Zdnet

O mercado de inteligência artificial-e todo o mercado de ações-foi abalado na segunda-feira pela repentina popularidade da Deepseek, o modelo de linguagem grande de código aberto desenvolvido por um fundo de hedge com sede na China que superou o melhor do OpenAI em algumas tarefas enquanto custa longe menos.

Também: Eu coloquei as habilidades de codificação do Deepseek AI à prova – aqui é onde ela se desfez

Como Radhika Rajkumar, da Zdnet, detalhou na segunda -feira, o sucesso do R1 destaca uma mudança no mar na IA que poderia capacitar laboratórios e pesquisadores menores para criar modelos competitivos e diversificar o campo das opções disponíveis.

O que você vai ler:

memória virtual

Por que o DeepSeek funciona tão bem?

Acontece que é uma abordagem ampla dentro de formas de aprendizado profundo de inteligência artificial para espremer mais chips de computador, explorando um fenômeno conhecido como “escassez”.

A escassez vem de várias formas. Às vezes, envolve a eliminação de partes dos dados que a IA usa quando esses dados não afetam materialmente a saída do modelo de IA.

Além disso: por que a China Deepseek poderia estourar nossa bolha de IA

Em outros momentos, pode envolver cortar partes inteiras de uma rede neural se isso não afetar o resultado final.

Deepseek é um exemplo deste último: uso parcimonioso de redes neurais.

O principal avanço que a maioria identificou na Deepseek é que ele pode ligar e desligar grandes seções da rede neural “pesos” ou “parâmetros”. Os parâmetros são o que moldam como uma rede neural pode transformar a entrada – o prompt que você digita – em texto ou imagens geradas. Os parâmetros têm um impacto direto em quanto tempo leva para executar cálculos. Mais parâmetros, mais esforço de computação, normalmente.

Sparsidade e seu papel na IA

A capacidade de usar apenas alguns dos parâmetros totais de um modelo de idioma grande e desligar o restante é um exemplo de esparsidade. Essa escassez pode ter um grande impacto no tamanho ou pequeno, o orçamento de computação é para um modelo de IA.

Os pesquisadores de IA da Apple, em um relatório na semana passada, explicam bem o quão profunda e abordagens semelhantes usam a escassez para obter melhores resultados para uma determinada quantidade de poder de computação.

A Apple não tem conexão com a Deepseek, mas a Apple faz sua própria pesquisa de IA regularmente e, portanto, os desenvolvimentos de empresas externas como a Deepseek fazem parte do envolvimento contínuo da Apple no campo de pesquisa da IA, em geral.

No artigo, intitulado “Parâmetros vs flops: escalando leis de escasso ideal para modelos de idiomas de mistura de especialistas”, publicado no servidor de pré-impressão Arxiv, o principal autor Samir Abnar, da Apple e outros pesquisadores da Apple, juntamente com o colaborador Harshay Shah do MIT, estudou como o desempenho variava ao explorar a escassez desligando partes da rede neural.

Também: DeepSeek’s new open-source AI model can outperform o1 for a fraction of

the cost

A ABNAR e a equipe conduziram seus estudos usando uma biblioteca de códigos lançada em 2023 por pesquisadores de IA na Microsoft, Google e Stanford, chamada Megablocks. No entanto, eles deixam claro que seu trabalho é aplicável à Deepseek e outras inovações recentes.

O ABNAR e a equipe perguntam se há um nível “ideal” para a escassez em modelos profundos e similares, significando, para uma determinada quantidade de poder de computação, existe um número ideal desses pesos neurais para ligar ou desligar?

Acontece que você pode quantificar completamente a escassez como a porcentagem de todos os pesos neurais que você pode desligar, com essa porcentagem se aproximando, mas nunca igual a 100% da rede neural “inativa”.

