Инструменты для профайлинга

Для профайлинга модели в `profiling.py` сделано несколько утилит:

- **ThroughputMeter** — счётчик пропускной способности (samples/sec).
- **make_profiler** — удобное создание профайлера.
- **print_top_tables** — удобный вывод топ-операторов по CPU/GPU времени.
- **export_trace** — сохранение трейсов
- **marked** — для собственных аннотаций для наглядности в трейсах.

Результаты профайлинга

Вот краткие результаты профайлинга базовой модели:

- [Throughput during profiled window] ~4.9 samples/sec
- Self CPU time total: 29.238s (время всего оверхеда на CPU)
- Self CUDA time total: 28.157s (время GPU затрачено ядрами)
  Основные проблемы:
- **aten::mul** — занимает ~64% времени на GPU. Вызвано раздуванием тензоров и дальнейшей работе с ними.
- **aten::copy_** — большое количество копирований тензоров из-за использования repeat и permute.

Логи

```
[W927 17:55:25.665631335 CPUAllocator.cpp:245] Memory block of unknown size was allocated

==== TOP CUDA ops (self_cuda_time_total) ====
-------------------------------------------------------  ------------  ------------  -----
                                                   Name    Self CPU %      Self CPU   CPU
-------------------------------------------------------  ------------  ------------  -----
                                                forward         0.00%       0.000us
void at::native::elementwise_kernel<128, 2, at::nati...         0.00%       0.000us
                                                aten::mul         0.16%      46.703ms
void at::native::elementwise_kernel<128, 2, at::nati...         0.00%       0.000us
                                               aten::copy_         0.08%      22.872ms
                                                 aten::sum         0.29%      84.059ms
                                               aten::addmm         0.26%      75.485ms
void at::native::reduce_kernel<128, 4, at::native::R...         0.00%       0.000us
                                    volta_sgemm_64x64_nn         0.00%       0.000us
                                                  aten::mm         0.25%      74.071ms
                                                aten::add         0.11%      33.205ms
void at::native::elementwise_kernel<128, 2, at::nati...         0.00%       0.000us
std::enable_if<!(false), void>::type internal::gemvx...         0.00%       0.000us
void gemv2N_kernel<int, int, float, float, float, fl...         0.00%       0.000us
void at::native::reduce_kernel<512, 1, at::native::R...         0.00%       0.000us
                        aten::native_layer_norm_backward         0.09%      26.662ms
                                                aten::bmm         0.08%      23.627ms
                                   aten::native_layer_norm         0.11%      33.520ms
void at::native::(anonymous namespace)::vectorized_l...         0.00%       0.000us
                                 Optimizer.step#AdamW.step         0.00%       0.000us
void at::native::(anonymous namespace)::GammaBetaBac...         0.00%       0.000us
void at::native::(anonymous namespace)::layer_norm_g...         0.00%       0.000us
                                    volta_sgemm_128x64_tn         0.00%       0.000us
                                     volta_sgemm_64x64_nt         0.00%       0.000us
void at::native::vectorized_elementwise_kernel<4, at...         0.00%       0.000us
void at::native::vectorized_elementwise_kernel<4, at...         0.00%       0.000us
                              aten::_softmax_backward_data         0.04%      12.866ms
void (anonymous namespace)::softmax_warp_backward<fl...         0.00%       0.000us
void at::native::vectorized_elementwise_kernel<4, at...         0.00%       0.000us
                                   aten::threshold_backward         0.04%      12.322ms
                                                   loss         0.00%       0.000us
                                          aten::clamp_min         0.04%      11.191ms
void at::native::vectorized_elementwise_kernel<4, at...         0.00%       0.000us
```

```
                                                aten::_softmax         0.04%      10.807ms
void at::native::(anonymous namespace)::multi_tensor...         0.00%       0.000us
void (anonymous namespace)::softmax_warp_forward<flo...         0.00%       0.000us
                                     volta_sgemm_32x128_tn         0.00%       0.000us
                                   aten::_foreach_addcdiv_         0.02%       6.749ms
void at::native::(anonymous namespace)::multi_tensor...         0.00%       0.000us
                                      aten::_foreach_mul_         0.04%      12.943ms
-------------------------------------------------------  ------------  ------------  -----
Self CPU time total: 29.238s
Self CUDA time total: 28.157s


==== TOP CPU ops (self_cpu_time_total) ====
-------------------------------------------------------  ------------  ------------  -----
                                                   Name    Self CPU %      Self CPU   CPU
-------------------------------------------------------  ------------  ------------  -----
                                     cudaStreamSynchronize        89.64%      26.210s
                                                  backward         3.69%       1.080s
                                       cudaLaunchKernel         0.78%     228.195ms
                                                   forward         0.53%     155.733ms
                                     cudaDeviceSynchronize         0.51%     149.004ms
                                Optimizer.step#AdamW.step         0.37%     108.590ms
                                                aten::sum         0.29%      84.059ms
                                              aten::addmm         0.26%      75.485ms
                                                 aten::mm         0.25%      74.071ms
                                              aten::empty         0.22%      63.355ms
                                                aten::mul         0.16%      46.703ms
    autograd::engine::evaluate_function: AddmmBackward0         0.13%      37.272ms
                                   aten::native_layer_norm         0.11%      33.520ms
                                                aten::add         0.11%      33.205ms
