HyMoERec: Hybrid Mixture-of-Experts for Sequential Recommendation (Student Abstract)

Kunrong Li[1], **Zhu Sun**[1], **Kwan Hui Lim**[1]

[1]Singapore University of Technology and Design, Singapore
kunrong_li@mymail.sutd.edu.sg, zhu_sun@sutd.edu.sg, kwanhui_lim@sutd.edu.sg

Abstract

We propose **HyMoERec**, a novel sequential recommendation framework that addresses the limitations of uniform Position-wise Feed-Forward Networks in existing models. Current approaches treat all user interactions and items equally, overlooking the heterogeneity in user behavior patterns and diversity in item complexity. HyMoERec initially introduces a hybrid mixture-of-experts architecture that combines shared and specialized expert branches with an adaptive expert fusion mechanism for the sequential recommendation task. This design captures diverse reasoning for varied users and items while ensuring stable training. Experiments on MovieLens-1M and Beauty datasets demonstrate that HyMoERec consistently outperforms state-of-the-art baselines.

Introduction

Sequential recommendation is a core task in modern recommendation systems, which predicts the next item based on historical behavior. Recent advances, ranging from recurrent neural networks (Jannach and Ludewig 2017), attention-based architectures (Sun et al. 2019; Li et al. 2017), to state-space models (Liu et al. 2024) and large language models (Li and Lim 2025), have greatly improved sequence modeling in this domain. However, a critical yet underexplored aspect lies with the Position-wise Feed-Forward Network (PFFN), which is the primary non-linear transformation module in most models. It applies the same FFN operation for all tokens and treats each item equally. This design neglects two key challenges: (1) User behavior heterogeneity: Users follow varied sequential patterns, with some interactions captured by simple dependencies and others requiring complex reasoning, which a uniform FFN cannot adaptively model; (2) Item complexity diversity: While popular items may be predicted from broad patterns, niche items often demand specialized representations, making identical computational pathways insufficient and less effective.

To address these challenges, we propose **HyMoERec**, a sequential recommendation framework that redefines PFFN processing via a hybrid mixture-of-experts architecture, which is illustrated in Fig. 1. Our approach introduces three innovations: (1) *Hybrid Mixture of Experts (HyMoE)*: A combination of one shared expert and multiple specialized experts. Unlike conventional MoE relying solely on sparse selection, this hybrid design ensures stable optimization through consistent learning signals while enabling specialization for domain-specific patterns; (2) *Adaptive Expert Fusion (AEF)*: A learnable gating parameter α that dynamically balances shared and specialized pathways. A progressive warm-up schedule begins with shared expert reliance and gradually integrates specialized knowledge, while load-balance regularization promotes diverse expert usage; (3) *Empirical validation*: Comprehensive experiments in MovieLens-1M and Amazon Beauty demonstrate that HyMoERec consistently exceeds state-of-the-art baselines.

Methodology

Hybrid Mixture-of-Experts (HyMoE). A common limitation in sequential recommendation is that all tokens are processed by a uniform PFFN, regardless of their semantic or temporal characteristics. This uniformity underutilizes the model capacity and fails to capture the heterogeneity of user behaviors. Therefore, we introduce Hybrid Mixture of Experts (HyMoE) block inspired by (Dai et al. 2024) for the sequential recommendation task. Unlike conventional MoE that routes tokens to a sparse set of experts, we employ a dual-branch structure. It consists of a dense, shared expert branch and a sparse, specialized expert branch, ensuring stable learning while providing adaptive capacity.

Given a token representation $\mathbf{x} \in \mathbb{R}^D$, we first calculate a dense FFN output $\mathbf{y}_{\text{dense}}$. In parallel, $\mathbf{x}$ is passed through a lightweight router $\pi(\cdot)$, parameterized as a two-layer linear computation. Let $\mathcal{E} = \{f_1, \ldots, f_E\}$ be the set of expert networks, the router produces $\pi(\mathbf{x}) \in \mathbb{R}^E$, where each dimension is a routing logit reflecting the affinity of the current token-context pair to one of the E FFN experts. Sparse gating weights are then obtained by retaining only the top-k expert logits and applying a softmax over them: $\mathbf{g} = \text{Softmax}\big(\text{TopK}(\pi(\mathbf{x}))\big)$, where TopK is the selection of the K experts with the highest probabilities, $\mathbf{g} = \{g_1, ..., g_K\}$ represents the set of weights for the K experts. The aggregated mixture of expert output is therefore computed as:

$$\mathbf{y}_{\text{MoE}} = \textstyle\sum_{i=1}^{K} g_i f_i(\mathbf{x}), \tag{1}$$

where only TopK experts are active. This design brings two benefits: (1) dense branch guarantees a stable representation for every token, while the sparse branch adaptively increases

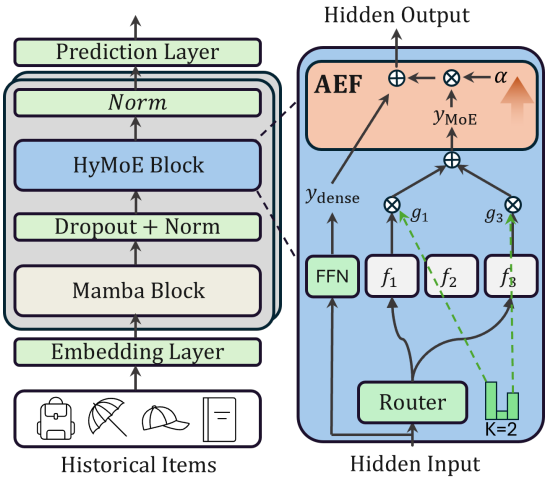

Figure 1: The overall framework of HyMoERec.

expressiveness only when necessary; and (2) the training of the model is stabilized, which avoids the expert collapse.

