Q1 Answer:

Given that:

Dimension: d = 768

No. of attention heads: h = 8

Dimension of Feed Forward Layer: d_{ffl} = 3072

No. of layers = 8

Max no. of tokens in input = 512

Size of vocabulary = 40000

Parameters in Transformer Layers:

For 8 attention heads, q/k/v matrices to be of dimensions d_q, d_k, d_v = d / h = 768 / 8 = 96

Parameters for q/k/v matrices per attention head = $d \times d_q + d \times d_k + d \times d_v = 768 \times 96 \times 3 = 221{,}184$

Parameters for q/k/v matrices for 8 such heads = $221{,}184 \times h = 221{,}184 \times 8 = 1{,}769{,}472$ $(= 3 \times d^2)$

Parameters for self-attention output projection = $d \times d = 768 \times 768 = 589{,}824$ $(= 1 \times d^2)$

Parameters in Feed Forward = $(d \times d_{ffl}) + (d_{ffl} \times d) = (768 \times 3072) + (3072 \times 768) = 4{,}718{,}592$ $(= 8 \times d^2)$

Total parameters per layer = 1,769,472 + 589,824 + 4,718,592 = 7,077,888 $(= 12 \times d^2)$

Total parameters for 8 such layers = 7,077,888 × 8 = 56,623,104

Parameters in Embedding Layers:

Parameters for token embeddings (token <—> vocab projection) = 768 × 40,000 = 30,720,000

Parameters for positional embeddings = 512 × 768 = 393,216

Total parameters for the model *ignoring positional embeddings* = 56,623,104 + 30,720,000

= 87,343,104 i.e. **~ 87.3 M**

Total parameters for the model *including positional embeddings* = 87,343,104 + 393,216

= 87,736,320 i.e. **~ 87.7 M**

Q2 Answer:

Given that:

$Input_{flying}$ = [0,1,1,1,1,0], $Input_{arrows}$ = [1,1,0,-1,-1,1]

q_{flying} = [0, 1], q_{arrows} = [1,1] #Considering 1st & 2nd dimensions from Input Embeddings

k_{flying} = [1, 1], k_{arrows} = [0, -1], d_k = 2 #Considering 3rd & 4th dimensions from Input Embeddings

v_{flying} = [1, 0], v_{arrows} = [-1, 1] #Considering 5th & 6th dimensions from Input Embeddings

Now:

Scaled dot product for q_{flying} with k_{flying} $= (q_{flying} \cdot k_{flying}) / \sqrt{d_k} = [0, 1] \cdot [1, 1]^T / \sqrt{2} = 1/\sqrt{2} \approx 0.707$

Scaled dot product for q_{flying} with k_{arrows} $= (q_{flying} \cdot k_{arrows}) / \sqrt{d_k} = [0, 1] \cdot [0, -1]^T / \sqrt{2} = -1/\sqrt{2} \approx -0.707$

Attention weights vector $[\lambda_{flying1}, \lambda_{flying2}]$ $= \text{softmax}([0.707, -0.707])$

$= [e^{0.707} / (e^{0.707} + e^{-0.707}), e^{-0.707} / (e^{0.707} + e^{-0.707})]$

$= [0.804, 0.196]$

Self-attention output for the word 'flying' corresponding to this attention head is:

$\lambda_{flying1} \times v_{flying} + \lambda_{flying2} \times v_{arrows}$ $= 0.804 \times [1, 0] + 0.196 \times [-1, 1]$

$= [0.804, 0] + [-0.196, 0.196]$

$= [0.804 - 0.196, 0 + 0.196]$

$=$ **[0.608, 0.196]**

Q3 Answer:

For Topic classification task:

BERT-base hidden dimension size = 768

No. of classes = 5

Therefore, number of task specific parameters (ignoring bias terms) = 768 × 5 = **3840**

Parameters including bias terms (1 per class) = 3840 + 5 = **3845**

For Language identification task:

BERT-base hidden dimension size = 768

No. of classes = No. of possible languages (English and Hindi) = 2

Therefore, number of task specific parameters (ignoring bias terms) = 768 × 2 = **1536**

Parameters including bias terms (1 per class) = 1536 + 2 = **1538**