开发者社区

Release Note

AI生态包

DLExamples

DCU 模型测试样例，包含训练及推理场景。

1. DLExamples for Training

20230606

请注意，后续该模块相关内容迁移到 model zoo 进行更新（https://sourcefind.cn/#/model-zoo/list）。

20230206

版本号：dlexamples_v2.2.0

1. 功能改进：
   - 增加各模型多机多卡测试方法；
   - 更新 openmmlab 测试用例，适配 mmcv1.6.1；
   - 添加 Keras,Vision_Transformer,以及 mmaction2 测试用例

20220901

版本号：dlexamples_v1.9.1

1. 功能改进：
   - 测试流程更新至 README 形式维护；
   - 添加 mmsegmentation 相关模型优化测试方法；

20220721

版本号：dlexamples_v1.7.1

1. 功能改进：
   - 删除了僵尸代码用例，修改了 tensorflow 和 pytorch 测试示例文档，同

时支持 dtk21.10 和 dtk22.04 测试；

- PyTorch 对应示例文档版本升级为《基于 pytorch 的 DCU 深度学习测试示例文档_V1.7.1.pdf》；
- TensorFlow 对应示例文档版本升级为《基于 tensorflow 的 DCU 深度学习测试示例文档_V1.7.1.pdf》；

20220712

版本号：dlexamples_v1.7.0

1. 功能改进：
   - 修改了 tensorflow 和 pytorch 测试示例文档，增加支持了 dtk22.04 测试；
   - PyTorch 对应示例文档版本升级为《基于 pytorch 的 DCU 深度学习测试示例文档_V1.7.0.pdf》；
   - TensorFlow 对应示例文档版本升级为《基于 tensorflow 的 DCU 深度学习测试示例文档_V1.7.0.pdf》；

20220506

版本号：dlexamples_v1.0.1

1. 功能改进：
   - 样例代码同时支持 TensorFlow 和 PyTorch 框架在标卡与集群上的测试；
   - GNMT(PyTorch)模型升级为 GNMT_V2 模型；
   - PyTorch 对应示例文档版本升级为《基于 pytorch 的 DCU 深度学习测试示例文档_V1.0.1.pdf》；
   - TensorFlow 对应示例文档版本升级为《基于 tensorflow 的 DCU 深度学习测试示例文档_V1.0.1.pdf》；

20220422

版本号：dlexamples_v1.0.0

1. 功能改进：
   - 建立 DCU 深度网络模型训练样例集；
   - 建立 pytorch 及 tensorflow 的对应使用参考文档；

2. DlExamples for Inferencing

20230606

请注意，后续该模块相关内容迁移到 model zoo 进行更新（https://sourcefind.cn/#/model-zoo/list）。

20221124

版本号：inferexamples-V2.4.1-V2.4.1 –> 76d9045

1. 功能改进：
   - 支持 dtk2210 系列版本基础软件栈；
   - 增加 nlp/bert-squad MIGraphX 应用案例；
   - MIGraphX 对应使用文档升级为《MIGraphX 教程_V2.4.1.pdf》；
   - 增加 MIGraphX 应用文档《MIGraphX 应用文档_v2.4.1.pdf》；

20220901

版本号：MIGraphX_Samples-V2.4.0

2. 功能改进：
   - 支持 dtk2204 系列版本基础软件栈；
   - MIGraphX 对应使用文档升级为《MIGraphX 教程_V2.4.0.pdf》；

20220615

版本号：MIGraphX_Samples_V2.3

1. 功能改进：
   - 支持 dtk2204 系列版本基础软件栈；
   - MIGraphX 对应使用文档升级为《MIGraphX 教程_V2.3.2.pdf》；

20220519

版本号：MIGraphX_Samples_V2.0.1

1. 功能改进：
   - 初始化阶段加了一次推理进行 warmup 操作，性能数据更准确；
   - MIGraphX 对应使用文档升级为《MIGraphX 教程_V2.3.0.pdf》；

Whl Packages

DCU 在人工智能领域的生态包，包含框架、三方组件、automl 组件及通信库等。

1. PyTorch

20230617

git: torch-1.13.1 –> 55d300e

存储位置：whl_dtk23.04

1. 功能改进
   - 支持 dtk23.04 系列软件栈；
   - 新增环境变量 ROCBLAS_ATOMICS_MOD 控制 rocblas_mode，默认开；
   - 调整 MIOpen 库的兼容性；
   - 还原 torch.__version__返回，增加 torch.__dcu_version__查询详细版本；
   - 添加环境变量 USE_MIOPEN_BATCHNORM 控制是否使用 MIOpen 加速库中的 batch norm，默认禁用；
   - 添加环境变量 ROCBLAS_COMPUTETYPE_FP16R 控制 fp16 gemm 计算类型，默认 fp16r；
   - 支持 magma 功能，自带 magma 相关 so 文件；
   - 添加 mkl 支持；
2. 问题修复
   - 修复 NLLLoss 计算错误；
   - 打包 MIOpen 相关头文件；
   - 解决使用 use_package_data=True 时的报错；
   - 修复部分 layer norm 相关计算精度问题；
3. 性能改进
   - Any, Max 等技术算子的性能优化；

