专题综述

大模型架构的下半场

华中科技大学 王兴刚 朱良辉

太长不看：我们花了十年去扩展层内的计算能力，却忘了扩展层间的通信能力。这件事亟需被改变。

过去十年，深度学习领域取得进展的方式出奇地一致：什么都往大了整。更多参数、更多数据、更长上下文；而且确实管用：loss 在降，能力在涨，scaling law（扩展定律）精确地告诉我们还需要投入多少。

但扩展的方向不同，差异也是巨大的。序列长度的扩展需要真正的创新，也确实催生了一整套机制研究和系统工程。数据的扩展则直截了当：数据越多，loss 越低。让模型变得更宽、更深，这看起来也和数据的扩展一样简单。但宽度和深度真的在同等地发挥作用吗？

并非如此。深度在数量上增长了，但在质量上却没有。层与层之间的通信机制几乎没有变化。接下来我们将解释这一点为什么重要，这不仅关乎网络的深度本身，更关乎我们设计神经网络架构时的一个集体盲区。

一、大模型架构上半场

要看清上半场做对了什么，就看看什么被成功地扩展了，以及是怎么做到的。

先看序列长度。早期 Transformer 只能处理几百个 token。要达到 128K+，需要多个方向上的持续创新：新的注意力模式（稀疏、线性、混合）、系统工程（FlashAttention）、位置编码的进步（RoPE scaling）。研究者和工程师们共同建造了一整个生态，持续改进 token 之间的通信方式。而回报颇丰，我们不止能够处理极其长的文档，还为 OpenAI-o1 和 DeepSeek-R1 的长链推理奠定了坚实的基础。这就是当我们认真投资于“信息在序列维度上的流动方式时”，所收获的斐然成果。

参数和数据的扩展是最符合人类直觉的部分。从深度学习的最早期开始，每本教科书都在教我们同一套配

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or “GPT-3”	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

GPT-3, 175B Params, over 300B Tokens

DeepSeekV3, 671B Params, 14.8T Tokens

GLM-5, 744B Params, 28.5T Tokens

Qwen3, 1T Params, 36T Tokens

Kimi K2, 1T Params, 15.5T Tokens

图 1　现代 LLM 中的参数与数据规模迅速增长

方：更多数据、更宽的层、更深的网络，自然带来更好的表征。从 GPT-2 的 15 亿参数到如今的数千亿，这套配方一直管用。我们不需要新机制，只需要持续拓展这些被验证了的方向。

只不过，对网络而言，更宽和更深往往并不是一回事。宽度的扩展是自然而然的：现代 GPU 天生擅长处理更宽的矩阵乘法，注意力机制的演进越来越高效，这使得更宽的网络可以无缝接入现有架构。

而深度则是另一个故事。模型确实变深了：我们将模型加到 32 层、64 层、甚至 100 层以上。但层间通信的机制本质上还是 ResNet 在 2015 年引入的深度残差“x + F(x)”。自它诞生以来，围绕它有过不少改良（归一化位置、残差缩放、跨层连接），但没有任何改良真正取代过那个深度残差中“+”的决定性地位。

残差连接可以说是深度学习中最重要的基石。没有它，就没有 100 层的 Transformer，没有现代 LLM，没有 scaling law。但基础性方案有一个特点：它们有时会变得太过隐形，以至于没人再去质疑它到底是最优解，还是仅仅是我们探索出的第一个能用的方案。

打个比方，想象一个有特殊规则的传话游戏。在标准版本里，第 1 个人对第 2 个人耳语，第 2 个人再对第 3 个人耳语。到第 18 个人的时候，消息已经面目全非了。这就是没有残差连接的深层网络：每一层只能看到上一层的输出。

残差连接修复了这个问题：每个人在传达自己的理解的同时，也把之前积累的原始信息原封不动地往下传。第 3 个人既能听到第 2 个人的新解读，也能听到之前的所有内容。原始信号始终被保留，它成为了不断壮大的合唱中的一个声部。

