1 Introduction

In 2017, Vaswani et al. introduced the Transformer architecture for the goal of machine translation, and since then, the Transformer has become the foundation for the modern landscape of machine learning and artificial intelligence. [Vaswani et al., 2017] To explore the Transformer architecture, we use it to solve the similar but simpler task of function word restoration wherein the input is a sentence with the function words – prepositions, articles, etc – removed, and the output is the same sentence with the function words reinserted. Like with machine translation, the input and expected output sequences are of differing length, and so, we use an encoder-decoder architecture using two Transformers to accomplish this task and perform three experiments to examine how changes in the architecture affects performance.

The first is comparing two different positional encoding strategies. Typically, positional encodings are done using sinusoidal functions, but we also use learnable positional embeddings as a point of comparison. The second is comparing two different decoding strategies: greedy decoding and beam search with various beam widths. The third experiments with the number of attention heads and the number of layers within each encoder and decoder.

2 Methods

2.1 Transformer and Attention

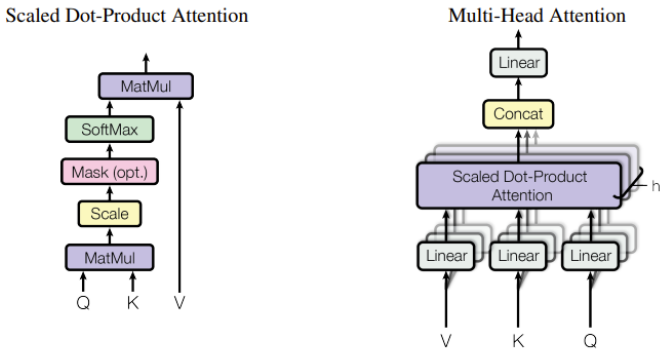


Figure 1: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel. [Vaswani et al., 2017]

Before describing the experiment implementation, we begin by explaining attention. To the left in Fig. 1, we have three matrices Q, K, V where Q and K are multiplied together and scaled to produce scores showing how much K attends to Q. Then these scores are softmaxed to create probabilities that serve as weights that are then multiplied with V. Optionally, a masking is applied to the upper triangle to the product of Q and K to disallow future tokens from attending to previous ones. For this task, $Q, K,$ and V are weight matrices that is multiplied with the input.

Then multi-head attention has multiple attention heads with their own Q, K, V matrices that can capture different relations, and these run in parallel.

With multi-head attention, we build the transformer architecture shown in Fig. 2, to the left is the encoder. First, each token embedding in the input is added with a positional encoding since the parallelization of the attention mechanism loses the original position of the tokens without positional encoding. For the encoder, the original input is passed into the multi-head attention explained earlier, and the result is then added and normalized by original input. This is then passed into a feed forward network, and again for stability, the new result is added and normalized by the previous result. This is done for N encoder layers, and the final result is used as a context vector for the decoder.

In the decoder, the outputs shifted right serve as the input for the multi-head attention, but since at inference time, the output is generated one token at a time, during training, the output is masked for the

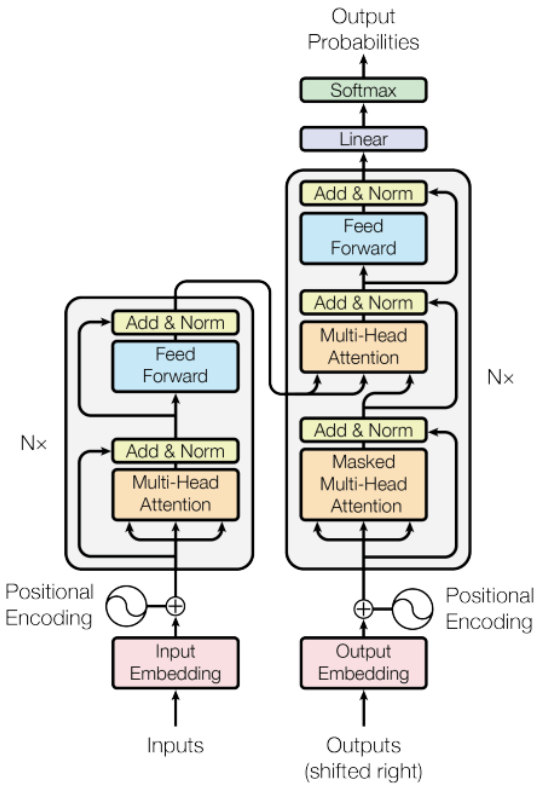


Figure 2: The Transformer - model architecture [Vaswani et al., 2017]

first layer of self attention. Then after the result of self attention is added and normalized with the previous input, it is multiplied with the V matrix of the next multi-head attention block while the context vector from the encoder is multiplied with the Q and K matrix. The added and normalized result is then passed into a feed-forward network. This process is then repeated N times before the final result of the decoder layer is passed into a linear layer, and a softmax is applied to give the output probabilities for the next token.

