大纲

1. Lora原理
2. sd 中如何使用 Lora
3. 如何使用 civitai 上的模型

Lora原理

1. 什么是 Lora

- Low-Rank Adaptation，主要用于大模型微调
- 利用矩阵低秩分解 $W = A \cdot B$，$W_{m \times n}$，$A_{m \times r}$，$B_{r \times n}$，$r \ll m$ 或 n

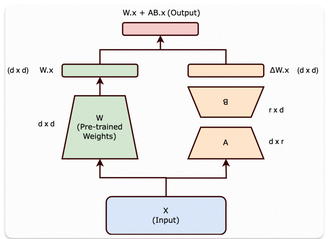
参数量：$m \times n \searrow r(m+n)$　1万 → 1K　10%

传统方法，对模型进行全参微调，即

$$W + \Delta W$$

Lora 方法假设 ΔW 低秩，可用 AB 近似

训练时增加一个旁路即可：



- 秩 r 的选择
  - 简单任务（如风格迁移）　r = 4～8
  - 复杂任务：r = 16～32
- 优点
  - 显存高效：仅需更新 1% 参数
  - 无推理延迟：训练完后 ΔW 可合并到 W_0 中，不增加模型计算量
  - 模块化：base 模型可自由切换多个 Lora 模型
- 库
  - Hugging Face PEFT 库
  - Unsloth 库

sd 中如何使用 Lora

资源

如何使用 Civitai 上的模型

1. Civitai 模型介绍
   - 围绕 Stable Diffusion / Flux
   - 模型类型：
     - Checkpoint
     - LoRA
2. 运行软件
   - Stable-diffusion-webui
   - Comfy UI
   - diffusers（适配不太好）
3. 网页操作演示
4. 如何在 Kaggle 平台上使用