## Question 1: Profiling Deep Learning Models for Computational Enhancement

Profiling is done in PyTorch using a simple API that analyzes the model's performance and measures the time and memory consumption of the model's operators.

The approach we followed during the task assigned:

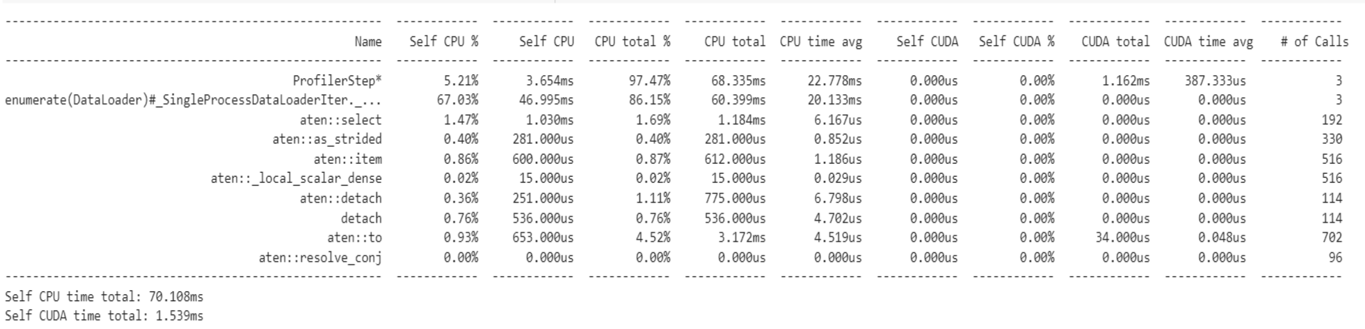
1. The first part relates to profiling, which measures time and memory usage.
2. The second part is related to profiling with automatic mixed precision.

For the profiling, we have used Tensorboard for visualization purposes, which helps visualize the graphs, model architecture visualization, training metrics monitoring, histograms, distribution, embeddings, etc.

### MLP

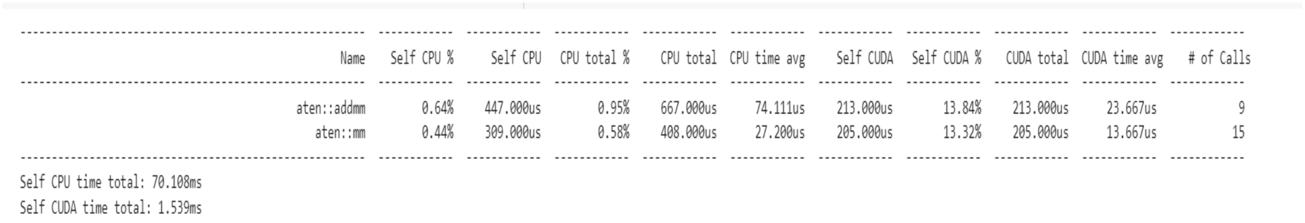
### *Profiling*

Fig 1 shows the stats of CPU memory usage for the model operations using profiling. The data loader iterator and profiler step take the most CPU time.



```
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                                   Name    Self CPU %      Self CPU   CPU total %     CPU total  CPU time avg     Self CUDA   Self CUDA %    CUDA total  CUDA time avg    # of Calls
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                          ProfilerStep*         5.21%       3.654ms        97.47%      68.335ms      22.778ms       0.000us         0.00%       1.162ms     387.333us             3
enumerate(DataLoader)#_SingleProcessDataLoaderIter._...        67.03%      46.995ms        86.15%      60.399ms      20.133ms       0.000us         0.00%       0.000us       0.000us             3
                                           aten::select         1.47%       1.030ms         1.69%       1.184ms       6.167us       0.000us         0.00%       0.000us       0.000us           192
                                       aten::as_strided         0.40%     281.000us         0.40%     281.000us       0.852us       0.000us         0.00%       0.000us       0.000us           330
                                             aten::item         0.86%     600.000us         0.87%     612.000us       1.186us       0.000us         0.00%       0.000us       0.000us           516
                              aten::_local_scalar_dense         0.02%      15.000us         0.02%      15.000us       0.029us       0.000us         0.00%       0.000us       0.000us           516
                                           aten::detach         0.36%     251.000us         1.11%     775.000us       6.798us       0.000us         0.00%       0.000us       0.000us           114
                                                 detach         0.76%     536.000us         0.76%     536.000us       4.702us       0.000us         0.00%       0.000us       0.000us           114
                                               aten::to         0.93%     653.000us         4.52%       3.172ms       4.519us       0.000us         0.00%      34.000us       0.048us           702
                                     aten::resolve_conj         0.00%       0.000us         0.00%       0.000us       0.000us       0.000us         0.00%       0.000us       0.000us            96
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
Self CPU time total: 70.108ms
Self CUDA time total: 1.539ms
```

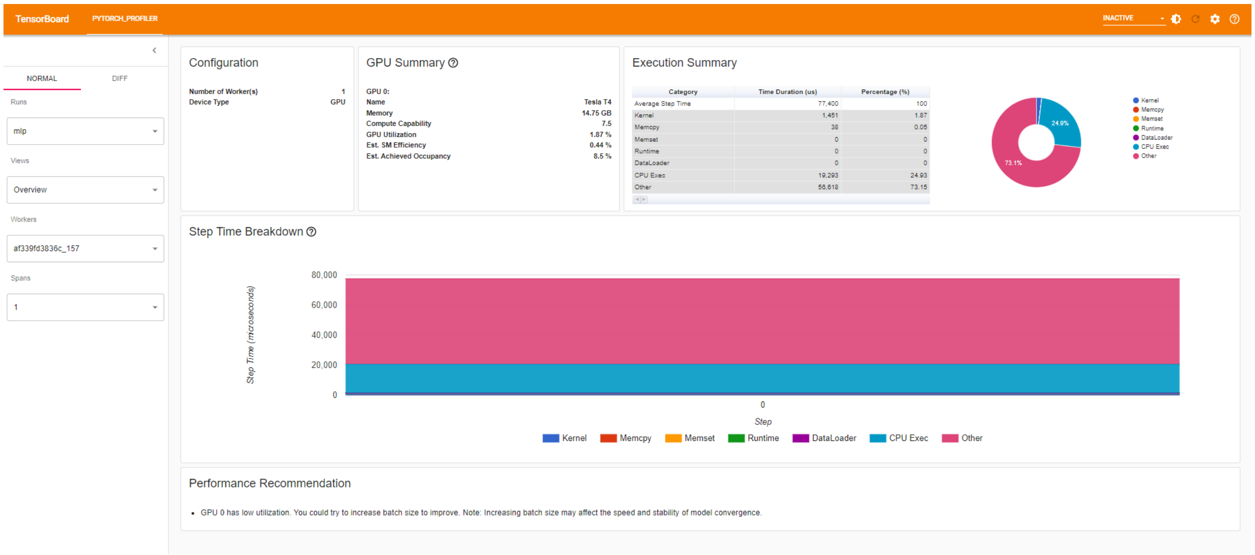
Fig 1: CPU memory usage and time using MLP



```
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                                   Name    Self CPU %      Self CPU   CPU total %     CPU total  CPU time avg     Self CUDA   Self CUDA %    CUDA total  CUDA time avg    # of Calls
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                           aten::addmm         0.64%     447.000us         0.95%     667.000us      74.111us     213.000us        13.84%     213.000us      23.667us             9
                                              aten::mm         0.44%     309.000us         0.58%     408.000us      27.200us     205.000us        13.32%     205.000us      13.667us            15
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
Self CPU time total: 70.108ms
Self CUDA time total: 1.539ms
```

Fig 2: CUDA memory usage and time using MLP

### *Tensorboard Visualizations*



Runs: mlp

Views: Overview

Workers: af339fd3836c_157

Spans: 1

Configuration

| | |
|---|---|
| Number of Worker(s) | 1 |
| Device Type | GPU |

GPU Summary

| | |
|---|---|
| GPU 0: | |
| Name | Tesla T4 |
| Memory | 14.75 GB |
| Compute Capability | 7.5 |
| GPU Utilization | 1.87 % |
| Est. SM Efficiency | 0.44 % |
| Est. Achieved Occupancy | 8.5 % |

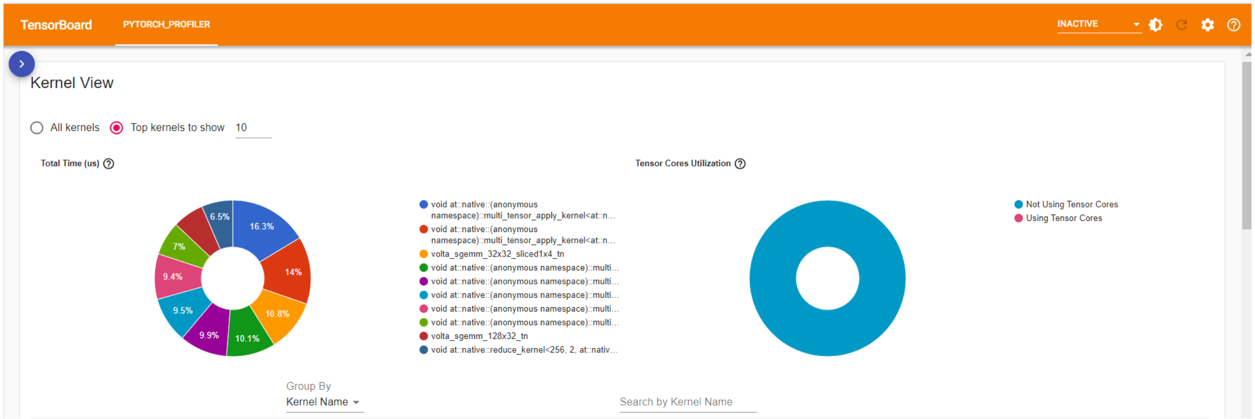
Execution Summary

| Category | Time Duration (us) | Percentage (%) |
|---|---|---|
| Average Step Time | 77,400 | 100 |
| Kernel | 1,451 | 1.87 |
| Memcpy | 38 | 0.05 |
| Memset | 0 | 0 |
| Runtime | 0 | 0 |
| DataLoader | 0 | 0 |
| CPU Exec | 19,293 | 24.93 |
| Other | 56,618 | 73.15 |

Step Time Breakdown

Performance Recommendation

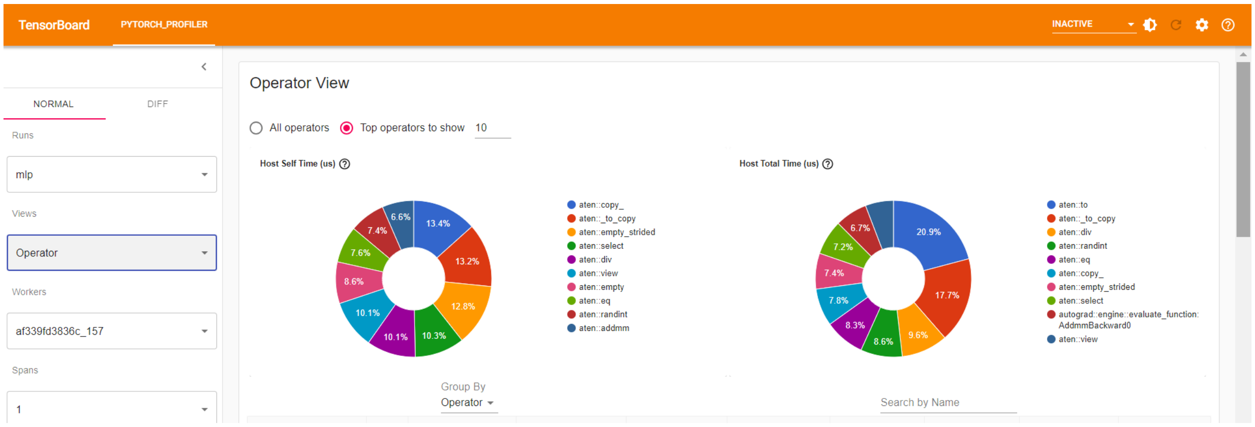
- GPU 0 has low utilization. You could try to increase batch size to improve. Note: Increasing batch size may affect the speed and stability of model convergence.


TensorBoard
PYTORCH_PROFILER
INACTIVE
Total Time (us)
16.3%
14%
10.8%
10.1%
9.9%
9.5%
9.4%
7%
6.5%
void at::native::(anonymous namespace)::multi_tensor_apply_kernel<at::n...
void at::native::(anonymous namespace)::multi_tensor_apply_kernel<at::n...
volta_sgemm_32x32_sliced1x4_tn
void at::native::(anonymous namespace)::multi...
void at::native::(anonymous namespace)::multi...
void at::native::(anonymous namespace)::multi...
void at::native::(anonymous namespace)::multi...
void at::native::(anonymous namespace)::multi...
volta_sgemm_128x32_tn
void at::native::reduce_kernel<256, 2, at::nativ...
Tensor Cores Utilization
Not Using Tensor Cores
Using Tensor Cores
Search by Kernel Name

Kernel View

○ All kernels ● Top kernels to show 10

Group By: Kernel Name



Runs: mlp

Views: Operator

Workers: af339fd3836c_157

Spans: 1

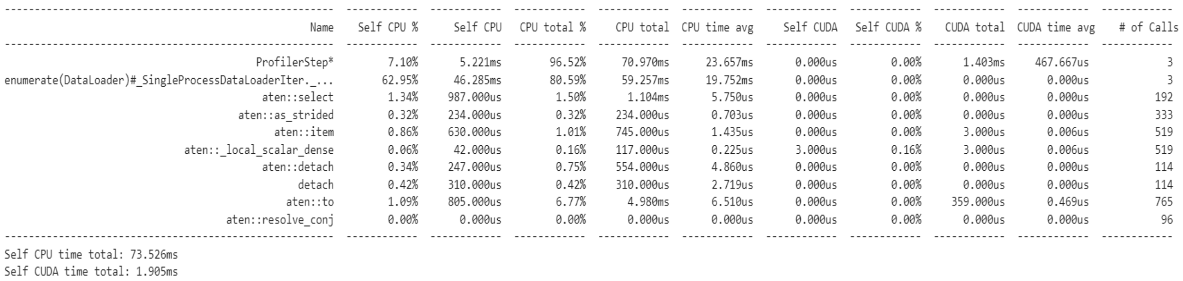
Operator View

○ All operators ● Top operators to show 10

Group By: Operator

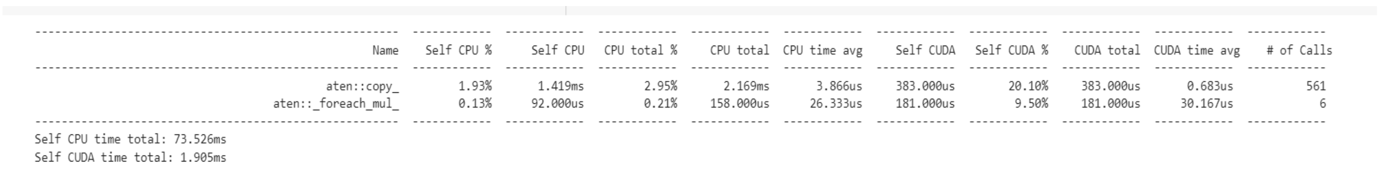
Fig 2 shows the two top operations of the model which used GPU and their memory usage.

