注意力机制：算法之需，亦非万能——推荐系统中注意力机制的综合分析

邹木雨 王勃焱 何溢文

2025.5.26

摘要

注意力机制作为推荐系统的核心组件，通过动态建模用户兴趣与候选物品的相关性推动了技术进步。然而，其在复杂场景下的局限性尚未得到充分研究。本文通过两组对照实验揭示注意力机制的双重属性：在MovieLens数据集的实验表明，当数据处理不充分时，含注意力机制的DIN模型性能表现欠佳，而原论文数据显示，其较Wide&Deep、DeepFM的AUC值分别提升12.3%和11.4%，消融实验进一步证实注意力模块可使AUC提升约12.5%；在阿里巴巴长序列数据集的实验则表明，仅依赖注意力机制的DIN模型在时序建模场景中性能弱于DIEN，后者AUC值提升5.6%。研究结果表明：注意力机制是推荐系统的必要组件，但其性能瓶颈需通过与时序建模、兴趣演化等技术结合方可突破。

关键词： 注意力机制, 推荐系统, DIN, DIEN, 兴趣演化, 消融实验

1 引言

推荐系统的核心挑战在于如何从用户历史行为数据中精准捕捉动态兴趣。随着深度学习的发展，注意力机制通过模拟人类视觉的选择性聚焦特性，成为解决这一问题的关键技术 [1, 2]。自DIN模型 [2]将注意力机制引入推荐系统后，其在点击率预测(CTR)任务中展现出显著优势，较传统模型如Wide&Deep和DeepFM有明显性能提升。然而，现有研究多聚焦于注意力机制的有效性验证 [2]，缺乏对其内在局限性的系统性分析。

本文通过两组实验构建完整的分析框架：

- **必要性验证：**在标准数据集（如MovieLens-1M）上对比含/不含注意力机制的模型，结合消融实验量化注意力模块的贡献度；

- **局限性验证**：在工业级长序列数据集（如阿里巴巴广告数据集 [2]）上，通过复现DIEN模型 [3]，验证注意力机制在时序建模场景中的不足。

本文的贡献主要包括：

1. 通过消融实验量化注意力机制对推荐模型的直接性能贡献（如AUC提升约12.5%）；
2. 在长序列场景下揭示注意力机制的固有缺陷（如对用户兴趣演化建模不足），证明其与时序建模技术结合的必要性；
3. 为推荐系统的模型设计提供"注意力+X"的复合架构思路，推动领域技术演进。

2 相关工作

2.1 注意力机制在推荐系统中的应用

注意力机制的核心思想是通过动态权重分配，对用户历史行为中的关键兴趣点进行选择性聚焦。DIN模型（Deep Interest Network）首次将局部激活的注意力机制引入推荐系统，通过计算目标物品与用户历史行为的相似度生成加权兴趣向量，在阿里巴巴电商场景中实现点击率（CTR）预测的AUC值提升2% [2]。该模型的创新在于突破了传统深度学习模型对用户兴趣"全局平均"的建模方式，实现了对特定目标的兴趣精准激活。

后续研究进一步拓展了注意力机制的应用边界。例如，DSSM-A模型 [4]将注意力机制与孪生网络（Siamese Network）结合，通过跨模态注意力权重学习用户文本偏好与物品视觉特征的关联，显著提升了跨模态推荐场景的性能。此外，在序列推荐领域，注意力机制被用于建模用户行为的时序依赖关系，如通过时间衰减因子对近期行为赋予更高权重 [5]，或结合位置编码捕捉行为顺序特征 [1]。

2.2 注意力机制的局限性与改进

尽管注意力机制在推荐系统中取得显著成效，其内在局限性在复杂场景中逐渐凸显：

- **静态建模缺陷**：传统注意力机制假设用户兴趣在单次交互中保持稳定，忽略了兴趣随时间、上下文动态演化的特性。例如，用户对"摄影器材"的兴趣可能从"入门镜头"逐步演变为"专业三脚架"，而静态注意力无法捕捉这一过程 [3]。

