LoRA

$h = W_0x + \Delta Wx = W_0x + BAx$

微调更新全连接：

推理延迟小，因为可以先把ABx算出来然后存起来

为什么要分解为AB？可以减少参数量，且前人研究发现W往往满足low rank