Elton Cardoso do Nascimento - 233840

IA024 - Redes Neurais Profundas para Processamento de Linguagem Natural - 1s2024

Leitura do Artigo "Retrieval-Augmented Generation for Large Language Models: A Survey" (Gao et al.)

O artigo realiza uma revisão de trabalhos relacionados a *RAG*, "Retrieval-Augmented Generation", que consiste em um conjunto de técnicas utilizadas para melhorar o funcionamento de LLMs, fornecendo informações relevantes sobre o prompt em questão, permitindo a redução de alucinação, traceabilidade de fontes e uso de conhecimento recente e especializado. O processo ocorre em três etapas principais: indexação de fontes, recuperação e geração. A indexação de fontes é onde são criadas representações das fontes a serem utilizadas, estas podendo estar desde um formato textual não-estruturado até um formato estruturado como um grafo, e com diferentes granularidades. A recuperação é onde fontes relevantes são procuradas, procurando similaridades semânticas entre o prompt e as fontes, estas podendo ser procuradas por similaridades de embedings, como similaridade de cosseno. Por fim, na geração, as fontes são combinadas junto com o prompt, não sendo considerado boa prática enviá-las de forma bruta para a LLM, passando então por processos como re-rankeamento, colocando informações mais relevantes antes; ou seleção e compressão de contexto, tantando reduzir a quantidade de informações forneciadas ao modelo através de sumarização, filtragem e deleção de informação irrelevante. Um ponto curioso, mas não citado pelo artigo, é a semelhança entre os processos de geração e processos cognitivos de atenção, colocando mais camadas de atenção sobre as arquiteturas já baseadas nessa habilidade.

As técnicas de RAG são divididas em 3 paradigmas, com melhorias ao custo de aumento de complexidade entre elas. O primeiro paradigma, "RAG simples", realiza o processo linear de indexing->retrieval->generation, sofrendo com recuperação de informações irrelavantes ou falta de informação e geração ainda suctível a alucinação. O paradigma de "RAG avançado" propõe etapas pré e pós-recuperação para reduzir estes problemas. Técnicas pré-recuperação incluem otimização da query, tentando melhorarar o prompt através de reescrita, expansão em múltiplos e uso de RAGs específicos para a query em questão. Já técnicas de pós-recuperação envolvem melhor preparação dos dados para envio a LLM, como re-ranqueamento e compressão de contexto. O terceiro paradigma, "RAG modular" propõe a criação de uma arquitetura modular, orquestrando estes módulos para melhor geração da resposta. Módulos adicionais podem incluir processos como memória, busca ativa em bases de dados e web, e roteamento entre diferentes módulos.

O processo em si do RAG também pode ser alterado, como execução iterativa, onde o processo de recuperação é executado mais de uma vez; execução recursiva, onde a recuperação é executada em camadas segundo a fonte e prompt; e adaptativo, onde a própria arquitetura é capaz de decidir quando o uso do RAG é necessário.

O trabalho também apresenta formas de avaliar RAGs e possíveis futuros desenvolvimentos, como RAGs multimodais e melhoria na robustez.