```
------------------------- 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:                                                 1.82 M
params of model = params per GPU * mp_size:                     1.82 M
fwd MACs per GPU:                                               2.6 GMACs
fwd flops per GPU:                                              5.25 G
fwd flops of model = fwd flops per GPU * mp_size:               5.25 G
fwd latency:                                                    228.3 ms
fwd FLOPS per GPU = fwd flops per GPU / fwd latency:            23.02 GFLOPS

----------------------------- Aggregated Profile per GPU
-----------------------------
Top 1 modules in terms of params, MACs or fwd latency at different model depths:
depth 0:
    params      - {'MobileNetV1Model': '1.82 M'}
    MACs        - {'MobileNetV1Model': '2.6 GMACs'}
    fwd latency - {'MobileNetV1Model': '228.3 ms'}
depth 1:
    params      - {'ModuleList': '1.82 M'}
    MACs        - {'ModuleList': '2.53 GMACs'}
    fwd latency - {'ModuleList': '218.77 ms'}
depth 2:
    params      - {'MobileNetV1ConvLayer': '1.82 M'}
    MACs        - {'MobileNetV1ConvLayer': '2.53 GMACs'}
    fwd latency - {'MobileNetV1ConvLayer': '218.77 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.

```
MobileNetV1Model(
  1.82 M, 100.00% Params, 2.6 GMACs, 100.00% MACs, 228.3 ms, 100.00% latency, 23.02
GFLOPS,
  (conv_stem): MobileNetV1ConvLayer(
    696, 0.04% Params, 65.03 MMACs, 2.50% MACs, 7.49 ms, 3.28% latency, 18.0 GFLOPS,

    (convolution): Conv2d(648, 0.04% Params, 65.03 MMACs, 2.50% MACs, 298.5 us,
0.13% latency, 435.7 GFLOPS, 3, 24, kernel_size=(3, 3), stride=(2, 2), bias=False)
    (normalization): BatchNorm2d(48, 0.00% Params, 0 MACs, 0.00% MACs, 4.86 ms,
2.13% latency, 990.51 MFLOPS, 24, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
    (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 1.96 ms, 0.86% latency,
0.0 FLOPS, )
  )
  (layer): ModuleList(
    (0): MobileNetV1ConvLayer(
      264, 0.01% Params, 21.68 MMACs, 0.83% MACs, 9.39 ms, 4.11% latency, 5.13
GFLOPS,
      (convolution): Conv2d(216, 0.01% Params, 21.68 MMACs, 0.83% MACs, 276.8 us,
0.12% latency, 156.62 GFLOPS, 24, 24, kernel_size=(3, 3), stride=(1, 1), groups=24,
bias=False)
      (normalization): BatchNorm2d(48, 0.00% Params, 0 MACs, 0.00% MACs, 6.84 ms,
3.00% latency, 703.96 MFLOPS, 24, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 1.94 ms, 0.85%
latency, 0.0 FLOPS, )
    )
    (1): MobileNetV1ConvLayer(
      1.25 k, 0.07% Params, 115.61 MMACs, 4.45% MACs, 9.98 ms, 4.37% latency, 24.13
GFLOPS,
      (convolution): Conv2d(1.15 k, 0.06% Params, 115.61 MMACs, 4.45% MACs, 245.57
us, 0.11% latency, 941.52 GFLOPS, 24, 48, kernel_size=(1, 1), stride=(1, 1),
bias=False)
      (normalization): BatchNorm2d(96, 0.01% Params, 0 MACs, 0.00% MACs, 7.57 ms,
3.32% latency, 1.27 GFLOPS, 48, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 1.93 ms, 0.85%
latency, 0.0 FLOPS, )
    )
    (2): MobileNetV1ConvLayer(
      528, 0.03% Params, 10.84 MMACs, 0.42% MACs, 4.04 ms, 1.77% latency, 5.97
GFLOPS,
      (convolution): Conv2d(432, 0.02% Params, 10.84 MMACs, 0.42% MACs, 266.08 us,
