```
------------------------- 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:                                               306.66 M
params of model = params per GPU * mp_size:                   306.66 M
fwd MACs per GPU:                                             712.52 GMACs
fwd flops per GPU:                                            1425.41 G
fwd flops of model = fwd flops per GPU * mp_size:             1425.41 G
fwd latency:                                                  213.16 ms
fwd FLOPS per GPU = fwd flops per GPU / fwd latency:          6.69 TFLOPS

----------------------------- Aggregated Profile per GPU
-----------------------------
Top 1 modules in terms of params, MACs or fwd latency at different model depths:
depth 0:
    params      - {'ViTModel': '306.66 M'}
    MACs        - {'ViTModel': '712.52 GMACs'}
    fwd latency - {'ViTModel': '213.16 ms'}
depth 1:
    params      - {'ViTEncoder': '302.31 M'}
    MACs        - {'ViTEncoder': '708.88 GMACs'}
    fwd latency - {'ViTEncoder': '209.93 ms'}
depth 2:
    params      - {'ModuleList': '302.31 M'}
    MACs        - {'ModuleList': '708.88 GMACs'}
    fwd latency - {'ModuleList': '209.22 ms'}
depth 3:
    params      - {'ViTLayer': '302.31 M'}
    MACs        - {'ViTLayer': '708.88 GMACs'}
    fwd latency - {'ViTLayer': '209.22 ms'}
depth 4:
    params      - {'ViTAttention': '100.76 M'}
    MACs        - {'ViTAttention': '241.81 GMACs'}
    fwd latency - {'ViTAttention': '94.12 ms'}
depth 5:
    params      - {'Linear': '201.45 M'}
    MACs        - {'Linear': '467.08 GMACs'}
    fwd latency - {'ViTSelfAttention': '76.02 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(
  306.66 M, 100.00% Params, 712.52 GMACs, 100.00% MACs, 213.16 ms, 100.00% latency,
6.69 TFLOPS,
  (embeddings): ViTEmbeddings(
    3.3 M, 1.07% Params, 3.62 GMACs, 0.51% MACs, 991.11 us, 0.46% latency, 7.31
TFLOPS,
    (patch_embeddings): ViTPatchEmbeddings(
      3.15 M, 1.03% Params, 3.62 GMACs, 0.51% MACs, 427.48 us, 0.20% latency, 16.96
TFLOPS,
      (projection): Conv2d(3.15 M, 1.03% Params, 3.62 GMACs, 0.51% MACs, 339.51 us,
0.16% latency, 21.35 TFLOPS, 3, 1024, kernel_size=(32, 32), stride=(32, 32))
    )
    (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 47.45 us, 0.02% latency,
0.0 FLOPS, p=0.0, inplace=False)
  )
  (encoder): ViTEncoder(
    302.31 M, 98.58% Params, 708.88 GMACs, 99.49% MACs, 209.93 ms, 98.48% latency,
6.76 TFLOPS,
    (layer): ModuleList(
      (0): ViTLayer(
        12.6 M, 4.11% Params, 29.54 GMACs, 4.15% MACs, 8.59 ms, 4.03% latency, 6.88
TFLOPS,
        (attention): ViTAttention(
          4.2 M, 1.37% Params, 10.08 GMACs, 1.41% MACs, 3.98 ms, 1.87% latency, 5.06
TFLOPS,
          (attention): ViTSelfAttention(
            3.15 M, 1.03% Params, 7.64 GMACs, 1.07% MACs, 3.21 ms, 1.51% latency,
4.76 TFLOPS,
            (query): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 661.85 us,
0.31% latency, 7.35 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (key): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 549.55 us,
0.26% latency, 8.85 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (value): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 567.67 us,
0.27% latency, 8.57 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 42.2 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
```

```
          )
          (output): ViTSelfOutput(
            1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 691.89 us, 0.32% latency,
7.03 TFLOPS,
            (dense): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 583.89 us,
0.27% latency, 8.33 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 34.09 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.32 ms, 0.62% latency, 14.74
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 645.64 us,
0.30% latency, 30.14 TFLOPS, in_features=1024, out_features=4096, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
595.57 us, 0.28% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 875.95 us, 0.41% latency,
22.22 TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 604.87 us,
0.28% latency, 32.17 TFLOPS, in_features=4096, out_features=1024, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 36.24 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.07
ms, 0.50% latency, 5.56 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.03
ms, 0.49% latency, 5.74 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
      )
      (1): ViTLayer(
        12.6 M, 4.11% Params, 29.54 GMACs, 4.15% MACs, 8.8 ms, 4.13% latency, 6.71
TFLOPS,
        (attention): ViTAttention(
          4.2 M, 1.37% Params, 10.08 GMACs, 1.41% MACs, 4.06 ms, 1.91% latency, 4.96
TFLOPS,
          (attention): ViTSelfAttention(
            3.15 M, 1.03% Params, 7.64 GMACs, 1.07% MACs, 3.22 ms, 1.51% latency,
4.75 TFLOPS,
            (query): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 563.14 us,
0.26% latency, 8.64 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (key): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 599.15 us,
0.28% latency, 8.12 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (value): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 548.36 us,
0.26% latency, 8.87 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 41.96 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 726.22 us, 0.34% latency,
6.7 TFLOPS,
            (dense): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 598.43 us,
0.28% latency, 8.13 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 35.76 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.29 ms, 0.61% latency, 15.04
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 552.42 us,
0.26% latency, 35.23 TFLOPS, in_features=1024, out_features=4096, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
643.49 us, 0.30% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 877.38 us, 0.41% latency,
22.18 TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 605.58 us,
0.28% latency, 32.14 TFLOPS, in_features=4096, out_features=1024, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 34.09 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.11
