轻量化微调

《大语言模型》编写团队：唐天一

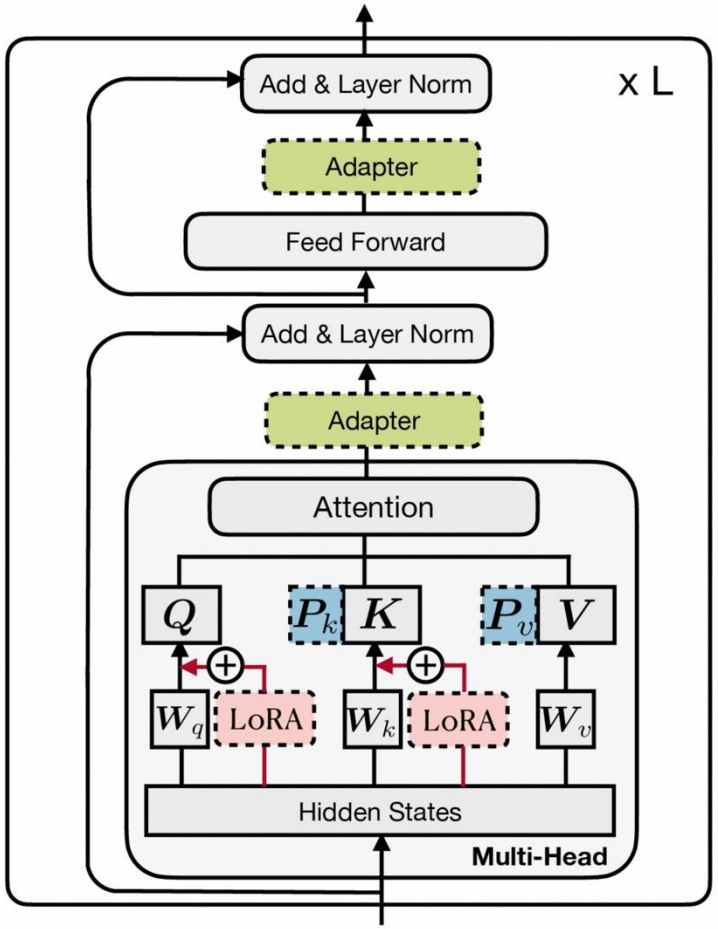
轻量化微调

AI Box

- 训练时每张GPU显存占用计算公式：

$$\underbrace{\frac{16P}{N}}_{\text{模型与优化器}} + \underbrace{2LBTH + 12BTV}_{\text{激活值}} + \underbrace{6}_{\text{其他}}$$

