Deepseek mHC技术解读

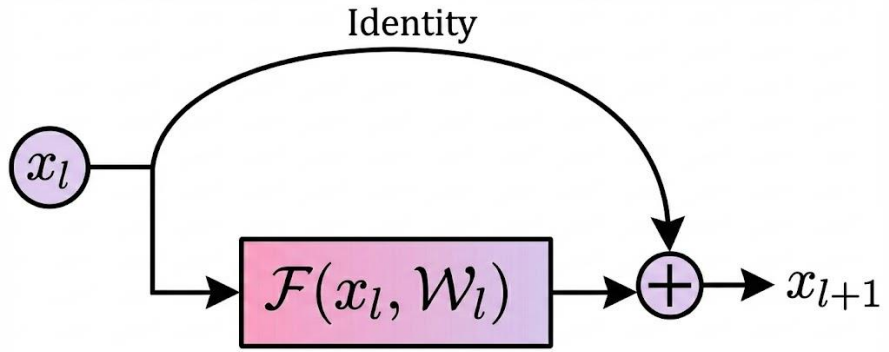
从**传统残差连接**到字节跳动的HC到Deepseek的mHC

东川路第一可爱猫猫虫

主要内容

- 传统残差连接
- 字节跳动的Hyper-Connections
  - 扩展了残差流的宽度
  - 但可能出现连乘爆炸的问题
- Deepseek的mHC
  - Birkhoff多胞形
  - Sinkhorn-Knopp 算法

残差连接



$$x_{l+1} = x_l + \mathcal{F}(x_l, \mathcal{W}_l)$$

- x_l表示第l层的输入， x_{l+1}表示第$l+1$层的输入
- $\mathcal{F}$是层函数，$\mathcal{W}_\ell$是第l层的权重参数
- 我们把残差连接公式递归展开，可以得到：$x_L = x_l + \sum_{i=l}^{L-1} \mathcal{F}(x_i, \mathcal{W}_i)$

where L and l correspond to deeper and shallower layers, respectively. The term identity mapping refers to the component $\mathbf{x}_l$ itself, which emphasizes the property that the signal from the shallower layer maps directly to the deeper layer without any modification.

- 恒等映射

  维持了大规模训练里的稳定性和效率

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

2024 字节跳动提出Hyper-Connections

HYPER-CONNECTIONS

Defa Zhu, Hongzhi Huang, Zihao Huang, Yutao Zeng, Yunyao Mao, Banggu Wu, Qiyang Min, Xun Zhou
Seed-Foundation-Model Team, ByteDance
{zhudefa,huanghongzhi.51,huangzihao.notabot,yutao.zeng,
maoyunyao.myy,wubanggu,minqiyang,zhouxun}@bytedance.com

- Hyper-Connections对传统残差连接做了修改
  - 将残差通道的宽度扩展n倍
  - 并引入可学习的线性变换来管理这些通道

HC扩展了残差流的宽度

- 传统残差连接里

  x_l表示第l层的输入，这是一个d维的列向量
- 我们觉得仅仅用一个向量在深层网络中传递信息，带宽太窄

  希望在这个高速公路上同时并行传输多个版本的信息
- 字节跳动的Hyper-Connections里面

  第l层的输入是一个$n \times d$维的矩阵H_l

  H_l是由n个不同的d维列向量拼接得来的

potential. The single-layer architecture of HC is illustrated in Fig. 1(b). By expanding the width of the residual stream and enhancing connection complexity, HC significantly increases topological

HC

the initial input $\mathbf{h}^0$ to the network. Initially, $\mathbf{h}^0 \in \mathbb{R}^d$ is replicated n times to form the initial *hyper hidden matrix* $\mathbf{H}^0 = \begin{pmatrix} \mathbf{h}^0 & \mathbf{h}^0 & \dots & \mathbf{h}^0 \end{pmatrix}^\intercal \in \mathbb{R}^{n \times d}$. Here, n is the expansion rate. For