ms, 0.52% latency, 5.37 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.12
ms, 0.53% latency, 5.29 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
      )
      (2): ViTLayer(
        12.6 M, 4.11% Params, 29.54 GMACs, 4.15% MACs, 8.48 ms, 3.98% latency, 6.97
TFLOPS,
        (attention): ViTAttention(
          4.2 M, 1.37% Params, 10.08 GMACs, 1.41% MACs, 3.86 ms, 1.81% latency, 5.22
TFLOPS,
          (attention): ViTSelfAttention(
            3.15 M, 1.03% Params, 7.64 GMACs, 1.07% MACs, 3.1 ms, 1.45% latency,
4.93 TFLOPS,
            (query): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 587.22 us,
0.28% latency, 8.29 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (key): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 527.14 us,
0.25% latency, 9.23 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (value): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 545.5 us,
0.26% latency, 8.92 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 41.96 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 687.36 us, 0.32% latency,
7.08 TFLOPS,
            (dense): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 580.55 us,
0.27% latency, 8.38 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 33.86 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.24 ms, 0.58% latency, 15.66
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 548.84 us,
0.26% latency, 35.46 TFLOPS, in_features=1024, out_features=4096, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
618.22 us, 0.29% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 914.1 us, 0.43% latency,
21.29 TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 564.34 us,
0.26% latency, 34.49 TFLOPS, in_features=4096, out_features=1024, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 34.57 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.09
ms, 0.51% latency, 5.47 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.05
ms, 0.49% latency, 5.64 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
      )
      (3): ViTLayer(
        12.6 M, 4.11% Params, 29.54 GMACs, 4.15% MACs, 10.52 ms, 4.94% latency, 5.62
TFLOPS,
        (attention): ViTAttention(
          4.2 M, 1.37% Params, 10.08 GMACs, 1.41% MACs, 4.95 ms, 2.32% latency, 4.07
TFLOPS,
          (attention): ViTSelfAttention(
            3.15 M, 1.03% Params, 7.64 GMACs, 1.07% MACs, 4.19 ms, 1.96% latency,
3.65 TFLOPS,
            (query): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 956.06 us,
0.45% latency, 5.09 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (key): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 835.42 us,
0.39% latency, 5.82 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (value): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 613.21 us,
0.29% latency, 7.93 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 46.25 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 666.86 us, 0.31% latency,
7.3 TFLOPS,
            (dense): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 561.48 us,
0.26% latency, 8.67 TFLOPS, in_features=1024, out_features=1024, bias=True)
```

```
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 33.86 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.34 ms, 0.63% latency, 14.49
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 547.89 us,
0.26% latency, 35.52 TFLOPS, in_features=1024, out_features=4096, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
717.16 us, 0.34% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 919.82 us, 0.43% latency,
21.16 TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 626.09 us,
0.29% latency, 31.08 TFLOPS, in_features=4096, out_features=1024, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 33.38 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.77
ms, 0.83% latency, 3.35 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.1
ms, 0.52% latency, 5.4 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
      )
      (4): ViTLayer(
        12.6 M, 4.11% Params, 29.54 GMACs, 4.15% MACs, 8.5 ms, 3.99% latency, 6.95
TFLOPS,
        (attention): ViTAttention(
          4.2 M, 1.37% Params, 10.08 GMACs, 1.41% MACs, 4.03 ms, 1.89% latency, 5.0
TFLOPS,
          (attention): ViTSelfAttention(
            3.15 M, 1.03% Params, 7.64 GMACs, 1.07% MACs, 3.33 ms, 1.56% latency,
4.59 TFLOPS,
            (query): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 591.99 us,
0.28% latency, 8.22 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (key): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 573.87 us,
0.27% latency, 8.48 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (value): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 616.31 us,
0.29% latency, 7.89 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 42.44 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 619.89 us, 0.29% latency,
7.85 TFLOPS,
            (dense): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 519.28 us,
0.24% latency, 9.37 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 32.19 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
```

```
          )
        )
        (intermediate): ViTIntermediate(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.18 ms, 0.56% latency, 16.43
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 520.23 us,
0.24% latency, 37.41 TFLOPS, in_features=1024, out_features=4096, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
591.75 us, 0.28% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 839.95 us, 0.39% latency,
23.17 TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 548.12 us,
0.26% latency, 35.51 TFLOPS, in_features=4096, out_features=1024, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 34.09 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.07
ms, 0.50% latency, 5.54 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.04
ms, 0.49% latency, 5.73 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
      )
      (5): ViTLayer(
        12.6 M, 4.11% Params, 29.54 GMACs, 4.15% MACs, 13.4 ms, 6.29% latency, 4.41
TFLOPS,
