Elton Cardoso do Nascimento - 233840

IA024 - Redes Neurais Profundas para Processamento de Linguagem Natural - 1s2024

Leitura do Artigo "LoRA: Low-Rank Adaptation of Large Language Models" (Hu et al)

O artigo apresenta a nova técnica LoRA, criada para adaptar um modelo pré-treinado para uma task downstream específica. A técnica se baseia em adicionar duas matrizes $A \in \mathbb{R}^{d \times r}$ e $B \in \mathbb{R}^{r \times k}$ para matrizes $W \in \mathbb{R}^{d \times k}$ existente no modelo original, de tal forma que o produto das duas matrizes novas e da matriz original possam ser somados na saída $h = BAx + Wx$. O treinamento é realizado com os pesos originais congelados.

O ponto chave do trabalho é que o rank r das matrizes adicionadas pode ser bem menor que o da matriz original ($r \ll \min(d, k)$), de tal forma que a quantidade de pesos a serem treinados é bem menor que a quantidade original, possuindo exemplos de testes com até 0,01% dos pesos originais. Isso permite um treinamento mais rápido e com menor uso de memória, enquanto que a performance se mostra igual ou superior a outras técnicas, incluindo o fine-tuning de todos os parâmetros.

Outra vantagem é que uma vez treinados as matrizes podem ser multiplicadas (BA) e adicionadas ao modelo original, reduzindo a latência de inferência, diferente de outras técnicas (adaptadores) que realizam a adaptação adicionando elementos à rede. É possível também em tempo de execução ficar trocando as matrizes a serem utilizadas, permitindo a adaptação para diferentes tarefas com baixo custo de memória comparada a trocar todos os pesos (da ordem de TB para GB).

A técnica é aplicada em modelos de NLP baseados em Transformers, concentrando-se apenas nos pesos das camadas de atenção. É demonstrado que possuir matrizes com rank menor, mas para várias matrizes do processo de atenção, é mais eficaz que adaptar apenas uma matriz com rank maior.

O trabalho também apresenta conclusões mais gerais sobre o processo. Primeiramente que as atualizações $\Delta W = BA$ costumam ocorrer com intensidade em poucas direções, e nas mesmas direções com várias inicializações aleatórias, apoiando a conclusão de que ranks r menores já são eficazes para melhorar o modelo original (em alguns casos) e mostrando uma robustez no processo. Também demonstra que as adaptações amplificam features importantes para a tarefa downstream que foram aprendidas, mas não enfatizadas, durante o processo de treinamento.

Concluindo, o trabalho apresenta uma técnica eficiente e eficaz para realizar a adaptação de modelos para tarefas específicas, ao mesmo tempo que apresenta conclusões interessantes para o entendimento geral dos processos de adaptação de modelos.