O Básico de Machine Unlearning em LLMs

Luca WB

2025-12-19

Table of contents

1 Ideia Geral

Machine Unlearning consiste na tecnica de tentar fazer modelos desaprenderem algo, seja um conhecimento especifico, uma lógica, ou uma habilidade inteira, esse post será especifico sobre bases de machine unlearning em LLMs

De maneira geral, ao tentar fazer uma llm esquecer de algo, passamos por um processo de fine tuning utilizando um uma loss especifica para isso junto com dois datasets.

A formula abaixo mostra como se calcula a Loss para um problema de MU (Machine Unlearning)

$$\min_{\theta} \underbrace{\mathbb{E}_{(x,y_f)\in D_f}[\ell(y_f|x;\theta)]}_{\text{Forget}} + \lambda \underbrace{\mathbb{E}_{(x,y)\in D_r}[\ell(y|x;\theta)]}_{\text{Retain}}$$

Vamos dissecar essa formula para não restar duvidas, a primeira parte, "Forget", é a função que loss que penaliza o modelo por ele dar a resposta original (ou qualquer uma que não seja a desejada) dado o input x e os pesos θ, isso é o que faz o Unlearning, note que ele usa o dataset D_f, que consiste em inputs e outputs desejaveis ***pós-unlearning***, ou seja, se eu quiser apagar o conhecimento do harry potter, a saida desejavel para a pergunta "Como Harry Potter encontrou a pedra filosofal no primeiro livro?" deve ser algo como "Não posso falar sobre isso", ou então "Não sei".

A segunda parte, "Retain", serve para que não ocorra um esquecimento generalizado do modelo (como por exemplo ele desaprender portugues), aqui usamos o dataset D_r, que tambem consiste em input output, porem aqui o output é a saida original do modelo, λ é um hiperparametro que regula o quanto o modelo deve priorizar manter o valor original dos pesos, ou seja, λ baixo, modelo pode esquecer de mais, λ alto, modelo pode não esquecer o suficente.

O E significa esperança (a média), no caso não usamos os resultados singulares de cada linha do dataset, e sim a média da loss deles.

$\min_\theta$ é simplesmente a notação que expressa que queremos minimizar isso modificando θ, em si não quer dizer nada matematicamente, isso é expresso subtraindo a loss dos valores dos pesos.

2 Otimizadores

Aparentemente otimizadoraes de segunda ordem que estimam a diagonal de uma hessiana (como a Sophia) funcionam melhor para unlearning do que otimizadores de primeira ordem (SGD, Adam, RMSprop etc)