        (attention): ViTAttention(
          4.2 M, 1.37% Params, 10.08 GMACs, 1.41% MACs, 4.15 ms, 1.95% latency, 4.85
TFLOPS,
          (attention): ViTSelfAttention(
            3.15 M, 1.03% Params, 7.64 GMACs, 1.07% MACs, 3.4 ms, 1.59% latency, 4.5
TFLOPS,
            (query): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 666.86 us,
0.31% latency, 7.3 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (key): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 577.45 us,
0.27% latency, 8.43 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (value): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 591.52 us,
0.28% latency, 8.23 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 42.2 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 679.97 us, 0.32% latency,
7.16 TFLOPS,
            (dense): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 578.4 us,
0.27% latency, 8.41 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 31.95 us, 0.01%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 4.53 ms, 2.13% latency, 4.29
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 2.31 ms,
1.09% latency, 8.41 TFLOPS, in_features=1024, out_features=4096, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
2.13 ms, 1.00% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.08 ms, 0.51% latency, 18.02
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 765.32 us,
0.36% latency, 25.43 TFLOPS, in_features=4096, out_features=1024, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 35.05 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.12
ms, 0.52% latency, 5.31 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 2.2
ms, 1.03% latency, 2.7 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
      )
      (6): ViTLayer(
        12.6 M, 4.11% Params, 29.54 GMACs, 4.15% MACs, 8.39 ms, 3.94% latency, 7.04
TFLOPS,
        (attention): ViTAttention(
          4.2 M, 1.37% Params, 10.08 GMACs, 1.41% MACs, 3.76 ms, 1.76% latency, 5.36
TFLOPS,
          (attention): ViTSelfAttention(
            3.15 M, 1.03% Params, 7.64 GMACs, 1.07% MACs, 2.99 ms, 1.40% latency,
5.12 TFLOPS,
            (query): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 552.42 us,
0.26% latency, 8.81 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (key): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 515.94 us,
0.24% latency, 9.43 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (value): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 537.87 us,
0.25% latency, 9.05 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 47.45 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 694.27 us, 0.33% latency,
7.01 TFLOPS,
            (dense): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 580.07 us,
0.27% latency, 8.39 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 41.72 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.3 ms, 0.61% latency, 15.0
```

```
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 581.26 us,
0.27% latency, 33.48 TFLOPS, in_features=1024, out_features=4096, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
638.01 us, 0.30% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 833.51 us, 0.39% latency,
23.35 TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 563.86 us,
0.26% latency, 34.51 TFLOPS, in_features=4096, out_features=1024, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 33.86 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.09
ms, 0.51% latency, 5.46 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.04
ms, 0.49% latency, 5.69 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
      )
      (7): ViTLayer(
        12.6 M, 4.11% Params, 29.54 GMACs, 4.15% MACs, 8.13 ms, 3.82% latency, 7.27
TFLOPS,
        (attention): ViTAttention(
          4.2 M, 1.37% Params, 10.08 GMACs, 1.41% MACs, 3.68 ms, 1.73% latency, 5.47
TFLOPS,
          (attention): ViTSelfAttention(
            3.15 M, 1.03% Params, 7.64 GMACs, 1.07% MACs, 2.97 ms, 1.39% latency,
5.15 TFLOPS,
            (query): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 538.11 us,
0.25% latency, 9.04 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (key): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 523.57 us,
0.25% latency, 9.29 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (value): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 535.49 us,
0.25% latency, 9.09 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 40.29 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 638.01 us, 0.30% latency,
7.63 TFLOPS,
            (dense): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 538.11 us,
0.25% latency, 9.04 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 31.95 us, 0.01%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.26 ms, 0.59% latency, 15.45
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 566.96 us,
0.27% latency, 34.33 TFLOPS, in_features=1024, out_features=4096, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
616.31 us, 0.29% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 847.82 us, 0.40% latency,
22.95 TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 559.81 us,
0.26% latency, 34.76 TFLOPS, in_features=4096, out_features=1024, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 34.09 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs,
985.15 us, 0.46% latency, 6.03 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.04
ms, 0.49% latency, 5.68 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
      )
      (8): ViTLayer(
        12.6 M, 4.11% Params, 29.54 GMACs, 4.15% MACs, 8.04 ms, 3.77% latency, 7.35
TFLOPS,
        (attention): ViTAttention(
          4.2 M, 1.37% Params, 10.08 GMACs, 1.41% MACs, 3.75 ms, 1.76% latency, 5.37
TFLOPS,
          (attention): ViTSelfAttention(
            3.15 M, 1.03% Params, 7.64 GMACs, 1.07% MACs, 2.96 ms, 1.39% latency,
5.17 TFLOPS,
            (query): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 531.91 us,
0.25% latency, 9.15 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (key): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 525.0 us,
