CUDA vs PyTorch MLP Inference Benchmarking with Batching

Abstract

This document captures a detailed and reproducible benchmarking discussion comparing a custom CUDA MLP implementation with a PyTorch equivalent. The benchmarks evaluate batched inference on a 4-5-1 feedforward MLP architecture, covering performance breakdowns across CUDA memory operations, kernel execution, and PyTorch inference. It includes full prompts and responses from the original dialogue.

Full Conversation Transcript with Implementation Details

Prompt 1

User:

> I would like you to code for me a simple multi-layer feedforward in CUDA...

Assistant:

```
__global__ void forward_pass_kernel(...) {
    __shared__ float hidden_activations[HIDDEN_NODES];
    ... // Each thread computes a hidden neuron activation
}
```

Prompt 11

User:

> Awesome! That worked. Now, lets repeat our CUDA code with PyTorch!

Assistant:

```
class MLP(nn.Module):
    def __init__(self):
        super(MLP, self).__init__()
        self.fc1 = nn.Linear(4, 5)
        self.relu = nn.ReLU()
        self.fc2 = nn.Linear(5, 1)

    def forward(self, x):
        return self.fc2(self.relu(self.fc1(x)))
```

Benchmark Results and Visualization

PyTorch Stacked Log-Scale Bar Chart

This plot shows the breakdown of timing components for each batch size using PyTorch.

CUDA vs PyTorch Total Time Comparison

Displays total inference time for CUDA and PyTorch for each batch.

Time Difference (CUDA - PyTorch)

Shows the performance delta with annotations on which framework is faster.

Selected Output CSV Format

CUDA

```
Batch,Malloc,H2D,Kernel,D2H,Free,Total
1,0.065,0.010,0.007,0.003,0.004,0.285
```

PyTorch

```
Batch,Init,H2D,Forward,D2H,Total
1,1.232,0.054,0.002,0.001,1.289
```

Reproducibility Checklist

1. Use mlp_cuda_batched.cu and update BATCH_SIZE
2. Use run_batch_benchmarks.sh with sed + sm_75 compile flag
3. Run mlp_pytorch_batched.py with GPU event timing
4. Generate CSVs mlp_timing_log.csv and mlp_timing_log_pytorch.csv
5. Plot using:
   - plot_pytorch_timing.py
   - plot_cuda_vs_pytorch_total.py
   - plot_cuda_vs_pytorch_diff.py

All steps were coded, tested, timed, and visualized in this reproducible study.