Adaptive Expert Fusion (AEF). To effectively balance dense and sparse experts while ensuring stable training dynamics, instead of a fixed combination strategy that may fail to adapt to the evolving expertise of different components (Dai et al. 2024), we design an Adaptive Expert Fusion (AEF) mechanism that dynamically combines the outputs of both branches. Our AEF employs a fusion parameter α combined with a progressive warm-up strategy:

$$\mathbf{y} = \mathbf{y}_{\text{dense}} + \alpha \cdot \mathbf{y}_{\text{MoE}}, \quad \alpha = \sigma(\alpha_{\text{param}}) \cdot w(t). \tag{2}$$

Here, $\sigma(\cdot)$ is the sigmoid function applied to the learnable parameter α_{param}, and $w(t)$ is the warm-up factor:

$$w(t) = \min(1.0, t/T_{\text{warmup}}) \tag{3}$$

where t is the current step and T_{warmup} is the warm-up duration. AEF ensures gradual integration of expert knowledge while maintaining training stability. To avoid a few experts domination, we apply a load-balance loss to encourage uniform expert usage. Thus, the full training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \lambda_{\text{lb}} \sum_e \bar{g}_e \log \bar{g}_e, \tag{4}$$

where $\bar{g}_e$ denotes the average gating probability of the expert e on all items in the sequence, λ_{lb} is the balancing parameter and $\mathcal{L}_{ce} = -\log \frac{\exp(\mathbf{y}^T \mathbf{r}_+)}{\sum_j \exp(\mathbf{y}^T \mathbf{r}_j)}$ with $\mathbf{r}_+$ the embedding of the target element and r_j as all embeddings. This regularization stabilizes training and promotes generalization by ensuring all experts contribute meaningfully to the predictions.

Experiments

Datasets and Setup. We experiment on the MovieLens-1M and Beauty datasets. We set the number of experts E=4 and select top K=2 experts. T_{warmup} is set to 500 steps, and λ_{lb}=0.02. We compare with NARM (Li et al. 2017), GRU4Rec (Jannach and Ludewig 2017), BERT4Rec (Sun et al. 2019), and Mamba4Rec (Liu et al. 2024).

Results. The results are shown in Tab. 1. On the ML-1M dataset, HyMoERec demonstrates superior performance

Dataset	Method	HR@5	HR@10	NDCG@5	NDCG@10
ML-1M	NARM	0.1727	0.2565	0.1135	0.1404
	GRU4Rec	0.1998	0.2815	0.1364	0.1628
	BERT4Rec	0.1879	0.2868	0.1249	0.1568
	Mamba4Rec	0.2106	0.3065	0.1447	0.1756
	HyMoERec	**0.2157**	**0.3103**	**0.1486**	**0.1790**
Beauty	NARM	0.0418	0.0630	0.0285	0.0352
	GRU4Rec	0.0402	0.0635	0.0274	0.0349
	BERT4Rec	0.0243	0.0402	0.0151	0.0202
	Mamba4Rec	0.0511	0.0731	0.0362	0.0432
	HyMoERec	**0.0518**	**0.0756**	**0.0365**	**0.0442**

Table 1: Recommendation performance on ML-1M and Beauty datasets. Bold scores represent the highest results.

with HR@5 of 0.2157, HR@10 of 0.3103, NDCG@5 of 0.1486, and NDCG@10 of 0.1790. These results represent superiority over the strongest baseline, Mamba4Rec, of 2.4% in HR@5, 1.2% in HR@10, 2.7% in NDCG@5, and 1.9% in NDCG@10. Compared to NARM and GRU4Rec, HyMoERec achieves 24.8% and 7.9% HR@5 improvements, demonstrating our effectiveness. In the Beauty dataset, HyMoERec also shows substantial improvements compared to BERT4Rec, improving from 0.0243 to 0.0518 in HR@5 and from 0.0151 to 0.0365 in NDCG@5. These consistent improvements demonstrate the effectiveness and generalizability of our approach across different domains.

Conclusion

We introduce HyMoERec, a sequential recommendation framework that directly addresses the limitations of PFFNs by modeling the heterogeneity of user behaviors and item complexities. We propose the Hybrid Mixture of Experts (HyMoE) and Adaptive Expert Fusion (AEF) mechanisms, which provide a dynamic and stable approach to sequential recommendation. Experiments on MovieLens-1M and Beauty datasets show that HyMoERec consistently outperforms state-of-the-art baselines.

Acknowledgements. This research is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award No. MOE-T2EP20123-0015).

References

Dai, D.; Deng, C.; Zhao, C.; Xu, R.; and et al. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv:2401.06066*.

Jannach, D.; and Ludewig, M. 2017. When recurrent neural networks meet the neighborhood for session-based recommendation. In *RecSys*, 306–310.

Li, J.; Ren, P.; Chen, Z.; Ren, Z.; Lian, T.; and Ma, J. 2017. Neural attentive session-based recommendation. In *CIKM*.

Li, K.; and Lim, K. H. 2025. RALLM-POI: Retrieval-Augmented LLM for Zero-shot Next POI Recommendation with Geographical Reranking. *arXiv:2509.17066*.

Liu, C.; Lin, J.; Wang, J.; Liu, H.; and Caverlee, J. 2024. Mamba4rec: Towards efficient sequential recommendation with selective state space models. *arXiv:2403.03900*.

Sun, F.; Liu, J.; Wu, J.; Pei, C.; Lin, X.; Ou, W.; and Jiang, P. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *CIKM*.