0.25% latency, 9.27 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (value): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 514.98 us,
0.24% latency, 9.45 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 41.72 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 717.64 us, 0.34% latency,
6.78 TFLOPS,
            (dense): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 596.52 us,
0.28% latency, 8.16 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 34.57 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.18 ms, 0.55% latency, 16.49
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 530.96 us,
0.25% latency, 36.65 TFLOPS, in_features=1024, out_features=4096, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
575.07 us, 0.27% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 787.26 us, 0.37% latency,
24.72 TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 518.8 us,
0.24% latency, 37.51 TFLOPS, in_features=4096, out_features=1024, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 33.14 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs,
974.18 us, 0.46% latency, 6.1 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.01
ms, 0.48% latency, 5.85 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
      )
      (9): ViTLayer(
        12.6 M, 4.11% Params, 29.54 GMACs, 4.15% MACs, 8.45 ms, 3.96% latency, 6.99
TFLOPS,
        (attention): ViTAttention(
          4.2 M, 1.37% Params, 10.08 GMACs, 1.41% MACs, 3.99 ms, 1.87% latency, 5.05
TFLOPS,
          (attention): ViTSelfAttention(
            3.15 M, 1.03% Params, 7.64 GMACs, 1.07% MACs, 3.25 ms, 1.53% latency,
4.7 TFLOPS,
            (query): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 541.69 us,
0.25% latency, 8.98 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (key): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 615.6 us,
0.29% latency, 7.9 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (value): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 607.73 us,
0.29% latency, 8.01 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 57.94 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 662.33 us, 0.31% latency,
7.35 TFLOPS,
            (dense): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 561.0 us,
0.26% latency, 8.67 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 31.71 us, 0.01%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.31 ms, 0.61% latency, 14.91
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 577.93 us,
0.27% latency, 33.67 TFLOPS, in_features=1024, out_features=4096, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
648.98 us, 0.30% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 809.43 us, 0.38% latency,
24.04 TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 550.99 us,
0.26% latency, 35.32 TFLOPS, in_features=4096, out_features=1024, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 33.14 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.02
ms, 0.48% latency, 5.83 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs,
999.69 us, 0.47% latency, 5.94 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
      )
      (10): ViTLayer(
        12.6 M, 4.11% Params, 29.54 GMACs, 4.15% MACs, 9.41 ms, 4.41% latency, 6.28
TFLOPS,
        (attention): ViTAttention(
          4.2 M, 1.37% Params, 10.08 GMACs, 1.41% MACs, 3.68 ms, 1.73% latency, 5.48
TFLOPS,
          (attention): ViTSelfAttention(
            3.15 M, 1.03% Params, 7.64 GMACs, 1.07% MACs, 2.94 ms, 1.38% latency,
5.19 TFLOPS,
            (query): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 567.44 us,
0.27% latency, 8.57 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (key): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 509.98 us,
0.24% latency, 9.54 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (value): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 517.61 us,
0.24% latency, 9.4 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 40.29 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 658.04 us, 0.31% latency,
7.39 TFLOPS,
            (dense): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 547.89 us,
0.26% latency, 8.88 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 32.19 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.73 ms, 0.81% latency, 11.23
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 673.29 us,
0.32% latency, 28.91 TFLOPS, in_features=1024, out_features=4096, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
957.97 us, 0.45% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.66 ms, 0.78% latency, 11.72
```

```
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 842.09 us,
0.40% latency, 23.11 TFLOPS, in_features=4096, out_features=1024, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 52.21 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.03
ms, 0.48% latency, 5.78 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs,
966.07 us, 0.45% latency, 6.15 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
      )
      (11): ViTLayer(
        12.6 M, 4.11% Params, 29.54 GMACs, 4.15% MACs, 8.36 ms, 3.92% latency, 7.07
TFLOPS,
        (attention): ViTAttention(
          4.2 M, 1.37% Params, 10.08 GMACs, 1.41% MACs, 3.73 ms, 1.75% latency, 5.41
TFLOPS,
          (attention): ViTSelfAttention(
            3.15 M, 1.03% Params, 7.64 GMACs, 1.07% MACs, 2.99 ms, 1.40% latency,
5.12 TFLOPS,
            (query): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 529.05 us,
0.25% latency, 9.2 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (key): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 537.87 us,
0.25% latency, 9.05 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (value): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 553.37 us,
0.26% latency, 8.79 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 41.25 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 661.13 us, 0.31% latency,
7.36 TFLOPS,
            (dense): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 520.71 us,
0.24% latency, 9.34 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 39.82 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.3 ms, 0.61% latency, 15.03
