

实验报告

课程名称: _____

实验类型: _____ 操作实验 _____

实验项目名称: _____ 实验一: 基于 Neurosurgeon 的云边协同推理实验 _____

姓名: _____ 学号: _____

QQ 号码: _____

(高校联合班成员填写: 学校_____姓名_____学号_____)

实验日期: _____ 年 _____ 月 _____ 日

了解云边协同计算的基本原理和架构

一、实验目的和要求:

- 了解云边协同计算的基本原理和架构
- 理解深度学习模型分割推理的机制和方法
- 掌握基于 AlexNet 模型的动态切分点选择技术
- 学习通过网络带宽控制工具进行性能测试的方法
- 分析不同网络条件下云边协同推理的性能表现

二、实验内容

基于 Neurosurgeon 框架以及 AlexNet 深度学习模型, 根据实验代码库 (<https://github.com/csmVic/edge-cloud-experiment>) 完成以下实验

云边协同推理性能测试实验

- ✓ 配置不同的边缘端 CPU 资源, 例如 (0.3, 0.5, 0.6, 0.7)
- ✓ 通过 tc 工具调节网络带宽, 例如 (5mbit/s, 7mbit/s, 10mbit/s)
- ✓ 分析模型结构, 测试不同切分点下的端到端推理延迟
- ✓ 分析本地计算时间、数据传输时间和云端计算时间
- ✓ 记录实验内容, 找到在哪一种设备以及网络带宽下, 云边协同的端到端延迟最低

三、实验背景

- DNN 云边协同工作

云边协同旨在充分利用云边端资源完成 DNN 任务的推理计算, 将整体模型进行划分后, 利用终端设备、边缘服务器以及云计算中心的计算资源, 将 DNN 划分为多个部

分，分别部署在不同设备上进行推理。

1. 充分利用系统中可用的计算资源

2. 降低输入数据的传输开销

- Neurosurgeon 框架介绍

Neurosurgeon 是一个用于深度神经网络(DNN)分区的预测框架,旨在优化移动设备上的 DNN 推理性能。该框架通过在移动设备和云端之间智能地分割神经网络层,来最小化端到端的推理延迟。Neurosurgeon 能够根据网络条件、设备性能和模型结构动态选择最优的切分点,实现云边协同推理的性能优化。

四、 主要仪器设备

- 边缘计算设备

高性能边缘设备香橙派

- 云端服务器

高性能边缘设备香橙派模拟

- 软件环境

代码运行环境: PyTorch

部署环境: Kubernetes, Docker.

五、 实验题目简答

a) 什么是云边协同计算? 它在物联网应用中有哪些优势?

b) AlexNet 模型的切分点选择对推理性能有什么影响? 请分析你会选择切分点 xxx 的原因?

c) 代码中云端和边缘端是使用什么方式来进行通信的?

六、 实验数据记录和处理

[tips](#) (实验文件 -> deployment.yaml 中) :

1. 调节网络带宽 (调节范围: 1-9mbit)

```
cpu: 0.6
requests:
  cpu: "0.6"
env:
- name: PYTHONUNBUFFERED
  value: "1"
args: ["/bin/bash", "-c", "tc qdisc add dev eth0 root tbf rate 5mbit burst 12kbit latency
codePath: usr/Demoac/.edge/edge-cloud/
ports:
- containerPort: 98
```

2. 调节网络资源（调节范围：0.2 - 0.8，一位小数点，注意 limits 和 requests 设置相同值）

```
privileged: true
resources:
  limits:
    cpu: "0.6"
  requests:
    cpu: "0.6"
env:
```

以下实验记录均需结合屏幕截图，进行文字标注和描述（看完请删除本句，以及以上tips）。

1) 实验环境配置截图

边缘资源配置：0.6

```
securityContext:
  privileged: true
resources:
  limits:
    cpu: "0.6"
  requests:
    cpu: "0.6"
env:
- name: PYTHONUNBUFFERED
  value: "1"
args: ["/bin/bash", "-c", "tc qdisc add dev eth0 root tbf rate 5mbit burst 32kb
codePath: usr/Demoac/.edge/edge-cloud/
ports:
```

边缘带宽配置：5mbit

```
cpu: 0.6
requests:
  cpu: "0.6"
env:
- name: PYTHONUNBUFFERED
  value: "1"
args: ["/bin/bash", "-c", "tc qdisc add dev eth0 root tbf rate 5mbit burst 12kbit latency
codePath: usr/Demoac/.edge/edge-cloud/
ports:
- containerPort: 98
```

2) 元物云平台终端运行截图

edge-test:

```
edge-test cloud-test

2025-07-06 16:38:58: websocket connect success
2025-07-06 16:38:58: get bandwidth value : 0.5857340159296119 MB/s
2025-07-06 16:38:58: 手动设置切分点: partition_point = 22
2025-07-06 16:38:58: 最终切分点: partition_point = 22
2025-07-06 16:38:58: short message , model type has been sent successfully
2025-07-06 16:38:58: short message , partition strategy has been sent successfully
2025-07-06 16:38:58: 设备预热中...
2025-07-06 16:39:16: CPU Warm Up : 1329.365ms
2025-07-06 16:39:16: =====
2025-07-06 16:39:56: alex_net 在边缘设备上推理完成 - 1331.246 ms
2025-07-06 16:39:56: 中间特征值占用字节 edge_output.nbytes = 4000 bytes
2025-07-06 16:39:56: 中间特征值序列化后大小 = 4396 bytes
2025-07-06 16:39:56: get yes , edge output has been sent successfully
2025-07-06 16:39:56: alex_net 传输完成 - 2.674 ms
2025-07-06 16:39:56: alex_net 在云端设备上推理完成 - 0.011 ms
2025-07-06 16:39:56: 端到端推理时间 - 1333.931 ms
2025-07-06 16:39:56: ===== DNN Collaborative Inference Finished. =====
```

cloud-test:

```
edge-test cloud-test

2025-07-16 22:15:52: websocket connect success
2025-07-16 22:15:52: successfully connection :<socket.socket fd=5, family=AddressFamily.AF_INET, t
'10.244.3.161', 53368]>
2025-07-16 22:15:52: get model type successfully.
2025-07-16 22:15:52: get partition point successfully.
2025-07-16 22:16:32: get edge_output and transfer latency successfully.
2025-07-16 22:16:32: short message , transfer latency has been sent successfully
2025-07-16 22:16:32: 设备预热中...
2025-07-16 22:16:32: CPU Warm Up : 0.007ms
2025-07-16 22:16:32: =====
2025-07-16 22:16:32: short message , cloud latency has been sent successfully
2025-07-16 22:16:32: ===== DNN Collaborative Inference Finished. =====
>>
```

3) 数据分析记录截图

cpu	带宽	切分点	end-end	本地	传输	云端
0.6	5mbit/s	0	1774.445	0.015	976.728	797.702
0.6	5mbit/s	8	1444.430	727.729	361.846	354.855
0.6	5mbit/s	13	1287.431	1137.941	36.436	113.054
0.6	5mbit/s	22	1327.931	1325.248	2.678	0.011

七、实验结果与分析

通过上述实验和相关资料学习，分别解答以下问题（看完请删除本句）：

- 基于该实验，谈谈你对云边协同推理计算的理解