2.2 Experiment Implementation

Grid Search

Before starting the experiments, a grid search was done testing the depth of the encoder/decoder layers, the dimension of the model, the number of attention heads, and the dimension of the feed-forward network with the following values given below in Table 1.

Hyperparameter	Values
Encoder/Decoder Depth	1, 2, 4
Model Dimension	256, 512
Attention Heads	2, 4, 8
Feed-Forward Dimension	256, 512

Table 1: Values used in grid search

Other hyperparameters considered were the max sequence length, dropout, and learning rate; however, the combinatorics of testing all these hyperparameters in combination was not feasible given the time and resource strengths, so these hyperparameters were fixed to the following values.

- Max Sequence Length: 50
- Dropout: 0.1
- Learning Rate: 1e-3

Each configuration was trained for three epochs, and the criteria used to find the best performing model was highest BLEU score on the validation set. The implementation of calculating BLEU score was done using `sentence_bleu` from `nltk`, and smoothing method 4 was used because it gives shorter translations smaller smoothed counts in order to reduce the inflation in precision shorter translations may yield due to having smaller denominators.

Experiment 1

To implement learnable positional encodings, another class `LearnablePosEncoding` was created, and within its initialization, an embedding was created using `nn.Embedding`. Then within the forward method of this class, a tensor consisting of the positions from one up to the max sequence length is passed into the embedding, and the result is then added to the input.

Then with this new class, two models are created using the best configuration found from the grid search but one with sinusoidal positional encodings and the other with learnable encodings. The evaluation loss was found using the validation set while the BLEU score was calculated using the test set.

Experiment 2

Using the best model configuration, the different decoding strategies were compared using the same pre-trained model and same test set. Due to time constraints, only the greedy decoding and beam search with width 3 were actually run.

Experiment 3

To examine the different parts of the Transformer architecture, two sub-experiments were done. The first is changing the number of attention heads, and the second is changing the encoder/decoder depth. For both sub-experiments, a model is created where all other hyperparameters are kept fixed, trained for 3 epochs, evaluated on the validation set, and had its BLEU score calculated using the test set.

3 Results

3.0 Grid Search

The results of the grid search are shown in Table 2. We see that the best performing model is the one with encoder/decoder depth of 4, 8 attention heads, and model and feed-forward dimension of 256. Out of the top five configurations, four of them have an encoder/decoder depth of 4, and we see a trend of larger depth yielding higher BLEU scores where the average BLEU score of encoder/decoder depth of 2 is 0.8341 while 1 is 0.8137. Similarly, there exists a positive correlation betweeen model dimension and performance as 256 always outperforms 128. The effects of the number of attention heads and feed-forward dimension on performance are not as pronounced.

Enc/Dec Depth	Model Dim	Attention Heads	FF Dim	BLEU Score
1	128	2	256	0.7997
1	128	2	512	0.7986
1	128	4	256	0.8060
1	128	4	512	0.8074
1	128	8	256	0.8077
1	128	8	512	0.8107
1	256	2	256	0.8150
1	256	2	512	0.8202
1	256	4	256	0.8227
1	256	4	512	0.8184
1	256	8	256	0.8260
1	256	8	512	0.8317
2	128	2	256	0.8277
2	128	2	512	0.8270
2	128	4	256	0.8249
2	128	4	512	0.8284
2	128	8	256	0.8268
2	128	8	512	0.8271
2	256	2	256	0.8371
2	256	2	512	0.8387
2	256	4	256	0.8423
2	256	4	512	0.8416
2	256	8	256	0.8416
2	256	8	512	0.8461
4	128	2	256	0.8329
4	128	2	512	0.8392
4	128	4	256	0.8387
4	128	4	512	0.8354
4	128	8	256	0.8335
4	128	8	512	0.8379
4	256	2	256	0.8390
4	256	2	512	0.8434
4	256	4	256	0.8452
4	256	4	512	0.8466
4	**256**	**8**	**256**	**0.8466**
4	256	8	512	0.8499

