Milestone 1 Report

Team: hybrid

Members:

Enyi Jiang(enyij2) rai_id: 5c78c48284318364bab96eaf

Xi Chen(xichen30) rai_id: 5c78c40d84318363f903701c

Yifan Chen(yifanc3) rai_id: 5c78c40e84318363f903701d

School Affiliation: On Campus

- ❏ Include a list of all kernels that collectively consume more than 90% of the program time.

  CUDA memcpy HtoD

  cudnn::detail::implicit_convolve_sgemm

  volta_cgemm_64x32_tn

  op_generic_tensor_kernel

  fft2d_c2r_32x32

  volta_sgemm_128x128_tn

  void cudnn::detail::pooling_fw_4d_kernel

  fft2d_r2c_32x32

- ❏ Include a list of all CUDA API calls that collectively consume more than 90% of the program time.

  cudaStreamCreateWithFlag

  cudaMemGetInfo

  cudaFree

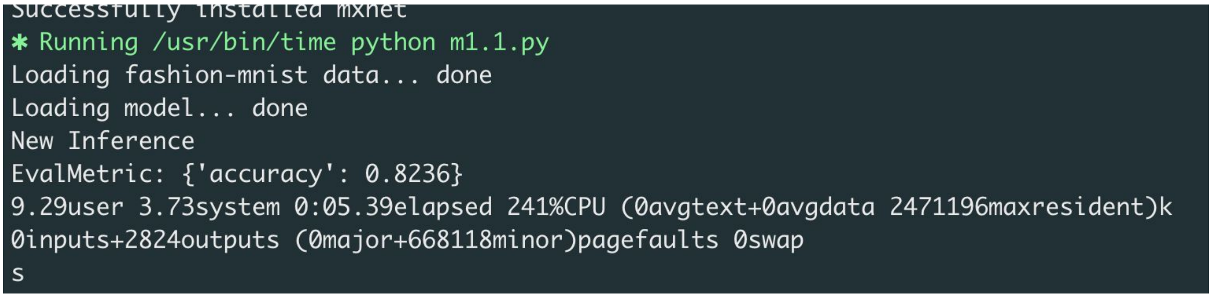
- ❏ Include an explanation of the difference between kernels and API calls

    - ❑ Kernels are C functions which are flagged to be run on a GPU (or a device). It is a more low-level program that instructs the performance of each thread. Kernels have no APIs as they are not libraries.
    - ❑ API calls are C function calls that executed by the host (CPU) to back up kernel codes, including transferring data, managing memory and so on. API calls certainly use some C libraries.
- ❑ Show output of rai running MXNet on the CPU
    - ❑ EvalMetric: {'accuracy': 0.8236}

      9.29 user 3.73 system 0:05.39 elapsed 241%CPU (0avgtext+0avgdata 2471196maxresident)k

      0inputs+2824outputs (0major+668118minor)pagefaults 0swap

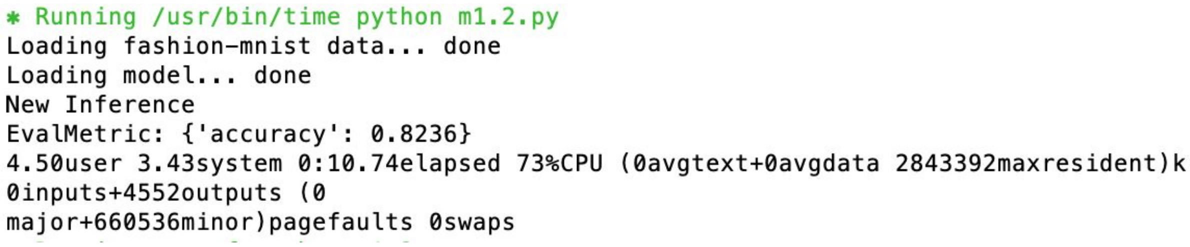
Screenshot of the Output:



