MegEngine Web Profile

User's Guide

杨小燕

`Version: 1.0.0`
`Date: 09.30.2021`

1 绪论

当深度学习模型完成训练开始部署、推理阶段，模型的推理速度、性能往往受到关注。常见的 profile 工具一般用于分析模型的执行时间、流程、内存消耗等。

构建 web profile 工具，可以让用户方便地通过网页，查看生动的 profile 结果，分析模型的运行状况；后台支持 x86、cuda、arm 等多种平台，便于用户分析模型在不同平台下的性能。

2 系统设计

2.1 总体架构设计

- 用户通过 web 前端提交 mge 模型和输入数据，选择后端平台和 MegEngine 版本；用户可选择使用具有权限的私有设备
- Server 后端为用户请求生成唯一的任务 ID，并通过数据库维护任务当前状态
- 任务流程：Server 后端生成-> 队列中等待运行资源->Worker 获取任务、运行并返回结果->Server 存储结果至数据库并等待前端获取
- 前端使用 Vue 实现
  - 使用`megengine.tools.network_visualize`的`visualize`函数,可实现模型结构的描述，结合开源项目`Netron`，可实现前端模型结构的展示
  - 使用`megengine.tools.profile_analyze`，可以生成模型的 profile json 结果，并根据用户需要在前端显示结果
- 后端（Server、Worker 等）使用 Python
  - 用户登录验证用户名密码匹配后，生成用户 token（用户令牌），并且设置登录时效性，将 token 返回给前端设置本地缓存
  - Worker 编译 MegEngine 的`load_and_run`运行模型
- 对象存储服务使用 MinIO

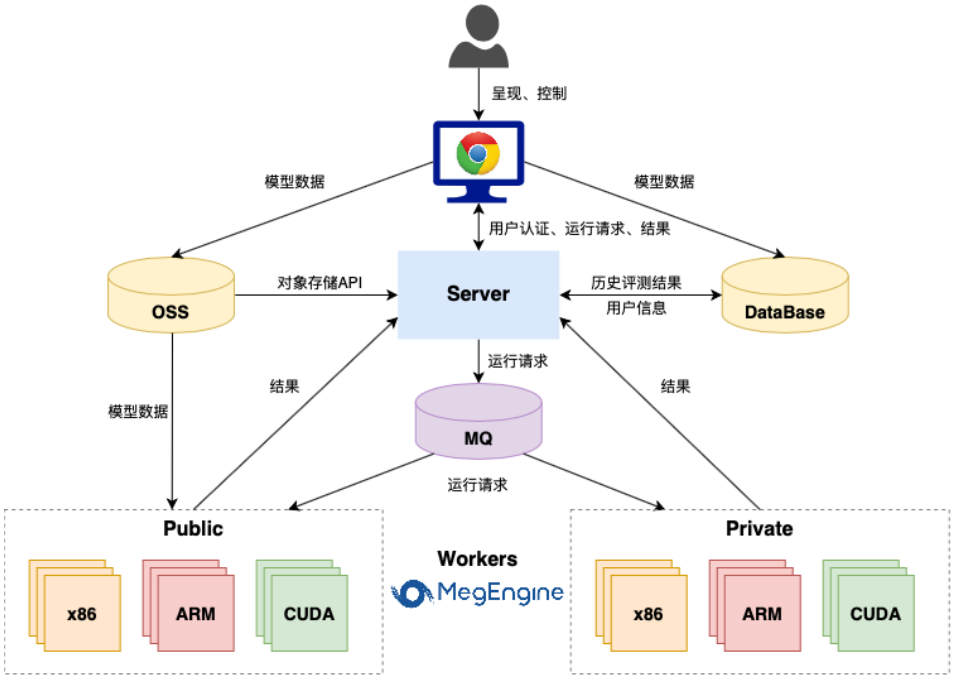

图 1: 系统总体架构设计

2.2 系统数据流程

- 用户进入 MegEngine Web Profile 网站，首先进行注册，填写相关信息，注册成功后使用其注册账号登录。Server 端数据库记录用户信息。
- 在“新建任务”界面，用户可以看到具有权限的所有 Worker 信息，用户从本地上传 mge 模型和数据文件，并选择 Worker 和 MegEngine 版本，Server 端数据库记录信息后，生成 MinIo 的预签名上传 URL 返回给前端，文件由前端直接上传至对象存储 MinIO
- 新建任务在 MQ 队列中等待运行资源。Worker 获取新任务后，从 Server 端获得相应的 MinIo 的预签名下载 URL，将模型和数据下载后并运行获得结果，并上传结果至 MinIo，更新 Server 端数据库中的任务信息
- 在“任务列表”界面，用户可以看到所有提交任务的信息及状态，信息由 Server 端读取数据库并返回至前端，查看结果或模型结构会跳转至相应界面
- 在“任务结果”界面，用户选择任务并设置属性，后端处理运行输出结果并返回信息至前端显示
- 在“模型预览”界面，用户选择任务，Server 端生成模型描述文件并存储至 MinIo，生成 MinIo 预签名 URL 返回前端，前端嵌入的 Netron 中显示可视化模型
- 管理员用户可查看所有 Worker 信息、用户信息，并有设置管理员、修改用户密码、给用户增加私有 Worker 权限等功能

