全连接层偏置项是否保留的研究

朱梦

初稿于 2025-06-17，修改于 2025-06-21

1. 问题定义及其分析

在神经网络中，全连接层（也称为仿射层或投影层）的偏置项在某些情况下可能变得“无用”，即可以被省略而不影响模型的表达能力或性能。这主要取决于下一层的类型。

2. 下一层类型

2.1 Q/K 投影层 + Softmax 层

设输入为 $\boldsymbol{X} \in \mathbb{R}^{s \times d_{\text{in}}}$，Q 投影层的矩阵参数为 $\boldsymbol{\Theta}_Q \in \mathbb{R}^{d_{\text{in}} \times d}$，偏置项参数为 $\boldsymbol{b}_Q \in \mathbb{R}^{1 \times d}$，K 投影层的矩阵参数为 $\boldsymbol{\Theta}_K \in \mathbb{R}^{d_{\text{in}} \times d}$，偏置项参数为 $\boldsymbol{b}_K \in \mathbb{R}^{1 \times d}$，那么注意力分数计算为：

$$
\begin{aligned}
\mathrm{Softmax}(\boldsymbol{q}_n \boldsymbol{K}^{\mathrm{T}}) &= \frac{e^{<\boldsymbol{q}_n, \boldsymbol{k}_i>}}{\sum_{j=1}^{s} e^{<\boldsymbol{q}_n, \boldsymbol{k}_j>}} = \frac{e^{<\boldsymbol{x}_n \boldsymbol{\Theta}_Q + \boldsymbol{b}_Q, \boldsymbol{x}_i \boldsymbol{\Theta}_K + \boldsymbol{b}_K>}}{\sum_{j=1}^{s} e^{<\boldsymbol{x}_n \boldsymbol{\Theta}_Q + \boldsymbol{b}_Q, \boldsymbol{x}_j \boldsymbol{\Theta}_K + \boldsymbol{b}_K>}} \\
&= \frac{e^{<\boldsymbol{x}_n \boldsymbol{\Theta}_Q, \boldsymbol{x}_i \boldsymbol{\Theta}_K>} e^{<\boldsymbol{x}_n \boldsymbol{\Theta}_Q, \boldsymbol{b}_K>} e^{<\boldsymbol{x}_i \boldsymbol{\Theta}_K, \boldsymbol{b}_Q>} e^{<\boldsymbol{b}_Q, \boldsymbol{b}_K>}}{\sum_{j=1}^{s} e^{<\boldsymbol{x}_n \boldsymbol{\Theta}_Q, \boldsymbol{x}_j \boldsymbol{\Theta}_K>} e^{<\boldsymbol{x}_n \boldsymbol{\Theta}_Q, \boldsymbol{b}_K>} e^{<\boldsymbol{x}_j \boldsymbol{\Theta}_K, \boldsymbol{b}_Q>} e^{<\boldsymbol{b}_Q, \boldsymbol{b}_K>}} \\
&= \frac{e^{<\boldsymbol{x}_n \boldsymbol{\Theta}_Q, \boldsymbol{x}_i \boldsymbol{\Theta}_K>} e^{<\boldsymbol{x}_i \boldsymbol{\Theta}_K, \boldsymbol{b}_Q>}}{\sum_{j=1}^{s} e^{<\boldsymbol{x}_n \boldsymbol{\Theta}_Q, \boldsymbol{x}_j \boldsymbol{\Theta}_K>} e^{<\boldsymbol{x}_j \boldsymbol{\Theta}_K, \boldsymbol{b}_Q>}} \\
&= \frac{e^{<\boldsymbol{q}_n, \boldsymbol{x}_i \boldsymbol{\Theta}_K>}}{\sum_{j=1}^{s} e^{<\boldsymbol{q}_n, \boldsymbol{x}_j \boldsymbol{\Theta}_K>}} \triangleq \frac{e^{<\boldsymbol{q}_n \boldsymbol{\Theta}_K^{\mathrm{T}}, \boldsymbol{x}_i>}}{\sum_{j=1}^{s} e^{<\boldsymbol{q}_n \boldsymbol{\Theta}_K^{\mathrm{T}}, \boldsymbol{x}_j>}} \\
&\triangleq \frac{e^{<\boldsymbol{x}_n (\boldsymbol{\Theta}_Q \boldsymbol{\Theta}_K^{\mathrm{T}}) + \boldsymbol{b}_Q \boldsymbol{\Theta}_K^{\mathrm{T}}, \boldsymbol{x}_i>}}{\sum_{j=1}^{s} e^{<\boldsymbol{x}_n (\boldsymbol{\Theta}_Q \boldsymbol{\Theta}_K^{\mathrm{T}}) + \boldsymbol{b}_Q \boldsymbol{\Theta}_K^{\mathrm{T}}, \boldsymbol{x}_j>}} \triangleq \frac{e^{<\widehat{\boldsymbol{x}}_n, \boldsymbol{x}_i>}}{\sum_{j=1}^{s} e^{<\widehat{\boldsymbol{x}}_n, \boldsymbol{x}_j>}}
\end{aligned}
\tag{1}
$$

