Ads Recommendation in a Collapsed and Entangled World

Junwei Pan, Wei Xue, Ximei Wang, Haibin Yu,
Xun Liu, Shijie Quan, Xueming Qiu,
Dapeng Liu, Lei Xiao, Jie Jiang

Tencent 腾讯

KDD2024
BARCELONA, SPAIN

Overview

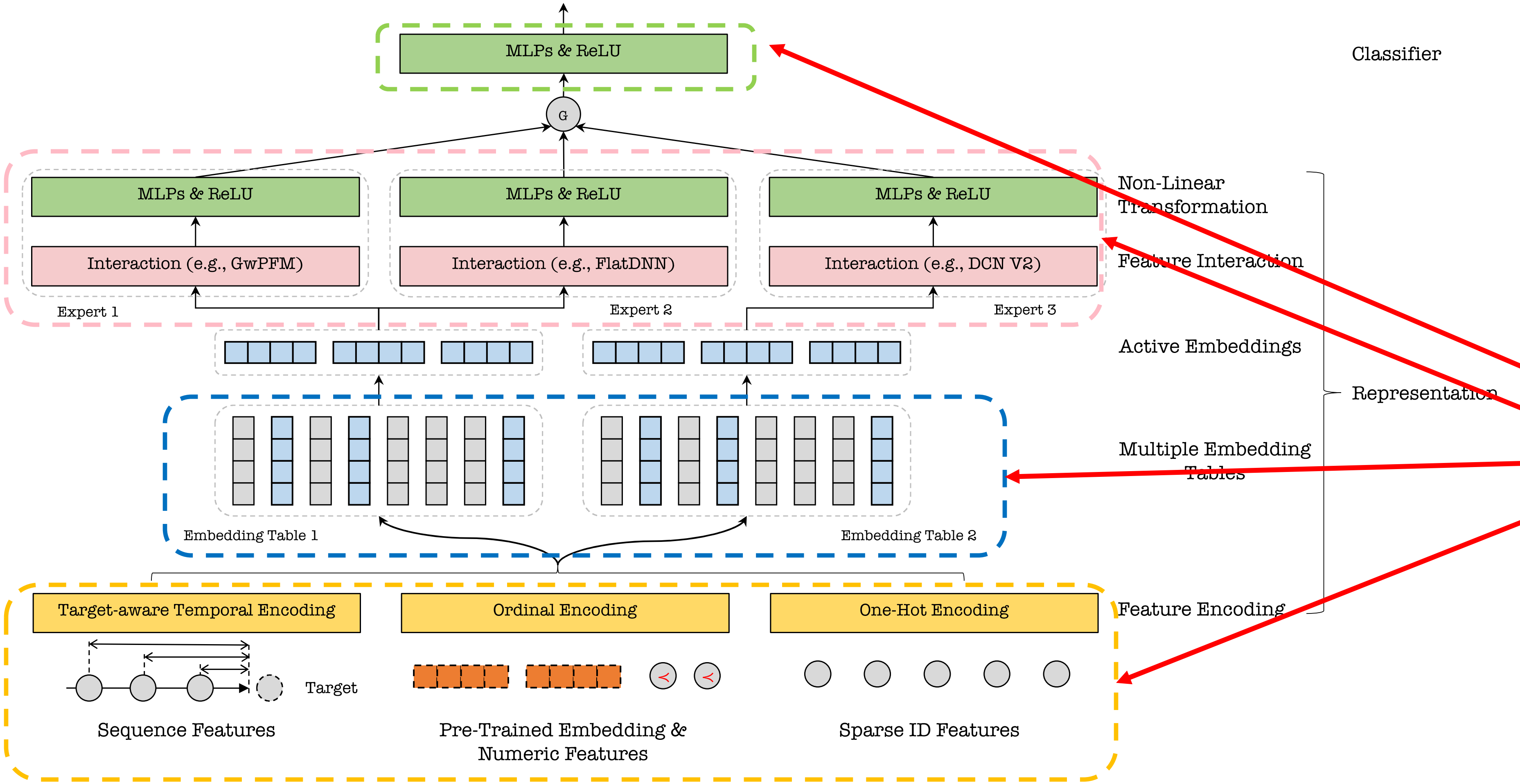
Three main challenges:

- **C1: Priors for Representation**
- **C2: Dimensional Collapse**
- **C3: Interest Entanglement**

This work is a summary of many advances of Tencent Ads Science team in the last several years, some of them are published as individual papers (e.g., in ICML 2024, KDD 2024, WWW 2024, AAAI 2024, etc.).