### *Profiling with Automatic Mixed Precision*



```
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                                   Name    Self CPU %      Self CPU   CPU total %     CPU total  CPU time avg     Self CUDA   Self CUDA %    CUDA total  CUDA time avg    # of Calls
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                           ProfilerStep*         7.10%       5.221ms        96.52%      70.970ms      23.657ms       0.000us         0.00%       1.403ms     467.667us             3
enumerate(DataLoader)#_SingleProcessDataLoaderIter._...        62.95%      46.285ms        80.59%      59.257ms      19.752ms       0.000us         0.00%       0.000us       0.000us             3
                                            aten::select         1.34%     987.000us         1.50%       1.104ms       5.750us       0.000us         0.00%       0.000us       0.000us           192
                                        aten::as_strided         0.32%     234.000us         0.32%     234.000us       0.703us       0.000us         0.00%       0.000us       0.000us           333
                                              aten::item         0.86%     630.000us         1.01%     745.000us       1.435us       0.000us         0.00%       3.000us       0.006us           519
                               aten::_local_scalar_dense         0.06%      42.000us         0.16%     117.000us       0.225us       3.000us         0.16%       3.000us       0.006us           519
                                            aten::detach         0.34%     247.000us         0.75%     554.000us       4.860us       0.000us         0.00%       0.000us       0.000us           114
                                                  detach         0.42%     310.000us         0.42%     310.000us       2.719us       0.000us         0.00%       0.000us       0.000us           114
                                                aten::to         1.09%     805.000us         6.77%       4.980ms       6.510us       0.000us         0.00%     359.000us       0.469us           765
                                      aten::resolve_conj         0.00%       0.000us         0.00%       0.000us       0.000us       0.000us         0.00%       0.000us       0.000us            96
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
Self CPU time total: 73.526ms
Self CUDA time total: 1.905ms
```

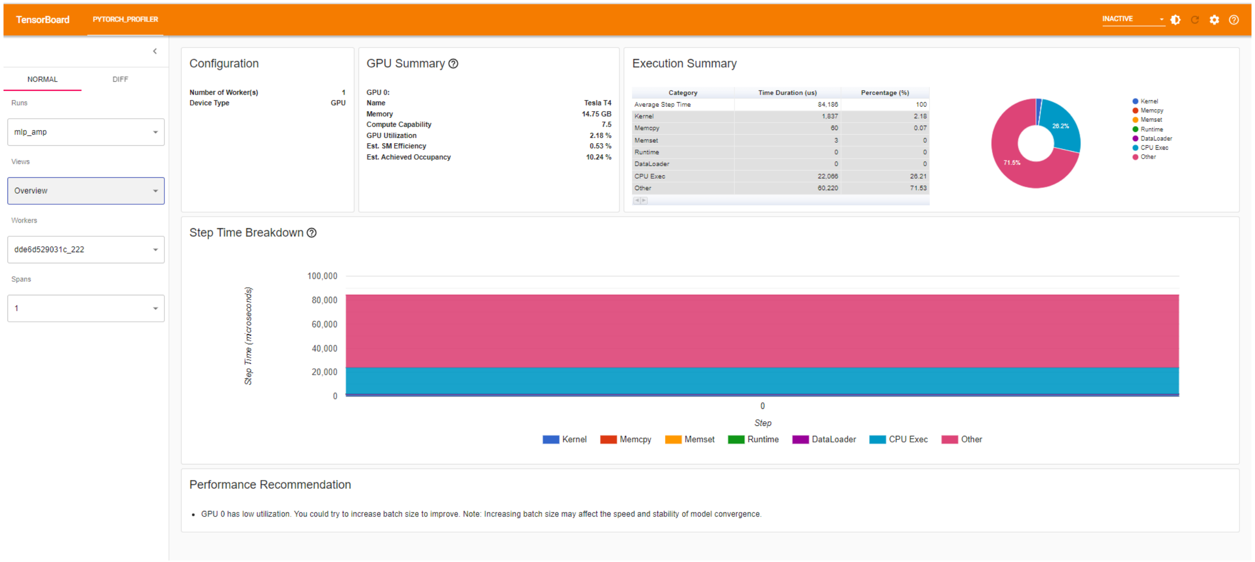
Fig 3: CPU memory usage and time using MLP with Automatic Mixed Precision



```
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                                   Name    Self CPU %      Self CPU   CPU total %     CPU total  CPU time avg     Self CUDA   Self CUDA %    CUDA total  CUDA time avg    # of Calls
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                             aten::copy_         1.93%       1.419ms         2.95%       2.169ms       3.866us     383.000us        20.10%     383.000us       0.683us           561
                                      aten::_foreach_mul_         0.13%      92.000us         0.21%     158.000us      26.333us     181.000us         9.50%     181.000us      30.167us             6
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
Self CPU time total: 73.526ms
Self CUDA time total: 1.905ms
```

Fig 4: CUDA memory usage and time for MLP with Automatic Mixed Precision

## *Tensorboard Visualizations*



Configuration

Number of Worker(s) 1
Device Type GPU

GPU Summary

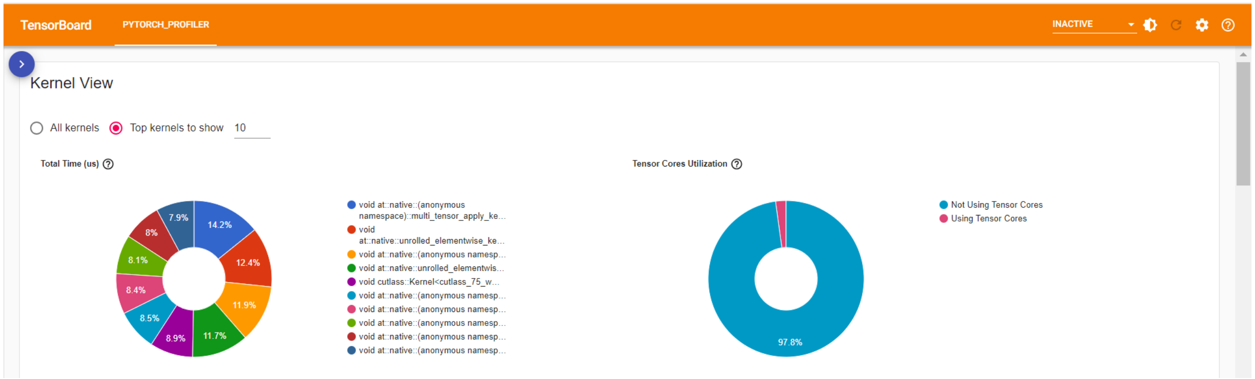
GPU 0:
Name Tesla T4
Memory 14.75 GB
Compute Capability 7.5
GPU Utilization 2.18 %
Est. SM Efficiency 0.53 %
Est. Achieved Occupancy 10.24 %

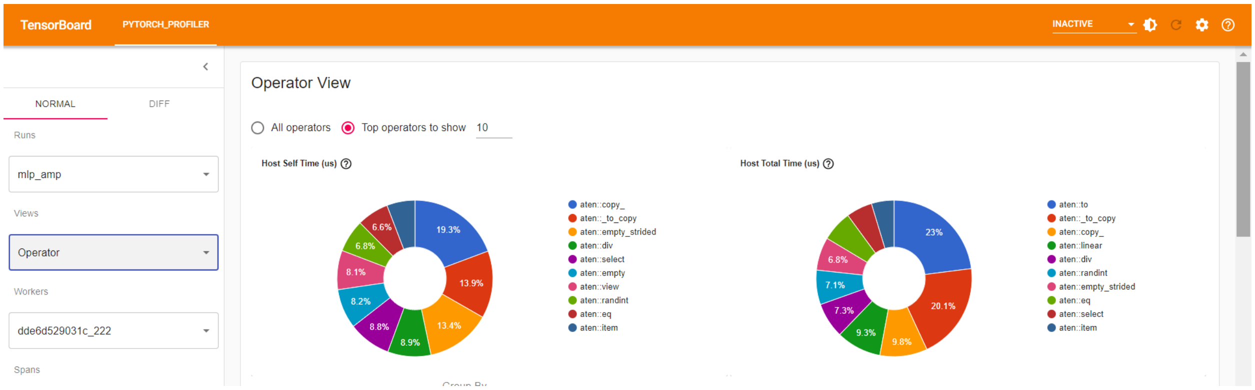
Execution Summary

| Category | Time Duration (us) | Percentage (%) |
|---|---|---|
| Average Step Time | 84,186 | 100 |
| Kernel | 1,837 | 2.18 |
| Memcpy | 60 | 0.07 |
| Memset | 3 | 0 |
| Runtime | 0 | 0 |
| DataLoader | 0 | 0 |
| CPU Exec | 22,066 | 26.21 |
| Other | 60,220 | 71.53 |

Performance Recommendation

- GPU 0 has low utilization. You could try to increase batch size to improve. Note: Increasing batch size may affect the speed and stability of model convergence.

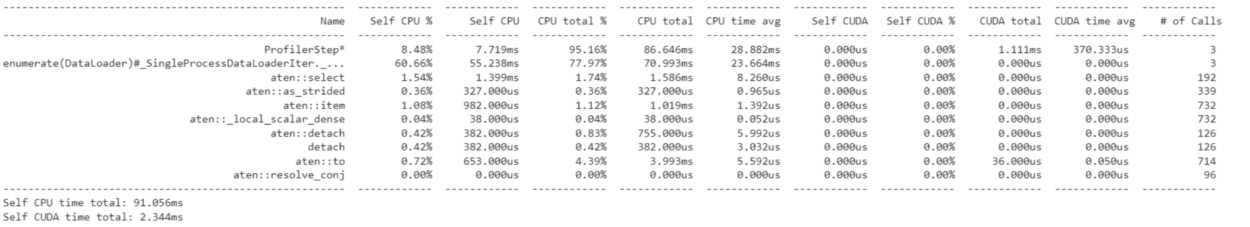

TensorBoard
PYTORCH_PROFILER
INACTIVE
Kernel View
○ All kernels
● Top kernels to show
10
Total Time (us)
7.9%
14.2%
8%
12.4%
8.1%
8.4%
11.9%
8.5%
8.9%
11.7%
void at::native::(anonymous namespace)::multi_tensor_apply_ke...
void at::native::unrolled_elementwise_ke...
void at::native::(anonymous namesp...
void at::native::unrolled_elementwis...
void cutlass::Kernel<cutlass_75_w...
void at::native::(anonymous namesp...
void at::native::(anonymous namesp...
void at::native::(anonymous namesp...
void at::native::(anonymous namesp...
void at::native::(anonymous namesp...
Tensor Cores Utilization
97.8%
Not Using Tensor Cores
Using Tensor Cores



Comparing the performance and memory usages using profiling with and without Automatic Mixed Precision from Fig 1-4 we can observe that there is an

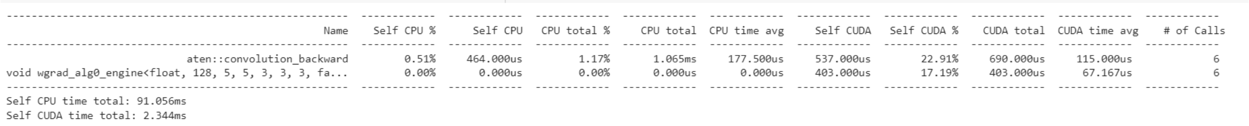
## LeNet

### *Profiling*



```
-----------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                                 Name    Self CPU %      Self CPU   CPU total %     CPU total  CPU time avg     Self CUDA   Self CUDA %    CUDA total  CUDA time avg    # of Calls
-----------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                        ProfilerStep*         8.48%       7.719ms        95.16%      86.646ms      28.882ms       0.000us         0.00%       1.111ms     370.333us             3
enumerate(DataLoader)#_SingleProcessDataLoaderIter._...        60.66%      55.238ms        77.97%      70.993ms      23.664ms       0.000us         0.00%       0.000us       0.000us             3
                                         aten::select         1.54%       1.399ms         1.74%       1.586ms       8.260us       0.000us         0.00%       0.000us       0.000us           192
                                     aten::as_strided         0.36%     327.000us         0.36%     327.000us       0.965us       0.000us         0.00%       0.000us       0.000us           339
                                           aten::item         1.08%     982.000us         1.12%       1.019ms       1.392us       0.000us         0.00%       0.000us       0.000us           732
                            aten::_local_scalar_dense         0.04%      38.000us         0.04%      38.000us       0.052us       0.000us         0.00%       0.000us       0.000us           732
                                         aten::detach         0.42%     382.000us         0.83%     755.000us       5.992us       0.000us         0.00%       0.000us       0.000us           126
                                               detach         0.42%     382.000us         0.42%     382.000us       3.032us       0.000us         0.00%       0.000us       0.000us           126
                                             aten::to         0.72%     653.000us         4.39%       3.993ms       5.592us       0.000us         0.00%      36.000us       0.050us           714
                                  aten::resolve_conj         0.00%       0.000us         0.00%       0.000us       0.000us       0.000us         0.00%       0.000us       0.000us            96
-----------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
Self CPU time total: 91.056ms
Self CUDA time total: 2.344ms
```

