Processamento de Linguagem Natural com Transformers

O Processo de Ajuste Fino (Fine-Tuning) de LLMs

No mundo em constante evolução da IA e do Processamento de Linguagem Natural (PLN), os Grandes Modelos de Linguagem (LLMs) e a IA Generativa tornaram-se ferramentas poderosas para diversas aplicações.

Alcançar os resultados desejados com esses modelos envolve diferentes abordagens que podem ser amplamente classificadas em três categorias: Engenharia de Prompts, Ajuste Fino e Criação de Um Novo Modelo. À medida que avançamos de um nível para outro, as exigências em termos de recursos e custos aumentam significativamente.

Nível Básico – Engenharia de Prompts nos Modelos Pré-Treinados

No nível básico, alcançar os resultados esperados dos grandes modelos de linguagem envolve uma engenharia cuidadosa de prompts, entradas para um LLM já treinado. Este processo envolve a elaboração de prompts e entradas adequadas para extrair as respostas desejadas do modelo. Prompt Engineering é uma técnica essencial para vários casos de uso, especialmente quando respostas gerais são suficientes. É a forma mais básica e simples de usar LLMs. O problema é que os LLMs são genéricos e podem não conseguir gerar texto para questões de uma área específica.

Nível Mais Alto – Criar Um Novo Modelo

No nível mais alto, a criação de um novo modelo envolve treinar um modelo do zero, especificamente adaptado para uma tarefa ou domínio específico. Essa abordagem fornece o mais alto nível de personalização, mas exige poder computacional substancial, quantidade considerável de dados e tempo (muito tempo) de treinamento do modelo.

Nível Intermediário – Ajuste Fino (a opção mais usada atualmente)

O ajuste fino nos permite aproveitar modelos pré-treinados existentes e adaptá-los a tarefas ou domínios específicos. Ao treinar o modelo em dados específicos de uma área ou assunto, podemos adaptá-lo para um bom desempenho em tarefas específicas.

No entanto, este processo pode consumir muitos recursos e ser dispendioso, pois modificaremos todos os milhões de parâmetros, como parte do treinamento. O ajuste fino do modelo ainda requer muitos dados de treinamento, enorme infraestrutura e esforço.

E no processo de ajuste fino completo de LLMs, existe ainda o risco de esquecimento catastrófico, onde se perde o conhecimento previamente adquirido no pré-treinamento. Ou seja, podemos ajustar um modelo por 8 horas para ao final descobrir que de fato ele “esqueceu” seu histórico de treinamento.

E a aplicação de ajuste fino completo a um único modelo para diferentes tarefas específicas de domínio geralmente resulta na criação de grandes modelos adaptados a tarefas específicas, sem modularidade.

Logo, o que necessitamos é de uma abordagem modular que evite a alteração de todos os parâmetros do modelo, ao mesmo tempo que exige menos recursos de infraestrutura e menos dados.

O Parameter Efficient Fine Tuning (PEFT) fornece uma maneira de realizar exatamente isso, o ajuste fino somente de parte dos parâmetros de um LLM, usando um volume de dados menor e também menor capacidade computacional. E, consequentemente, em menos tempo!

Vamos definir o PEFT na aula a seguir.