开源软件供应链点亮计划 -- 暑期 2020 -- 项目申请书

“将 xgboost 与 prophet 等数据科学软件库集成到 Debian GNU/Linux”

申请人：周默

2020 年 6 月 19 日

1. 项目背景：

xgboost （梯度提升），prophet （时间序列预测）以及 Stan （高性能统计建模），pytorch（深度学习)等是重要的数据科学软件。Debian 作为用户基数庞大的 Linux 发行版之一，用户群中不乏数据科学用户群体，而上述软件尚未集成到该系统中，或者缺乏一部分计算性能相关的依赖库。为帮助用户获得更佳使用体验，方便安装及部署，该项目旨在为 Debian 官方仓库引入这些软件及其对应的依赖库，并将仓库中已有的依赖更新到理想状态。

2. 项目详细方案：

项目中涉及到的软件集成到 Debian GNU/Linux 主要有以下几方面问题：（1）部分上游使用了嵌入的快照版本三方库，不利于集成 （2）部分上游的编译系统不支持使用本地依赖库，需要改进这部分的代码（3）这些项目缺少对应的用于集成到 Debian APT/DPKG 系统中的脚本（4）这些软件在集成之后缺少持续可靠的测试脚本。

对于（1）和（2）问题，我们可以采用对上游项目的 CMake 等文件打补丁并提交 PR 的方式进行修复。对于（3）我们可以通过编写基于 debhelper 的一系列脚本来完成。对于（4）则可以根据具体的项目选择玩具数据集进行模型训练，并作为功能性测试：例如，对于 Pytorch 的测试，我们可以编写测试案例使用 MNIST 一类数据集进行分类器训练，看模型是否能够收敛，并且达到合理的分类正确率。

3.项目开发时间计划：

7 月 1 日-7 月 7 日（1 周）：将现有的 Pytorch 软件包更新到 1.5.1 版本。并添加在 Fashion-MNIST 上的分类器训练脚本作为测试脚本。
7 月 8 日-7 月 14 日（1 周）：集成 pytorch-text 文本预处理库，并编写相应的测试脚本。
7 月 15 日-7 月 21 日（1 周）：集成 pytorch-audio 音频预处理库，并编写相应的测试脚本。
7 月 22 日-7 月 28 日（1 周）：集成 pytorch 下的 FBGEMM 与 tensorpipe 库。
7 月 29 日-8 月 4 日（1 周）：集成 ideep 库和 dmlc/rabit 库。
8 月 5 日-8 月 11 日（1 周）：集成 xgboost 库，并编写相应的测试脚本。
8 月 12 日-8 月 25 日（2 周）：集成 stan 的基础 math 库等 stan 的 C++依赖。
8 月 26 日-9 月 1 日（1 周）：集成 cmdstan 以及 stan 的 python 接口。
9 月 2 日-9 月 8 日（1 周）：集成 prophet 时间序列预测库。
9 月 9 日-9 月 29 日（3 周）：启用已经集成的 pytorch 的更多功能，并改进上游代码，使用系统提供的依赖库进行编译。同时确认 ideep，fbgemm 等依赖库与 pytorch 的兼容性，并在 pytorch 中启用。
9 月 30 日-9 月 31 日（2 日）：修复剩余 bug，整理总结。