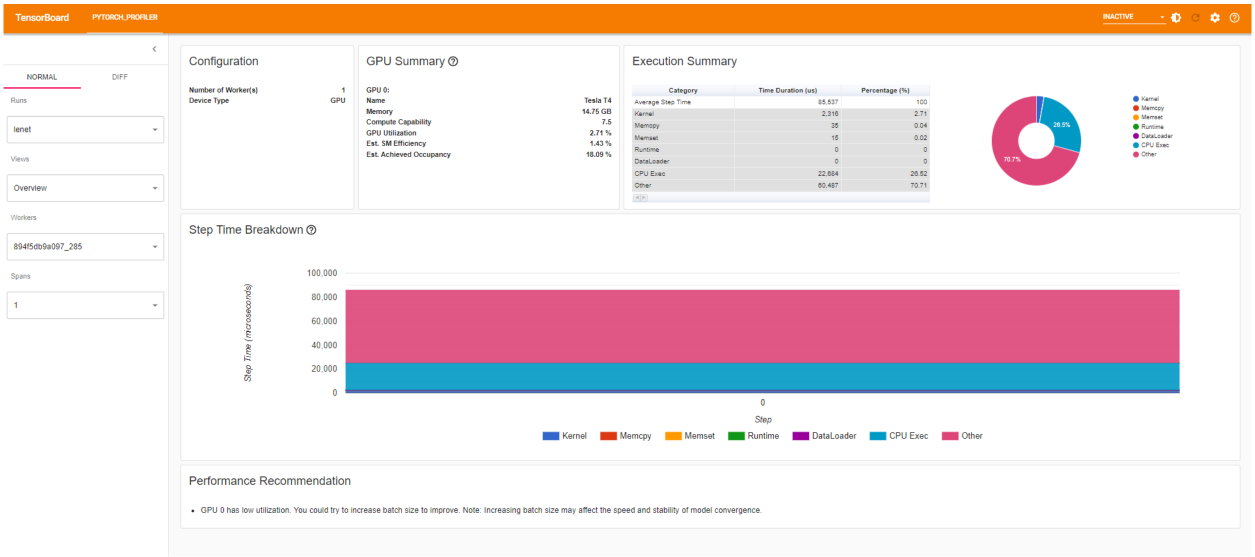
Fig 5: CPU memory usage and time for LeNet



```
-----------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                                 Name    Self CPU %      Self CPU   CPU total %     CPU total  CPU time avg     Self CUDA   Self CUDA %    CUDA total  CUDA time avg    # of Calls
-----------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                             aten::convolution_backward         0.51%     464.000us         1.17%       1.065ms     177.500us     537.000us        22.91%     690.000us     115.000us             6
void wgrad_alg0_engine<float, 128, 5, 5, 3, 3, 3, fa...         0.00%       0.000us         0.00%       0.000us       0.000us     403.000us        17.19%     403.000us      67.167us             6
-----------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
Self CPU time total: 91.056ms
Self CUDA time total: 2.344ms
```

Fig 6: CUDA memory usage and time for LeNet

## *Tensorboard Visualizations*



Configuration

Number of Worker(s) 1
Device Type GPU

GPU Summary

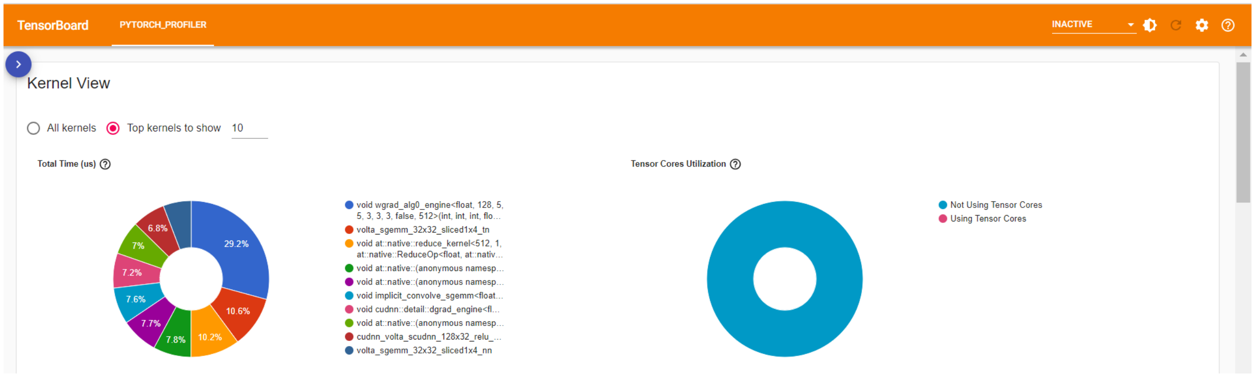
GPU 0:
Name Tesla T4
Memory 14.75 GB
Compute Capability 7.5
GPU Utilization 2.71 %
Est. SM Efficiency 1.43 %
Est. Achieved Occupancy 18.09 %

Execution Summary

| Category | Time Duration (us) | Percentage (%) |
|---|---|---|
| Average Step Time | 85,537 | 100 |
| Kernel | 2,316 | 2.71 |
| Memcpy | 35 | 0.04 |
| Memset | 15 | 0.02 |
| Runtime | 0 | 0 |
| DataLoader | 0 | 0 |
| CPU Exec | 22,684 | 26.52 |
| Other | 60,487 | 70.71 |

Performance Recommendation

- GPU 0 has low utilization. You could try to increase batch size to improve. Note: Increasing batch size may affect the speed and stability of model convergence.

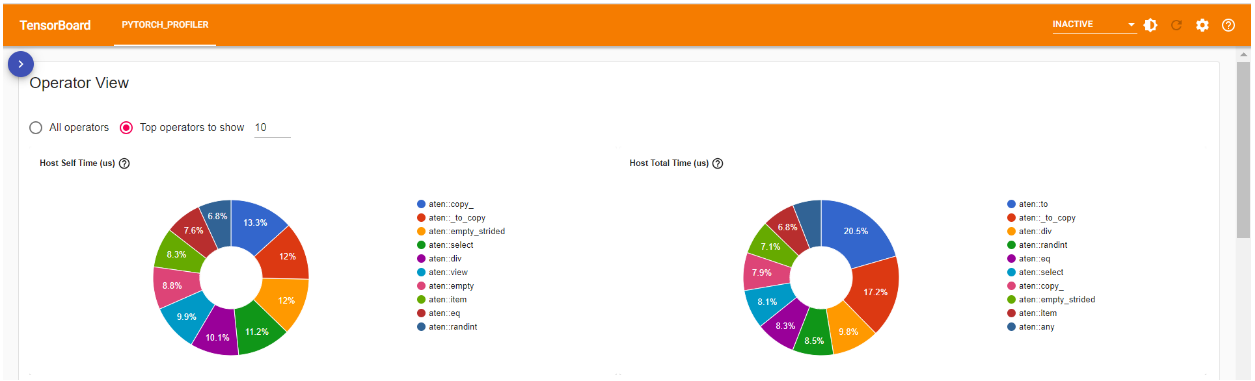


TensorBoard PYTORCH_PROFILER INACTIVE

Kernel View

○ All kernels ● Top kernels to show 10

Total Time (us)
29.2%
10.6%
10.2%
7.8%
7.7%
7.6%
7.2%
7%
6.8%
void wgrad_alg0_engine<float, 128, 5, 5, 3, 3, 3, false, 512>(int, int, int, flo...
volta_sgemm_32x32_sliced1x4_tn
void at::native::reduce_kernel<512, 1, at::native::ReduceOp<float, at::nativ...
void at::native::(anonymous namesp...
void at::native::(anonymous namesp...
void implicit_convolve_sgemm<float...
void cudnn::detail::dgrad_engine<fl...
void at::native::(anonymous namesp...
cudnn_volta_scudnn_128x32_relu_...
volta_sgemm_32x32_sliced1x4_nn
Tensor Cores Utilization
Not Using Tensor Cores
Using Tensor Cores

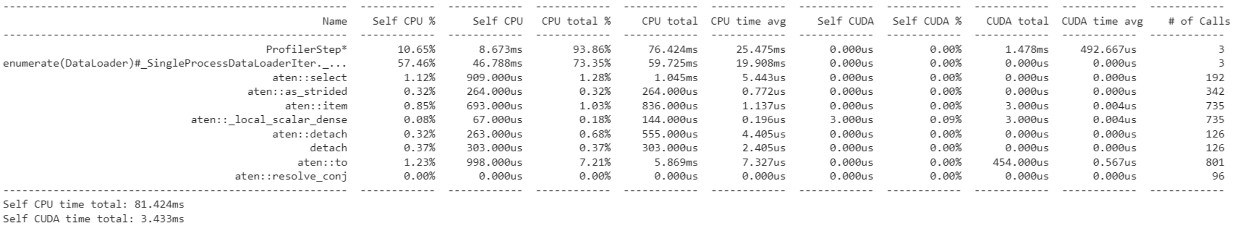


TensorBoard PYTORCH_PROFILER INACTIVE

Operator View

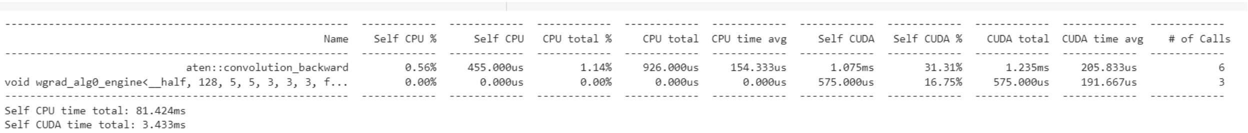
○ All operators ● Top operators to show 10

## *Profiling with Automatic Mixed Precision*



```
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                                   Name    Self CPU %      Self CPU   CPU total %     CPU total  CPU time avg     Self CUDA   Self CUDA %    CUDA total  CUDA time avg    # of Calls
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                          ProfilerStep*        10.65%       8.673ms        93.86%      76.424ms      25.475ms       0.000us         0.00%       1.478ms     492.667us             3
enumerate(DataLoader)#_SingleProcessDataLoaderIter._...        57.46%      46.788ms        73.35%      59.725ms      19.908ms       0.000us         0.00%       0.000us       0.000us             3
                                           aten::select         1.12%     909.000us         1.28%       1.045ms       5.443us       0.000us         0.00%       0.000us       0.000us           192
                                       aten::as_strided         0.32%     264.000us         0.32%     264.000us       0.772us       0.000us         0.00%       0.000us       0.000us           342
                                             aten::item         0.85%     693.000us         1.03%     836.000us       1.137us       0.000us         0.00%       3.000us       0.004us           735
                              aten::_local_scalar_dense         0.08%      67.000us         0.18%     144.000us       0.196us       3.000us         0.09%       3.000us       0.004us           735
                                           aten::detach         0.32%     263.000us         0.68%     555.000us       4.405us       0.000us         0.00%       0.000us       0.000us           126
                                                 detach         0.37%     303.000us         0.37%     303.000us       2.405us       0.000us         0.00%       0.000us       0.000us           126
                                               aten::to         1.23%     998.000us         7.21%       5.869ms       7.327us       0.000us         0.00%     454.000us       0.567us           801
                                     aten::resolve_conj         0.00%       0.000us         0.00%       0.000us       0.000us       0.000us         0.00%       0.000us       0.000us            96
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
Self CPU time total: 81.424ms
Self CUDA time total: 3.433ms
```

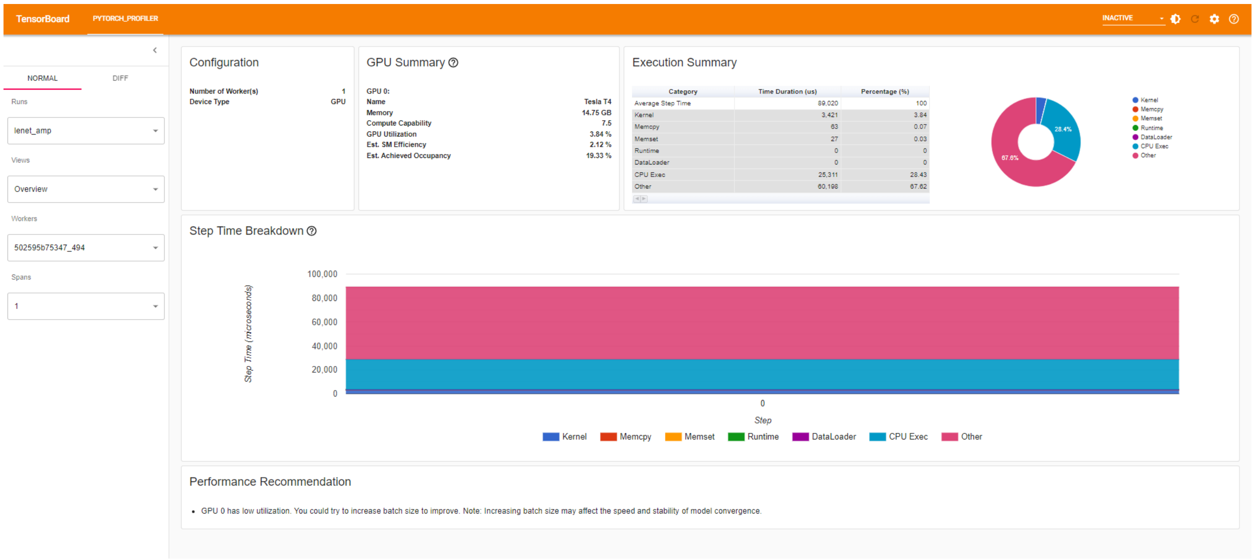
Fig 7: CPU memory usage and time for LeNet with automatic mixed precision



```
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                                   Name    Self CPU %      Self CPU   CPU total %     CPU total  CPU time avg     Self CUDA   Self CUDA %    CUDA total  CUDA time avg    # of Calls
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                               aten::convolution_backward         0.56%     455.000us         1.14%     926.000us     154.333us       1.075ms        31.31%       1.235ms     205.833us             6
void wgrad_alg0_engine<__half, 128, 5, 5, 3, 3, 3, f...         0.00%       0.000us         0.00%       0.000us       0.000us     575.000us        16.75%     575.000us     191.667us             3
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
Self CPU time total: 81.424ms
Self CUDA time total: 3.433ms
```

Fig 8: CUDA memory usage and time for LeNet with automatic mixed precision

## *Tensorboard Visualizations*



Configuration

|  |  |
|---|---|
| Number of Worker(s) | 1 |
| Device Type | GPU |

GPU Summary

|  |  |
|---|---|
| GPU 0: |  |
| Name | Tesla T4 |
| Memory | 14.75 GB |
| Compute Capability | 7.5 |
| GPU Utilization | 3.84 % |
| Est. SM Efficiency | 2.12 % |
| Est. Achieved Occupancy | 19.33 % |

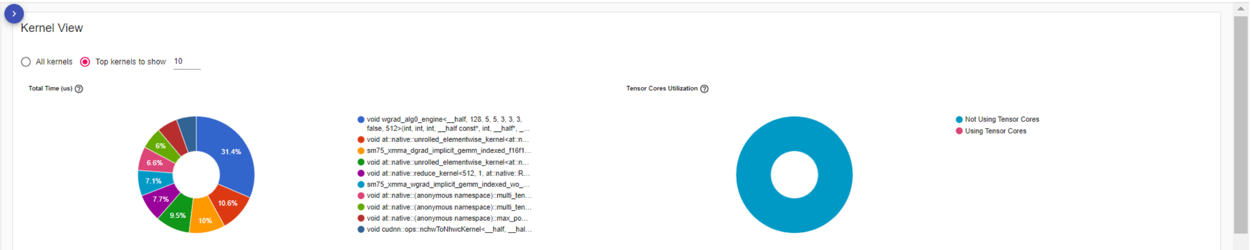
Execution Summary

| Category | Time Duration (us) | Percentage (%) |
|---|---|---|
| Average Step Time | 89,020 | 100 |
| Kernel | 3,421 | 3.84 |
| Memcpy | 63 | 0.07 |
| Memset | 27 | 0.03 |
| Runtime | 0 | 0 |
| DataLoader | 0 | 0 |
| CPU Exec | 25,311 | 28.43 |
| Other | 60,198 | 67.62 |

Step Time Breakdown

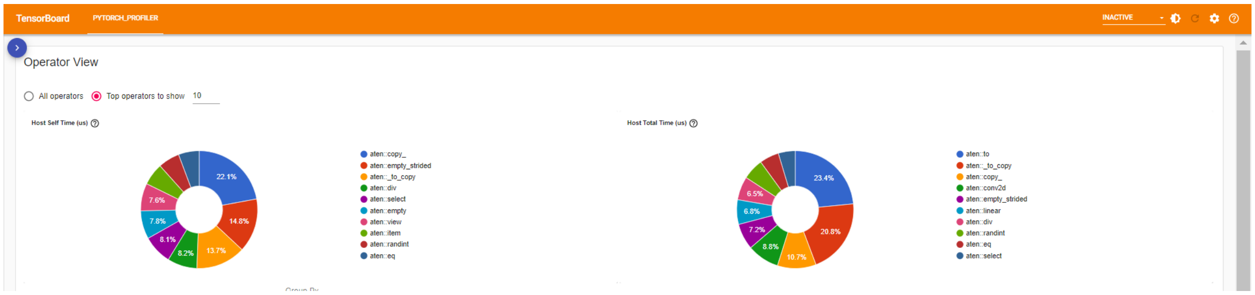
Performance Recommendation

- GPU 0 has low utilization. You could try to increase batch size to improve. Note: Increasing batch size may affect the speed and stability of model convergence.