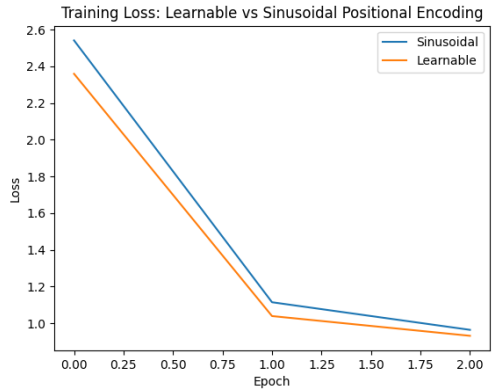
Table 2: Grid Search Results

3.1 Experiment 1

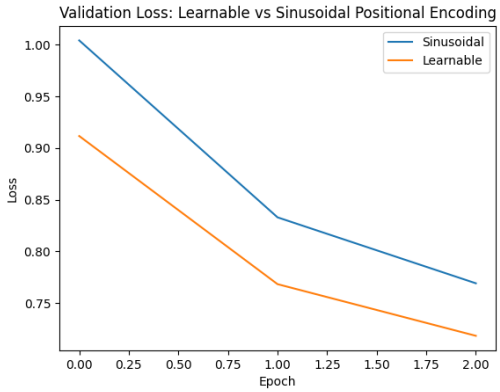
As we can see from Fig 3, the learnable positional encodings consistently have a lower training and validation loss than the sinusoidal ones, and both models follow a similar loss curve. We also see the same behavior when evaluating the decoding on the test set where the learnable positional encodings had a marginally higher score. In the example outputs, we see that both models produce similar or even the same output as shown in Example 2, but the alternative tends to produce longer sequences.

The BLEU score for each model reflects this as well where since the BLEU scores for each are close in range, the outputs are also relatively similar.

Ex 1
- **Input**: at least if be otherwise , there be not want other motif much more influential with him .
- **Ground Truth**: or at least if this were otherwise , there were not wanting other motives much more influential with him .



(a) Training loss



(b) Validation loss

Figure 3: Loss curves for both models

Model	Trainable Parameters	BLEU Score
Sinusoidal	31863657	0.8351
Learnable	31889257	0.8399

Table 3: Number of trainable parameters and BLEU scores for each model

- **Sinusoidal Prediction**: at least if being otherwise , there was not wanted other motives much more often with him .
- **Learnable Prediction**: at least if be otherwise , there is not wanting to the other motives much more to the wast with him .

Ex 2
- **Input**: turn from thy fierce wrath , repent of evil against thy people .
- **Ground Truth**: turn from thy fierce wrath , and repent of this evil against thy people .
- **Sinusoidal Prediction**: turn from thy fierce wrath , repent of evil against thy people .
- **Learnable Prediction**: turn from thy fierce wrath , repent of evil against thy people .

Ex 3
- **Input**: 6 : 22 now be make free from sin , become servant god , ye have your fruit unto holiness , end everlasting life .
- **Ground Truth**: 6 : 22 but now being made free from sin , and become servants to god , ye have your fruit unto holiness , and the end everlasting life .
- **Sinusoidal Prediction**: 6 : 22 now this is made free from sin , and became servant god , ye have your fruit unto holiness , and end everlasting life .
- **Learnable Prediction**: 6 : 22 now are made free from sin , and become the servant god , ye have your fruit unto holiness , and the ends everlasting lives .

Ex 4
- **Input**: then master of house be angry say his servant , go out quickly into street lane of city , bring in hither poor , maim , halt , blind .
- **Ground Truth**: then the master of the house being angry said to his servant , go out quickly into the streets and lanes of the city , and bring in hither the poor , and the maimed , and the halt , and the blind .
- **Sinusoidal Prediction**: then the master of the house was angry to say his servant , went out quickly into the street lane of the city , and brought in hither the poor , and maimed , and blind.
- **Learnable Prediction**: then the master of the house was angry said his servant , go out quickly into the streets lane of the city , bring in hither and the poor , maimed , and halted .