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 577.69 us,
0.27% latency, 33.69 TFLOPS, in_features=1024, out_features=4096, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
605.11 us, 0.28% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 788.69 us, 0.37% latency,
24.68 TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 547.17 us,
0.26% latency, 35.57 TFLOPS, in_features=4096, out_features=1024, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 31.71 us, 0.01%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.1
ms, 0.51% latency, 5.42 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.13
ms, 0.53% latency, 5.26 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
      )
      (12): ViTLayer(
        12.6 M, 4.11% Params, 29.54 GMACs, 4.15% MACs, 8.1 ms, 3.80% latency, 7.3
TFLOPS,
        (attention): ViTAttention(
          4.2 M, 1.37% Params, 10.08 GMACs, 1.41% MACs, 3.81 ms, 1.79% latency, 5.29
TFLOPS,
          (attention): ViTSelfAttention(
            3.15 M, 1.03% Params, 7.64 GMACs, 1.07% MACs, 3.07 ms, 1.44% latency,
4.98 TFLOPS,
            (query): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 591.28 us,
0.28% latency, 8.23 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (key): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 528.57 us,
0.25% latency, 9.2 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (value): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 566.96 us,
0.27% latency, 8.58 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 40.29 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 664.71 us, 0.31% latency,
7.32 TFLOPS,
            (dense): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 561.48 us,
0.26% latency, 8.67 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 32.19 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.18 ms, 0.55% latency, 16.56
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 556.47 us,
0.26% latency, 34.97 TFLOPS, in_features=1024, out_features=4096, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
544.55 us, 0.26% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 818.25 us, 0.38% latency,
23.78 TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 573.4 us,
0.27% latency, 33.94 TFLOPS, in_features=4096, out_features=1024, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 32.42 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.02
ms, 0.48% latency, 5.82 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs,
981.33 us, 0.46% latency, 6.05 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
      )
      (13): ViTLayer(
        12.6 M, 4.11% Params, 29.54 GMACs, 4.15% MACs, 8.65 ms, 4.06% latency, 6.83
TFLOPS,
        (attention): ViTAttention(
          4.2 M, 1.37% Params, 10.08 GMACs, 1.41% MACs, 4.1 ms, 1.92% latency, 4.92
TFLOPS,
          (attention): ViTSelfAttention(
            3.15 M, 1.03% Params, 7.64 GMACs, 1.07% MACs, 3.39 ms, 1.59% latency,
4.51 TFLOPS,
            (query): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 617.03 us,
0.29% latency, 7.89 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (key): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 512.36 us,
0.24% latency, 9.5 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (value): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 543.83 us,
0.26% latency, 8.95 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 42.68 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 628.71 us, 0.29% latency,
7.74 TFLOPS,
            (dense): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 527.62 us,
0.25% latency, 9.22 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 31.71 us, 0.01%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.23 ms, 0.58% latency, 15.87
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 549.55 us,
0.26% latency, 35.41 TFLOPS, in_features=1024, out_features=4096, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
600.58 us, 0.28% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 921.25 us, 0.43% latency,
21.13 TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 584.84 us,
0.27% latency, 33.28 TFLOPS, in_features=4096, out_features=1024, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 42.44 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.05
ms, 0.49% latency, 5.66 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.02
ms, 0.48% latency, 5.85 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
      )
      (14): ViTLayer(
        12.6 M, 4.11% Params, 29.54 GMACs, 4.15% MACs, 8.39 ms, 3.94% latency, 7.04
TFLOPS,
        (attention): ViTAttention(
          4.2 M, 1.37% Params, 10.08 GMACs, 1.41% MACs, 3.94 ms, 1.85% latency, 5.12
TFLOPS,
          (attention): ViTSelfAttention(
            3.15 M, 1.03% Params, 7.64 GMACs, 1.07% MACs, 3.15 ms, 1.48% latency,
4.86 TFLOPS,
            (query): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 588.18 us,
0.28% latency, 8.27 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (key): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 501.87 us,
0.24% latency, 9.69 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (value): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 608.44 us,
0.29% latency, 8.0 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 41.01 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 711.2 us, 0.33% latency,
6.84 TFLOPS,
            (dense): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 588.42 us,
0.28% latency, 8.27 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 33.38 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.17 ms, 0.55% latency, 16.62
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 551.7 us,
0.26% latency, 35.28 TFLOPS, in_features=1024, out_features=4096, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
546.46 us, 0.26% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 819.44 us, 0.38% latency,
23.75 TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 540.97 us,
