Documento de Briefing: Temas Chave em IA - Tokens, Janelas de Contexto e Geração Aumentada por Recuperação

Data: 25 de Maio de 2024
Assunto: Revisão dos principais conceitos que impulsionam a IA moderna, com foco em tokens, janelas de contexto e RAG, com base em informações da NVIDIA, Anthropic e IBM, juntamente com uma fonte acadêmica sobre RAG.
Fontes:

- Excertos de "Explaining Tokens — the Language and Currency of AI | NVIDIA Blog" (NVIDIA)
- Excertos de "Introducing the next generation of Claude | Anthropic" (Anthropic)
- Excertos de "What is a context window? | IBM" (IBM)
- Excertos de "[2005.11401] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks" (arXiv)

Resumo Executivo

Este briefing explora três conceitos interligados fundamentais para a operação e avanço dos modelos de Inteligência Artificial (IA), particularmente os Grandes Modelos de Linguagem (LLMs): tokens, janelas de contexto e a técnica de Geração Aumentada por Recuperação (RAG). Os tokens são as unidades básicas de dados que os modelos de IA processam, atuando como sua linguagem e moeda. A janela de contexto define a quantidade de tokens que um modelo pode "lembrar" e considerar de uma vez, impactando diretamente a capacidade de processamento de informações longas e a coerência da resposta. Finalmente, a Geração Aumentada por Recuperação é uma abordagem que combina memória paramétrica (o próprio modelo pré-treinado) com memória não paramétrica (uma base de dados externa pesquisável) para melhorar a factualidade e especificidade das respostas, mitigando algumas limitações dos modelos puramente paramétricos. As fontes da NVIDIA e IBM detalham o que são tokens e janelas de contexto e sua importância econômica e computacional, enquanto a Anthropic ilustra o uso prático de janelas de contexto maiores e o documento do arXiv descreve a técnica RAG e seus benefícios.

Tokens: A Linguagem e Moeda da IA

Os tokens são a base fundamental do processamento de dados para modelos de IA, tanto durante o treinamento quanto na inferência. Eles representam "unidades minúsculas de dados que vêm da quebra de pedaços maiores de informação" (NVIDIA). Essa tokenização de dados

— seja texto, imagens, áudio ou vídeo — permite que os modelos processem e aprendam as relações entre essas unidades.

- Função: Os modelos processam tokens para habilitar capacidades como predição, geração e raciocínio. "Quanto mais rápido os tokens puderem ser processados, mais rápido os modelos poderão aprender e responder" (NVIDIA).
- Tokenização: É o processo de traduzir dados em tokens. Uma tokenização eficiente "ajuda a reduzir a quantidade de poder computacional necessário para treinamento e inferência" (NVIDIA).
- Variabilidade do Token: Um token pode representar um único caractere, uma parte de uma palavra, uma palavra inteira ou até mesmo uma curta frase multi-palavras (IBM). A mesma palavra pode ser representada por tokens diferentes dependendo do contexto para capturar nuances de significado (NVIDIA).
- Importância Econômica: No contexto das "fábricas de IA" (centros de dados otimizados para IA), o processamento eficiente de tokens se traduz em "inteligência manufaturada", que é o "ativo mais valioso na nova revolução industrial impulsionada pela IA" (NVIDIA). O custo e a receita de serviços de IA estão cada vez mais sendo medidos pelo número de tokens consumidos e gerados (NVIDIA).

Janela de Contexto: A Memória de Trabalho dos Modelos de IA

A janela de contexto de um LLM refere-se à quantidade de texto (medida em tokens) que o modelo pode considerar ou "lembrar" em um dado momento (IBM). É equivalente à "memória de trabalho" de um modelo e impacta crucialmente seu desempenho e capacidades.

- Função: Uma janela de contexto maior permite que um modelo "processe entradas mais longas e incorpore uma quantidade maior de informação em cada saída" (IBM). Isso se traduz em "maior precisão, menos alucinações, respostas de modelo mais coerentes, conversas mais longas e uma capacidade aprimorada de analisar sequências de dados mais longas" (IBM).
- Limitações: Quando um prompt, conversa, documento ou base de código excede a janela de contexto do modelo, a informação deve ser "truncada ou resumida para o modelo prosseguir" (IBM).
- Tradeoffs Computacionais: Aumentar o tamanho da janela de contexto "geralmente acarreta maiores requisitos de poder computacional — e, portanto, maiores custos — e um potencial aumento na vulnerabilidade a ataques adversários" (IBM). O custo computacional escala quadraticamente com o comprimento da sequência de tokens (IBM).
- Avanços Recentes: Modelos recentes, como a família Claude 3 da Anthropic e os modelos Gemini 1.5 do Google, demonstram um aumento significativo no tamanho da janela de contexto, com a Anthropic oferecendo 200K tokens como padrão e a capacidade de aceitar entradas excedendo 1 milhão de tokens para clientes selecionados, e o Google Gemini 1.5 Pro oferecendo até 2 milhões de tokens (Anthropic, IBM).
- Desafios da Janela de Contexto Longa: Apesar dos benefícios, modelos podem ter dificuldade em utilizar informações no meio de longos contextos ("Lost in the Middle") e o aumento da janela de contexto pode apresentar uma superfície de ataque maior para "jailbreaking" (IBM).

