BANGC 算子与风格迁移实验手册（下）

一、实时风格迁移在线推理与离线部署

1.实验内容：

a) 模型量化：将单精度模型量化为 INT8 模型。

b) 在线推理，完成 MLU 推理代码的移植工作，分别比较三种不同模型（原始模型、PowerDifference 算子模型，Numpy 算子模型）在 CPU 和 MLU 上的性能和精度。

c) 离线推理，完成 MLU CNRT 离线推理代码，并完成部署测试。

2.实验步骤：

a) 模型量化：MLU270 对于 mlp 和 conv 等算子可使用 int 数据类型进行计算以提高效率，所以先要将原始的 float32 数据类型的 pb 模型量化成为 int 类型。在 /opt/AICSE-demo-student/demo/style_transfer_bcl/tools/fppb_to_intpb 目录中执行：python fppb_to_intpb.py udnie_int8.ini。生成的模型会保存在 /opt/AICSE-demo-student/demo/style_transfer_bcl/models/int_pb_models 目录下。

b) 在线推理程序：分别补全/opt/AICSE-demo-student/demo/style_transfer_bcl/src/online_mlu/transform_mlu.py 和 /opt/AICSE-demo-student/demo/style_transfer_bcl/src/online_cpu/transform_cpu.py。比较多种不同推理方式的性能精度差异。

c) 离线推理程序：在执行 MLU 在线推理时，通过配置文件生成并保存两种离线模型（原始模型、PowerDifference 算子模型）至 /opt/AICSE-demo-student/demo/style_transfer_bcl/models/offline_models 文件夹下。补全 CNRT 离线推理模型代码：/opt/AICSE-demo-student/demo/style_transfer_bcl/src/offline/src/inference.cpp 然后编译并执行离线推理。