Topics

My posts

More

CATEGORIES

General

Site Feedback

All categories

Skip to main content

2025 USA-NA-AIO Round 2, Problem 2, Part 13

USAAIO

May 2025

Part 13 (5 points, coding task)

Do the following tasks:

1. Define a function called `reduced_matrices`.
    - Input arguments
        - `W_DKV, W_UK, W_UV, W_Q, W_O, H`
    - Outputs
        - `W_K_MLA_hat, W_V_MLA_hat, W_Q_MLA_hat, W_O_MLA_hat`
    - Requirment of your code
        - The code of computing each output must be in one line
        - Loop is not allowed
2. Set your device as `gpu`:

device = torch.device(“cuda” if torch.cuda.is_available() else “cpu”)

```
3. Construct the following synthetic data:
```

```
D = 1024
H = 32
D_qkv = D // H
r = 50

W_DKV = torch.randn(r, D)
W_UK = torch.randn(D, r)
W_UV = torch.randn(D, r)
W_Q = torch.randn(D, D)
W_O = torch.randn(D, D)

B = 32
L_1 = 100
L_2 = 300

x = torch.randn(B, L_1, D).to(device)
y = torch.randn(B, L_2, D).to(device)
```

~~~~
 4. Study a vanilla attention model

  * Initialize the model

    ```
 model_MHA_vanilla = MyMHA(D, D, D_qkv, D_qkv, H)
    ```

  * Update model paramteres

      * `model_MHA_vanilla.W_K.weight, model_MHA_vanilla.W_V.weight, model_MHA_var

      * Compute the output

        ```
        output_vanilla = model_MHA_vanilla(x, y)
~~~~

Skip to main content

~~~~
           ```

5. Study a reduced attention model

 * Initialize the model

   ```
model_MHA_reduced = MyMHA(D, D, r, r, H)
   ```

 * Update model paramteres

     * `model_MHA_reduced.W_K.weight, model_MHA_reduced.W_V.weight, model_MHA_rec

     * Compute the output

         ```
         output_reduced = model_MHA_reduced(x, y)
         ```

6. Check the correctness of the reduced model by computing and printing a relativ
~~~~

relative_error = mse_output**.5 / torch.mean(output_vanilla**2).**5

USAAIO May 2025

```
### WRITE YOUR SOLUTION HERE ###

# Function

def reduced_matrices(W_DKV, W_UK, W_UV, W_Q, W_O, H):
    r = W_DKV.shape[0]
```

Skip to main content

```
    D = W_DKV.shape[1]

    W_K_MLA_hat = W_DKV
    W_V_MLA_hat = W_DKV
    W_Q_MLA_hat = (W_UK.reshape(H, -1, r).transpose(-2, -1) @ W_Q.reshape(H, -1,
    W_O_MLA_hat = (W_O.reshape(D, H, -1).transpose(0, 1) @ W_UV.reshape(H, -1, r)

    return W_K_MLA_hat, W_V_MLA_hat, W_Q_MLA_hat, W_O_MLA_hat

# Device
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")

# Data
D = 1024
H = 32
D_qkv = D // H
r = 50

W_DKV = torch.randn(r, D)
W_UK = torch.randn(D, r)
W_UV = torch.randn(D, r)
W_Q = torch.randn(D, D)
W_O = torch.randn(D, D)

B = 32
L_1 = 100
L_2 = 300

x = torch.randn(B, L_1, D).to(device)
y = torch.randn(B, L_2, D).to(device)


# Vanilla model
model_MHA_vanilla = MyMHA(D, D, D_qkv, D_qkv, H)
model_MHA_vanilla.W_K.weight = nn.Parameter(W_UK @ W_DKV)
model_MHA_vanilla.W_V.weight = nn.Parameter(W_UV @ W_DKV)
model_MHA_vanilla.W_Q.weight = nn.Parameter(W_Q)
model_MHA_vanilla.W_O.weight = nn.Parameter(W_O)

model_MHA_vanilla.to(device)
```

Skip to main content

```
output_vanilla = model_MHA_vanilla(x, y)

# Reduced model
model_MHA_reduced = MyMHA(D, D, r, r, H)
W_K_MLA_hat, W_V_MLA_hat, W_Q_MLA_hat, W_O_MLA_hat = reduced_matrices(W_DKV, W_UK
model_MHA_reduced.W_K.weight = nn.Parameter(torch.concatenate([W_K_MLA_hat] * H,
model_MHA_reduced.W_V.weight = nn.Parameter(torch.concatenate([W_V_MLA_hat] * H,
model_MHA_reduced.W_Q.weight = nn.Parameter(W_Q_MLA_hat)
model_MHA_reduced.W_O.weight = nn.Parameter(W_O_MLA_hat)

model_MHA_reduced.to(device)
output_reduced = model_MHA_reduced(x, y)

# Check the correctness of the reduced model
mse_output = torch.mean((output_vanilla - output_reduced)**2)
relative_error = mse_output**.5 / torch.mean(output_vanilla**2)**.5

print(f"Relative error: {relative_error.item()}")

""" END OF THIS PART """
```

Related topics

Topic	Replies	Activity
2025 USA-NA-AIO Round 2, Problem 2, Part 12	1	May 2025
2025 USA-NA-AIO Round 2, Problem 2, Part 10	1	May 2025

Skip to main content

Topic	Replies	Activity
2025 USA-NA-AIO Round 2, Problem 2, Part 5	1	May 2025
2025 USA-NA-AIO Round 2, Problem 2, Part 1	2	Dec 2025
2025 USA-NA-AIO Round 2, Problem 2, Part 7	1	May 2025