Ex 5
- **Input**: just so with head ; with difference : about head envelope , though not so thick , be of boneless toughness , inestimable by man who have not handle it .

- **Ground Truth**: just so with the head ; but with this difference : about the head this envelope , though not so thick , is of a boneless toughness , inestimable by any man who has not handled it .
- **Sinusoidal Prediction**: just so with head ; and with the difference : about head envelope , though not so thick , is of the envy , and the marquis by man who has not handled it .
- **Learnable Prediction**: just so with head ; with the difference : about the head and envelope , though not so thick , was of the interpretation <unk> , and vexation by a man who had not handle it .

Ex 6
- **Input**: jesus say unto her , do i condemn thee : go , sin more .
- **Ground Truth**: and jesus said unto her , neither do i condemn thee : go , and sin no more .
- **Sinusoidal Prediction**: jesus said unto her , do i condemn thee : go , sin more .
- **Learnable Prediction**: and jesus said unto her , did i condemn thee : go , sin more .

3.2 Experiment 2

Due to restrictions in time, only the greedy decoding strategy and beam search with width 3 were run fully, and the beam search with widths 5 and 10 were run on the first 100 examples within the test set from which we infer behavior.

We see that in the full run, greedy decoding and beam search produced nearly the same BLEU score with greedy performing better while beam search with beam width 3 took over twice as long.

When examining the partial results, we see that roughly doubling the beam width leads to doubling the time to decode each sample. Furthermore, beam search with width 5 outperforms width 3, but width 10 performed the worst out of all the beam search widths, but again, this is not a true reflection of the results as with this small sample size, beam search with width 3 outperforms greedy decoding when it does not when evaluating on the full test set.

Strategy	Time per Sample (s)	Avg Seq Length	BLEU Score
Greedy	0.5139	20.18	0.7299
Beam Width 3	2.4445	20.16	0.7298

Table 4: Full results of each decoding strategy

Strategy	Time per Sample (s)	Avg Seq Length	BLEU Score
Greedy	0.5339	22.91	0.8699
Beam Width 3	2.5916	22.90	0.8728
Beam Width 5	4.7212	22.88	0.8734
Beam Width 10	9.7749	22.84	0.8723

Table 5: Partial results of each decoding strategy

Ex 1
- **Input**: at least if be otherwise , there be not want other motif much more influential with him .
- **Ground Truth**: or at least if this were otherwise , there were not wanting other motives much more influential with him .
- **Greedy Prediction**: at least if was otherwise , and there was not wanted other motives much more influential with him .
- **Beam 3 Prediction**: at least if was otherwise , there was not wanting other motives much more influential with him .
- **Beam 5 Prediction**: at least if was otherwise , there was not wanting other motives much more influential with him .

- **Beam 10 Prediction**: at least if was otherwise , there was not wanting other motives much more influential with him .

Ex 2
- **Input**: 29 : 8 young men saw me , hide themselves : age arose , stand up .
- **Ground Truth**: 29 : 8 the young men saw me , and hid themselves : and the aged arose , and stood up .
- **Greedy Prediction**: 29 : 8 the young men saw me , and hide themselves : and the age arose , and stood up .
- **Beam 3 Prediction**: 29 : 8 the young men saw me , and hide themselves : and the age arose , and stood up .
- **Beam 5 Prediction**: 29 : 8 the young men saw me , and hide themselves : the age arose , and stood up .
- **Beam 10 Prediction**: 29 : 8 the young men saw me , and hide themselves : the age arose , and stood up .

Ex 3
- **Input**: in fact , felix have till now profess himself his firm ally , have on his part receive from franklin unequivocal proof of friendship ; for it must be tell that other morning , when it be felix ' s turn get breakfast , felix
- **Ground Truth**: in fact , felix had till now professed himself his firm ally , and had on his part received from franklin unequivocal proofs of friendship ; for it must be told that every other morning , when it was felix ' s turn to get breakfast , felix
- **Greedy Prediction**: in fact , felix had till now professed himself his firm firm , ally had on his part received from franklin to the proof proof of friendship ; for it must be told that other morning , when it was felix ' s turn got breakfast breakfast ,
- **Beam 3 Prediction**: in fact , felix had till now professed himself his firm firm , ally had on his part received from franklin to the proof proof of friendship ; for it must be told that other morning , when it was felix ' s turning to get breakfast ,
- **Beam 5 Prediction**: in fact , felix had till now professed himself his firm firm , ally had on his part received from franklin to the proof proof of friendship ; for it must be told that other morning , when it was felix ' s turning to get breakfast ,
- **Beam 10 Prediction**: in fact , felix had till now professed himself his firm firm , ally had on his part received from franklin to the proof proof of friendship ; for it must be told that other morning , when it was felix ' s turning to get breakfast ,