- 初始化的时候

  矩阵H（ X_l）由d维向量复制n份而来

  $$H_{l+1} = \mathcal{H}_l^{res} \cdot H_l + (\text{新信息})$$

  $$H_0 = \begin{bmatrix} x_0^T \\ x_0^T \\ \vdots \\ x_0^T \end{bmatrix}$$

- HC里的残差项不再是直接相加

  而是乘上了一个可以动态调整的矩阵

- HC的新信息

  对n个通道的输入进行加权聚合到一个通道通过层函数

  然后再广播回n个通道成为新信息

$$\mathbf{x}_{l+1} = \mathcal{H}_l^{\text{res}}\mathbf{x}_l + \mathcal{H}_l^{\text{post}\,\top}\mathcal{F}(\mathcal{H}_l^{\text{pre}}\mathbf{x}_l, \mathcal{W}_l)$$

- 这里的X_l , X_{l+1}行数为n，列数为隐藏层维度d
- HC里的三个H矩阵 res post pre

  都是可训练的
- 递归展开一下：

$$X_{l+1} = \underbrace{\mathcal{H}_l^{res} X_l}_{\text{线性残差项}} + \underbrace{\mathcal{H}_l^{post}\mathcal{F}(\mathcal{H}_l^{pre} X_l)}_{\text{非线性更新项}} \qquad X_{l+1} = \mathcal{H}_l^{res} X_l + \Phi_l(X_l)$$

$$X_1 = \mathcal{H}_0^{res} X_0 + \Phi_0$$

$$X_2 = \mathcal{H}_1^{res} X_1 + \Phi_1 \qquad X_2 = \mathcal{H}_1^{res}(\mathcal{H}_0^{res} X_0 + \Phi_0) + \Phi_1$$

$$X_2 = (\mathcal{H}_1^{res} \cdot \mathcal{H}_0^{res}) X_0 + (\mathcal{H}_1^{res}\Phi_0 + \Phi_1)$$

$$X_3 = (\mathcal{H}_2^{res} \cdot \mathcal{H}_1^{res} \cdot \mathcal{H}_0^{res}) X_0 + \dots$$

传统残差连接与HC的区别：累加与累乘

$$x_L = x_l + \sum_{i=l}^{L-1} \mathcal{F}(x_i, \mathcal{W}_i)$$

$$X_L = \left(\prod_{k=l}^{L-1} \mathcal{H}_k^{res} \right) X_l + \sum_{k=l}^{L-1} \left(\prod_{j=k+1}^{L-1} \mathcal{H}_j^{res} \right) \Phi_k$$

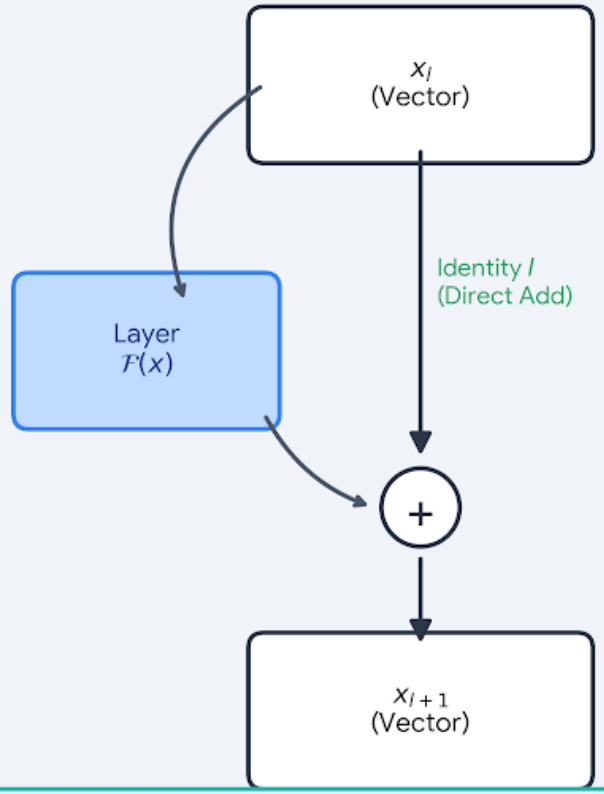
- 传统残差连接里

  恒等映射
- Hyper-Connections

  复合映射

Standard Residual Connection

(ResNet / Transformer)

