

Actions Speak Louder than Words: Trillion-Parameter Sequential Transducers for Generative Recommendations (2024)

Challenges

- **DLRMs do not scale with compute.** The abstract notes that despite huge data and features, "most Deep Learning Recommendation Models (DLRMs) in industry fail to scale with compute." This motivates a new formulation.
- **High-cardinality, heterogeneous features at industrial scale.** The paper targets systems that must handle "tens of billions of user actions" with highly heterogeneous features.
- **Need a unified formulation for ranking and retrieval.** The authors argue for reformulating recommendation as a sequential transduction problem within a generative modeling framework (GRs) to enable scaling laws similar to LLMs.

Key Initiatives

- **Generative Recommenders (GRs).** The work "reformulate[s] recommendation problems as sequential transduction tasks" under a generative modeling framework.
- **HSTU encoder.** Introduces a new encoder design, *Hierarchical Sequential Transduction Unit (HSTU)*, a stack of identical layers that replaces heterogeneous DLRM modules with a single modular block.
- **Efficiency + scaling.** HSTU is reported as **5.3x–15.2x faster** than state-of-the-art Transformers on length-8192 sequences, enabling training/inference at scale.
- **Production impact.** The paper states that GRs/HSTU "have led to **12.4%** metric improvements in production" and allow models "285x more complex" with less inference compute.
- **M-FALCON inference algorithm.** New inference algorithm to reduce target-aware cross-attention cost and improve serving throughput (Figure 11/12).

Methodology

- **Sequential transduction formulation.** Ranking/retrieval are cast as sequence-to-sequence prediction of interleaved item/action tokens.
- **HSTU layer design.** Each layer contains: (1) Pointwise Projection, (2) Spatial Aggregation, (3) Pointwise Transformation.
- **Core equations.** The paper defines:
 - $U(X), V(X), Q(X), K(X) = \text{Split}(\phi_1(f_1(X)))$ (Eq. 1)
 - $A(X)V(X) = \phi_2(Q(X)K(X)^T + rab_{p,t})V(X)$ (Eq. 2)
 - $Y(X) = f_2(\text{Norm}(A(X)V(X)) \odot U(X))$ (Eq. 3)
- **Pointwise aggregated attention.** Replaces softmax normalization with a pointwise aggregated mechanism to better capture engagement intensity and handle non-stationary vocabularies.
- **Model figure.** The key model comparison is shown in **Figure 3** (DLRMs vs GRs, simplified HSTU block).

Experiments

- **Public dataset gains (Table 4).** HSTU and HSTU-large improve HR@10 / NDCG@10 across ML-1M, ML-20M, Books. Example results from Table 4:
 - **ML-1M:** HSTU HR@10 **0.3097** (+8.6%), NDCG@10 **0.1720** (+7.3%); HSTU-large HR@10 **0.3294** (+15.5%), NDCG@10 **0.1893** (+18.1%).
 - **ML-20M:** HSTU HR@10 **0.3252** (+11.9%), NDCG@10 **0.1878** (+15.9%); HSTU-large HR@10 **0.3567** (+22.8%), NDCG@10 **0.2106** (+30.0%).
 - **Books:** HSTU HR@10 **0.0404** (+38.4%), NDCG@10 **0.0219** (+40.6%); HSTU-large HR@10 **0.0469** (+60.6%), NDCG@10 **0.0257** (+65.8%).
- **Ablations (Table 5).** Comparing HSTU variants (with/without $rab_{p,t}$, softmax) shows clear quality drops in ablations, supporting the design choices.
- **Efficiency/scaling.** HSTU is up to **15.2x** more efficient and supports longer sequences (Figure 5/6/7), enabling scaling behavior similar to LLMs.
- **Production impact.** The paper claims GRs/HSTU deliver **12.4%** production metric improvements.

References (selected)

- Mudigere et al. (2022). DLRM at scale.
- Raffel et al. (2020). T5 / relative attention bias.
- Dao (2023). FlashAttention-2.
- Kang & McAuley (2018). SASRec.
- Zhou et al. (2018). DIN.