                                             aten::transpose         0.11%      31.507ms
                                           aten::as_strided         0.10%      28.479ms
                                                    aten::t         0.09%      27.529ms
                                      aten::empty_strided         0.09%      27.072ms
                          aten::native_layer_norm_backward         0.09%      26.662ms
autograd::engine::evaluate_function: NativeLayerNorm...         0.09%      26.147ms
                                               aten::view         0.09%      25.462ms
autograd::engine::evaluate_function: torch::autograd...         0.08%      24.691ms
                                                aten::bmm         0.08%      23.627ms
                                              aten::copy_         0.08%      22.872ms
```

```
                                                  aten::add_         0.08%      22.234ms
                                               aten::unsqueeze         0.08%      22.197ms
                                                AddmmBackward0         0.07%      21.669ms
                                                          loss         0.06%      17.851ms
      autograd::engine::evaluate_function: MulBackward0         0.06%      16.480ms
autograd::engine::evaluate_function: UnsqueezeBackwa...         0.05%      15.321ms
     autograd::engine::evaluate_function: ViewBackward0         0.05%      14.745ms
                                               aten::squeeze         0.05%      14.674ms
                                               aten::reshape         0.05%      14.235ms
                                                aten::repeat         0.05%      13.935ms
        autograd::engine::evaluate_function: TBackward0         0.05%      13.912ms
                          torch::autograd::AccumulateGrad         0.05%      13.719ms
                                               aten::linear         0.05%      13.417ms
                                               aten::expand         0.04%      13.014ms
                                     aten::_foreach_add_         0.04%      12.948ms
                                     aten::_foreach_mul_         0.04%      12.943ms
-------------------------------------------------------  ------------  ------------  -----
Self CPU time total: 29.238s
Self CUDA time total: 28.157s


==== TOP CUDA ops (group_by_input_shape) ====
-------------------------------------------------------  ------------  ------------  -----
                                                   Name    Self CPU %      Self CPU   CPU
-------------------------------------------------------  ------------  ------------  -----
                                                forward         0.00%       0.000us
void at::native::elementwise_kernel<128, 2, at::nati...         0.00%       0.000us
                                                  aten::mul         0.02%       5.569ms
void at::native::elementwise_kernel<128, 2, at::nati...         0.00%       0.000us
                                                 aten::copy_         0.01%       3.378ms
                                                 aten::addmm         0.07%      21.310ms
void at::native::reduce_kernel<128, 4, at::native::R...         0.00%       0.000us
                                                  aten::mul         0.01%       3.563ms
                                                  aten::sum         0.06%      16.592ms
                                       volta_sgemm_64x64_nn         0.00%       0.000us
void at::native::elementwise_kernel<128, 2, at::nati...         0.00%       0.000us
                                                  aten::add         0.02%       6.431ms
std::enable_if<!(false), void>::type internal::gemvx...         0.00%       0.000us
                                                   aten::mm         0.03%       8.479ms
                                                  aten::mul         0.02%       5.192ms
```

```
void gemv2N_kernel<int, int, float, float, float, fl...         0.00%       0.000us         0
                                                 aten::mm         0.02%       5.823ms
void at::native::reduce_kernel<512, 1, at::native::R...         0.00%       0.000us
                                                aten::sum         0.03%       7.490ms
                        aten::native_layer_norm_backward         0.04%      11.657ms
void at::native::(anonymous namespace)::vectorized_l...         0.00%       0.000us
                                  aten::native_layer_norm         0.06%      17.221ms
                                Optimizer.step#AdamW.step         0.00%       0.000us
void at::native::(anonymous namespace)::GammaBetaBac...         0.00%       0.000us
void at::native::(anonymous namespace)::layer_norm_g...         0.00%       0.000us
                                                aten::bmm         0.02%       5.182ms
                                                aten::bmm         0.06%      18.445ms
                                     volta_sgemm_128x64_tn         0.00%       0.000us
                                      volta_sgemm_64x64_nt         0.00%       0.000us
void at::native::vectorized_elementwise_kernel<4, at...         0.00%       0.000us
void at::native::vectorized_elementwise_kernel<4, at...         0.00%       0.000us
                                                aten::mul         0.01%       3.436ms
void (anonymous namespace)::softmax_warp_backward<fl...         0.00%       0.000us
                               aten::_softmax_backward_data         0.02%       4.776ms
                                                  aten::copy_         0.01%       4.127ms
                                                aten::sum         0.03%       9.170ms
void at::native::vectorized_elementwise_kernel<4, at...         0.00%       0.000us
                                                     loss         0.00%       0.000us
                                  aten::threshold_backward         0.02%       4.648ms
                                                aten::add         0.02%       4.806ms
-------------------------------------------------------  ------------  ------------  -----
Self CPU time total: 29.238s
Self CUDA time total: 28.157s

Не удалось сохранить trace: Trace is already saved.

[Throughput during profiled window] ~4.9 samples/sec
```