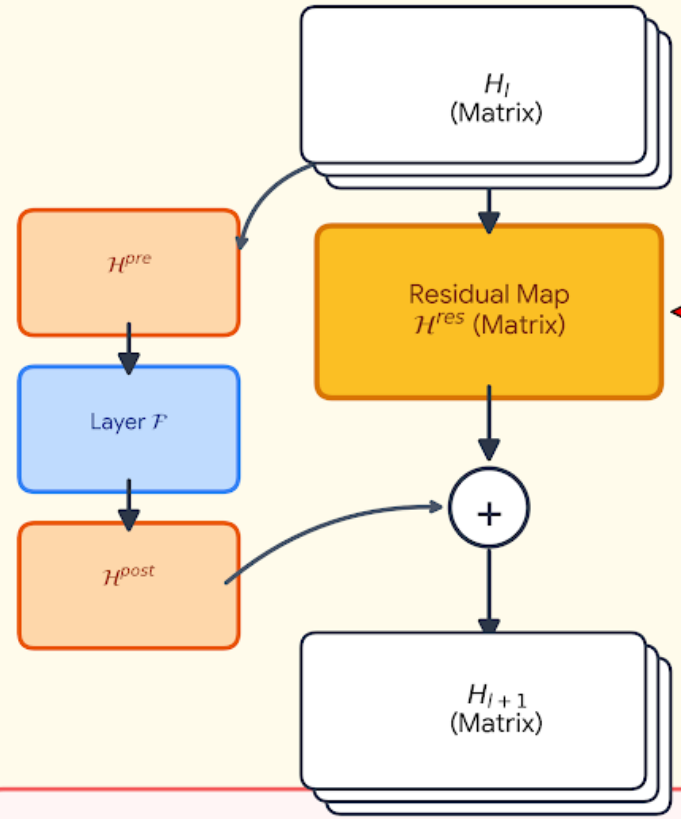


Mathematical Nature: Summation

$$x_{final} = x_{input} + \sum \mathcal{F}(x_i)$$

Hyper-Connections (HC)

(Matrix Multiplication Chain)



Mathematical Nature: Product Chain

$$H_{final} = (\prod \mathcal{H}_i^{res}) H_{input} + \ \dots$$

the composite mapping $\prod_{i=1}^{L-l} \mathcal{H}_{L-i}^{\text{res}}$ in HC fails to preserve the global mean of the features. This discrepancy leads to unbounded signal amplification or attenuation, resulting in instability during large-scale training. A further consideration is that, while HC preserves computational

- deepseek认为这样的复合映射不好

  可能会导致无界的信号放大和衰减
- 无约束的矩阵连乘会闯祸

  需要对$\mathcal{H}^{res}$加上约束

  让他守规矩，避免连乘爆炸

Birkhoff多胞形

$$\mathcal{P}_{\mathcal{M}^{res}}(\mathcal{H}_l^{res}) := \{\mathcal{H} \in \mathbb{R}^{n\times n} \mid \mathcal{H}\mathbf{1}_n = \mathbf{1}_n, \mathbf{1}_n^\top \mathcal{H} = \mathbf{1}_n^\top, \mathcal{H} \geq 0\}$$

- 每一行加起来是1
- 每一列加起来是1
- 元素均非负

we restrict $\mathcal{H}_l^{\text{res}}$ to be a doubly stochastic matrix, which has non-negative entries where both the rows and columns sum to 1. Formally, let $\mathcal{M}^{\text{res}}$ denote the manifold of doubly stochastic

- 把原本无约束的$\mathcal{H}^{res}$强行投影到Birkhoff多胞形

  这样的数学限制就解决了连乘爆炸的问题

为什么是Birkhoff多胞形?

