东吴证券 SOOCHOW SECURITIES

汽车行业点评报告

AI 系列深度 06 期: Transformer 如何从序列任务走向基座模型?

2026 年 07 月 30 日

证券分析师 黄细里
执业证书: S0600520010001
021-60199793
huangxl@dwzq.com.cn

增持（维持）

投资要点

- **RNN 与 Transformer 是序列建模的两代核心架构，解决文字、语音、视频和轨迹等有先后顺序的数据如何形成表征的问题。**序列建模的难点在于变长、有序和长程依赖：模型既要保留上下文，又要把相距较远的相关信息联系起来。
- **RNN 的基本思路是把历史压缩进一个随时间更新的隐藏状态。**该结构推理成本相对恒定，适合早期机器翻译、语音识别和时间序列任务；LSTM、GRU 通过门控机制缓解长程遗忘，Seq2Seq 和注意力机制进一步推动模型更迭。但 RNN 必须逐步串行处理，难以充分利用并行算力，长距离信息也容易在压缩状态中衰减。
- **Transformer 采取不同路径:** 不再把历史压缩成单一状态，而是用自注意力让每个位置直接与所有位置交互。多头注意力、位置编码和前馈层共同构成其结构基础，使训练可以并行、长程依赖路径显著缩短，并为模型规模化扩展提供架构基础。GPT、BERT 以及后续 LLM、VLM、VLA 大多以 Transformer 为核心底座，反映其在云端通用基础模型中的主导地位。
- **两者分歧可归纳为记忆、计算和长程建模三组权衡:** RNN 以定长状态换取相对恒定的推理成本，但训练串行、记忆容量受限；Transformer 全量缓存历史，训练并行、长程建模更强，但注意力计算随序列长度平方增长，超长上下文推理成本较高。该约束也是 Mamba、RWKV、线性注意力和混合架构重新受到关注的重要原因。
- **Transformer 替代传统 RNN 主干，并不意味着循环思想消失。**2023 年以来，状态空间模型、选择性状态更新和线性注意力在长上下文、端侧低功耗、低时延场景中体现价值；同时，存在混合架构的技术路径，用少数注意力层保留精确回忆能力，用多数高效层降低计算成本。整体看，Transformer 仍是通用基础模型主线，新循环更像特定约束下的重要补充。
- **我们认为,数字 AI 底座模型发展路径已逐渐清晰、价值在于确定性，但其难以实现物理世界的建模闭环；我们认为物理 AI 将成为物理世界的 AI 解决方案，从而在 AI 的当前发展阶段成为增量更大的方向，智能驾驶作为最先跑通闭环的代表场景，我们认为应重点关注。**
- **风险提示:** 技术迭代不及预期的风险，大模型厂商 ARR 增速放缓的风险，云服务商资本开支不及预期的风险，智能驾驶及机器人商业化落地不及预期的风险。

行业走势

相关研究

《AI 系列深度 05 期：模型如何学习表征并实现对齐？》

2026-07-29

《AI 系列深度 04 期：什么是表征？》

2026-07-29

内容目录

1. RNN 与 Transformer：序列建模的两代答案

RNN 与 Transformer 是序列建模从状态压缩走向全局交互的两代核心架构。CNN/ViT 主要解决图像表征问题，RNN 与 Transformer 则面向文字、语音、视频、车辆轨迹等有顺序的数据，核心难点在于上下文保留和长程依赖建模。我们认为，Transformer 已在云端通用基础模型中显著替代传统 RNN 主干，并成为 GPT、BERT 以及多数 LLM、VLM、VLA 的核心底座；原因在于 RNN 依赖逐步递推和状态压缩，训练难以并行，长程信息容易衰减，而 Transformer 通过自注意力实现全局交互和高并行训练，能够承接更大规模数据与算力。但循环、状态压缩与递推记忆并未消失，正在长上下文、端侧和低时延场景中以新形式回归。

当下大模型、智能体与具身智能的能力基础，均离不开序列建模网络。理解人工智能为何在 2017 年后加速演进，需要比较循环神经网络与 Transformer 两类架构的替代关系。

Transformer 是当今云端通用基础模型的核心底座。生成式预训练模型，名称中的“T”即 Transformer；从大语言模型（LLM）到视觉语言模型（VLM），再到机器人领域的视觉语言动作模型（VLA），主流云端基础模型多以 Transformer 为核心骨架。但在车端、端侧和低时延系统中，混合架构、状态空间模型或压缩机制仍具备应用空间。厘清 Transformer，有助于理解序列模型为何能够高效并行训练，并将模型和数据规模推向新的量级。

