```
------------------------- DeepSpeed Flops Profiler -------------------------
Profile Summary at step 10:
Notations:
data parallel size (dp_size), model parallel size(mp_size),
number of parameters (params), number of multiply-accumulate operations(MACs),
number of floating-point operations (flops), floating-point operations per second
(FLOPS),
fwd latency (forward propagation latency), bwd latency (backward propagation
latency),
step (weights update latency), iter latency (sum of fwd, bwd and step latency)

params per gpu:                                                 86.39 M
params of model = params per GPU * mp_size:                     86.39 M
fwd MACs per GPU:                                               274.37 GMACs
fwd flops per GPU:                                              548.94 G
fwd flops of model = fwd flops per GPU * mp_size:               548.94 G
fwd latency:                                                    167.59 ms
fwd FLOPS per GPU = fwd flops per GPU / fwd latency:            3.28 TFLOPS

----------------------------- Aggregated Profile per GPU
-----------------------------
Top 1 modules in terms of params, MACs or fwd latency at different model depths:
depth 0:
    params      - {'ViTModel': '86.39 M'}
    MACs        - {'ViTModel': '274.37 GMACs'}
    fwd latency - {'ViTModel': '167.59 ms'}
depth 1:
    params      - {'ViTEncoder': '85.05 M'}
    MACs        - {'ViTEncoder': '273.44 GMACs'}
    fwd latency - {'ViTEncoder': '163.37 ms'}
depth 2:
    params      - {'ModuleList': '85.05 M'}
    MACs        - {'ModuleList': '273.44 GMACs'}
    fwd latency - {'ModuleList': '162.78 ms'}
depth 3:
    params      - {'ViTLayer': '85.05 M'}
    MACs        - {'ViTLayer': '273.44 GMACs'}
    fwd latency - {'ViTLayer': '162.78 ms'}
depth 4:
    params      - {'ViTAttention': '28.35 M'}
    MACs        - {'ViTAttention': '94.96 GMACs'}
    fwd latency - {'ViTAttention': '75.26 ms'}
depth 5:
    params      - {'Linear': '56.67 M'}
    MACs        - {'Linear': '178.48 GMACs'}
    fwd latency - {'ViTSelfAttention': '60.51 ms'}

------------------------------ Detailed Profile per GPU
------------------------------
```

```
Each module profile is listed after its name in the following order:
params, percentage of total params, MACs, percentage of total MACs, fwd latency,
percentage of total fwd latency, fwd FLOPS

Note: 1. A module can have torch.nn.module or torch.nn.functional to compute logits
(e.g. CrossEntropyLoss). They are not counted as submodules, thus not to be printed
out. However they make up the difference between a parent's MACs (or latency) and
the sum of its submodules'.
2. Number of floating-point operations is a theoretical estimation, thus FLOPS
computed using that could be larger than the maximum system throughput.
3. The fwd latency listed in the top module's profile is directly captured at the
module forward function in PyTorch, thus it's less than the fwd latency shown above
which is captured in DeepSpeed.