Ex 4
- **Input**: 19 : 10 men put forth their hand , pull lot into house them , shut door .
- **Ground Truth**: 19 : 10 but the men put forth their hand , and pulled lot into the house to them , and shut to the door .
- **Greedy Prediction**: 19 : 10 and the men put forth their hand , and pulled lot into the house to them , and shut the door .
- **Beam 3 Prediction**: 19 : 10 and the men put forth their hand , pulling lot into the house to them , and shut the door .
- **Beam 5 Prediction**: 19 : 10 and the men put forth their hand , pulling lot into the house to them , and shut the door .
- **Beam 10 Prediction**: 19 : 10 and the men put forth their hand , pulling lot into the house to them , and shut the door .

Ex 5
- **Input**: pass , if pas , void profound of unessential night receive him next , wide - gaping , with utter loss of be threatens him , plunge in abortive gulf .
- **Ground Truth**: these passed , if any pass , the void profound of <unk> night receives him next , wide - gaping , and with utter loss of being <unk> him , plunged in that abortive gulf .
- **Greedy Prediction**: pass , if pass , void profound of <unk> night received him next , wide - gaping , with uttered the loss of being <unk> him , plunged in a discontented gulf .

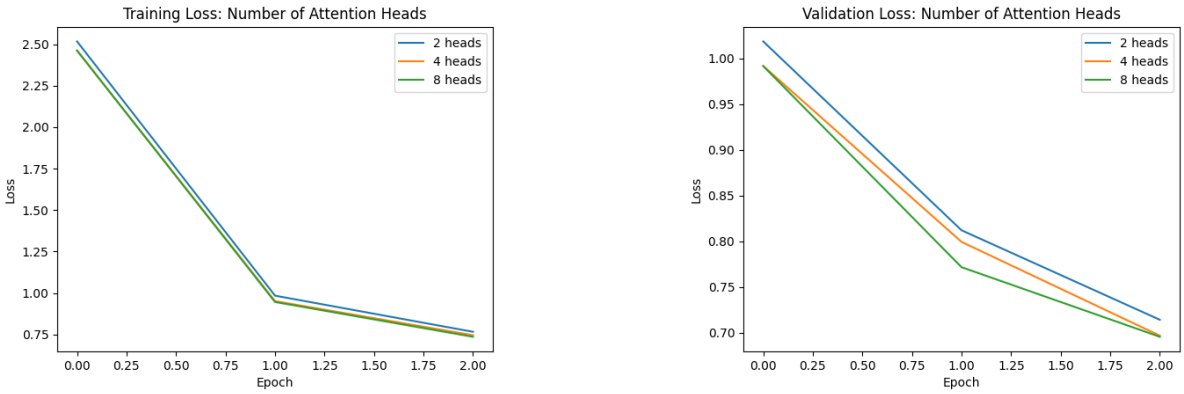
- **Beam 3 Prediction**: pass , if pass , void profound of <unk> night received him next , the wide - kick , with uttered the loss of being <unk> him , plunged in a discontented gulf .
- **Beam 5 Prediction**: pass , if pass , void profound of <unk> night received him next , the wide - kick , with uttered the loss of being <unk> him , plunged in a discontented gulf .
- **Beam 10 Prediction**: pass , if pass , void profound of <unk> night received him next , the wide - kick , with uttered the loss of being <unk> him , plunged in a discontented gulf .

Ex 6
- **Input**: 39 : 2 i be dumb with silence , i hold my peace , even from good ; my sorrow be stir .
- **Ground Truth**: 39 : 2 i was dumb with silence , i held my peace , even from good ; and my sorrow was stirred .
- **Greedy Prediction**: 39 : 2 i am dumb with silence , i hold my peace , even from good ; my sorrows is stir .
- **Beam 3 Prediction**: 39 : 2 and i was dumb with silence , and i hold my peace , even from good ; my sorrow was stir .
- **Beam 5 Prediction**: 39 : 2 i am dumb with silence , and i hold my peace , even from good ; my sorrow is stir .
- **Beam 10 Prediction**: 39 : 2 i am dumb with silence , i hold my peace , even from good ; my sorrow is stir .

