**TECHNICAL REPORT WRITING FOR CA2 EXAMINATION**



**NAME**- SAPNADARSHI SINGHA

**STERAM-** COMPUTER SCIENCE AND BUSINESS SYSTEM

**ROLL NO** – 18731122014

**SUBJECT NAME**- MACHINE LEARNING

**SUBJECT CODE**- ESC CSBS 502

**YEAR**- THIRD

**SEMESTER**- FIFTH

**TOPIC-** Advances in NLP using transformer models and attention mechanisms

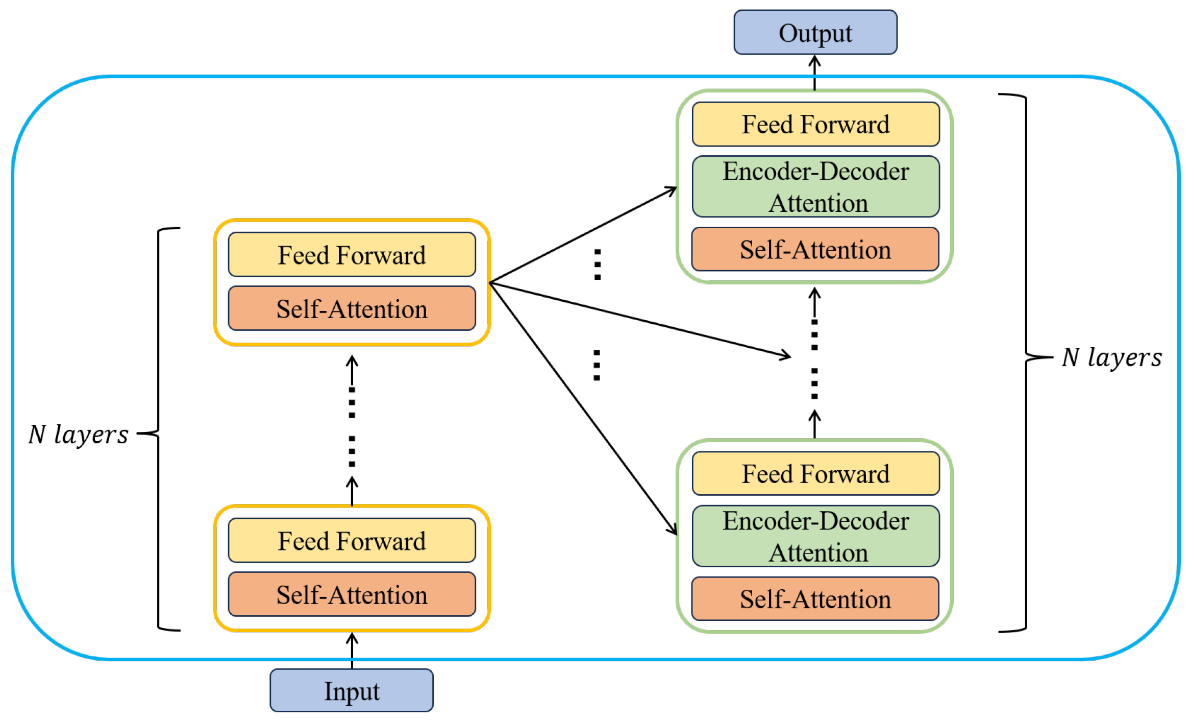
* **INTRODUCTION**

Natural Language Processing (NLP) is a branch of artificial intelligence that studies the interaction between computers and human language. It entails creating algorithms and models that allow machines to understand, interpret, generate, and respond to human language in a meaningful and practical manner. NLP covers a wide range of tasks, including language translation, sentiment analysis, text summarization, speech recognition, and question answering. NLP aims to bridge the gap between human communication and computer understanding by combining linguistic rules, statistical methods, and advanced machine learning techniques such as transformers and deep learning.

WHAT IS TRANSFORMER MODEL AND ATTENTION MECHANISM?

Vaswani et al. introduced the Transformer model, a deep learning architecture, which has revolutionized Natural Language Processing (NLP) by replacing traditional convolutional networks with self-attention mechanisms. The model consists of an encoder-decoder structure, with the encoder processing input data and the decoder producing output. Each component is composed of layers that include self-attention and feedforward neural networks.

**The attention mechanism,** particularly self-attention, is central to the Transformer's success. It allows the model to weigh the importance of each word in a sentence relative to others, regardless of their position, enabling it to capture long-range dependencies and complex relationships within the text. This is achieved through scaled dot-product attention, which compares queries, keys, and values of the input data, and multi-head attention, which performs this process multiple times in parallel to capture different aspects of the relationships. The Transformer’s architecture enables faster training and better handling of long sequences compared to previous models.



