1. What are transformers? How can you use them in summarisation?

Transformers are a type of neural network architecture that uses self-attention to process sequential data like text. This makes them highly effective at handling long sequences of text, which has traditionally been a challenge for other types of neural networks. Transformers have revolutionized the field of natural language processing (NLP) and can be used in various applications, such as text summarization.

Summarization is the process of distilling the most important information from a longer text and presenting it in a shorter format. Transformers can be used in a technique called "encoder-decoder architecture" for summarization. The input text is first encoded into a fixed-length vector representation using a transformer encoder. Then, a transformer decoder is used to generate the summary based on the encoded vector. The decoder is trained to produce a summary that accurately reflects the most important information in the original text.

Using transformers in summarization has several advantages. They can handle long sequences of text and learn to identify important information in a text, making them highly effective at producing summaries that capture the most important information in a given text.

2. What is Attention ? Why is it so important ?

Attention is a mechanism used in deep learning that allows neural networks to selectively focus on specific parts of input data during processing. It works by assigning weights to different parts of the input data, allowing the network to prioritize important features and ignore irrelevant ones. This can significantly improve the accuracy and efficiency of a neural network.

Attention has become an important tool in many deep learning applications, particularly in natural language processing (NLP) and computer vision. In NLP, attention can help a model understand the relationships between words in a sentence, improving its ability to accurately process and generate natural language. In computer vision, attention can help models identify important features in images, allowing for more accurate object recognition and image classification.

In addition to improving accuracy and efficiency, attention can also provide valuable insights into how a neural network is processing information. By visualizing the attention weights assigned to different parts of the input data, researchers can gain a better understanding of how the network is making its predictions and identify areas where it may need further refinement.

3. Can we use some other forms of machine learning for the problem ?

While transformers are currently the state-of-the-art method for text summarization, there are other machine learning approaches that can be used for this NLP problem. The choice of method depends on the specific requirements and constraints of the application.One such method is the use of convolutional neural networks (CNNs). CNNs are a type of neural network that are commonly used for image recognition tasks, but they can also be adapted for text summarization. In this approach, a CNN is trained to scan a given input text and identify the most important sentences or phrases.

Another method is the use of recurrent neural networks (RNNs). RNNs are designed to process sequential data, such as text, and can be used for text summarization by training the network to generate a summary based on the input text.

There are also unsupervised learning methods, such as clustering and topic modeling, which can be used for summarization. These methods do not require labeled data and instead identify patterns and relationships within the input text to generate a summary.