但到了第 152 个人，你同时在听 152 个声音：原始信息加上 151 层叠加上去的内容，全部混在一句耳语里。理论上，前面那些人的声音依然存在，但它们已经被淹没了。如果第 152 个人需要知道第 3 个人具体说了什么，他得费力地从这首宏大的合唱声中把它挑出来。通常而言，第 152 个人是做不到这一点的。

这就是信息稀释。每一层都面临两难：倘若该层贡献新信息就可能会掩盖之前的内容，但保守不动则能保留之前层传过来的已有信息。这种状况下，很多层学会了保守不动，它们几乎不往残差流里写入任何东西。这样的深度网络在纸面上很深，实际上却很浅。我们堆了 152 层，但其中很多层却只学会了保持沉默。

这里的瓶颈不在于 152 层网络所需求的算力，而在于信息穿过这些层的通信能力。CPU 的发展在几十年前就撞过同样的墙：处理器越来越快，直到内存带宽跟不上了，逼得整个行业转向缓存和通信。组织管理也一样：一群聪明人所能发挥出的创造力，也受限于他们之间的沟通、组织方式。深度学习正在经历自己的版本：十年来不断增强每一层的能力，而层与层之间的通道始终是 2015 年那条单车道公路。

那么，有没有更好的机制？

二、大模型架构配方

在我之前已经有很多研究者注意到了深度学习的瓶颈。多年来，修补方案越来越巧妙：获评 CVPR best paper 的 DenseNet 保留了每一层的输出，但代价是平方级的开销。使用可学习加权的方案 DenseFormer、LIMe 降低了成本，但训练完成后权重就固定了，每个 token、每套上下文都用同样的权重。字节跳动的 Hyper-Connections 和 DeepSeek 的 mHC 另辟蹊径，它们把管道拓宽到 N 个通道，层间用混合矩阵连接，这相当于信息高速公路上同时多了好几条车道。但坏消息是，信息仍然在逐层流动，第 152 层没有办法直接回溯到第 3 层。

彩云公司的 MUDDFormer 让混合每层输出这件事变成动态的，它会根据每个 token 的表征来生成权重。这在根本方向上是对的：从每一层汲取多少信息本就应该取决于你正在处理的内容。但同样有个坏消息，第 152 层在决定从第 3 层汲取多少时，只依赖第 152 层本身的状态，它并不知道第 3 层实际包含了什么。它是在预测哪些层有用，而不是在查看。

以上的每一步都修复了一个真实存在的缺陷，但却鲜有哪一个方法质疑过深度残差的框架本身。

不难发现这些方法都有一个共同点。从 DenseNet

到 Hyper-Connections，每个方法都在回答同一个隐含的问题：“怎样才能更好地混合各层的输出？”更好的系数，更多的通道，自适应的权重。但自始至终都是混合，自始至终都是累加。ELMo 早就表明，不同的层编码的是截然不同的信息：浅层编码句法，深层编码语义。所有人得出的结论都是“学习更好的混合权重用来平衡句法和语义”。但还有一条被主流忽视的道路：如果不同层持有不同信息，也许每一层应该能够根据内容而非位置，从持有所需信息的那一层直接检索。

这就是范畴谬误：把层间通信当作累加（用学习到的或生成的系数来组合信号）而非检索（通过基于内容的匹配来选择信息）。在累加框架下，即使是动态方法也只从当前层的状态生成混合权重，而不去查看信息的来源层实际包含了什么。在检索框架下，Query（查询）编码的是“我需要什么”，Key（键）编码的是“我有什么”，而它们之间的运算决定了相关性。Query 和 Key 双方都应该有发言权。

合优化策略，实现几何与外观质量的协同提升。

回到传话游戏。之前所有的方法都在试图产生一个更清晰的合唱：更好的发音、更多的中继通道、自适应的音量。没有一个质疑过这个根本约束：所有声音必须累加成一个声音吗？也没有人问过：咱是否可以直接走回去，跟之前的任何一个人当面对话呢？

我觉得这种范畴谬误在架构设计中无处不在。当某个东西足够好用的时候，你不会去质疑它的概念框架，而是只会在框架内改进。经历了多年越来越巧妙的修补之后，我才看清：深度维度的残差连接需要的不是更好的系数，而是被一种根本不同的操作所替代：一种在序列维度上已经成功解决了同样问题的操作。