0.25% latency, 35.98 TFLOPS, in_features=4096, out_features=1024, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 32.9 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.16
ms, 0.54% latency, 5.12 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
```

```
        (layernorm_after): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs,
994.44 us, 0.47% latency, 5.97 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
      )
      (15): ViTLayer(
        12.6 M, 4.11% Params, 29.54 GMACs, 4.15% MACs, 9.09 ms, 4.26% latency, 6.5
TFLOPS,
        (attention): ViTAttention(
          4.2 M, 1.37% Params, 10.08 GMACs, 1.41% MACs, 3.89 ms, 1.83% latency, 5.18
TFLOPS,
          (attention): ViTSelfAttention(
            3.15 M, 1.03% Params, 7.64 GMACs, 1.07% MACs, 3.08 ms, 1.44% latency,
4.97 TFLOPS,
            (query): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 575.54 us,
0.27% latency, 8.45 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (key): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 556.47 us,
0.26% latency, 8.74 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (value): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 504.49 us,
0.24% latency, 9.64 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 42.44 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 722.41 us, 0.34% latency,
6.73 TFLOPS,
            (dense): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 615.6 us,
0.29% latency, 7.9 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 33.86 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.86 ms, 0.87% latency, 10.49
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.16 ms,
0.55% latency, 16.74 TFLOPS, in_features=1024, out_features=4096, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
600.58 us, 0.28% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 854.97 us, 0.40% latency,
22.76 TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 563.38 us,
0.26% latency, 34.54 TFLOPS, in_features=4096, out_features=1024, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 34.57 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.07
ms, 0.50% latency, 5.54 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.08
ms, 0.51% latency, 5.51 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
```

```
      )
      (16): ViTLayer(
        12.6 M, 4.11% Params, 29.54 GMACs, 4.15% MACs, 8.25 ms, 3.87% latency, 7.17
TFLOPS,
        (attention): ViTAttention(
          4.2 M, 1.37% Params, 10.08 GMACs, 1.41% MACs, 3.74 ms, 1.75% latency, 5.39
TFLOPS,
          (attention): ViTSelfAttention(
            3.15 M, 1.03% Params, 7.64 GMACs, 1.07% MACs, 3.04 ms, 1.43% latency,
5.02 TFLOPS,
            (query): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 556.95 us,
0.26% latency, 8.74 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (key): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 569.82 us,
0.27% latency, 8.54 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (value): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 501.63 us,
0.24% latency, 9.7 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 39.82 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 619.65 us, 0.29% latency,
7.85 TFLOPS,
            (dense): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 519.99 us,
0.24% latency, 9.36 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 31.95 us, 0.01%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.3 ms, 0.61% latency, 14.99
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 556.71 us,
0.26% latency, 34.96 TFLOPS, in_features=1024, out_features=4096, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
640.15 us, 0.30% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 802.76 us, 0.38% latency,
24.24 TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 509.74 us,
0.24% latency, 38.18 TFLOPS, in_features=4096, out_features=1024, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 31.95 us, 0.01%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.07
ms, 0.50% latency, 5.55 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.02
ms, 0.48% latency, 5.85 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
      )
      (17): ViTLayer(
        12.6 M, 4.11% Params, 29.54 GMACs, 4.15% MACs, 8.24 ms, 3.87% latency, 7.17
TFLOPS,
        (attention): ViTAttention(
          4.2 M, 1.37% Params, 10.08 GMACs, 1.41% MACs, 3.9 ms, 1.83% latency, 5.16
TFLOPS,
          (attention): ViTSelfAttention(
            3.15 M, 1.03% Params, 7.64 GMACs, 1.07% MACs, 3.15 ms, 1.48% latency,
4.85 TFLOPS,
            (query): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 517.13 us,
0.24% latency, 9.41 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (key): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 529.77 us,
0.25% latency, 9.18 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (value): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 499.73 us,
0.23% latency, 9.74 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 42.68 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 677.82 us, 0.32% latency,
7.18 TFLOPS,
            (dense): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 563.86 us,
0.26% latency, 8.63 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 32.9 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.2 ms, 0.56% latency, 16.28
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 529.29 us,
0.25% latency, 36.77 TFLOPS, in_features=1024, out_features=4096, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
589.61 us, 0.28% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 817.78 us, 0.38% latency,
23.8 TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 575.54 us,
0.27% latency, 33.81 TFLOPS, in_features=4096, out_features=1024, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 32.19 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.04
ms, 0.49% latency, 5.72 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs,
977.99 us, 0.46% latency, 6.07 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
      )
      (18): ViTLayer(
        12.6 M, 4.11% Params, 29.54 GMACs, 4.15% MACs, 8.66 ms, 4.06% latency, 6.82
TFLOPS,
        (attention): ViTAttention(