Norm Preservation: The spectral norm of a doubly stochastic matrix is bounded by 1 (i.e., $\|\mathcal{H}_l^{\text{res}}\|_2 \leq 1$). This implies that the learnable mapping is non-expansive, effectively mitigating the gradient explosion problem.

Compositional Closure: The set of doubly stochastic matrices is closed under matrix multiplication. This ensures that the composite residual mapping across multiple layers, $\prod_{i=1}^{L-l} \mathcal{H}_{L-i}^{\text{res}}$, remains doubly stochastic, thereby preserving stability throughout the entire depth of the model.

- Birkhoff多胞形谱范数不大于1

  信号不会无限放大

- Birkhoff多胞形对矩阵乘法封闭

  连乘后的矩阵依旧在我们的安全区里

为什么是Birkhoff多胞形?

Geometric Interpretation via the Birkhoff Polytope: The set $\mathcal{M}^{\text{res}}$ forms the Birkhoff polytope, which is the convex hull of the set of permutation matrices. This provides a clear geometric interpretation: the residual mapping acts as a convex combination of permutations. Mathematically, the repeated application of such matrices tends to increase the mixing of information across streams monotonically, effectively functioning as a robust feature fusion mechanism.

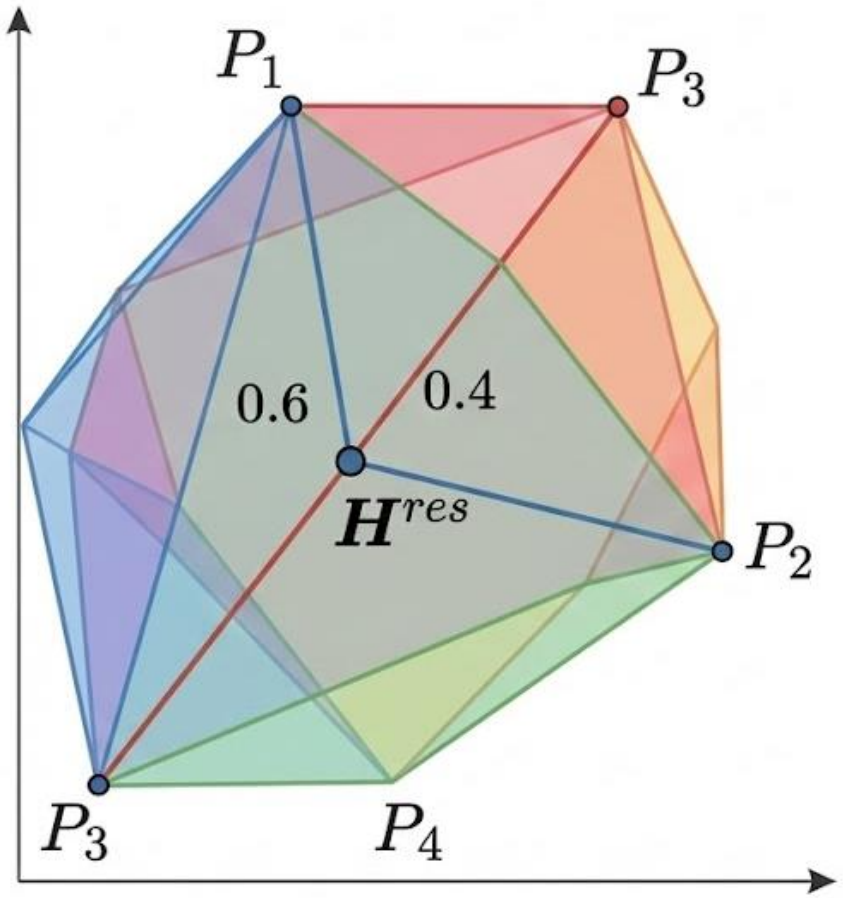
- Birkhoff 多胞形是置换矩阵的凸包

  左乘置换矩阵就是只做行与行的交换，不融合

  Birkhoff 多胞形里的$\mathcal{H}^{res}$就是多个置换矩阵的加权平均

  这意味着$\mathcal{H}^{res}$是在做一种能量守恒的特征融合

能量守恒的特征融合



Birkhoff多胞形是置换矩阵的凸包。

$$\mathcal{H}^{res}\text{Input} = (0.6P_1 + 0.4P_2)\text{Input}$$

$$= 0.6 \begin{bmatrix} \mathbf{D} \\ \mathbf{A} \\ \mathbf{B} \\ \mathbf{C} \end{bmatrix} + 0.4 \begin{bmatrix} \mathbf{B} \\ \mathbf{C} \\ \mathbf{D} \\ \mathbf{A} \end{bmatrix}$$

凸组合（加权平均）：H^{res} 是 P_1 和 P_2 的加权和。

$$P_1 \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \\ \mathbf{C} \\ \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{D} \\ \mathbf{A} \\ \mathbf{B} \\ \mathbf{C} \end{bmatrix}$$

置换矩阵 P_1 重新排列输入向量的元素。

Sinkhorn-Knopp 算法

where $\sigma(\cdot)$ denotes the Sigmoid function. The Sinkhorn-Knopp$(\cdot)$ operator firstly makes all elements to be positive via an exponent operator and then conducts iterative normalization process that alternately rescales rows and columns to sum to 1. Specifically, given a positive