4. 已知问题
   - import torch 前，需要导入 torch 安装目录下 lib，例如: export LD_LIBRARY_PATH=/usr/local/lib/python${PY_VERSION}/site-packages/torch/lib:$LD_LIBRARY_PATH（或可通过 yum install intel-mkl-2020.0-088 -y --nogpgchec 安装 mkl）
   - 在较低版本驱动上运行程序，可能会出现进程卡住问题；

git: torch-1.10.0 –> e378c3c

存储位置：whl_dtk23.04

5. 功能改进
   - 支持 dtk23.04 系列软件栈；
   - 调整 MIOpen 库的兼容性；
   - 还原 torch.__version__返回，增加 torch.__dcu_version__查询详细版本；
   - 添加环境变量 USE_MIOPEN_BATCHNORM 控制是否使用 MIOpen 加速库中的 batch norm，默认禁用；
   - 添加环境变量 ROCBLAS_COMPUTETYPE_FP16R 控制 fp16 gemm 计算类型，默认 fp16r；
   - 支持 magma 功能，自带 magma 相关 so 文件；
6. 问题修复
   - 修复 NLLLoss 计算错误；
   - 打包 MIOpen 相关头文件；
   - 解决使用 use_package_data=True 时的报错；
   - 修复部分 layer norm 相关计算精度问题；
7. 性能改进
   - Any, Max 等技术算子的性能优化；
8. 已知问题
   - import torch 前，需要导入 torch 安装目录下 lib，例如: export LD_LIBRARY_PATH=/usr/local/lib/python${PY_VERSION}/site-packages/torch/lib:$LD_LIBRARY_PATH（或可通过 yum install intel-mkl-2020.0-088 -y --nogpgchec 安装 mkl）

- 在较低版本驱动上运行程序，可能会出现进程卡住问题；

20230606

git: torch-1.13.0a0 –> efb907c

存储位置：whl_dtk22.10

9. 功能改进
   - 支持 torch1.13 python3.8；
10. 问题修复
    - 修复部分 DCU Kernel profiler 功能；
11. 已知问题
    - Jit 相关功能检测到非法指令

20221115

git: torch-1.10.0a0 –> 2040069

存储位置：whl_dtk22.04.2

1. 功能改进
   - 新增支持 DCU Kernel profiler 功能；
   - 新增环境变量 ROCBLAS_ATOMICS_MOD 控制 rocblas_mode；
   - 恢复 rocblas atomic 默认状态，仍由 torch.use_deterministic_algorithms 控制；
2. 问题修复
   - 修复部分 layer norm 相关计算精度问题；

20220804

git: torch-1.10.0a0 –> c7f69d6

存储位置：whl_dtk22.04.2

1. 功能改进
   - gitc7f69d6_magma 支持 magma 功能，自带 magma 相关 so 文件

2. 问题修复
   - 修复 apex 支持 mulithead_attn 时 undefined symbol 报错；
3. 性能改进
   - 针对 NLP 领域 fp16 精度进行了优化；
   - upsample backward 优化（默认关，HIP_UPSAMPLE_OPTIMIZE=1 使用优化 kernel）
   - 支持超大规模训练（超 20000 卡测试验证）；

20220604

git: 450cdd1/8da4652

存储位置：whl_dtk22.04.1

1. 功能改进
   - 支持对应官方的 pytorch1.10.0 版本；
   - 支持 python3.7/3.8/3.9 版本；
   - 支持 dtk22.04 系列 dtk 软件栈；
   - whl 命名增加对 dtk 大版本支持编号；
   - 关闭 rocblas atomic 操作，支持 PyTorch 的 Reproducibility 功能；
   - git8da4652 支持 magma 功能，自带 magma 相关 so 文件
2. 问题修复
   - 解决上一版本只能使用/opt/dtk 路径问题，支持 dtk 的不同安装路径；
   - 解决在调用 miopen 产生的部分尺寸 batchnormal 计算准确性问题；
   - 修复在网络训练中潜在的内存溢出问题；
3. 性能改进
   - 提升了部分网络在调用 miopen 卷积计算初次加载的性能；
   - git8da4652 版本针对 NLP 领域 fp16 精度进行了优化；
   - git8da4652 版本支持超大规模训练（超 20000 卡测试验证）；
4. 已知问题
   - 稀疏矩阵相关算子中，多线程存在竞争，可能会导致计算失败；
   - 部分 Prof 性能分析接口无法抓取 DCU 相关信息；

- 部分 jit 功能存在 bug；

2. TensorFlow

20221114

git: tf-1.15 -> b34e9207、512656b2/ tf-2.7 -> b748c90e/ tf2.9 -> 251a1066

存储位置：whl_dtk2210

1. 功能改进
   - 支持 dtk22.10 系列 dtk 软件栈；
   - 新增支持对应官方的 tensorflow2.9.0 版本；
   - 支持 tf2.9 的 python3.7/3.8/3.9 版本；
   - 拉齐 tf2.7 的代码，更新至最新的 2.7.0 的 commit 版本
2. 问题修复
   - 解决 tf1.15 调用 xla 的 bitcode module not found at ./hc.amdgcn.bc 的 bug；
   - 解决 tf1.5 的 gfx21687 is not reconnized processor for this target(ignoring processor)的 bug；
   - 解决 tf1.15 的 py3.8 和 py3.9 版本的_logger_find_caller() takes from 0 to 1 positional arguments but 2 were given；
3. 已知问题
   - MobileNet 系列网络，使用 fp16 计算，bs 大于等于 128 时，可能存在 loss nan；

20220809

git: tf-1.15 -> 3a9cf119/ tf-2.7 -> 4015ec59/ tf2.8 -> fe67d51c

存储位置：whl_dtk22.04.2

4. 功能改进
   - 新增支持对应官方的 tensorflow2.8.0 版本；
   - 支持 tf2.8 的 python3.7/3.8/3.9 版本；

- tf2.7、tf2.8 的 Prof 性能分析接口支持抓取 DCU 相关信息；

5. 问题修复
   - 解决 tf1.15 调用 C API 的 could not find hipfftCreate in rocfft DSO 的 bug；
   - 解决 tf1.5 的 rocBLAS does not curremtly supprot the SCAL operation for the complex<> datatype 的 bug；
6. 已知问题
   - MobileNet 系列网络，使用 fp16 计算，bs 大于等于 128 时，可能存在 loss nan；