0.12% latency, 81.47 GFLOPS, 48, 48, kernel_size=(3, 3), stride=(2, 2), groups=48,
bias=False)
      (normalization): BatchNorm2d(96, 0.01% Params, 0 MACs, 0.00% MACs, 1.65 ms,
0.72% latency, 1.46 GFLOPS, 48, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
```

```
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 1.8 ms, 0.79%
latency, 0.0 FLOPS, )
    )
    (3): MobileNetV1ConvLayer(
      4.8 k, 0.26% Params, 115.61 MMACs, 4.45% MACs, 4.05 ms, 1.77% latency, 58.27
GFLOPS,
      (convolution): Conv2d(4.61 k, 0.25% Params, 115.61 MMACs, 4.45% MACs, 228.64
us, 0.10% latency, 1.01 TFLOPS, 48, 96, kernel_size=(1, 1), stride=(1, 1),
bias=False)
      (normalization): BatchNorm2d(192, 0.01% Params, 0 MACs, 0.00% MACs, 1.67 ms,
0.73% latency, 2.89 GFLOPS, 96, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 1.93 ms, 0.85%
latency, 0.0 FLOPS, )
    )
    (4): MobileNetV1ConvLayer(
      1.06 k, 0.06% Params, 21.68 MMACs, 0.83% MACs, 4.22 ms, 1.85% latency, 11.42
GFLOPS,
      (convolution): Conv2d(864, 0.05% Params, 21.68 MMACs, 0.83% MACs, 236.51 us,
0.10% latency, 183.3 GFLOPS, 96, 96, kernel_size=(3, 3), stride=(1, 1), groups=96,
bias=False)
      (normalization): BatchNorm2d(192, 0.01% Params, 0 MACs, 0.00% MACs, 1.76 ms,
0.77% latency, 2.74 GFLOPS, 96, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 1.89 ms, 0.83%
latency, 0.0 FLOPS, )
    )
    (5): MobileNetV1ConvLayer(
      9.41 k, 0.52% Params, 231.21 MMACs, 8.90% MACs, 11.19 ms, 4.90% latency, 41.74
GFLOPS,
      (convolution): Conv2d(9.22 k, 0.51% Params, 231.21 MMACs, 8.90% MACs, 7.28 ms,
3.19% latency, 63.55 GFLOPS, 96, 96, kernel_size=(1, 1), stride=(1, 1), bias=False)
      (normalization): BatchNorm2d(192, 0.01% Params, 0 MACs, 0.00% MACs, 1.69 ms,
0.74% latency, 2.85 GFLOPS, 96, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 1.96 ms, 0.86%
latency, 0.0 FLOPS, )
    )
    (6): MobileNetV1ConvLayer(
      1.06 k, 0.06% Params, 5.42 MMACs, 0.21% MACs, 11.94 ms, 5.23% latency, 1.01
GFLOPS,
      (convolution): Conv2d(864, 0.05% Params, 5.42 MMACs, 0.21% MACs, 238.18 us,
0.10% latency, 45.5 GFLOPS, 96, 96, kernel_size=(3, 3), stride=(2, 2), groups=96,
bias=False)
      (normalization): BatchNorm2d(192, 0.01% Params, 0 MACs, 0.00% MACs, 8.66 ms,
3.79% latency, 139.11 MFLOPS, 96, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 2.67 ms, 1.17%
latency, 0.0 FLOPS, )
    )
```

```
    (7): MobileNetV1ConvLayer(
      18.82 k, 1.04% Params, 115.61 MMACs, 4.45% MACs, 4.1 ms, 1.80% latency, 56.94
GFLOPS,
      (convolution): Conv2d(18.43 k, 1.01% Params, 115.61 MMACs, 4.45% MACs, 264.64
us, 0.12% latency, 873.67 GFLOPS, 96, 192, kernel_size=(1, 1), stride=(1, 1),
bias=False)
      (normalization): BatchNorm2d(384, 0.02% Params, 0 MACs, 0.00% MACs, 1.7 ms,
0.74% latency, 1.42 GFLOPS, 192, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 1.9 ms, 0.83%
latency, 0.0 FLOPS, )
    )
    (8): MobileNetV1ConvLayer(
      2.11 k, 0.12% Params, 10.84 MMACs, 0.42% MACs, 11.44 ms, 5.01% latency, 2.11
GFLOPS,
      (convolution): Conv2d(1.73 k, 0.10% Params, 10.84 MMACs, 0.42% MACs, 234.6 us,