由式(1)可知：K 投影层的偏置项参数严格冗余。

2.2　投影层 + 标准化层

设输入为 $\boldsymbol{X} \in \mathbb{R}^{d_{\text{in}}\times s}$，投影层矩阵参数为 $\boldsymbol{\Theta} \in \mathbb{R}^{d_{\text{out}}\times d_{\text{in}}}$，偏置项参数为 $\boldsymbol{b} \in \mathbb{R}^{d_{\text{out}}\times 1}$，那么投影层有：

$$z_{j,k} = b_j + \sum_{i=1}^{d_{\text{in}}} \theta_{j,i} x_{i,k} \tag{2}$$

（一）下一层为 BN 层。设 BN 层的缩放平移向量为 $\boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^{d_{\text{out}}\times 1}$，那么 BN 层有：

$$\begin{gathered} y_{j,k} = \widehat{z}_{j,k}\gamma_j + \beta_j,\ \widehat{z}_{j,k} = \frac{z_{j,k} - \mu_j}{\sqrt{\sigma_j^2 + \epsilon}}, \\ \sigma_j^2 = \frac{1}{s}\sum_{n=1}^{s}(z_{j,n} - \mu_j)^2,\ \mu_j = \frac{1}{s}\sum_{n=1}^{s} z_{j,n} \end{gathered} \tag{3}$$

将式(2)代入式(3)有：

$$\begin{aligned} & y_{j,k} = \widehat{z}_{j,k}\gamma_j + \beta_j,\ \widehat{z}_{j,k} = \frac{\sum_{i=1}^{d_{\text{in}}} \theta_{j,i}x_{i,k} - \frac{1}{s}\sum_{k=1}^{s}\sum_{i=1}^{d_{\text{in}}} \theta_{j,i}x_{i,k}}{\sqrt{\sigma_j^2 + \epsilon}}, \\ & \sigma_j^2 = \frac{1}{s}\sum_{n=1}^{s}\left(\sum_{i=1}^{d_{\text{in}}} \theta_{j,i}x_{i,n} - \frac{1}{s}\sum_{n=1}^{s}\sum_{i=1}^{d_{\text{in}}} \theta_{j,i}x_{i,n}\right)^2, \\ & \mu_j = \frac{1}{s}\sum_{n=1}^{s}\left(b_j + \sum_{i=1}^{d_{\text{in}}} \theta_{j,i}x_{i,n}\right) = b_j + \frac{1}{s}\sum_{n=1}^{s}\sum_{i=1}^{d_{\text{in}}} \theta_{j,i}x_{i,n} \end{aligned} \tag{4}$$

在式(4)中，$\widehat{z}_{j,k}$ 的计算并没有出现 b_j，因此下一层为 BN 层时，投影层的偏置项参数严格冗余。

（二）下一层为 LN 层。LN 层有：

$$\begin{gathered} y_{j,k} = \widehat{z}_{j,k}\gamma_j + \beta_j,\ \widehat{z}_{j,k} = \frac{z_{j,k} - \mu_k}{\sqrt{\sigma_k^2 + \epsilon}}, \\ \sigma_k^2 = \frac{1}{d_{\text{out}}}\sum_{m=1}^{d_{\text{out}}}(z_{m,k} - \mu_k)^2,\ \mu_k = \frac{1}{d_{\text{out}}}\sum_{m=1}^{d_{\text{out}}} z_{m,k} \end{gathered} \tag{5}$$

将式(2)代入式(5)有：

$$y_{j,k}=\widehat{z}_{j,k}\gamma_j+\beta_j,\ \widehat{z}_{j,k}=\frac{z_{j,k}-\mu_k}{\sqrt{\sigma_k^2+\epsilon}},$$
$$\sigma_k^2=\frac{1}{d_{\text{out}}}\sum_{m=1}^{d_{\text{out}}}\left(\left(b_m+\sum_{i=1}^{d_{\text{in}}}\theta_{m,i}x_{i,k}\right)-\left(\frac{1}{d_{\text{out}}}\sum_{m=1}^{d_{\text{out}}}b_m+\frac{1}{d_{\text{out}}}\sum_{m=1}^{d_{\text{out}}}\sum_{i=1}^{d_{\text{in}}}\theta_{m,i}x_{i,k}\right)\right)^2,$$
$$\mu_k=\frac{1}{d_{\text{out}}}\sum_{m=1}^{d_{\text{out}}}\left(b_m+\sum_{i=1}^{d_{\text{in}}}\theta_{m,i}x_{i,k}\right)=\frac{1}{d_{\text{out}}}\sum_{m=1}^{d_{\text{out}}}b_m+\frac{1}{d_{\text{out}}}\sum_{m=1}^{d_{\text{out}}}\sum_{i=1}^{d_{\text{in}}}\theta_{m,i}x_{i,k} \tag{6}$$

在式(6)中，$\widehat{z}_{j,k}$ 的计算是出现 b_j 的，因此下一层为 LN 层时，投影层的偏置项参数非严格冗余。

2.3 工程实践约定俗成

对于 pre-norm 结构，K 投影层的偏置项参数严格冗余，其它投影层的偏置项参数非冗余。对于 post-norm 结构，自注意力机制中 $\boldsymbol{Q}/\boldsymbol{K}/\boldsymbol{V}$ 的偏置项通常无用，因为 LN 的偏移参数 $\boldsymbol{\beta}$ 已经具备了学习数据分布偏移的能力。对于 post-norm 结构，FFN 中第二层投影层的偏置项参数也通常无用。