20220604

git: tf-1.15 -> 2f0f5d2a/ tf-2.7 -> 50328a05

存储位置：whl_dtk22.04.1

7. 功能改进
   - 支持对应官方的 tensorflow1.15/2.7.0 版本；
   - 支持 python3.7/3.8/3.9 版本；
   - 支持 dtk22.04 系列 dtk 软件栈；
   - whl 命名增加对 dtk 大版本支持编号；
   - 设备检测部分显示 DCU；
   - 补充 tensorflow 的 c/c++接口库；
8. 问题修复
   - 解决在调用 miopen 产生的部分尺寸 batchnormal 计算准确性问题；
   - 修复在网络训练中潜在的内存溢出问题；
9. 性能改进
   - 提升 Depthwise 卷积在 Group>1 时的计算性能；
   - 提升 MobileNet 系列网络计算性能；
   - 提升了部分网络在调用 miopen 卷积计算初次加载的性能；
10. 已知问题
    - 部分 Prof 性能分析接口无法抓取 DCU 相关信息；

- 部分 xla/mlir 算子无法支持；
- 加速卡部分不支持 NHWC 数据格式；

3. Paddle

20230614

git：paddle-2.3.2 - > dbe08e9b

存储位置：whl_dtk23.04

1. 功能改进
   - Gemm fp16 计算时的 compute type 默认为 fp16，在不降低精度的同时提高性能，export FLAGS_gemm_use_half_precision_compute_type=0 可以关闭。
   - 解决 hipfft 运行结束时段错误。
   - 对应官方的 paddle-2.4.2 版本。
   - 支持高可复用算子库 PHI 的升级。
2. 问题修复
   - 解决 rccl 版本判断有误导致使用到 nccl 部分较新 api 的 bug。
   - 解决版本号不在系统里显示的问题。
3. 已知问题
   - 部分 GPU 算子功能不支持。
   - 静态图 sequence_reverse 在 rocm 上有正确性问题。

20221115

git：paddle-2.3.2 - > 0195561

存储位置：whl_dtk22.10

4. 功能改进
   - Gemm fp16 计算时的 compute type 改为 fp16，在不降低精度的同时提高性能。
   - 编译选项加上--gpu-max-threads-per-block=1024，避免 kernel 大 block

size 导致计算错误。

- 对应官方的 paddle-2.3.2 版本。
- 支持高可复用算子库 PHI 的升级。

5. 问题修复
   - 解决 rccl 版本判断有误导致使用到 nccl 部分较新 api 的 bug。
   - 解决 whl 包安装偶尔出现 not surpport on this platform。
6. 已知问题
   - 最新版本在系统显示 0.0.0。
   - 部分 GPU 算子功能不支持。

20220806

git：paddle-2.3.0 - > ddf4287

存储位置：whl_dtk22.04.2

7. 功能改进
   - 新增 paddle profiler 功能，支持 op，kernel，memory，占用率等更全面的分析，新增 json 可视化分析。
   - 支持对应官方的 paddle-2.3.0 版本。
   - 支持高可复用算子库 PHI。
8. 问题修复
   - 解决 data_type_transform.cu 与 tensor_util.cu 软连接不能编译问题。
   - 解决编译时第三方包下载不能连接 github 问题，避免库编译问题。
   - 解决 manylinux 导入 openssl 版本不匹配问题，各系统请直接安装默认 openssl，无需额外编译 openssl1.1.1。
9. 已知问题
   - 部分 GPU 算子功能不支持。

20220527

git：f1bf906b

存储位置：whl_dtk22.04.1

10. 功能改进

- 支持对应官方的 paddle-2.2.2 版本；
- 支持 python3.7/3.8/3.9 版本；
- 支持 dtk22.04 系列 dtk 软件栈；
- whl 命名增加对 dtk 大版本支持编号；
- 新增 paddle.nn.Mish 和 paddle.nn.functional.mish，支持逐元素计算 mish 激活函数；

11. 问题修复

- 解决链接 libxxhash.a 时出错问题；
- 解决第三方包下载问题，brpc 与 protobuf 版本冲突问题；
- 解决 paddle 自定义迭代器未重载-运算符引起的 rocprim 类型不匹配问题；

12. 性能改进

- 提升了部分网络在调用 miopen 卷积计算初次加载的性能；

13. 已知问题

- 部分 GPU 算子功能不支持；

4. KALDI

20221115

git：kaldi5.5->82095abd
存储位置：whl_dtk2210

1. 功能改进
   - 支持 dtk22.10 系列版本
   - 放开 hipStreamPerThread feature
2. 已知问题
   - 某些 kernel 在 512size 大小下会出现 VMFault

20220810

git: kaldi5.5 -> cd8752a4

存储位置：whl_dtk22.04.2

3. 功能改进
   - 支持 dtk22.04.2 版本；

20220613

版本号：kaldi5.5-dtk2204.tar.gz

git: cd8752a4

存储位置：whl_dtk22.04.1

4. 功能改进
   - 支持 dtk2204 系列版本；

20220429

版本号：kaldi5.5-21.10.1.tar.gz

git: cd8752a4

存储位置：whl_dtk21.10.1

5. 功能改进
   - 支持 dtk21.10.1 版本 Runtime；
   - 对齐 kaldi 官方 5.5 版本；
6. 已知问题：
   - hipStreamPerThread 特性不支持；
   - 目前只支持 linux 操作系统；

5. RCCL

20221115

版本号：rccl-2.12.12-dtk_rel_22.10.el7.x86_64.rpm

git: 6e3cc8e5

1. 功能改进
   - RCCL 升级至 2.12.12 版本，功能对标 NCCL 2.12.12；
   - 添加 timeline 功能
   - 支持 dtk22.10 软件栈