KDD2024
BARCELONA, SPAIN

Architecture



Our model follows the widely adopted Embedding & Explicit Interaction framework, which consists of four key modules:

- Classification towers
- Experts
- Multi-embedding lookup
- Feature encoding

Outline

Three main challenges:

- **C1: Priors for Representation**
- **C2: Dimensional Collapse**
- **C3: Interest Entanglement**

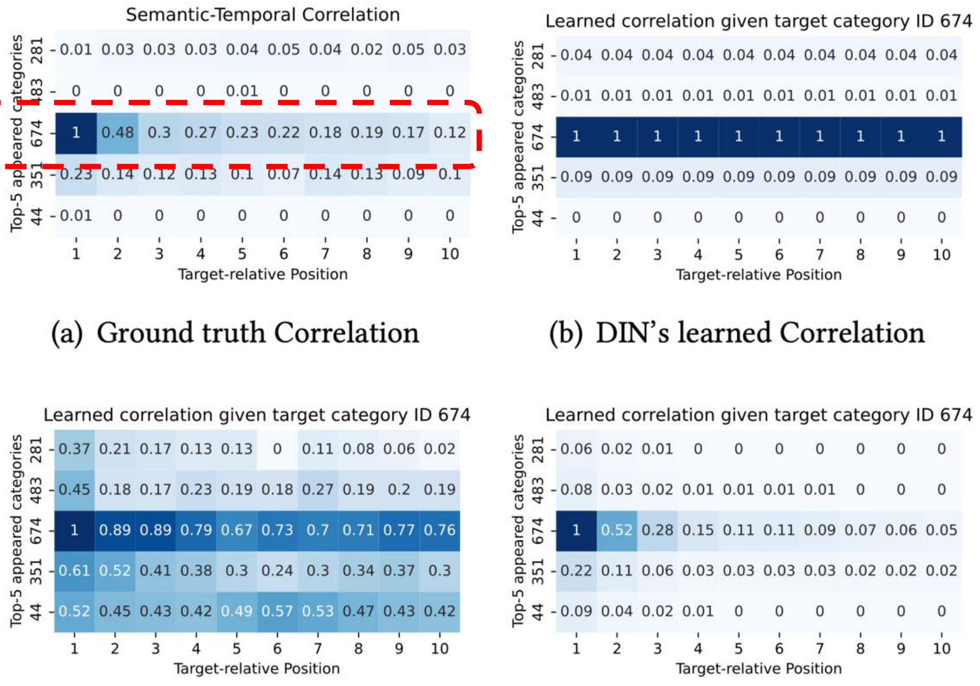
KDD2024
BARCELONA, SPAIN

C1: Priors of Sequence Features

Semantic correlation: behaviors with the same category as the target (674 here) are more correlated.

Temporal correlation: there is a strong temporal decaying pattern.

