

Finetune

1. 全量微调(Full Fine-tuning)

- 全量参数微调
- 模型结构和参数均进行调整
- 训练时间长，消耗大

2. PEFT(PEFT - Parameter-Efficient Fine-Tuning)

LoRA (Low-Rank Adaptation)

- 低秩适应
- 通过低秩矩阵更新权重
- 提高训练效率

数学公式

$$W \in R^{r \times k} \text{ Lora } h = Wx + \triangle Wx = Wx + BAx \quad \dots$$

$$B \in R^{d \times r} \quad A \in R^{r \times k} \quad r \leq \min(d, k)$$

- 通过低秩矩阵 $\triangle W$ 更新权重
- 通过 $\triangle W$ 的梯度计算 $\nabla \triangle W$
- 通过 $\nabla \triangle W = \Sigma V^T r$ 更新 $\triangle W$

代码示例

- $A @ B @ B^T @ W @ 0$

Adapter

- Transformer模型的适配器
- Adapter
- Adapter

模型

Transformer模型的适配器

模型

Adapter模型

- 通过线性层 $d @ r(r \leq d)$
- 通过激活函数RELU或GELU
- 通过线性层 $r @ d$

模型

Adapter(x) = $x + W_{\text{up}} \cdot \text{ReLU}(W_{\text{down}} \cdot x)$

- $W_{\text{down}} \in R^{r \times d}$, $W_{\text{up}} \in R^{d \times r}$
- 通过线性层 $r @ d$

Bias-only

模型

- 偏向量(bias)的适配器

- 1%
1%
1%