Изменения в модели

Все изменения и комментарии к ним есть в файле `model.py`

Результат после изменений внутри модели

Стало сильно лучше, время загрузки GPU и CPU заметно сократилось.

- [Throughput during profiled window] ~7.8 samples/sec
- Self CPU time total: 2.541s
- Self CUDA time total: 506.149ms

Логи

```
[W930 17:08:15.984771851 CPUAllocator.cpp:245] Memory block of unknown size was allocated

==== TOP CUDA ops (self_cuda_time_total) ====
-------------------------------------------------------  ------------  ------------  -----
                                                   Name    Self CPU %      Self CPU   CPU
-------------------------------------------------------  ------------  ------------  -----
                                             forward         0.00%       0.000us
                                           aten::bmm         1.79%      45.371ms
                                Optimizer.step#AdamW.step         0.00%       0.000us
                                     volta_sgemm_64x64_nn         0.00%       0.000us
                          aten::native_layer_norm_backward         0.87%      22.090ms
                                           aten::sum         2.54%      64.434ms
                                  volta_sgemm_128x64_tn         0.00%       0.000us
                                           aten::mul         1.74%      44.341ms
                                   volta_sgemm_64x64_nt         0.00%       0.000us
                                 aten::native_layer_norm         1.29%      32.908ms
void at::native::(anonymous namespace)::vectorized_l...         0.00%       0.000us
void at::native::(anonymous namespace)::layer_norm_g...         0.00%       0.000us
void at::native::reduce_kernel<128, 4, at::native::R...         0.00%       0.000us
void at::native::elementwise_kernel<128, 2, at::nati...         0.00%       0.000us
void at::native::reduce_kernel<512, 1, at::native::R...         0.00%       0.000us
void at::native::(anonymous namespace)::GammaBetaBac...         0.00%       0.000us
                                            aten::mm         2.24%      56.952ms
void at::native::vectorized_elementwise_kernel<4, at...         0.00%       0.000us
                                           aten::add         1.19%      30.333ms
void at::native::elementwise_kernel<128, 2, at::nati...         0.00%       0.000us
                             aten::_softmax_backward_data         0.38%       9.605ms
void (anonymous namespace)::softmax_warp_backward<fl...         0.00%       0.000us
                                        aten::addmm         2.12%      53.949ms
                                aten::threshold_backward         0.35%       8.801ms
void at::native::vectorized_elementwise_kernel<4, at...         0.00%       0.000us
                                      aten::_softmax         0.54%      13.667ms
void (anonymous namespace)::softmax_warp_forward<flo...         0.00%       0.000us
                                     aten::clamp_min         0.31%       7.923ms
void at::native::vectorized_elementwise_kernel<4, at...         0.00%       0.000us
                                  volta_sgemm_32x128_tn         0.00%       0.000us
                                aten::_foreach_addcdiv_         0.25%       6.390ms
void at::native::(anonymous namespace)::multi_tensor...         0.00%       0.000us
                                   aten::_foreach_mul_         0.47%      11.985ms
```

```
void at::native::(anonymous namespace)::multi_tensor...         0.00%       0.000us
                                    volta_sgemm_128x64_nt         0.00%       0.000us
                                     aten::_foreach_lerp_         0.06%       1.461ms
void at::native::(anonymous namespace)::multi_tensor...         0.00%       0.000us
                                  aten::_foreach_addcmul_         0.26%       6.552ms
void at::native::(anonymous namespace)::multi_tensor...         0.00%       0.000us
                                    volta_sgemm_128x64_nn         0.00%       0.000us
-------------------------------------------------------  ------------  ------------  -----
Self CPU time total: 2.541s
Self CUDA time total: 506.149ms


==== TOP CPU ops (self_cpu_time_total) ====
