大小模型协同进化

> 主要参考材料：达摩院报告

1. 开头

核心解决一个问题：大模型怎么应用落地？

达摩院打出的这个概念，明确表明他们并不希望重复百度、newbing和chatgpt所走的道路，企图寻找新赛道。他们基于过去阿里成熟的移动互联、分布式和智能家居基础，提出将大模型部署应用于移动设备，i.e. 手机、汽车和家具。但他们面临如下问题：

1. 达摩院M6大模型参数量庞大，是目前参数量最大的大模型。这意味着不能直接部署。
    - GPT-3: 1750亿
    - 达摩院M6: 100000亿
2. 智能上：达摩院M6的智能并没有随参数量的暴增有明显优势，对比GPT-3。
    - 一个结论，增加参数量对智能的提升似乎又到了瓶颈。
    - 一些质疑：M6只对标了图文的多模态能力，而且训练集太小，仅292GB的文本（其他至少1TB），所以部分人质疑达摩院对M6智能的定义有所隐瞒。
    - 题外话，理论界还在继续优化数学部分 任何深度ReLU都可被写为浅层MLP 优化Attention从而大大减少模型参数量。
3. 移动端计算性能有限。限制了部署的模型的参数大小。
4. 移动端上的小模型如何更新，且保证更新后依旧适配指定任务。

现有的大模型工业化应用主要集中在搜索引擎和多模态交流，前者如支付宝的智能搜索框和newbing，后者如chatgpt-4。但它们的做法都是基于api前后端交互式：移动端上传需求，云端并行计算得到结果再返回给移动端。

在当前大模型冷静期，大小模型协同进化是一个重要应用方向，但协同演化理论一直处于发展阶段，没有形成成熟理论体系。其中，理论上有一座大山 遗忘性灾难 也被发现存在于大模型微调上，目前方法只能做到延缓遗忘速度。

2. 协同进化技术流程与难点挑战

部署移动端是一个很具有挑战性的赛道，整个流程如下：

1. 将大模型压缩为小模型；
2. 将小模型根据具体指定任务进行部署，将大模型中的知识在指定任务上迁移；
3. 设备上的小模型从指定任务中学习新样本，微调小模型；
4. 将小模型中被微调的参数向大模型进行升维；
5. 对升维后的参数进行清洗与治理；
6. 对升维参数按 *是否信任* 与大模型参数进行融合； *note: 有很多个小模型*
7. 大模型将融合后的参数降维到小模型维度，并下发更新结果；
8. 小模型按 *是否信任* 与下发参数进行融合；
9. 回到步骤3；实现报告中所说的 *终生学习*

而大小模型协同进化这点分如下几步技术难点（个人理解）：

1. 如何把大模型，通过蒸馏压缩和参数共享等技术，压缩为轻量级的小模型（1GB以下），并尽可能保留更多性能；
    1. 达摩院M6成功被压缩到10MB，且保留90%性能，但鉴于M6的智能，实际性能还有提升前景。
2. 如何将小模型在不同指定任务中进行部署和应用；
    1. 涉及 Few-shot Learning 等微调技术。
3. 【协同机制】小模型在大模型的知识常识上把知识迁移到具体任务上，并对新知识进行执行与学习。但是新学习到的知识会微调小模型的参数，如何把新参数和大模型的参数进行协同，再把协同结果降维到各个小模型是技术难点（也是理论难点）。
    1. 关键在于全文提到的 信任 问题。小模型的升维参数必定遗忘了部分常识信息，而大模型的降维参数必定遗忘部分指定任务相关知识。怎么融合，怎么清洗，怎么信任，是协同进化机制的难点。