ViTModel(
  86.39 M, 100.00% Params, 274.37 GMACs, 100.00% MACs, 167.59 ms, 100.00% latency,
3.28 TFLOPS,
  (embeddings): ViTEmbeddings(
    742.66 k, 0.86% Params, 924.84 MMACs, 0.34% MACs, 1.08 ms, 0.64% latency, 1.72
TFLOPS,
    (patch_embeddings): ViTPatchEmbeddings(
      590.59 k, 0.68% Params, 924.84 MMACs, 0.34% MACs, 404.36 us, 0.24% latency,
4.58 TFLOPS,
      (projection): Conv2d(590.59 k, 0.68% Params, 924.84 MMACs, 0.34% MACs, 271.08
us, 0.16% latency, 6.83 TFLOPS, 3, 768, kernel_size=(16, 16), stride=(16, 16))
    )
    (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 56.27 us, 0.03% latency,
0.0 FLOPS, p=0.0, inplace=False)
  )
  (encoder): ViTEncoder(
    85.05 M, 98.45% Params, 273.44 GMACs, 99.66% MACs, 163.37 ms, 97.48% latency,
3.35 TFLOPS,
    (layer): ModuleList(
      (0): ViTLayer(
        7.09 M, 8.20% Params, 22.79 GMACs, 8.30% MACs, 12.21 ms, 7.28% latency, 3.73
TFLOPS,
        (attention): ViTAttention(
          2.36 M, 2.73% Params, 7.91 GMACs, 2.88% MACs, 5.55 ms, 3.31% latency, 2.85
TFLOPS,
          (attention): ViTSelfAttention(
            1.77 M, 2.05% Params, 6.05 GMACs, 2.21% MACs, 4.44 ms, 2.65% latency,
2.73 TFLOPS,
            (query): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 729.8
us, 0.44% latency, 5.09 TFLOPS, in_features=768, out_features=768, bias=True)
            (key): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 739.1 us,
0.44% latency, 5.03 TFLOPS, in_features=768, out_features=768, bias=True)
            (value): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 877.62
us, 0.52% latency, 4.24 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 56.98 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
```

```
          )
          (output): ViTSelfOutput(
            590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 986.1 us, 0.59% latency,
3.77 TFLOPS,
            (dense): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 837.56
us, 0.50% latency, 4.44 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 46.01 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 1.93 ms, 1.15% latency, 7.71
TFLOPS,
          (dense): Linear(2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 813.01 us,
0.49% latency, 18.29 TFLOPS, in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
1.0 ms, 0.60% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 1.26 ms, 0.75% latency,
11.85 TFLOPS,
          (dense): Linear(2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 875.71 us,
0.52% latency, 16.98 TFLOPS, in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 48.88 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.44
ms, 0.86% latency, 4.21 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.56
ms, 0.93% latency, 3.88 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
      )
      (1): ViTLayer(
        7.09 M, 8.20% Params, 22.79 GMACs, 8.30% MACs, 12.58 ms, 7.51% latency, 3.62
TFLOPS,
        (attention): ViTAttention(
          2.36 M, 2.73% Params, 7.91 GMACs, 2.88% MACs, 5.83 ms, 3.48% latency, 2.72
TFLOPS,
          (attention): ViTSelfAttention(
            1.77 M, 2.05% Params, 6.05 GMACs, 2.21% MACs, 4.69 ms, 2.80% latency,
2.58 TFLOPS,
            (query): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 816.35
us, 0.49% latency, 4.55 TFLOPS, in_features=768, out_features=768, bias=True)
            (key): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 895.26 us,
0.53% latency, 4.15 TFLOPS, in_features=768, out_features=768, bias=True)
            (value): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 910.52
us, 0.54% latency, 4.08 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 58.89 us, 0.04%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
```

```
            590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 1.03 ms, 0.61% latency,
3.62 TFLOPS,
            (dense): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 876.67
us, 0.52% latency, 4.24 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 46.73 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 1.91 ms, 1.14% latency, 7.78
TFLOPS,
          (dense): Linear(2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 847.34 us,
0.51% latency, 17.55 TFLOPS, in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
952.72 us, 0.57% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 1.24 ms, 0.74% latency,
11.97 TFLOPS,
          (dense): Linear(2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 845.19 us,
0.50% latency, 17.6 TFLOPS, in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 64.85 us, 0.04%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.49
ms, 0.89% latency, 4.06 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.66
ms, 0.99% latency, 3.64 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
      )
      (2): ViTLayer(
        7.09 M, 8.20% Params, 22.79 GMACs, 8.30% MACs, 12.87 ms, 7.68% latency, 3.54
TFLOPS,
        (attention): ViTAttention(
