问题:神经网络加上注意力机制，精度反而下降，为什么会这样呢?

回答一

提供了一些创新的注意力结构

- 软注意力机制与硬注意力机制
- 分层注意力机制
- 层次注意力机制
- 自顶向下注意力机制
- 多步注意力机制
- 多维注意力机制
- 方向型注意力机制
- 双向分块自注意力机制
- 强化学习自注意力机制
- 结构化自注意力机制

问题:深度学习中有什么非常惊艳或者轻量级的 Attention 操作?

回答一

这方面的资料非常多了 ~~~ 拿视觉应用来说，处理的数据一般是 BCHW 四个维度，都可以加 attention，最早是在 C 维度加，就是 SENet；在 HW 维度加，就是 non-local neural network；将 C、HW 两个串行或者并行起来，就是 CBAM。将 C、H、W 三个维度并行起来，就是 TripletAttnetion

最近看到在 batch 上加 attention 的工作，就是 CVPR2022 上的 BatchFormer

回答二

这里仅仅讨论视觉中的 attention

attention 的核心思想是根据全图的特征突出 feature map 中的某一核心部分，使得模型更加集中关注有效信息。所以前期的模型设计中更多采用一种类似于 mask 的方式，产生逐通道或逐像素的 mask 并与原 feature map 乘积，详见 senet，cbam

这种 attention 方式在全局池化的时候确实借鉴了全图的信息，但是仅用一种全图向量来增强原 feature map 本来就是受限的，于是参考 transformer 的 kqv，non local 横空出世，这种逐像素的 attention 扩大全局感受野的同时，权重的计算更加精细。但是由于要计算相关度矩阵，可能要消耗大量资源，所以后续也有一些轻量化方案，比如分块等等（待补充）

时间来到 2021，传统的卷积网络定式开始被打破，纯 transformer 开始进入 cv 领域，这类方法采用分 patch 的方式切分原图，并将其放入 transformer 中，也有无数的实例证明，在数据量极大的情况下，其性能能超越卷积。transformer 在攻克了 nlp 的大量任务之后，再次为 cv 带来了新的曙光。