Details of the GPT-2 Model

(**G**enerative **P**re-trained **T**ransformer)

By: Hiva Mohammadzadeh

Structure of the model

- Pre-trained deep learning model that uses unidirectional transformers to generate one token at a time.
- A decoder only model - Generates output in an autoregressive fashion. It learns to predict the next word in a sequence of text, given previous words in the sequence.
- Consists of 12 Transformer Decoder blocks where each block has two sub-layers: a Multi-Head Masked self-attention mechanism and a position-wise fully connected Feed-Forward Network.
- Use GPT-2 small with 128 tokens and an embedding dimension of 768 and 3072 feed-forward filter size.

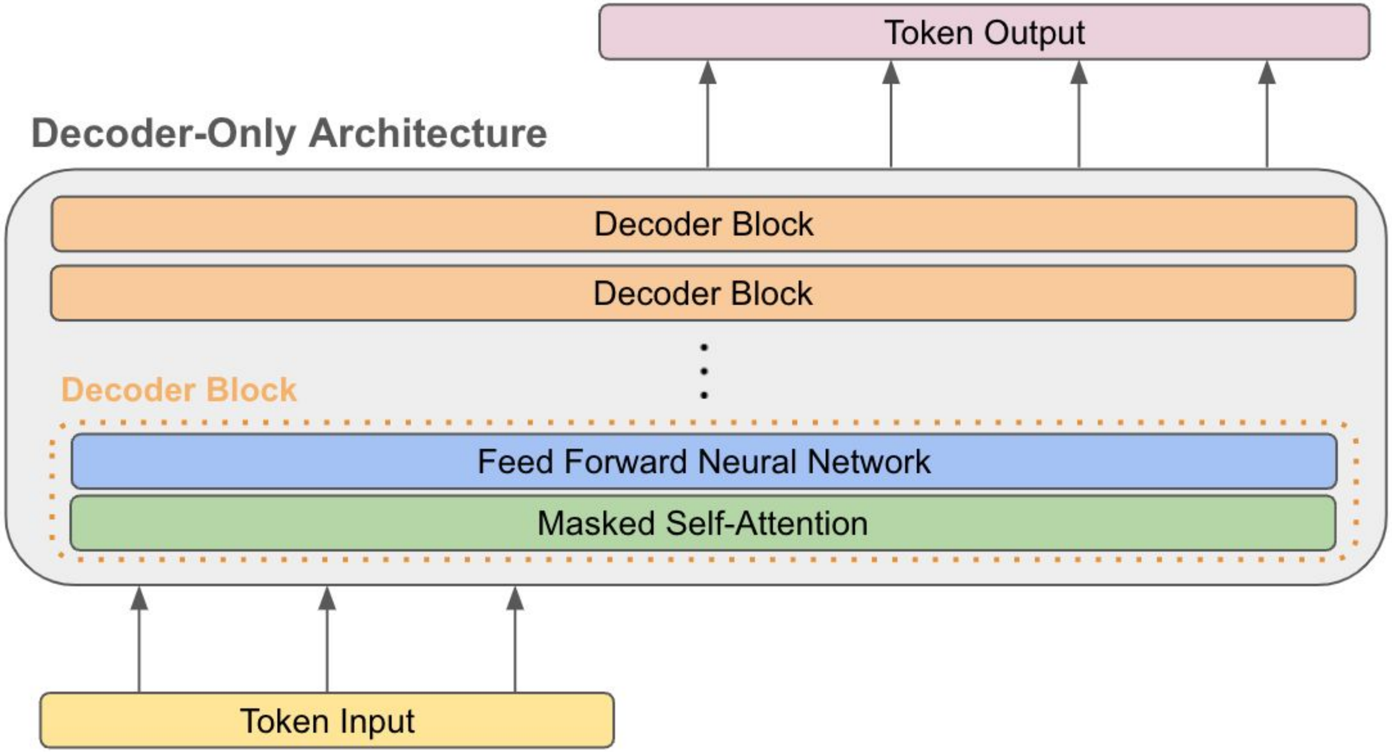
Training Scheme

- Pre-Training:
  a. Pre-trained on large, diverse text using an unsupervised learning approach. The model is trained to predict the next toke in the sequence give the past and present tokens.
- Fine-Tuning:
  a. Fine-Tune the model on specific downstream tasks by adding a task-specific output layer.