0.10% latency, 92.39 GFLOPS, 192, 192, kernel_size=(3, 3), stride=(1, 1),
groups=192, bias=False)
      (normalization): BatchNorm2d(384, 0.02% Params, 0 MACs, 0.00% MACs, 8.87 ms,
3.88% latency, 271.55 MFLOPS, 192, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 2.0 ms, 0.88%
latency, 0.0 FLOPS, )
    )
    (9): MobileNetV1ConvLayer(
      37.25 k, 2.05% Params, 231.21 MMACs, 8.90% MACs, 11.3 ms, 4.95% latency, 41.12
GFLOPS,
      (convolution): Conv2d(36.86 k, 2.03% Params, 231.21 MMACs, 8.90% MACs, 261.55
us, 0.11% latency, 1.77 TFLOPS, 192, 192, kernel_size=(1, 1), stride=(1, 1),
bias=False)
      (normalization): BatchNorm2d(384, 0.02% Params, 0 MACs, 0.00% MACs, 2.79 ms,
1.22% latency, 864.51 MFLOPS, 192, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 8.02 ms, 3.51%
latency, 0.0 FLOPS, )
    )
    (10): MobileNetV1ConvLayer(
      2.11 k, 0.12% Params, 2.71 MMACs, 0.10% MACs, 7.03 ms, 3.08% latency, 857.04
MFLOPS,
      (convolution): Conv2d(1.73 k, 0.10% Params, 2.71 MMACs, 0.10% MACs, 206.47 us,
0.09% latency, 26.25 GFLOPS, 192, 192, kernel_size=(3, 3), stride=(2, 2),
groups=192, bias=False)
      (normalization): BatchNorm2d(384, 0.02% Params, 0 MACs, 0.00% MACs, 1.39 ms,
0.61% latency, 432.81 MFLOPS, 192, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 5.11 ms, 2.24%
latency, 0.0 FLOPS, )
    )
    (11): MobileNetV1ConvLayer(
      74.5 k, 4.10% Params, 115.61 MMACs, 4.45% MACs, 4.32 ms, 1.89% latency, 53.76
```

```
GFLOPS,
      (convolution): Conv2d(73.73 k, 4.06% Params, 115.61 MMACs, 4.45% MACs, 243.43
us, 0.11% latency, 949.82 GFLOPS, 192, 384, kernel_size=(1, 1), stride=(1, 1),
bias=False)
      (normalization): BatchNorm2d(768, 0.04% Params, 0 MACs, 0.00% MACs, 1.9 ms,
0.83% latency, 635.09 MFLOPS, 384, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 1.93 ms, 0.84%
latency, 0.0 FLOPS, )
    )
    (12): MobileNetV1ConvLayer(
      4.22 k, 0.23% Params, 5.42 MMACs, 0.21% MACs, 9.9 ms, 4.34% latency, 1.22
GFLOPS,
      (convolution): Conv2d(3.46 k, 0.19% Params, 5.42 MMACs, 0.21% MACs, 243.9 us,
0.11% latency, 44.44 GFLOPS, 384, 384, kernel_size=(3, 3), stride=(1, 1),
groups=384, bias=False)
      (normalization): BatchNorm2d(768, 0.04% Params, 0 MACs, 0.00% MACs, 7.32 ms,
3.21% latency, 164.55 MFLOPS, 384, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 1.99 ms, 0.87%
latency, 0.0 FLOPS, )
    )
    (13): MobileNetV1ConvLayer(
      148.22 k, 8.16% Params, 231.21 MMACs, 8.90% MACs, 11.22 ms, 4.91% latency,
41.34 GFLOPS,
      (convolution): Conv2d(147.46 k, 8.12% Params, 231.21 MMACs, 8.90% MACs, 269.17
us, 0.12% latency, 1.72 TFLOPS, 384, 384, kernel_size=(1, 1), stride=(1, 1),
bias=False)
      (normalization): BatchNorm2d(768, 0.04% Params, 0 MACs, 0.00% MACs, 1.68 ms,
0.73% latency, 717.76 MFLOPS, 384, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 9.03 ms, 3.96%
latency, 0.0 FLOPS, )
    )
    (14): MobileNetV1ConvLayer(
      4.22 k, 0.23% Params, 5.42 MMACs, 0.21% MACs, 4.25 ms, 1.86% latency, 2.83
GFLOPS,
      (convolution): Conv2d(3.46 k, 0.19% Params, 5.42 MMACs, 0.21% MACs, 247.48 us,
0.11% latency, 43.79 GFLOPS, 384, 384, kernel_size=(3, 3), stride=(1, 1),
