# 项目计划书

(2021218)在 openEuler aarch64 架构上完成 mlpack 基于公开数据集完成 mnist 训练过程

## 简历


**基本信息** **姓名：**刘文卓

**Email：**lwzbill@foxmail.com

**GitHub 账号：**Enter-tainer

**网站：**https://margatroid.xyz/


**教育经历** **华中科技大学**

计算机科学与技术学院，卓越工程师实验班，大一

**获奖经历** **全国青少年信息学奥林匹克 NOIP 2017**

省级一等奖

**相关经验** 实现过基于 BERT 和卷积神经网络的新闻话题分类

实现过基于 Linux 系统调用的支持重定向，管道等功能的 Shell

实现过基于 ucontext 的 C 语言有栈协程库

长期使用 Linux，熟悉 Linux 系统

熟悉 Git 使用，有协作开发的经验

## 项目

### 项目名称

在 openEuler aarch64 架构上完成 mlpack 基于公开数据集完成 mnist 训练过程

### 项目分析

mlpack 是一个 C++机器学习库，支持决策树，KNN，SVM，神经网络，朴素贝叶斯，adaboost 等机器学习算法。本项目计划在 openEuler aarch64 架构上编译 mlpack，并使用其中的支持向量机，神经网络等算法来完成基于 mnist 数据集的手写数字识别模型训练。

**项目大纲**

1. 配置 openEuler aarch64

   首先，我会去配置虚拟机并安装 openEuler aarch64。然后通过查阅 openEuler 文档，熟悉 openEuler 的使用并配置开发工具链，为接下来的工作进行做好准备。

2. 编译 mlpack

   在进行这部分工作之前，我会先检查一下其它基于 CentOS aarch64 的发行版对 mlpack 的打包和处理，争取与业界的普遍实践接轨。然后，我会分析 mlpack 的依赖，接着编译并打包需要的依赖库。由于 aarch64 架构有一些特殊，因此在编译的过程中可能会遇到一些困难，如果我遇到了这种情况，我会通过查阅编译器和依赖库的文档来尝试调整编译参数，对代码进行修改，进而解决问题。

   编译完成后，我会运行测试，确保 mlpack 工作正常，性能符合预期。我会尝试链接 openBLAS 等加速库来提高 mlpack 的运行性能。

3. 打包 rpm 软件包并编写文档

   在编译完成后，我会整理编译流程，然后通过阅读 rpm 等包管理器的文档，和 openEuler 社区的打包规范，把构建产物打包成易于使用的软件包，并对其进行测试，确保其工作正常，依赖关系正确，可以正确的安装、升级、卸载。然后总结经验，编写软件包文档。

4. 训练模型

   在进行这部分工作之前，我会首先找到适合训练使用的 MNIST 数据集，然后查阅 mlpack 文档，学习 mlpack 的基本使用方法。接下来，我会使用 mlpack 导入数据集，将其分为训练集和测试集，然后使用 mlpack 建立并训练模型。我将会使用多种不同的模型，并对比它们的效果。我目前计划使用支持向量机，全连接神经网络和卷积神经网络三种模型。并对比模型的收敛速度，准确率，F1-score 和内存占用，推理速度等指标。在训练完成后，我还会编写一些胶水代码，使得我们可以导入模型，并对图片进行推理，从而识别数字。

5. 总结项目并编写文档

   在以上的部分完成之后，我会总结整个项目，将代码整理后推送至指定的仓库和平台。然后根据我在项目中的参与经验，写一写我在项目进行中遇到的困难和得到的收获，并发布到知乎等技术平台上，从而推广 openEuler 社区，来让大家熟悉 openEuler 上的 AI 软件栈和开发流程，让更多的人参与到开源社区的建设中来。

**项目难点**

1. mlpack 的编译

   由于 openEuler 是 aarch64 架构的，有可能会遇到兼容性的问题。目前，CentOS，Arch Linux 等其它发行版已经有 mlpack 的在 aarch64 平台上的软件包，这证明了在 aarch64 架构的 Linux 系统上编译 mlpack 的可行性。但我仅在 x86_64 的 Linux 系统上编译过 mlpack，在 aarch64 架构上仍需要探索。

2. rpm 软件包的打包

   我仅尝试过 Arch Linux 的软件打包，不是很了解 openEuler 的软件包管理方式，对 rpm 的打包流程也不是很了解。需要查阅文档学习。我会通过阅读 openEuler 社区的打包规范，和 rpm 的文档和指南来进行打包。

3. mlpack 的使用

   现在广泛流行的深度学习框架是 Python 语言的 Tensorflow 和 Torch。因此，相对而言，mlpack 的资料会少一点，这可能会对项目的进行带来一定的阻碍。

我相信，这些问题虽然复杂，但是都是可以通过阅读文档和刻苦钻研解决的。

**预期时间表**

| 日期 | 工作 |
| --- | --- |
| 7 月 1 日至 7 月 7 日 | 安装虚拟机并配置工具链 |
| 7 月 8 日至 7 月 14 日 | 检查其它发行版对 mlpack 的处理，学习最佳实践 |
| 7 月 15 日至 7 月 21 日 | 分析 mlpack 的所有依赖，列出依赖列表并检查兼容性，如果出现不兼容的情况，需要进一步解决兼容性问题。 |
| 7 月 22 日至 7 月 28 日 | 编译 mlpack 和它的所有依赖，并进行测试 |
| 7 月 29 日至 8 月 4 日 | 阅读 rpm 文档，学习打包方式 |
| 8 月 5 日至 8 月 10 日 | 将构建产物打包成软件包并进行测试 |
| 8 月 11 日至 8 月 15 日 | 编写软件包文档并准备中期报告 |
| 8 月 16 日至 8 月 22 日 | 阅读 mlpack 文档，学习使用方法 |

| | |
|---|---|
| 8 月 23 日至 8 月 29 日 | 建立支持向量机模型并进行训练 |
| 8 月 30 日至 9 月 5 日 | 建立全连接神经网络模型并进行训练 |
| 9 月 6 日至 9 月 12 日 | 建立卷积神经网络模型并进行训练 |
| 9 月 13 日至 9 月 18 日 | 对比三种模型的效果 |
| 9 月 19 日至 9 月 25 日 | 编写胶水代码，使得可以模型可以推理手写数字的图片 |
| 9 月 25 日至 9 月 30 日 | 总结项目并编写技术文档，心得体会，准备最终项目报告 |

# 额外信息

**参与原因**

自从华为发布 openEuler 以来，我一直对它有着浓厚的兴趣，希望有一天可以体验到它。感谢 OSPP 提供了这一机会，让我能得到 openEuler 社区的资深导师的指导，来对社区做出贡献。如果我有机会能参与到这个项目中，我一定会不遗余力地把项目做好。同时，在项目的进行过程中，我会保持与导师和社区的联系，积极听取经验，接受指导，参与社区的建设。

我在大数据课程的课程设计中，使用 BERT 和卷积神经网络完成了简单的新闻话题分类任务。虽然我没有使用过 mlpack，但我相信框架的思想都是相通的，我目前也正在熟悉 mlpack；我一直在使用 Arch Linux，并基于系统调用实现了简单的 Shell 和一部分工具。我相信我对 Linux 的理解和对 AI 的经验能让我胜任这个项目。