- 轻量化微调的目的
  - 减少模型训练参数量，从而降低显存占用
  - 同时尽可能接近全量微调的性能



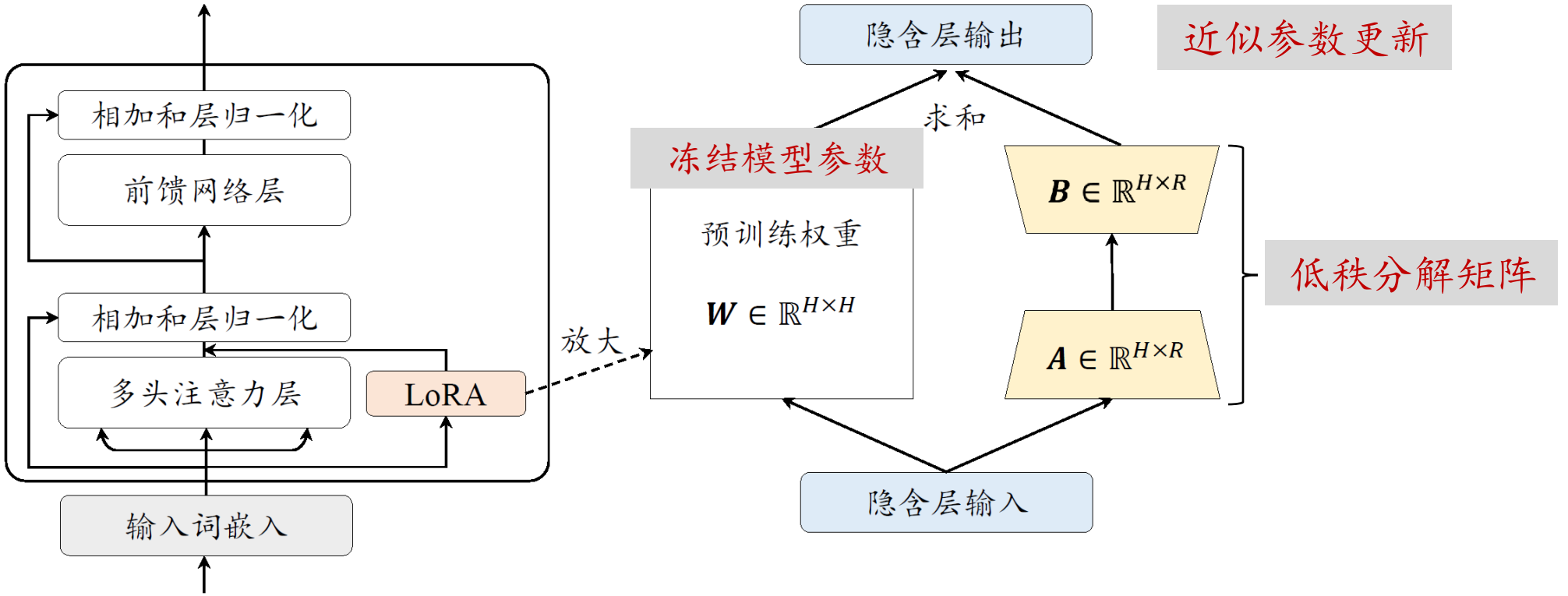
冻结LLM参数，只微调极少额外参数

Towards a Unified View of Parameter-Efficient Transfer Learning, **ICLR 2022**

低秩适配微调方法（LoRA）