-------------------------------------------------------  ------------  ------------  -----
                                                   Name    Self CPU %      Self CPU   CPU
-------------------------------------------------------  ------------  ------------  -----
                                             backward        37.41%     950.695ms
                                        cudaLaunchKernel         8.05%     204.544ms
                                              forward         5.88%     149.381ms
                                 Optimizer.step#AdamW.step         4.06%     103.253ms
                                               aten::sum         2.54%      64.434ms
                                                aten::mm         2.24%      56.952ms
                                             aten::empty         2.21%      56.080ms
                                             aten::addmm         2.12%      53.949ms
                                               aten::bmm         1.79%      45.371ms
                                               aten::mul         1.74%      44.341ms
                                  aten::native_layer_norm         1.29%      32.908ms
                                               aten::add         1.19%      30.333ms
    autograd::engine::evaluate_function: AddmmBackward0         1.19%      30.279ms
                                           aten::transpose         1.15%      29.225ms
                                      aten::empty_strided         1.05%      26.626ms
                                              aten::add_         0.96%      24.443ms
                                              aten::view         0.93%      23.665ms
autograd::engine::evaluate_function: torch::autograd...         0.91%      23.226ms
                         aten::native_layer_norm_backward         0.87%      22.090ms
                                                 aten::t         0.85%      21.643ms
autograd::engine::evaluate_function: NativeLayerNorm...         0.84%      21.410ms
                                           aten::as_strided         0.82%      20.749ms
      autograd::engine::evaluate_function: MulBackward0         0.69%      17.455ms
                                           AddmmBackward0         0.64%      16.319ms
```

```
                                          aten::unsqueeze         0.62%      15.697ms
      autograd::engine::evaluate_function: AddBackward0         0.55%      13.903ms
                                           aten::_softmax         0.54%      13.667ms
     autograd::engine::evaluate_function: ViewBackward0         0.53%      13.373ms
                                           aten::squeeze         0.51%      13.008ms
                         torch::autograd::AccumulateGrad         0.51%      12.913ms
                                                 detach         0.49%      12.514ms
                                     aten::_foreach_add_         0.48%      12.306ms
                                     aten::_foreach_mul_         0.47%      11.985ms
                                     aten::_foreach_sqrt         0.45%      11.465ms
                                                    loss         0.45%      11.379ms
                                            aten::linear         0.44%      11.249ms
        autograd::engine::evaluate_function: TBackward0         0.44%      11.233ms
                                           aten::expand         0.42%      10.790ms
                                          aten::reshape         0.42%      10.674ms
                                            aten::copy_         0.42%      10.671ms
-------------------------------------------------------  ------------  ------------  -----
Self CPU time total: 2.541s
Self CUDA time total: 506.149ms


==== TOP CUDA ops (group_by_input_shape) ====
-------------------------------------------------------  ------------  ------------  -----
                                                  Name    Self CPU %      Self CPU   CPU
-------------------------------------------------------  ------------  ------------  -----
                                               forward         0.00%       0.000us
                             Optimizer.step#AdamW.step         0.00%       0.000us