Kernel View

○ All kernels ● Top kernels to show 10

Total Time (us)
31.4%
10.6%
10%
9.5%
7.7%
7.1%
6.6%
6%
void wgrad_alg0_engine<__half, 128, 5, 5, 3, 3, 3, false, 512>(int, int, int, __half const*, int, __half*, _...
void at::native::unrolled_elementwise_kernel<at::n...
sm75_xmma_dgrad_implicit_gemm_indexed_f16f1...
void at::native::unrolled_elementwise_kernel<at::n...
void at::native::reduce_kernel<512, 1, at::native::R...
sm75_xmma_wgrad_implicit_gemm_indexed_wo_...
void at::native::(anonymous namespace)::multi_ten...
void at::native::(anonymous namespace)::multi_ten...
void at::native::(anonymous namespace)::max_po...
void cudnn::ops::nchwToNhwcKernel<__half, __hal...
Tensor Cores Utilization
Not Using Tensor Cores
Using Tensor Cores



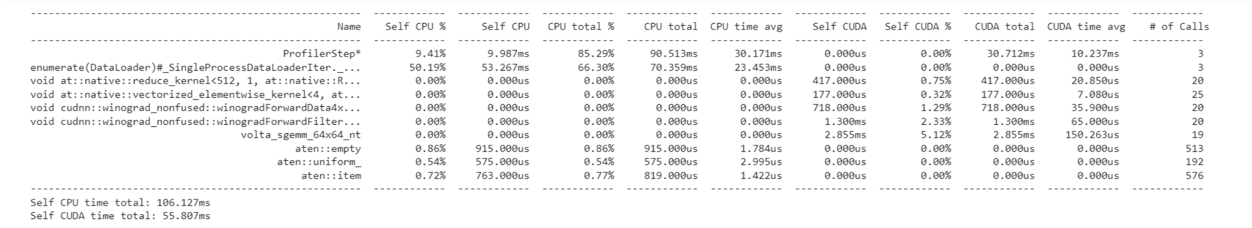
Operator View

○ All operators ● Top operators to show 10

Comparing the performance and memory usages using profiling with and without Automatic Mixed Precision from Fig 5-8 we can observe that there is an improvement in CPU performance and more tensor cores have been used to perform CUDA based operations.

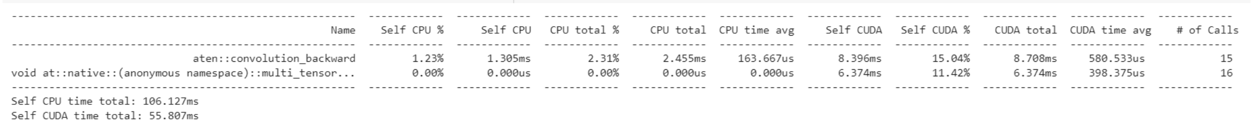
## AlexNet

### *Profiling*



```
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                                   Name    Self CPU %      Self CPU   CPU total %     CPU total  CPU time avg     Self CUDA   Self CUDA %    CUDA total  CUDA time avg    # of Calls
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                          ProfilerStep*         9.41%       9.987ms        85.29%      90.513ms      30.171ms       0.000us         0.00%      30.712ms      10.237ms             3
enumerate(DataLoader)#_SingleProcessDataLoaderIter._...        50.19%      53.267ms        66.30%      70.359ms      23.453ms       0.000us         0.00%       0.000us       0.000us             3
void at::native::reduce_kernel<512, 1, at::native::R...         0.00%       0.000us         0.00%       0.000us       0.000us     417.000us         0.75%     417.000us      20.850us            20
void at::native::vectorized_elementwise_kernel<4, at...         0.00%       0.000us         0.00%       0.000us       0.000us     177.000us         0.32%     177.000us       7.080us            25
void cudnn::winograd_nonfused::winogradForwardData4x...         0.00%       0.000us         0.00%       0.000us       0.000us     718.000us         1.29%     718.000us      35.900us            20
void cudnn::winograd_nonfused::winogradForwardFilter...         0.00%       0.000us         0.00%       0.000us       0.000us       1.300ms         2.33%       1.300ms      65.000us            20
                                      volta_sgemm_64x64_nt         0.00%       0.000us         0.00%       0.000us       0.000us       2.855ms         5.12%       2.855ms     150.263us            19
                                            aten::empty         0.86%     915.000us         0.86%     915.000us       1.784us       0.000us         0.00%       0.000us       0.000us           513
                                         aten::uniform_         0.54%     575.000us         0.54%     575.000us       2.995us       0.000us         0.00%       0.000us       0.000us           192
                                             aten::item         0.72%     763.000us         0.77%     819.000us       1.422us       0.000us         0.00%       0.000us       0.000us           576
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
Self CPU time total: 106.127ms
Self CUDA time total: 55.807ms
```

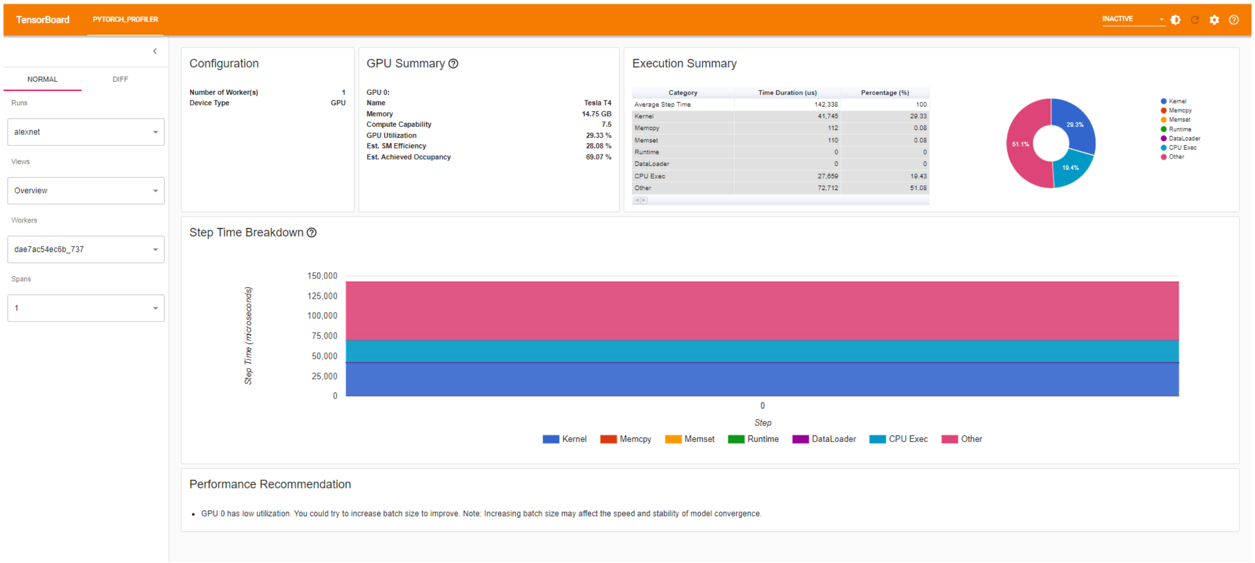
Fig 9: CPU memory usage and time for AlexNet



```
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                                   Name    Self CPU %      Self CPU   CPU total %     CPU total  CPU time avg     Self CUDA   Self CUDA %    CUDA total  CUDA time avg    # of Calls
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                               aten::convolution_backward         1.23%       1.305ms         2.31%       2.455ms     163.667us       8.396ms        15.04%       8.708ms     580.533us            15
void at::native::(anonymous namespace)::multi_tensor...         0.00%       0.000us         0.00%       0.000us       0.000us       6.374ms        11.42%       6.374ms     398.375us            16
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
Self CPU time total: 106.127ms
Self CUDA time total: 55.807ms
```

Fig 10: CUDA  memory usage and time for AlexNet

## *Tensorboard Visualizations*



Configuration

| | |
|---|---|
| Number of Worker(s) | 1 |
| Device Type | GPU |

GPU Summary

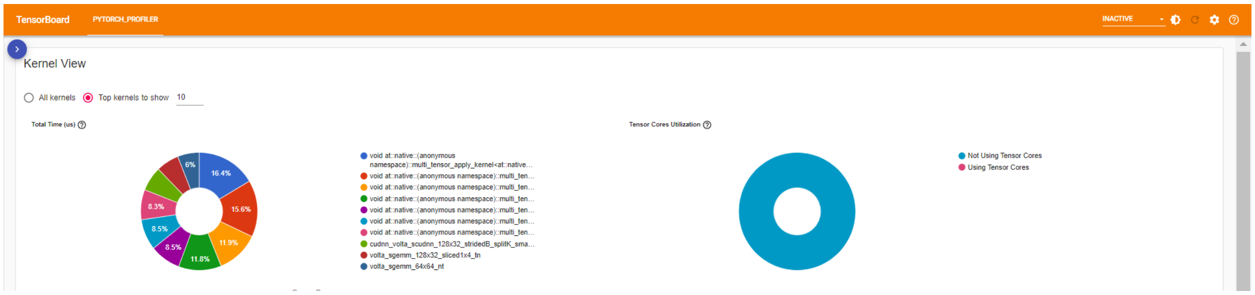
| | |
|---|---|
| GPU 0: | |
| Name | Tesla T4 |
| Memory | 14.75 GB |
| Compute Capability | 7.5 |
| GPU Utilization | 29.33 % |
| Est. SM Efficiency | 28.08 % |
| Est. Achieved Occupancy | 69.07 % |

Execution Summary

| Category | Time Duration (us) | Percentage (%) |
|---|---|---|
| Average Step Time | 142,338 | 100 |
| Kernel | 41,745 | 29.33 |
| Memcpy | 112 | 0.08 |
| Memset | 110 | 0.08 |
| Runtime | 0 | 0 |
| DataLoader | 0 | 0 |
| CPU Exec | 27,659 | 19.43 |
| Other | 72,712 | 51.08 |

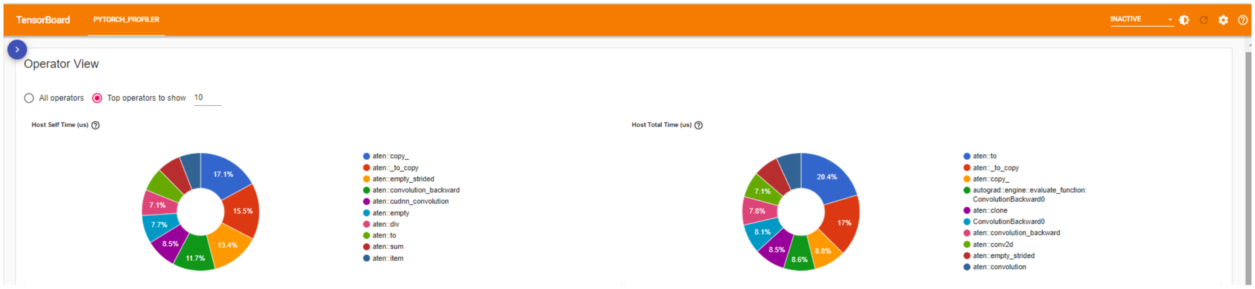
Performance Recommendation

- GPU 0 has low utilization. You could try to increase batch size to improve. Note: Increasing batch size may affect the speed and stability of model convergence.