AI Box

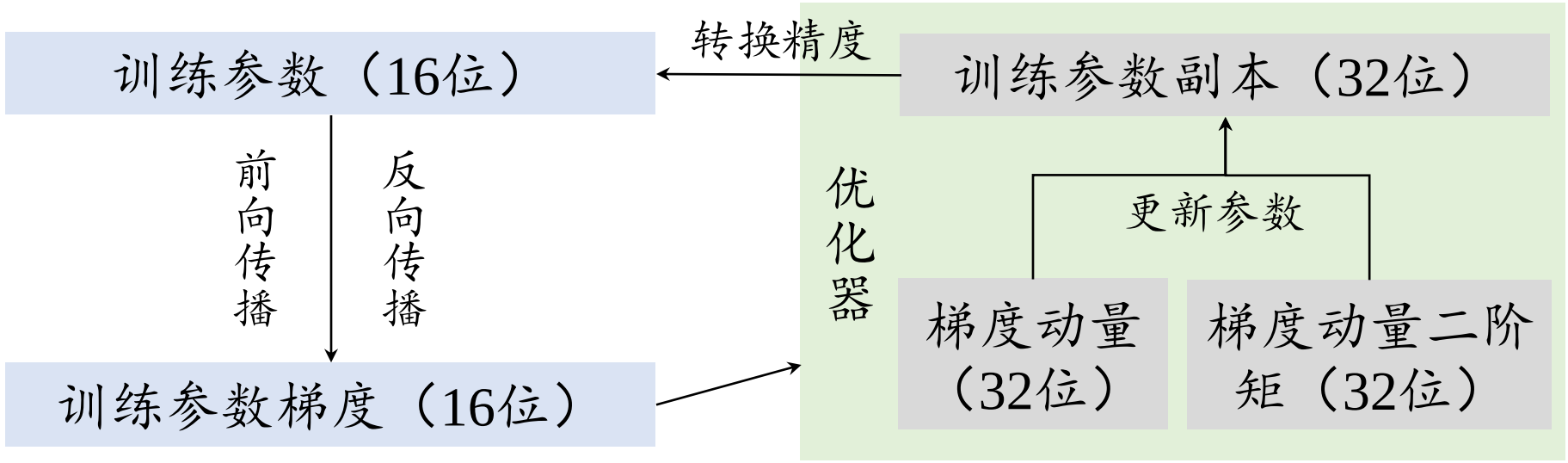
- LoRA 更新参数$\boldsymbol{W}$ 过程如下
  - $\boldsymbol{W} \leftarrow \boldsymbol{W} + \Delta\boldsymbol{W} = \boldsymbol{W} + \boldsymbol{A} \cdot \boldsymbol{B}^T$