三、大模型架构下半场

一旦我们把层间的通信理解为检索而非累加，一个很自然的答案就是在深度维度上引入注意力机制。很多团队都独立地收敛到了这个想法：谷歌提出的 DCA、华为的 MRLA、Hessian.AI 的 Dreamer、Kimi 的 AttnRes，大家都尝试在层间应用注意力机制。这种独立趋同本身就是一个信号：方向走对了！

但找对方向和做出成品是两回事。说实话，我第一次用 Pytorch 实现运行深度注意力的时候，前向和反向传播共计耗时达到了 44,924 ms。44 秒啊！朋友们！这个时间都够我喝完一瓶 500 毫升的冰红茶了！也就是说，在深度维度上应用注意力机制的想法本身没问题，但工程现实却残酷到了极点。现代 GPU 为大规模的矩阵乘法做了大量优化，却不擅长数千个跨深度的极小规模的注意力操作。深度注意力作为一个计算量不大的算法，跑起来却可能慢得要命。

至此，之前的方法都陷入了两难：要么简化深度注意力来换速度，这种方式丢掉了完整的选择性检索这一核心价值；要么保持完整的表达能力，但运算代价变得不可接受。我们找到了一条出路：不是简化算法，而是重新组织参与计算的数据布局，从而适配 GPU 硬件。如图 2，Flash Depth Attention（https://github.com/hustvl/MoDA）让具备完整表达能力的深度检索快到可以参与实际训练。

有了高效的深度检索之后，我们注意到网络的主干流水线变成了：深度注意力→序列注意力→深度注意力→ FFN（前馈网络）。这三个连续的注意力操作作用于不同的 Key（键，缩写作 K） 和 Value（值，缩写作

```
================================================================================
Comprehensive Benchmark [fwd_bwd]  B=1 H=2 K=64 V=64 L=64 dtype=torch.float16
================================================================================
    T_kv     G        T_q |         DepthRef |            FDAv4 |   DepthRef/FDAv4
--------------------------------------------------------------------------------
     512     8       4096 |      4341.213 ms |         0.686 ms |      6328.299x
    1024     8       8192 |      9560.738 ms |         0.674 ms |     14185.071x
    4096     8      32768 |     44924.228 ms |         0.948 ms |     47388.426x

     512    16       8192 |      9199.153 ms |         0.639 ms |     14396.171x
    1024    16      16384 |     19010.542 ms |         0.869 ms |     21876.343x
```

