```
------------------------- 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:                                                 85.98 M
params of model = params per GPU * mp_size:                     85.98 M
fwd MACs per GPU:                                               803.26 GMACs
fwd flops per GPU:                                              1607.32 G
fwd flops of model = fwd flops per GPU * mp_size:               1607.32 G
fwd latency:                                                    118.7 ms
fwd FLOPS per GPU = fwd flops per GPU / fwd latency:            13.54 TFLOPS

----------------------------- Aggregated Profile per GPU
-----------------------------
Top 1 modules in terms of params, MACs or fwd latency at different model depths:
depth 0:
    params      - {'BeitModel': '85.98 M'}
    MACs        - {'BeitModel': '803.26 GMACs'}
    fwd latency - {'BeitModel': '118.7 ms'}
depth 1:
    params      - {'BeitEncoder': '85.38 M'}
    MACs        - {'BeitEncoder': '800.54 GMACs'}
    fwd latency - {'BeitEncoder': '116.36 ms'}
depth 2:
    params      - {'ModuleList': '85.38 M'}
    MACs        - {'ModuleList': '800.54 GMACs'}
    fwd latency - {'ModuleList': '115.95 ms'}
depth 3:
    params      - {'BeitLayer': '85.38 M'}
    MACs        - {'BeitLayer': '800.54 GMACs'}
    fwd latency - {'BeitLayer': '115.95 ms'}
depth 4:
    params      - {'BeitAttention': '28.66 M'}
    MACs        - {'BeitAttention': '277.79 GMACs'}
    fwd latency - {'BeitAttention': '51.1 ms'}
depth 5:
    params      - {'Linear': '56.67 M'}
    MACs        - {'Linear': '522.74 GMACs'}
    fwd latency - {'BeitSelfAttention': '41.41 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.

BeitModel(
  85.98 M, 100.00% Params, 803.26 GMACs, 100.00% MACs, 118.7 ms, 100.00% latency,
13.54 TFLOPS,
  (embeddings): BeitEmbeddings(
    591.36 k, 0.69% Params, 2.72 GMACs, 0.34% MACs, 741.24 us, 0.62% latency, 7.34
TFLOPS,
    (patch_embeddings): BeitPatchEmbeddings(
      590.59 k, 0.69% Params, 2.72 GMACs, 0.34% MACs, 305.89 us, 0.26% latency,
17.78 TFLOPS,
      (projection): Conv2d(590.59 k, 0.69% Params, 2.72 GMACs, 0.34% MACs, 221.01
us, 0.19% latency, 24.61 TFLOPS, 3, 768, kernel_size=(16, 16), stride=(16, 16))
    )
    (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 48.4 us, 0.04% latency,
0.0 FLOPS, p=0.0, inplace=False)
  )
  (encoder): BeitEncoder(
    85.38 M, 99.31% Params, 800.54 GMACs, 99.66% MACs, 116.36 ms, 98.03% latency,
13.77 TFLOPS,
    (layer): ModuleList(
      (0): BeitLayer(
        7.12 M, 8.28% Params, 66.71 GMACs, 8.31% MACs, 9.32 ms, 7.85% latency, 14.32
TFLOPS,
        (attention): BeitAttention(
          2.39 M, 2.78% Params, 23.15 GMACs, 2.88% MACs, 4.16 ms, 3.51% latency,
11.13 TFLOPS,
          (attention): BeitSelfAttention(
            1.8 M, 2.09% Params, 17.7 GMACs, 2.20% MACs, 3.36 ms, 2.83% latency,
10.56 TFLOPS,
            (query): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 613.21
us, 0.52% latency, 17.76 TFLOPS, in_features=768, out_features=768, bias=True)
            (key): Linear(589.82 k, 0.69% Params, 2.72 GMACs, 0.34% MACs, 149.25 us,
0.13% latency, 36.48 TFLOPS, in_features=768, out_features=768, bias=False)
            (value): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 617.27
us, 0.52% latency, 17.64 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 43.87 us, 0.04%
latency, 0.0 FLOPS, p=0.0, inplace=False)
```

```
            (relative_position_bias): BeitRelativePositionBias(26.54 k, 0.03%
Params, 0 MACs, 0.00% MACs, 216.96 us, 0.18% latency, 0.0 FLOPS, )
          )
          (output): BeitSelfOutput(
            590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 726.46 us, 0.61%
latency, 14.99 TFLOPS,
            (dense): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 603.2
us, 0.51% latency, 18.05 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 34.57 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): BeitIntermediate(
          2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 1.26 ms, 1.06% latency,
34.68 TFLOPS,
          (dense): Linear(2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 553.37 us,
0.47% latency, 78.72 TFLOPS, in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