Geração Aumentada por Recuperação (RAG): Combinando Memória Paramétrica e Não Paramétrica

A Geração Aumentada por Recuperação (RAG) é uma técnica descrita no artigo do arXiv que aborda as limitações dos modelos de linguagem pré-treinados puramente paramétricos em tarefas que exigem acesso preciso a conhecimento específico.

- Problema Abordado: Modelos pré-treinados armazenam conhecimento factual em seus parâmetros, mas têm "habilidade limitada para acessar e manipular conhecimento de forma precisa" (arXiv). Isso leva a desempenho inferior em tarefas intensivas em conhecimento e dificuldades em fornecer proveniência para suas decisões e atualizar seu conhecimento (arXiv).
- Abordagem RAG: RAG combina a "memória paramétrica é um modelo seq2seq pré-treinado" com "memória não paramétrica é um índice vetorial denso da Wikipedia, acessado com um recuperador neural pré-treinado" (arXiv).
- Benefícios: Modelos RAG podem superar as limitações dos modelos puramente paramétricos em tarefas intensivas em conhecimento. No artigo do arXiv, os modelos RAG alcançaram resultados de ponta em três tarefas de Perguntas e Respostas de domínio aberto e geraram "linguagem mais específica, diversa e factual do que uma linha de base seq2seq puramente paramétrica de ponta" (arXiv).
- Mecanismo: O RAG permite que o modelo acesse um banco de dados externo (memória não paramétrica) para recuperar informações relevantes que são então usadas para condicionar a geração da resposta. O documento da IBM também menciona a recuperação de informações adicionais de fontes de dados externas para Geração Aumentada por Recuperação (RAG), armazenada na janela de contexto durante a inferência.

Interconexões e Implicações

Tokens, janelas de contexto e RAG estão profundamente interligados:

- A eficiência da tokenização é crucial para maximizar a quantidade de informação que cabe dentro de uma janela de contexto (IBM, NVIDIA).
- Uma janela de contexto maior permite que modelos processem mais tokens de uma vez, o que é essencial para tarefas que exigem compreensão de longas sequências, como resumir um romance ou uma hora de podcast (NVIDIA, IBM).
- A técnica RAG se beneficia de janelas de contexto maiores, pois a informação recuperada de uma base de dados externa precisa ser incluída na janela de contexto do modelo para influenciar a geração da resposta (IBM).

As implicações desses conceitos são significativas para o desenvolvimento e aplicação da IA:

- Economia da IA: O processamento de tokens é um fator de custo e receita, com modelos de preços sendo baseados no uso de tokens (NVIDIA). A otimização do uso de tokens é vital para maximizar o valor das aplicações de IA (NVIDIA).
- Desempenho do Modelo: O tamanho e a gestão da janela de contexto afetam diretamente a precisão, coerência e capacidade de raciocínio dos modelos (IBM, NVIDIA). Técnicas como RAG aprimoram o desempenho em tarefas intensivas em conhecimento (arXiv).

Experiência do Usuário: Métricas como "time to first token" e "inter-token latency" (velocidade de geração de tokens) determinam a fluidez e a responsividade das aplicações de IA (NVIDIA).

Conclusão

A compreensão de tokens, janelas de contexto e Geração Aumentada por Recuperação é fundamental para entender o funcionamento e o potencial dos modelos de IA atuais. Tokens formam a linguagem subjacente, a janela de contexto determina a capacidade de "memória" e processamento de informações longas, enquanto RAG melhora a factualidade e especificidade ao integrar conhecimento externo. O avanço contínuo nessas áreas, conforme evidenciado pelas capacidades dos modelos Claude 3 e Gemini 1.5 e a adoção de RAG, é crucial para desbloquear novas aplicações e impulsionar a "revolução industrial impulsionada pela IA" (NVIDIA). A otimização desses elementos não apenas melhora o desempenho e a precisão, mas também impacta diretamente a economia da IA e a experiência do usuário final.