图 2　Pytorch 实现的深度注意力（DepthRef）很慢；Flash Depth Attention（FDA）很快

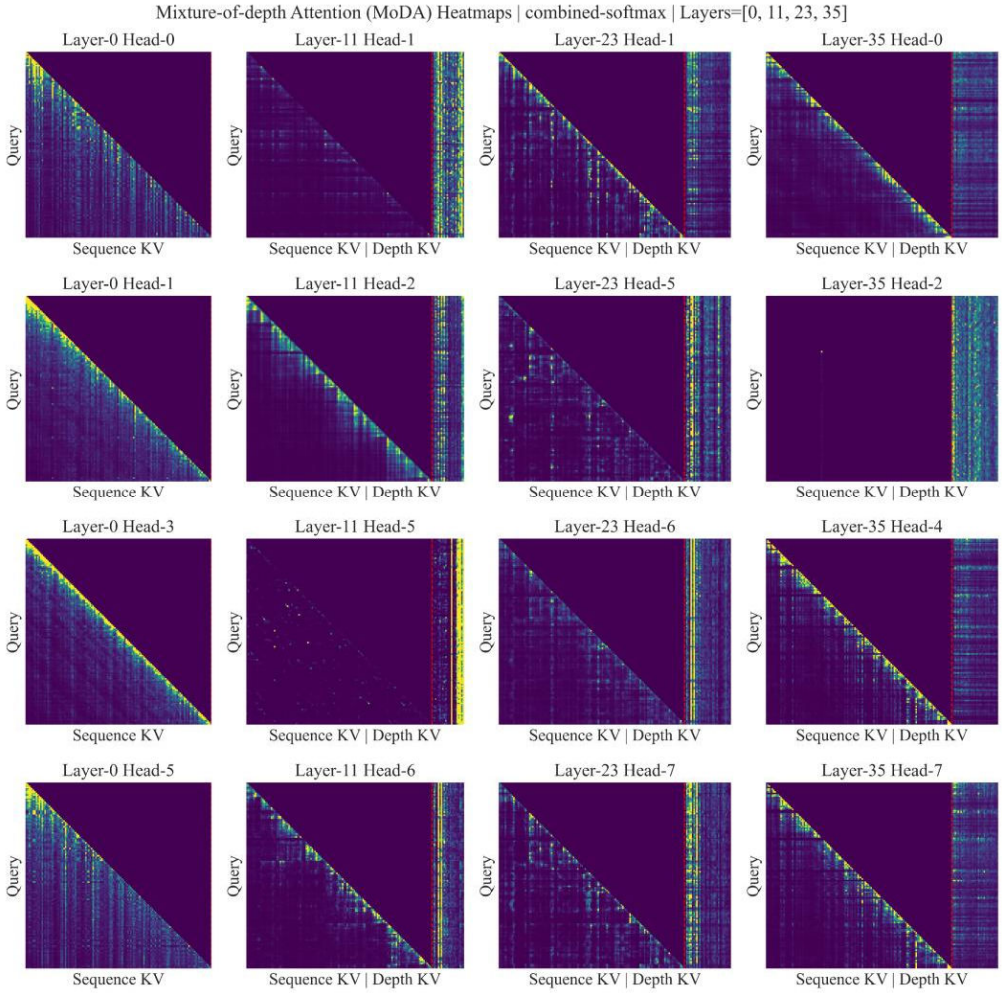


图 3　左侧区域是序列 KV，右侧区域是深度 KV。颜色越黄，注意力越强。

V），却共享着近乎相同的 Query（查询）。一个很自然的做法就是把它们融合。我们提出了混合深度注意力（Mixture-of-depths Attention, MoDA）将深度检索和序列检索合并到一个统一的 softmax 中。如图 3，每个注意力头同时关注当前层的序列 KV 对（键值对）和所有前序层的深度 KV 对（键值对）。在同一个 softmax 下，模型可以自由决定何时关注序列中的其他 token，何时跨层检索自身的历史信息。通过一次操作，我们完成了两个维度的检索。

回到传话游戏。在残差连接的版本里，第 152 个人费力地从累加的合唱中辨认第 3 个人的声音。有了深度检索，第 152 个人拍拍第 3 个人的肩膀直接问："你刚才说了什么？"没有中间人，没有累积的噪音。可视化的实验结果也印证了这个类比所预测的现象：当模型获得了通过深度 KV 从特定层进行选择性检索的能力时，它会持续且主动地使用这种能力。之前困扰模型架构研究员们的 Attention Sink（注意力沉没）现象，即模型把概率质量堆积在少数固定 token 上的行为，也随之减弱。这就是当我们尝试发展层之间而非仅仅层之内的信息流动时，所取得的有趣成果。

大模型架构的上半场是关于扩展组件的。我们扩展出更长的序列，更多的数据，更大的模型。这个阶段最关键的问题是"我们怎么把一切都做大？"。在上半场，这是正确且关键的问题，它把我们从 GPT-2 带到了 GPT-4。下半场是关于扩展通信的。新的问题是："组件之间的通信质量如何？"

深度是最明显的例子，因为现有方案（累加）和可能的方案（选择性检索）之间的差距是巨大的。而我相信这个原则是可以推广的。凡是神经网络使用静态的、与数据无关的通道来传递信息的地方，包括层与层之间、

模态与模态之间、时间步与时间步之间等等，很可能都会有一个检索机制等着替代那个累加操作。

我们花了十年掌握 token 之间如何对话，现在是时候掌握层与层之间如何对话了。而最终，我们将掌握神经网络中每个组件如何与其他任意组件对话。深度残差的“+”带我们跑过了一段极为精彩的旅程，但现在，是时候升级这座阶梯了。

