Principais contribuições do artigo “LoRA: Low-rank adaptation of large language models”, de Hu et al

Aluno: Leandro Carísio Fernandes

- A ideia do artigo é muito simples:
  - Grandes modelos de linguagem normalmente passam por duas fases. A primeira fase cria um modelo geral, é o pré-treino. Na segunda fase é feita a adaptação do modelo (fine-tuning) para atividades específicas.
  - O treinamento para as atividades específicas exige o ajuste de todos os parâmetros do modelo.
  - Normalmente esses parâmetros estão espalhados em diversas matrizes W de tamanho (i x o), com d muito grande.
  - A proposta é, durante o fine-tuning, travar as matrizes (i x o) e treinar novas matrizes A e B de tamanho (i x r) e (r x o), com r << min(i, o).
- Os ganhos são:
  - Redução no número de parâmetros que devem ser treinados, o que possibilita que o fine-tuning de modelos muito grandes sejam feitos com hardware mais modestos.
  - Ao final do fine-tuning, o conteúdo da matriz original é atualizado com o conteúdo das matrizes A e B. Com isso, não há aumento de delay durante a inferência.
  - Em casos de redes muito grandes, pode ocorrer até mesmo aumento de eficiência de treino. O artigo aplicou o método no GPT-3 175B e verificou redução de memória do checkpoint de 10.000x (de 350GB para 35MB) e melhoria de 25% na velocidade de treinamento.
    - Ao inserir as matrizes A e B no fine-tuning, espera-se um aumento do tempo de treinamento. Entretanto, dependendo da complexidade da rede, isso pode ser compensada pela redução no cálculo dos gradientes (são treinados bem menos parâmetros).