**Per layer latency and bottleneck analysis full pytorch**

Model: Resnet20

Dataset: CIFAR10

Batch size = 32

Time: In Seconds

Set up: 100 training iterations on Jetson Xavier, TensorRT to do conv2d operations and all other operations done on CPU.

**Layer 1**

Input size: torch.Size([32, 3, 224, 224])

Stride, padding= (1, 1) (1, 1)

Weight size= torch.Size([32, 3, 3, 3])

Output size =torch.Size([32, 16, 224, 224])

Total: 27.585

| Operations | Latency |
| --- | --- |
| 1. Forward    1. Load input and Weight refitting to GPU    2. Compute    3. Load\_output to CPU | 2.740  0.319  0.044  2.368 |
| 1. Back Prop    1. Input gradient       1. Load input and Weight refitting to GPU       2. Compute       3. Load\_output to CPU    2. weight gradient       1. Load input and Weight refitting to GPU       2. Compute       3. Load\_output to CPU | 14.189  2.6390  1.438  0.040  1.139  11.443  4.423  0.033  5.259 |

**Layer 2-6**

Input size: torch.Size([32, 16, 32, 32])

Stride, padding= (1, 1) (1, 1)

Weight size= torch.Size([16, 16, 3, 3])

Output size =torch.Size([32, 16, 32, 32])

Total = 4.783

| Operations | Latency |
| --- | --- |
| 1. Forward    1. Load input and Weight refitting to GPU    2. Compute    3. Load\_output to CPU | 0.458  0.120  0.050  0.279 |
| 1. Back Prop    1. Input gradient       1. Load input and Weight refitting to GPU       2. Compute       3. Load\_output to CPU    2. weight gradient       1. Load input and Weight refitting to GPU       2. Compute       3. Load\_output to CPU | 2.821  0.557  0.157  0.081  0.291  2.211  0.844  0.037  0.694 |

**Layer 7**

Input size= torch.Size([32, 16, 32, 32])

stride , padding= (2, 2) (1, 1)

Weight size = torch.Size([32, 16, 3, 3])

Output size= torch.Size([32, 32, 16, 16])

Total : 4.325

| Operations | Latency |
| --- | --- |
| 1. Forward    1. Load input and Weight refitting to GPU    2. Compute    3. Load\_output to CPU | 0.506  0.134  0.048  0.324 |
| 1. Back Prop    1. Input gradient       1. Load input and Weight refitting to GPU       2. Compute       3. Load\_output to CPU    2. weight gradient       1. Load input and Weight refitting to GPU       2. Compute       3. Load\_output to CPU | 2.144  0.674  0.134  0.094  0.413  1.452  0.681  0.041  0.471 |

**Layer 8-12**

Input size = torch.Size([32, 32, 16, 16])

stride ,, padding = (1, 1) (1, 1)

Weight size = torch.Size([32, 32, 3, 3])

Output size = torch.Size([32, 32, 16, 16])

Total: 3.342

| Operations | Latency |
| --- | --- |
| 1. Forward    1. Load input and Weight refitting to GPU    2. Compute    3. Load\_output to CPU | 0.250  0.110  0.050  0.090 |
| 1. Back Prop    1. Input gradient       1. Load input and Weight refitting to GPU       2. Compute       3. Load\_output to CPU    2. weight gradient       1. Load input and Weight refitting to GPU       2. Compute       3. Load\_output to CPU | 2.125  0.267  0.114  0.046  0.087  1.808  0.725  0.024  0.545 |

**Layer 13**

Input size=torch.Size([32, 32, 16, 16])

stride , padding = (2, 2) (1, 1)

torch.Size([64, 32, 3, 3])

torch.Size([32, 64, 8, 8])

Total : 3.343

| Operations | Latency |
| --- | --- |
| 1. Forward    1. Load input and Weight refitting to GPU    2. Compute    3. Load\_output to CPU | 0.445  0.128  0.072  0.245 |
| 1. Back Prop    1. Input gradient       1. Load input and Weight refitting to GPU       2. Compute       3. Load\_output to CPU    2. weight gradient       1. Load input and Weight refitting to GPU       2. Compute       3. Load\_output to CPU | 1.793  0.467  0.110  0.074  0.258  1.311  0.625  0.030  0.411 |

**Layer 14-18**

Input size= torch.Size([32, 64, 8, 8])

stride , padding = (1, 1) (1, 1)

Weight size = torch.Size([64, 64, 3, 3])

Output size = torch.Size([32, 64, 8, 8])

Total = 3.162

| Operations | Latency |
| --- | --- |
| 1. Forward    1. Load input and Weight refitting to GPU    2. Compute    3. Load\_output to CPU | 0.234  0.112  0.061  0.061 |
| 1. Back Prop    1. Input gradient       1. Load input and Weight refitting to GPU       2. Compute       3. Load\_output to CPU    2. weight gradient       1. Load input and Weight refitting to GPU       2. Compute       3. Load\_output to CPU | 2.146  0.295  0.095  0.036  0.147  1.802  0.823  0.027  0.502 |

**Vertical Axis: time in seconds**