624.66 us, 0.53% latency, 0.0 FLOPS, )
        )
        (output): BeitOutput(
          2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 682.83 us, 0.58% latency,
63.8 TFLOPS,
          (dense): Linear(2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 576.02 us,
0.49% latency, 75.63 TFLOPS, in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 34.09 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, 1.14
ms, 0.96% latency, 15.58 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
        (drop_path): Identity(0, 0.00% Params, 0 MACs, 0.00% MACs, 52.45 us, 0.04%
latency, 0.0 FLOPS, )
        (layernorm_after): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.14
ms, 0.96% latency, 15.49 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
      )
      (1): BeitLayer(
        7.12 M, 8.28% Params, 66.71 GMACs, 8.31% MACs, 10.03 ms, 8.45% latency, 13.3
TFLOPS,
        (attention): BeitAttention(
          2.39 M, 2.78% Params, 23.15 GMACs, 2.88% MACs, 4.87 ms, 4.10% latency,
9.52 TFLOPS,
          (attention): BeitSelfAttention(
            1.8 M, 2.09% Params, 17.7 GMACs, 2.20% MACs, 4.04 ms, 3.41% latency,
8.76 TFLOPS,
            (query): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 615.6
us, 0.52% latency, 17.69 TFLOPS, in_features=768, out_features=768, bias=True)
            (key): Linear(589.82 k, 0.69% Params, 2.72 GMACs, 0.34% MACs, 155.21 us,
0.13% latency, 35.08 TFLOPS, in_features=768, out_features=768, bias=False)
            (value): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 767.47
us, 0.65% latency, 14.19 TFLOPS, in_features=768, out_features=768, bias=True)
```

```
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 42.92 us, 0.04%
latency, 0.0 FLOPS, p=0.0, inplace=False)
            (relative_position_bias): BeitRelativePositionBias(26.54 k, 0.03%
Params, 0 MACs, 0.00% MACs, 324.01 us, 0.27% latency, 0.0 FLOPS, )
          )
          (output): BeitSelfOutput(
            590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 749.83 us, 0.63%
latency, 14.52 TFLOPS,
            (dense): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 649.45
us, 0.55% latency, 16.77 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 31.95 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): BeitIntermediate(
          2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 1.3 ms, 1.09% latency,
33.59 TFLOPS,
          (dense): Linear(2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 548.36 us,
0.46% latency, 79.44 TFLOPS, in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
669.48 us, 0.56% latency, 0.0 FLOPS, )
        )
        (output): BeitOutput(
          2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 725.98 us, 0.61% latency,
60.0 TFLOPS,
          (dense): Linear(2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 618.22 us,
0.52% latency, 70.46 TFLOPS, in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 34.81 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, 1.06
ms, 0.89% latency, 16.76 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
        (drop_path): BeitDropPath(0, 0.00% Params, 0 MACs, 0.00% MACs, 53.64 us,
0.05% latency, 0.0 FLOPS, p=0.00909090880304575)
        (layernorm_after): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.16
ms, 0.98% latency, 15.31 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
      )
      (2): BeitLayer(
        7.12 M, 8.28% Params, 66.71 GMACs, 8.31% MACs, 9.06 ms, 7.63% latency, 14.73
TFLOPS,
        (attention): BeitAttention(
          2.39 M, 2.78% Params, 23.15 GMACs, 2.88% MACs, 3.97 ms, 3.35% latency,
11.67 TFLOPS,
          (attention): BeitSelfAttention(
            1.8 M, 2.09% Params, 17.7 GMACs, 2.20% MACs, 3.25 ms, 2.74% latency,
10.91 TFLOPS,
            (query): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 601.53
us, 0.51% latency, 18.1 TFLOPS, in_features=768, out_features=768, bias=True)
            (key): Linear(589.82 k, 0.69% Params, 2.72 GMACs, 0.34% MACs, 139.95 us,
0.12% latency, 38.91 TFLOPS, in_features=768, out_features=768, bias=False)
```

```
            (value): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 628.71