3.3 Experiment 3

Experiment 3a

As shown in Fig. 4, there is an inverse relation between the number of heads and loss, but there is a less of difference between 4 and 8 heads than 2 and 4 heads. This is also reflected in the BLEU score of each model where 4 and 8 heads only differ by 0.001 while there is a near 0.4 increase in BLEU from 2 heads to 4 heads. Between all three models, the outputs from decoding have minimal difference.



(a) Training loss

(b) Validation loss

Figure 4: Loss curves for models with 2, 4, and 8 attention heads

Heads	BLEU Score
2	0.8426
4	0.8461
8	0.8462

Table 6: BLEU scores for different number of attention heads

Ex 1
- **Input**: at least if be otherwise , there be not want other motif much more influential with him .
- **Ground Truth**: or at least if this were otherwise , there were not wanting other motives much more influential with him .
- **2 Heads Prediction**: at least if was otherwise , there was not wanting to the other motives much more with him .
- **4 Heads Prediction**: at least if being otherwise , there is not wanted to other motives much more influential with him .
- **8 Heads Prediction**: at least if be otherwise , there be not want other motif more influential more influential with him .

Ex 2
- **Input**: turn from thy fierce wrath , repent of evil against thy people .
- **Ground Truth**: turn from thy fierce wrath , and repent of this evil against thy people .
- **2 Heads Prediction**: turn from thy fierce wrath , repented of evil against thy people .
- **4 Heads Prediction**: turn from thy fierce wrath , and repented of evil against thy people .
- **8 Heads Prediction**: turn from thy fierce wrath , repenting of evil against thy people .

Ex 3
- **Input**: 6 : 22 now be make free from sin , become servant god , ye have your fruit unto holiness , end everlasting life .
- **Ground Truth**: 6 : 22 but now being made free from sin , and become servants to god , ye have your fruit unto holiness , and the end everlasting life .
- **2 Heads Prediction**: 6 : 22 now these are made free from sins , and become the servants god , ye have your fruit unto holiness , the ends everlasting life .
- **4 Heads Prediction**: 6 : 22 now is made free from sins , and become servants and god , and ye have your fruit unto holiness , and the end everlasting life .
- **8 Heads Prediction**: 6 : 22 now be made free from sin , and became servants god , ye have your fruit unto holiness , and end to everlasting life .

Ex 4
- **Input**: then master of house be angry say his servant , go out quickly into street lane of city , bring in hither poor , maim , halt , blind .
- **Ground Truth**: then the master of the house being angry said to his servant , go out quickly into the streets and lanes of the city , and bring in hither the poor , and the maimed , and the halt , and the blind .
- **2 Heads Prediction**: then the master of the house was angry said his servants , went out quickly into the street of the city , brought in hither poor , and the almighty , and halted , and the blind , blind .
- **4 Heads Prediction**: then the master of the house was angry said to his servants , went out quickly into the street of the city , and bring in hither poor , and the maim , halt , blind .
- **8 Heads Prediction**: then the masters of the house was angry said to his servant , go out quickly into the streets lanes lane of the city , brought in hither poor , maim , halted , blind , blind .

Ex 5
- **Input**: just so with head ; with difference : about head envelope , though not so thick , be of boneless toughness , inestimable by man who have not handle it .
- **Ground Truth**: just so with the head ; but with this difference : about the head this envelope , though not so thick , is of a boneless toughness , inestimable by any man who has not handled it .
- **2 Heads Prediction**: just so with the head ; with the difference : about the head envelope , though not so thick , was of the naomi <unk> , and the <unk> by man who had not handled it .
- **4 Heads Prediction**: just so with head ; with difference : about head envelope , though not so thick , be of foregoing , inestimable by man who have not it .