Target category: 674

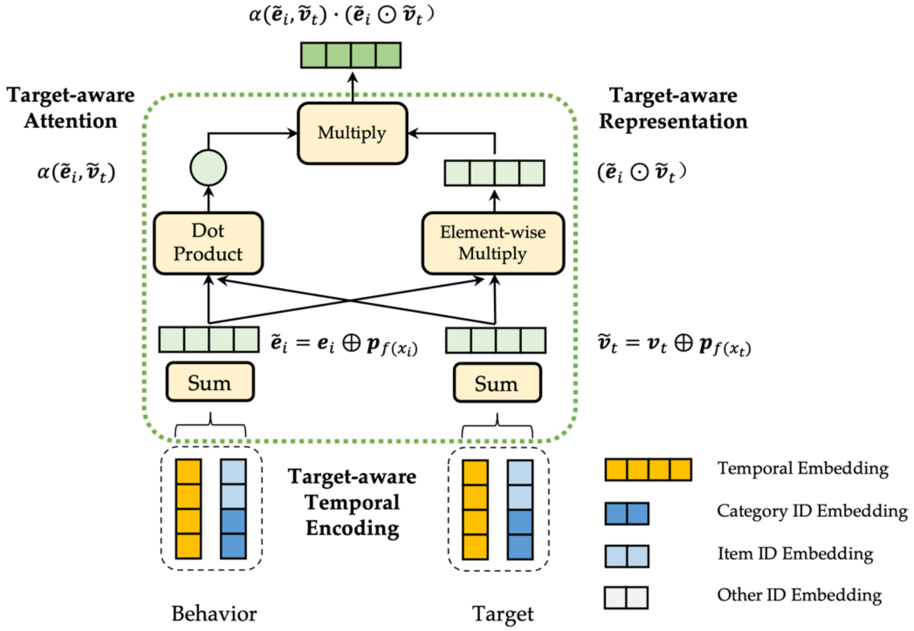


(a) Ground truth Correlation

(b) DIN's learned Correlation

(c) SASRec's learned Correlation

(d) TIM's learned Correlation

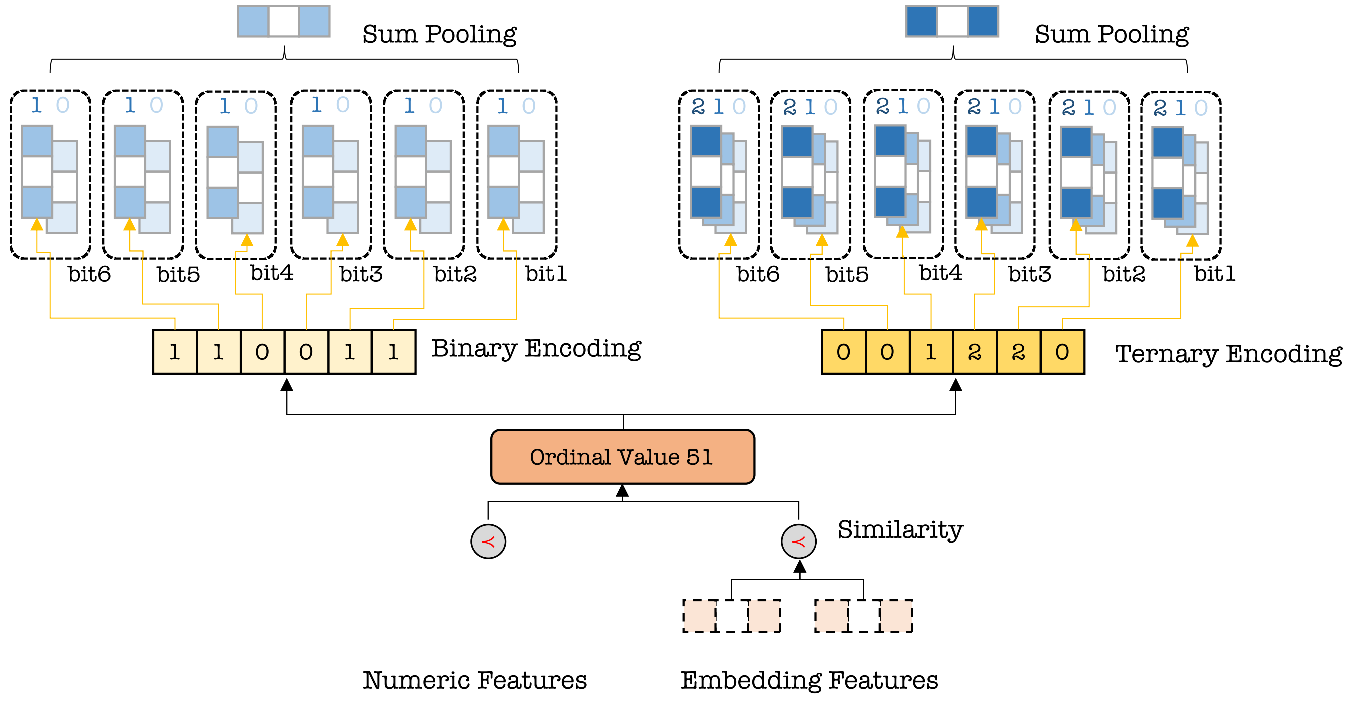


- There are strong semantic and temporal correlation between behaviors and the target.
- Existing SOTA methods fail to capture such correlation.

Our proposed TIN composes 3 key components: target-aware temporal encoding, target-aware attention and target-aware representation.