低秩适配微调方法（LoRA）

AI Box

- LoRA 微调显存占用情况
  - 模型：$2P + 2P_{\text{LoRA}}$，梯度：$2P_{\text{LoRA}}$，优化器：$12P_{\text{LoRA}}$
  - 从 $16P$ 减为 $2P + 16P_{\text{LoRA}}$



低秩适配微调方法（LoRA）

AI Box

- LoRA 用于注意力层的线性变换 $\boldsymbol{W}^K$和$\boldsymbol{W}^V$
  - $P_{\text{LoRA}} = 2 * 2LHR + 2 * 2LHR = 8LHR$（R通常取16）
- 以 LLaMA 7B 为例，$P \approx 6.7 \times 10^9$，$P_{\text{LoRA}} \approx 1.7 \times 10^7$，$P_{\text{LoRA}} \ll P$
  - 模型和优化器占用从 $16P$ 降至 $2P$
  - 3090 24G 可以微调 7B 模型
- QLoRA：量化参数矩阵，用 4 比特存储模型参数
  - 模型占用从 $2P$ 降至 $0.5P$
  - A6000 48G 可以微调 65B 模型

指令微调的资源开销

AI Box

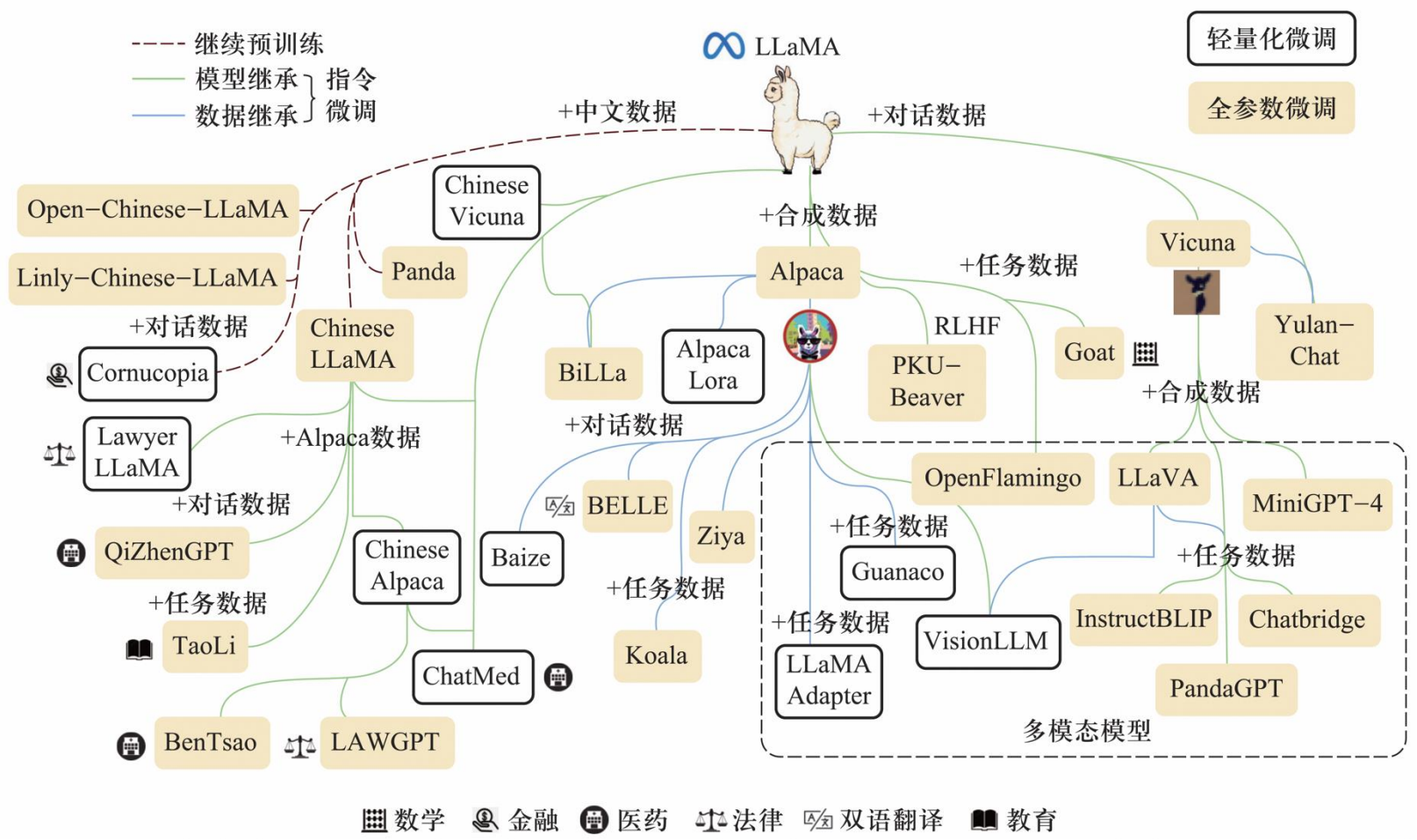
- LoRA微调 Alpaca-52K 所需的 A800 (80 G) 数量、批次大小和微调时间
  - 使用数据并行、ZeRO-3、BF16 和激活重计算技术

模型	GPU 数量	批次大小	时间
LLaMA (7 B)	1	16	2.3 h
LLaMA (13 B)	1	8	3.8 h
LLaMA (33 B)	1	1	10.2 h
LLaMA (65 B)	2	1	26.0 h

