1. Total Number of Parameters in BERT Model

To determine the number of parameters for the BERT model with 8 layers, a hidden state dimension of 768, 8 attention heads, and a vocabulary size of 40,000, we compute several components:

A. Embedding Layer

The embedding layer consists of:

- **Token Embeddings**:

$\text{Parameter Count} = V \times H = 40{,}000 \times 768 = 30{,}720{,}000$

B. Transformer Layers

Each transformer layer consists of:

1. **Multi-Head Attention**:

    - Each attention head has parameters for computing queries, keys, and values:

$H \times (H/A) = 768 \times (768/8) = 768 \times 96 = 73{,}728$ (for each head)

    - For all heads (3 heads per layer):

$3 \times A \times 73{,}728 = 3 \times 8 \times 73{,}728 = 1{,}764{,}864$

    - Output projection:

$H \times H = 768 \times 768 = 589{,}824$

    - Total for attention per layer:

$1{,}764{,}864 + 589{,}824 = 2{,}354{,}688$

2. **Feed-Forward Network**:

    - With a hidden dimension of 3072:

$2 \text{ (for 2 linear transformations)} \times H \times 3072 = 2 \times 768 \times 3072 = 4{,}732{,}160$

C. Total for One Layer

Combining both contributions from the attention and feed-forward networks gives:

$2{,}354{,}688 + 4{,}732{,}160 = 7{,}086{,}848$ (per layer)

D. Total for All Layers

For all 8 layers:

$8 \times 7{,}086{,}848 = 56{,}694{,}784$

E. Final Parameter Count

Combining the embedding parameters and the total parameters from all transformer layers gives:

30,720,000+56,694,784=87,414,784

Thus, the total number of parameters in the BERT model is **87,414,784**.

---

2. Self-Attention Output for 'Flying'

Considering the input embeddings for the words **flying** and **arrows** as [0, 1, 1, 1, 1, 0] and [1, 1, 0, -1, -1, 1], and using only the first 2 dimensions for the self-attention calculation, we proceed to calculate the attention output for **flying**.

Step 1: Query, Key, and Value Vectors

For the first attention head:

- **Query Vector (Q)**: [0,1]
- **Key Vector (K)**: [0,1]
- **Value Vector (V)**: [0,1]

Step 2: Scaled Dot Product Attention

Using the scaled dot product:

$\text{Attention}(Q,K,V)=\text{softmax}(Q.K^T/(d_k)^{1/2})V$

Where:

- d_k is the dimension of the key vectors (which is 2 here).

Step 3: Calculating Inputs

1. **Dot Product**:

$Q\cdot K^T=[0,1]\cdot[0,1]^T=1$

2. **Scaling**:

$\text{scaled}=1/2^{1/2}=0.707$

3. **Softmax** (assuming compatible inputs; 2 inputs):

   - For simplicity, assuming uniform output:

$\text{softmax}(1,1)=[0.5,0.5]$

Step 4: Final Attention Calculation

$\text{Output}=[0.5,0.5]\cdot V=[0.5,0.5]\cdot[0,1]=0.5$

The self-attention output for the word **flying** is **0.5**.

---

3. Task-Specific Parameters in BERT-base

A. Topic Classification with 5 Classes

In the case of topic classification with 5 classes, we need an additional classification layer on top of the BERT model. Thus, the number of task-specific parameters for this output layer is:

$\text{Parameters}=\text{Number of classes}\times\text{Hidden size}=5\times768=3{,}840$

B. Language Identification in Code-Switched Dataset

For language identification of two languages (English and Hindi):

$\text{Parameters}=\text{Number of classes}\times\text{Hidden size}=2\times768=1{,}536$