传统 RNN 主干已被显著替代，但对照 RNN 才能看清本轮跃迁。RNN 是 2010 年代序列建模的主力，机器翻译、语音识别和早期对话系统多以其为核心。Transformer 相对 RNN 的关键变化，是从逐步压缩历史转向全局交互与并行训练。

但注意力机制的计算成本随序列长度呈平方增长，在超长上下文与端侧推理场景中代价高昂。2023 年以来，以状态空间模型 Mamba、RWKV 和线性注意力为代表的新循环路线重新活跃，循环思想以新的形态回归。

2. 共同问题：序列如何保留上下文

RNN 与 Transformer 的结构差异较大，但二者回答的是同一问题：模型如何处理一段有先后顺序、长度可变且前后相互依赖的数据。理解这一共同问题，是理解两类架构分歧的前提。

2.1. 序列数据的本质：变长、有序、长程依赖

序列建模指对按顺序排列且长度不固定的元素进行建模，用于预测、生成或理解整段内容。文本、语音、视频和车辆轨迹均属于序列数据，其共同特征是存在明确先后关

系，且前后元素相互影响。

顺序携带信息：序列的顺序本身即为信息。“狗咬人”与“人咬狗”用词相同、顺序不同，含义截然相反。模型必须对位置与先后敏感，而不能把输入当作无序集合。

长程依赖是核心难点：序列建模最难的地方在于长程依赖：决定当前输出的关键信息可能出现在很久以前。例如“小张昨天把那本从图书馆借来的、封面已经磨损的书还了，它……”中的“它”指代哪本书，须回看一长串修饰之前的名词。模型能否跨越很长的间隔把相关信息联系起来，直接决定其能力上限。

可用的序列模型还需满足两类工程诉求：训练阶段能否并行，决定其是否能够在大规模数据上高效训练；推理阶段成本是否可控，决定其能否低成本部署并处理长输入。即长程记忆、训练并行和推理成本之间存在权衡。

2.2. 两类核心机制：状态压缩与全量检索

RNN 路线将历史压缩为定长状态。模型在每个时间步将当前输入与上一时间步的隐藏状态共同更新为新的隐藏状态，用该状态概括截至当前的历史信息。该设计的优势是无论序列多长，状态维度保持不变；约束也来自这一点，历史信息越多，早期细节越可能在连续压缩中衰减。

Transformer 路线保留历史并按需检索。模型不把全部历史压缩为单一状态，而是保留各位置表示，并通过注意力机制让当前位置与历史位置直接交互。该方式降低了长程依赖路径长度，提升了信息保留能力，但也需要存储和计算更长的上下文。

3. RNN 机制：把历史压缩为隐藏状态

循环神经网络是序列建模 AI1.0 时代的主流架构。其核心思路是用一个随时间不断更新的状态概括历史。

3.1. 核心机制：循环、隐状态与时间展开

循环与隐状态构成 RNN 的基本机制。RNN 按时间顺序处理序列元素，每一步将当前输入与上一隐藏状态输入同一组网络，得到新的隐藏状态和输出。隐藏状态承担历史信息载体的角色，通过逐步更新把上下文传递到后续时间步。

时间展开与参数共享是 RNN 适配变长序列的基础。将循环沿时间轴展开后，RNN 等价于一条按时间连接的深层网络，每个时间步使用同一组参数。参数共享使其能够处理不同长度的序列。

RNN 通过沿时间反向传播完成训练。该方法将展开后的时间链条视为深层网络进行误差回传；链条越长，梯度传播越困难，也为长程依赖建模带来约束。

3.2. 门控的引入：LSTM 与 GRU 如何缓解长程遗忘

早期 RNN 难以学习长程依赖，核心原因是梯度消失和梯度爆炸。训练误差需要沿时间链条逐步回传，距离越远，信号可能快速衰减或放大，导致模型难以保留较早时间步的信息。Bengio 等和 Hochreiter 的研究均指出了这一问题。

LSTM 通过门控机制缓解长程遗忘。长短期记忆网络（LSTM）引入贯穿时间的记忆单元，并通过输入门、遗忘门和输出门控制信息写入、保留和输出。该设计提升了长期信息保留能力，成为 2010 年代序列建模的重要标准结构。

GRU：更精简的门控：门控循环单元（GRU）把门简化为更新门与重置门两个，参数更少、计算更省，在许多任务上与 LSTM 相当，是常用的轻量替代。

Seq2Seq 与注意力机制推动 RNN 应用扩展。2014 年的序列到序列框架（Seq2Seq），Sutskever 等，用一个 RNN 编码输入序列，再用另一个 RNN 解码输出，支撑了早期神经机器翻译。但将整句压缩为单一向量会形成信息瓶颈，Bahdanau 等在 2014 年提出注意力机制，使解码端能够回看输入各位置，这也成为 Transformer 的重要思想来源。