GPT-2 Architecture

Two key ingredients:

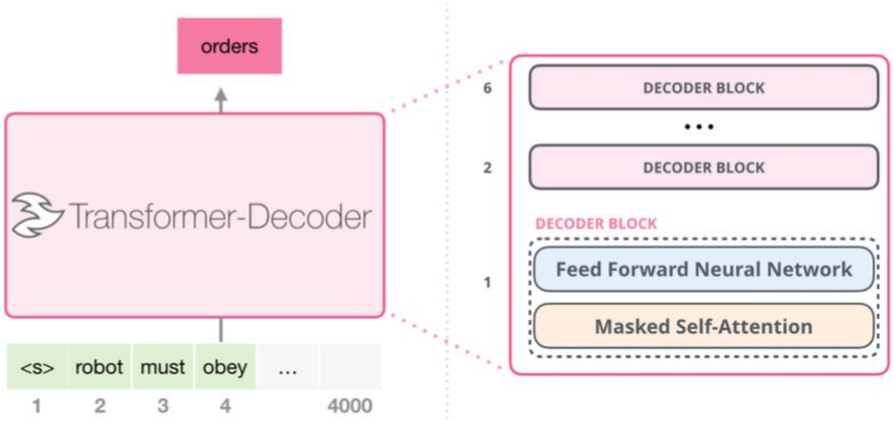
- Transformer Decoder
- Task specific output layer



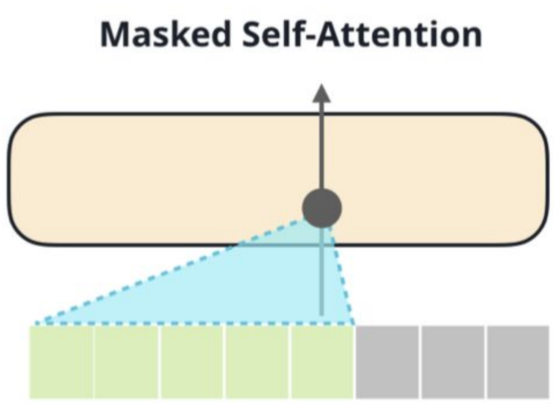
Transformer Decoder

Two components: Attention and FFN

- Masked self-attention is used to allow the model to attend to different parts of the input sequence.
- Feed Forward Network (FNN) module is used in every transformer block to process the output of the normalization layer in a way to better fit it to the next attention layer.



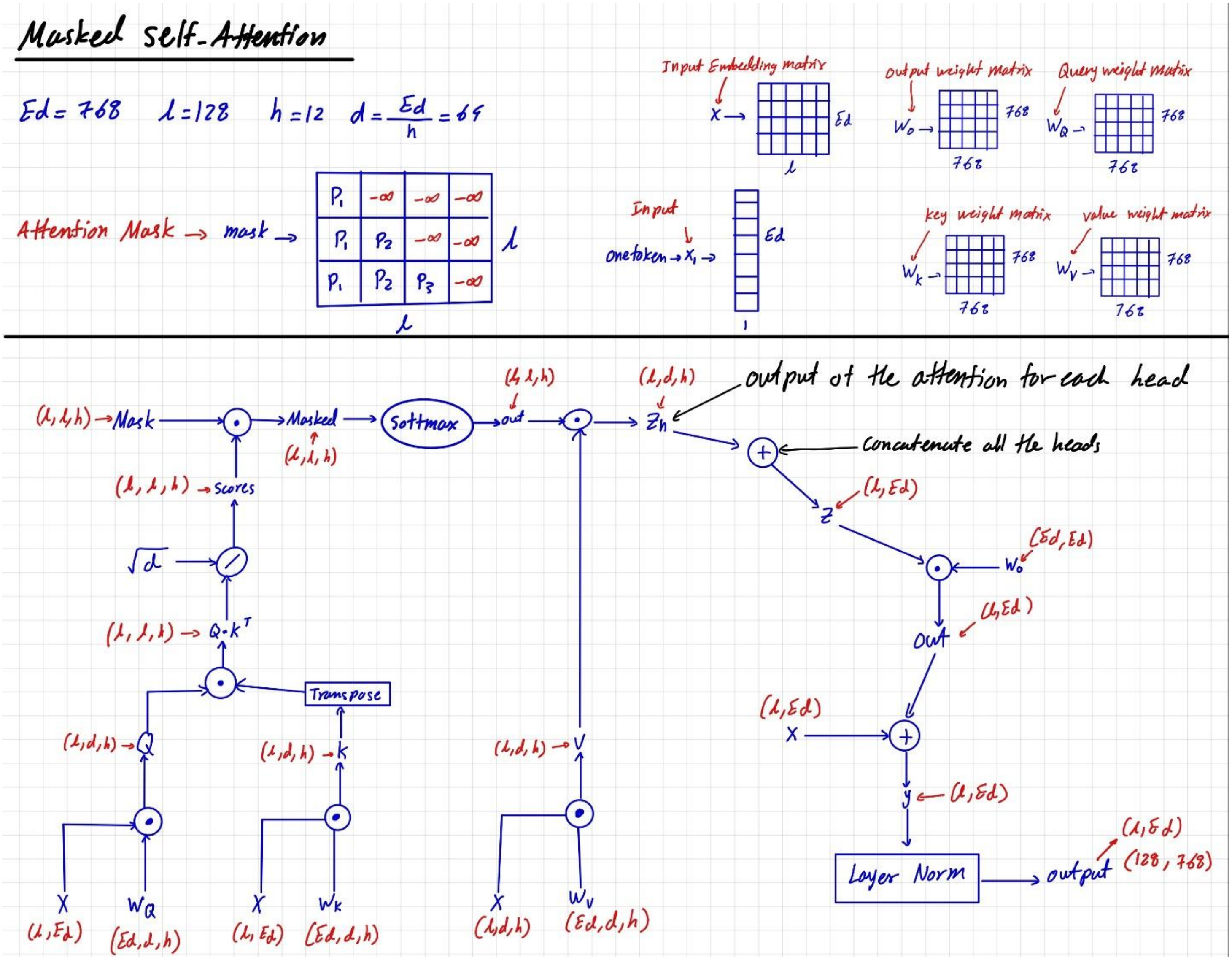
- Residual connections are used to avoid vanishing gradient problem.
- The Layer Normalizations are used to improve the model's convergence speed.
- GELU is the non-linearity used because it has been found to perform better than the other activation functions.