us, 0.53% latency, 17.32 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 41.01 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
            (relative_position_bias): BeitRelativePositionBias(26.54 k, 0.03%
Params, 0 MACs, 0.00% MACs, 191.93 us, 0.16% latency, 0.0 FLOPS, )
          )
          (output): BeitSelfOutput(
            590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 645.64 us, 0.54%
latency, 16.87 TFLOPS,
            (dense): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 544.31
us, 0.46% latency, 20.01 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 32.42 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): BeitIntermediate(
          2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 1.31 ms, 1.11% latency,
33.13 TFLOPS,
          (dense): Linear(2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 552.42 us,
0.47% latency, 78.86 TFLOPS, in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
684.02 us, 0.58% latency, 0.0 FLOPS, )
        )
        (output): BeitOutput(
          2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 784.64 us, 0.66% latency,
55.52 TFLOPS,
          (dense): Linear(2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 663.76 us,
0.56% latency, 65.63 TFLOPS, in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 36.95 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, 1.03
ms, 0.87% latency, 17.18 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
        (drop_path): BeitDropPath(0, 0.00% Params, 0 MACs, 0.00% MACs, 65.09 us,
0.05% latency, 0.0 FLOPS, p=0.0181818176060915)
        (layernorm_after): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.04
ms, 0.87% latency, 17.08 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
      )
      (3): BeitLayer(
        7.12 M, 8.28% Params, 66.71 GMACs, 8.31% MACs, 8.86 ms, 7.47% latency, 15.06
TFLOPS,
        (attention): BeitAttention(
          2.39 M, 2.78% Params, 23.15 GMACs, 2.88% MACs, 3.9 ms, 3.28% latency,
11.88 TFLOPS,
          (attention): BeitSelfAttention(
            1.8 M, 2.09% Params, 17.7 GMACs, 2.20% MACs, 3.16 ms, 2.67% latency,
11.2 TFLOPS,
            (query): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 567.91
us, 0.48% latency, 19.18 TFLOPS, in_features=768, out_features=768, bias=True)
```

```
            (key): Linear(589.82 k, 0.69% Params, 2.72 GMACs, 0.34% MACs, 146.87 us,
0.12% latency, 37.08 TFLOPS, in_features=768, out_features=768, bias=False)
            (value): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 581.26
us, 0.49% latency, 18.74 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 75.82 us, 0.06%
latency, 0.0 FLOPS, p=0.0, inplace=False)
            (relative_position_bias): BeitRelativePositionBias(26.54 k, 0.03%
Params, 0 MACs, 0.00% MACs, 237.46 us, 0.20% latency, 0.0 FLOPS, )
          )
          (output): BeitSelfOutput(
            590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 633.0 us, 0.53% latency,
17.2 TFLOPS,
            (dense): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 533.34
us, 0.45% latency, 20.42 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 31.23 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): BeitIntermediate(
          2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 1.22 ms, 1.03% latency,
35.65 TFLOPS,
          (dense): Linear(2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 528.81 us,
0.45% latency, 82.38 TFLOPS, in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
588.18 us, 0.50% latency, 0.0 FLOPS, )
        )
        (output): BeitOutput(
          2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 649.21 us, 0.55% latency,
67.1 TFLOPS,
          (dense): Linear(2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 549.79 us,
0.46% latency, 79.23 TFLOPS, in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 32.9 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, 1.23
ms, 1.04% latency, 14.37 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
        (drop_path): BeitDropPath(0, 0.00% Params, 0 MACs, 0.00% MACs, 52.93 us,
0.04% latency, 0.0 FLOPS, p=0.027272727340459824)
        (layernorm_after): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.02
ms, 0.86% latency, 17.42 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
      )