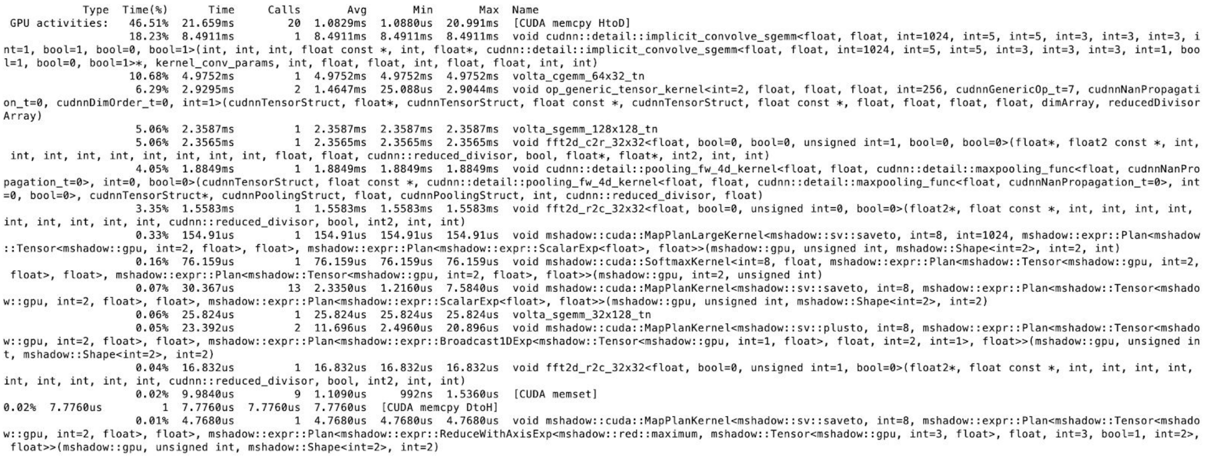
```
Successfully installed mxnet
✱ Running /usr/bin/time python m1.1.py
Loading fashion-mnist data... done
Loading model... done
New Inference
EvalMetric: {'accuracy': 0.8236}
9.29user 3.73system 0:05.39elapsed 241%CPU (0avgtext+0avgdata 2471196maxresident)k
0inputs+2824outputs (0major+668118minor)pagefaults 0swap
s
```

- ❑ List program run time
    - ❑ 9.29 - user 3.73 - system 0:05.39 elapsed 241%CPU
- ❑ Show output of rai running MXNet on the GPU
    - ❑ EvalMetric: {'accuracy': 0.8236}