* **METHODOLOGY OF ATTENTION MECHANISM**

The attention mechanism works by allowing a model to focus on different parts of the input sequence while processing data, allowing it to capture word relationships regardless of position. Here's how it works, step by step:

1. Input Representations.

Each word in the input sequence is represented by three vectors: queries (Q), keys (K), and values (V). These vectors are obtained by linearly transforming the input embeddings.

1. Getting Attention Scores

The attention mechanism uses the dot product of a word's query vector and the key vectors of all the words in the sequence to calculate the relevance or similarity of each word to the other words. The gradients are stabilized during training by scaling these dot products by dividing by the square root of the key vectors' dimension.

**Attention Scores=QKT where (Q= Query Matrix, KT = transpose of key matrix)**

1. Normalization with Softmax

The scaled scores are transformed into probabilities, or attention weights, by running them through a softmax function. These weights show the relative importance of each word to the word that is currently being processed.

**Attention Weights=softmax (QKT​/sqrt(dk))**

1. Weighted Sum of Values

A weighted sum of the value vectors is calculated using the attention weights. Each word is given a new representation by this sum, which takes into account context from the words in the sequence that are most pertinent.

**Output=Attention Weights × V**

1. Multi-Head Attention

Multi-head attention operates the attention mechanism multiple times in parallel (multiple heads), each focusing on a different aspect of the relationships in the sequence, as opposed to applying attention just once. The final result is obtained by concatenating and linearly transforming the outputs from these heads.