TensorBoard
PYTORCH_PROFILER
INACTIVE
Kernel View
Total Time (us)
6%
16.4%
15.6%
11.9%
11.8%
8.5%
8.5%
8.3%
void at::native::(anonymous namespace)::multi_tensor_apply_kernel<at::native...
void at::native::(anonymous namespace)::multi_ten...
void at::native::(anonymous namespace)::multi_ten...
void at::native::(anonymous namespace)::multi_ten...
void at::native::(anonymous namespace)::multi_ten...
void at::native::(anonymous namespace)::multi_ten...
void at::native::(anonymous namespace)::multi_ten...
cudnn_volta_scudnn_128x32_stridedB_splitK_sma...
volta_sgemm_128x32_sliced1x4_tn
volta_sgemm_64x64_nt
Tensor Cores Utilization
Not Using Tensor Cores
Using Tensor Cores

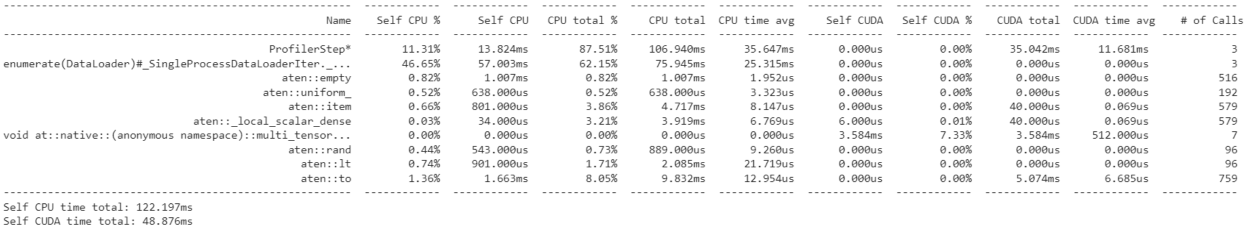
○ All kernels ● Top kernels to show 10



Operator View

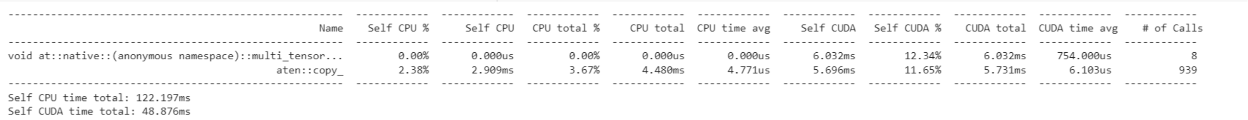
○ All operators ● Top operators to show 10

## Profiling with Automatic Mixed Precision



```
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                                   Name    Self CPU %      Self CPU   CPU total %     CPU total  CPU time avg     Self CUDA   Self CUDA %    CUDA total  CUDA time avg    # of Calls
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                          ProfilerStep*        11.31%      13.824ms        87.51%     106.940ms      35.647ms       0.000us         0.00%      35.042ms      11.681ms             3
enumerate(DataLoader)#_SingleProcessDataLoaderIter._...        46.65%      57.003ms        62.15%      75.945ms      25.315ms       0.000us         0.00%       0.000us       0.000us             3
                                            aten::empty         0.82%       1.007ms         0.82%       1.007ms       1.952us       0.000us         0.00%       0.000us       0.000us           516
                                         aten::uniform_         0.52%     638.000us         0.52%     638.000us       3.323us       0.000us         0.00%       0.000us       0.000us           192
                                             aten::item         0.66%     801.000us         3.86%       4.717ms       8.147us       0.000us         0.00%      40.000us       0.069us           579
                              aten::_local_scalar_dense         0.03%      34.000us         3.21%       3.919ms       6.769us       6.000us         0.01%      40.000us       0.069us           579
void at::native::(anonymous namespace)::multi_tensor...         0.00%       0.000us         0.00%       0.000us       0.000us       3.584ms         7.33%       3.584ms     512.000us             7
                                             aten::rand         0.44%     543.000us         0.73%     889.000us       9.260us       0.000us         0.00%       0.000us       0.000us            96
                                               aten::lt         0.74%     901.000us         1.71%       2.085ms      21.719us       0.000us         0.00%       0.000us       0.000us            96
                                               aten::to         1.36%       1.663ms         8.05%       9.832ms      12.954us       0.000us         0.00%       5.074ms       6.685us           759
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
Self CPU time total: 122.197ms
Self CUDA time total: 48.876ms
```

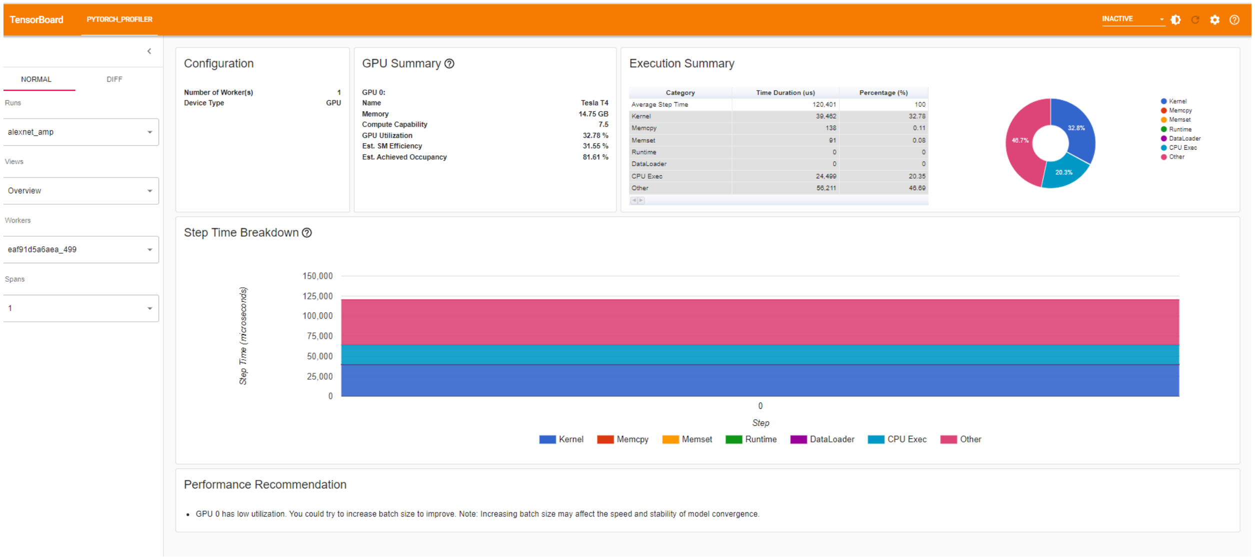
Fig 11: CPU memory usage and time for AlexNet with automatic mixed precision



```
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                                   Name    Self CPU %      Self CPU   CPU total %     CPU total  CPU time avg     Self CUDA   Self CUDA %    CUDA total  CUDA time avg    # of Calls
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
void at::native::(anonymous namespace)::multi_tensor...         0.00%       0.000us         0.00%       0.000us       0.000us       6.032ms        12.34%       6.032ms     754.000us             8
                                            aten::copy_         2.38%       2.909ms         3.67%       4.480ms       4.771us       5.696ms        11.65%       5.731ms       6.103us           939
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
Self CPU time total: 122.197ms
Self CUDA time total: 48.876ms
```

Fig 12: CUDA memory usage and time for AlexNet with automatic mixed precision

## ***Tensorboard Visualizations***



Runs: alexnet_amp

Views: Overview

Workers: eaf91d5a6aea_499

Spans: 1

Configuration

| | |
|---|---|
| Number of Worker(s) | 1 |
| Device Type | GPU |

GPU Summary

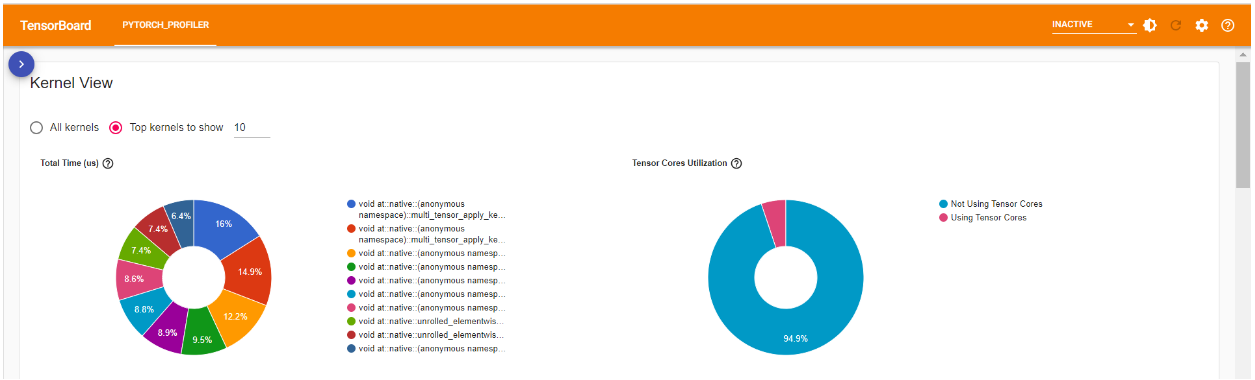
| | |
|---|---|
| GPU 0: | |
| Name | Tesla T4 |
| Memory | 14.75 GB |
| Compute Capability | 7.5 |
| GPU Utilization | 32.78 % |
| Est. SM Efficiency | 31.55 % |
| Est. Achieved Occupancy | 81.61 % |

Execution Summary

| Category | Time Duration (us) | Percentage (%) |
|---|---|---|
| Average Step Time | 120,401 | 100 |
| Kernel | 39,462 | 32.78 |
| Memcpy | 138 | 0.11 |
| Memset | 91 | 0.08 |
| Runtime | 0 | 0 |
| DataLoader | 0 | 0 |
| CPU Exec | 24,499 | 20.35 |
| Other | 56,211 | 46.69 |

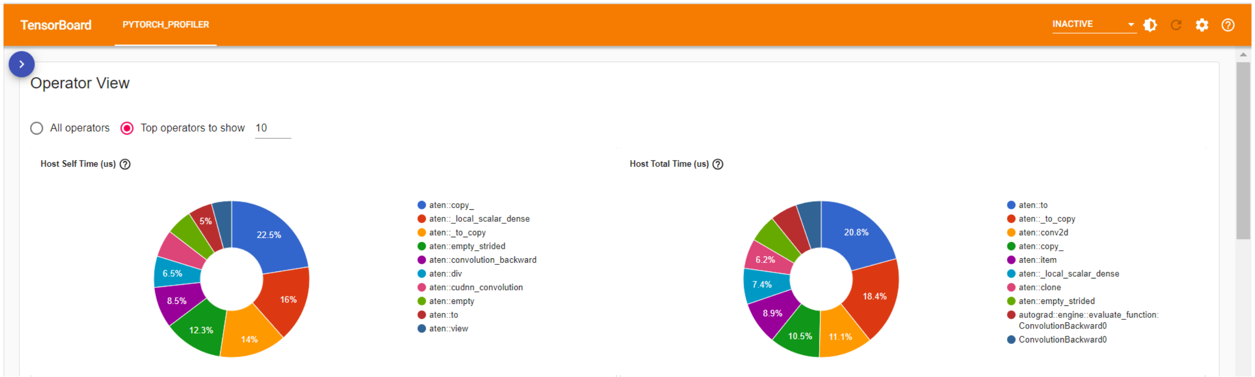
Performance Recommendation

- GPU 0 has low utilization. You could try to increase batch size to improve. Note: Increasing batch size may affect the speed and stability of model convergence.


TensorBoard
PYTORCH_PROFILER
INACTIVE
Total Time (us)
16%
14.9%
12.2%
9.5%
8.9%
8.8%
8.6%
7.4%
7.4%
6.4%
void at::native::(anonymous namespace)::multi_tensor_apply_ke...
void at::native::(anonymous namespace)::multi_tensor_apply_ke...
void at::native::(anonymous namesp...
void at::native::(anonymous namesp...
void at::native::(anonymous namesp...
void at::native::(anonymous namesp...
void at::native::(anonymous namesp...
void at::native::unrolled_elementwis...
void at::native::unrolled_elementwis...
void at::native::(anonymous namesp...
Tensor Cores Utilization
94.9%
Not Using Tensor Cores
Using Tensor Cores

Kernel View

○ All kernels ● Top kernels to show 10

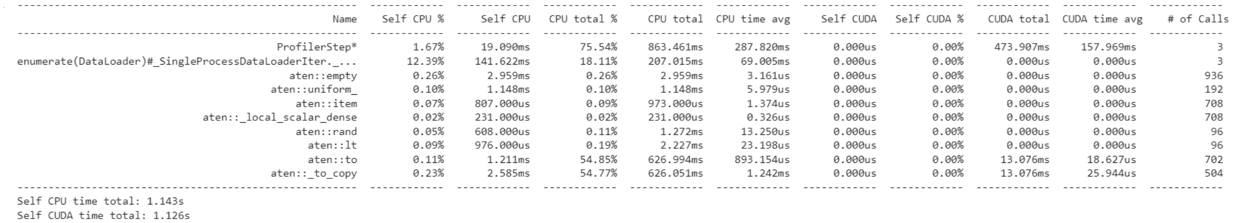


Operator View

○ All operators ● Top operators to show 10

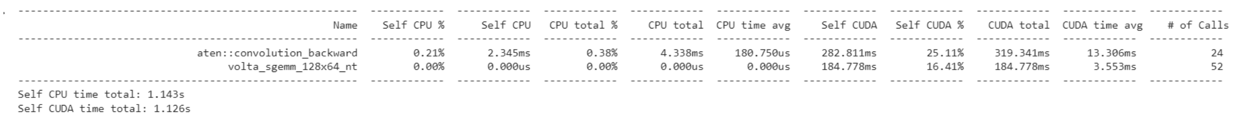
# VGG

## *Profiling*



```
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                                   Name    Self CPU %      Self CPU   CPU total %     CPU total  CPU time avg     Self CUDA   Self CUDA %    CUDA total CUDA time avg    # of Calls
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                          ProfilerStep*         1.67%      19.090ms        75.54%     863.461ms     287.820ms       0.000us         0.00%     473.907ms     157.969ms             3
enumerate(DataLoader)#_SingleProcessDataLoaderIter._...        12.39%     141.622ms        18.11%     207.015ms      69.005ms       0.000us         0.00%       0.000us       0.000us             3
                                            aten::empty         0.26%       2.959ms         0.26%       2.959ms       3.161us       0.000us         0.00%       0.000us       0.000us           936
                                         aten::uniform_         0.10%       1.148ms         0.10%       1.148ms       5.979us       0.000us         0.00%       0.000us       0.000us           192
                                             aten::item         0.07%     807.000us         0.09%     973.000us       1.374us       0.000us         0.00%       0.000us       0.000us           708
                              aten::_local_scalar_dense         0.02%     231.000us         0.02%     231.000us       0.326us       0.000us         0.00%       0.000us       0.000us           708
                                             aten::rand         0.05%     608.000us         0.11%       1.272ms      13.250us       0.000us         0.00%       0.000us       0.000us            96
                                               aten::lt         0.09%     976.000us         0.19%       2.227ms      23.198us       0.000us         0.00%       0.000us       0.000us            96
                                               aten::to         0.11%       1.211ms        54.85%     626.994ms     893.154us       0.000us         0.00%      13.076ms      18.627us           702
                                         aten::_to_copy         0.23%       2.585ms        54.77%     626.051ms       1.242ms       0.000us         0.00%      13.076ms      25.944us           504
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
Self CPU time total: 1.143s
Self CUDA time total: 1.126s
```