          2.36 M, 2.73% Params, 7.91 GMACs, 2.88% MACs, 5.66 ms, 3.38% latency, 2.8
TFLOPS,
          (attention): ViTSelfAttention(
            1.77 M, 2.05% Params, 6.05 GMACs, 2.21% MACs, 4.53 ms, 2.71% latency,
2.67 TFLOPS,
            (query): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 862.6
us, 0.51% latency, 4.31 TFLOPS, in_features=768, out_features=768, bias=True)
            (key): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 844.24 us,
0.50% latency, 4.4 TFLOPS, in_features=768, out_features=768, bias=True)
            (value): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 849.01
us, 0.51% latency, 4.38 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 56.27 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 961.07 us, 0.57%
latency, 3.87 TFLOPS,
```

```
            (dense): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 794.17
us, 0.47% latency, 4.68 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 46.25 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 2.13 ms, 1.27% latency, 6.99
TFLOPS,
          (dense): Linear(2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 931.26 us,
0.56% latency, 15.97 TFLOPS, in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
1.04 ms, 0.62% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 1.26 ms, 0.75% latency,
11.81 TFLOPS,
          (dense): Linear(2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 854.02 us,
0.51% latency, 17.42 TFLOPS, in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 51.5 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.63
ms, 0.97% latency, 3.7 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.66
ms, 0.99% latency, 3.64 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
      )
      (3): ViTLayer(
        7.09 M, 8.20% Params, 22.79 GMACs, 8.30% MACs, 13.34 ms, 7.96% latency, 3.42
TFLOPS,
        (attention): ViTAttention(
          2.36 M, 2.73% Params, 7.91 GMACs, 2.88% MACs, 6.25 ms, 3.73% latency, 2.53
TFLOPS,
          (attention): ViTSelfAttention(
            1.77 M, 2.05% Params, 6.05 GMACs, 2.21% MACs, 5.03 ms, 3.00% latency,
2.41 TFLOPS,
            (query): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 929.59
us, 0.55% latency, 4.0 TFLOPS, in_features=768, out_features=768, bias=True)
            (key): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 913.14 us,
0.54% latency, 4.07 TFLOPS, in_features=768, out_features=768, bias=True)
            (value): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 890.25
us, 0.53% latency, 4.18 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 61.04 us, 0.04%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 1.1 ms, 0.65% latency,
3.39 TFLOPS,
            (dense): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 912.9
us, 0.54% latency, 4.07 TFLOPS, in_features=768, out_features=768, bias=True)
```

```
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 63.42 us, 0.04%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 1.93 ms, 1.15% latency, 7.72
TFLOPS,
          (dense): Linear(2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 855.45 us,
0.51% latency, 17.39 TFLOPS, in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
926.73 us, 0.55% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 1.32 ms, 0.79% latency, 11.3
TFLOPS,
          (dense): Linear(2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 903.84 us,
0.54% latency, 16.46 TFLOPS, in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 52.69 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.71
ms, 1.02% latency, 3.53 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.65
ms, 0.99% latency, 3.66 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
      )
      (4): ViTLayer(
        7.09 M, 8.20% Params, 22.79 GMACs, 8.30% MACs, 13.3 ms, 7.93% latency, 3.43
TFLOPS,
        (attention): ViTAttention(
          2.36 M, 2.73% Params, 7.91 GMACs, 2.88% MACs, 6.11 ms, 3.65% latency, 2.59
TFLOPS,
          (attention): ViTSelfAttention(
            1.77 M, 2.05% Params, 6.05 GMACs, 2.21% MACs, 4.94 ms, 2.95% latency,
2.45 TFLOPS,
            (query): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 891.45
us, 0.53% latency, 4.17 TFLOPS, in_features=768, out_features=768, bias=True)
            (key): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 865.94 us,
0.52% latency, 4.29 TFLOPS, in_features=768, out_features=768, bias=True)
            (value): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 972.51
us, 0.58% latency, 3.82 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 61.51 us, 0.04%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 1.05 ms, 0.63% latency,
3.54 TFLOPS,
            (dense): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 882.63
us, 0.53% latency, 4.21 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 52.45 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
```

```
          )
        )
        (intermediate): ViTIntermediate(
