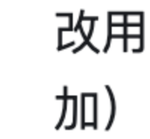
《Attention Is All You Need》的底层逻辑



扫帚的影子
弱水三千 只取一瓢

已关注

4 人赞同了该文章 ›

这篇论文的核心理念可以用一个比喻来理解：**将信息处理从“逐字扫描”升级为“全局检索”。**

1. 核心矛盾：传统序列模型的“视野局限”

- **传统RNN的困境**：
  想象你在读一本书，但只能通过一个狭窄的视窗逐字移动（时间步依赖）。要理解第100页的内容，必须从第1页开始一步步“爬”到第100页（梯度消失/爆炸问题）。
  → **问题本质**：顺序计算的“信息传递瓶颈”。
- **CNN的妥协**：
  改用卷积核像放大镜一样扫描局部区域，但若想看到整页内容，需要堆叠多层放大镜（深度增加），效率低下且难以捕捉远距离依赖。
  → **问题本质**：局部感受野与长程依赖的矛盾。

Transformer的突破：
直接抛弃“逐字扫描”和“局部放大”，改用**全局检索机制**——自注意力✦允许模型在任何位置“一眼看到”整个序列，并通过权重动态聚焦关键信息。

---

2. 底层逻辑一：注意力即“动态信息检索”

- **检索过程的三要素**：
  每个词（Query）主动向序列中所有词（Key）发起“提问”，根据匹配程度（相似度）分配权重，最终汇总值（Value）得到新表示。
  - **数学表达**：
    Attention(Q,K,V)=softmax(QKTdk)VAttention(*Q*,*K*,*V*)=softmax(*dkQKT*)*V*
  - **物理意义**：
    每个词通过“提问-匹配-汇总”动态构建上下文感知的表示，而非静态的局部特征。

- **为何用点积+缩放?**
  - 点积高效计算相似度，但维度高时点积值过大，导致softmax梯度饱和（类似“强光下看不清细节”）。
  - 缩放因子dk*dk*将数值拉回合理范围，保持梯度稳定。

---

3. 底层逻辑二：多头机制✦——“分治法”增强表达能力

- **单一注意力的局限**：
  若只用一个注意力头，模型可能陷入“单一视角”的局限（例如只关注语法或语义）。
- **多头注意力的设计**：
  将输入投影到多个子空间（如8个头），并行执行独立检索，最后拼接结果。
  - **类比**：
    让8个专家同时阅读同一段文字，各自关注语法结构、情感倾向、指代关系等不同方面，再综合意见。
  - **数学表达**：
    MultiHead(Q,K,V)=Concat(head1,...,headh)WOMultiHead(*Q*,*K*,*V*)=Concat(head1,...,head*h*)*WO*

- **效果**：
  模型能同时捕获多种依赖模式（如局部语法、长距离指代），避免信息混合后的“平均化稀释”。

---

4. 底层逻辑三：位置编码✦——为无状态模型注入“顺序感知”

- **自注意力的缺陷**：
  虽然能捕捉任意位置的关系，但**天生无视顺序**（排列不变性）。例如，“猫追狗”和“狗追猫”在原始注意力中可能无法区分。
- **解决方案**：
  通过**位置编码**（Positional Encoding）为每个位置赋予独特的“坐标信号”。
  - **正弦/余弦函数的选择**：
    固定编码能让模型轻松学习相对位置关系（例如位置pos+k*pos*+*k*的编码可由pos*pos*线性变换得到）。
  - **可视化意义**：
    将词嵌入从“词袋空间”映射到“词序空间”，类似为黑白照片添加色彩通道。

---

5. 底层逻辑四：残差与归一化——深层网络的“稳定器”

- **深层网络的挑战**：
  随着层数增加，梯度可能消失或爆炸，导致训练不稳定。

- **残差连接✦（Residual Connection）**：
  让每一层的输出为x+Sublayer(x)*x*+Sublayer(*x*)，相当于保留原始信号的同时叠加新特征。
  作用：梯度可直接回传，缓解深层网络的退化问题。
- **层归一化（LayerNorm）**：
  对每层的输出进行标准化，避免数值随深度累积失控。
  作用：稳定训练动态，加速收敛。

---

6. 底层逻辑五：从“序列 transduction”到“通用计算框架”

- **编码器-解码器架构的泛化性**：
  Transformer的模块化设计（自注意力、交叉注意力、前馈网络）使其不局限于翻译任务。
- **关键洞察**：
  通过堆叠相同的层（如6层编码器），模型能逐步提炼抽象特征——低层捕捉局部语法，高层整合语义逻辑。
- **与RNN/CNN的本质区别**：
  - RNN：时间步间强耦合，信息流动受限；
  - CNN：层次化局部归纳偏置；
  - Transformer：完全数据驱动的全局关系建模，依赖注意力权重动态调整信息流。

总结

Transformer的成功源于对传统模型的**降维打击**：

1. **抛弃归纳偏置**：不依赖局部性或顺序性先验，完全由数据驱动关系建模；
2. **拥抱并行计算**：自注意力的矩阵运算天然适合GPU加速；
3. **解耦信息与位置**：通过位置编码分离内容与顺序，使模型更灵活。

它的底层逻辑启示我们：**复杂问题的解法可能藏在对“常识”的颠覆中**。正如爱因斯坦所说：“我们不能用制造问题时的同一思维水平来解决问题。”

发布于 2025-02-24 09:54 · IP 属地北京