Center For Artificial Intelligence

Dr. B.R. Ambedkar National Institute of Technology, Jalandhar

Report on Expert Talk Lecture

Lecture Topic- Information Retrieval using Large Language Models (LLMs)

Date: 14/08/2024

Time: 5:00-6:00pm

Topic: Information Retrieval using Large Language Models (LLMs)

Speaker: Dr Kushal Shah, Professor Computer Science Sitare University

The expert lecture on "Information Retrieval using Large Language Models (LLMs)" was organized by Center for Artificial Intelligence. The talk was given by Dr Kushal Shah ,Professor of Computer Science (Sitare University).

Dr. Kushal Shah is a seasoned data science professional with extensive experience in solution design and the management of Large Language Models (LLMs). As the founder of *Building Self-Shiksha*, an initiative focused on AI/ML education, Dr. Shah is dedicated to advancing knowledge in these fields. He is currently engaged in several open-source projects related to Machine Learning and Natural Language Processing. Additionally, Dr. Shah is developing an innovative self-learning platform for Data Science and Machine Learning, which can be accessed at www.bekushal.com.

Information Retrieval using Large Language Models (LLMs)

Introduction to Information Retrieval-

In the digital age, information retrieval (IR) is essential for assisting users in locating pertinent information from vast datasets such as the internet or specialized databases. Simple keyword matching has given way to more complex machine learning and deep learning techniques in information retrieval (IR). Due to these developments, search results are now much more relevant and accurate, which makes IR systems essential for applications like recommendation engines, search engines, and natural language processing (NLP).

Word Embeddings Models-

The use of word embeddings, which depict words as vectors in a continuous vector space, is a fundamental advance in contemporary IR and NLP. This method makes it possible to record the semantic connections between words. A novel model called Word2Vec creates word vectors by forecasting context words. Another well-known technique, called GloVe (Global Vectors for Word Representation), concentrates on gathering global word co-occurrence statistics in order to provide more context-aware embeddings. By adding subword information, FastText improves on Word2Vec and strengthens it against uncommon terms and misspellings. More sophisticated models like BERT and GPT overcome this shortcoming, as these models only generate static word representations that do not take context-dependent meanings into account.

LSTM (Long Short-Term Memory) Networks-

LSTM networks marked a significant advancement in handling sequential data, overcoming the limitations of traditional Recurrent Neural Networks (RNNs), particularly the vanishing gradient problem. LSTMs maintain information over long sequences through memory cells, making them suitable for tasks like speech recognition and language modeling. Despite their strengths in learning long-term dependencies, LSTMs require substantial computational resources and are prone to overfitting, particularly with smaller datasets.

BERT (Bidirectional Encoder Representations from Transformers)-

By enabling bidirectional text processing, BERT transformed natural language processing (NLP) by enabling simultaneous extraction of context from a word's left and right sides. Because of this skill, BERT performs incredibly well in jobs requiring extensive contextual awareness, such sentiment analysis and question answering. The

transformer model, which uses self-attention mechanisms to rank words in order of priority, is the foundation of BERT's architecture. BERT is a strong tool, but it requires a lot of processing power and works best when optimized on high-quality information, which presents problems in contexts with limited resources.

GPT (Generative Pre-trained Transformer)-

Unlike BERT models, GPT models anticipate the subsequent word in a series by using a unidirectional context. GPT is especially useful for creative jobs like dialogue simulation and content creation because of this method. The most sophisticated version, GPT-3, shows incredible adaptability and can handle a variety of tasks with little adjustment. But there are ethical questions because GPT models can produce incorrect or biased information, reflecting the biases in their training data, and they need a lot of processing power.