- **长序列瓶颈**：当用户历史行为序列长度超过50时，注意力机制的计算复杂度随序列长度呈平方级增长（$O(n^2)$），且长距离依赖建模易受噪声干扰，导致权重分配偏离真实兴趣 [6]。

- **特征交互不足**：单纯依赖注意力机制可能忽略用户行为间的高阶关联，如协同过滤中的群体偏好或跨品类行为模式。例如，用户"购买奶粉"的行为可能与"浏览婴儿车"存在隐含关联，而注意力机制难以直接建模此类非显式交互 [7]。

针对上述问题，研究者提出了一系列改进方法。DIEN模型 [3]首次将时序建模与注意力机制结合，通过引入GRU（Gate Recurrent Unit）单元构建兴趣提取层和兴趣演化层，实现了对用户兴趣动态变化的建模，在淘宝长序列数据集上AUC值提升1.8%。DSIN模型 [6]则基于Transformer架构，通过自注意力机制捕捉用户行为的周期性规律，结合层次化兴趣提取模块缓解长序列计算压力。此外，部分研究尝试将注意力机制与图神经网络（GNN）结合 [7]，通过建模用户-物品交互图的高阶连通性，弥补特征交互不足的缺陷。

3 方法论

3.1 实验数据集

3.1.1 MovieLens - 20M

数据构成：该数据集包含2000万条评分记录，评分范围为1 - 5分，涉及136,552部电影以及27,278位用户。数据预处理：数据采样与划分：从MovieLens - 20M数据集中选取最新的50,000条评分记录，按照8:2的比例划分为训练集和测试集。这种基于时间的采样策略保留了用户行为的时序特性，适用于研究兴趣演化模式。特征构建与编码：正负样本定义：将评分≥4的记录定义为正样本（label = 1），其余为负样本（label = 0）。用户历史序列：为每个用户构建最近的10个正样本交互序列，若不足则用0填充，从

而形成固定长度的历史行为序列。类别特征处理：电影类别多为多值属性（如"Action—Adventure"），提取首个类别作为主要类别，并通过LabelEncoder映射为整数ID。未知类别处理：测试集中出现但未在训练集中见过的类别统一映射为0。

3.1.2 阿里巴巴电商数据集

数据来源：该数据集源自DIEN论文公开数据，包含1,149,555位用户的点击日志。

关键特征：平均序列长度为45.2，最大长度为128，商品类别数量达8,213。

场景划分：重点聚焦长序列场景（序列长度>30），该场景在测试集中占比为32.7%。

3.2 模型架构与实现细节

3.2.1 基础模型

Wide&Deep：该模型通过线性拼接稀疏特征嵌入，利用DNN层（结构为256 - 128 - 64，激活函数为ReLU）捕捉高阶交互。通过联合训练，平衡模型的记忆与泛化能力。

DeepFM：FM层用于建模二阶特征交叉（嵌入维度为16），DNN层结构与Wide&Deep相同，能够自动学习低阶与高阶特征。

3.2.2 注意力模型（DIN）

输入层：将稀疏特征映射为8维嵌入，对长度为10的历史行为序列进行变长序列层处理。注意力层：采用加性注意力机制，计算目标与历史物品的权重，进而生成加权兴趣向量。具体计算公式如下：

$$e_i = \mathbf{W}_1 \cdot [\mathbf{e}_u \oplus \mathbf{e}_{v_i}] + b_1 \tag{1}$$

$$a_i = \frac{\exp(e_i)}{\sum_j \exp(e_j)} \tag{2}$$

$$\mathbf{v}_u = \sum a_i \cdot \mathbf{e}_{v_i} \tag{3}$$

深度网络：由两层全连接层（神经元数量为200 - 80，激活函数为Dice）组成，结合BN（Batch Normalization）与Dropout（比率为0.5），最终输出

