Parameter-efficient fine-tuning (PEFT)

↳ high-level: we want to avoid modifying most of the pretrained model's parameters during fine-tuning.

↳ prompting: requires adjusting zero params to solve a downstream task

What is the sentiment of the below sentence? Answer w/ either "pos" or "neg".
<input sentence>
output: pos

↳ prompt engineering

↳ limitations:

→ hard to solve very complex reasoning / understanding tasks

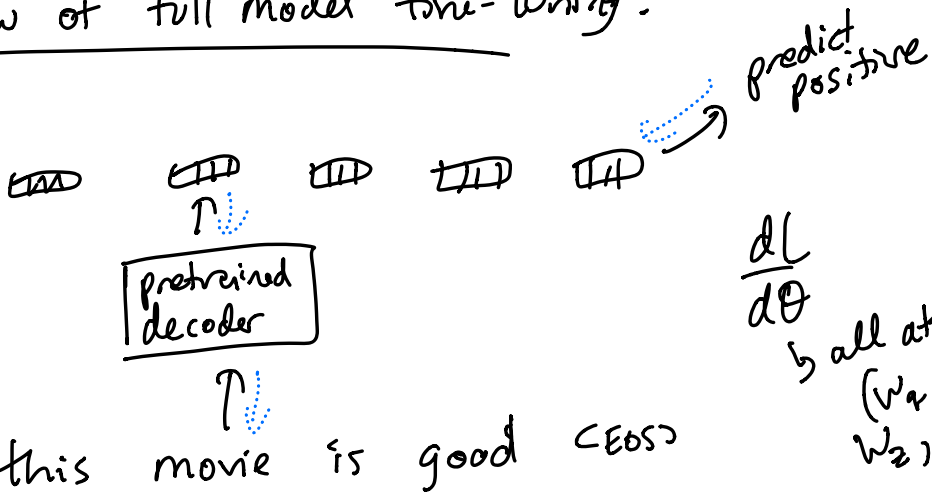
→ requirements for the pretrained model are immense

→ huge-scale pretraining

→ high quality large scale instruction tuning

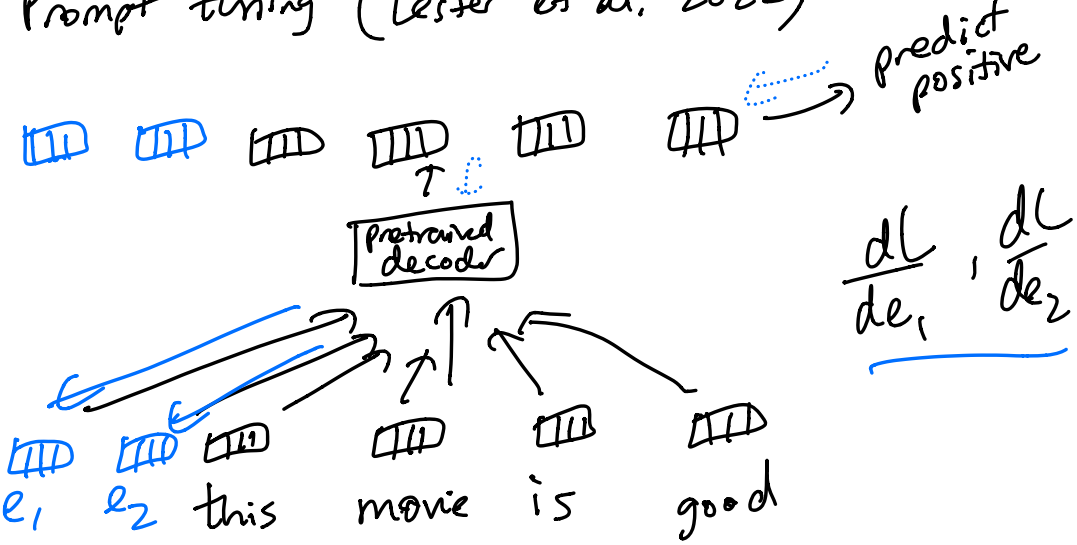
→ RLHF, requires access to very expensive human pref datasets

Review of full model fine-tuning:



$\frac{dL}{d\theta}$ to all attn params (W_Q, W_K, W_V) $W_2, b_1 \ldots N$

Prompt tuning (Lester et al. 2022)



$\frac{dL}{de_1}, \frac{dL}{de_2}$

update: keep all pretrained params frozen,

only do $e_{1_{new}} = e_{1_{old}} - \eta \frac{dL}{de_1}$

$e_{2_{new}} = \ldots$

LoRA (low-rank adaptation):

$$h = f(Wx) \qquad \frac{dL}{dW}$$

W is an $m \times n$ matrix

$\frac{dL}{dW}$ is also $m \times n$

$$W_{new} = W_{old} - \eta \underbrace{\frac{dL}{dW}}_{m \times n}$$

having two low-rank matrices A and B

A: $m \times r$ $\quad$ B: $n \times r$

r = rank parameter

want $r \ll m, n$

product $\underbrace{AB^T}$, $m \times n$

in LoRA:

$$h = f\left(\left(W_{pre} + \underbrace{AB^T}\right) x\right)$$

we compute $\frac{dL}{dA}$, $\frac{dL}{dB}$, much smaller than $\frac{dL}{dW}$

$\frac{dL}{dA}$: $m \times r$; $\frac{dL}{dB}$: $n \times r$

at the end of LoRA fine-tuning,
we have a separate A, B for each tuned weight matrix

$$W_{new} = W_{pre} + AB^T$$

$$f(W_{new} x)$$

QLoRA: quantized LoRA

normal models: FP32

↓

4 bit, 8 bit integer quantization