3 功能介绍

3.1 用户功能

3.1.1 用户登录

在用户没有登录账号的情况下，打开网站首先进入登录页面 (图2)。

图 2: 登录页面

用户通过注册时填写的邮箱和密码进行登录，若邮箱密码错误或未注册，仍停留在登录页面并弹出相应提示。若用户未注册，需点击页面上的注册按钮先注册，再登录。

3.1.2 用户注册

注册时用户需填写其常用邮箱，且是未注册过的。在填写过程中会对邮箱和密码的格式进行校验和提醒。填写完点击注册后，注册成功则跳转回登陆页面；注册失败会返回相应的错误信息。(图3)

3.1.3 新建任务

在上方为“新建任务”处选择本地 mge 模型文件、数据文件、具体 Worker 及 MegEngine 版本，数据文件可上传多个，但文件名需与 npy 中的变量名一致。点击创建按钮创建任务 (图4)。提示：mge 模型较大时，上传需要一定时间，请等待“创建”按钮的缓冲结束。

页面下方显示了当前用户所有可用的 Worker 信息，包括平台上所有公有（public）Worker 及该用户拥有权限的私有（private）Worker（图5）。

图 3: 注册页面

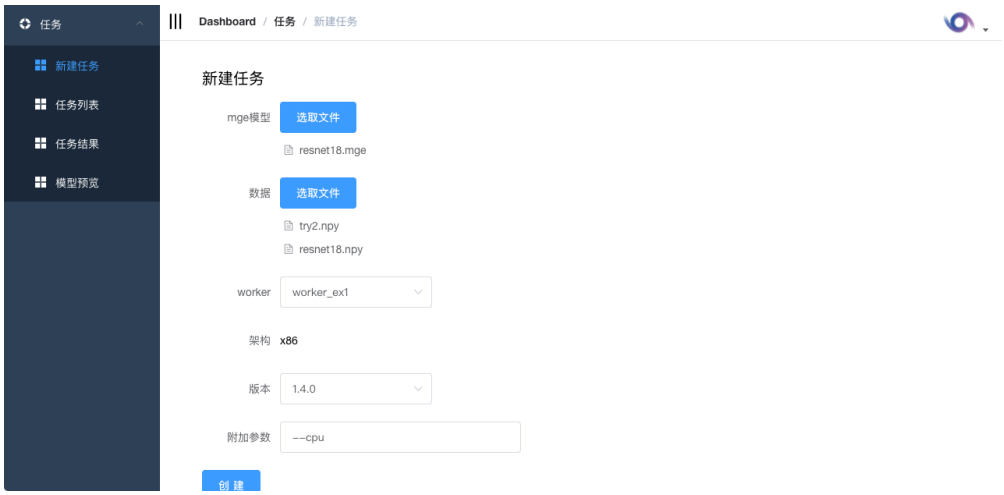

图 4: 新建任务

3.1.4 任务列表

页面展示了当前用户所有任务的信息，包括任务 ID、MGE 模型名称、数据 DATA 名称、Worker 名称、MegEngine 版本、任务更新时间、任务状态。其中任务状态 “initiated” 表示任务信息初次提交成功；“waiting” 表示任务文件已上传成功，正在队列中等待资源；“running” 表示任务正在被 Worker 运行；“succeeded” 表示任务运行结束且成功；“failed” 表示任务结束但失败。(图6)

点击相应行 “结果” 列的 “查看” 按钮，对于运行完成且失败的任务，可查看错误日志 (图7)；对于运行完成且成功的任务，将跳转至任务结果页面，待用户补充其余信息。

