Projeto de Formatura – 2025

PCS - Departamento de Engenharia de Computação e Sistemas Digitais

Engenharia de Computação

Tema:

Seleção de Camadas para o Ajuste Fino Eficiente de LLMs

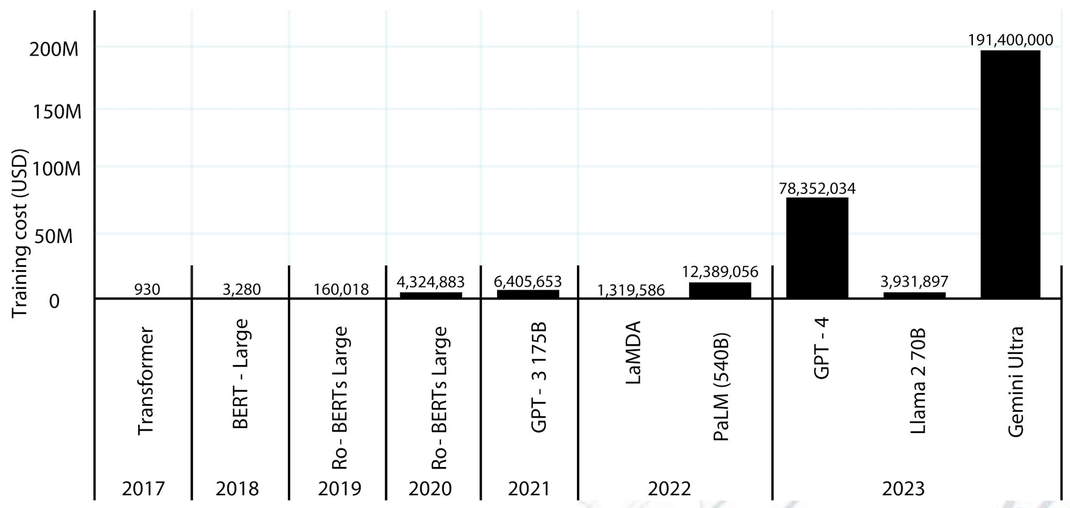
Introdução

Grandes modelos de linguagem pré-treinados apresentam desempenho não visto antes em processamento de linguagem natural e moldam uma nova onda de progresso em tarefas de raciocínio complexo. Entretanto, esta efetividade implica um alto custo computacional nas fases de treino e ajuste fino.

Para mitigar esse problema, métodos de ajuste fino eficiente, como o LoRA, propõem a adaptação de uma porção menor de parâmetros de toda matriz de pesos:

$$W = W_0 + \Delta W = W_0 + BA.$$

Em nosso trabalho, exploramos a pergunta: Os módulos LoRA são necessários em todas as camadas?

