基于推理时间复杂度与推理空间复杂度比较深度卷积、线性 RNN、多头自注意力

朱梦

初稿于 2025-07-05，修改于 2025-07-06

1. 深度卷积推理的时间复杂度与空间复杂度

设输入为 $\boldsymbol{X} \in \mathbb{R}^{C\times L}$。设卷积核大小为 k（k 被设置为奇数），填充 p 为 $\lfloor k/2 \rfloor$（对称填充），步长 s 为 1。第一步通过 im2col 将输入 $\boldsymbol{X}$ 形状变换为 $\boldsymbol{Z} \in \mathbb{R}^{C\times k\times L}$：

$$z_{i,t,m} = x_{i,m+t-p} \tag{1}$$

其中，$t \in \{1, 2 \dots, k\}$ 表示卷积核内的位置索引，$m \in \{1, 2, \dots, L\}$ 表示输出位置索引，$(m+t-p)$ 是输入位置索引，需满足 $1 \le m+t-p \le L$（越界时视为填充值，通常为 0）。设深度卷积所需的参数为 $\boldsymbol{\Theta} \in \mathbb{R}^{C\times k}$。第二步对 $\boldsymbol{Z}$ 做矩阵运算：

$$y_{i,m} = \sum_{t=1}^{k} \theta_{i,t} z_{i,t,m} \tag{2}$$

基于式(1)，(2)，深度卷积推理的时间复杂度为 $\mathcal{O}(CkL)$（由 im2col 变换和矩阵运算主导，各 $\mathcal{O}(CkL)$），空间复杂度为 $\mathcal{O}(CkL)$（由 im2col 输出 $\boldsymbol{Z}$ 主导）。

2. 线性 RNN 推理的时间复杂度与空间复杂度

设如下极简 RNN：

$$\boldsymbol{y}_t = \sum_{k=1}^{t} \boldsymbol{\Lambda}^{t-k} \boldsymbol{x}_k \tag{3}$$

其中，$\boldsymbol{x}_k \in \mathbb{R}^{D\times T}$，$\boldsymbol{\Lambda} \in \mathbb{R}^{D\times D}$ 表示对角矩阵。

基于式(3)和 Prefix Sum 算法，时间复杂度为 $\mathcal{O}(DT)$（并行深度为 $\mathcal{O}(\log T)$，需 $\mathcal{O}(DT)$ 个处理器），空间复杂度为 $\mathcal{O}(DT)$。

3. 多头自注意力推理的时间复杂度与空间复杂度

多头自注意力（MHA）为：

$$
\begin{aligned}
&\text{MHA}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Concat}(\mathbf{head}_1,\mathbf{head}_2,\dots,\mathbf{head}_h),\\
&\mathbf{head}_i = \text{Attn}(\boldsymbol{X}\boldsymbol{\Theta}_{1,i},\boldsymbol{X}\boldsymbol{\Theta}_{2,i},\boldsymbol{X}\boldsymbol{\Theta}_{3,i})\\
&\qquad = \text{Softmax}(\kappa(\boldsymbol{X}\boldsymbol{\Theta}_{1,i})(\boldsymbol{X}\boldsymbol{\Theta}_{2,i})^{\mathrm{T}})(\boldsymbol{X}\boldsymbol{\Theta}_{3,i})
\end{aligned}
\tag{4}
$$

其中，$\boldsymbol{X}\in\mathbb{R}^{s\times d}$，$\mathbf{head}_i\in\mathbb{R}^{s\times(d/h)}$，$\boldsymbol{\Theta}_{1,i},\boldsymbol{\Theta}_{2,i}\in\mathbb{R}^{d\times d_K}$，$\boldsymbol{\Theta}_{3,i}\in\mathbb{R}^{d\times(d/h)}$，$\kappa$ 表示缩放因子。

基于式(4)，MHA 推理的时间复杂度为 $6sd^2+2s^2d+3hs^2+2s^2d\approx\mathcal{O}(sd^2+s^2d)$，空间复杂度为 $3sd+hs^2+hs^2+sd\approx\mathcal{O}(hs^2+sd)$。

4. 总结

当固定模型宽度时，深度卷积推理为线性时间复杂度，即正比于序列长度。线性 RNN 推理为线性时间复杂度，即正比于序列长度。MHA 推理为二次方时间复杂度，即正比于序列长度的二次方。

当固定模型宽度时，深度卷积推理为线性空间复杂度，即正比于序列长度。线性 RNN 推理为线性空间复杂度，即正比于序列长度。MHA 推理为二次方空间复杂度，即正比于序列长度的二次方。

因此，如何降低 MHA 推理复杂度既为难点也为热点。