点击相应行 “模型” 列的 “查看” 按钮，将跳转至 “模型预览界面” 并展示模型结构。

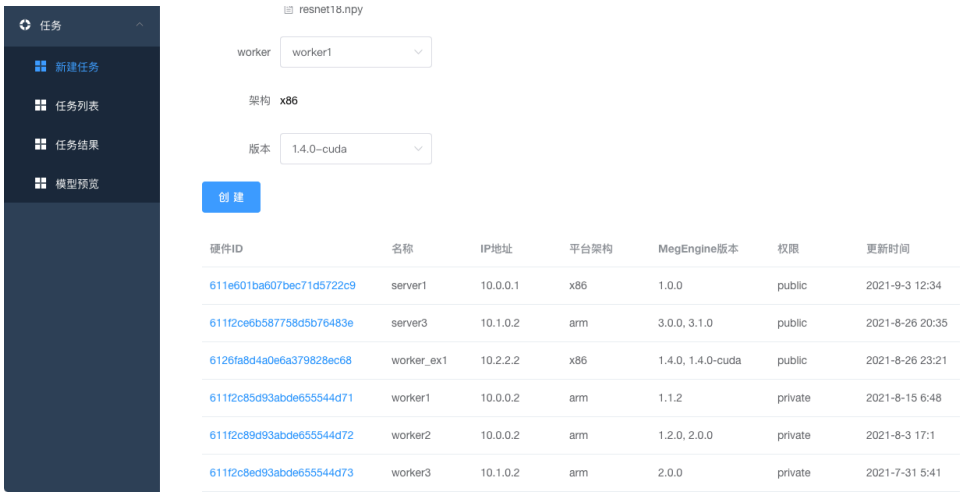

图 5: 用户 Worker 信息

图 6: 任务列表

3.1.5 任务结果

点击“查看结果”按钮，在弹出框中选择任务并设定其他配置（图8），点击“查看”即可看到模型 Profile 结果（图9）。

3.1.6 模型预览

选择任务 ID 并点击“显示”按钮，即可查看模型结构，点击算子可在页面右边弹框中查看算子具体信息（图10）。

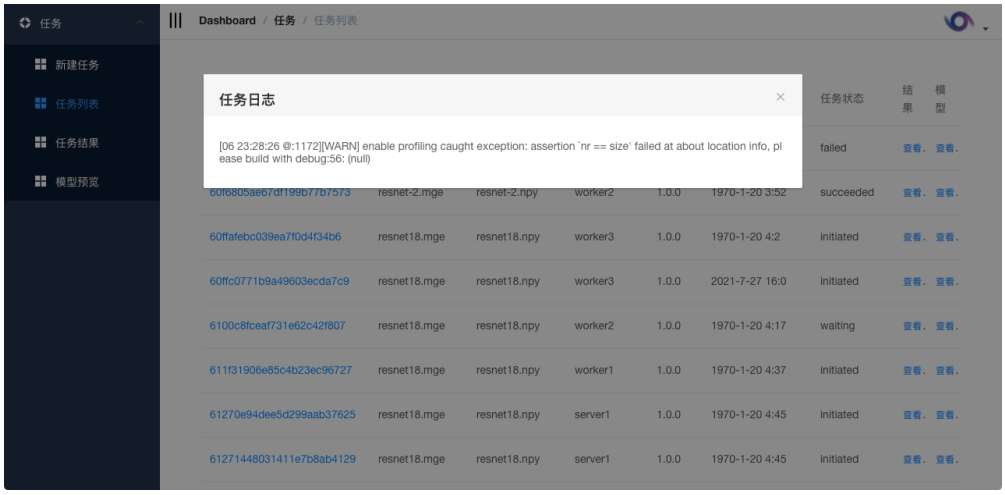

图 7: 失败任务日志

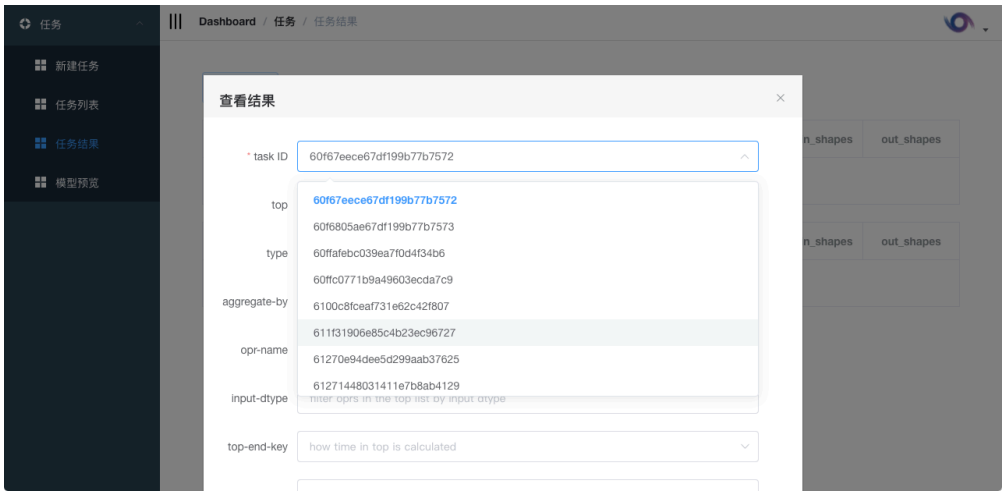

图 8: 查看结果

3.2 管理员功能

3.2.1 硬件信息