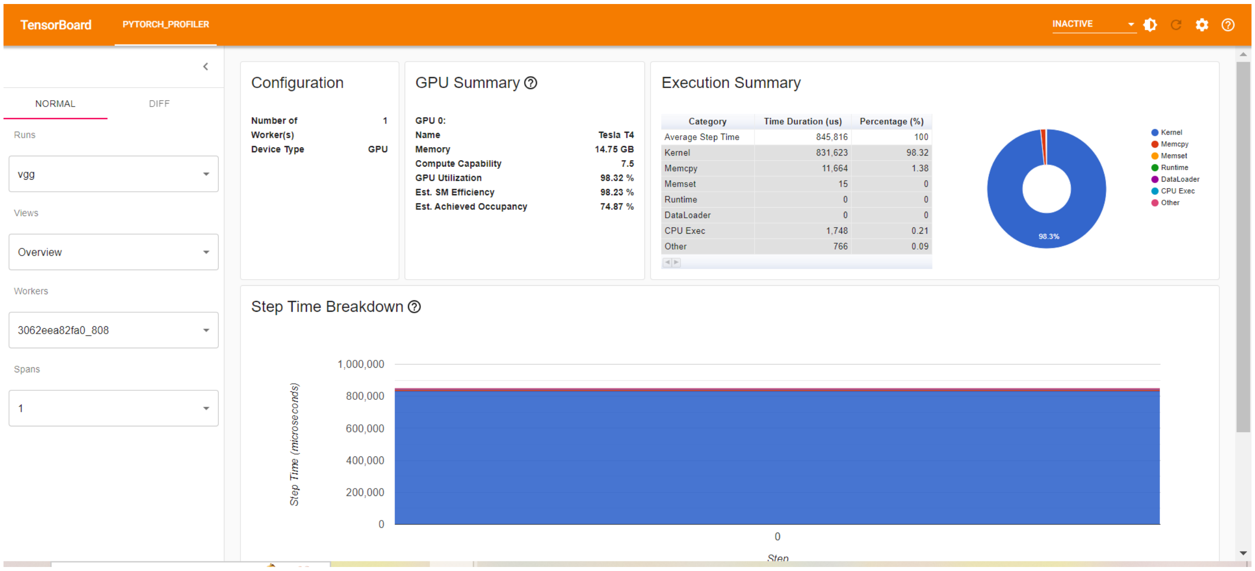
Fig 13: CPU memory usage and time for VGG



```
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                                   Name    Self CPU %      Self CPU   CPU total %     CPU total  CPU time avg     Self CUDA   Self CUDA %    CUDA total CUDA time avg    # of Calls
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                             aten::convolution_backward         0.21%       2.345ms         0.38%       4.338ms     180.750us     282.811ms        25.11%     319.341ms      13.306ms            24
                                  volta_sgemm_128x64_nt         0.00%       0.000us         0.00%       0.000us       0.000us     184.778ms        16.41%     184.778ms       3.553ms            52
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
Self CPU time total: 1.143s
Self CUDA time total: 1.126s
```

Fig 14: CUDA memory usage and time for VGG

## *Tensorflow Visualizations*



Runs: vgg

Views: Overview

Workers: 3062eea82fa0_808

Spans: 1

Configuration

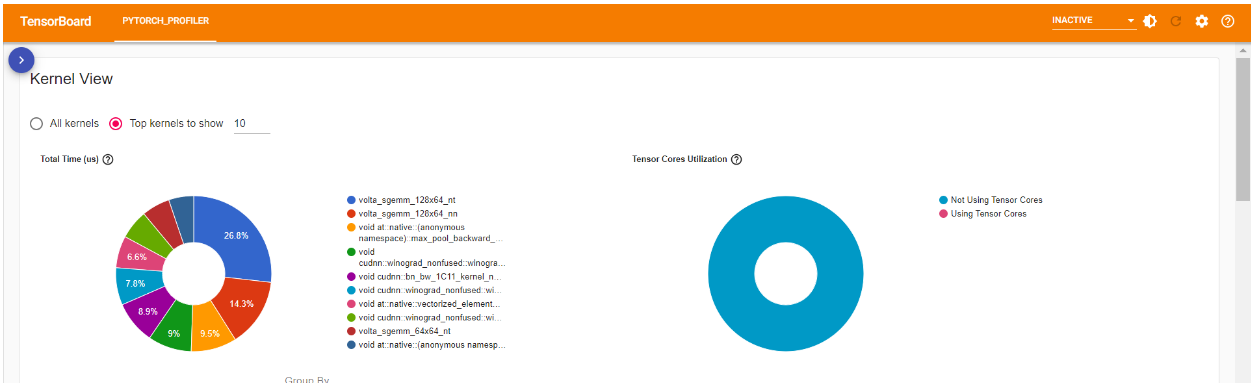
|  |  |
|---|---|
| Number of Worker(s) | 1 |
| Device Type | GPU |

GPU Summary

| GPU 0: |  |
|---|---|
| Name | Tesla T4 |
| Memory | 14.75 GB |
| Compute Capability | 7.5 |
| GPU Utilization | 98.32 % |
| Est. SM Efficiency | 98.23 % |
| Est. Achieved Occupancy | 74.87 % |

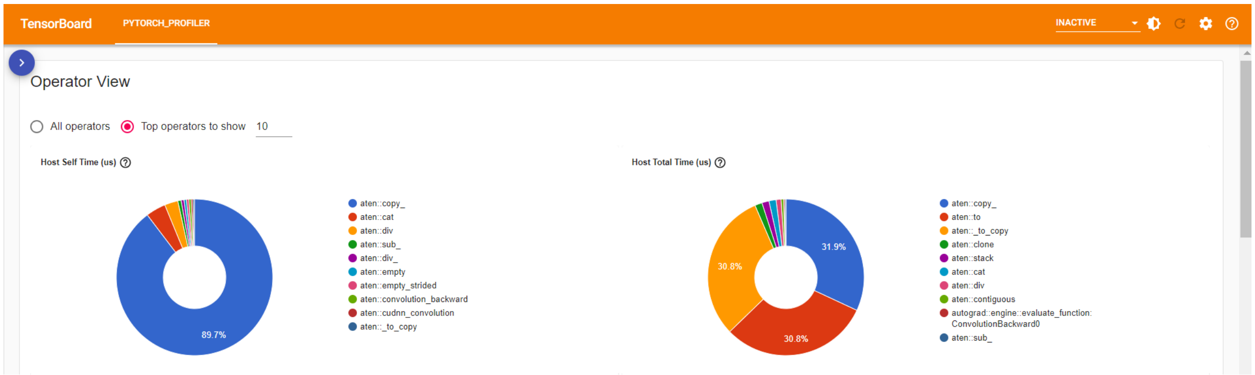
Execution Summary

| Category | Time Duration (us) | Percentage (%) |
|---|---|---|
| Average Step Time | 845,816 | 100 |
| Kernel | 831,623 | 98.32 |
| Memcpy | 11,664 | 1.38 |
| Memset | 15 | 0 |
| Runtime | 0 | 0 |
| DataLoader | 0 | 0 |
| CPU Exec | 1,748 | 0.21 |
| Other | 766 | 0.09 |


TensorBoard
PYTORCH_PROFILER
INACTIVE
Total Time (us)
26.8%
14.3%
9.5%
9%
8.9%
7.8%
6.6%
volta_sgemm_128x64_nt
volta_sgemm_128x64_nn
void at::native::(anonymous namespace)::max_pool_backward_...
void cudnn::winograd_nonfused::winogra...
void cudnn::bn_bw_1C11_kernel_n...
void cudnn::winograd_nonfused::wi...
void at::native::vectorized_element...
void cudnn::winograd_nonfused::wi...
volta_sgemm_64x64_nt
void at::native::(anonymous namesp...
Tensor Cores Utilization
Not Using Tensor Cores
Using Tensor Cores
Group By

Kernel View

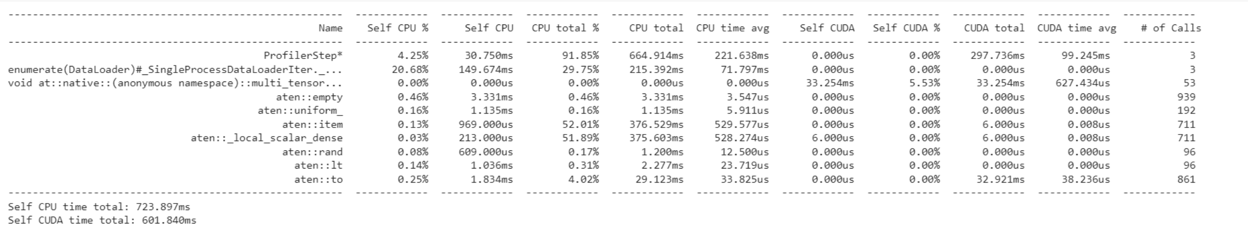
○ All kernels ● Top kernels to show 10



Operator View

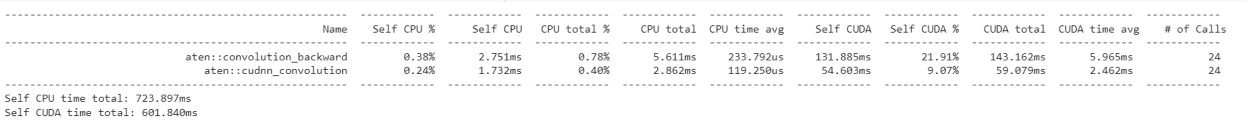
○ All operators ● Top operators to show 10

## *Profiling with Automatic Mixed Precision*



```
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                                   Name    Self CPU %      Self CPU   CPU total %     CPU total  CPU time avg     Self CUDA   Self CUDA %    CUDA total  CUDA time avg    # of Calls
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                          ProfilerStep*         4.25%      30.750ms        91.85%     664.914ms     221.638ms       0.000us         0.00%     297.736ms      99.245ms             3
enumerate(DataLoader)#_SingleProcessDataLoaderIter._...        20.68%     149.674ms        29.75%     215.392ms      71.797ms       0.000us         0.00%       0.000us       0.000us             3
void at::native::(anonymous namespace)::multi_tensor...         0.00%       0.000us         0.00%       0.000us       0.000us      33.254ms         5.53%      33.254ms     627.434us            53
                                            aten::empty         0.46%       3.331ms         0.46%       3.331ms       3.547us       0.000us         0.00%       0.000us       0.000us           939
                                         aten::uniform_         0.16%       1.135ms         0.16%       1.135ms       5.911us       0.000us         0.00%       0.000us       0.000us           192
                                             aten::item         0.13%     969.000us        52.01%     376.529ms     529.577us       0.000us         0.00%       6.000us       0.008us           711
                              aten::_local_scalar_dense         0.03%     213.000us        51.89%     375.603ms     528.274us       6.000us         0.00%       6.000us       0.008us           711
                                             aten::rand         0.08%     609.000us         0.17%       1.200ms      12.500us       0.000us         0.00%       0.000us       0.000us            96
                                               aten::lt         0.14%       1.036ms         0.31%       2.277ms      23.719us       0.000us         0.00%       0.000us       0.000us            96
                                               aten::to         0.25%       1.834ms         4.02%      29.123ms      33.825us       0.000us         0.00%      32.921ms      38.236us           861
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
Self CPU time total: 723.897ms
Self CUDA time total: 601.840ms
```

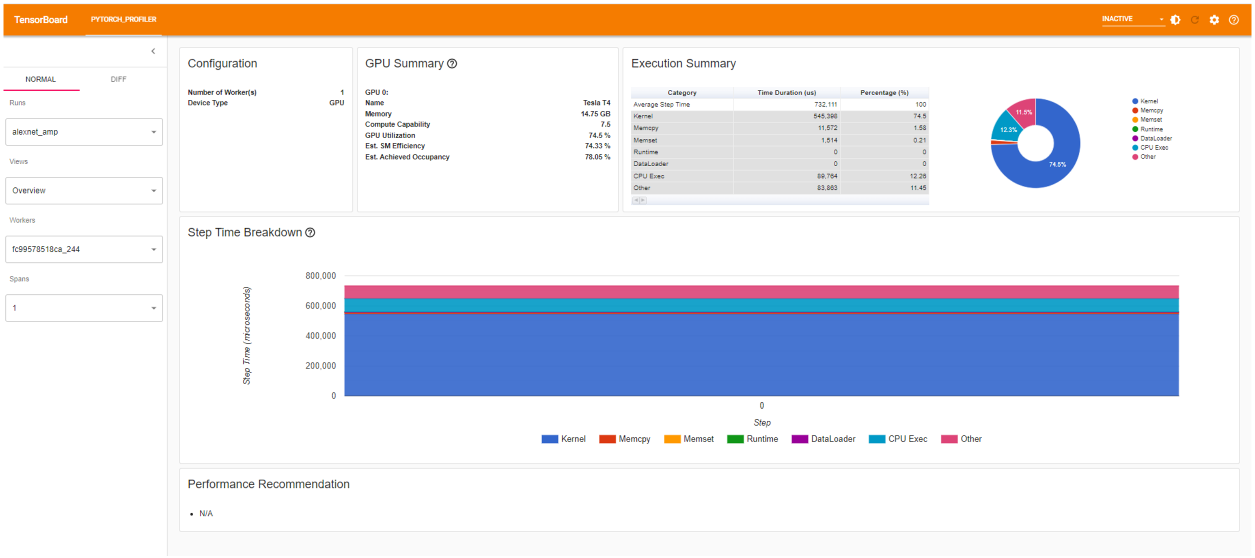
Fig 15: CPU memory usage and time for VGG with automatic mixed precision



```
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                                   Name    Self CPU %      Self CPU   CPU total %     CPU total  CPU time avg     Self CUDA   Self CUDA %    CUDA total  CUDA time avg    # of Calls
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                               aten::convolution_backward         0.38%       2.751ms         0.78%       5.611ms     233.792us     131.885ms        21.91%     143.162ms       5.965ms            24
                                  aten::cudnn_convolution         0.24%       1.732ms         0.40%       2.862ms     119.250us      54.603ms         9.07%      59.079ms       2.462ms            24
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
Self CPU time total: 723.897ms
Self CUDA time total: 601.840ms
```

Fig 16: CUDA memory usage and time for VGG with automatic mixed precision

## *Tensorflow Visualizations*



Runs: alexnet_amp

Views: Overview

Workers: fc99578518ca_244

Spans: 1

Configuration

| | |
|---|---|
| Number of Worker(s) | 1 |
| Device Type | GPU |

GPU Summary

| | |
|---|---|
| GPU 0: | |
| Name | Tesla T4 |
| Memory | 14.75 GB |
| Compute Capability | 7.5 |
| GPU Utilization | 74.5 % |
| Est. SM Efficiency | 74.33 % |
| Est. Achieved Occupancy | 78.05 % |

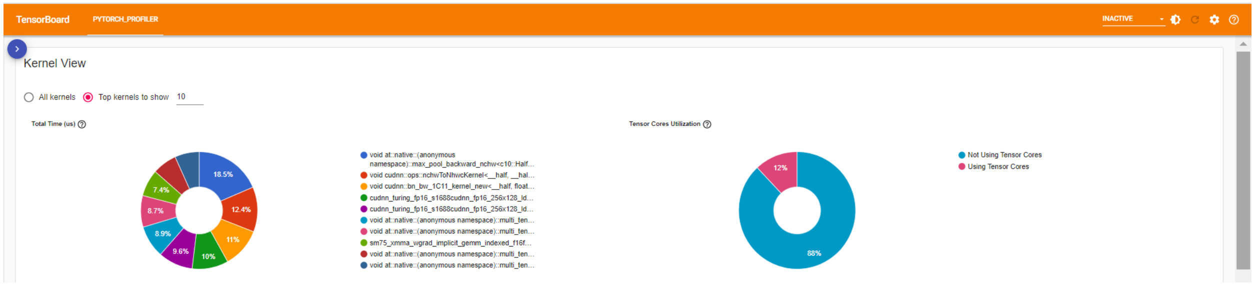
Execution Summary

| Category | Time Duration (us) | Percentage (%) |
|---|---|---|
| Average Step Time | 732,111 | 100 |
| Kernel | 545,398 | 74.5 |
| Memcpy | 11,572 | 1.58 |
| Memset | 1,514 | 0.21 |
| Runtime | 0 | 0 |
| DataLoader | 0 | 0 |
| CPU Exec | 89,764 | 12.26 |
| Other | 83,863 | 11.45 |