20220810

版本号：rccl-2.9.9-dtk_rel_22.04.2.el7.x86_64

git：aec273b9

1. 功能改进
   - 支持 dtk22.04.2 软件栈；
2. 问题修复
   - 解决在多进程执行模式下随机卡主的问题（需升级 dtk22.04.2）；
3. 已知问题
   - 基于 clique 内核当前不支持托管内存；

20220527

版本号：rccl-2.9.9-dtk_rel_22.04.el7.x86_64

git：aec273b9

4. 功能改进
   - 改进拓扑检测函数的实现；
   - 支持 dtk22.04 系列软件栈；
   - RCCL 升级到 2.9.9 版本，功能对标 NCCL 2.9.9；
   - 按设备名称对 IB 设备进行排序；

- 增加支持的 dtk 大版本号说明；

5. 问题修复
   - 解决 prof 分析出来的 kernel 函数名都是 sendrecv 问题；
   - 规避 hsa_amd_pointer_info 和 dtk2204.1 版本冲突问题；
   - 修复在同一进程中重新创建多个通信器时潜在的内存泄漏问题；
   - 解决在 device 内存拷贝时，在不进行 device 切换可能存在内存泄露问题；
6. 已知问题
   - 基于 clique 内核当前不支持托管内存；

6. TorchVision

20230617

git: torchvision-0.14.1 -> 9134838

存储位置：whl_dtk23.04

1. 功能改进
   - 支持 dtk23.04 软件栈；
   - 支持 python3.7/3.8/3.9 版本，配合 1.13 版本 torch 使用；
   - 还原 torchvision.__version__返回，增加 torchvision.__dcu_version__查询详细版本；
   - 其他版本需求可到 AIComponent / vision · GitLab (hpccube.com)工程下载，通过源码编译；

git: torchvision-0.10.0 -> 48e6bbb

存储位置：whl_dtk23.04

2. 功能改进
   - 支持 dtk23.04 软件栈；
   - 支持 python3.7/3.8/3.9 版本，配合 1.10 版本 torch 使用；
   - 还原 torchvision.__version__返回，增加 torchvision.__dcu_version__查询详细版本；

- 其他版本需求可到 AIComponent / vision · GitLab (hpccube.com)工程下载，通过源码编译；

20230606

git: torchvision-0.14.1a0 -> d36dd76

存储位置：whl_dtk22.10

3. 功能改进
   - 支持 torchvision0.14.1 python3.8 版本，配合 1.13 版本 torch 使用；
   - 其他版本需求可到 AIComponent / vision · GitLab (hpccube.com)工程下载，通过源码编译；

20221115

git: torchvision-0.10.0a0 -> e04d001

存储位置：whl_dtk22.10

4. 功能改进
   - 支持 dtk22.10 软件栈；
   - 支持 python3.7/3.8/3.9 版本
   - 适配 image 相关算子；

20220809

git: torchvision-0.10.0a0 -> e17f5ea

存储位置：whl_dtk22.04.2

5. 功能改进
   - 版本支持 0.10.0；
   - 支持 python3.7/3.8/3.9 版本；
   - 支持 dtk22.04.2 软件栈；

20220525

git: vision-0.10.0 -> e17f5ea

存储位置：whl_dtk22.04.1

1. 功能改进
   - 版本支持 0.10.0/0.12.0；
   - 支持 python3.7/3.8/3.9 版本；
   - 支持 dtk22.04 系列 dtk 软件栈；
   - whl 命名增加对 dtk 大版本支持编号；
   - nms 和 roi_align 的前向传递支持量化的 dtype 张量；
   - 新增 KITTI 数据集；
   - 增加了 SSD & SSDlite 目标检测模型；

7. MMCV

20230616

git：dtk-23.04_v2.0.0 –>c7118e58

1．功能改进
   - 支持 dtk23.04 系列 dtk 软件栈；
   - 支持对应官方的 mmcv-2.0.0 版本;
   - 添加添加”__dcu_vewrsion__”用于区分 dcu 的版本；

2．已知问题
   - 部分 mmseg 分割类网络在 DCU 上的性能较差

git：dtk-23.04_v1.6.1 –>0d20119a

1．功能改进
   - 支持 dtk23.04 系列 dtk 软件栈；
   - 添加”__dcu_vewrsion__”用于区分 dcu 的版本；

2．已知问题
   - 部分 mmseg 分割类网络在 DCU 上的性能较差

20221110

git：mmcv_full-1.6.1 –>debbc801

1．功能改进

- 支持 dtk22.10 系列 dtk 软件栈；

2．问题解决

- 解决 python3.9 环境 git 版本号，获取路径错误；
- 解决 dcu blocksize 支持 1024 的问题；

3．已知问题

- 部分 mmseg 分割类网络在 DCU 上的性能较差

20220816

git：mmcv_full-1.6.1 –>32fec19

存储位置：whl_dtk22.04.2

1．问题修复

- 修复 correlation 测试用例不过的 bug；

20220816

git：mmcv_full-1.6.1 –>32fec19

存储位置：whl_dtk22.04.2

2．问题修复

- 修复 correlation 测试用例不过的 bug；

20220810

git：mmcv_full-1.6.1 –>59b0fb6

存储位置：whl_dtk22.04.2

1．功能改进

- 支持对应官方的 mmcv-1.6.1 版本；

- whl 包版本号添加对应的 git 信息；
- 支持 dtk22.04.2 系列的 dtk 软件栈；

2．问题修复

- 修改 rocblas 和 torch 调用冲突；
- 修改 mpirun 启动的模式；
- 修复 rocm 的编译转码方式；
- 修复 DCU 不支持__shfl_down_sync 的问题；