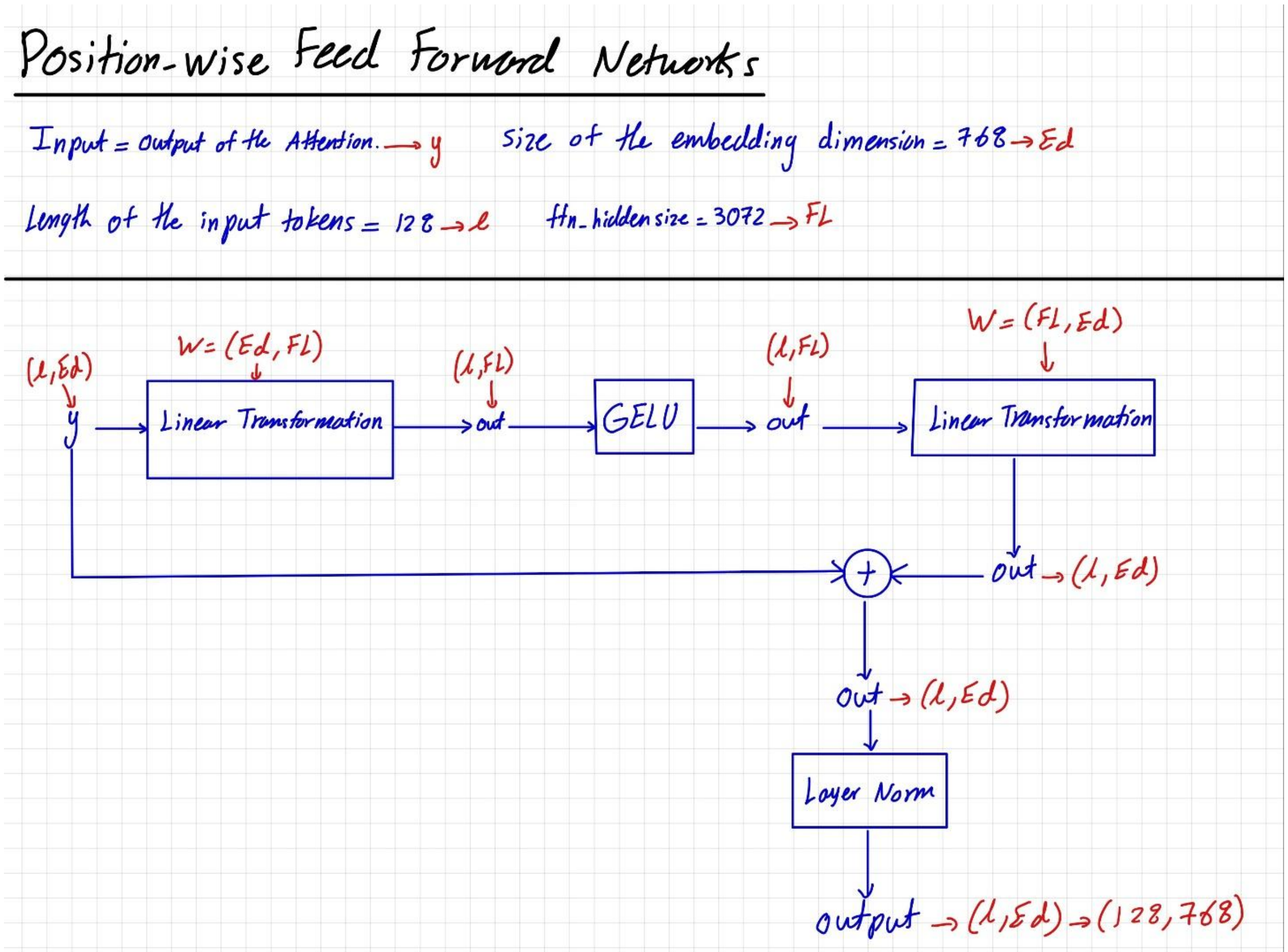
Dimensions of Weight Matrices

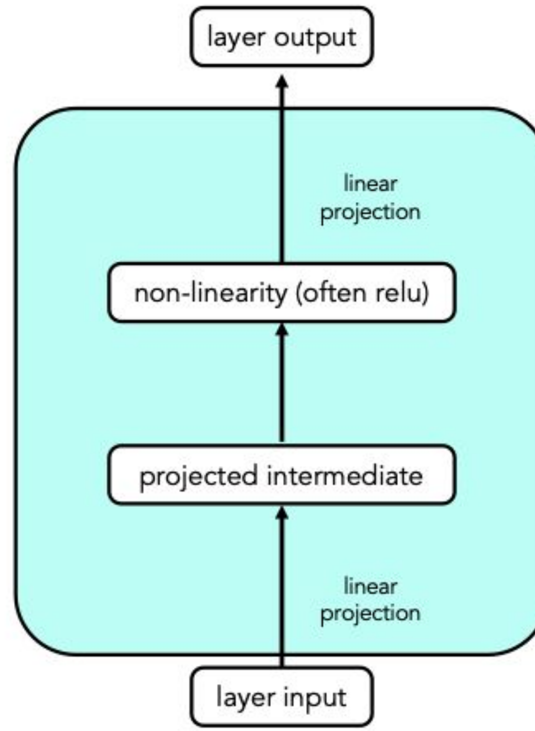
- The input embedding matrix has a dimension of (batch_size, sequence_length, embedding_size) = (batch_size, 128, 768) = (batch_size, 128, 64, 12)
- The Wk, Wv, and Wq weight matrices used in the **masked self-attention mechanism** have dimensions of:
  - Query weight matrix: (batch_size, hidden_size , hidden_size) = (batch_size, 768, 768)
  - Key weight matrix: (batch_size, hidden_size, hidden_size) = (batch_size, 768, 768)
  - Value weight matrix: (batch_size, hidden_size, hidden_size) = (batch_size, 768, 768)
  - Output weight matrix: (batch_size, hidden_size, hidden_size) = (batch_size, 768, 768)
- The K, V, Q matrices also have dimensions of:
  - Query matrix: (batch_size, 1 token, hidden_size) = (batch_size, 128, 768) or (batch_size, 128, 64, 12)
  - Key matrix: (batch_size, sequence_length, hidden_size) = (batch_size, 128, 768) or (batch_size, 128, 64, 12)
  - Value matrix: (batch_size, sequence_length, hidden_size) = (batch_size, 128, 768) or (batch_size, 128, 64, 12)
- The weight matrices used in the **feed-forward neural network** have dimensions of:
  - First dense layer: (hidden_size, 4*hidden_size) = (768, 3072).
  - Second dense layer: (4*hidden_size, hidden_size) = (3072, 768).

