**Question 1: The Transformer – An Overview (5 Points)**

Write an essay:

* To present an overview of the Transformer neural network.
* The overview should include some history of its advent, general information about its AI-related technology and its applications

**Answer:**

The transformer is the backbone for the new age Large language Models like ChatGPT and Gemini. The neural network Transformer is relatively new with its origin marking to 2017. Some ground-breaking Transformer models developed by many researchers over the past 5 years since the development of the first Transformer in 2017.

**History and Advent:**

The below diagram shows the timeline of origin and evolution of the transformer.

A graph of company logos

AI-generated content may be incorrect.

Google released its Transformer on June 12th 2017. The architecture of the Transformer is shown in the diagram below.

LSTM and RNNs were the predecessors for the transformer. However, there were many challenges with LSTMs and RNNs and they were not providing appropriate support for processing Natural Language efficiently.

A diagram of a network

AI-generated content may be incorrect.

**General information about its AI-related technology and its applications**

The transformer is a ground breaking part of Artificial intelligence. The Transformer is not just a neural network. It represents a disruptive gigantic shift in the vast landscape of Natural Language Processing (NLP) and Artificial Intelligence (AI).

The Transformer’s revolutionary features have led to disruptive stunning achievements and advancements in many AI fields such as Natural Language Understanding (NLU), Language Generation, and Conversational AI, to name a few.

The Transformer will remain a foundational neural network that continues shaping the future of language-based artificial intelligence (AI), especially the generative AI and conversational AI. Its impacts on society will be deep and broad.

“**Attention Is All You Need**!”

The Transformer and Attention-Based Neural Networks: A revolution in the world of Artificial Intelligence (AI) Natural Language Processing (NLP).

**Question 2: The Transformer – The Neural Network Architecture (10 Points)**

Write an essay:

* To present an overview of the architecture of the Transformer neural network.
* To describe in detail each operational block of the architecture, including how each block works
* To provide a complete, simple example – end to end: from inputs to outputs – to demonstrate the overall operation process of the Transformer, including the role of each block of the architecture.

**Answer:**

In this section we will present an overview of the architecture of the Transformer neural network. We also describe in detail each operational block of the architecture, including how each block works. We will also provide a complete, simple example – end to end: from inputs to outputs – to demonstrate the overall operation process of the Transformer, including the role of each block of the architecture.

The diagram below shows the model architecture of the transformer.

A diagram of a software algorithm

AI-generated content may be incorrect.

One of the key components of the transformer’s architecture is encoder-decoder.

**Encoder-Decoder**

The encoder decoder has the following foundational revolutionary core features:

**Parallelize Computations:** Transformers discard the sequential processing paradigm of earlier recurrent neural networks (RNNs) in favor of parallel computation facilitated by attention mechanisms. This dramatically speeds up training and inference.

**Handle Long-Range Dependencies:** RNNs tend to struggle with long-range dependencies within sequences. The Transformer, through its attention mechanism, directly calculates relationships between any two elements in the sequence, regardless of distance, fostering a better grasp of context.

**Encoder:** The encoder's job is to process the input sequence (for example, a sentence of text) and generate a rich, contextualized representation of that input.

**Decoder:** The decoder takes the encoder's output and, in an auto-regressive manner (one step at a time), generates the target sequence (for example, a translation of the sentence).

**Multi-Head Attention Mechanism:** This is the Heart and Mind of the Transformer. It works as follows:

**Queries, Keys, and Values:** For each word in the input sequence, the Transformer creates three vectors: a query vector (Q), a key vector (K), and a value vector (V). These vectors are learned representations of the word itself.

**Scaled Dot-Product Attention:** The query vector of a word is compared to the key vectors of all words in the sequence, including itself. Similarity scores (via dot product) are calculated, and softmax is applied to normalize these scores into probabilities. These represent how much "attention" a word should pay to every word.

**Weighted Sum:** Finally, the value vectors are multiplied by their corresponding attention scores and summed up. This becomes the attention-enhanced output representation of the current word.

**Multi-Head:** Multiple sets of these Q, K, V transformations happen in parallel (forming "heads"), creating multiple diverse representations of each word. These representations are concatenated and linearly projected to yield a final unified representation for each word.

**Attention:** The model can weigh the importance of different words in an input sequence and transform them effectively to understand the interconnected nuances of the statement.