欢迎来到大模型架构的下半场。

本论文内容主要来自于华中科技大学(HUST)电子信息与通信学院视觉实验室(Vision Lab)的研究工作。HUST Vision Lab研究主要集中在计算机视觉和深度学习领域，尤其关注以下方向：多模态基础模型，视觉表征学习，目标检测、分割与跟踪，端到端自动驾驶，和新型神经网络架构。

HUST Vision Lab突破视觉智能边界的代表性工作：CCNet（TPAMI 2020，4300+引用，1.5K Star）、Mask Scoring R-CNN（CVPR 2019，1400+引用，1.9K Star）、FairMOT（IJCV 2021，2200+引用，4.2K Star）、ByteTrack（ECCV 2022，3400+引用，6.2K Star）、EVA（CVPR 2023，1100+引用，2.7K Star）、MapTR（ICLR 2023，400+引用，1.5K Star）、Vectorized Autonomous Driving（ICCV 2023，600+引用，1.3K Star）、DiffusionDrive（CVPR 2025，200+引用，1.3K Star）、Vision Mamba（ICML 2024，3100+引用，3.8K Star）、4D Gaussian Splatting (4DGS)（CVPR 2024，1400+引用，3.5K Star）、YOLOS（NeurIPS 2021，500+引用，900+ Star）、YOLO-World（CVPR 2024，1000+引用，6.3K Star），及 LightningDiT & VA-VAE（CVPR 2025，200+引用，1.4K Star）。

责任编委 王金甲

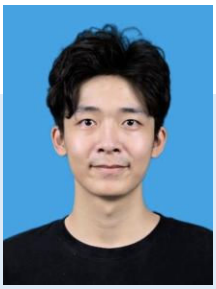
参考文献

[1] https://arxiv.org/abs/2603.15619.
[2] https://github.com/hustvl/MoDA.
[3] https://github.com/hustvl.

王兴刚

华中科技大学电子信息与通信学院教授，博士生导师，国家“万人计划”青年拔尖人才，现任 Image and Vision Computing 期刊（Elsevier, IF 4.2）主编，分别于 2009 年和 2014 年在华中科技大学获得了本科和博士学位，期间在美国 UCLA 和天普大学访问研究。主要研究方向为计算机视觉、深度学习、多模态基础模型、视觉表征学习、目标检测分割跟踪、自动驾驶等，发表包括 IEEE TPAMI、IJCV、CVPR、ICCV、NeurIPS 等顶级期刊会议在内的论文 100 余篇，谷歌学术引用 5 万余次，其中一作/通讯 1000+引用论文 6 篇，入选 Elsevier 中国高被引学者、斯坦福前 2%科学家。研究成果应用于地平线征程系列芯片、华为海思昇腾芯片、Deepmind 蛋白质结构预测模型 AlphaFold；获得了华为、vivo 等公司优秀合作项目奖。担任 CVPR、ICCV、NeurIPS 等顶级会议领域主席，IEEE TPAMI、Machine Vision and Application 等期刊编委。入选了中国科协青年人才托举工程，获湖北青年五四奖章、CSIG 青年科学家奖、吴文俊人工智能优秀青年奖、CVMJ 2021 年度最佳论文奖、MIR 期刊年度最高引用论文奖、湖北省自然科学二等奖等。

Email: xgwang@hust.edu.cn



朱良辉

华中科技大学电子信息与通信学院博士研究生，中共党员，师从王兴刚教授和刘文予教授，入选首届字节跳动校企联培博士。主要研究方向为基础模型架构、弱监督视觉感知和视觉语言大模型等，已发表 ICML、ICLR、CVPR、ICCV、AAAI、ACM MM、IJCV、TIP 等顶级期刊会议论文 11 篇，其中一作 8 篇，谷歌学术引用 4500 余次，开源代码获 GitHub Stars 4600+。代表作 Vision Mamba 提出首个基于 Mamba 的线性复杂度通用视觉骨干网络，入选 ICML 2024 最具影响力论文和 ArXiv 2024 计算机视觉方向最具影响力论文。主持国家自然科学基金青年学生基础研究项目和中国科协青年人才托举工程博士生专项，获博士研究生国家奖学金、华中科技大学“学术新星”、ICLR Notable Reviewer 等荣誉。

Email: lhzh@hust.edu.cn