3.3. 固有约束：串行、难并行与长程上限

串行计算是 RNN 被替代的关键工程约束。RNN 第 t 步依赖第 t-1 步输出，因此必须按时间顺序逐步计算，难以并行处理同一序列中的不同位置。这限制了其对 GPU 等并行算力的利用效率，也使长序列训练速度较慢。当模型规模和数据规模成为核心变量后，该约束被进一步放大。

长程依赖仍受限：门控缓解但未根除长程问题。信息要穿过很多步的反复改写才能抵达远处，仍可能被稀释；定长状态的容量也限制了能携带的信息量。

4. Transformer 机制：基于自注意力实现全局交互

2017 年，谷歌团队的论文《Attention Is All You Need》提出 Transformer，彻底舍弃循环，仅用注意力机制建模序列。

4.1. 自注意力与 QKV：全局检索的基础

自注意力（self-attention）是 Transformer 的核心机制。处理某一位置时，模型让该位置与序列中所有位置计算相关性，并按相关性权重汇总其他位置的信息，从而在同一层内获得全局上下文。相较 RNN 逐步传递信息，自注意力显著缩短了远距离信息交互路径。

QKV 对应查询、键和值三类向量。每个位置生成 Query、Key 和 Value：Query 表示当前需要检索的信息，Key 表示该位置可被匹配的特征，Value 承载实际内容。模型用 Query 与所有 Key 计算相关度，再按权重汇总对应 Value，从而完成按需检索。

与循环结构相比，自注意力的关键变化在于任意两个位置可以直接相连。RNN 需要

通过多步状态更新传递远处信息，Transformer 则可在单层中建立远距离关联，使长程依赖从多步传递问题转化为全局匹配问题。

4.2. 核心设计：多头注意力、位置编码与前馈层

多头注意力并行设置多组 QKV，每组关注不同关系模式，例如语法搭配、指代关系或局部结构。多头机制使模型能够在同一层捕捉多种依赖关系，并在后续层中进一步组合。

位置编码：自注意力本身不区分先后顺序，因此须额外注入“位置编码”告诉模型每个元素原本的位置，弥补“顺序携带信息”这一序列特性。

残差、归一化与前馈层：每个 Transformer 块由注意力子层与前馈网络（FFN）子层组成，辅以残差连接与层归一化以稳定深层训练。

4.3. 规模化能力与计算代价

Transformer 的核心优势是并行训练和规模化能力。不同于 RNN 必须按时间步串行计算，Transformer 可以同时计算序列中各位置之间的注意力关系，训练并行度显著提升。该特性使海量数据和大规模模型训练在工程上更可行，并推动规模法则（Scaling Law）成为大模型时代的重要经验规律。

注意力计算随长度平方增长。自注意力需要让每个位置与其他位置两两匹配，序列长度翻倍时，注意力计算量约扩大为原来的四倍。处理超长文本、长视频或长轨迹时，该成本会迅速上升，也成为后续高效注意力和状态空间模型重点优化的对象。

KV 缓存带来显存和带宽压力。生成阶段为避免重复计算，模型会缓存已生成内容的 Key 和 Value；该机制降低了单步重复计算，但缓存大小随上下文长度线性增长。上下文越长，显存占用和内存带宽压力越高，这是端侧、车端和长上下文部署的重要瓶颈。

因此，Transformer 以全量保留和按需检索换取长程建模与并行训练能力，同时承担随序列长度增长的计算和缓存成本。这一代价，构成后续新循环路线和高效注意力方法重新受到关注的基础。

5. 路线分歧：记忆、计算与新循环回潮

在云端通用基础模型中，Transformer 已经成为主导架构，并显著替代传统 RNN 主干。近年 Mamba、RWKV 等新循环路线重新活跃，并非已替代 Transformer，而是在效率约束、长上下文和端侧低功耗场景中的重要补充与部署探索。

5.1. 三组关键分歧

记忆方式：定长状态对全量缓存。RNN 以固定大小的状态概括历史，内存相对恒定，但远处信息可能被稀释；Transformer 保留全部历史并按需检索，信息保留更充分，但内存随长度增长。这是二者最根本的差异。

