PIDformer 的设计与理论分析

雍征彼
北京理工大学

日期：2024 年 11 月 18 日

引言

现实世界中不只有真实的物理系统可以被当成被控对象，还有很多虚拟的系统也可以被当成被控对象。这些虚拟的系统是由计算机模拟的，比如计算机游戏、虚拟现实、仿真系统等。这些虚拟系统的特点是可以通过软件的方式对其进行修改，从而改变系统的行为。这种特性使得虚拟系统可以被当成一个黑盒子，通过输入输出的方式来控制系统的行为。在这种情况下，控制系统的设计就变成了一个数据驱动的问题。而在这其中，深度学习可能是一个很有研究价值的被控对象，围绕深度学习模型构建复杂采样系统很有研究的价值。

深度学习模型不同于现实世界物理模型的是，其内部状态完全存储于模型的参数中，这就意味着如果我们能对其建立状态空间模型，则该离散时间系统模型的所有状态都是可观测的，这天生有利于控制系统的设计，因为在现实中几乎无法实现的状态观测问题在这里是不存在的。因此，深度学习模型的控制问题是一个很有研究价值的问题。

随着深度学习的快速发展，Transformer 成为处理序列数据的主流模型。然而，Transformer 的自注意力机制存在两个主要问题：对输入扰动敏感和输出表示的秩坍塌问题。这限制了其在复杂任务中的表现。《PIDformer: Transformer Meets Control Theory》[1] 介绍了 PIDformer，这是一种新型 Transformer 架构，通过引入比例-积分-微分（PID）控制器克服了上述问题。PIDformer 将控制理论融入深度学习，增强了模型的鲁棒性和稳定性。

1. 自注意力机制

Transformer 的核心在于自注意力机制，其基本思想是为每个输入位置计算与其他位置的注意力权重。假设输入令牌序列为：

$$\mathrm{X}^{\ell} := \left[\mathrm{x}^{\ell}(1), \cdots, \mathrm{x}^{\ell}(N)\right]^{\top}, \quad \mathrm{X}^{\ell} \in \mathbb{R}^{N \times D_x}.$$

第 ℓ 层的查询、键和值矩阵定义为：

$$\mathrm{Q}^{\ell} = \mathrm{X}^{\ell}\mathrm{W}_{Q}^{\ell^{\top}}, \quad \mathrm{K}^{\ell} = \mathrm{X}^{\ell}\mathrm{W}_{K}^{\ell^{\top}}, \quad \mathrm{V}^{\ell} = \mathrm{X}^{\ell}\mathrm{W}_{V}^{\ell^{\top}},$$

其中，$\mathrm{W}_{Q}^{\ell}, \mathrm{W}_{K}^{\ell} \in \mathbb{R}^{D_{qk} \times D_x}$，$\mathrm{W}_{V}^{\ell} \in \mathbb{R}^{D \times D_x}$ 为可训练权重矩阵。

自注意力机制计算第 i 个输出：

$$u^{\ell}(i) = \sum_{j=1}^{N} \mathrm{softmax}\left(\frac{\mathrm{q}^{\ell}(i)^{\top}\mathrm{k}^{\ell}(j)}{\sqrt{D_{qk}}}\right)\mathrm{v}^{\ell}(j),$$

其中 $\mathrm{q}^{\ell}(i)$ 是查询矩阵 Q^{ℓ} 的第 i 行，$\mathrm{k}^{\ell}(j)$ 和 $\mathrm{v}^{\ell}(j)$ 分别是键矩阵和值矩阵的第 j 行。Softmax 函数确保注意力权重的归一化。

尽管这种机制有效，但对输入噪声或干扰的敏感性以及输出表示的秩坍塌问题限制了其性能。具体而言，随着模型层数的增加，令牌嵌入趋于相似，从而降低表示能力。

2. 自注意力机制的控制框架

为了从控制理论的角度理解自注意力机制，文章将其建模为一种状态空间模型 (SSM)。给定值矩阵：

$$\mathrm{V}^{\ell} := \left[\mathrm{v}^{\ell}(1), \cdots, \mathrm{v}^{\ell}(N)\right]^{\top}, \quad \mathrm{V}^{\ell} \in \mathbb{R}^{N \times D}.$$

状态信号 $\mathrm{v}(x,t)$ 定义为：

$$\frac{\partial \mathrm{v}(x,t)}{\partial t} = \int_{\Omega} (\mathrm{v}(y,t) - \mathrm{v}(x,t)) K(x,y,t) dy + \mathrm{z}(x,t),$$

其中，$\mathrm{z}(x,t)$ 是控制输入，$K(x,y,t)$ 是核函数。若无控制输入（$\mathrm{z}(x,t) = 0$），此模型是自治的，通过非局部全变差最小化实现信号平滑。然而，这种机制导致了表示信息的丢失和秩坍塌。

2.1 状态空间模型与自注意力的联系

状态空间模型的梯度流可表示为：

$$\frac{\partial \mathrm{v}(x,t)}{\partial t} = -\nabla_{\mathrm{v}} J(\mathrm{v}),$$

其中：

$$J(\mathrm{v}) = \frac{1}{2} \int_{\Omega \times \Omega} \|\mathrm{v}(x) - \mathrm{v}(y)\|_2^2 k(x,y) dx dy.$$