低秩适配微调方法（LoRA）

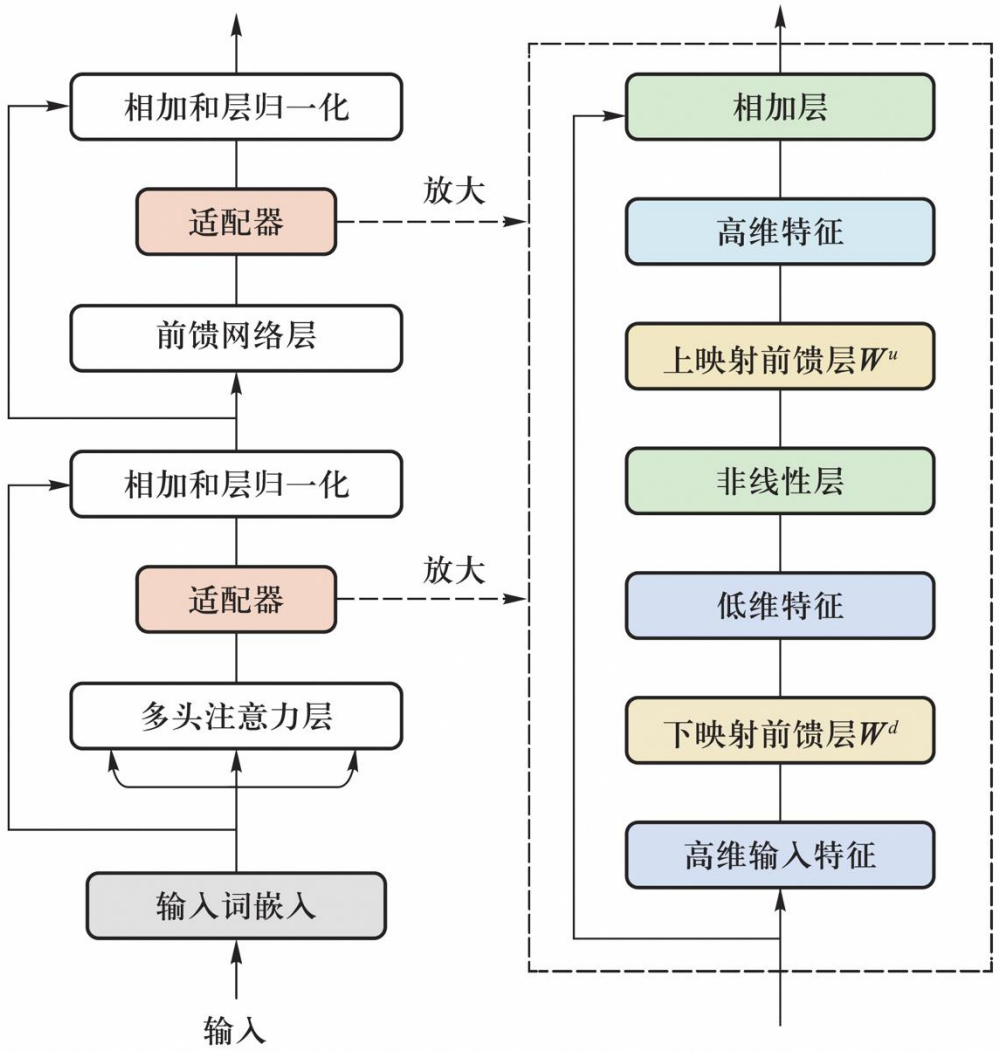
AI Box

- SFT 以及 LoRA 在下游应用中的适配



其他轻量化微调方法

- 适配器微调（Adapter tuning）
- 引入小型神经网络模块
- $\boldsymbol{h} = \boldsymbol{h} + \sigma(\boldsymbol{h} \cdot \boldsymbol{W}^d) \cdot \boldsymbol{W}^u$
  - $\boldsymbol{W}^d \in \mathbb{R}^{H \times R}$，$\boldsymbol{W}^u \in \mathbb{R}^{R \times H}$，$R \ll H$