      (4): BeitLayer(
        7.12 M, 8.28% Params, 66.71 GMACs, 8.31% MACs, 9.55 ms, 8.05% latency, 13.98
TFLOPS,
        (attention): BeitAttention(
          2.39 M, 2.78% Params, 23.15 GMACs, 2.88% MACs, 3.94 ms, 3.32% latency,
11.76 TFLOPS,
          (attention): BeitSelfAttention(
            1.8 M, 2.09% Params, 17.7 GMACs, 2.20% MACs, 3.2 ms, 2.70% latency,
11.07 TFLOPS,
```

```
            (query): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 639.68
us, 0.54% latency, 17.03 TFLOPS, in_features=768, out_features=768, bias=True)
            (key): Linear(589.82 k, 0.69% Params, 2.72 GMACs, 0.34% MACs, 152.59 us,
0.13% latency, 35.69 TFLOPS, in_features=768, out_features=768, bias=False)
            (value): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 532.87
us, 0.45% latency, 20.44 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 41.96 us, 0.04%
latency, 0.0 FLOPS, p=0.0, inplace=False)
            (relative_position_bias): BeitRelativePositionBias(26.54 k, 0.03%
Params, 0 MACs, 0.00% MACs, 222.68 us, 0.19% latency, 0.0 FLOPS, )
          )
          (output): BeitSelfOutput(
            590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 655.89 us, 0.55%
latency, 16.6 TFLOPS,
            (dense): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 531.44
us, 0.45% latency, 20.49 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 54.84 us, 0.05%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): BeitIntermediate(
          2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 1.47 ms, 1.24% latency,
29.68 TFLOPS,
          (dense): Linear(2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 748.16 us,
0.63% latency, 58.23 TFLOPS, in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
638.96 us, 0.54% latency, 0.0 FLOPS, )
        )
        (output): BeitOutput(
          2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 706.43 us, 0.60% latency,
61.66 TFLOPS,
          (dense): Linear(2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 586.75 us,
0.49% latency, 74.24 TFLOPS, in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 37.43 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, 1.38
ms, 1.16% latency, 12.83 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
        (drop_path): BeitDropPath(0, 0.00% Params, 0 MACs, 0.00% MACs, 58.17 us,
0.05% latency, 0.0 FLOPS, p=0.036363635212183)
        (layernorm_after): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.11
ms, 0.93% latency, 15.99 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
      )
      (5): BeitLayer(
        7.12 M, 8.28% Params, 66.71 GMACs, 8.31% MACs, 9.14 ms, 7.70% latency, 14.6
TFLOPS,
        (attention): BeitAttention(
          2.39 M, 2.78% Params, 23.15 GMACs, 2.88% MACs, 4.02 ms, 3.39% latency,
11.52 TFLOPS,
          (attention): BeitSelfAttention(
```

```
            1.8 M, 2.09% Params, 17.7 GMACs, 2.20% MACs, 3.24 ms, 2.73% latency,
10.94 TFLOPS,
            (query): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 581.26
us, 0.49% latency, 18.74 TFLOPS, in_features=768, out_features=768, bias=True)
            (key): Linear(589.82 k, 0.69% Params, 2.72 GMACs, 0.34% MACs, 158.79 us,
0.13% latency, 34.29 TFLOPS, in_features=768, out_features=768, bias=False)
            (value): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 559.09
us, 0.47% latency, 19.48 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 41.01 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
            (relative_position_bias): BeitRelativePositionBias(26.54 k, 0.03%
Params, 0 MACs, 0.00% MACs, 208.14 us, 0.18% latency, 0.0 FLOPS, )
          )
          (output): BeitSelfOutput(
            590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 703.81 us, 0.59%
latency, 15.47 TFLOPS,
            (dense): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 599.86
us, 0.51% latency, 18.16 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 33.14 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): BeitIntermediate(
          2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 1.34 ms, 1.13% latency,
32.45 TFLOPS,
          (dense): Linear(2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 594.62 us,