通过离散化，状态空间模型可近似为：

$$\mathrm{v}(x, t+1) \approx \int_{\Omega} K(x,y,t) \mathrm{v}(y,t) dy,$$

其中：

$$K(x,y,t) := \frac{\exp(\mathrm{q}(x,t)^{\top} \mathrm{k}(y,t) / \sqrt{D_{qk}})}{\int_{\Omega} \exp(\mathrm{q}(x,t)^{\top} \mathrm{k}(y',t) / \sqrt{D_{qk}}) dy'}.$$

这恰好对应于自注意力机制公式：

$$\mathrm{v}^{\ell+1}(i) \approx \sum_{j=1}^{N} \mathrm{softmax}\left(\frac{\mathrm{q}^{\ell}(i)^{\top} \mathrm{k}^{\ell}(j)}{\sqrt{D_{qk}}}\right) \mathrm{v}^{\ell}(j).$$

2.2 状态空间模型的稳定性分析

以下引理揭示了状态空间模型在时间趋于无穷时的行为：

引理 0.1. *给定矩阵 $K - I$ 的复数谱 $\{\alpha_1, \alpha_2, \ldots, \alpha_M\}$，其解为：*

$$\mathrm{V}(t) = P \exp(Jt) P^{-1} \mathrm{V}^0,$$

其中 $J = diag(J_{\alpha_1, m_1}, \cdots, J_{\alpha_M, m_M})$ 是 Jordan 分解。

根据上述结论，状态空间模型的解会收敛到一个低秩状态，从而导致表示能力不足。

3. PID 控制器的引入

为了解决上述问题，PIDformer 在状态空间模型中引入了 PID 控制器。其动态方程为：

$$\frac{\partial \mathrm{v}(x,t)}{\partial t} = \int_{\Omega} (\mathrm{v}(y,t) - \mathrm{v}(x,t)) K(x,y,t) dy + \mathrm{z}(x,t),$$

其中控制输入为：

$$\mathrm{z}(x,t) = \lambda_P e(x,t) + \lambda_I \int_0^t e(x,\tau) d\tau + \lambda_D \frac{de(x,t)}{dt},$$

且误差项定义为：

$$e(x,t) = f(x) - \mathrm{v}(x,t).$$

PID 控制器包含三部分：- **比例控制 (P)**：直接调整误差。- **积分控制 (I)**：累积历史误差。- **微分控制 (D)**：预测误差变化趋势。

3.1 PID 控制的优化视角

引入 PID 控制器后，状态空间模型隐式最小化了以下泛函：

$$E(\mathrm{v}, f) = J(\mathrm{v}) + G(\mathrm{v}, f),$$

其中：

$$G(\mathrm{v}, f) = \frac{\lambda}{2} \int_{\Omega} \|\mathrm{v}(x) - f(x)\|_2^2 dx.$$

3.2 PID 控制器的稳定性分析

PID 控制器通过调整参数 $\lambda_P, \lambda_I, \lambda_D$，实现对输入扰动的鲁棒性和稳定性。以下引理和命题说明了其行为：

引理 0.2. *设 $B = K - (\lambda_P + 1)I$，则解为：*

$$\mathrm{V}(t) = \exp\left(\frac{1}{1+\lambda_D} Bt\right)\left(\mathrm{V}^0 + B^{-1}\mathrm{F}\right) - \lambda_P B^{-1}\mathrm{F}.$$

命题 0.3. *对于任意正值 $\lambda_P, \lambda_I, \lambda_D$，PID 控制的状态空间模型是稳定的。*

4. PIDformer 的实现

PIDformer 将 PID 控制器集成到 Transformer 的自注意力机制中，其第 ℓ 层的输出计算公式为：

$$u^{\ell}(i) = \sum_{j=1}^{N} \mathrm{softmax}\left(\frac{\mathrm{q}^{\ell}(i)^{\top}\mathrm{k}^{\ell}(j)}{\sqrt{D_{qk}}}\right)\mathrm{v}^{\ell}(j) + \lambda_P e^{\ell}(i) + \lambda_I \sum_{m=1}^{\ell} e^{m}(i) + \lambda_D\left(e^{\ell}(i) - e^{\ell-1}(i)\right),$$

其中误差定义为：

$$e^{\ell}(i) = v^0(i) - v^{\ell}(i).$$

结论

PIDformer 通过引入 PID 控制器，解决了 Transformer 中存在的输入扰动敏感性和秩坍塌问题。其核心思想是将控制理论与深度学习相结合，为模型的鲁棒性提供了全新的设计视角。但在状态空间的建模方面仍然存在一些问题，由于 Transformer 或者其他深度学习模型的结构基本上是通过模仿或者经验的方式设计的，例如 Transformer 的 QKV 来自于数据库系统，而 U-Net 或者深层卷积网络是通过做实验调参得到的，因此模型本身的物理意义并不明确。而所有的深度学习模型的 backbone 都是为了拟合某种函数或者分布，这种拟合过程的稳定性与模型参数的稳定性并不是完全等价的，如何更好地对现有的深度学习模型进行控制理论的建模，仍然是一个极具挑战并且值得研究的问题。

参考文献

[1] Tam Nguyen et al. *PIDformer: Transformer Meets Control Theory*. Feb. 2024. DOI: 10.48550/arXiv.2402.15989. arXiv: 2402.15989. (Visited on 11/11/2024).