          4.2 M, 1.37% Params, 10.08 GMACs, 1.41% MACs, 4.19 ms, 1.97% latency, 4.81
TFLOPS,
          (attention): ViTSelfAttention(
            3.15 M, 1.03% Params, 7.64 GMACs, 1.07% MACs, 3.43 ms, 1.61% latency,
4.46 TFLOPS,
            (query): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 541.93 us,
0.25% latency, 8.98 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (key): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 550.75 us,
0.26% latency, 8.83 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (value): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 525.71 us,
0.25% latency, 9.25 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 41.01 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 682.59 us, 0.32% latency,
7.13 TFLOPS,
            (dense): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 581.03 us,
0.27% latency, 8.37 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 32.42 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.29 ms, 0.60% latency, 15.11
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 621.32 us,
0.29% latency, 31.32 TFLOPS, in_features=1024, out_features=4096, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
591.28 us, 0.28% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 815.15 us, 0.38% latency,
23.87 TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 543.59 us,
0.26% latency, 35.8 TFLOPS, in_features=4096, out_features=1024, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 34.33 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs,
975.13 us, 0.46% latency, 6.09 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.08
ms, 0.51% latency, 5.48 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
      )
      (19): ViTLayer(
        12.6 M, 4.11% Params, 29.54 GMACs, 4.15% MACs, 8.04 ms, 3.77% latency, 7.35
TFLOPS,
        (attention): ViTAttention(
          4.2 M, 1.37% Params, 10.08 GMACs, 1.41% MACs, 3.72 ms, 1.75% latency, 5.41
```

```
TFLOPS,
          (attention): ViTSelfAttention(
            3.15 M, 1.03% Params, 7.64 GMACs, 1.07% MACs, 3.0 ms, 1.41% latency,
5.09 TFLOPS,
            (query): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 566.72 us,
0.27% latency, 8.59 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (key): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 547.65 us,
0.26% latency, 8.88 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (value): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 519.99 us,
0.24% latency, 9.36 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 40.53 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 643.25 us, 0.30% latency,
7.56 TFLOPS,
            (dense): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 542.64 us,
0.25% latency, 8.97 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 31.95 us, 0.01%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.19 ms, 0.56% latency, 16.35
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 541.21 us,
0.25% latency, 35.96 TFLOPS, in_features=1024, out_features=4096, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
575.3 us, 0.27% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 819.68 us, 0.38% latency,
23.74 TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 531.67 us,
0.25% latency, 36.6 TFLOPS, in_features=4096, out_features=1024, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 32.19 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.0
ms, 0.47% latency, 5.92 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs,
986.34 us, 0.46% latency, 6.02 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
      )
      (20): ViTLayer(
        12.6 M, 4.11% Params, 29.54 GMACs, 4.15% MACs, 8.63 ms, 4.05% latency, 6.85
TFLOPS,
        (attention): ViTAttention(
          4.2 M, 1.37% Params, 10.08 GMACs, 1.41% MACs, 4.04 ms, 1.89% latency, 4.99
TFLOPS,
          (attention): ViTSelfAttention(
            3.15 M, 1.03% Params, 7.64 GMACs, 1.07% MACs, 3.24 ms, 1.52% latency,
4.72 TFLOPS,
            (query): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 572.92 us,
0.27% latency, 8.49 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (key): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 598.43 us,
0.28% latency, 8.13 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (value): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 602.01 us,
0.28% latency, 8.08 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 40.53 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 716.92 us, 0.34% latency,
6.79 TFLOPS,
            (dense): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 613.45 us,
0.29% latency, 7.93 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 32.19 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.21 ms, 0.57% latency, 16.05
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 568.87 us,
0.27% latency, 34.21 TFLOPS, in_features=1024, out_features=4096, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
567.2 us, 0.27% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 797.75 us, 0.37% latency,
24.4 TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 550.27 us,
0.26% latency, 35.37 TFLOPS, in_features=4096, out_features=1024, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 31.95 us, 0.01%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.14
ms, 0.53% latency, 5.21 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.13
ms, 0.53% latency, 5.27 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
      )
      (21): ViTLayer(
        12.6 M, 4.11% Params, 29.54 GMACs, 4.15% MACs, 8.06 ms, 3.78% latency, 7.33
TFLOPS,
        (attention): ViTAttention(
          4.2 M, 1.37% Params, 10.08 GMACs, 1.41% MACs, 3.84 ms, 1.80% latency, 5.24
TFLOPS,
          (attention): ViTSelfAttention(
            3.15 M, 1.03% Params, 7.64 GMACs, 1.07% MACs, 3.07 ms, 1.44% latency,
4.99 TFLOPS,
            (query): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 563.14 us,
0.26% latency, 8.64 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (key): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 516.18 us,
