ADARLAB AI Training Course

Lec7 Homework Report

110511118 陳孟頡

Part I. Build a ViT model for image classification on FashionMNIST

- **Overall test accuracy screenshot**

```
Test Loss: 0.437977

Test Accuracy of Class     0: 83.90% (839/1000)
Test Accuracy of Class     1: 96.30% (963/1000)
Test Accuracy of Class     2: 73.30% (733/1000)
Test Accuracy of Class     3: 85.60% (856/1000)
Test Accuracy of Class     4: 73.90% (739/1000)
Test Accuracy of Class     5: 93.20% (932/1000)
Test Accuracy of Class     6: 61.40% (614/1000)
Test Accuracy of Class     7: 94.70% (947/1000)
Test Accuracy of Class     8: 95.50% (955/1000)
Test Accuracy of Class     9: 93.90% (939/1000)

Test Accuracy (Overall): 85.17% (8517/10000)
```

Test loss: 0.437977

Overall test accuracy: **85.17%**

- **Number of MACs and parameters**

```
MACs: 9983872.0, Params: 201290.0.
```

Number of MACs: 9,983,872

Number of parameters: 201,290

Part II. Report #MACs and #parameters of a single transformer encoder layer by manual calculations

- **Overall model architecture**

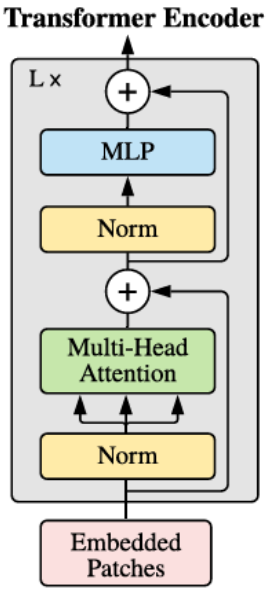
```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Linear-1              [-1, 49, 64]           1,088
         LayerNorm-2              [-1, 50, 64]             128
              Norm-3              [-1, 50, 64]               0
            Linear-4              [-1, 50, 64]           4,160
            Linear-5              [-1, 50, 64]           4,160
            Linear-6              [-1, 50, 64]           4,160
            Linear-7              [-1, 50, 64]           4,160
Multi_Head_Attention-8              [-1, 50, 64]               0
         LayerNorm-9              [-1, 50, 64]             128
             Norm-10              [-1, 50, 64]               0
           Linear-11             [-1, 50, 256]          16,640
             GELU-12             [-1, 50, 256]               0
           Linear-13              [-1, 50, 64]          16,448
          Dropout-14              [-1, 50, 64]               0
              MLP-15              [-1, 50, 64]               0
Transformer_Encoder-16              [-1, 50, 64]               0
         LayerNorm-17              [-1, 50, 64]             128
             Norm-18              [-1, 50, 64]               0
           Linear-19              [-1, 50, 64]           4,160
           Linear-20              [-1, 50, 64]           4,160
           Linear-21              [-1, 50, 64]           4,160
           Linear-22              [-1, 50, 64]           4,160
Multi_Head_Attention-23              [-1, 50, 64]               0
         LayerNorm-24              [-1, 50, 64]             128
             Norm-25              [-1, 50, 64]               0
           Linear-26             [-1, 50, 256]          16,640
             GELU-27             [-1, 50, 256]               0
           Linear-28              [-1, 50, 64]          16,448
          Dropout-29              [-1, 50, 64]               0
              MLP-30              [-1, 50, 64]               0
Transformer_Encoder-31              [-1, 50, 64]               0
         LayerNorm-32              [-1, 50, 64]             128
```

Norm-33	[-1, 50, 64]	0
Linear-34	[-1, 50, 64]	4,160
Linear-35	[-1, 50, 64]	4,160
Linear-36	[-1, 50, 64]	4,160
Linear-37	[-1, 50, 64]	4,160
Multi_Head_Attention-38	[-1, 50, 64]	0
LayerNorm-39	[-1, 50, 64]	128
Norm-40	[-1, 50, 64]	0
Linear-41	[-1, 50, 256]	16,640
GELU-42	[-1, 50, 256]	0
Linear-43	[-1, 50, 64]	16,448
Dropout-44	[-1, 50, 64]	0
MLP-45	[-1, 50, 64]	0
Transformer_Encoder-46	[-1, 50, 64]	0
LayerNorm-47	[-1, 50, 64]	128
Norm-48	[-1, 50, 64]	0
Linear-49	[-1, 50, 64]	4,160
Linear-50	[-1, 50, 64]	4,160
Linear-51	[-1, 50, 64]	4,160
Linear-52	[-1, 50, 64]	4,160
Multi_Head_Attention-53	[-1, 50, 64]	0
LayerNorm-54	[-1, 50, 64]	128
Norm-55	[-1, 50, 64]	0
Linear-56	[-1, 50, 256]	16,640
GELU-57	[-1, 50, 256]	0
Linear-58	[-1, 50, 64]	16,448
Dropout-59	[-1, 50, 64]	0
MLP-60	[-1, 50, 64]	0
Transformer_Encoder-61	[-1, 50, 64]	0
LayerNorm-62	[-1, 64]	128
Linear-63	[-1, 10]	650

==========

Total params: 201,802

- #MACs and #parameters calculation of encoder block



#MACs

Layer		#MACs
Multi_Head _Attention	Q, K, V Linear	3 * batch_size × seq_len × embed_size × embed_size = 3 * 1 * (49+1) * 64 * 64 = **614,400**
	Attention score (alpha)	batch_size × seq_len × seq_len × head_dim = 1 * 50 * 50 * 64 = **160,000**
	Output dot product	batch_size × seq_len × seq_len × head_dim = 1 * 50 * 50 * 64 = **160,000**
	Output Linear	seq_len ×head_dim ×head_dim = 50 * 64 * 64 = **204,800**
MLP	Fc1	seq_len × head_dim × embed_dim = 50 * 64 * 256 = **819,200**
	Fc2	seq_len × embed_dim ×head_dim = 50 * 256 * 64 = **819,200**
Total		**2,777,600**

#Parameters

Layer		#Parameters
LayerNorm 1		64 * 2 = **128**
Multi_Head _Attention	Q, K, V Linear	3 * (64^2 + 64) = **12,480**
	Output Linear	64^2 + 64 = **4,160**
LayerNorm 2		64 * 2 = **128**
MLP	Fc1	64 * (64*4) + (64*4) = **16,640**
	Fc2	64 * (64*4) + 64 = **16,448**
Total		**49,984**