**Multi-Head Self-Attention:**

Self-attestation enables every word in the input sequence to “attend” to all other words and identify the relevance of each word in providing the context. This component performs this operation multiple times parallelly with each head focusing on different scenarios of the relationship between each word, hence enriching the representation of each word.

**Positional Encoding:** as transformers do not process sequences recurrently, the placement of words in a sequence needs to be encoded explicitly. Vectors called positional encoding is used to facilitate the same.

**Feed-Forward Networks:** pointwise, fully-connected feed-forward neural networks are present in the encoder and decoder, adding further non-linear transformations to the output which is based on “attention”. This improvises the expressiveness of the model.

**Main Layers of the Transformer: Encoder**

**Input Embeddings:** The input words are converted into word embeddings (dense vector representations).

**Multi-Head Attention (Self-Attention):** Words "attend" to each other within the input sequence.

**Layer Normalization:** Help stabilization during training.

**Feed-Forward Network:** Further refine the representations of words

**Residual Connections:** Shortcut connections add the input to the output of each sub-layer, making training more manageable.

A Transformer has stacked encoder layers (usually 6 or more), with each layer refining the representations of words or contents from the previous.

**Main Layers of the Transformer: Decoder**

**Masked Multi-Head Attention:** Ensure, during output generation, that dependencies only consider previously generated words (auto-regressive property).

**Encoder-Decoder Attention:** Connect decoder to encoder output, helping it focus on relevant areas of the input sequence.

**Layer Normalization**

**Feed-Forward Network**

**Residual Connections**

**Final Linear Layer and Softmax:** Transform the decoder's output into logits (raw predictions) and further into a probability distribution over the vocabulary.

**Complete, simple example – end to end: from inputs to outputs – to demonstrate the overall operation process of the Transformer, including the role of each block of the architecture.**

**Example:**

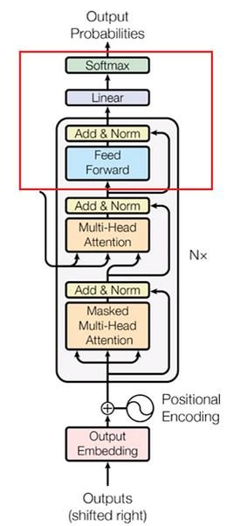
**TO DO:** Find out how relevant a particular word (French) is w.r.t to other words in that sentence.

**OUTPUT:** Attention vectors

* Each vector represents the relevance of a word (French) w.r.t other words.
* Every word is an attention vector
* Capture the contextual relationship between the word (French) and other words in that sentence.
* “Masked”: Learning mechanism in artificial neural networks (ANN):
* Give an English word.
* First, the transformer translates the English word into the French version by itself using previous results.
* Then the transformer matches and compares the predicted French word with the actual correct French word that is fed into the decoder block.
* After comparing both, the transformer updates its matrix value, through which the transformer can learn the correct translation after several iterations.
* “Masked”: Learning mechanism in artificial neural networks (ANN):
* To teach the transformer and make it learn better: Hide or mask the next French word.
* At first the transformer predicts the next word by itself using previous results, without knowing the real translated word.
* Of course, it makes no sense if it is shown the next French word →Hide or mask the actual French word.
* “Masked”: Learning mechanism in artificial neural networks (ANN):
* ENCODER:
* It is possible to take any word from the English sentence.
* DECODER:
* It is possible to only take the previous word of the French sentence.
* Therefore, while performing the parallelization with matrix operation, it should ensure the matrix should mask the words appearing later by transforming them into 0’s so that the attention network (decoder) cannot use them.
* Decoder: Multi-Head Attention (Self-Attention)
* a.k.a. Encoder-Decoder Attention Block:
* The resulting attention vectors from the previous decoder layer are passed into the Decoder Multi-Head Attention Block.
* The attention vectors from the Encoder Block are also passed into the Decoder Multi-Head Attention Block.
* The results from the encoder block comes into the picture.
* That’s why it is called Encoder-Decoder Attention Block.
* Decoder: Multi-Head Attention (Self-Attention)
* a.k.a. Encoder-Decoder Attention Block:
* There is one vector of every word for each English and French sentence.
* This block (Multi-Head Attention in Decoder)
* Do the mapping of English and French words
* Find out the relation between English and French words.
* Therefore, Multi-Head Attention in Decoder:
* Where the main English to French word mapping happens
* Decoder: Multi-Head Attention (Self-Attention)
* a.k.a. Encoder-Decoder Attention Block:

