**Execution Time Analysis (per batch)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Execution Time (Baseline, ms)**  **Baseline** | **Execution Time (layer and kernel fusion, ms)** | **Speed Up %** | **Execution Time (layer fused + TensorRT + batch concurrency, ms)** | **Speed Up %** |
| **ResNet-50** | **11.76** | **4.759** |  | **5.260** |  |
| **EfficientNet-B0** | **28.438** | **25.041** |  | **7.316** |  |
| **MobileNetV2** | **19.895** | **11.551** |  | **2.902** |  |

Data Transfer Time Analysis between Host to Device

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | H2D Transfer Time (ms) | H2D Transfer Time (Layer and kernel fusion(ms) | Speed Up % | H2D Transfer Time (Layer Fusion and Kernel Fusion + Batch concurrency) | Speed Up % |
|  | Baseline | Fusion |  | Fusion + Batch concurrency |  |
| ResNet-50 | 1.48 | 0.199 |  | - |  |
| EfficientNet-B0 | 9.03 | 1.846 |  | - |  |
| MobileNetV2 | 7.86 | 1.530 |  | - |  |

Data Transfer Time Analysis between Device to Host

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | D2H Transfer Time (ms) | D2H Transfer Time (Layer and kernel fusion(ms) | Speed Up % | D2H Transfer Time (Layer Fusion and Kernel Fusion + Batch concurrency) | Speed Up % |
|  | Baseline | Fusion |  | Fusion + Batch concurrency |  |
| ResNet-50 | 8.36 | 0.159 |  | - |  |
| EfficientNet-B0 | 0.21 | 0.240 |  | - |  |
| MobileNetV2 | 0.27 | 0.132 |  | - |  |