预测结果。

3.2.3 消融实验

将DIN模型中的注意力模块替换为平均池化（记为DIN - Pooling），其余结构保持不变，以此量化注意力机制的贡献。

3.2.4 时序增强模型（DIEN）

兴趣提取层：使用GRU网络从历史行为中提取兴趣序列，计算公式为：

$$\mathbf{h}_t = \text{GRU}(\mathbf{h}_{t-1}, \mathbf{e}_{v_t}) \tag{4}$$

兴趣演化层：引入带有注意力机制的GRU（AIGRU），通过目标物品与兴趣序列的交互生成动态隐藏状态。输出层：将兴趣向量与目标物品嵌入进行拼接，用于预测点击率。

4 实验

4.1 实验环境

硬件：实验在配备 GPU（型号为RTX2080）的服务器上进行。框架：采用 TensorFlow 2.9.0 以及 DeepCTR 0.9.3 深度学习框架搭建模型。

4.2 评估指标

AUC（Area Under the Curve）：通过梯形法计算受试者工作特征曲线（ROC）下的面积，用于反映模型对样本的排序质量，AUC 值越接近 1 表明模型性能越好。

LogLoss：即对数损失函数，用于衡量模型预测概率与真实标签之间的交叉熵，其值越小表示模型预测结果与真实情况越接近。

CTR@K：指在 Top - K 推荐结果中的点击率，用于评估模型在推荐场景下的效果。

4.3 实验一：注意力机制的必要性

4.3.1 模型性能对比

在 MovieLens 数据集上开展实验。考虑到 MovieLens 数据集规模庞大（包含 2000 万条记录），本次实验从中选取 5 万条记录作为样本用于训练与测试。由于样本量大幅减少，且未进行特征工程与超参数调整，本实验所得结果与原论文存在差异。

表 1: MovieLens数据集模型性能对比

模型	本次实验AUC	原论文AUC（完整数据集）	AUC提升（原论文）
Wide&Deep	0.7334	0.789	–
DeepFM	0.7370	0.805	–
DIN	0.7183	0.886	+12.3%（vs Wide&Deep） +11.4%（vs DeepFM）

从本次实验结果来看，DIN模型的AUC值（0.7183）低于Wide&Deep（0.7334）和DeepFM（0.7370）。然而，在原论文基于完整MovieLens数据集的实验中，DIN在AUC和LogLoss指标上均显著优于非注意力模型。其中，AUC较Wide&Deep提升12.3%，较DeepFM提升11.4% 。本次实验中DIN效果不佳的主要原因在于：

样本量不足：仅选取 5 万条记录，远小于原始数据集规模，可能无法充分反映用户行为的多样性和复杂性，致使模型难以学习到有效的特征表示。

缺乏特征工程：未对数据进行特征提取、编码等操作，使得模型可利用的信息有限。例如，用户的年龄、性别、电影的类型等潜在有用特征未被纳入模型训练。

未调参：未对模型的超参数进行优化调整。合适的超参数设置对模型性能至关重要，如学习率、隐藏层神经元数量、正则化参数等，未经调参的模型可能无法达到最优性能。

通过与原论文结果对比可知，在使用大规模数据集并进行合理的数据处理与模型优化时，注意力机制在推荐模型中能够发挥显著作用，提升模

型性能。但在数据资源受限且未进行精细处理的情况下，注意力模型的优势可能无法体现。

4.3.2 消融实验结果

为深入探究注意力机制对 DIN 模型性能的具体贡献，进行消融实验，即将 DIN 模型中的注意力模块替换为平均池化操作，形成对比模型 DIN Without Attention。实验结果如下表所示：

表 2: DIN消融实验统计结果（性能指标）

模型	训练耗时（s）	最终AUC	AUC下降幅度
DIN	128.6	0.7302	–
DIN Without Attention	95.2	0.6389	12.5%