0.50% latency, 73.26 TFLOPS, in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
664.71 us, 0.56% latency, 0.0 FLOPS, )
        )
        (output): BeitOutput(
          2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 710.49 us, 0.60% latency,
61.31 TFLOPS,
          (dense): Linear(2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 590.09 us,
0.50% latency, 73.82 TFLOPS, in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 36.72 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, 1.07
ms, 0.91% latency, 16.5 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
        (drop_path): BeitDropPath(0, 0.00% Params, 0 MACs, 0.00% MACs, 56.27 us,
0.05% latency, 0.0 FLOPS, p=0.045454543083906174)
        (layernorm_after): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.09
ms, 0.92% latency, 16.31 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
      )
      (6): BeitLayer(
        7.12 M, 8.28% Params, 66.71 GMACs, 8.31% MACs, 11.09 ms, 9.34% latency,
12.04 TFLOPS,
        (attention): BeitAttention(
          2.39 M, 2.78% Params, 23.15 GMACs, 2.88% MACs, 3.87 ms, 3.26% latency,
```

```
11.98 TFLOPS,
          (attention): BeitSelfAttention(
            1.8 M, 2.09% Params, 17.7 GMACs, 2.20% MACs, 3.07 ms, 2.59% latency,
11.53 TFLOPS,
            (query): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 584.6
us, 0.49% latency, 18.63 TFLOPS, in_features=768, out_features=768, bias=True)
            (key): Linear(589.82 k, 0.69% Params, 2.72 GMACs, 0.34% MACs, 147.1 us,
0.12% latency, 37.02 TFLOPS, in_features=768, out_features=768, bias=False)
            (value): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 531.2
us, 0.45% latency, 20.5 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 41.96 us, 0.04%
latency, 0.0 FLOPS, p=0.0, inplace=False)
            (relative_position_bias): BeitRelativePositionBias(26.54 k, 0.03%
Params, 0 MACs, 0.00% MACs, 193.6 us, 0.16% latency, 0.0 FLOPS, )
          )
          (output): BeitSelfOutput(
            590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 717.88 us, 0.60%
latency, 15.17 TFLOPS,
            (dense): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 612.97
us, 0.52% latency, 17.77 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 33.62 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): BeitIntermediate(
          2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 2.3 ms, 1.93% latency,
18.97 TFLOPS,
          (dense): Linear(2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 928.16 us,
0.78% latency, 46.93 TFLOPS, in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
1.24 ms, 1.04% latency, 0.0 FLOPS, )
        )
        (output): BeitOutput(
          2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 1.13 ms, 0.95% latency,
38.48 TFLOPS,
          (dense): Linear(2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 950.57 us,
0.80% latency, 45.83 TFLOPS, in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 61.04 us, 0.05%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.06
ms, 0.89% latency, 16.74 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
        (drop_path): BeitDropPath(0, 0.00% Params, 0 MACs, 0.00% MACs, 73.43 us,
0.06% latency, 0.0 FLOPS, p=0.054545458406209946)
        (layernorm_after): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.47
ms, 1.24% latency, 12.08 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
      )
      (7): BeitLayer(
        7.12 M, 8.28% Params, 66.71 GMACs, 8.31% MACs, 11.9 ms, 10.03% latency,
11.21 TFLOPS,
```

```
        (attention): BeitAttention(
          2.39 M, 2.78% Params, 23.15 GMACs, 2.88% MACs, 5.82 ms, 4.90% latency,
7.96 TFLOPS,
          (attention): BeitSelfAttention(
            1.8 M, 2.09% Params, 17.7 GMACs, 2.20% MACs, 4.61 ms, 3.88% latency,
7.69 TFLOPS,
            (query): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 810.62
us, 0.68% latency, 13.43 TFLOPS, in_features=768, out_features=768, bias=True)
            (key): Linear(589.82 k, 0.69% Params, 2.72 GMACs, 0.34% MACs, 206.95 us,
0.17% latency, 26.31 TFLOPS, in_features=768, out_features=768, bias=False)
            (value): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 781.06