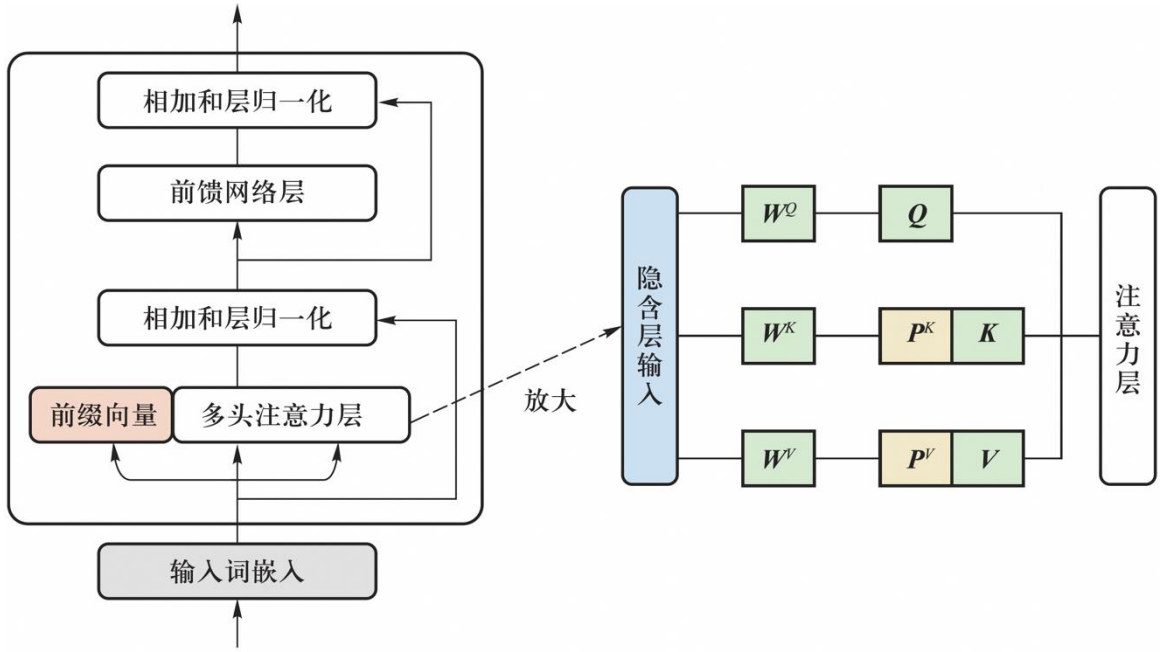


Parameter-Efficient Transfer Learning for NLP, **ICML 2019**

其他轻量化微调方法

AI Box

- 前缀微调（Prefix tuning）
  - 在多头注意力层中添加前缀参数
- Attention($\boldsymbol{XW}^Q, \boldsymbol{P}^K \oplus \boldsymbol{XW}^K, \boldsymbol{P}^V \oplus \boldsymbol{XW}^V$)$\boldsymbol{P}^K, \boldsymbol{P}^V \in \mathbb{R}^{L \times H}$， L常取10-100

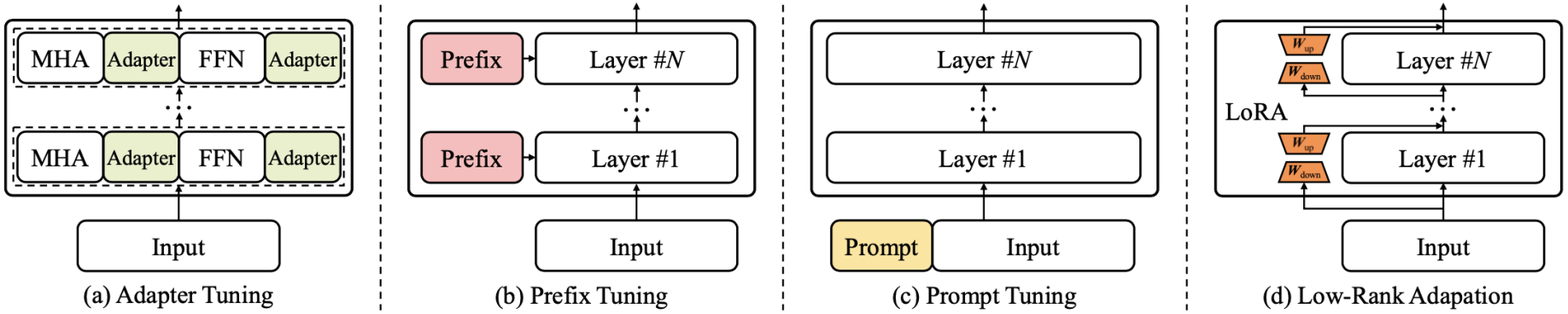


Prefix-Tuning: Optimizing Continuous Prompts for Generation, **ACL 2021**

其他轻量化微调方法

AI Box

- 不同轻量化方法的对比



- 适配器微调：每层都加一个小网络，加在两个核心组件后面
- 前缀微调：每层输入都加入prompt向量
- 提示微调：只在输入层加入prompt向量
- LoRA 微调：低秩分解学习权重增量部分

谢谢