          2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 1.97 ms, 1.18% latency, 7.55
TFLOPS,
          (dense): Linear(2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 895.98 us,
0.53% latency, 16.6 TFLOPS, in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
955.34 us, 0.57% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 1.34 ms, 0.80% latency,
11.06 TFLOPS,
          (dense): Linear(2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 909.57 us,
0.54% latency, 16.35 TFLOPS, in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 52.69 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.63
ms, 0.97% latency, 3.71 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.75
ms, 1.04% latency, 3.46 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
      )
      (5): ViTLayer(
        7.09 M, 8.20% Params, 22.79 GMACs, 8.30% MACs, 13.19 ms, 7.87% latency, 3.46
TFLOPS,
        (attention): ViTAttention(
          2.36 M, 2.73% Params, 7.91 GMACs, 2.88% MACs, 5.97 ms, 3.56% latency, 2.65
TFLOPS,
          (attention): ViTSelfAttention(
            1.77 M, 2.05% Params, 6.05 GMACs, 2.21% MACs, 4.75 ms, 2.83% latency,
2.55 TFLOPS,
            (query): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 847.82
us, 0.51% latency, 4.39 TFLOPS, in_features=768, out_features=768, bias=True)
            (key): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 822.78 us,
0.49% latency, 4.52 TFLOPS, in_features=768, out_features=768, bias=True)
            (value): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 874.28
us, 0.52% latency, 4.25 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 58.65 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 1.08 ms, 0.64% latency,
3.44 TFLOPS,
            (dense): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 913.14
us, 0.54% latency, 4.07 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 53.17 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
```

```
        (intermediate): ViTIntermediate(
          2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 2.09 ms, 1.25% latency, 7.1
TFLOPS,
          (dense): Linear(2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 897.88 us,
0.54% latency, 16.56 TFLOPS, in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
1.07 ms, 0.64% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 1.39 ms, 0.83% latency,
10.72 TFLOPS,
          (dense): Linear(2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 944.38 us,
0.56% latency, 15.75 TFLOPS, in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 54.12 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.58
ms, 0.94% latency, 3.84 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.66
ms, 0.99% latency, 3.64 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
      )
      (6): ViTLayer(
        7.09 M, 8.20% Params, 22.79 GMACs, 8.30% MACs, 13.17 ms, 7.86% latency, 3.46
TFLOPS,
        (attention): ViTAttention(
          2.36 M, 2.73% Params, 7.91 GMACs, 2.88% MACs, 6.03 ms, 3.60% latency, 2.63
TFLOPS,
          (attention): ViTSelfAttention(
            1.77 M, 2.05% Params, 6.05 GMACs, 2.21% MACs, 4.83 ms, 2.88% latency,
2.51 TFLOPS,
            (query): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 889.78
us, 0.53% latency, 4.18 TFLOPS, in_features=768, out_features=768, bias=True)
            (key): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 873.09 us,
0.52% latency, 4.26 TFLOPS, in_features=768, out_features=768, bias=True)
            (value): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 878.81
us, 0.52% latency, 4.23 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 58.89 us, 0.04%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 1.07 ms, 0.64% latency,
3.47 TFLOPS,
            (dense): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 908.14
us, 0.54% latency, 4.09 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 52.69 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 1.96 ms, 1.17% latency, 7.6
```

```
TFLOPS,
          (dense): Linear(2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 843.52 us,
0.50% latency, 17.63 TFLOPS, in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
996.11 us, 0.59% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 1.31 ms, 0.78% latency,
11.33 TFLOPS,
          (dense): Linear(2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 897.65 us,
0.54% latency, 16.57 TFLOPS, in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 52.69 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.7
ms, 1.02% latency, 3.55 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.67
ms, 1.00% latency, 3.62 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
      )
      (7): ViTLayer(
        7.09 M, 8.20% Params, 22.79 GMACs, 8.30% MACs, 15.54 ms, 9.27% latency, 2.93
TFLOPS,
        (attention): ViTAttention(
          2.36 M, 2.73% Params, 7.91 GMACs, 2.88% MACs, 7.53 ms, 4.49% latency, 2.1
TFLOPS,
          (attention): ViTSelfAttention(