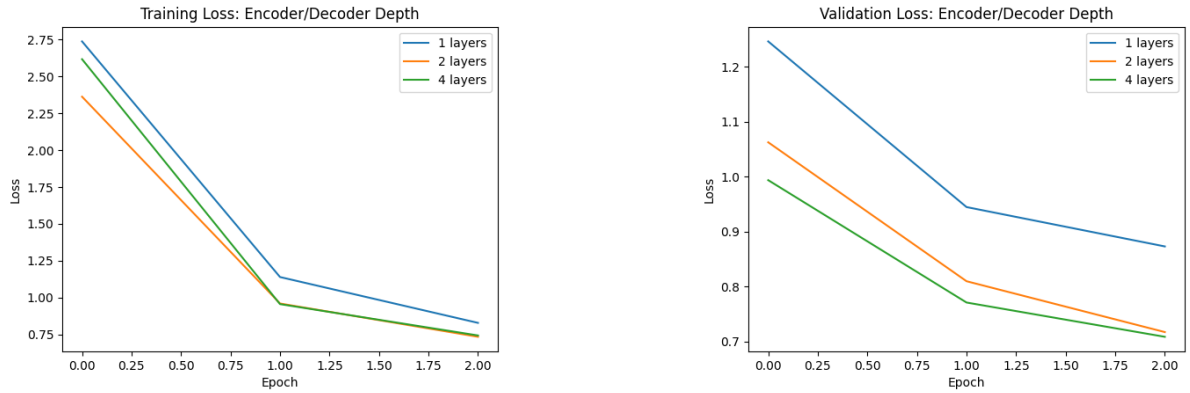
- **8 Heads Prediction**: just so with head ; with difference : about head envelope , though not so thick , be of port , encamp by man who have not it .

Ex 6

- **Input**: jesus say unto her , do i condemn thee : go , sin more .
- **Ground Truth**: and jesus said unto her , neither do i condemn thee : go , and sin no more .
- **2 Head Prediction**: and jesus said unto her , do i condemn thee : go , and sins more .
- **4 Head Prediction**: jesus said unto her , do i condemn thee : go , sin no more .
- **8 Head Prediction**: jesus said unto her , do i condemn thee : go , sin more .

Experiment 3b

Unlike Experiment 3a, increasing the number of layers does not always lead to a decrease in loss as seen in the training loss in Fig. 5. Furthermore, the BLEU score of the model with 2 layers outperforms the 4 layer model by 0.003.



(a) Training loss

(b) Validation loss

Figure 5: Loss curves for models with 1, 2, and 4 encoder/decoder layers

Layers	BLEU Score
1	0.8253
2	0.8454
4	0.8412

Table 7: BLEU scores for each model

Ex 1

- **Input**: at least if be otherwise , there be not want other motif much more influential with him .
- **Ground Truth**: or at least if this were otherwise , there were not wanting other motives much more influential with him .
- **1 Layer Prediction**: at least if be otherwise , there was not wanted the other motives much more influential with him .
- **2 Layer Prediction**: at least if that was otherwise , there was not wanting other motives much more influential with him .
- **4 Layer Prediction**: at least if this was otherwise , there was not wanting the other motives much more and with him .

Ex 2

- **Input**: turn from thy fierce wrath , repent of evil against thy people .
- **Ground Truth**: turn from thy fierce wrath , and repent of this evil against thy people .
- **1 Layer Prediction**: turn from thy fierce wrath , repent of evil against thy people .

- **2 Layer Prediction**: turn from thy fierce wrath , repent of evil against thy people .
- **4 Layer Prediction**: turning from thy fierce wrath , repented of evil against thy people .

Ex 3
- **Input**: 6 : 22 now be make free from sin , become servant god , ye have your fruit unto holiness , end everlasting life .
- **Ground Truth**: 6 : 22 but now being made free from sin , and become servants to god , ye have your fruit unto holiness , and the end everlasting life .
- **1 Layer Prediction**: 6 : 22 now be made free from sin , and become servants god , ye have your fruit unto holiness , ending life .
- **2 Layer Prediction**: 6 : 22 now is made free from sin , and became the servant god , ye have your fruit unto holiness , and end everlasting life .
- **4 Layer Prediction**: 6 : 22 now is made free from sin , and become servant god , and ye have your fruit unto holiness , and end everlasting life .

Ex 4
- **Input**: then master of house be angry say his servant , go out quickly into street lane of city , bring in hither poor , maim , halt , blind .
- **Ground Truth**: then the master of the house being angry said to his servant , go out quickly into the streets and lanes of the city , and bring in hither the poor , and the maimed , and the halt , and the blind .
- **1 Layer Prediction**: then master of the house was angry and said his servants , went out quickly into the streets and lanes of the city , and brought in hither poor , halt , halt , and blind .
- **2 Layer Prediction**: then the master of the house was angry said to his s ervants , going out quickly into the street of the city , brought in hither poor , and maimed , halting , and halting
- **4 Layer Prediction**: then the master of the house was angry said his servant , go out quickly into the streets lane of the city , brought in hither poor , and maimed , halt , blinds .