us, 0.66% latency, 13.94 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 62.7 us, 0.05%
latency, 0.0 FLOPS, p=0.0, inplace=False)
            (relative_position_bias): BeitRelativePositionBias(26.54 k, 0.03%
Params, 0 MACs, 0.00% MACs, 303.03 us, 0.26% latency, 0.0 FLOPS, )
          )
          (output): BeitSelfOutput(
            590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 1.11 ms, 0.93% latency,
9.83 TFLOPS,
            (dense): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 982.28
us, 0.83% latency, 11.09 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 34.57 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): BeitIntermediate(
          2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 1.36 ms, 1.15% latency,
31.94 TFLOPS,
          (dense): Linear(2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 603.44 us,
0.51% latency, 72.19 TFLOPS, in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
671.15 us, 0.57% latency, 0.0 FLOPS, )
        )
        (output): BeitOutput(
          2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 715.26 us, 0.60% latency,
60.9 TFLOPS,
          (dense): Linear(2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 575.78 us,
0.49% latency, 75.66 TFLOPS, in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 63.66 us, 0.05%
latency, 0.0 FLOPS, p=0.0, inplace=False)
        )
        (layernorm_before): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.93
ms, 1.63% latency, 9.17 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
        (drop_path): BeitDropPath(0, 0.00% Params, 0 MACs, 0.00% MACs, 59.37 us,
0.05% latency, 0.0 FLOPS, p=0.06363636255264282)
        (layernorm_after): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.12
ms, 0.95% latency, 15.78 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
      )
      (8): BeitLayer(
```

```
        7.12 M, 8.28% Params, 66.71 GMACs, 8.31% MACs, 9.55 ms, 8.05% latency, 13.98
TFLOPS,
        (attention): BeitAttention(
          2.39 M, 2.78% Params, 23.15 GMACs, 2.88% MACs, 4.23 ms, 3.56% latency,
10.95 TFLOPS,
          (attention): BeitSelfAttention(
            1.8 M, 2.09% Params, 17.7 GMACs, 2.20% MACs, 3.44 ms, 2.90% latency,
10.3 TFLOPS,
            (query): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 630.62
us, 0.53% latency, 17.27 TFLOPS, in_features=768, out_features=768, bias=True)
            (key): Linear(589.82 k, 0.69% Params, 2.72 GMACs, 0.34% MACs, 151.63 us,
0.13% latency, 35.91 TFLOPS, in_features=768, out_features=768, bias=False)
            (value): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 578.17
us, 0.49% latency, 18.84 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 43.63 us, 0.04%
latency, 0.0 FLOPS, p=0.0, inplace=False)
            (relative_position_bias): BeitRelativePositionBias(26.54 k, 0.03%
Params, 0 MACs, 0.00% MACs, 270.61 us, 0.23% latency, 0.0 FLOPS, )
          )
          (output): BeitSelfOutput(
            590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 706.91 us, 0.60%
latency, 15.41 TFLOPS,
            (dense): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 595.81
us, 0.50% latency, 18.28 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 35.05 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): BeitIntermediate(
          2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 1.39 ms, 1.17% latency,
31.38 TFLOPS,
          (dense): Linear(2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 653.98 us,
0.55% latency, 66.61 TFLOPS, in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
653.03 us, 0.55% latency, 0.0 FLOPS, )
        )
        (output): BeitOutput(
          2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 697.61 us, 0.59% latency,
62.44 TFLOPS,
          (dense): Linear(2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 579.36 us,
0.49% latency, 75.19 TFLOPS, in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 35.52 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, 1.11
ms, 0.93% latency, 15.98 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
        (drop_path): BeitDropPath(0, 0.00% Params, 0 MACs, 0.00% MACs, 57.46 us,
0.05% latency, 0.0 FLOPS, p=0.0727272778749466)