            1.77 M, 2.05% Params, 6.05 GMACs, 2.21% MACs, 6.24 ms, 3.72% latency,
1.94 TFLOPS,
            (query): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 900.75
us, 0.54% latency, 4.13 TFLOPS, in_features=768, out_features=768, bias=True)
            (key): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 958.44 us,
0.57% latency, 3.88 TFLOPS, in_features=768, out_features=768, bias=True)
            (value): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 1.05 ms,
0.63% latency, 3.54 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 70.33 us, 0.04%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 1.14 ms, 0.68% latency,
3.26 TFLOPS,
            (dense): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 970.36
us, 0.58% latency, 3.83 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 57.22 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 2.35 ms, 1.40% latency, 6.33
TFLOPS,
          (dense): Linear(2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 993.25 us,
```

```
0.59% latency, 14.97 TFLOPS, in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
1.23 ms, 0.73% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 1.48 ms, 0.88% latency,
10.06 TFLOPS,
          (dense): Linear(2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 1.01 ms,
0.60% latency, 14.71 TFLOPS, in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 52.21 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.75
ms, 1.04% latency, 3.46 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.9
ms, 1.13% latency, 3.19 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
      )
      (8): ViTLayer(
        7.09 M, 8.20% Params, 22.79 GMACs, 8.30% MACs, 15.45 ms, 9.22% latency, 2.95
TFLOPS,
        (attention): ViTAttention(
          2.36 M, 2.73% Params, 7.91 GMACs, 2.88% MACs, 7.28 ms, 4.34% latency, 2.17
TFLOPS,
          (attention): ViTSelfAttention(
            1.77 M, 2.05% Params, 6.05 GMACs, 2.21% MACs, 5.56 ms, 3.32% latency,
2.18 TFLOPS,
            (query): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 991.11
us, 0.59% latency, 3.75 TFLOPS, in_features=768, out_features=768, bias=True)
            (key): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 955.34 us,
0.57% latency, 3.89 TFLOPS, in_features=768, out_features=768, bias=True)
            (value): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 1.07 ms,
0.64% latency, 3.46 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 64.37 us, 0.04%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 1.58 ms, 0.94% latency,
2.35 TFLOPS,
            (dense): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 1.05 ms,
0.62% latency, 3.55 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 346.18 us, 0.21%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 2.24 ms, 1.33% latency, 6.65
TFLOPS,
          (dense): Linear(2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 987.05 us,
0.59% latency, 15.07 TFLOPS, in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
```

```
1.1 ms, 0.66% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 1.44 ms, 0.86% latency,
10.34 TFLOPS,
          (dense): Linear(2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 971.08 us,
0.58% latency, 15.32 TFLOPS, in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 56.27 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.87
ms, 1.12% latency, 3.23 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 2.06
ms, 1.23% latency, 2.94 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
      )
      (9): ViTLayer(
        7.09 M, 8.20% Params, 22.79 GMACs, 8.30% MACs, 15.55 ms, 9.28% latency, 2.93
TFLOPS,
        (attention): ViTAttention(
          2.36 M, 2.73% Params, 7.91 GMACs, 2.88% MACs, 7.25 ms, 4.33% latency, 2.18
TFLOPS,
          (attention): ViTSelfAttention(
            1.77 M, 2.05% Params, 6.05 GMACs, 2.21% MACs, 5.95 ms, 3.55% latency,
2.03 TFLOPS,
            (query): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 957.25
us, 0.57% latency, 3.88 TFLOPS, in_features=768, out_features=768, bias=True)
            (key): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 997.07 us,
0.59% latency, 3.73 TFLOPS, in_features=768, out_features=768, bias=True)
            (value): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 1.14 ms,
0.68% latency, 3.27 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 74.63 us, 0.04%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 1.17 ms, 0.70% latency,
3.19 TFLOPS,
            (dense): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 954.39
us, 0.57% latency, 3.9 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 74.15 us, 0.04%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 2.38 ms, 1.42% latency, 6.26
TFLOPS,
          (dense): Linear(2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 1.03 ms,
0.62% latency, 14.39 TFLOPS, in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
1.21 ms, 0.72% latency, 0.0 FLOPS, )
