LAB 6: STUDENT WORKSHEET

# Transformer

**Name:**  **Student ID:**  **Date:**

# PART 1: Basic Operations in Transformer

Q1: **Attention**: What is the definition of “attention”, and what are required by an attention mechanism?

Q2: **Scaled Dot Product Attention**: In the scaled\_dot\_product function, explain why the attention logits are divided by math.sqrt(d\_k). What would happen if this scaling factor were omitted?

Q3: **Multi-Head Attention**: What is the crucial characteristic of the multi-head attention?

Q4: **Transformer Encoder**: Why are the residual connection and layer normalization crucial in the Transformer architecture?

Q5: **Positional Encoding**: What do we choose as positional encoding? How do they help the Transformer model understand the position of input tokens?

# PART 2: Vision Transformer

Complete the table below with the numerical values from your experiments:

1. **Image Preprocessing**

|  |  |
| --- | --- |
|  | size |
| Original image |  |
| Patch |  |

Q1: What do we need to do on image preprocessing for Vision transformer?

1. **ViT inference**

**2.1 Inference latency V.S. layer numbers**

|  |  |
| --- | --- |
| Layer numbers | Inference latency (s) |
| 1 |  |
| 2 |  |
| 3 |  |
| 4 |  |
| 5 |  |
| 6 |  |
| 7 |  |
| 8 |  |
| 9 |  |
| 10 |  |
| 11 |  |
| 12 |  |

**2.2 Inference latency V.S. head numbers**

|  |  |
| --- | --- |
| Head numbers | Inference latency (s) |
| 2 |  |
| 4 |  |
| 6 |  |
| 8 |  |
| 10 |  |

Q1: How does the inference latency change when increase the number of layers, and number of heads.

1. **Transformer Training**

Q1: On small dataset like CIFAR10, ResNet50 can reach over 90% accuracy, while ViT can reach about 70%, what may cause this? What if we change to a bigger dataset, such as imagenet?

Q2: Why don’t we train ViT in class?

# PART 4: REFLECTION (100 words)

Write a short reflection on model selection criteria for different applications based on your experiment results.

# INSTRUCTOR COMMENTS

Grade: /