Screenshot of the output:



```
✱ Running /usr/bin/time python m1.2.py
Loading fashion-mnist data... done
Loading model... done
New Inference
EvalMetric: {'accuracy': 0.8236}
4.50user 3.43system 0:10.74elapsed 73%CPU (0avgtext+0avgdata 2843392maxresident)k
0inputs+4552outputs (0
major+660536minor)pagefaults 0swaps
```



```
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   46.51%  21.659ms        20  1.0829ms  1.0880us  20.991ms  [CUDA memcpy HtoD]
                   18.23%  8.4911ms         1  8.4911ms  8.4911ms  8.4911ms  void cudnn::detail::implicit_convolve_sgemm<float, float, int=1024, int=5, int=5, int=3, int=3, int=3, i
nt=1, bool=1, bool=0, bool=1>(int, int, int, float const *, int, float*, cudnn::detail::implicit_convolve_sgemm<float, float, int=1024, int=5, int=5, int=3, int=3, int=3, int=1, boo
l=1, bool=0, bool=1>*, kernel_conv_params, int, float, float, int, float, float, int, int)
                   10.68%  4.9752ms         1  4.9752ms  4.9752ms  4.9752ms  volta_cgemm_64x32_tn
                    6.29%  2.9295ms         2  1.4647ms  25.088us  2.9044ms  void op_generic_tensor_kernel<int=2, float, float, float, int=256, cudnnGenericOp_t=7, cudnnNanPropagati
on_t=0, cudnnDimOrder_t=0, int=1>(cudnnTensorStruct, float*, cudnnTensorStruct, float const *, cudnnTensorStruct, float const *, float, float, float, float, dimArray, reducedDivisor
Array)
                    5.06%  2.3587ms         1  2.3587ms  2.3587ms  2.3587ms  volta_sgemm_128x128_tn
                    5.06%  2.3565ms         1  2.3565ms  2.3565ms  2.3565ms  void fft2d_c2r_32x32<float, bool=0, bool=0, unsigned int=1, bool=0, bool=0>(float*, float2 const *, int,
 int, int, int, int, int, int, int, int, float, float, cudnn::reduced_divisor, bool, float*, float*, int2, int, int)
                    4.05%  1.8849ms         1  1.8849ms  1.8849ms  1.8849ms  void cudnn::detail::pooling_fw_4d_kernel<float, float, cudnn::detail::maxpooling_func<float, cudnnNanPro
pagation_t=0>, int=0, bool=0>(cudnnTensorStruct, float const *, cudnn::detail::pooling_fw_4d_kernel<float, float, cudnn::detail::maxpooling_func<float, cudnnNanPropagation_t=0>, int
=0, bool=0>, cudnnTensorStruct*, cudnnPoolingStruct, float, cudnnPoolingStruct, int, cudnn::reduced_divisor, float)
                    3.35%  1.5583ms         1  1.5583ms  1.5583ms  1.5583ms  void fft2d_r2c_32x32<float, bool=0, unsigned int=0, bool=0>(float2*, float const *, int, int, int, int,
int, int, int, int, int, cudnn::reduced_divisor, bool, int2, int, int)
                    0.33%  154.91us         1  154.91us  154.91us  154.91us  void mshadow::cuda::MapPlanLargeKernel<mshadow::sv::saveto, int=8, int=1024, mshadow::expr::Plan<mshadow
::Tensor<mshadow::gpu, int=2, float>, float>, mshadow::expr::Plan<mshadow::expr::ScalarExp<float>, float>>(mshadow::gpu, unsigned int, mshadow::Shape<int=2>, int=2, int)
                    0.16%  76.159us         1  76.159us  76.159us  76.159us  void mshadow::cuda::SoftmaxKernel<int=8, float, mshadow::expr::Plan<mshadow::Tensor<mshadow::gpu, int=2,
 float>, float>, mshadow::expr::Plan<mshadow::Tensor<mshadow::gpu, int=2, float>, float>>(mshadow::gpu, int=2, unsigned int)
                    0.07%  30.367us        13  2.3350us  1.2160us  7.5840us  void mshadow::cuda::MapPlanKernel<mshadow::sv::saveto, int=8, mshadow::expr::Plan<mshadow::Tensor<mshado
w::gpu, int=2, float>, float>, mshadow::expr::Plan<mshadow::expr::ScalarExp<float>, float>>(mshadow::gpu, unsigned int, mshadow::Shape<int=2>, int=2)
                    0.06%  25.824us         1  25.824us  25.824us  25.824us  volta_sgemm_32x128_tn
                    0.05%  23.392us         2  11.696us  2.4960us  20.896us  void mshadow::cuda::MapPlanKernel<mshadow::sv::plusto, int=8, mshadow::expr::Plan<mshadow::Tensor<mshado
w::gpu, int=2, float>, float>, mshadow::expr::Plan<mshadow::expr::Broadcast1DExp<mshadow::Tensor<mshadow::gpu, int=1, float>, float, int=2, int=1>, float>>(mshadow::gpu, unsigned in
t, mshadow::Shape<int=2>, int=2)
                    0.04%  16.832us         1  16.832us  16.832us  16.832us  void fft2d_r2c_32x32<float, bool=0, unsigned int=1, bool=0>(float2*, float const *, int, int, int, int,
int, int, int, int, int, cudnn::reduced_divisor, bool, int2, int, int)