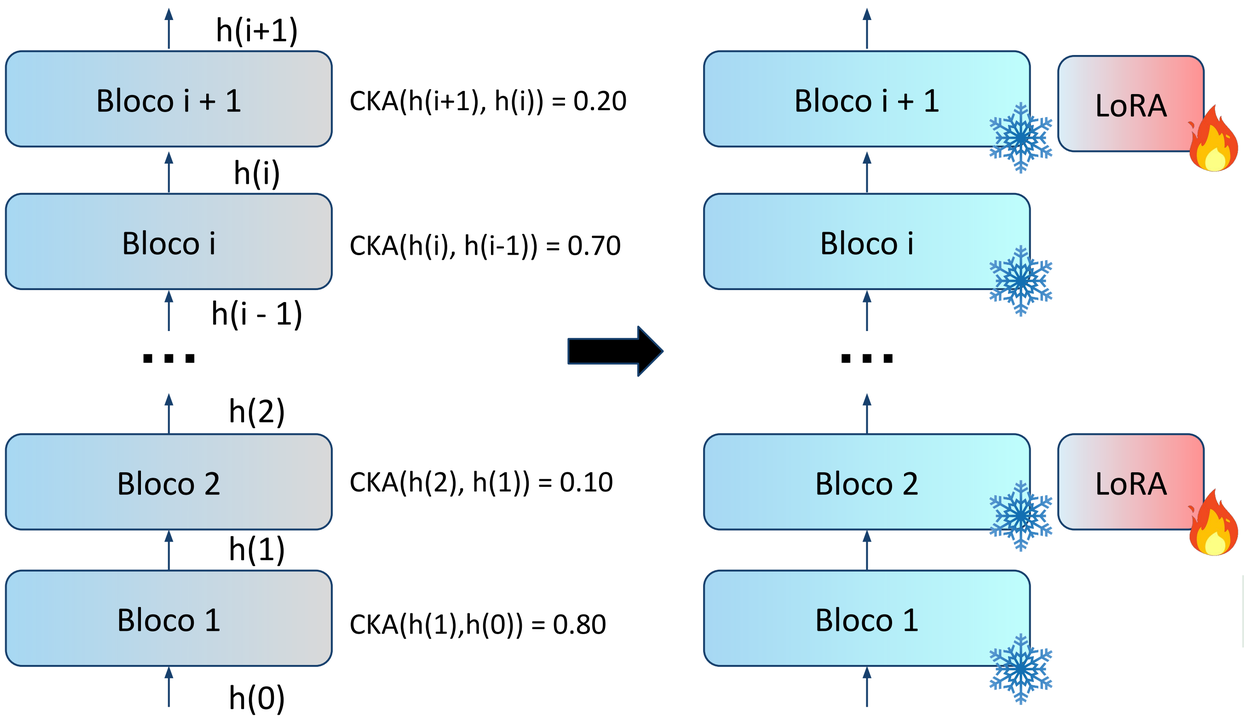


Método Proposto

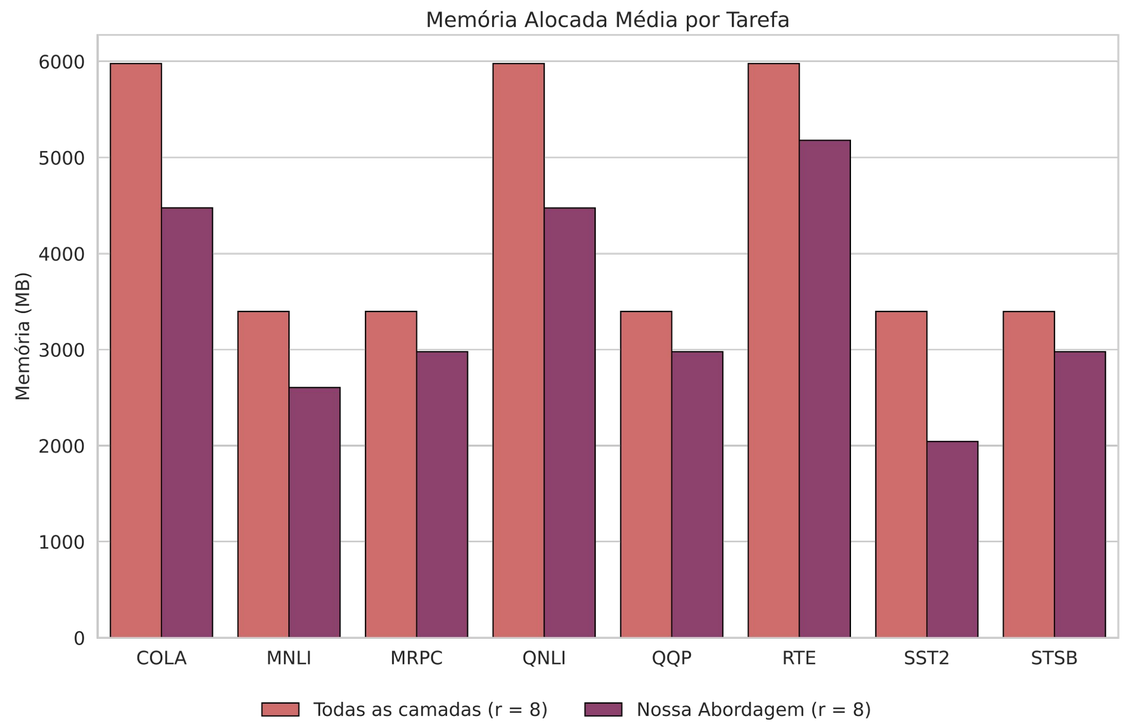
Com base em evidências experimentais e da literatura, construímos nosso método partindo da hipótese de que não só existem camadas mais importantes para o ajuste fino de um modelo de linguagem, mas também que elas são específicas à tarefa.

