Carrera:

Licenciatura en Sistemas de Información

Tema:

Retrieval-Augmented Generation (RAG)

Asignatura:

Inteligencia Artificial

Alumnos:

Díaz, Bricia Candela

Guevara Gonzalez, Johan Sebastian

2025

Justificación de la elección del Tema:

La tecnología Retrieval-Augmented Generation (RAG) es una solución emergente muy innovadora. RAG aborda una limitación clave de los “large language models” (LLMs): su tendencia a "alucinar" (generar información incorrecta) y la dificultad para justificar sus respuestas. Esta tecnología RAG fusiona el conocimiento paramétrico (almacenado en los pesos del modelo) con el no-paramétrico (bases de datos externas). Esta integración mejora significativamente el Procesamiento del Lenguaje Natural (NLP), permitiendo aplicaciones de IA más fiables y verificables.

Relevancia Social: En un contexto global donde la desinformación es un desafío, la capacidad de RAG para proporcionar respuestas fundamentadas y atribuibles es invaluable. Incrementa la confianza en los sistemas de IA utilizados en sectores críticos como la educación, la atención médica y los medios de comunicación, ya que los usuarios pueden verificar directamente la fuente de la información.

Aspectos Teóricos que lo Sustentan:

La propuesta es una arquitectura que combina un modelo de lenguaje generativo pre-entrenado paramétrico (como BART) con una memoria no-paramétrica (un índice vectorial denso de Wikipedia). La innovación teórica radica en hacer el proceso de recuperación diferenciable de extremo a extremo. Esto significa que el sistema aprende no solo a generar texto, sino también a seleccionar activamente la información más relevante de su memoria externa para informar su generación, basándose en el concepto de recuperación de pasajes densos (DPR).

Métodos y Herramientas Seleccionadas:

RAG-Sequence: El modelo utiliza el mismo documento recuperado para condicionar la generación de toda la secuencia de salida.

RAG-Token: Permite que el modelo recupere un nuevo documento para cada token que genera.

Impacto de estas Tecnologías Emergentes de la IA:

La tecnología RAG ha sentado las bases para el diseño de arquitecturas de IA híbridas que combinan las fortalezas de los paradigmas generativos y basados en la recuperación. Subraya la importancia de fundamentar los modelos de IA en bases de conocimiento externas y actualizadas, empujando los límites de los LLMs al reducir las "alucinaciones" y aumentar la confianza.

Palabras Clave:

Retrieval-Augmented Generation (RAG)

Large Language Models (LLMs)

Natural Language Processing (NLP)

Bibliografía

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel*, D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks" in Proc. of the 34th Conf. Neural Information Processing Systems (NeurIPS 2020)**, Vancouver, Canada, 2020, pp. 9459-9474. [Online]. Disponible: Retrieval-augmented generation for knowledge-intensive NLP tasks