Sequence-to-Sequence Transformer for Text Reconstruction: An Experimental Analysis

Introduction

This work addresses recovering original sentences from preprocessed versions where determiners, conjunctions, and prepositions have been removed and remaining words lemmatized. For example, transforming "little man begin his recital" to "the little man began his recital".
The Transformer architecture (Vaswani et al., 2017) uses self-attention to capture long-range dependencies. This project implements an encoder-decoder Transformer and evaluates positional encoding strategies (sinusoidal vs. learnable) and architecture variants (attention heads and model depth).

Methods

Transformer Architecture

The implementation follows the standard encoder-decoder architecture:
Encoder: Token embeddings, positional encoding, and stacked layers with multi-head self-attention and feedforward networks.
Decoder: Masked self-attention, cross-attention to encoder outputs, and feedforward networks.
Multi-Head Attention: Attention(Q, K, V) = softmax(QK^T / √d_k)V

Positional Encoding

Sinusoidal: Fixed functions with no trainable parameters.
Learnable: Trainable embeddings adding positional information through learned parameters.

Implementation

Training uses teacher forcing, CrossEntropyLoss, Adam (lr=0.001), batch size 32, 3 epochs. Fixed hyperparameters: d_model=128, d_ff=512, dropout=0.1, max_len=50.

Results

Experiment 1: Positional Encoding

Configuration	Parameters	Test Loss	BLEU
Sinusoidal (2L-4H)	14,816,889	0.8364	0.7555
Learnable (2L-4H)	14,777,833	0.9646	0.7433

Epoch	Sinusoidal Train	Sinusoidal Val	Learnable Train	Learnable Val
1	~3.88	~2.16	3.8816	2.1569
2	~2.01	~1.27	2.0105	1.2646
3	~1.35	~0.95	1.3524	0.9511

Experiment 3: Architecture Variants

Config	L	H	Loss	BLEU	Params	Time
3a	1	2	1.1463	0.7354	11.9M	26s
3b	1	4	1.0916	0.7365	12.5M	73s
3c	1	8	1.1019	0.7367	13.7M	26s
3d	2	2	0.8519	0.7526	13.1M	84s
3e	2	4	0.8259	0.7611	14.3M	43s
3f	2	8	0.8229	0.7595	16.7M	87s
3g	4	2	0.7662	0.7722	14.3M	81s
3h	4	4	0.7447	0.7715	15.7M	110s
3i	4	8	0.7731	0.7657	18.4M	113s

Analysis

The sinusoidal positional encoding outperformed learnable embeddings across all metrics. Depth provided larger gains than increasing attention heads. Training time scaled strongly with number of layers, while BLEU improvements showed diminishing returns beyond 2 layers.

Conclusion

Sinusoidal encoding outperforms learnable embeddings. Depth offers consistent improvements; width offers minimal benefit. 2L-4H is the best efficiency-accuracy configuration. Future work includes longer training, larger models, hybrid encodings, and broader evaluation.