Os gráficos mostram que, para uma determinada rede neural, em uma determinada quantidade de orçamento de computação, há uma quantidade ideal da rede neural que pode ser desligada para atingir um nível de precisão. É a mesma regra econômica que tem sido verdadeira para todas as novas gerações de computadores pessoais: um resultado melhor para o mesmo dinheiro ou o mesmo resultado por menos dinheiro.

Maçã

E acontece que, para uma rede neural de um determinado tamanho em parâmetros totais, com uma determinada quantidade de computação, você precisa de cada vez menos parâmetros para obter a mesma ou melhor precisão em um determinado teste de referência de IA, como matemática ou resposta a perguntas .

Em outras palavras, seja qual for o seu poder de computação, você pode desligar cada vez mais partes da rede neural e obter os mesmos ou melhores resultados.

Otimizando a IA com menos parâmetros

Como o Abnar e a equipe o colocam em termos técnicos, “aumentando a escassez e a expansão proporcionalmente o número total de parâmetros consistentemente leva a uma perda de pré -treinamento mais baixa, mesmo quando restringida por um orçamento de computação de treinamento fixo”. O termo “perda de pré -treinamento” é o termo da IA para a precisão de uma rede neural. Menor perda de treinamento significa resultados mais precisos.

Essa descoberta explica como o DeepSeek poderia ter menos poder de computação, mas atingir o mesmo ou melhor resultado simplesmente desligando cada vez mais partes da rede.

Além disso: a melhor IA para codificar em 2025 (e o que não usar)

A escassez é um tipo de mostrador mágico que encontra a melhor correspondência do modelo de IA que você tem e a computação que você tem disponível.

É a mesma regra econômica que tem sido verdadeira para todas as novas gerações de computadores pessoais: um resultado melhor para o mesmo dinheiro ou o mesmo resultado por menos dinheiro.

Existem outros detalhes a serem considerados sobre o DeepSeek. Por exemplo, outra inovação da Deepseek, como bem explicada por Ege Erdil, da Epoch Ai, é um truque matemático chamado “atenção latente de várias cabeças”. Sem se aprofundar muito nas ervas daninhas, a atenção latente de várias cabeças é usada para comprimir um dos maiores consumidores de memória e largura de banda, o cache de memória que contém o texto mais recentemente de entrada de um prompt.

O futuro da pesquisa de escassez

Detalhes à parte, o ponto mais profundo sobre tudo isso é que a escassez como fenômeno não é nova na pesquisa de IA, nem é uma nova abordagem na engenharia.

Os pesquisadores de IA exibem há muitos anos que eliminarem partes de uma rede neural pode alcançar uma precisão comparável ou ainda melhor com menos esforço.

O concorrente da NVIDIA Intel, há anos, identifica a escassez como uma avenida importante de pesquisa para mudar o estado da arte no campo. As abordagens de startups baseadas na esparsidade também obtiveram pontuações altas nos benchmarks do setor nos últimos anos.

O mostrador mágico da esparsidade não apenas reduz os custos de computação, como no caso da DeepSeek – também funciona na outra direção: também pode tornar os computadores de IA maiores e maiores mais eficientes.

Maçã

O mostrador mágico da esparsidade é profundo, pois não apenas melhora a economia para um orçamento pequeno, como no caso da Deepseek, mas também funciona na outra direção: gaste mais e você obterá benefícios ainda melhores por escassez. À medida que você aumenta seu poder de computação, a precisão do modelo de IA melhora, a ABNAR e a equipe descobriram.

Como eles disseram: “À medida que a escassez aumenta, a perda de validação diminui para todos os orçamentos de computação, com orçamentos maiores atingindo perdas mais baixas em cada nível de escassez”.

Em teoria, então, você pode criar modelos maiores e maiores, em computadores maiores e maiores, e obter melhor retorno ao seu dinheiro.

Todo esse trabalho de escassez significa que o Deepseek é apenas um exemplo de uma ampla área de pesquisa que muitos laboratórios já estão seguindo, e que muitos outros agora saltarão para replicar o sucesso de Deepseek.