Assignment 10 - [Transformers]

Q1)
ANS:-
The following are components:

1. Embedding Layers:
- Token Embeddings (40k vocab): 40,000 x 768 = 30,720,000
- Positional Embeddings (512 max tokens): 512 x 768 = 393,216
- Segment Embeddings (2 segments): 2 x 768 = 1,536
Total Embedding Parameters: 30,720,000 + 393,216 + 1,536 ≈ 31,114,752

2. Encoder Layers (8 layers):
- Self-Attention Mechanism (8 attention heads):
  - Query/Key/Value Weight Matrices: 3 x (768 x 768) x 8 = 14,155,008
  - Output Weight Matrix: 768 x 768 x 8 = 4,569,984
Total Self-Attention Parameters: 18,725,992
- Feed-Forward Network (FFN):
  - Input Weight Matrix: 768 x 3072 = 2,359,296
  - Output Weight Matrix: 3072 x 768 = 2,359,296
Total FFN Parameters: 4,718,592
Total Encoder Parameters (per layer): 18,725,992 + 4,718,592 = 23,444,584
Total Encoder Parameters (8 layers): 23,444,584 x 8 = 187,556,672

Total Model Parameters:
31,114,752 (Embeddings) + 187,556,672 (Encoder) ≈ 218,671,424
Total parameters in BERT model ≈ 218 million.

Q2)
ANS:-
To calculate self-attention output for the word 'flying', we'll follow these steps:
Input Embeddings:
- Flying: [0, 1, 1, 1, 1, 0]
- Arrows: [1, 1, 0, -1, -1, 1]

Query (Q), Key (K), and Value (V) Matrices:
Since we're using 2 dimensions for each, we'll extract the relevant parts:
- Q (Flying): [0, 1]
- K (Flying): [1, 1]
- V (Flying): [1, 1]
- Q (Arrows): [1, 1]
- K (Arrows): [0, -1]
- V (Arrows): [-1, -1]

Scaled Dot-Product Attention:
1. Compute attention scores:
   - Flying-Flying: (Q * K) / sqrt(d) = ([0, 1] * [1, 1]) / sqrt(2) = 1 / sqrt(2)
   - Flying-Arrows: (Q * K) / sqrt(d) = ([0, 1] * [0, -1]) / sqrt(2) = -1 / sqrt(2)

2. Apply softmax to attention scores:
   - Softmax(Flying-Flying) = exp(1 / sqrt(2)) / (exp(1 / sqrt(2)) + exp(-1 / sqrt(2))) ≈ 0.6225
   - Softmax(Flying-Arrows) = exp(-1 / sqrt(2)) / (exp(1 / sqrt(2)) + exp(-1 / sqrt(2))) ≈ 0.3775

3. Compute weighted sum:
   - Flying: 0.6225 * [1, 1] + 0.3775 * [-1, -1] ≈ [0.245, 0.245]
   - This is the self-attention output for 'flying'.

So, the self-attention output for the word 'flying' corresponding to this attention head is approximately:
[0.245, 0.245]

Q3)
ANS:-

Topic Classification (5 classes):
- Input: [CLS] token representation from BERT
- Output: Probability distribution over 5 classes
- Classification Head:
  - Weight Matrix: 768 (BERT hidden size) x 5 (number of classes) = 3,840
  - (Optional) Bias Term: 5 = 5

Total task-specific parameters: 3,840 + 5 ≈ 3,845

Language Identification (2 languages, English and Hindi):
- Input: [CLS] token representation from BERT (or word-level representation)
- Output: Probability distribution over 2 languages
- Classification Head:
  - Weight Matrix: 768 (BERT hidden size) x 2 (number of languages) = 1,536
  - (Optional) Bias Term: 2 = 2

Total task-specific parameters: 1,536 + 2 ≈ 1,538

In both cases, the number of task-specific parameters is relatively small compared to the total number of BERT parameters (~218 million). This is one of the advantages of using pre-trained language models like BERT.