计算特性：训练与推理存在权衡。RNN 训练串行、并行度较低，但推理每步成本相对恒定；Transformer 训练高度并行，但推理时注意力计算随长度增加，KV 缓存也随上下文增长。二者在不同阶段承受不同类型的成本。

长程依赖：理论容量与现实成本并存。Transformer 任意两位置可直接交互，长程建模路径更短；RNN 受定长状态和多步更新约束，远距离信息更易衰减。但在超长序列中，全量注意力的平方成本会削弱其经济性，两类架构均存在现实边界。

5.2. 循环思想以新形态融入模型

线性注意力通过改写注意力形式，将平方复杂度降为线性复杂度。Katharopoulos 等在 2020 年提出的《Transformers are RNNs》显示，特定注意力形式可等价改写为循环结构，从而在数学上建立注意力与循环之间的联系。

状态空间模型与 Mamba 代表新循环的重要方向。状态空间模型源自控制理论，用连续状态方程概括历史，并可在线性时间内计算。S4 将其引入深度学习；Mamba 进一步加入选择性机制，使状态更新依赖输入内容，从而具备更强的信息筛选能力。Mamba 训练随长度线性增长，推理每步成本相对恒定且无需 KV 缓存，在长序列场景中展现出与 Transformer 竞争的潜力。

RWKV 尝试结合 Transformer 训练并行性与 RNN 推理效率。其训练阶段可实现较高并行度，推理阶段则以接近 RNN 的恒定成本逐步生成。相关论文将模型规模扩展至 140 亿参数，是大规模循环类架构的重要探索。

混合架构是当前更具落地性的折中路径。一类方案在多数层使用线性注意力或状态空间结构降低成本，在少数层保留全注意力以保留精确回忆能力。AI21 的 Jamba 将 Mamba 层、Transformer 层与混合专家（MoE）组合；MiniMax-01 采用多层线性闪电注意力后接 softmax 注意力的混合方案。混合架构体现的是效率约束下的再平衡，而非两类范式的简单合一。

5.3. 结语：以记忆与计算权衡理解序列智能

从 RNN 到 Transformer，序列建模演化的主线是记忆方式与计算方式的再平衡：用定长状态压缩历史以换取恒定成本，还是全量保留历史并按需检索以换取精确记忆。我们认为，Transformer 凭借并行训练和规模化能力，成为云端通用基础模型时代的核心底座，从 GPT 到 VLA 的多数主流基础模型仍以其为主。

RNN 的核心价值并未完全消失，其每步推理成本相对恒定、无需 KV 缓存的特征，正对应 Transformer 在长上下文和端侧部署中的成本压力。Mamba、RWKV 等新循环路线在两个场景中更具价值：1）上下文极长，需要控制计算和缓存成本；2）端侧、车端等功耗和时延约束更强的部署环境。产业层面看，通用大模型仍以 Transformer 为默认底座；物理 AI、端侧和车端更可能较早采用循环类、状态空间或混合架构。因此，Transformer 在通用基础模型中维持主导地位，循环和状态空间类架构在效率约束场景中

成为重要补充。

6. 风险提示

技术迭代不及预期的风险：数字 AI、物理 AI 技术进展低于预期，任务或应用场景拓宽速度低于预期。

大模型厂商 ARR 增速放缓的风险：若客户增长、续费率或客单价不及预期，模型厂商 ARR 增速可能放缓。

云服务商资本开支不及预期的风险：若自由现金流承压，或 AI 算力需求及投资回报低于预期，云服务商可能下调资本开支。

智能驾驶及机器人商业化落地不及预期的风险：若产品可靠性、成本下降或用户需求低于预期，相关产品商业化进程可能放缓。

免责声明

东吴证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

在法律许可的情况下，东吴证券及其所属关联机构可能会持有报告中提到的公司所发行的证券并进行交易，还可能为这些公司提供投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

东吴证券投资评级标准

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的），北交所基准指数为北证 50 指数），具体如下：

公司投资评级：

买入：预期未来 6 个月个股涨跌幅相对基准在 15%以上；

增持：预期未来 6 个月个股涨跌幅相对基准介于 5%与 15%之间；

中性：预期未来 6 个月个股涨跌幅相对基准介于-5%与 5%之间；

减持：预期未来 6 个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来 6 个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持： 预期未来 6 个月内，行业指数相对强于基准 5%以上；

中性： 预期未来 6 个月内，行业指数相对基准-5%与 5%；

减持： 预期未来 6 个月内，行业指数相对弱于基准 5%以上。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

东吴证券研究所
苏州工业园区星阳街 5 号
邮政编码：215021
传真：（0512）62938527
公司网址： http://www.dwzq.com.cn