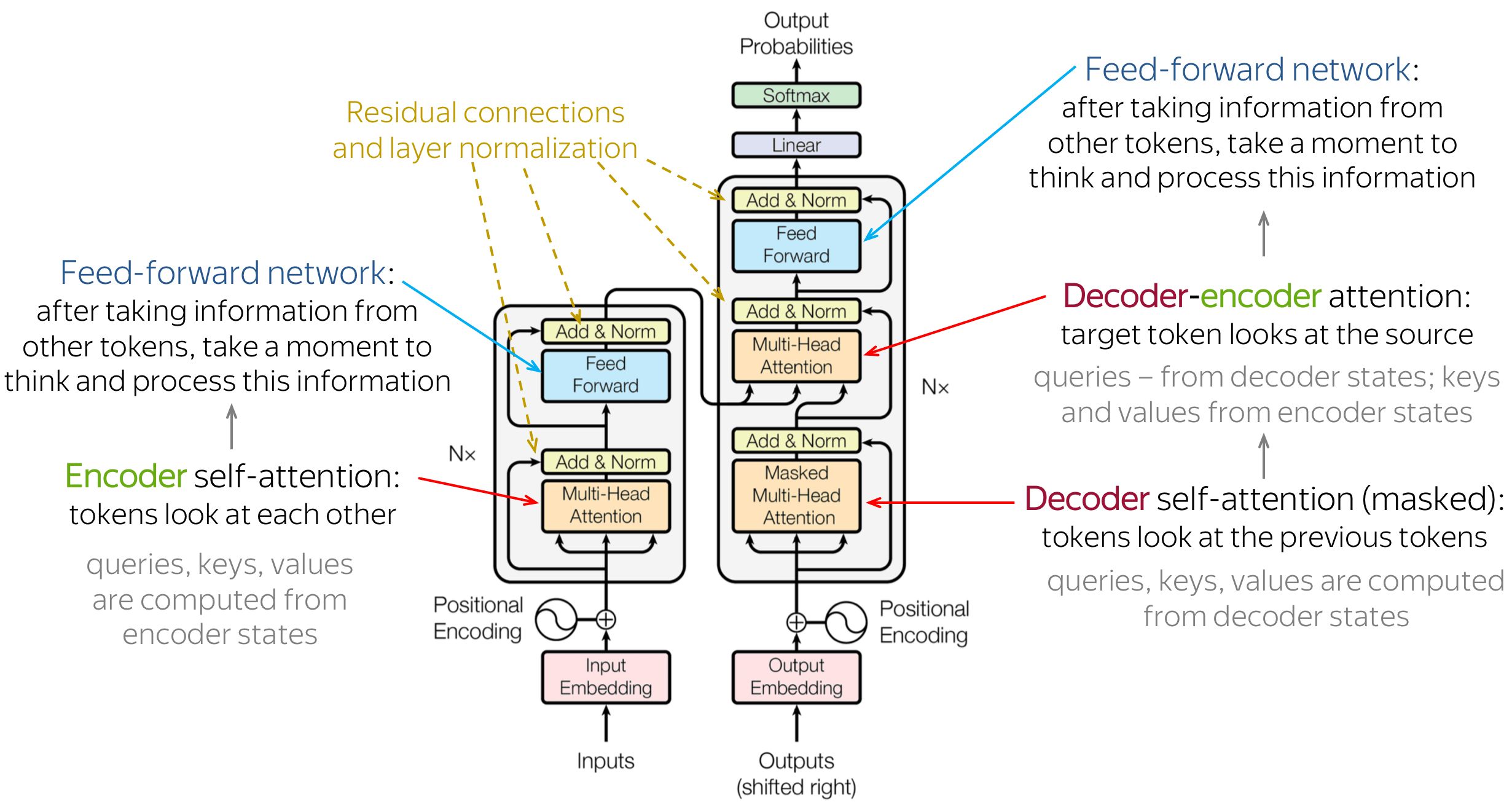
1. Position Encoding

Since self-attention processes input data in parallel, positional encodings are added to the input embeddings to provide the model with information about the order of tokens in the sequence.

* **TRANSFORMER MODEL**

The Transformer model is a deep learning architecture with an encoder-decoder structure that is used to process sequences, such as text

1. Embedding: To capture the order of the tokens, positional encodings are added after the input tokens are converted into continuous vectors.
2. Encoder: Layers of feed forward neural networks and self-attention are present in the encoder. By using self-attention, the model is able to capture dependencies throughout the sequence by evaluating the relative importance of each token.
3. Self-Attention: This technique calculates the connections among every token in the input, enabling the model to concentrate on pertinent segments of the sequence.
4. Decoder: Produces output sequences by utilizing cross-attention to absorb data from the encoder's outputs and masked self-attention to focus only on previous tokens. It then refines predictions using feedforward layers.

WORKFLOW DIAGRAM OF TRANSFORMER MODEL

* **How transformer models and attention mechanism advances NLP**

The attention mechanism has significantly advanced Natural Language Processing (NLP), especially in the context of Transformer models.

1. Improved comprehension of context  
   Global Context: Attention mechanisms enable the model to take into account the entire sequence at once, in contrast to conventional models that process text sequentially. This implies that every word is able to attend to every other word, capturing contextual subtleties and long-range dependencies that are essential for comprehending meaning.
2. Efficiency and Parallelization  
   Parallel Processing: Attention mechanisms allow input sequences to be processed in parallel, in contrast to sequential models. This greatly accelerates training and inference, enabling the handling of complex models and large-scale datasets.
3. Adaptability and Versatility  
   Adaptability: Transformers can be used for a wide range of natural language processing (NLP) tasks, including question-answering, summarization, translation, and text generation. Their capacity to produce detailed, context-aware representations accounts for their adaptability.
4. Flexibility  
   Big Models: Attention mechanisms make it easier to scale up models, which enables the development of extremely strong and large models like BERT and GPT-3. These large models make use of enormous amounts of data and computational power to achieve state-of-the-art performance on a variety of tasks.

* **CONCLUSION**

In conclusion, the Transformer model's creative application of parallel processing and self-attention mechanisms has completely transformed the field of natural language processing. It gets around the drawbacks of more conventional models like RNNs and LSTMs by effectively capturing complex dependencies in sequences. With the help of self-attention, the encoder-decoder structure, the Transformer can process context and long-range dependencies more quickly and accurately. Because of its adaptability and scalability, NLP tasks have advanced significantly, setting new performance standards for text generation, translation, and summarization, among other applications. The Transformer is a key component of many cutting-edge models, including BERT andGPT, and it keeps advancing the field of machine learning and natural language generation.

* **REFERENCE**

1. [**https://toloka.ai/blog/transformer-architecture/**](https://toloka.ai/blog/transformer-architecture/)
2. [**https://medium.com/machine-intelligence-and-deep-learning-lab/transformer-the-self-attention-mechanism-d7d853c2c621**](https://medium.com/machine-intelligence-and-deep-learning-lab/transformer-the-self-attention-mechanism-d7d853c2c621)
3. [**https://www.linkedin.com/pulse/transformer-model-neural-network-which-uses-attention-tejas-bankar/**](https://www.linkedin.com/pulse/transformer-model-neural-network-which-uses-attention-tejas-bankar/)
4. [**https://machinelearningmastery.com/the-transformer-model/**](https://machinelearningmastery.com/the-transformer-model/)
5. [**https://www.mdpi.com/1999-4893/17/1/37**](https://www.mdpi.com/1999-4893/17/1/37)