Step Time Breakdown

Performance Recommendation

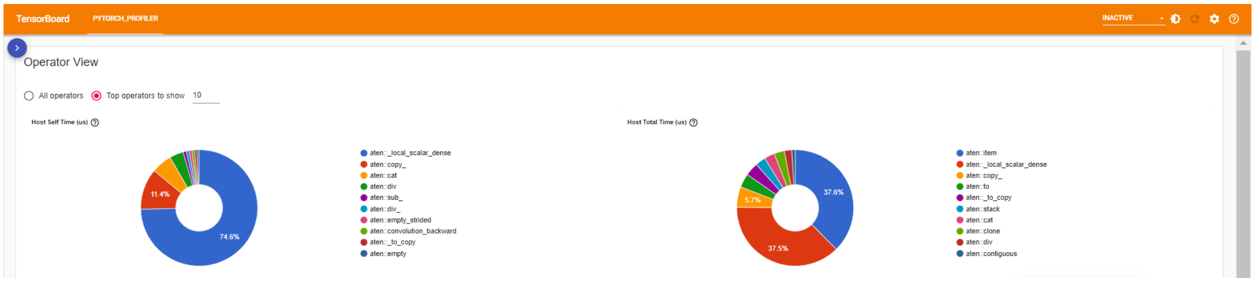
- N/A



Kernel View

○ All kernels ● Top kernels to show 10

- void at::native::(anonymous namespace)::max_pool_backward_nchw<c10::Half...
- void cudnn::ops::nchwToNhwcKernel<__half, __hal...
- void cudnn::bn_bw_1C11_kernel_new<__half, float...
- cudnn_turing_fp16_s1688cudnn_fp16_256x128_ld...
- cudnn_turing_fp16_s1688cudnn_fp16_256x128_ld...
- void at::native::(anonymous namespace)::multi_ten...
- void at::native::(anonymous namespace)::multi_ten...
- sm75_xmma_wgrad_implicit_gemm_indexed_f16f...
- void at::native::(anonymous namespace)::multi_ten...
- void at::native::(anonymous namespace)::multi_ten...

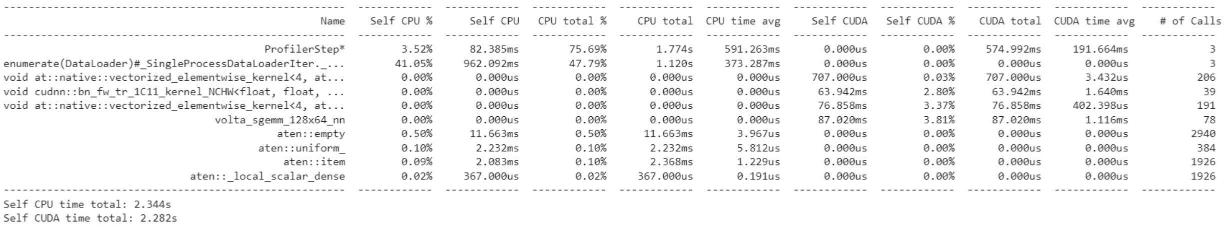


Operator View

○ All operators ● Top operators to show 10

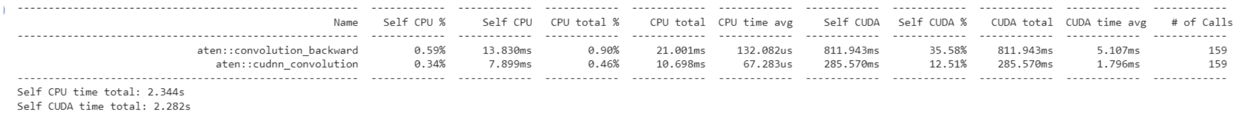
# ResNet

## *Profiling*



```
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                                   Name    Self CPU %      Self CPU   CPU total %     CPU total  CPU time avg     Self CUDA   Self CUDA %    CUDA total  CUDA time avg    # of Calls
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                          ProfilerStep*         3.52%      82.385ms        75.69%        1.774s     591.263ms       0.000us         0.00%     574.992ms     191.664ms             3
enumerate(DataLoader)#_SingleProcessDataLoaderIter._...        41.05%     962.092ms        47.79%        1.120s     373.287ms       0.000us         0.00%       0.000us       0.000us             3
void at::native::vectorized_elementwise_kernel<4, at...         0.00%       0.000us         0.00%       0.000us       0.000us     707.000us         0.03%     707.000us       3.432us           206
void cudnn::bn_fw_tr_1C11_kernel_NCHW<float, float, ...         0.00%       0.000us         0.00%       0.000us       0.000us      63.942ms         2.80%      63.942ms       1.640ms            39
void at::native::vectorized_elementwise_kernel<4, at...         0.00%       0.000us         0.00%       0.000us       0.000us      76.858ms         3.37%      76.858ms     402.398us           191
                                   volta_sgemm_128x64_nn        0.00%       0.000us         0.00%       0.000us       0.000us      87.020ms         3.81%      87.020ms       1.116ms            78
                                            aten::empty         0.50%      11.663ms         0.50%      11.663ms       3.967us       0.000us         0.00%       0.000us       0.000us          2940
                                         aten::uniform_         0.10%       2.232ms         0.10%       2.232ms       5.812us       0.000us         0.00%       0.000us       0.000us           384
                                             aten::item         0.09%       2.083ms         0.10%       2.368ms       1.229us       0.000us         0.00%       0.000us       0.000us          1926
                              aten::_local_scalar_dense         0.02%     367.000us         0.02%     367.000us       0.191us       0.000us         0.00%       0.000us       0.000us          1926
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
Self CPU time total: 2.344s
Self CUDA time total: 2.282s
```

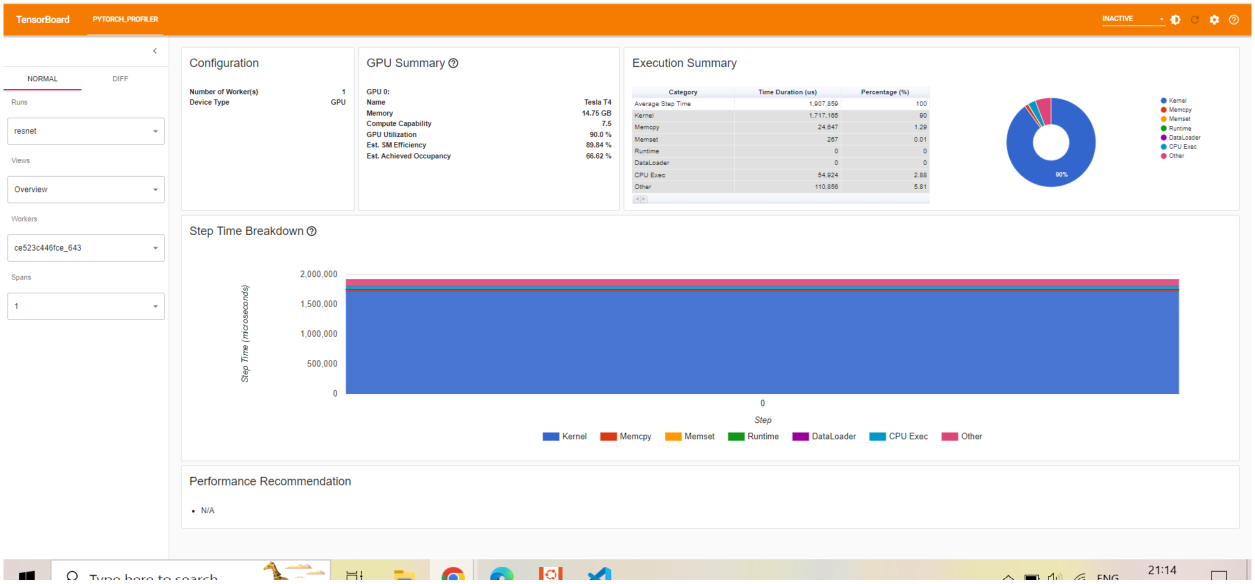
Fig 17: CPU memory usage and time for ResNet



```
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                                   Name    Self CPU %      Self CPU   CPU total %     CPU total  CPU time avg     Self CUDA   Self CUDA %    CUDA total  CUDA time avg    # of Calls
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                aten::convolution_backward         0.59%      13.830ms         0.90%      21.001ms     132.082us     811.943ms        35.58%     811.943ms       5.107ms           159
                                  aten::cudnn_convolution         0.34%       7.899ms         0.46%      10.698ms      67.283us     285.570ms        12.51%     285.570ms       1.796ms           159
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
Self CPU time total: 2.344s
Self CUDA time total: 2.282s
```

Fig 18: CUDA memory usage and time for ResNet

## *Tensorboard Visualizations*



Runs: resnet

Views: Overview

Workers: ce523c446fce_643

Spans: 1

Configuration

| | |
|---|---|
| Number of Worker(s) | 1 |
| Device Type | GPU |

GPU Summary

| GPU 0: | |
|---|---|
| Name | Tesla T4 |
| Memory | 14.75 GB |
| Compute Capability | 7.5 |
| GPU Utilization | 90.0 % |
| Est. SM Efficiency | 89.84 % |
| Est. Achieved Occupancy | 66.62 % |

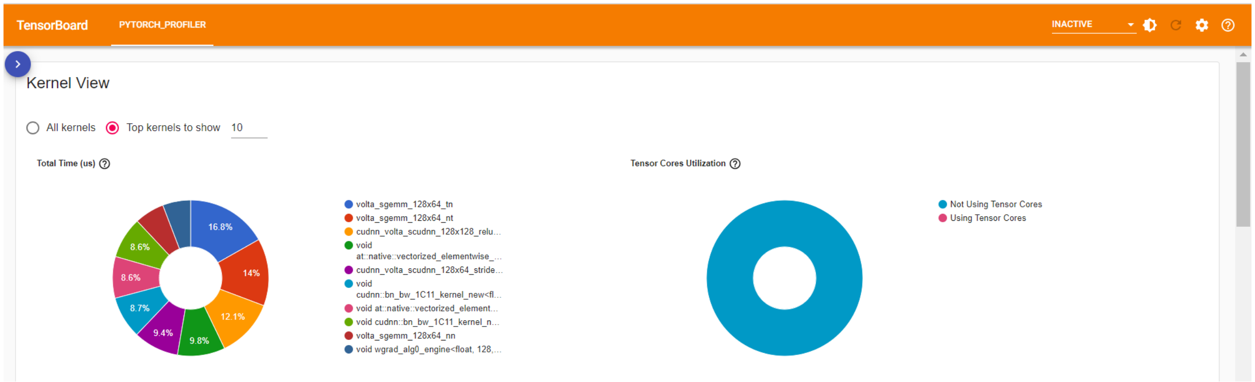
Execution Summary

| Category | Time Duration (us) | Percentage (%) |
|---|---|---|
| Average Step Time | 1,907,859 | 100 |
| Kernel | 1,717,165 | 90 |
| Memcpy | 24,647 | 1.29 |
| Memset | 267 | 0.01 |
| Runtime | 0 | 0 |
| DataLoader | 0 | 0 |
| CPU Exec | 54,924 | 2.88 |
| Other | 110,856 | 5.81 |

Step Time Breakdown

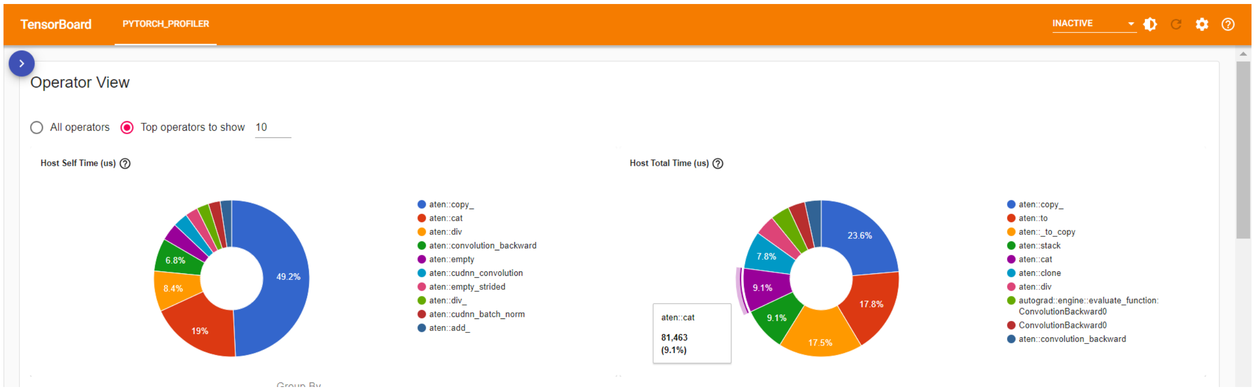
Performance Recommendation

- N/A


TensorBoard
PYTORCH_PROFILER
INACTIVE
Total Time (us)
16.8%
14%
12.1%
9.8%
9.4%
8.7%
8.6%
8.6%
volta_sgemm_128x64_tn
volta_sgemm_128x64_nt
cudnn_volta_scudnn_128x128_relu…
void at::native::vectorized_elementwise_…
cudnn_volta_scudnn_128x64_stride…
void cudnn::bn_bw_1C11_kernel_new<fl…
void at::native::vectorized_element…
void cudnn::bn_bw_1C11_kernel_n…
volta_sgemm_128x64_nn
void wgrad_alg0_engine<float, 128,…
Tensor Cores Utilization
Not Using Tensor Cores
Using Tensor Cores