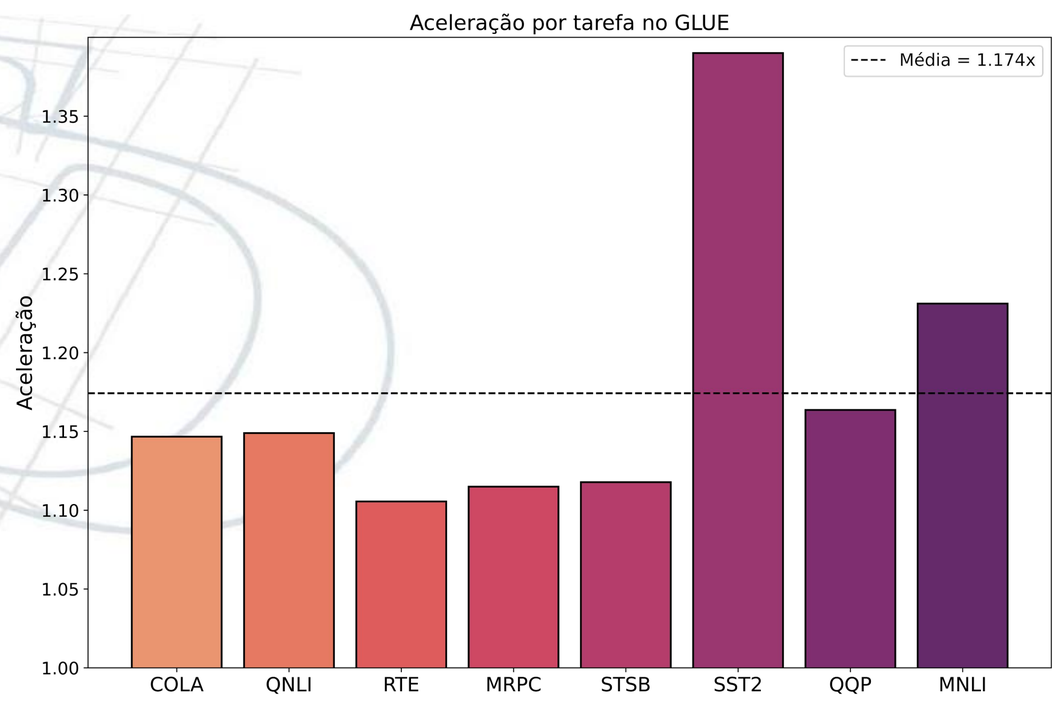
Nesse sentido, associamos a importância de uma camada com o seu impacto nas representações internas do modelo por meio da métrica CKA de similaridade.

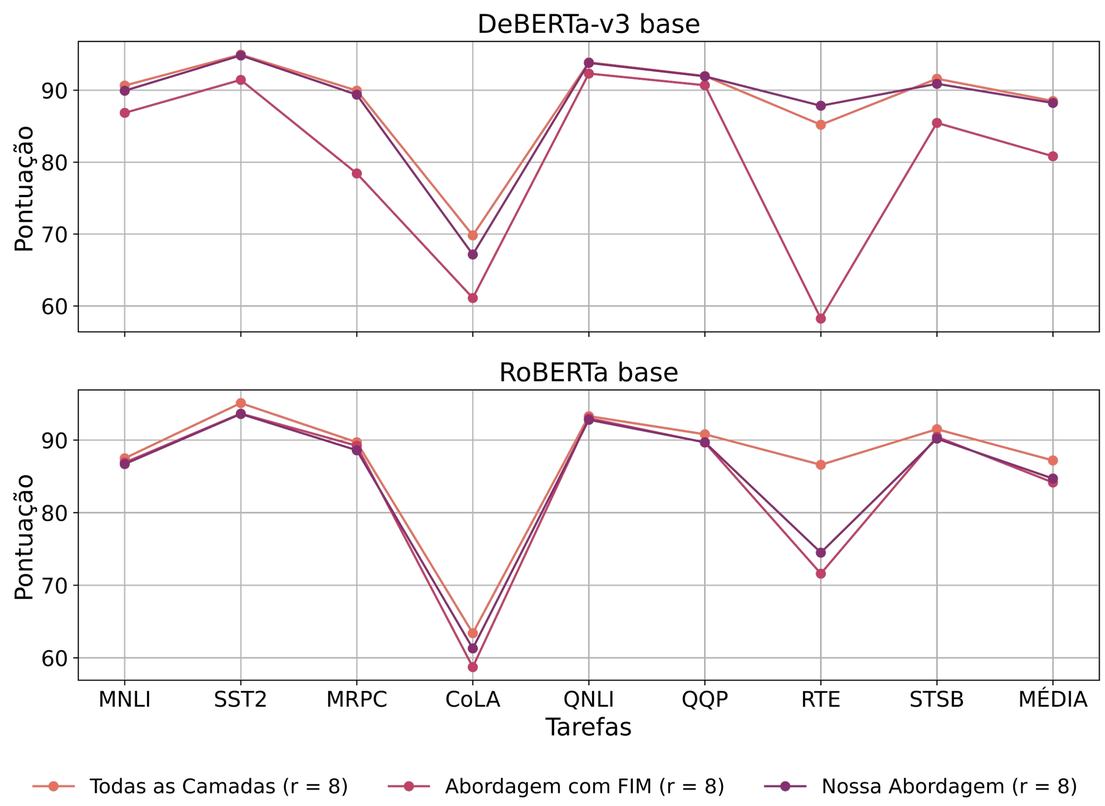
Assim, com uma medida de importância para cada camada, adaptamos apenas as N mais importantes, mantendo as demais congeladas durante o processo de treino.



Resultados







Agradecimentos

Este trabalho foi realizado com apoio do Itaú Unibanco S.A, por meio do Programa de Bolsas Itaú (PBI).

Integrante: Keith Ando Ogawa

Professor Orientador: Prof. Dr. Artur Jordão Lima Correia

< c²d_ centro de ciência de dados >

Universidade de São Paulo · Escola Politécnica