        )
```

```
        (output): ViTOutput(
          2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 1.46 ms, 0.87% latency,
10.19 TFLOPS,
          (dense): Linear(2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 1.02 ms,
0.61% latency, 14.6 TFLOPS, in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 51.02 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 2.05
ms, 1.22% latency, 2.96 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.9
ms, 1.13% latency, 3.19 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
      )
      (10): ViTLayer(
        7.09 M, 8.20% Params, 22.79 GMACs, 8.30% MACs, 12.74 ms, 7.60% latency, 3.58
TFLOPS,
        (attention): ViTAttention(
          2.36 M, 2.73% Params, 7.91 GMACs, 2.88% MACs, 5.68 ms, 3.39% latency, 2.79
TFLOPS,
          (attention): ViTSelfAttention(
            1.77 M, 2.05% Params, 6.05 GMACs, 2.21% MACs, 4.59 ms, 2.74% latency,
2.64 TFLOPS,
            (query): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 901.46
us, 0.54% latency, 4.12 TFLOPS, in_features=768, out_features=768, bias=True)
            (key): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 781.77 us,
0.47% latency, 4.76 TFLOPS, in_features=768, out_features=768, bias=True)
            (value): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 790.6
us, 0.47% latency, 4.7 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 58.17 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 981.33 us, 0.59%
latency, 3.79 TFLOPS,
            (dense): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 829.46
us, 0.49% latency, 4.48 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 46.97 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 1.96 ms, 1.17% latency, 7.59
TFLOPS,
          (dense): Linear(2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 908.61 us,
0.54% latency, 16.37 TFLOPS, in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
933.89 us, 0.56% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 1.41 ms, 0.84% latency,
```

```
10.54 TFLOPS,
          (dense): Linear(2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 933.41 us,
0.56% latency, 15.93 TFLOPS, in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 54.6 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.65
ms, 0.98% latency, 3.68 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.51
ms, 0.90% latency, 4.01 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
      )
      (11): ViTLayer(
        7.09 M, 8.20% Params, 22.79 GMACs, 8.30% MACs, 12.85 ms, 7.67% latency, 3.55
TFLOPS,
        (attention): ViTAttention(
          2.36 M, 2.73% Params, 7.91 GMACs, 2.88% MACs, 6.12 ms, 3.65% latency, 2.59
TFLOPS,
          (attention): ViTSelfAttention(
            1.77 M, 2.05% Params, 6.05 GMACs, 2.21% MACs, 4.96 ms, 2.96% latency,
2.44 TFLOPS,
            (query): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 909.33
us, 0.54% latency, 4.09 TFLOPS, in_features=768, out_features=768, bias=True)
            (key): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 931.26 us,
0.56% latency, 3.99 TFLOPS, in_features=768, out_features=768, bias=True)
            (value): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 877.86
us, 0.52% latency, 4.24 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 54.6 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 988.48 us, 0.59%
latency, 3.76 TFLOPS,
            (dense): Linear(590.59 k, 0.68% Params, 1.86 GMACs, 0.68% MACs, 825.88
us, 0.49% latency, 4.5 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 46.49 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 1.92 ms, 1.15% latency, 7.73
TFLOPS,
          (dense): Linear(2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 905.75 us,
0.54% latency, 16.42 TFLOPS, in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
906.71 us, 0.54% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 1.2 ms, 0.71% latency, 12.41
TFLOPS,
          (dense): Linear(2.36 M, 2.73% Params, 7.44 GMACs, 2.71% MACs, 826.12 us,
```

```
0.49% latency, 18.0 TFLOPS, in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 46.73 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.7
ms, 1.02% latency, 3.55 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.42
ms, 0.85% latency, 4.25 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
      )
    )
  )
  (layernorm): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.64 ms, 0.98%
latency, 3.68 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
  (pooler): ViTPooler(
    590.59 k, 0.68% Params, 9.44 MMACs, 0.00% MACs, 1.2 ms, 0.72% latency, 15.67
GFLOPS,
    (dense): Linear(590.59 k, 0.68% Params, 9.44 MMACs, 0.00% MACs, 825.17 us, 0.49%
latency, 22.87 GFLOPS, in_features=768, out_features=768, bias=True)
    (activation): Tanh(0, 0.00% Params, 0 MACs, 0.00% MACs, 206.71 us, 0.12%
latency, 0.0 FLOPS, )
  )
)
------------------------------------------------------------------------------
```