**LAYERS & Params**

A screenshot of a computer program

Description automatically generated

**encoder\_input**

A batch of sequences with shape (batch\_size, sequence\_length) is input into the encoder\_input layer.

**decoder\_input**

A batch of sequences with shape (batch\_size, sequence\_length) is input into the decoder\_input layer.

**Token and Position Embedding Layers (Encoder)**:

The encoder input is embedded into a (batch\_size, sequence\_length, 256) representation.

Input Shape: (None, None) Output Shape: (None, None, 256) Param #: 1,548,544

This layer combines token embedding and positional encoding for the encoder input.

Parameter Calculation:

Vocabulary size (English) = 6033

Embedding dimension = 256

Total parameters = 6033 \* 256 = 1,548,544

**token\_and\_position\_embedding\_1 (Decoder)**:

The decoder input is embedded into a (batch\_size, sequence\_length, 256) representation.

Input Shape: (None, None) Output Shape: (None, None, 256) Param #: 3,126,528

This layer combines token embedding and positional encoding for the decoder input.

Parameter Calculation:

Vocabulary size (French) = 12197

Embedding dimension = 256

Total parameters = 12197 \* 256 = 3,126,528

**Transformer Encoder**

The embedded encoder input is processed, resulting in an output of shape (batch\_size, sequence\_length, 256).

Input Shape: (None, None, 256) Output Shape: (None, None, 256) Param #: 389,614

This layer processes the embedded encoder input using self-attention and feed-forward networks. The input and output shapes remain the same, as the transformation happens internally.

Parameter Calculation:

**Transformer Encoder Parameters:**

**Multi-Head Self-Attention:**

* + Embedding Dimension: 256
  + Number of Heads: 10
  + Parameters for Query, Key, and Value Projections:
    - Weight matrices: 10×(256×256)=2,560,000
    - Biases: 10×256=2,56010
    - Total for projections: 2,560,000+2,560=2,562,560
  + OutputLinear Layer:
    - Weight matrix: 256×256=65,536
    - Bias vector: 256
    - Total: 65,536+256=65,792

**Feed-Forward Network (FFN):**

* + Hidden Size: 1024
  + Total Parameters:
    - First Layer: 256×1024+1024=262,144+1024=263,168
    - Second Layer: 1024×256+256=262,144+256=262,400
    - Total: 263,168+262,400=525,568

**Total Parameters for One Encoder Layer:**

* Self-Attention: 2,562,560+65,792=2,628,352
* Feed-Forward Network: 525,568
* Overall Total**:** 2,628,352+525,568=3,153,920

**Transformer Decoder**

The embedded decoder input and the encoder output are processed together, resulting in an output of shape (batch\_size, sequence\_length, 256).

Input Shape: (None, None, 256) Output Shape: (None, None, 256)

Param #: 647,132

This layer processes the embedded decoder input using self-attention, encoder-decoder attention, and feed-forward networks.

Parameter Calculation:

This includes parameters for self-attention, encoder-decoder attention, and the feed-forward neural network. With num\_heads=10, embed\_dim=256, the parameters include query, key, value projections and the feed-forward network's parameters for both self-attention and encoder-decoder attention.

**Dropout Layer**

Dropout is applied to the decoder output, maintaining the shape (batch\_size, sequence\_length, 256).

Input Shape: (None, None, 256) Output Shape: (None, None, 256)

Param #: 0

Applies dropout to the output of the transformer decoder to prevent overfitting. No additional parameters are required.

**Dense Layer**

The final output is projected to the target vocabulary size, resulting in a shape of (batch\_size, sequence\_length, 12197).

Input Shape: (None, None, 256) Output Shape: (None, None, 12197)

Param: 3,134,629

This layer projects the 256-dimensional output of the decoder to the target vocabulary size (12197). The shape (None, None, 12197) indicates variable sequence length, each token represented by a 12197-dimensional vector.

Embedding dimension = 256

Vocabulary size (French) = 12197

Total parameters = 256 \* 12197 (weights) + 12197 (bias) = 3,134,629