Kernel View

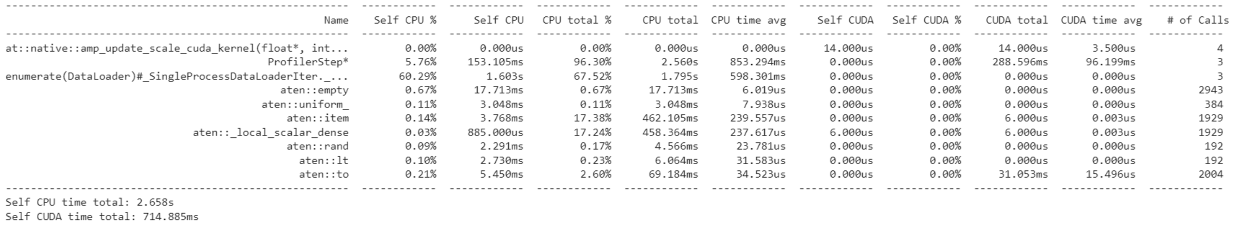
○ All kernels ● Top kernels to show 10



Operator View

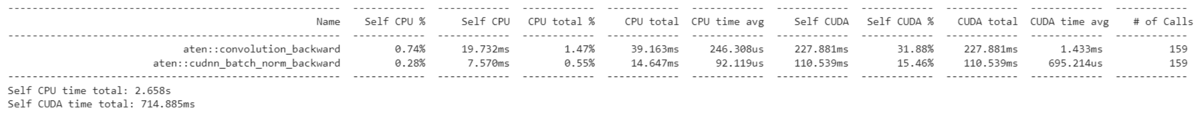
○ All operators ● Top operators to show 10

## *Profiling with Automatic Mixed Precision*



```
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                                   Name    Self CPU %      Self CPU   CPU total %     CPU total  CPU time avg     Self CUDA   Self CUDA %    CUDA total  CUDA time avg    # of Calls
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
at::native::amp_update_scale_cuda_kernel(float*, int...         0.00%       0.000us         0.00%       0.000us       0.000us      14.000us         0.00%      14.000us       3.500us             4
                                          ProfilerStep*         5.76%     153.105ms        96.30%       2.560s     853.294ms       0.000us         0.00%     288.596ms      96.199ms             3
enumerate(DataLoader)#_SingleProcessDataLoaderIter._...        60.29%       1.603s        67.52%       1.795s     598.301ms       0.000us         0.00%       0.000us       0.000us             3
                                            aten::empty         0.67%      17.713ms         0.67%      17.713ms       6.019us       0.000us         0.00%       0.000us       0.000us          2943
                                         aten::uniform_         0.11%       3.048ms         0.11%       3.048ms       7.938us       0.000us         0.00%       0.000us       0.000us           384
                                             aten::item         0.14%       3.768ms        17.38%     462.105ms     239.557us       0.000us         0.00%       6.000us       0.003us          1929
                              aten::_local_scalar_dense         0.03%     885.000us        17.24%     458.364ms     237.617us       6.000us         0.00%       6.000us       0.003us          1929
                                             aten::rand         0.09%       2.291ms         0.17%       4.566ms      23.781us       0.000us         0.00%       0.000us       0.000us           192
                                               aten::lt         0.10%       2.730ms         0.23%       6.064ms      31.583us       0.000us         0.00%       0.000us       0.000us           192
                                               aten::to         0.21%       5.450ms         2.60%      69.184ms      34.523us       0.000us         0.00%      31.053ms      15.496us          2004
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
Self CPU time total: 2.658s
Self CUDA time total: 714.885ms
```

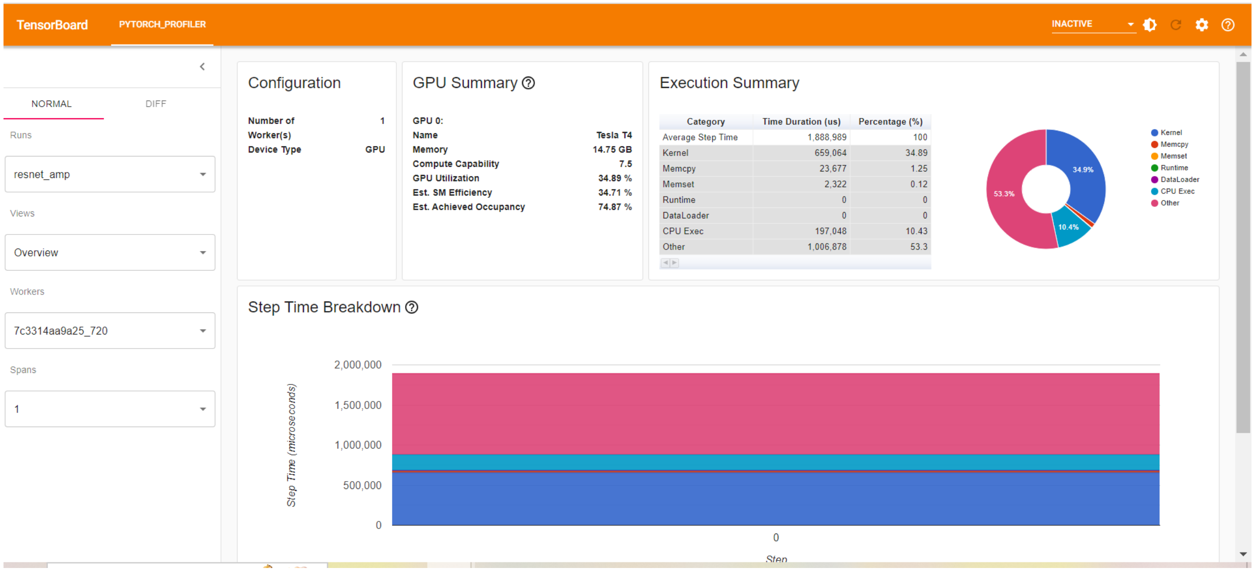
Fig 19: CPU memory usage and time for ResNet with automatic mixed precision



```
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                                                   Name    Self CPU %      Self CPU   CPU total %     CPU total  CPU time avg     Self CUDA   Self CUDA %    CUDA total  CUDA time avg    # of Calls
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
                               aten::convolution_backward         0.74%      19.732ms         1.47%      39.163ms     246.308us     227.881ms        31.88%     227.881ms       1.433ms           159
                          aten::cudnn_batch_norm_backward         0.28%       7.570ms         0.55%      14.647ms      92.119us     110.539ms        15.46%     110.539ms     695.214us           159
-------------------------------------------------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------  ------------
Self CPU time total: 2.658s
Self CUDA time total: 714.885ms
```

Fig 20: CUDA memory usage and time for ResNet with automatic mixed precision

## *Tensorboard Visualizations*



Runs: resnet_amp

Views: Overview

Workers: 7c3314aa9a25_720

Spans: 1

Configuration

| | |
|---|---|
| Number of Worker(s) | 1 |
| Device Type | GPU |

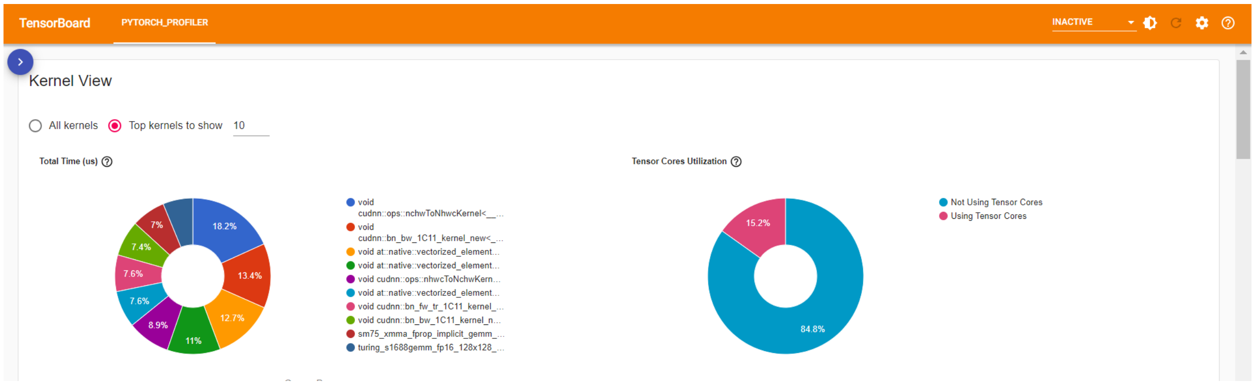
GPU Summary

| GPU 0: | |
|---|---|
| Name | Tesla T4 |
| Memory | 14.75 GB |
| Compute Capability | 7.5 |
| GPU Utilization | 34.89 % |
| Est. SM Efficiency | 34.71 % |
| Est. Achieved Occupancy | 74.87 % |

Execution Summary

| Category | Time Duration (us) | Percentage (%) |
|---|---|---|
| Average Step Time | 1,888,989 | 100 |
| Kernel | 659,064 | 34.89 |
| Memcpy | 23,677 | 1.25 |
| Memset | 2,322 | 0.12 |
| Runtime | 0 | 0 |
| DataLoader | 0 | 0 |
| CPU Exec | 197,048 | 10.43 |
| Other | 1,006,878 | 53.3 |

34.9%
10.4%
53.3%
Kernel
Memcpy
Memset
Runtime
DataLoader
CPU Exec
Other
Step Time Breakdown
2,000,000
1,500,000
1,000,000
500,000
0
Step Time (microseconds)
0
Step

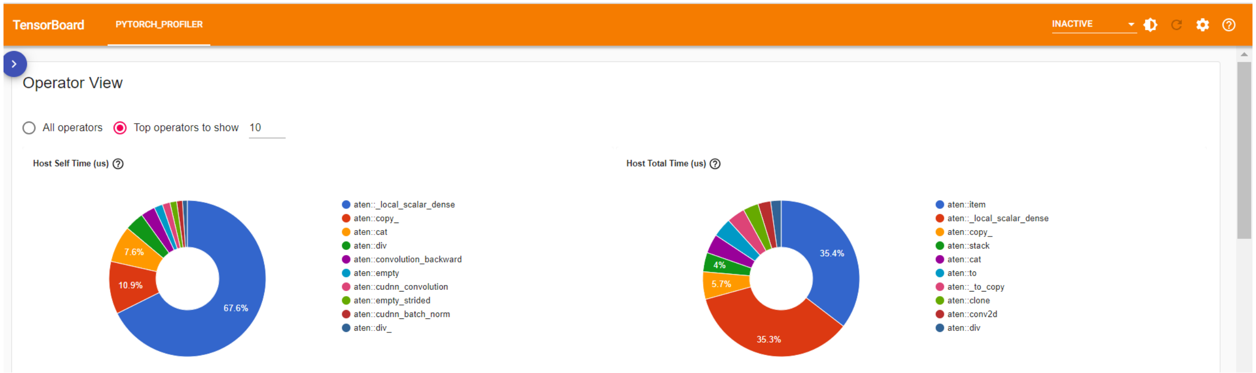


TensorBoard
PYTORCH_PROFILER
INACTIVE

Kernel View

○ All kernels ● Top kernels to show 10

Total Time (us)
18.2%
13.4%
12.7%
11%
8.9%
7.6%
7.6%
7.4%
7%
void cudnn::ops::nchwToNhwcKernel<__...
void cudnn::bn_bw_1C11_kernel_new<_...
void at::native::vectorized_element...
void at::native::vectorized_element...
void cudnn::ops::nhwcToNchwKern...
void at::native::vectorized_element...
void cudnn::bn_fw_tr_1C11_kernel_...
void cudnn::bn_bw_1C11_kernel_n...
sm75_xmma_fprop_implicit_gemm_...
turing_s1688gemm_fp16_128x128_...
Tensor Cores Utilization
15.2%
84.8%
Not Using Tensor Cores
Using Tensor Cores



Operator View

○ All operators ● Top operators to show 10

## ***Observation from the above images:***

## *MLP:*

*The CPU memory usage and time using MLP is the same as the CUDA memory usage and time using MLP, only without the APEX library. It consumes 14.75GB of memory with an estimated achieved occupancy of around 8.5%. Regarding the execution summary, the CPU execution takes around 24.93% of the time. The other is taking 73.15% of the time using Tesla T4 and not using tensor cores.*
*With APEX,  The CPU memory usage and time using MLP increased by approximately 3ms. We can observe that it is using a Tesla T4. Regarding GPU usage, Memory is 14.75GB with an achieved occupancy of 10.24%. When it comes to the execution, the CPU execution takes 26.2% of the total execution time, 2% by the kernel, and the remaining by the others. And it is using both tensor cores and non-tensor cores.*

## *LeNet:*

*The CPU memory usage and time for Lenet are the same as the CUDA memory usage and time for Lenet, only without the APEX. It consumes 14.75GB of memory usage with 18.09% Est Achieved occupancy, and when it comes to execution time, CPU execution takes around 26.5% of the total time. The kernel is also taking around 2% of the total time. It uses a Tesla T4 and does not use tensor cores at all.*
*With APEX:*
*The time for the execution for lenet with APEX took 10ms less when compared to without APEX.*
*Regarding the GPU summary, it is using Tesla T4 and consuming 14.75GB memory with an estimated achieved Occupancy of 19.33%, and when it comes to the Execution time, CPU execution is taking 28.4%. The kernel is taking around 4% of the total execution time. It is not using tensor cores at all.*

## *Alex:*

*The CPU memory usage and time for Alex are the same as the CUDA memory and usage and time for Alex, only without the APEX. It consumes 14.75GB of memory usage with 69.07% Est achieved occupancy when it comes to execution time, CPU execution is 19.4% of the total time kernel execution takes around 29.3% of the total execution time and it does not use Tensor cores at all.*
*With APEX:*
*The time for the execution for Alex with APEX took 16ms more when compared to the execution without APEX.*
*Regarding the GPU summary, it is using Tesla T4 and 14.75GB of total memory with an Est Achieved occupancy of 81.61%, and when it comes to the execution time, Kernel is taking 32% of the total execution time, and CPU execution time is 20% of the total time and it is using both tensor cores and nontensor cores.*

## *VGG:*

*The CPU memory usage and time for VGG are the same as CUDA memory usage and time for VGG without APEX.*
*Regarding GPU summary, it uses Tesla T4, with 14.75GB memory with an estimated achieved occupancy of 74.87%.*
*Regarding execution time, kernel execution takes 98.3% of the total execution time and does not use tensor cores.*
*With APEX:*
*The CPU memory usage and time for VGG with APEX took more time when compared to without APEX usage.*
*Regarding GPU summary, it occupies 14.75GB of memory with an estimated achieved occupancy of 78.05%.*
*Regarding the execution time, the kernel takes 74% of the total execution and 12 and 11 % of the total estimation time by other and CPU execution time. It is using both tensor cores and nontensor cores.*

## *ResNet:*

*The CPU memory usage and time for ResNet are the same as CUDA memory usage and time for ResNet without APEX. When it comes to the GPU summary, it is using Tesla T4 with 14.75GB of memory and 66.62% Est Achieved occupancy, and when it comes to the execution time, 90% of the total time is with the Kernel execution, and 7% memory execution*
*With APEX:*
*The CPU memory usage and time for Resnet is less when compared to the CPU memory usage of Resnet without APEX. Regarding the GPU summary, it uses Tesla 14 with 14.75GB of memory with 78.05% Est Achieved occupancy. Regarding execution time, kernel execution takes 74.5% of the total execution time, and 12.3% of the total execution time is taken at CPU execution.*