Computation Diagram of Attention Block

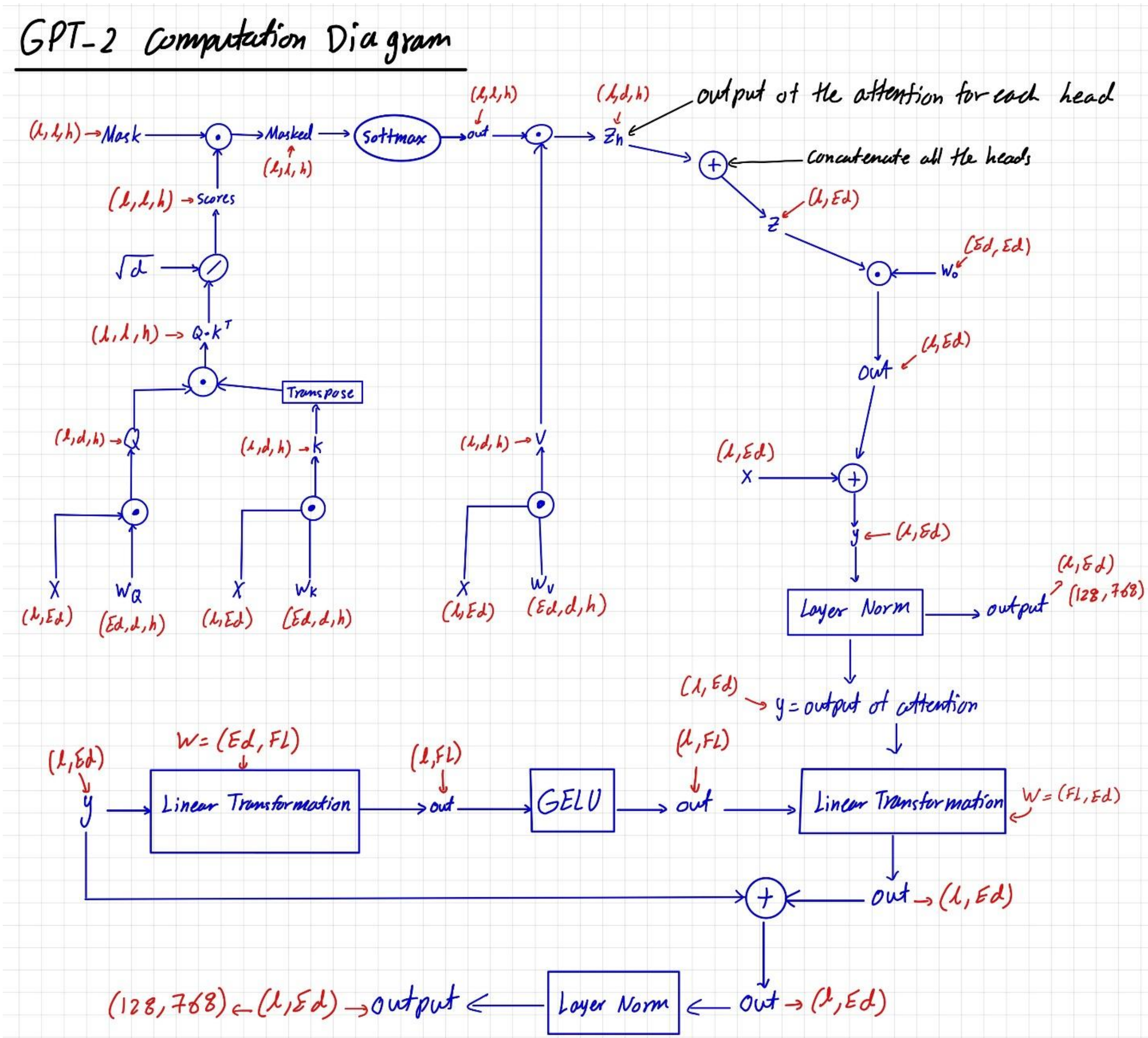


Computation Diagram of Feed Forward Layer





Computation Diagram of the whole GPT-2 architecture



FLOPS

Assuming that every dot product requires 1 multiplications and 1 addition, I calculated the following FLOPS:

- Flops for attention block:
  - Multiplication between embedding matrix and the weight matrices of Q, K and V : 3 x 128 tokens x 768 embeddings x 768 embeddings x 2 FLOPs = 301,989,888 FLOPs
  - Masked Multi-Head Self-Attention for 12 layers: 2 x 12 x L(depends on where we are in the sequence) x 64 FLOPs (dot product of Q and K) + 1 x 128 x 128 x 2 (dot product between mask and scores) + 2 x 12 x 128 x 64 FLOPs (scaled dot-product attention between output of softmax and V) + 2 x 768 x 768 FLOPs (concatenation of heads and matrix multiplication of W0) = 1,605,632
- FLOPS for Feed Forward Network Module:
  - Feed-Forward Network: FLOPs = 128 x 768 x 3072 x 2 FLOPs x 2 FLOPs + 128 FLOPs = 1,207,959,552
- Total FLOPs for 12 transformer block: FLOPs = 1,207,959,552 FLOPs (FFN) + 1,605,632 FLOPs (MHA) + (1,179,648 FLOPs + 301,989,888 FLOPs)(weight matrix calculations) = 1,512,734,848 FLOPs