该页面显示所有 Worker 的信息，包括 WorkerID、名称、IP 地址、平台架构、所有 MegEngine 版本、上次更新时间及权限（图11）。其中上次更新时间为上次 Server 接收到该 Worker 心跳的时间，若该项长时间未更新，极有可能是 Worker 掉线，请及时查看并处理。

3.2.2 用户权限

该页面显示所有用户的信息，包括用户 ID、邮箱、角色、私有 Worker 列表（图12）。点击对应行的“新增”（“删除”）按钮，在弹框中选择 Worker 名称，即可给该用户增（删）此

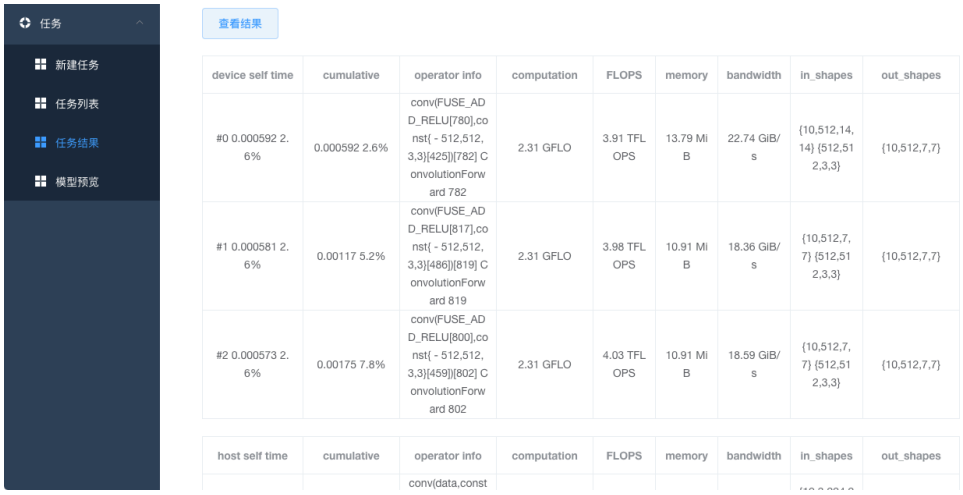

图 9: 模型 Profile 结果

图 10: 模型结构预览

私有 Worker 的使用权。

点击页面上方“添加管理员”（“删除管理员”）按钮，在弹框中输入邮箱，即可将邮箱对应用户添加（去除）管理员权限。

点击“修改密码”按钮，在弹框中输入用户邮箱和新密码，即可给该用户修改密码。

图 11: 所有 Worker 信息

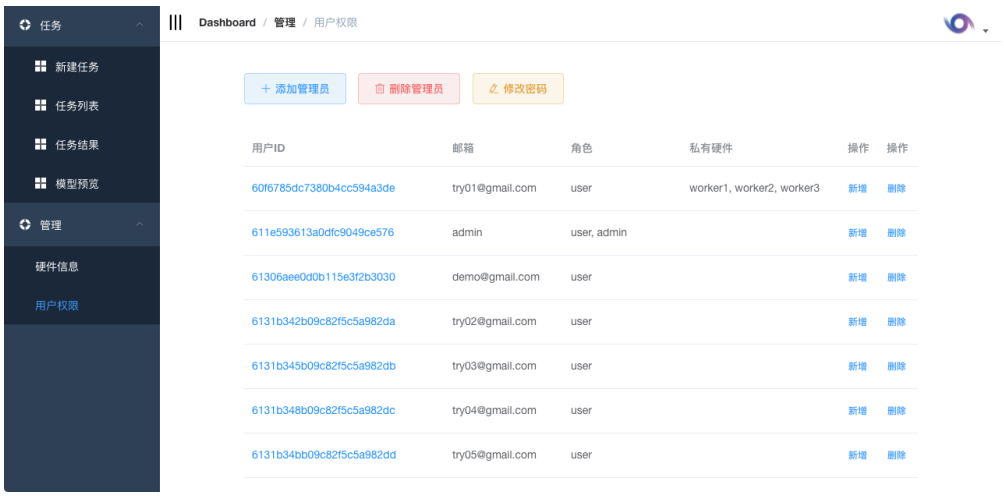

图 12: 所有用户信息