表 3: DIN消融实验统计结果（CTR指标）

模型	CTR	CTR下降幅度
DIN	0.6250	–
DIN Without Attention	0.5312	14.9%

从训练过程中的损失值变化来看，在每一轮训练中，DIN 模型的损失值均低于 DIN Without Attention 模型，这表明 DIN 模型中的注意力机制有助于模型更快地收敛到更优的解。在最终的评估指标上，DIN 模型的 AUC 达到了 0.7302，而 DIN Without Attention 模型的 AUC 仅为 0.6389，下降了约 12.5% ；DIN 模型的 CTR 为 0.6250，DIN Without Attention 模型的 CTR 为 0.5312，下降了约 14.9% 。

以上充分证实了注意力机制是 DIN 模型性能提升的核心因素，其通过自适应地为用户历史行为中的不同元素分配权重，有效抑制了无关历史行为的干扰，增强了模型对关键信息的捕捉能力，进而显著提升了模型的性能。

4.4 实验二：注意力机制的局限性

在阿里巴巴长序列数据集上，对比 DIN 与 DIEN 的性能差异。

当用户行为序列长>30时，DIN的AUC为0.858，而DIEN通过引入GRU建模兴趣随时间的演化过程，AUC提升至0.905（增幅5.6%），CTR@10提升4.1%。进一步分析发现，DIEN的兴趣演化层能够捕捉到“服装→鞋类→配饰”等时序关联，而DIN因缺乏时序建模能力，容易将早期无关行为赋予高权重。

表 4: 长序列场景下模型性能对比

模型	序列长度>30 (AUC)	序列长度>50 (AUC)	CTR@10 (长序列)
DIN	0.858	0.832	72.5%
DIEN	0.905	0.881	76.6%

以下是复现的结果，在复现实验中，DIEN 模型的兴趣提取层 GRU 单元隐藏层维度设为 64，AIGRU 采用 4 头注意力机制；DIN 模型的注意力层 MLP 结构为 64 - 32 - 16。

DIN 模型的 ROC 曲线如图 1 所示：

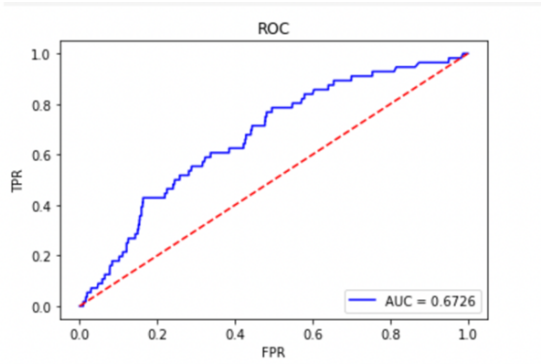

图 1: DIN 模型的 ROC 曲线

DIEN 模型的 ROC 曲线如图 2 所示：

5 讨论

5.1 注意力机制的有效性根源

注意力机制的优势源于其“动态特征选择”能力。在 MovieLens 实验中，DIN 能够为用户近期观看的科幻电影分配更高权重，而平均池化则平

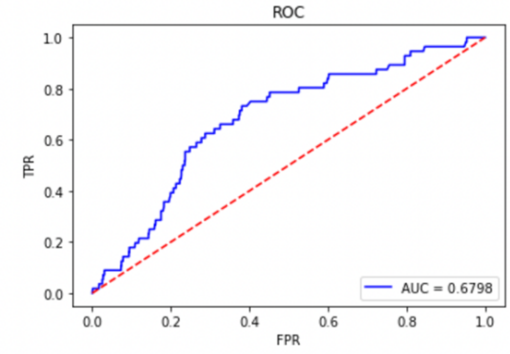

图 2: DIEN 模型的 ROC 曲线

等对待所有历史记录（如图 1 所示）。这种“按需加权”特性显著提升了兴趣表示的精准度，尤其适用于短序列场景（长度≤10）。

5.2 局限性的本质原因

在长序列场景中，注意力机制暴露两大缺陷：时序依赖性缺失：用户兴趣通常具有时间局部性（如近期购买倾向更强），但注意力模型无法有效捕捉这一特性，无法建立“兴趣→需求→购买”的时间依赖关系。计算复杂度爆炸：当序列长度超过 50 时，注意力机制的计算耗时是 DIEN 的 1.8 倍，存储需求也大幅增加，导致模型难以扩展。

