قسم علوم الحاسب
كلية الحاسبات والذكاء الاصطناعى
جامعة بنى سويف

مجلة قسم علوم الحاسب للعلوم المتقدمة

Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) is a natural language processing (NLP) paradigm that combines the strengths of both retrieval-based and generation-based models. In traditional NLP tasks, retrieval models retrieve relevant information from a large corpus, while generation models generate text based on given prompts or inputs. RAG integrates these approaches by enhancing generation models with retrieval capabilities.

In RAG, a large database of text documents or passages is pre-processed and indexed for efficient retrieval. During inference, the model can retrieve relevant passages from this database based on the input query or prompt. These retrieved passages are then used to augment the generation process, providing additional context or information for generating more coherent and relevant responses.

RAG models are particularly useful in tasks where both contextual understanding and creative generation are required, such as question answering, summarization, and dialogue generation. By leveraging both retrieval and generation capabilities, RAG models aim to produce more accurate and informative outputs compared to traditional generation models alone.

RAG comprises two main components — Retrieval and Generation:

Retrieval Models

Retrieval models act as the information gatekeepers in the RAG architecture. Their primary function is to search through a large corpus of data to find relevant pieces of information that can be used for text generation. Think of them as specialized librarians who know exactly which 'books' to pull off the 'shelves' when you ask a question. These models use algorithms to rank and select the most pertinent data, offering a way to introduce external knowledge into the text generation process.

Address: New Beni-Suef City. Beni-Suef. 62111 Web Site: WWW.fci.bsu.edu.eg
Email: fci@fci.bsu.edu.eg Telephone/Fax: 082 2246796

قسم علوم الحاسب
كلية الحاسبات والذكاء الاصطناعى
جامعة بنى سويف

مجلة قسم علوم الحاسب للعلوم المتقدمة

Generative Models

Once the retrieval model has sourced the appropriate information, generative models come into play. These models act as creative writers, synthesizing the retrieved information into coherent and contextually relevant text. Usually built upon Large Language Models (LLMs), generative models can create text that is grammatically correct, semantically meaningful, and aligned with the initial query or prompt.

RAG represents a promising direction in NLP research, offering a way to bridge the gap between retrieval-based and generation-based approaches and achieve more robust and effective language understanding and generation.

إعداد

د / إبراهيم الدسوقى

دكتور بقسم علوم الحاسب -كلية الحاسبات والذكاء الاصطناعى – جامعة بنى سويف