**Module 13**

**Name: Upadhyay Sachin Naresh**

**Batch id: AIWDEOB 300821**

**Topic: Sequence to sequence model, Transformers and BERT**

Compare BERT, GPT-2 and XLNET. Write down the differences between them.

**Answer:**

**BERT:**

It stands for Bidirectional Encoder Representations from Transformers, is a neural network-based technique for natural language processing pre-training. In plain English, it can be used to help Google better discern the context of words in search queries.

BERT’s key technical innovation is applying the bidirectional training of Transformer, a popular attention model, to language modelling. This is in contrast to previous efforts which looked at a text sequence either from left to right or combined left-to-right and right-to-left training. The paper’s results show that a language model which is bidirectionally trained can have a deeper sense of language context and flow than single-direction language models. In the paper, the researchers detail a novel technique named Masked LM (MLM) which allows bidirectional training in models in which it was previously impossible.

**Working of BERT: -**

BERT makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. In its vanilla form, Transformer includes two separate mechanisms — an encoder that reads the text input and a decoder that produces a prediction for the task. Since BERT’s goal is to generate a language model, only the encoder mechanism is necessary.

When training language models, there is a challenge of defining a prediction goal. Many models predict the next word in a sequence (e.g., “The child came home from \_\_\_”), a directional approach which inherently limits context learning. To overcome this challenge, BERT uses two training strategies:

**1. Masked LM (MLM): -**

Before feeding word sequences into BERT, 15% of the words in each sequence are replaced with a [MASK] token. The model then attempts to predict the original value of the masked words, based on the context provided by the other, non-masked, words in the sequence. In technical terms, the prediction of the output words requires:

(i) Adding a classification layer on top of the encoder output.

(ii) Multiplying the output vectors by the embedding matrix, transforming them into the vocabulary dimension.

(iii) Calculating the probability of each word in the vocabulary with SoftMax.

**2. Next Sentence Prediction (NSP): -**

In the BERT training process, the model receives pairs of sentences as input and learns to predict if the second sentence in the pair is the subsequent sentence in the original document. During training, 50% of the inputs are a pair in which the second sentence is the subsequent sentence in the original document, while in the other 50% a random sentence from the corpus is chosen as the second sentence. The assumption is that the random sentence will be disconnected from the first sentence.

To help the model distinguish between the two sentences in training, the input is processed in the following way before entering the model:

(i) A [CLS] token is inserted at the beginning of the first sentence and a [SEP] token is inserted at the end of each sentence.

(ii) A sentence embedding indicating Sentence A or Sentence B is added to each token. Sentence embeddings are similar in concept to token embeddings with a vocabulary of 2.

(iii) A positional embedding is added to each token to indicate its position in the sequence. The concept and implementation of positional embedding are presented in the Transformer paper.

**GPT2:**

Generative Pre-trained Transformer 2 (GPT-2) is an open-source artificial intelligence created by Open AI in February 2019. GPT-2 translates text, answers question, summarizes passages, and generates text output on a level that, while sometimes indistinguishable from that of humans, can become repetitive or nonsensical when generating long passages. It is a general-purpose learner; it was not specifically trained to do any of these tasks, and its ability to perform them is an extension of its general ability to accurately synthesize the next item in an arbitrary sequence. GPT-2 was created as a "direct scale-up" of Open Ai’s 2018 GPT model, with a ten-fold increase in both its parameter count and the size of its training dataset.

The GPT architecture implements a deep neural network, specifically a transformer model,[9] which uses attention in place of previous recurrence- and convolution-based architectures. Attention mechanisms allow the model to selectively focus on segments of input text it predicts to be the most relevant. This model allows for greatly increased parallelization, and outperforms previous benchmarks for RNN/CNN/LSTM-based models.

Open AI released the complete version of the GPT-2 language model (with 1.5 billion parameters) in November 2019. GPT-2 was to be followed by the 175-billion-parameter GPT-3, revealed to the public in 2020 (whose source code has never been made available). Access to GPT-3 is provided exclusively through an API offered by Microsoft

**XLNET:**

XLNet is an autoregressive Transformer that leverages the best of both autoregressive language modelling and autoencoding while attempting to avoid their limitations. Instead of using a fixed forward or backward factorization order as in conventional autoregressive models, XLNet maximizes the expected log likelihood of a sequence w.r.t. all possible permutations of the factorization order. Thanks to the permutation operation, the context for each position can consist of tokens from both left and right. In expectation, each position learns to utilize contextual information from all positions, i.e., capturing bidirectional context.

Additionally, inspired by the latest advancements in autogressive language modelling, XLNet integrates the segment recurrence mechanism and relative encoding scheme of Transformer-XL into pretraining, which empirically improves the performance especially for tasks involving a longer text sequence.

**Comparison of BERT, GPT-2 and XLNET**