0.24% latency, 9.43 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (value): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 545.26 us,
0.26% latency, 8.92 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 51.98 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 700.71 us, 0.33% latency,
6.94 TFLOPS,
            (dense): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 597.72 us,
0.28% latency, 8.14 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 31.95 us, 0.01%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.2 ms, 0.56% latency, 16.28
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 539.3 us,
0.25% latency, 36.09 TFLOPS, in_features=1024, out_features=4096, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
579.36 us, 0.27% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 778.91 us, 0.37% latency,
24.99 TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 535.96 us,
0.25% latency, 36.31 TFLOPS, in_features=4096, out_features=1024, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 32.66 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs,
968.46 us, 0.45% latency, 6.13 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 973.7
us, 0.46% latency, 6.1 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
      )
      (22): ViTLayer(
        12.6 M, 4.11% Params, 29.54 GMACs, 4.15% MACs, 8.07 ms, 3.79% latency, 7.32
TFLOPS,
        (attention): ViTAttention(
          4.2 M, 1.37% Params, 10.08 GMACs, 1.41% MACs, 3.64 ms, 1.71% latency, 5.53
TFLOPS,
          (attention): ViTSelfAttention(
            3.15 M, 1.03% Params, 7.64 GMACs, 1.07% MACs, 2.91 ms, 1.37% latency,
5.25 TFLOPS,
            (query): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 548.84 us,
0.26% latency, 8.86 TFLOPS, in_features=1024, out_features=1024, bias=True)
```

```
            (key): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 521.66 us,
0.24% latency, 9.33 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (value): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 517.37 us,
0.24% latency, 9.4 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 39.82 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 654.94 us, 0.31% latency,
7.43 TFLOPS,
            (dense): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 544.79 us,
0.26% latency, 8.93 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 32.9 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.19 ms, 0.56% latency, 16.29
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 541.69 us,
0.25% latency, 35.93 TFLOPS, in_features=1024, out_features=4096, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
577.93 us, 0.27% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 979.42 us, 0.46% latency,
19.87 TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 728.85 us,
0.34% latency, 26.7 TFLOPS, in_features=4096, out_features=1024, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 33.14 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs,
974.89 us, 0.46% latency, 6.09 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs,
975.13 us, 0.46% latency, 6.09 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
      )
      (23): ViTLayer(
        12.6 M, 4.11% Params, 29.54 GMACs, 4.15% MACs, 7.98 ms, 3.74% latency, 7.4
TFLOPS,
        (attention): ViTAttention(
          4.2 M, 1.37% Params, 10.08 GMACs, 1.41% MACs, 3.67 ms, 1.72% latency, 5.49
TFLOPS,
          (attention): ViTSelfAttention(
            3.15 M, 1.03% Params, 7.64 GMACs, 1.07% MACs, 2.95 ms, 1.39% latency,
5.18 TFLOPS,
            (query): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 540.02 us,
0.25% latency, 9.01 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (key): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 553.61 us,
0.26% latency, 8.79 TFLOPS, in_features=1024, out_features=1024, bias=True)
```

```
            (value): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 519.51 us,
0.24% latency, 9.37 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 40.05 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
          (output): ViTSelfOutput(
            1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 642.06 us, 0.30% latency,
7.58 TFLOPS,
            (dense): Linear(1.05 M, 0.34% Params, 2.43 GMACs, 0.34% MACs, 542.16 us,
0.25% latency, 8.97 TFLOPS, in_features=1024, out_features=1024, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 31.47 us, 0.01%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): ViTIntermediate(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 1.2 ms, 0.56% latency, 16.2
TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 547.17 us,
0.26% latency, 35.57 TFLOPS, in_features=1024, out_features=4096, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
579.12 us, 0.27% latency, 0.0 FLOPS, )
        )
        (output): ViTOutput(
          4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 771.28 us, 0.36% latency,
25.23 TFLOPS,
          (dense): Linear(4.2 M, 1.37% Params, 9.73 GMACs, 1.37% MACs, 527.62 us,
0.25% latency, 36.89 TFLOPS, in_features=4096, out_features=1024, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 33.38 us, 0.02%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs,
984.43 us, 0.46% latency, 6.03 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
        (layernorm_after): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.04
ms, 0.49% latency, 5.74 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
      )
    )
  )
  (layernorm): LayerNorm(2.05 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.01 ms, 0.47%
latency, 5.91 GFLOPS, (1024,), eps=1e-12, elementwise_affine=True)
  (pooler): ViTPooler(
    1.05 M, 0.34% Params, 16.78 MMACs, 0.00% MACs, 977.99 us, 0.46% latency, 34.31
GFLOPS,
    (dense): Linear(1.05 M, 0.34% Params, 16.78 MMACs, 0.00% MACs, 673.29 us, 0.32%
latency, 49.84 GFLOPS, in_features=1024, out_features=1024, bias=True)
    (activation): Tanh(0, 0.00% Params, 0 MACs, 0.00% MACs, 169.04 us, 0.08%
latency, 0.0 FLOPS, )
  )
)
------------------------------------------------------------------------------
```