                                           aten::bmm         0.47%      11.924ms
                                volta_sgemm_64x64_nn         0.00%       0.000us
                                           aten::bmm         0.98%      24.781ms
                               volta_sgemm_128x64_tn         0.00%       0.000us
                                volta_sgemm_64x64_nt         0.00%       0.000us
                      aten::native_layer_norm_backward         0.19%       4.763ms
void at::native::(anonymous namespace)::vectorized_l...         0.00%       0.000us
                                aten::native_layer_norm         0.34%       8.547ms
                                           aten::sum         0.57%      14.450ms
void at::native::(anonymous namespace)::layer_norm_g...         0.00%       0.000us
void at::native::reduce_kernel<128, 4, at::native::R...         0.00%       0.000us
void at::native::elementwise_kernel<128, 2, at::nati...         0.00%       0.000us
void at::native::reduce_kernel<512, 1, at::native::R...         0.00%       0.000us
```

```
void at::native::(anonymous namespace)::GammaBetaBac...         0.00%       0.000us
                                                aten::bmm         0.34%       8.666ms
void at::native::vectorized_elementwise_kernel<4, at...         0.00%       0.000us
void at::native::elementwise_kernel<128, 2, at::nati...         0.00%       0.000us
                                                aten::mul         0.15%       3.843ms
                            aten::_softmax_backward_data         0.19%       4.854ms
void (anonymous namespace)::softmax_warp_backward<fl...         0.00%       0.000us
void at::native::vectorized_elementwise_kernel<4, at...         0.00%       0.000us
                                                aten::sum         1.02%      25.864ms
                                aten::threshold_backward         0.18%       4.468ms
                                                aten::add         0.20%       5.173ms
                                             aten::_softmax         0.20%       5.065ms
void (anonymous namespace)::softmax_warp_forward<flo...         0.00%       0.000us
void at::native::vectorized_elementwise_kernel<4, at...         0.00%       0.000us
                                           aten::clamp_min         0.14%       3.592ms
                                               aten::addmm         0.49%      12.576ms
                                  volta_sgemm_32x128_tn         0.00%       0.000us
                                                aten::mul         0.17%       4.300ms
                                aten::_foreach_addcdiv_         0.25%       6.390ms
void at::native::(anonymous namespace)::multi_tensor...         0.00%       0.000us
                                     aten::_foreach_mul_         0.47%      11.985ms
void at::native::(anonymous namespace)::multi_tensor...         0.00%       0.000us
                                                aten::mm         0.35%       8.870ms
                                  volta_sgemm_128x64_nt         0.00%       0.000us
                                   aten::_foreach_lerp_         0.06%       1.461ms
-------------------------------------------------------  ------------  ------------  -----
Self CPU time total: 2.541s
Self CUDA time total: 506.149ms

Не удалось сохранить trace: Trace is already saved.

[Throughput during profiled window] ~7.8 samples/sec
```

Настройка даталоадера

- batch_size: 1024
- num_workers: 2

- pin_memory: True
- отключил профайлинг для честной скорости

В результате **[Throughput epoch] ~964.4 samples/sec**

torch.compile

Использование torch.compile дало [**Throughput epoch] ~1326.6 samples/sec**

```
cuda:0
  0%|                                                                | 0/156 [00:00<?, ?it/s]W0930 18:59:33.080000
9025 torch/_inductor/utils.py:1137] [0/0] Not enough SMs to use max_autotune_gemm mode
AUTOTUNE addmm(1024x128, 1024x128, 128x128)
  addmm 0.0348 ms 100.0%
  bias_addmm 0.0430 ms 80.9%
SingleProcess AUTOTUNE benchmarking takes 0.2580 seconds and 0.0003 seconds precompiling for 2 choice
s
100%|████████████████████████████████████████| 156/156 [04:24<00:00,  1.70s/it]

[Throughput epoch] ~1326.6 samples/sec
```