OUTPUTS:

* Attention vectors for every word in English and French sentences.
* Each vector represents the relationship with other words in both languages.
* Decoder: Feed Forward Neural Networks (FFNN): Linear Layers
* Each attention vector is passed into a feed-forward neural network.
* FFNN transforms output vectors into some form which is easily acceptable by another decoder block or a linear layer.
* A Linear Layer, a.k.a. Flattened Layer, is a feed-forward layer belonging to the feed forward neural network.
* It is used to expand the dimensions into numbers of words in the French language after translation.
* Decoder: Softmax Layer
* The outputs from the last Feed Forward Linear Layer are passed through a Softmax Layer,
* Transform the input into a probability distribution.
* The final outputs are resulting words produced with the highest probability after translation.

Source: Arxiv: 1706.03762

**Conclusion:**

In this section we presented an overview of the architecture of the Transformer neural network. We also described in detail each operational block of the architecture, including how each block works. We will also provided a complete, simple end to end example from inputs to outputs to demonstrate the overall operation process of the Transformer, including the role of each block of the architecture.

**Question 3: The Transformer – A Revolutionary Achievement in AI NLP (15 Points)**

Write an essay:

* To present the revolutionaryfeatures of the Transformer’s architecture that made it a disruptive technology leap in the NLP field.
* To discuss in detail how Transformer-based LLMs like OpenAI’s GPT, Anthropic’s Claude, or Alphabet/Google’s Gemini can have significant impacts on human life **with respect to** the student’s **selected domain expertise field**.

**Answer:**

The transformer did set the groundwork for building the high performing LLMS like ChatGPT, Claude, Gemini etc.

Powerful feature like “attention mechanism” make the transformers understand the context. It also enables it to access past information.

Another feature is “positional embedding” which stores the information of the position of each word in the sentence.

These features differentiate transformers for other networks and make them suitable for processing Large language Models.

The below figure shows the GPT transformer by OpenAI.

A diagram of a program

AI-generated content may be incorrect.

When OpenAI released ChatGPT in 2022, it took the world by storm. Anyone and everyone with access to smartphone or a computer wanted to try it out! ChatGPT and OpenAI became household names as the news was going into primetime shows. Since then, there has been no looking back for these power LLM applications. All the Big-Tech and non Big-Tech organizations want to do something with LLMs in order to gain the “first mover’s advantage”.

The organizations working in the domain of Logistics (Warehouse management) are also trying to incorporate LLMs in their systems reap the benefits and maximize profits.

In the field of Supply chain management there have been various implementations of LLMs and some of them got into headlines. Microsoft started a pilot project based on LLM from March 2023 to October 2023. This system proved to be increasing productivity and efficiency.

There was a significant decrease in incident response time and decision making became faster. Finally, the system was deployed in November 2023 with an expectation for refined outputs.

The system and hardware utilization reports which earlier used to go through multiple teams and divisions to get the data and signoff and took about a week to complete can be done in a matter of minutes with the LLM system.

There is a boost in productivity, however, there are challenges as well. We need to train the workforce to use the LLM system efficiently. For Example, the question/prompt “Can we utilize factory A better?” can be interpreted in multiple ways. We need to train the people to ask the questions in the format that is interpreted correctly by LLM.

More and more organizations are training their LLM on their Domain specific data. If any unsupported question is asked, the LLMs responds in a generic manner like “I do not understand the question. Here are a few suggestions.” The system needs to be trained to handle these kinds of questions effectively.

**Conclusion:**

The LLM based systems have brought higher efficiency and productivity. However, to maximize the benefits, we need to train the workforce to adapt to the LLM. On the other hand, LLM also needs to be continuously trained to overcome the limitations and errors.

**References**:

<https://sanchman21.medium.com/evolution-of-transformers-part-1-faac3f19d780#:~:text=Transformer%2C%20a%20revolutionary%20model%2C%20was,RNNs%20take%20in%20sequential%20data>.

<https://hbr.org/2025/01/how-generative-ai-improves-supply-chain-management>

<https://hbr.org/2024/09/embracing-gen-ai-at-work>

<https://hbr.org/2024/03/how-machine-learning-will-transform-supply-chain-management>

Course Material: llm\_transfomers\_introduction.pdf

Course Material: llm\_transfomers\_architecture\_PART\_I.pdf

Course Material: llm\_transfomers\_architecture\_PART\_II.pdf