

《智能计算系统》实验考核标准

一、实验 1: BANGC 算子实现与 TensorFlow 的集成

- 60 分标准: 完成 PowerDifference BCL 算子的基本实现以及基于 CNRT 的测试, 在测试数据中精度误差在 1%以内, 延时在 100ms 以内;
- 70 分标准: 在 60 分的基础上, CNRT 精度误差在 0.1%以内, 延时在 50ms 以内;
- 80 分标准: 在 70 分的基础上, 完成 TensorFlow 的算子集成, 包括 cnplugin 集成与 TensorFlow 的编译; 补全 python CPU、MLU 单算子测试程序 (测试数据与 CNRT 测试不同, 规模更大-256*256*3)。使用 python 在 CPU 端精度误差在 0.1%以内, 延时不做要求; 在 MLU 端精度误差在 10%以内, 平均延时在 50ms 以内;
- 90 分标准: 在 80 分基础上, 使用 python 在 MLU 端测试大规模数据时 (256*256*3) 精度误差在 1%以内, 平均延时在 50ms 以内 (提示: BangC 中需要使用多核拆分计算);
- 100 分标准: 在 90 分基础上, 使用 python 在 MLU 端测试超大规模数据时 (例如 256*256*16、512*512*32 且幂指数大于 2) 精度误差在 0.1%以内, 延时显著优于 CPU。

二、实验 2: 实时风格迁移在线推理与离线部署

- 60 分标准: 完成模型量化与 TensorFlow 在线推理程序, 输出 CPU 端、MLU 端精度正确的风格迁移图片, 输出正确的离线模型;
- 80 分标准: 在 60 分的基础上, 完成离线推理程序的编写, 执行离线推理时风格迁移图片精度正常;
- 100 分标准: 在 80 分基础上, 执行离线推理时在图片精度正常的同时使用 BangC 实现 PowerDifference 算子生成的离线模型相比使用纯 CNML 算子库生成的离线模型, 推理时延时更低, 性能更好。

三、综合实验-目标检测网络-YOLOv3

三个综合实验有不同的难度系数分, 该实验难度中等, 难度系数分为 1.1, 学生最终得分为难度系数分乘以相应的评判标准得分。

- 60 分标准: 补全 nms_detection.h 文件, cnplugin 可正常编过;
- 70 分标准: 在 60 分的基础上完成 TensorFlow 的集成编译, 完成 pb 模型添加后处理大算子的操作, 执行测试脚本时 map 值高于 30%, 单 batch 延时 (包含后处理) 低于 300ms;

- 80 分标准：在 70 分的基础上，执行测试脚本时 map 值高于 50%，单 batch 延时（包含后处理）低于 100ms；
- 90 分标准：在 80 分基础上，执行测试脚本时 map 值高于 54%，单 batch 延时（包含后处理）低于 50ms；
- 100 分标准：在 90 分基础上，执行测试脚本时 map 值高于 56%，单 batch 延时（包含后处理）低于 25ms。

四、综合实验-文本识别 OCR-EAST

三个综合实验有不同的难度系数分，该实验难度一般，难度系数分为 1.0，学生最终得分为难度系数分乘以相应的评判标准得分。

- 60 分标准：完成 BANGC 算子的实现与 CNRT 测试，CNRT 测试时延时低于 30ms；
- 70 分标准：在 60 分的基础上完成 TensorFlow 的算子集成，包括 cnplugin 集成与 TensorFlow 的编译，完成 pb 模型的修改操作。执行测试脚本时，修改后的模型精度与修改前误差在 5%以内，且延时不高于原始模型；
- 80 分标准：在 70 分的基础上，执行测试脚本时，修改后的模型精度与修改前误差在 1%以内，且延时至少优于原始模型 10ms；
- 90 分标准：在 80 分基础上，执行测试脚本时，修改后的模型精度与修改前误差在 0.1%以内，且延时至少优于原始模型 25ms；
- 100 分标准：在 90 分基础上，执行测试脚本时，修改后的模型精度与修改前完全一致，且延时至少优于原始模型 40ms。

五、综合实验-自然语言处理-BERT

三个综合实验有不同的难度系数分，该实验难度偏高，难度系数分为 1.2，学生最终得分为难度系数分乘以相应的评判标准得分。

- 60 分标准：完成 BatchMatMulV2 BANGC 算子的实现与 cnplugin 的集成，可以跑通单算子测试程序；
- 70 分标准：在 60 分的基础上，执行单算子测试脚本时，错误率在 1%以内；
- 80 分标准：在 70 分的基础上，完成 TensorFlow 算子集成，执行完整的 BERT 模型推理验证脚本，f1 值大于 80、exact_match 值大于 70，单 Batch 单次推理平均延时小于 100ms；
- 90 分标准：在 80 分基础上，执行测试脚本时，f1 值大于 82、exact_match 值大于 73，单 Batch 单次推理平均延时小于 60ms；
- 100 分标准：在 90 分基础上，执行测试脚本时，f1 值大于 82.5、exact_match 值大于 74，单 Batch 单次推理平均延时小于 50ms。

六、最终课程成绩的建议评分标准

我们建议全部实验考核，对于必须笔试和必须参考考勤考核的学校，请老师们依照实际情况调整。

如果可以全部实验考核，我们建议如下：

考核内容	所占比重
实验 1（BANGC 算子实现与 TensorFlow 的集成）	20%
BANG 算子 4 选 1 实验	5%
实验 2：实时风格迁移在线推理与离线部署	20%
综合实验 3 选 1	50%
实验报告	5%