3．已知问题

- 部分 mmseg 分割类网络在 DCU 上的性能较差

20220527

版本号：mmcv_full-1.3.16+dtk.22.4-cp37-cp37m-linux_x86_64

git: 59fd5bd5

存储位置：whl_dtk22.04.1

1. 功能改进

- 支持对应官方的 mmcv-1.3.16 版本；
- 支持 python3.7 版本；
- 支持 dtk22.04 系列 dtk 软件栈；
- 添加 4 个 detection3d 相关的 DCU 算子，分别是 voxel op，roiaware pool3d op，iou3d op 和 group points op；
- whl 包版本号添加对应的大版本 dtk 编号；

2. 问题修复

- 修复安装 whl 后无法加载 model_zoo 组件中部分模块的问题；
- 修复 carafe op 在 fp16 精度下计算准确性问题；
- 修复非 ascii 字符导致的 config 解析错误问题；

3. 已知问题

- 部分 mmseg 分割类网络在 DCU 上的性能较差；

8. APEX

20230614

git: apex-0.1->f49ddd4

1. 功能改进
   - 优化 multi_tensor_apply 性能，减小了 chunk_size，增大了 depth_to_max_blocks
   - 添加了 FusedLARS 优化器
   - 添加环境变量 APEX_ROCBLAS_GEMM_ALLOW_HALF 用于控制是否使用 fp16r
   - 添加 dcu 版本信息和 dtk 信息
2. 已知问题
   - multihead_attn_norm_add 计算 fp16 时会出现部分数据精度问题

20221115

git：apex-0.1 - >db7007a

存储位置：whl_dtk2210

1. 功能改进：
   - 支持 python3.7/3.8/3.9 版本；
   - 新增 focal_loss、index_mul_2d、transducer 模块
   - 新增基于 fused_layer_norm 的算子 FusedRMSNorm
   - Fused kernels 支持梯度裁剪（不计算 L2 范数时回退到 PyTorch impl）
   - DistributeFusedAdam 支持 ZeRO-2
2. 问题修复
   - 修改 rocblas_gemm_ex 计算 fp16 时的计算精度
3. 已知问题：
   - multihead_attn_norm_add 计算 fp16 时会出现部分数据精度问题

20220805

git：apex-0.1 - > a38191a

存储位置：whl_dtk22.04.2

1. 功能改进：
   - 增加 distributed_lamb_cuda 算子
   - 增加 distributed_adam_cuda 算子
2. 问题修复：
   - 修复 distributed_fused_lamb 算子中冗余函数的问题

20220527

版本号：apex-0.1_dtk22.04-cp37-cp37m-linux_x86_64.whl

git: e27c7452

存储位置：whl_dtk22.04.1

1. 功能改进：
   - 支持对应官方的 apex-0.1 版本；
   - 支持 python3.7 版本；
   - 支持 dtk22.04 系列 dtk 软件栈；
   - whl 命名增加对 dtk 大版本支持编号；
   - 增加 multihead atten 模块的支持；
   - 增加分布式算子 FusedLAMB；
2. 问题修复：
   - 修复了 self_multihead_attn_norm_add 的计算准确性问题；
3. 已知问题
   - 不支持 nhwc batch normal 算子；
   - FusedAdam 算子 float16 计算结果偶尔会有 NAN 出现；
   - cached_cast 会引起 dtk runtime 错误；

9. NNI

20230617

git：nni2.9 –> git83609d8
存储位置：whl_dtk23.04

1. 功能改进
   - nni 版本升级至 2.9；
   - 支持 dtk23.04 系列 dtk 软件栈；
   - 添加 dcu 版本信息查询；
2. 已知问题
   - 存在_locl_file 相关的 AttributeError，需要安装低版本的 filelock，如 pip install filelock==3.10；
   - 存在 dict 相关的 AttributeError，需要安装低版本的 typeguard，如 pip install "typeguard<3"；

20221114

版本号：nni-2.6+gitc6ca2d03.dtk2210-py3-none-manylinux1_x86_64.whl

git: c6ca2d03

存储位置：whl_dtk22.10

1. 功能改进：
   - 支持 dtk22.10 系列 dtk 软件栈

20220813

git: nni-2.6 –> git10113773

存储位置：whl_dtk21.10.2

2. 功能改进：
   - 增加 dtk 版本标识；
   - NNI 支持 dtk22.04.2 版本软件栈；

20220519

版本号：nni-2.6-py3-none-manylinux1_x86_64

git: 10113773

存储位置：whl_dtk21.10.1

1. 功能改进：
   - NNI 版本升级至 2.6；
   - 支持 dtk21.10.1 软件栈；
   - NNI 支持 Tensorflow 和 Pytorch；
   - NNI 增加 DCU 相关信息抓取，可实现相关 DCU 信息的可视化；

10. DeepSpeed

20230621

git：deepspeed-v0.9.2->25d5540b4434a97cab39b2dd06575fd605c6676f

1. 功能改进
   - 支持对应官方的 deepspeed-0.9.2 版本；
   - 支持 python3.7/3.8/3.9 版本，支持 torch1.13 /torch1.10 版本，支持 abi0/abi1。
   - 支持两种版本查询方式，deepspeed.__version__返回主版本号，deepspeed.__dcu_version__查询基于 dtk 适配的详细版本；
   - 其他版本需求或者问题可到 http://developer.hpccube.com/codes/aicomponent/deepspeed.git 工程查看与反馈。
2. 已知问题
   - 暂时不支持 quantizer_op、random_ltd、transformer_inference_op 等 op 模块，待 DTK 增加 cg 等函数接口后再做完善修正；
   - triton 暂未适配 ROCm，相关依赖项均未支持；
   - accelerators 组件相关的 op 计算会存在一部分计算误差，该问题为

gemm fp16r 计算参数类型导致，综合考量性能，暂时默认该设置。屏蔽该误差需要关闭 torch 端对应接口该类型的调用。

20230530

git：deepspeed-v0.8.2-> ac5fbab439765b122ce43b9185c9053bd2c15372

1．功能改进

- 支持对应官方的 deepspeed-0.8.2、deepspeed-0.9.2 版本；
- 支持 python3.7、torch1.10 版本，其他版本需求可通过开发者社区自主编，两个版本编译方式基本一致，具体可参考以下链接说明：https://developer.hpccube.com/codes/aicomponent/deepspeed/-/blob/ds-v0.8.2-rocm/README_HIP.md，；