                    0.02%  9.9840us         9  1.1090us     992ns  1.5360us  [CUDA memset]
0.02%  7.7760us         1  7.7760us  7.7760us  7.7760us  [CUDA memcpy DtoH]
                    0.01%  4.7680us         1  4.7680us  4.7680us  4.7680us  void mshadow::cuda::MapPlanKernel<mshadow::sv::saveto, int=8, mshadow::expr::Plan<mshadow::Tensor<mshado
w::gpu, int=2, float>, float>, mshadow::expr::Plan<mshadow::expr::ReduceWithAxisExp<mshadow::red::maximum, mshadow::Tensor<mshadow::gpu, int=3, float>, float, int=3, bool=1, int=2>,
 float>>(mshadow::gpu, unsigned int, mshadow::Shape<int=2>, int=2)
```

```
      API calls:   47.29%  3.80438s        22  172.93ms  14.300us  1.97833s  cudaStreamCreateWithFlags
                   29.61%  2.38157s        24  99.232ms  78.559us  2.37638s  cudaMemGetInfo
                   20.05%  1.61322s        19  84.906ms  1.0110us  434.72ms  cudaFree
                    1.06%  85.180ms       216  394.35us     905ns  48.894ms  cudaEventCreateWithFlags
                    0.78%  62.434ms       912  68.458us     318ns  25.580ms  cudaFuncSetAttribute
                    0.55%  43.933ms         9  4.8814ms  47.899us  21.108ms  cudaMemcpy2DAsync
                    0.28%  22.323ms        29  769.75us  3.3850us  10.576ms  cudaStreamSynchronize
                    0.18%  14.253ms        68  209.60us  7.3620us  2.1641ms  cudaMalloc
                    0.06%  4.6346ms        12  386.21us  8.2800us  4.0283ms  cudaMemcpy
                    0.06%  4.6207ms         4  1.1552ms  409.61us  1.7090ms  cudaGetDeviceProperties
                    0.03%  2.3209ms       375  6.1890us     293ns  325.92us  cuDeviceGetAttribute
                    0.01%  987.73us         8  123.47us  14.321us  772.82us  cudaStreamCreateWithPriority
                    0.01%  812.76us         2  406.38us  51.531us  761.23us  cudaHostAlloc
                    0.01%  724.93us         9  80.548us  11.550us  488.79us  cudaMemsetAsync
                    0.01%  654.04us         4  163.51us  96.749us  282.29us  cuDeviceTotalMem
                    0.01%  633.87us        30  21.128us  8.3030us  96.220us  cudaLaunchKernel
                    0.01%  556.26us         4  139.07us  87.316us  214.50us  cudaStreamCreate
                    0.00%  322.49us       210  1.5350us     545ns  5.9370us  cudaDeviceGetAttribute
                    0.00%  272.82us         4  68.205us  48.044us  107.61us  cuDeviceGetName
                    0.00%  191.07us        32  5.9700us  1.0030us  44.003us  cudaSetDevice
                    0.00%  119.00us       564     210ns      82ns     610ns  cudaGetLastError
                    0.00%  66.913us         6  11.152us  1.9910us  41.214us  cudaEventCreate
                    0.00%  57.658us         6  9.6090us  1.3020us  39.701us  cudaEventRecord
                    0.00%  50.209us        18  2.7890us     720ns  4.8710us  cudaGetDevice
                    0.00%  27.096us         2  13.548us  4.8480us  22.248us  cudaHostGetDevicePointer
                    0.00%  14.600us         1  14.600us  14.600us  14.600us  cudaBindTexture
                    0.00%  6.6720us         2  3.3360us  2.2730us  4.3990us  cudaDeviceGetStreamPriorityRange
                    0.00%  6.5120us         3  2.1700us  1.4580us  2.8380us  cudaStreamWaitEvent
                    0.00%  6.2300us         6  1.0380us     569ns  1.8760us  cuDeviceGetCount
                    0.00%  5.0380us        18     279ns     114ns     685ns  cudaPeekAtLastError
                    0.00%  4.4180us         1  4.4180us  4.4180us  4.4180us  cuDeviceGetPCIBusId
                    0.00%  4.3120us         5     862ns     401ns  1.4530us  cuDeviceGet
                    0.00%  4.1910us         3  1.3970us     808ns  2.4080us  cuInit
                    0.00%  3.5660us         1  3.5660us  3.5660us  3
.5660us  cudaUnbindTexture
                    0.00%  3.0820us         4     770ns     356ns  1.2920us  cuDeviceGetUuid
                    0.00%  2.9870us         1  2.9870us  2.9870us  2.9870us  cudaEventQuery
                    0.00%  2.8870us         4     721ns     439ns  1.4220us  cudaGetDeviceCount
                    0.00%  2.0230us         3     674ns     334ns  1.1830us  cuDriverGetVersion
```

- ❏ List program run time
  - ❏ 4.50 - user 3.43 - system 0:10.74 elapsed 73%CPU