DeepSeek desafia gigantes da IA: 50% de custos e cortes de API - Against

Data: 2025-10-06 04:55:28

Autor: Inteligência Against Invaders

Redazione RHC:**6 outubro 2025 06:54**

A empresa chinesa DeepSeek apresentou uma versão experimental de seu modelo de linguagem, **DeepSeek-V3.2-Exp,** que pela primeira vez **implementa sua própria versão de atenção esparsa,** uma técnica que *reduz significativamente o custo computacional do processamento de sequências de texto longas* . O novo mecanismo, chamado **DeepSeek Atenção Esparsa,** é dito ser capaz de **reduzir os custos operacionais do modelo quase pela metade** . Para demonstrar essas economias, **a empresa reduziu o preço de sua API em 50%.**

O problema da sobrecarga computacional em grandes modelos de linguagem é particularmente agudo para diálogos longos. A arquitetura clássica do Transformer, desenvolvida em 2017, compara cada palavra na sequência de entrada com todas as outras palavras, resultando em um aumento quadrático no número de operações. *Para mil palavras, isso se traduz em um milhão de comparações e, para dez mil palavras, em cem milhões.* Essa sobrecarga **aumenta o uso de recursos em sessões longas e diminui o desempenho,** pois o sistema é forçado a reanalisar todo o histórico de diálogo para cada nova solicitação.

A tecnologia de atenção esparsa funciona de maneira diferente. **Ele não combina todas as palavras com todas as outras, mas seleciona um conjunto limitado das conexões mais significativas.** O DeepSeek usa um mecanismo proprietário chamado **o Lightning Indexer,** um pequeno *Unidade de rede neural adicional* que avalia a significância dos pares de palavras e seleciona até *2.048 das conexões mais relevantes para cada posição* . A empresa não divulgou detalhes sobre como o indexador toma suas decisões, mas diz *não compromete a qualidade da compreensão do texto.*

Testes internos mostraram que **o novo modelo fornece resultados comparáveis à versão anterior, DeepSeek-V3.1-Terminus** , mantendo alta precisão e capacidade de processar sequências longas. Notavelmente, o DeepSeek **abriu o código de seus componentes sob a licença do MIT** e forneceu pesos acessíveis ao público, *permitindo que outros pesquisadores testem e desenvolvam as soluções propostas.*

O DeepSeek ganhou as manchetes pela primeira vez em janeiro, quando seu modelo R1 correspondeu ao desempenho o1 da OpenAI **com um custo de treinamento de apenas US$ 6 milhões** . Além disso, o aplicativo de bate-papo da empresa liderou brevemente **a loja de aplicativos do iPhone, superando o ChatGPT.** Desde então, a atenção da indústria se concentrou no laboratório chinês, *que foi forçado a encontrar maneiras de otimizar seus cálculos devido ao acesso limitado a GPUs modernas e outros chips especializados devido a restrições de exportação.*

Embora essa abordagem tenha recebido pouca atenção por muito tempo e tenha sido usada pela primeira vez no GPT-3 e em vários outros modelos por desenvolvedores ocidentais, o DeepSeek afirma que **Sua implementação permitiu um ajuste preciso e uma redução significativa nos custos computacionais sem qualquer perda perceptível de qualidade.** Especialistas independentes ainda não confirmaram esses resultados, mas se as conclusões da empresa estiverem corretas, esses métodos podem mudar significativamente a economia do uso de modelos de IA a longo prazo.

Redação
A equipe editorial da Red Hot Cyber é composta por um grupo de indivíduos e fontes anônimas que colaboram ativamente para fornecer informações e notícias antecipadas sobre segurança cibernética e computação em geral.

Lista degli articoli