2. 已知问题

- 暂时不支持 quantizer_op、random_ltd、transformer_inference_op 等 op 模块，待 DTK 增加 cg 等函数接口后再做完善修正；
- triton 暂未适配 ROCm，相关依赖项均未支持；
- accelerators 组件相关的 op 计算会存在一部分计算误差，该问题为 gemm fp16r 计算参数类型导致，综合考量性能，暂时默认该设置。屏蔽该误差需要关闭 torch 端对应接口该类型的调用。

20221117

git：deepspeed-0.6.3 ->1b2721a

1. 功能改进

- 支持 dtk22.10 系列 dtk 软件栈；
- 适配多 python 版本，py3.7/py3.8/py3.9

20220809

git：deepspeed-0.6.3 ->1b2721a

存储位置：whl_dtk22.04.2

1. 功能改进
   - 增加 dtk 版本号标识
   - 支持 dtk22.04.2 版本

20220606

版本号： deepspeed-0.6.3+d335bff-cp37-cp37m-linux_x86_64.whl

存储位置：whl_dtk22.04.1

3. 功能改进
   - 支持对应官方的 deepspeed-0.6.3 版本，并且支持 transformer inference；
   - 支持 python3.7 版本；
   - 支持 dtk22.04 系列 dtk 软件栈；
   - 添加 3 个 op module 的 DCU 支持，分别是 cpu_adagrad，quantizer 和 transformer_inference；
   - whl 包版本号添加 git 号识别；
2. 问题修复
   - 修复安装 whl 后无法加载 sparse attention 组件中部分模块的问题；
3. 已知问题
   - triton 包暂未适配 ROCm；
   - sparse_attention 中的 softmax、matmul 计算暂不支持，缺省 triton 相关模块支持；
   - async_io_op 暂不支持。

11. OnnxRuntime

20220620

git：onnxruntime-lite-1.14.0 ->81e68c5

存储位置：whl_dtk2304

1. 功能改进
   - 支持 dtk2304 版本；
   - 更新 ort Conv 系列、BN、pool 系列、Reduce 系列算子实现方式，解决推理应用首帧瓶颈问题；

20220606

git：onnxruntime-1.14.0 ->bf238cdd

存储位置：whl_dtk2210.1

2. 功能改进
   - 支持 dtk22.10.1 版本；
   - 增加了 ROCMExecutionProvider 推理支持，支持实现动态推理；
   - 关闭卷积相关融合算子调用。

20221115

git：onnxruntime-1.8.0 ->27635174

存储位置：whl_dtk2210

3. 功能改进
   - 支持 dtk22.10 版本

20220809

版本号：onnxruntime-1.8.0

存储位置：whl_dtk22.04.2

1. 功能改进
   - 支持 dtk22.04.2 版本

12. MIGraphX

请注意：MIGraphX 后续版本随 dtk 发布，不再单独维护安装包；

20230606

git: MIGraphX3.0.0 –> 153649ca

1. 功能改进
   - 优化了 yolov5 系列以及 gpt 系列动态推理性能，平均性能提高 10%；
2. 问题修复
   - 修复了动态 shape 兼容性问题，能够兼容目前主流 AI 模型的动态推理；
   - 修复 NMS 算子运行异常问题；
   - 修复 inception_v3 的动态推理精度问题。

20230412

git: MIGraphX2.5.2 –> 8533e115

2. 功能改进
   - 支持了 GPT2 模型的动态推理；
2. 问题修复
   - 修复了 Transformer 的结构动态 shape 问题；
   - 修复 selu 算子的精度问题；
   - 修复了 vit(vision transformer)模型 FP16 模式下的精度问题。

20230215

git: MIGraphX2.5.0 –> f2f7ae8

存储位置：whl_dtk2210

3. 功能改进
   - 添加加载 mxr 模型后可以修改 device id 的功能；
2. 问题修复
   - 修复 LSTM 算子不支持动态 shape 的问题，目前可以支持 H,W 维度动态，但是 batchsize 只能设置为 1；
   - 修复卷积和反卷积在某些 shape 中出现的 workspace 的问题；

- 修改全局池化算子在动态 shape 中存在的正确性问题；
- 修复 CRNN 不支持动态 shape 的问题；
- 修复加载 mxr 模型时动态 shape 运行时超出最大 shape 的问题；
- 修复 FCN 模型不能支持动态 shape 的问题；
- 修复 vit(vision Transformer)在静态 shape 和动态 shape 的精度问题；
- 修复 yolov5 系列模型动态 shape 模式下的精度问题；
- 修复 LSTM 在 FP16 中的程序异常问题；
- 修复 efficientnet 在静态 shape 下精度的问题；

20221115

git: MIGraphX2.4.1 –> 17ff0721

存储位置：whl_dtk2210

1. 功能改进
   - 支持 dtk22.10 系列 dtk 软件栈；
2. 问题修复
   - 修复了 CRNN 中的 cip 和 reshape 算子动态 shape 问题；
   - 修复卷积算子动态 shape 问题；
   - 修改官方 find_gemm_add 的实现减少计算，提高性能；
   - 修复 rocblas_gemm 的动态 shape 问题；