Ex 5
- **Input**: just so with head ; with difference : about head envelope , though not so thick , be of boneless toughness , inestimable by man who have not handle it .
- **Ground Truth**: just so with the head ; but with this difference : about the head this envelope , though not so thick , is of a boneless toughness , inestimable by any man who has not handled it .
- **1 Layer Prediction**: just so with head ; with difference difference : about head envelope envelope , though not so thick , be of mischievous , he by man who have not handle it .
- **2 Layer Prediction**: just so with head ; with difference : about head env elope , though not so thick , be of steep <unk> , inestimable by man who have not handle it .
- **4 Layer Prediction**: just so with head ; with difference : about head env.elope , though not so thick , be of hal , divert by man wh.o have not handle it .

Ex 6
- **Input**: jesus say unto her , do i condemn thee : go , sin more .
- **Ground Truth**: and jesus said unto her , neither do i condemn thee : go , and sin no more .
- **1 Layer Prediction**: jesus say unto her , do i condemn thee : go , sin no more .
- **2 Layer Prediction**: jesus said unto her , do i condemn thee : go , sin no more .
- **4 Layer Prediction**: jesus said unto her , do i condemn thee : go , sin more .

4 Analysis

4.0 Grid Search

In the grid search, we see that the largest contribution to increasing performance was increasing the encoder/decoder depth and the model dimension. While the Transformer is a black box architecture where we cannot fully understand the outputs of a single pass of the encoder and decoder, it can be inferred that each pass through each layer gives the model more information and allows for deeper learning. Similarly, increaseing the model dimension allows for more information to be retained. The instances of increasing attention heads leading to a slight drop in performance may be due to overfitting the training data.

4.1 Experiment 1

The learnable positional encodings performing better is expected as even though sinusoidal works well, there may be more information encoded in the relative position of words that it cannot learn and therefore not capture. With an increase of around 20,000 parameters, the training time did not significantly increase

4.2 Experiment 2

In theory, because beam search maintains the best w (w being the width) paths, it should produce results that are globally more correct compared to greedily choosing the immediate most correct token, but this was not reflected in the full runs of Experiment 2. Even in the partial runs, the largest beam search width performed the worst out of all the beam searches. This may be due to the max length being capped at 50, or the average length of sequences being near to 10. However, because the gains in BLEU score are marginal while the increase in decoding time per sample is exponential, the experiment does not support using beam search over greedy decoding.

4.3 Experiment 3

Experiment 3a: Number of Attention Heads

As mentioned in the grid search analysis, the number of attention heads did not produce a significant change in BLEU score with the difference between the best and worst performing models being 0.004. Another possibility for why increasing the number of attention heads does not significantly increase performance is that the task is simple enough where there are not many different relations for each attention head to learn.

Experiment 3b: Encoder/Decoder Depth

From the grid search 4 layers was expected to perform the best; however, it performed marginally worse than 2 layers. We see in Fig. 5 that in the final epoch, the 2 layer model has a lower training loss than the 4 layer one, so it is possible that by this epoch, the 4 layer one started to overfit the training data in a way that did not affect evaluation on the validation set.

5 Conclusion

When using an encoder-decoder transformer for the task of function word restoration, we see that learnable position encodings outperform static sinusoidal positional encodings; greedy decoding is comparable and in some instances outperform beam search while vastly saving decoding time; and larger models usually outperform smaller models with few instances of overfitting. For this project, the largest limitation was time required to compute the BLEU scores. Unlike in the previous project, access to a GPU was available, so training the models was greatly sped up. But the time it took to compute BLEU did not change depending on hardware. Furthermore, in Experiment 2, a preloaded model from the grid search was used rather than truly training a model to convergence, so the BLEU scores may actually be higher, and the difference between the different decoding strategies could be more pronounced. Some future improvements would be to rerun Experiment 2 to obtain true values for the different decoding strategies and examining other aspects of the Transformer architecture such as the attention mechanism and residual connection.

References

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.