智源学者成果展示——智源青年科学家

作者　兰艳艳（中国科学院计算技术研究所网络数据科学与技术重点实验室）

2020 年 6 月

项目负责人：兰艳艳

单位 / 团队：中国科学院计算技术研究所网络数据科学与技术重点实验室

成果名称：On Layer Normalization in the Transformer Architecture

成果简介：

与北京大学、微软亚洲研究院联合研究，该论文已被 ICML2020 录用。

Transformer 是自然语言处理中最常用的神经网络架构之一，而层归一化（Layer Normalization）在其中发挥了关键的作用。最初设计的 Transformer 将层归一化放置在残差块（Residual Blocks）之间，这种架构通常被称为 Post-Layer Normalization (Post-LN)。与传统的神经网络相比，训练 Post-LN Transformer 更加困难，特别是，要从头开始训练 Post-LN Transformer，任何基于梯度的优化方法都需要一个学习率预热阶段。这样的预热阶段减缓了优化，并带来了更多的超参数调整。研究表明，最终的模型性能对预热迭代次数的设置非常敏感。

我们利用平均场理论研究了初始时的优化行为。我们发现在 Post-LN Transformer 中，输出层附近的期望梯度非常大，若不进行预热，这些不平衡的梯度会使得优化过程不稳定。此外，我们发现层归一化对于控制梯度的尺度具有重要作用，这促使我们研究是否有其他放置层归一化的方式可以解决这一问题。特别地，我们研究了 Pre-LN Transformer，该架构将层归一化放置于残差块的内部，并在预测前额外增加了一个层归一化。我们从理论上和经验上表明，在初始阶段，Pre-LN Transformer 的梯度表现良好，因而不再需要学习率预热。

我们在多个任务上进行了实验，结果显示，Pre-LN Transformer 训练更加稳定，不再需要学习率预热，且具有更快的收敛速度。例如在无监督预训练上，Pre-LN Transformer 带来了 40% 的训练加速，这一点对于训练大规模模型来说尤其重要。

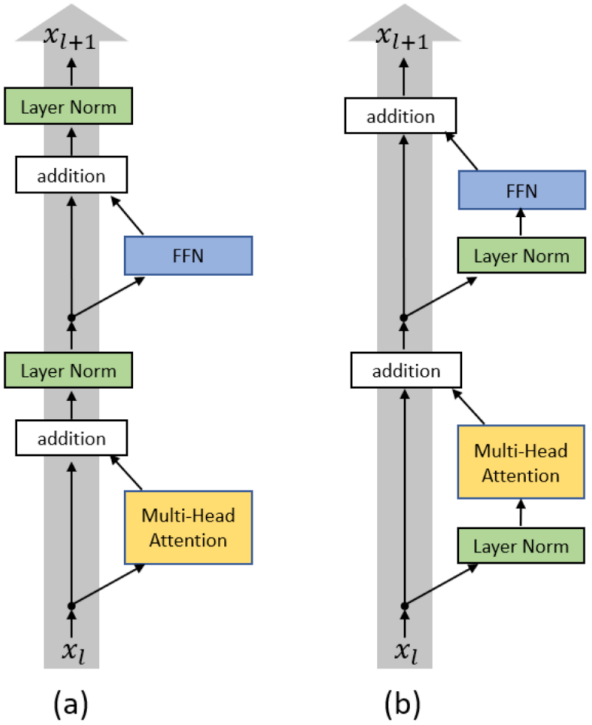

图　(a) Post-LN Transformer Layer　(b) Pre-LN Transformer Layer

Beijing Academy of Artificial Intelligence

微信关注

北京智源人工智能研究院