Transformer Language Models

Ofek Kirshenboim

1 English Text

1.1 Training Parameters

Parameter	**Value**
Sequence Length	128
Batch Size	64
Number of Layers	6
Number of Heads	7
Embedding Size	469
Learning Rate	0.00023
Gradient Clipping	0.30435
Weight Decay	0.00031
Number of Batches to Train	50000

Table 1: Training Parameters for English Text

1.2 Results

Result	**Value**
Last Loss	0.1964
Parameter Count	16.00M
Total Training Sequences	3,200,000

Table 2: Results for English Text

2 Hebrew Text

2.1 Training Parameters

Parameter	**Value**
Sequence Length	128
Batch Size	64
Number of Layers	8
Number of Heads	8
Embedding Size	192
Learning Rate	0.0005
Gradient Clipping	1.0
Weight Decay	0.0001
Number of Batches to Train	50000

Table 3: Training Parameters for Hebrew Text

2.2 Results

Result	Value
Last Loss	0.1650
Parameter Count	3.63M
Total Training Sequences	3,200,000

Table 4: Results for Hebrew Text

3 Modifications and Optimizations

3.1 Dropout

A dropout layer with a dropout rate of 0.1 was added in the `TransformerDecoderBlock(nn.Module)` after the causal attention.

3.2 Hyperparameter Optimization

Optuna was used to optimize the hyperparameters. This involved running multiple trials to find the best set of hyperparameters for the model. The parameters tuned included the number of layers, number of heads, embedding size, learning rate, gradient clipping, and weight decay.

3.3 Optimizer

The AdamW optimizer was chosen for training the model.

4 Hebrew Model Performance

The model demonstrates an ability to generate words in Hebrew. However, when examining longer sequences of words that form a sentence, it becomes evident that the sentences lack coherence and meaningful content. Despite this, it's clear that the model has learned to produce Hebrew words to some extent.

4.1 Example

For the word - "say" (in Hebrew), the model output -

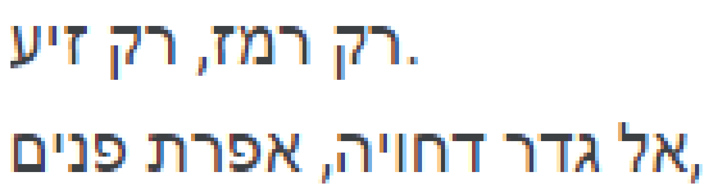

Figure 1: The output model for the word "say" in Hebrew

for the evaluation, I used a temperature of 0.5.

5 Attention Analysis

The color scale indicates the strength of attention, with white representing high attention (value close to 1.0) and black representing low attention (value close to 0.0). The near-perfect white on the one-below-diagonal positions shows strong attention to the previous token. This suggests that each token pays the most attention to the previous token in the sequence.

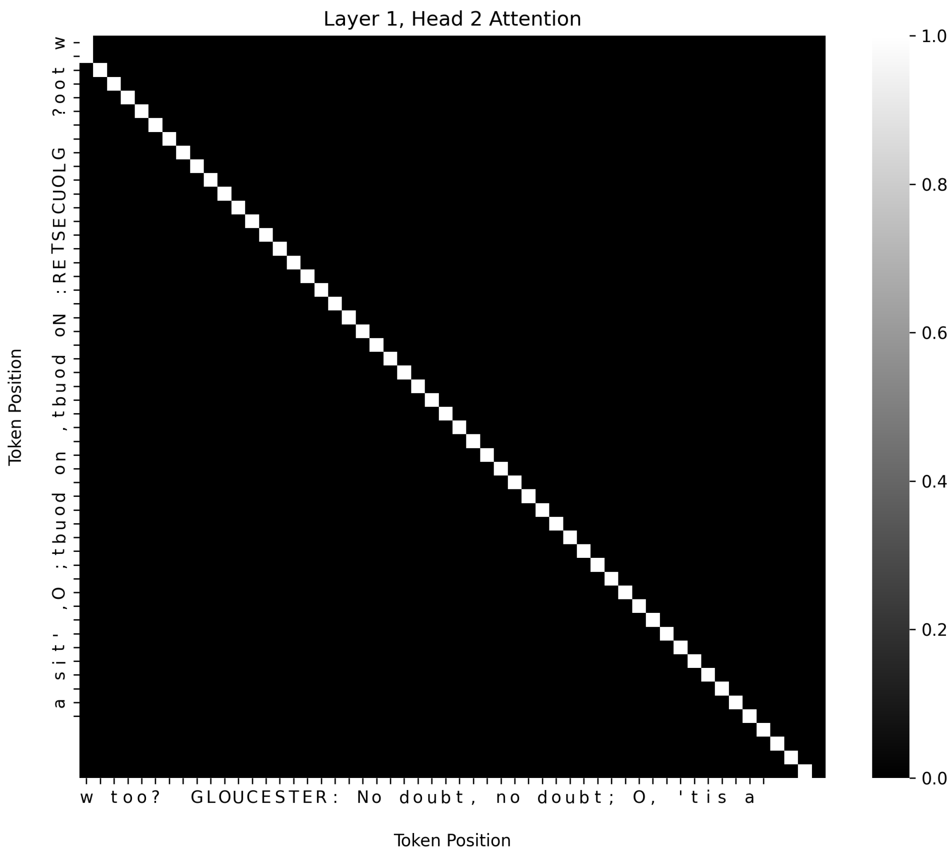

Figure 2: Heatmap of Attention Matrix for Layer 1, Head 2