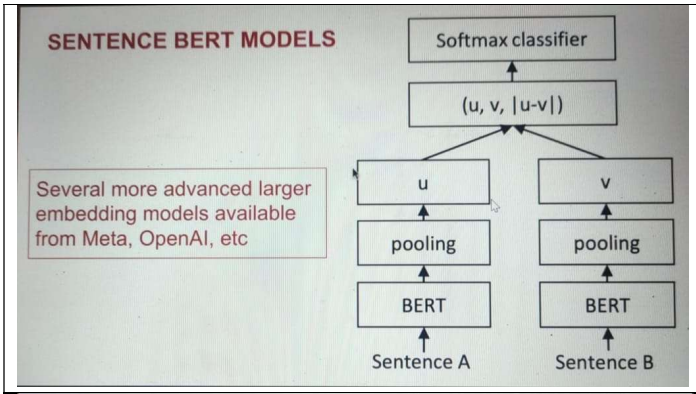
Advanced Models and Future Directions-

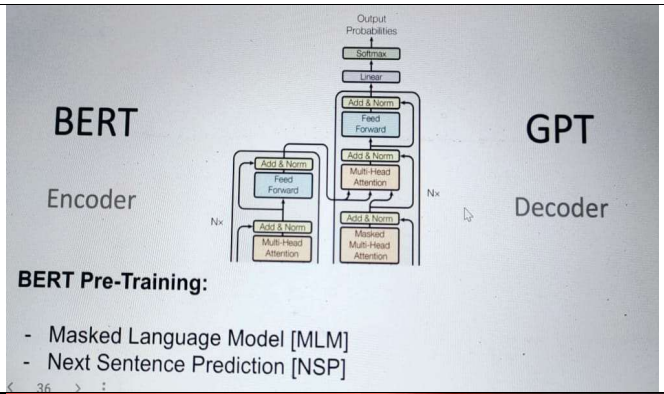
Different activities require different qualities from BERT and GPT. BERT is better at tasks requiring deep contextual awareness, whereas GPT is better at text production. The creation of ever more sophisticated models, like GPT-3, which can carry out difficult tasks with a high degree of accuracy, holds the key to the future of IR and NLP. Addressing the ethical ramifications of these models—bias and the requirement for openness, among other issues—is also becoming more and more important. With further development, these technologies should incorporate additional modalities—like pictures and videos—to produce AI systems that are more complete and all-encompassing.

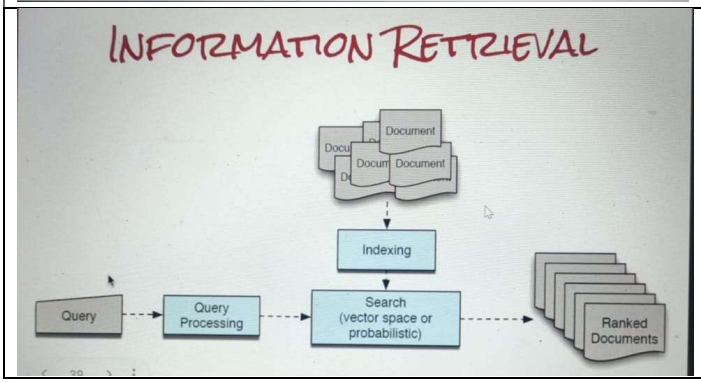
Conclusion-

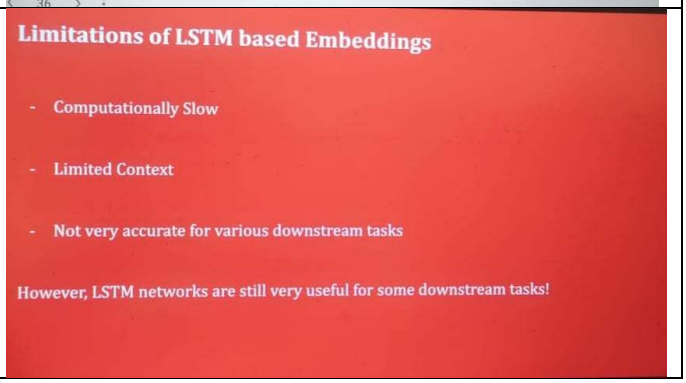
The advancement of word embeddings, transformer-based models like BERT and GPT, and LSTM networks has greatly influenced the development of information retrieval. The accuracy and relevance of IR systems have greatly increased as a result of these advancements. Ongoing model development and a focus on ethical AI are crucial if we are to maximize the potential of these models while reducing risks.

Below are some slides of the Expert Talk









Submitted By -Mohit Lohani

M.Tech-AI (24901312)