20220901

git: MIGraphX2.4.0 –> c547cceb

版本号： migraphx-2.4.0-.el7.x86_64.rpm

存储位置：whl_dtk22.04.2

1. 功能改进
   - 支持 dtk22.04.2 版本；
   - 添加了保存加载编译后模型的功能，缩短模型启动时间；

2. 问题修复
   - 修复 migraphx::save 无法保存 device_id 的问题；
   - 修复 RetinaFace_ResNet50 的动态 shape 问题；
   - 修复 jit 不支持动态 shape 的问题，通过设置环境变量 MIGRAPHX_DYNAMIC_SHAPE 来屏蔽动态 shape 中不支持的 jit 优化；
   - 修复网易和京东模型问题；
3. 性能改进
   - 修改 profiling 工具中算子耗时占比的显示格式，去掉 ceil 操作，让结果更加准确；
4. 说明
   - 使用 MIGraphX 进行程序开发需要再安装对应的 devel 包；

20220606

版本号： migraphx-2.3.0-b55e78e.el7.x86_64.rpm

存储位置：whl_dtk22.04.1

1. 功能改进
   - 支持 dtk22.04 系列 dtk 软件栈；
   - 添加了直接使用 gpu 数据的功能，同时支持 C++和 Python 接口；
   - 2.3.0 版本不再区分 onnxruntime 版本；
   - 新增算子：HardSigmoid, Softplus, Softsign, GreaterOrEqual, HardSwish, Mean, IsNaN, ScatterND, EyeLike, Size, Celu, LpNormalization, ScatterElements, LpPool, GlobalLpPool, ReverseSequence, GatherND；
2. 问题修复
   - 修复性能分析工具中 batchsize 显示不准确的问题；
   - 修复了动态 shape 中 sync_stream 算子的 do_reshape()返回值的问题；
   - 修复 compile_ops pass 在 dtk 中 resnet50 编译报错的问题；
3. 性能改进
   - 优化 hip_copy_from_gpu 算子，将内存分配从运行期移动到了编译期；

- 优化了 parse_gemm，删除了不必要的常量传播，便于动态 shape 的实现；
- 为 offload_copy 为 false 加入流同步，避免额外的同步操作；

4. 性能改进
   - 使用 MIGraphX 进行程序开发需要再安装对应的 devel 包；

13. JAX

20221114

git：dtk-22.10_jax0.2.21 –>6c419e3

1. 功能改进
   - 支持 dtk22.10 系列 dtk 软件栈；
2. 已知问题
   - ROCSolver 中未实现对称特征分解

14. Horovod

20230529

git：0.26.1->5675b001
存储位置：whl_dtk2210

1. 功能改进
   - cp37 和 cp39 版本更新至 0.26.1
   - 解决部分卡死现象

20221114

git：0.21.3->6b25a989/ 0.22.1->6b25a989

存储位置：whl_dtk2210

1. 功能改进
   - 支持 dtk22.10 系列 dtk 软件栈；

- 支持 tensorflow1.15、tensorflow2.7 和 torch1.10 版本；

15. OneFlow

20230606

git：oneflow-0.9 - > 5be579db

存储位置：whl_dtk22.10.1/ whl_dtk23.04

1. 功能改进
   - 支持 AMP（自动混合精度）
   - 支持模型并行，流水并行，3D 并行。
   - Layer_norm 算子性能提升
   - 支持 bn_add_relu 融合
2. 已知问题
   - 部分 cutlass 相关的融合算子功能不支持。
   - dtk23.04 上 matmul_add_bias_fused 算子有正确性问题，关闭 rocblas 原子操作解决。

20221115

git：oneflow-0.8 - > 3f56062

存储位置：whl_dtk22.10

1. 功能改进
   - 支持 dtk21.10.1 系列 dtk 软件栈
   - 支持 dtk22.04.2 系列 dtk 软件栈。
   - 支持 dtk22.10 系列 dtk 软件栈。
   - 对应官方的 oneflow-0.8 版本。
   - 支持 Libai-0.2，OneEmbedding。
2. 已知问题
   - 部分 GPU 算子功能不支持。

- AMP 尚不支持

16. Faiss

20221117

版本号：faiss-1.7.2_dtk22.10_gitb7348e7df780

1. 功能改进
   - 新增支持 dtk22.10；
2. 已知问题
   - 不支持 IVFFlat.UnifiedMemory 及 IVFPQ.UnifiedMemory

17. Colossalai

20230616

git：dtk-23.04_colossalai0.1.13 –>43ff1d4

1. 功能改进
   - 新增支持 dtk23.04；
2. 已知问题
   - 不支持 zero_level_3

20230531

git：dtk-22.10.1_colossalai0.1.13 –>632b5a1

1. 功能改进
   - 新增支持 dtk22.10.1；
   - 支持 python3.8、torch1.10 版本;
2. 已知问题
   - 不支持 zero_level_3

20221117

git：colossalai-0.0.2_git364b418428e07_dtk2210

1. 功能改进
   - 新增支持 dtk22.10；
2. 已知问题
   - 不支持 zero_level_3

18. FastFold

20230616

git：dtk-23.04_fastfold0.2.1 –>9f6252f

1. 功能改进
   - 支持 dtk23.04 系列 dtk 软件栈；

20230307

git：dtk-22.10_fastfold0.2.1 –>dd424ec

1. 功能改进
   - 支持保存.pkl 模型；
   - 支持最小力场化操作；
   - 支持保存最小力场化后的 releax 模型；
2. 问题修复
   - 修复不包含模板的模型以及包含 ptm 模型的加载问题；
   - 修复 openmm 加载后端问题；
   - 修复模板匹配时的无效错误处理；
3. 性能改进
   - 可推理长度最高为 3071 的蛋白质单体序列；
   - 可推理长度最高为 2030 的蛋白质多体序列；