- 神经网络可能输出负数

  Birkhoff多胞形的硬性要求是非负

  因此Sinkhorn-Knopp首先通过指数运算**把所有元素变为正值**

- 然后进行迭代归一化

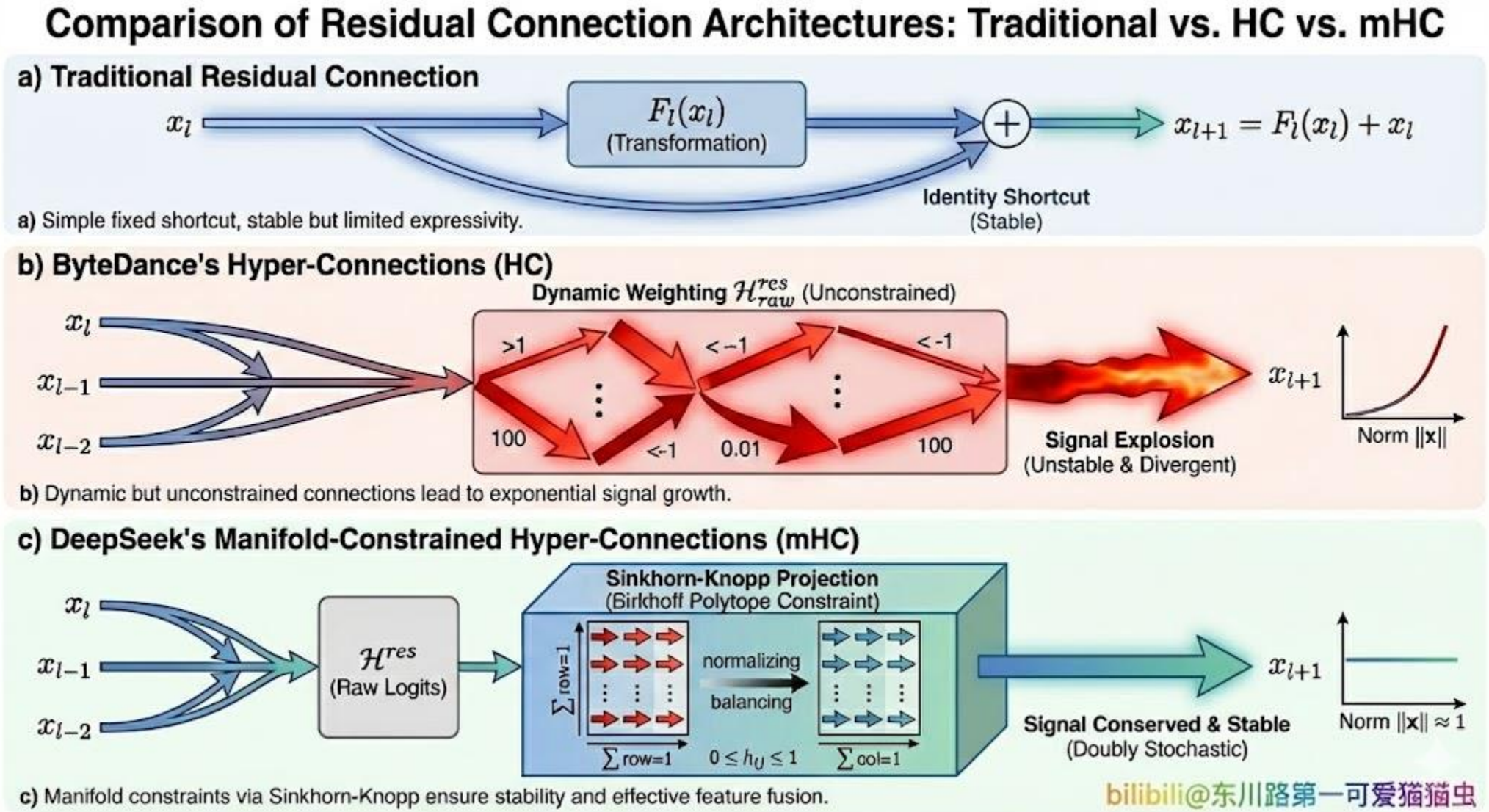
  交替缩放行和列

  使其和为1

$$\mathbf{M}^{(t)} = \mathcal{T}_r\left(\mathcal{T}_c(\mathbf{M}^{(t-1)})\right)$$

where $\mathcal{T}_r$ and $\mathcal{T}_c$ denote row and column normalization, respectively. This process converges to a doubly stochastic matrix $\mathcal{H}_l^{\text{res}} = \mathbf{M}^{(t_{\max})}$ as $t_{\max} \to \infty$. We choose $t_{\max} = 20$ as a practical value in our experiments.

对比总结


Comparison of Residual Connection Architectures: Traditional vs. HC vs. mHC
a) Traditional Residual Connection
x_l
$F_l(x_l)$
(Transformation)
$x_{l+1} = F_l(x_l) + x_l$
Identity Shortcut
(Stable)
a) Simple fixed shortcut, stable but limited expressivity.
b) ByteDance's Hyper-Connections (HC)
Dynamic Weighting $\mathcal{H}^{res}_{raw}$ (Unconstrained)
x_l
x_{l-1}
x_{l-2}
>1
100
<-1
<-1
0.01
<-1
100
Signal Explosion
(Unstable & Divergent)
x_{l+1}
Norm ||x||
b) Dynamic but unconstrained connections lead to exponential signal growth.
c) DeepSeek's Manifold-Constrained Hyper-Connections (mHC)
x_l
x_{l-1}
x_{l-2}
$\mathcal{H}^{res}$
(Raw Logits)
Sinkhorn-Knopp Projection
(Birkhoff Polytope Constraint)
$\sum$ row=1
normalizing
balancing
$\sum$ row=1
$0 \le h_{ij} \le 1$
$\sum$ col=1
Signal Conserved & Stable
(Doubly Stochastic)
x_{l+1}
Norm ||x|| ≈ 1
c) Manifold constraints via Sinkhorn-Knopp ensure stability and effective feature fusion.
bilibili@东川路第一可爱猫猫虫