paddle的框架在异构部署方面难度更高，因为paddle本身的定位是对算法更具亲和性，所以引入很多非工程常规的算子，例如yolo\_box, deformable\_conv, multiclass\_nms等。这些算子属于paddle中的原子粒度，无法通过普通的图优化或者转义方式进行面向TensorRT的映射，需要特殊处理，挑战比较大。
2. Paddle目前的使用度快速上升，尤其是对于中国用户和算法工程人员，未来可能会成为一个主要趋势，而paddle对tensorrt的支持目前不多，我们希望借此机会做一个探索
3. 此项目使用人数较多，且有大量预训练的模型，在算法工程人员中使用广泛，有一定影响力

***具体选择的模型为：yolo v3***，组网模式选择使用paddledetection这个项目的yolo v3的configure：

<https://github.com/PaddlePaddle/PaddleDetection/blob/release/2.0-rc/configs/dcn/yolov3_r50vd_dcn_db_obj365_pretrained_coco.yml>

这个变种里面的plugin有不同解决方案，可以进行除了手写cuda之外的探索，这个yolo 系列用的也比较多，有通用性

人员组成：

队长：孟通，阿里安全AI平台基础架构组软件开发工程师，负责模型编译相关工作

队员：刘彬，阿里安全AI平台基础架构组软件开发工程师，负责算法平台整体架构，runtime框架开发维护等工作

队员：田惠舟，阿里安全AI平台基础架构组软件开发工程师