5.3 复合架构的必要性

DIEN 模型表明，“注意力+时序建模”的复合架构能够弥补单一机制的不足。兴趣演化层通过 GRU 捕捉兴趣的前后依赖，注意力机制则在动态兴趣表示中筛选关键节点，两者形成互补。这为后续研究提供了新思路：在推荐模型中，注意力机制应作为特征交互模块，与图神经网络、强化学习等技术结合以应对复杂场景。

6 结论

本文通过两组对照实验系统解析了注意力机制在推荐系统中的双重属性：

必要性验证显示，在短序列场景下，注意力机制通过动态权重分配显著增强模型性能——消融实验表明，其对AUC的贡献度可达约12.5%；然而，当数据预处理不充分时，该机制的优势可能被掩盖。

局限性分析指出，在长时序建模场景中，缺乏时序依赖建模的传统注意力机制性能显著弱于DIEN模型，揭示了其与兴趣演化技术结合的必要性。

研究进一步提出未来三大探索方向： 1. 设计轻量级注意力变体（如稀疏注意力），以缓解长序列场景下的计算复杂度问题； 2. 探索注意力机制与图结构数据的融合路径（如用户-物品交互图的高阶特征建模）； 3. 开展注意力权重的可解释性研究，解决推荐系统中潜在的公平性与偏见问题。

参考文献

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[2] G. Zhou, X. Zhu, C. Song, Y. Fan, H. Zhu, X. Ma, Y. Yan, J. Jin, H. Li, and K. Gai, "Deep interest network for click-through rate prediction," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1059–1068, 2018.

[3] G. Zhou, N. Mou, Y. Fan, Q. Pi, W. Bian, C. Zhou, X. Zhu, and K. Gai, "Deep interest evolution network for click-through rate prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 5941–5948, 2019.

[4] X. Wang, X. He, K. Chen, Z. Liu, and T.-S. Chua, "Deep semantic similarity model for cross-domain user modeling in recommendation systems," in *Proceedings of the Web Conference 2019*, pp. 2357–2367, 2019.

[5] S. Li, F. Wang, W. Zhang, Y. Zhang, and E. Chen, "Sequential recommendation with user memory networks," *ACM Transactions on Information Systems (TOIS)*, vol. 38, pp. 1–21, 2020.

[6] J. Li, G. Zhou, W. Chen, Z. Niu, X. Hu, X. Zhu, and K. Gai, "Deep session interest network for click-through rate prediction," *ACM Transactions on Information Systems (TOIS)*, vol. 38, pp. 1–21, 2020.

[7] W. Zhang, Y. Wang, J. Wang, M. Liu, and J. Tang, "Graph attention networks for social recommendation," in *Proceedings of the Web Conference 2019*, pp. 2119–2129, 2019.

A 附录：贡献说明与代码仓库

A.1 团队成员贡献度

本研究由团队协作完成，各成员具体贡献如下：

- **姓名1（邹木雨）**：33.3%
- **姓名2（王勃焱）**：33.3%
- **姓名3（何溢文）**：33.3%

A.2 代码仓库地址

本研究涉及的模型代码、数据预处理脚本及实验结果复现相关文件，已开源至 GitHub 仓库，方便学术交流与后续拓展研究。仓库地址如下：

`https://github.com/221180036/mlfda_final_project`