HW7 Report

STuser19 賴昱凱

Training result:

```
Test Loss: 0.369930

Test Accuracy of Class     0: 84.70% (847/1000)
Test Accuracy of Class     1: 97.80% (978/1000)
Test Accuracy of Class     2: 76.60% (766/1000)
Test Accuracy of Class     3: 86.70% (867/1000)
Test Accuracy of Class     4: 83.20% (832/1000)
Test Accuracy of Class     5: 96.50% (965/1000)
Test Accuracy of Class     6: 68.40% (684/1000)
Test Accuracy of Class     7: 97.20% (972/1000)
Test Accuracy of Class     8: 96.40% (964/1000)
Test Accuracy of Class     9: 95.30% (953/1000)

Test Accuracy (Overall): 88.28% (8828/10000)
```

Model:

```
ViT(
  (patch_embedding): Linear(in_features=16, out_features=64, bias=True)
  (transformer_encoders): ModuleList(
    (0-3): 4 x Transformer_Encoder(
      (norm1): Norm(
        (norm): LayerNorm((64,), eps=1e-05, elementwise_affine=True)
      )
      (mha): Multi_Head_Attention(
        (qkv): Linear(in_features=64, out_features=192, bias=True)
        (fc_out): Linear(in_features=64, out_features=64, bias=True)
      )
      (norm2): Norm(
        (norm): LayerNorm((64,), eps=1e-05, elementwise_affine=True)
      )
      (mlp): MLP(
        (mlp): Sequential(
          (0): Linear(in_features=64, out_features=256, bias=True)
          (1): ReLU()
          (2): Linear(in_features=256, out_features=64, bias=True)
        )
      )
    )
  )
  (mlp_head): Sequential(
    (0): LayerNorm((64,), eps=1e-05, elementwise_affine=True)
    (1): Linear(in_features=64, out_features=10, bias=True)
  )
)
```

參數量 (a single transformer encoder layer)

Transformer Encoder

Multi-Head Attention:

- **norm1**: 每個 channel 中皆有 gamma, beta 兩個參數。64 個 channels
  $Parameters = 64 + 64 = 128$

- **qkv linear layers**: 64 input channels, 192 output channels
  $Parameters = 64 \times 192 + 192 = 12480$

- **fc_out linear layers**: 64 input channels, 64 output channels
  $Parameters = 64 \times 64 + 64 = 4160$

- **norm2**: 每個 channel 中皆有 gamma, beta 兩個參數。64 個 channels
  $Parameters = 64 + 64 = 128$

MLP:

- **Linear Layer 1**: 64 input channels, 256 output channels
  $Parameters = 64 \times 256 + 256 = 16640$

- **Linear Layer 2**: 256 input channels, 64 output channels
  $Parameters = 256 \times 64 + 64 = 16448$

$$\boldsymbol{Total\ parameters = 128 + 12480 + 128 + 4160 + 16640 + 16448}$$
$$\boldsymbol{= 49984}$$

計算量 (calculate linear layer only for a single transformer encoder layer)

一個$n \times m$矩陣與一個$m \times n$矩陣做矩陣乘法，共需要$n \times m^2$個乘法以及$(n-1) \times m^2(sum\ of\ product) + m^2(bias) = n \times m^2$個加法，共**$n \times m$^2個乘加計算(MACs)**。

Transformer Encoder

Multi-Head Attention:

- **qkv linear layers**: 64 input channels, 192 output channels

$$MACs = B \times N \times input\ channel \times output\ channel$$
$$= 100 \times 50 \times 64 \times 192 = 61{,}440{,}000$$

- **Calculate attention score alpha = Q @ K^T**

$$MACs = B \times num\ heads \times N \times (head\ dim \times N)$$
$$= 100 \times 4 \times 50 \times (16 \times 50) = 16{,}000{,}000$$

- **Calculate output out = attention @ V**

$$MACs = B \times num\ heads \times N \times (N \times head\ dim)$$
$$= 100 \times 4 \times 50 \times (50 \times 16) = 16{,}000{,}000$$

- **fc_out linear layers**: 64 input channels, 64 output channels

$$MACs = B \times N \times input\ channel \times output\ channel$$
$$= 100 \times 50 \times 64 \times 64 = 20{,}480{,}000$$

MLP:

- **Linear Layer 1**: 64 input channels, 256 output channels

$$MACs = B \times N \times input\ channel \times output\ channel$$
$$= 100 \times 50 \times 64 \times 256 = 81{,}920{,}000$$

- **Linear Layer 2**: 256 input channels, 64 output channels

$$MACs = B \times N \times input\ channel \times output\ channel$$
$$= 100 \times 50 \times 256 \times 64 = 81{,}920{,}000$$

$$\boldsymbol{Total\ MACs = 61,440,000 + 16,000,000 + 16,000,000 + 20,480,000 + 81,920,000 + 81,920,000 = 277,760,000}$$