20221118

git：dtk-22.10_fastfold0.2.0 –>52919c6

1. 功能改进
   - 支持 dtk22.10 系列 dtk 软件栈；

19. OpenFold

20230616

git：dtk-23.04_openfold1.0.1 –>ae4af8f

1. 功能改进
   - 支持 dtk23.04 系列 dtk 软件栈；

20230112

git：dtk-22.10_openfold1.0.1 –>ed252f7

1. 功能改进
   - 支持最小力场化操作；
2. 问题修复
   - 修复 openmm 加载 HIP 后端；

20221118

git：dtk-22.10_openfold1.0.0 –>f32f248

1. 功能改进
   - 支持 dtk22.10 系列 dtk 软件栈；

20. MMdeploy

20230616

git：dtk22101_v1.0.0–>0618e08a

1. 功能改进
   - 支持 dtk23.04 系列 dtk 软件栈；
   - 支持 python3.7/3.8/3.9 版本;
   - 支持对应官方的 1.0.0 版本;
   - 添加”__dcu_vewrsion__”用于区分 dcu 的版本；
2. Whl 使用
   - Abi0 环境 python3.8 单独的 cp38-cp38 的 whl，python3.7 和 python3.9 使用 py3-none 的 whl；Abi1 的环境使用共同的 py3-none 的 whl，不分区 python 本本
3. 已知问题
   - 推理后端只支持 OnnxRuntime1.14.0

20230530

git：dtk22101_v1.0.0–>4be2b872

1. 功能改进
   - 支持 dtk21.10.1 系列 dtk 软件栈；
   - 支持 python3.8 版本;
2. 已知问题
   - 推理后端只支持 OnnxRuntime1.14.0

21. FastMoE

20230606

git： dtk-22.10_fastmoe0.3.0–>acf8bec

2. 功能改进

- 支持 dtk22.10 系列 dtk 软件栈；

22. Dgl

20230606

git： git822caeb6.dtk2210

1. 功能改进

- 支持 dtk22.10 系列 dtk 软件栈；
- 支持 python3.9 版本;

2. 已知问题

- Fp16 类型暂时不支持；

23. multi_scale_deformable_attention

20230606

git： 8f6c15bc

1. 功能改进

- 支持 dtk22.10 系列 dtk 软件栈；
- 支持 python3.7、torch1.10 版本；

24. TorchAudio

20230614

git: 0.13.1-dtk23.04->8bfbf473

1. 功能改进
   - 适配 PyTorch1.13.1 使用，对齐官方 0.13.1 版本
   - 添加 dcu 版本号及 dtk 信息

git: 0.10.0-dtk23.04->a9847c38

2. 功能改进
   - 适配 PyTorch1.10.0 使用，对齐官方 0.10.0 版本
     添加 dcu 版本号及 dtk 信息

25. Dlib

20230615

git: 4abf0a6

1. 功能改进
   - 支持 dtk23.04 系列 dtk 软件栈
   - 可用 torch.__dcu_version__查询详细版本
   - 单元测试中测试 layer_norm_中有报错，可能有潜在 bug。但已排除算子计算错误，目前为止尚未发现使用时的问题

26. Warpctc

20230615

git: 010367b

1.功能改进
   - 支持 dtk23.04 系列 dtk 软件栈

- 可用 torch.__dcu_version__查询详细版本
- 添加 dcu 版本号及 dtk 信息

27. Deepbench

20230615

git: d327324

1.功能改进

- 支持 dtk23.04 系列 dtk 软件栈，可以 whl 包方式安装
- 为了降低安装 deepbench 的难度，去掉了 mpirun 方式测试 RCCL。在测试 RCCL 时不再需要额外安装 mpich 库

28. lightop

20230616

git:90afe939 (torch1.10)

git: 1e1af5a1 (torch1.13)

功能改进

- 支持 dtk23.04 系列 dtk 软件栈，可以 whl 包方式安装
- 新增对 torch1.13 的支持
- 增加 droppath 和 ln_droppath 算子
- 增加 c 方式的 autograd，加快 cpu 端算子加载时间

29. BladeDISC

20230616

git:5cedf37 (tensorflow 2.7)

git:12c84b8 (torch 1.10)

功能改进

- 支持 dtk22.04.2 系列 dtk 软件栈，可以 whl 包方式安装

其他说明

20230113

1. 更新 DCU 镜像使用手册，增加 dtk22.10 系列镜像，以光源为镜像主要下载方式，并增加 Notebook 镜像制作方法及上传使用方式。

20221123

1. 基于 dtk22.10 的 tensorflow 和 pytorch 在 ubuntu22.04 系统上会出现 segmentation fault，manylinux 的 whl 暂不支持 ubuntu22.04。

20220813

1. 增加 whl 的 manylinux 模式支持，操作系统包括 centos7.6、centos7.9、ubuntu18.04、ubuntu22.04、nfs3.2、v2101.ky10、UOS 20 1020e

20220606

1. 删除 dockerfiles 模块，以百度云盘的 docker 镜像方式提供，在技术文档目录中增加《DCU 镜像使用手册-v1.6.0.pdf》文档说明，其中包含镜像的下载地址；

保密声明

在此声明，该文档展示的全部技术信息及其相关内容，版权皆属于开发者社区（https://developer.hpccube.com/） 所有。未经允许，严禁截屏、大规模传播及转发。另外，对使用该技术文档而导致任何侵犯第三方专利或其他权利的行为，开发者社区不承担任何责任。

感谢您的理解与支持。