        (layernorm_after): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.16
ms, 0.97% latency, 15.34 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
```

```
      )
      (9): BeitLayer(
        7.12 M, 8.28% Params, 66.71 GMACs, 8.31% MACs, 8.93 ms, 7.52% latency, 14.95
TFLOPS,
        (attention): BeitAttention(
          2.39 M, 2.78% Params, 23.15 GMACs, 2.88% MACs, 3.97 ms, 3.34% latency,
11.67 TFLOPS,
          (attention): BeitSelfAttention(
            1.8 M, 2.09% Params, 17.7 GMACs, 2.20% MACs, 3.18 ms, 2.68% latency,
11.14 TFLOPS,
            (query): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 551.94
us, 0.47% latency, 19.73 TFLOPS, in_features=768, out_features=768, bias=True)
            (key): Linear(589.82 k, 0.69% Params, 2.72 GMACs, 0.34% MACs, 157.36 us,
0.13% latency, 34.6 TFLOPS, in_features=768, out_features=768, bias=False)
            (value): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 551.46
us, 0.46% latency, 19.75 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 43.39 us, 0.04%
latency, 0.0 FLOPS, p=0.0, inplace=False)
            (relative_position_bias): BeitRelativePositionBias(26.54 k, 0.03%
Params, 0 MACs, 0.00% MACs, 208.62 us, 0.18% latency, 0.0 FLOPS, )
          )
          (output): BeitSelfOutput(
            590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 709.53 us, 0.60%
latency, 15.35 TFLOPS,
            (dense): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 602.25
us, 0.51% latency, 18.08 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 34.57 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): BeitIntermediate(
          2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 1.22 ms, 1.03% latency,
35.69 TFLOPS,
          (dense): Linear(2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 596.28 us,
0.50% latency, 73.06 TFLOPS, in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
550.99 us, 0.46% latency, 0.0 FLOPS, )
        )
        (output): BeitOutput(
          2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 700.0 us, 0.59% latency,
62.23 TFLOPS,
          (dense): Linear(2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 580.79 us,
0.49% latency, 75.01 TFLOPS, in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 33.86 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, 1.05
ms, 0.89% latency, 16.85 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
        (drop_path): BeitDropPath(0, 0.00% Params, 0 MACs, 0.00% MACs, 56.51 us,
0.05% latency, 0.0 FLOPS, p=0.08181818574666977)
```

```
        (layernorm_after): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.09
ms, 0.92% latency, 16.19 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
      )
      (10): BeitLayer(
        7.12 M, 8.28% Params, 66.71 GMACs, 8.31% MACs, 9.5 ms, 8.01% latency, 14.04
TFLOPS,
        (attention): BeitAttention(
          2.39 M, 2.78% Params, 23.15 GMACs, 2.88% MACs, 4.38 ms, 3.69% latency,
10.57 TFLOPS,
          (attention): BeitSelfAttention(
            1.8 M, 2.09% Params, 17.7 GMACs, 2.20% MACs, 3.6 ms, 3.04% latency, 9.84
TFLOPS,
            (query): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 893.59
us, 0.75% latency, 12.19 TFLOPS, in_features=768, out_features=768, bias=True)
            (key): Linear(589.82 k, 0.69% Params, 2.72 GMACs, 0.34% MACs, 154.73 us,
0.13% latency, 35.19 TFLOPS, in_features=768, out_features=768, bias=False)
            (value): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 558.85
us, 0.47% latency, 19.49 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 43.39 us, 0.04%
latency, 0.0 FLOPS, p=0.0, inplace=False)
            (relative_position_bias): BeitRelativePositionBias(26.54 k, 0.03%
Params, 0 MACs, 0.00% MACs, 200.99 us, 0.17% latency, 0.0 FLOPS, )
          )
          (output): BeitSelfOutput(
            590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 700.71 us, 0.59%
latency, 15.54 TFLOPS,
            (dense): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 593.66
us, 0.50% latency, 18.34 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 34.09 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): BeitIntermediate(
          2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 1.27 ms, 1.07% latency,