groups=384, bias=False)
      (normalization): BatchNorm2d(768, 0.04% Params, 0 MACs, 0.00% MACs, 1.73 ms,
0.76% latency, 694.09 MFLOPS, 384, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 1.89 ms, 0.83%
latency, 0.0 FLOPS, )
    )
    (15): MobileNetV1ConvLayer(
      148.22 k, 8.16% Params, 231.21 MMACs, 8.90% MACs, 13.46 ms, 5.89% latency,
34.45 GFLOPS,
      (convolution): Conv2d(147.46 k, 8.12% Params, 231.21 MMACs, 8.90% MACs, 7.32
```

```
ms, 3.21% latency, 63.15 GFLOPS, 384, 384, kernel_size=(1, 1), stride=(1, 1),
bias=False)
      (normalization): BatchNorm2d(768, 0.04% Params, 0 MACs, 0.00% MACs, 1.84 ms,
0.80% latency, 655.96 MFLOPS, 384, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 4.06 ms, 1.78%
latency, 0.0 FLOPS, )
    )
    (16): MobileNetV1ConvLayer(
      4.22 k, 0.23% Params, 5.42 MMACs, 0.21% MACs, 11.48 ms, 5.03% latency, 1.05
GFLOPS,
      (convolution): Conv2d(3.46 k, 0.19% Params, 5.42 MMACs, 0.21% MACs, 252.01 us,
0.11% latency, 43.01 GFLOPS, 384, 384, kernel_size=(3, 3), stride=(1, 1),
groups=384, bias=False)
      (normalization): BatchNorm2d(768, 0.04% Params, 0 MACs, 0.00% MACs, 8.87 ms,
3.89% latency, 135.74 MFLOPS, 384, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 2.01 ms, 0.88%
latency, 0.0 FLOPS, )
    )
    (17): MobileNetV1ConvLayer(
      148.22 k, 8.16% Params, 231.21 MMACs, 8.90% MACs, 4.09 ms, 1.79% latency,
113.39 GFLOPS,
      (convolution): Conv2d(147.46 k, 8.12% Params, 231.21 MMACs, 8.90% MACs, 270.13
us, 0.12% latency, 1.71 TFLOPS, 384, 384, kernel_size=(1, 1), stride=(1, 1),
bias=False)
      (normalization): BatchNorm2d(768, 0.04% Params, 0 MACs, 0.00% MACs, 1.65 ms,
0.72% latency, 730.64 MFLOPS, 384, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 1.95 ms, 0.85%
latency, 0.0 FLOPS, )
    )
    (18): MobileNetV1ConvLayer(
      4.22 k, 0.23% Params, 5.42 MMACs, 0.21% MACs, 11.68 ms, 5.12% latency, 1.03
GFLOPS,
      (convolution): Conv2d(3.46 k, 0.19% Params, 5.42 MMACs, 0.21% MACs, 407.46 us,
0.18% latency, 26.6 GFLOPS, 384, 384, kernel_size=(3, 3), stride=(1, 1), groups=384,
bias=False)
      (normalization): BatchNorm2d(768, 0.04% Params, 0 MACs, 0.00% MACs, 1.82 ms,
0.80% latency, 661.98 MFLOPS, 384, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 2.05 ms, 0.90%
latency, 0.0 FLOPS, )
    )
    (19): MobileNetV1ConvLayer(
      148.22 k, 8.16% Params, 231.21 MMACs, 8.90% MACs, 4.23 ms, 1.85% latency,
109.7 GFLOPS,
      (convolution): Conv2d(147.46 k, 8.12% Params, 231.21 MMACs, 8.90% MACs, 249.62
us, 0.11% latency, 1.85 TFLOPS, 384, 384, kernel_size=(1, 1), stride=(1, 1),
bias=False)
```

```
      (normalization): BatchNorm2d(768, 0.04% Params, 0 MACs, 0.00% MACs, 1.66 ms,
0.73% latency, 725.39 MFLOPS, 384, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 2.08 ms, 0.91%
latency, 0.0 FLOPS, )
    )
    (20): MobileNetV1ConvLayer(
      4.22 k, 0.23% Params, 5.42 MMACs, 0.21% MACs, 15.53 ms, 6.80% latency, 775.4
MFLOPS,
      (convolution): Conv2d(3.46 k, 0.19% Params, 5.42 MMACs, 0.21% MACs, 239.85 us,
0.11% latency, 45.19 GFLOPS, 384, 384, kernel_size=(3, 3), stride=(1, 1),
groups=384, bias=False)
