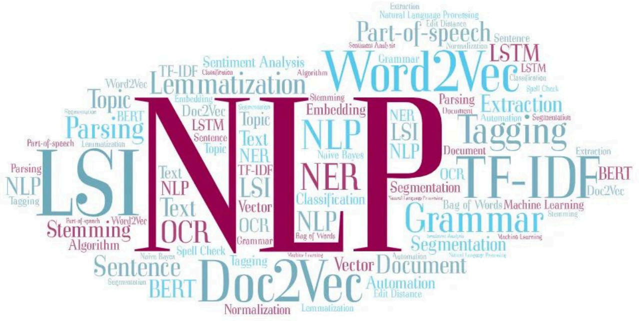
Transformers in NLP

Ayushi Bawari (9920103216), Nipun Singh Rathore (9920103205)



INTRODUCTION

In the ever-evolving landscape of Natural Language Processing (NLP), the utilisation of pre-trained transformer models has emerged as a cornerstone for various language understanding tasks. These models, such as BERT & RoBERT have demonstrated remarkable effectiveness in capturing rich contextual embeddings, thereby significantly advancing the state-of-the-art in NLP applications. Yet, the practical implementation and optimisation of these models for specific projects pose unique challenges, from computational resource constraints to domain-specific requirements.

PROBLEM STATEMENT

This project aims to comprehensively study transformer models in natural language processing, focusing on architectural design, pre-training methods, fine-tuning, and model training from scratch. By grasping core concepts like self-attention and positional encoding, we seek to enhance performance in tasks such as text classification, sentiment analysis, and language translation. Additionally, we plan to apply a custom-trained model for question answering within a web application, exploring practical implications for NLP applications.

SOLUTION APPROACH

Scaled Dot-Product Attention: Employed for computing attention scores, enabling the model to focus on relevant parts of the input sequence.

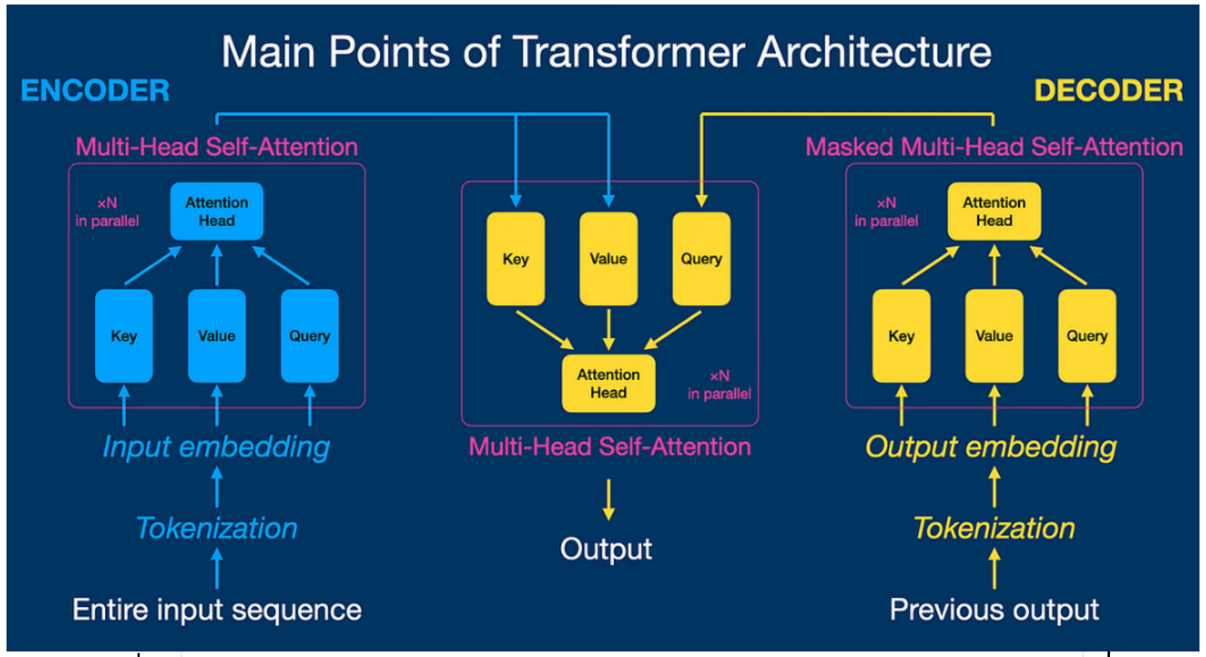
Encoder: Incorporates multi-head self-attention and feed-forward networks, with dropout and layer normalisation for regularisation and stability.

Decoder: Utilises multi-head self-attention in two stages, feed-forward networks, and layer normalisation to capture complex patterns in the data.

Positional and Word Embeddings: Provide spatial and semantic information to the model, enhancing understanding of token order and meaning.

Masked Language Model and Next Sentence Prediction: Training objectives for learning contextual representations and relationships between sentences.

Web Application: Developed to showcase Transformer



Main Points of Transformer Architecture

ENCODER

Multi-Head Self-Attention

×N in parallel

Attention Head

Key Value Query

Input embedding

Tokenization

Entire input sequence

Key Value Query

Attention Head

×N in parallel

Multi-Head Self-Attention

Output

DECODER

Masked Multi-Head Self-Attention

×N in parallel

Attention Head

Key Value Query

Output embedding

Tokenization

Previous output

ISSUES FACED

1. Model Fine-tuning: Pre-trained BERT fine-tuned on SQuAD, primarily for question-answering, potentially limiting performance for other use cases.
2. Performance Variability: Application performance may vary based on task complexity, with slower response times for complex questions or larger models.

FUTURE WORK

1. Custom Transformer Model Completion: Refine and finalize the custom transformer model, addressing challenges and optimizing architecture and training procedures.
2. Web Application Integration: Prioritize integrating the custom transformer model into the web application for seamless performance and user experience.
3. Explore Additional NLP Tasks: Extend the project to include tasks like sentiment analysis, named entity recognition, and text generation using the custom transformer model.

REFERENCES

[1] Devlin, Jacob et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." North American Chapter of the Association for Computational Linguistics (2019).

[2] Salazar, Julian et al. "Masked Language Model Scoring." Annual Meeting of the Association for Computational Linguistics (2019).