Comentário crítico: *LORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS*

Aluno: Maria Fernanda Bosco
RA: 183544

O artigo introduz o LoRA, uma técnica de *fine-tuning* em que os parâmetros originais do modelo pré-treinado são congelados, e matrizes de decomposição de baixo rank são adicionadas em paralelo às camadas lineares já existentes. Durante o ajuste, apenas essas matrizes adicionais são treinadas, reduzindo drasticamente o número de parâmetros atualizados.

Na época da publicação, modelos de linguagem de larga escala, como o GPT-3 com bilhões de parâmetros, já haviam sido lançados, e um desafio central era o *fine-tuning* para novas tarefas. Ajustar todos os parâmetros do modelo completo era extremamente custoso em GPU e memória, além de pouco prático quando se desejava criar múltiplas variantes adaptadas a diferentes aplicações. Os autores partem da hipótese de que, durante o *fine-tuning*, o modelo não precisa atualizar toda a matriz de pesos, pois as mudanças relevantes residem em um subespaço de baixa dimensionalidade e podem ser representadas por poucas direções principais.

Outro ponto levantado pelos autores é a questão da *inference latency*, isto é, o tempo adicional necessário para que o modelo produza previsões após ser adaptado. Na época, soluções existentes como *adapters* ou *prefix-tuning* demonstravam bom desempenho, mas introduziam um custo na inferência. Portanto, o LoRA diferenciou-se também por manter praticamente inalterado o tempo de inferência, já que não insere novas camadas nem prolonga as sequências, oferecendo assim uma alternativa mais eficiente.

Os resultados mostram que o LoRA alcança desempenho muito próximo ao *fine-tuning* completo, mesmo treinando menos de 1% dos parâmetros do modelo, e supera alternativas como *adapters* e *prefix-tuning* em termos de custo e latência.

Logo, o artigo do LoRA se destaca por propor uma abordagem de *fine-tuning* prática e de grande impacto, combinando economia de recursos com resultados robustos em múltiplas tarefas. Seu sucesso influenciou fortemente pesquisas posteriores e consolidou o LoRA como uma das técnicas mais relevantes para a adaptação de modelos de linguagem de larga escala.