      (normalization): BatchNorm2d(768, 0.04% Params, 0 MACs, 0.00% MACs, 9.62 ms,
4.21% latency, 125.16 MFLOPS, 384, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 5.32 ms, 2.33%
latency, 0.0 FLOPS, )
    )
    (21): MobileNetV1ConvLayer(
      148.22 k, 8.16% Params, 231.21 MMACs, 8.90% MACs, 4.05 ms, 1.77% latency,
114.5 GFLOPS,
      (convolution): Conv2d(147.46 k, 8.12% Params, 231.21 MMACs, 8.90% MACs, 237.46
us, 0.10% latency, 1.95 TFLOPS, 384, 384, kernel_size=(1, 1), stride=(1, 1),
bias=False)
      (normalization): BatchNorm2d(768, 0.04% Params, 0 MACs, 0.00% MACs, 1.66 ms,
0.73% latency, 726.75 MFLOPS, 384, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 1.88 ms, 0.82%
latency, 0.0 FLOPS, )
    )
    (22): MobileNetV1ConvLayer(
      4.22 k, 0.23% Params, 1.35 MMACs, 0.05% MACs, 7.15 ms, 3.13% latency, 421.09
MFLOPS,
      (convolution): Conv2d(3.46 k, 0.19% Params, 1.35 MMACs, 0.05% MACs, 3.27 ms,
1.43% latency, 828.01 MFLOPS, 384, 384, kernel_size=(3, 3), stride=(2, 2),
groups=384, bias=False)
      (normalization): BatchNorm2d(768, 0.04% Params, 0 MACs, 0.00% MACs, 1.65 ms,
0.72% latency, 182.55 MFLOPS, 384, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 1.9 ms, 0.83%
latency, 0.0 FLOPS, )
    )
    (23): MobileNetV1ConvLayer(
      296.45 k, 16.32% Params, 115.61 MMACs, 4.45% MACs, 10.14 ms, 4.44% latency,
22.85 GFLOPS,
      (convolution): Conv2d(294.91 k, 16.23% Params, 115.61 MMACs, 4.45% MACs,
275.37 us, 0.12% latency, 839.63 GFLOPS, 384, 768, kernel_size=(1, 1), stride=(1,
1), bias=False)
      (normalization): BatchNorm2d(1.54 k, 0.08% Params, 0 MACs, 0.00% MACs, 1.69
ms, 0.74% latency, 355.4 MFLOPS, 768, eps=0.001, momentum=0.9997, affine=True,
```

```
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 7.95 ms, 3.48%
latency, 0.0 FLOPS, )
    )
    (24): MobileNetV1ConvLayer(
      8.45 k, 0.47% Params, 2.71 MMACs, 0.10% MACs, 9.23 ms, 4.04% latency, 652.4
MFLOPS,
      (convolution): Conv2d(6.91 k, 0.38% Params, 2.71 MMACs, 0.10% MACs, 241.28 us,
0.11% latency, 22.46 GFLOPS, 768, 768, kernel_size=(3, 3), stride=(1, 1),
groups=768, bias=False)
      (normalization): BatchNorm2d(1.54 k, 0.08% Params, 0 MACs, 0.00% MACs, 1.67
ms, 0.73% latency, 360.31 MFLOPS, 768, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 6.99 ms, 3.06%
latency, 0.0 FLOPS, )
    )
    (25): MobileNetV1ConvLayer(
      591.36 k, 32.55% Params, 231.21 MMACs, 8.90% MACs, 9.36 ms, 4.10% latency,
49.45 GFLOPS,
      (convolution): Conv2d(589.82 k, 32.47% Params, 231.21 MMACs, 8.90% MACs,
242.47 us, 0.11% latency, 1.91 TFLOPS, 768, 768, kernel_size=(1, 1), stride=(1, 1),
bias=False)
      (normalization): BatchNorm2d(1.54 k, 0.08% Params, 0 MACs, 0.00% MACs, 1.68
ms, 0.74% latency, 357.56 MFLOPS, 768, eps=0.001, momentum=0.9997, affine=True,
track_running_stats=True)
      (activation): ReLU6(0, 0.00% Params, 0 MACs, 0.00% MACs, 7.2 ms, 3.15%
latency, 0.0 FLOPS, )
    )
  )
  (pooler): AdaptiveAvgPool2d(0, 0.00% Params, 0 MACs, 0.00% MACs, 623.94 us, 0.27%
latency, 482.51 MFLOPS, output_size=(1, 1))
)
--------------------------------------------------------------------------------
```