34.36 TFLOPS,
          (dense): Linear(2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 567.44 us,
0.48% latency, 76.77 TFLOPS, in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
622.99 us, 0.52% latency, 0.0 FLOPS, )
        )
        (output): BeitOutput(
          2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 724.08 us, 0.61% latency,
60.16 TFLOPS,
          (dense): Linear(2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 595.81 us,
0.50% latency, 73.11 TFLOPS, in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 37.67 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, 1.07
ms, 0.90% latency, 16.61 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
```

```
        (drop_path): BeitDropPath(0, 0.00% Params, 0 MACs, 0.00% MACs, 57.22 us,
0.05% latency, 0.0 FLOPS, p=0.09090909361839294)
        (layernorm_after): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.14
ms, 0.96% latency, 15.49 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
      )
      (11): BeitLayer(
        7.12 M, 8.28% Params, 66.71 GMACs, 8.31% MACs, 9.0 ms, 7.58% latency, 14.83
TFLOPS,
        (attention): BeitAttention(
          2.39 M, 2.78% Params, 23.15 GMACs, 2.88% MACs, 3.97 ms, 3.34% latency,
11.67 TFLOPS,
          (attention): BeitSelfAttention(
            1.8 M, 2.09% Params, 17.7 GMACs, 2.20% MACs, 3.25 ms, 2.74% latency,
10.91 TFLOPS,
            (query): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 639.2
us, 0.54% latency, 17.04 TFLOPS, in_features=768, out_features=768, bias=True)
            (key): Linear(589.82 k, 0.69% Params, 2.72 GMACs, 0.34% MACs, 151.16 us,
0.13% latency, 36.02 TFLOPS, in_features=768, out_features=768, bias=False)
            (value): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 577.93
us, 0.49% latency, 18.84 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 43.87 us, 0.04%
latency, 0.0 FLOPS, p=0.0, inplace=False)
            (relative_position_bias): BeitRelativePositionBias(26.54 k, 0.03%
Params, 0 MACs, 0.00% MACs, 204.56 us, 0.17% latency, 0.0 FLOPS, )
          )
          (output): BeitSelfOutput(
            590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 640.63 us, 0.54%
latency, 17.0 TFLOPS,
            (dense): Linear(590.59 k, 0.69% Params, 5.45 GMACs, 0.68% MACs, 537.4
us, 0.45% latency, 20.27 TFLOPS, in_features=768, out_features=768, bias=True)
            (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 32.42 us, 0.03%
latency, 0.0 FLOPS, p=0.0, inplace=False)
          )
        )
        (intermediate): BeitIntermediate(
          2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 1.29 ms, 1.09% latency,
33.65 TFLOPS,
          (dense): Linear(2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 546.46 us,
0.46% latency, 79.72 TFLOPS, in_features=768, out_features=3072, bias=True)
          (intermediate_act_fn): GELUActivation(0, 0.00% Params, 0 MACs, 0.00% MACs,
669.0 us, 0.56% latency, 0.0 FLOPS, )
        )
        (output): BeitOutput(
          2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 677.82 us, 0.57% latency,
64.27 TFLOPS,
          (dense): Linear(2.36 M, 2.75% Params, 21.78 GMACs, 2.71% MACs, 576.73 us,
0.49% latency, 75.53 TFLOPS, in_features=3072, out_features=768, bias=True)
          (dropout): Dropout(0, 0.00% Params, 0 MACs, 0.00% MACs, 32.9 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, 1.11
ms, 0.93% latency, 15.98 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
        (drop_path): BeitDropPath(0, 0.00% Params, 0 MACs, 0.00% MACs, 53.17 us,
0.04% latency, 0.0 FLOPS, p=0.10000000149011612)
        (layernorm_after): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.07
ms, 0.90% latency, 16.55 GFLOPS, (768,), eps=1e-12, elementwise_affine=True)
      )
    )
  )
  (layernorm): Identity(0, 0.00% Params, 0 MACs, 0.00% MACs, 21.22 us, 0.02%
latency, 0.0 FLOPS, )
  (pooler): BeitPooler(
    1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.36 ms, 1.14% latency, 22.66 MFLOPS,
    (layernorm): LayerNorm(1.54 k, 0.00% Params, 0 MACs, 0.00% MACs, 1.03 ms, 0.87%
latency, 29.84 MFLOPS, (768,), eps=1e-12, elementwise_affine=True)
  )
)
--------------------------------------------------------------------------------
```