

Finetune

1. 全量微调(Full Fine-tuning)

- 所有参数都进行更新
- 训练速度较慢
- 需要大量的显存

2. 参数高效微调(PEFT - Parameter-Efficient Fine-Tuning)

LoRA (Low-Rank Adaptation)

- 将权重矩阵分解为低秩矩阵
- 只训练低秩矩阵
- 保持原始权重矩阵不变

原理

假设 $W \in \mathbb{R}^{r \times k}$ 为 Lora 权重矩阵 $h = Wx + \triangle Wx = Wx + BAx$

其中 $B \in \mathbb{R}^{d \times r}$ $A \in \mathbb{R}^{r \times k}$ $r \ll \min(d, k)$

- 将权重矩阵分解为低秩矩阵
- 训练过程中只更新低秩矩阵
- 推理时不需要更新低秩矩阵
- 增量 ΔW 的计算
 - $\triangle W$ 的计算不需要存储完整的 W 矩阵
 - $\triangle W = U \Sigma V^T$ 其中 $U \in \mathbb{R}^{d \times r}$ $\Sigma \in \mathbb{R}^{r \times r}$ $V \in \mathbb{R}^{r \times k}$

初始化

- A 和 B 的初始化方式为 W_0

Adapter

- Transformer 的每个层都添加 Adapter
- Adapter 的初始化方式为 W_0
- Adapter 的推理不需要更新

原理

Transformer 的每个层都添加 Adapter

原理

Adapter 的推理

- 推理时不需要更新 d 维的 r ($r \ll d$)
- 推理时不需要更新 $ReLU$ 和 $GELU$
- 推理时不需要更新 r 维的 d

原理

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

- $W_{down} \in \mathbb{R}^{r \times d}$ $W_{up} \in \mathbb{R}^{d \times r}$ 的初始化方式为 W_0
- 推理时不需要更新 W_{down} 和 W_{up}

Bias-only

原理

- 推理时不需要更新 $(bias)$ 的初始化方式为 W_0