KDD2024 BARCELONA, SPAIN

C1: Priors of Numeric & Embedding Features



Numeric Features

Multiple Numeral Systems Encoding (MNSE):

$$f_{\mathrm{MNS}}(v) = \left[\sum_{k=1}^{K_2} \mathrm{X}^{(2)}_{2k+\mathbb{B}_k}, \sum_{k=1}^{K_3} \mathrm{X}^{(3)}_{3k+\mathbb{C}_k}, \ldots, \sum_{k=1}^{K_n} \mathrm{X}^{(n)}_{nk+\mathbb{N}_k}\right]$$

$$\mathbb{B} = \mathrm{func\_binary}(v),\ \mathbb{C} = \mathrm{func\_ternary}(v), \ldots$$

Embedding Features

Compute a similarity score based on the embeddings

$$w_{\mathrm{sim}}(u, i) = \mathrm{sim}(\bar{\boldsymbol{e}}_u, \bar{\boldsymbol{e}}_i).$$

Treat the score as a numeric feature, use the MNSE method

Outline

Three main challenges:

- **C1: Priors for Representation**
- **C2: Dimensional Collapse**
- **C3: Interest Entanglement**

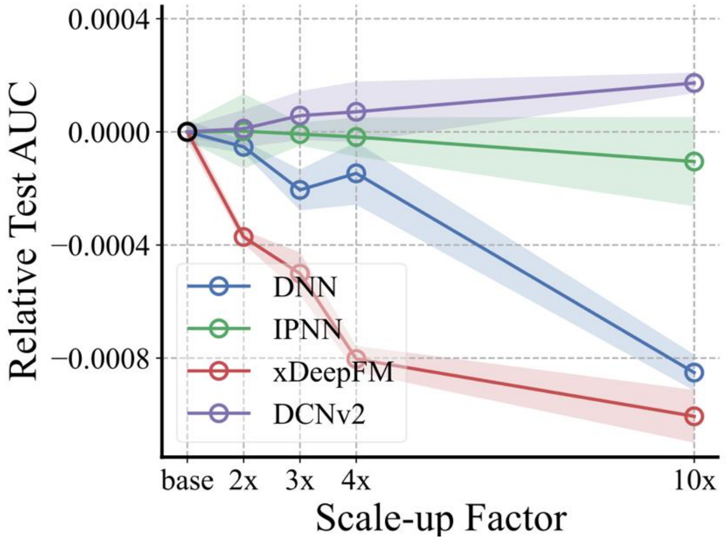
KDD2024
BARCELONA, SPAIN

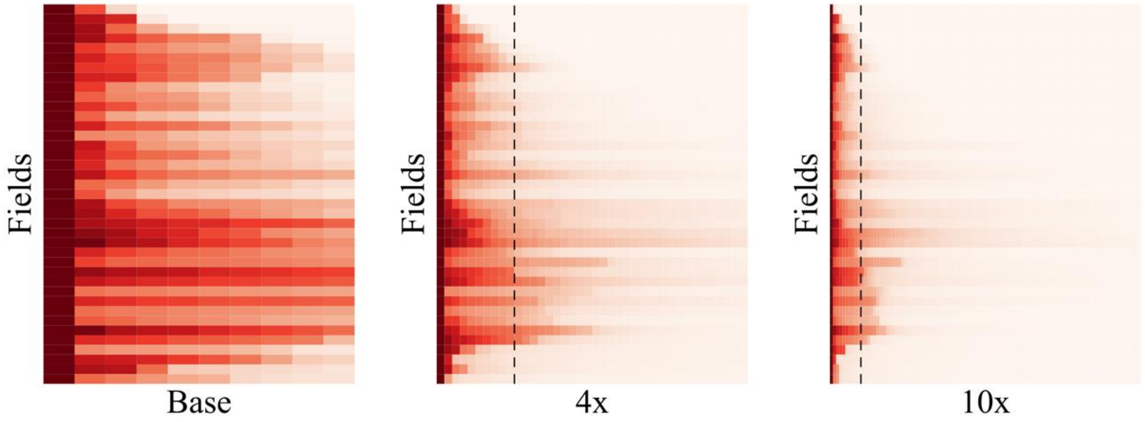
C2: Dimensional Collapse of Embeddings

Large Recommendation System

How to scale up recommendation models w.r.t. #parameters?

- Most parameters in RecSys are embeddings (**99.999%**)
- Simply enlarge the embedding dimension does not always work.
- Singular spectrum analysis show dimensional collapse of embeddings





Many singular values are very small, indicating that embeddings of many fields end up spanning a lower-dimensional subspace instead of the entire available embedding space.

Finding 1 (Interaction-Collapse Theory). In feature interaction of recommendation models, fields with low-information-abundance embeddings constrain the information abundance of other fields, resulting in collapsed embedding matrices.

KDD2024 BARCELONA, SPAIN

On the Embedding Collapse when Scaling up Recommendation Models. ICML 2024. Tencent Ads.

C2: Dimensional Collapse - Architecture

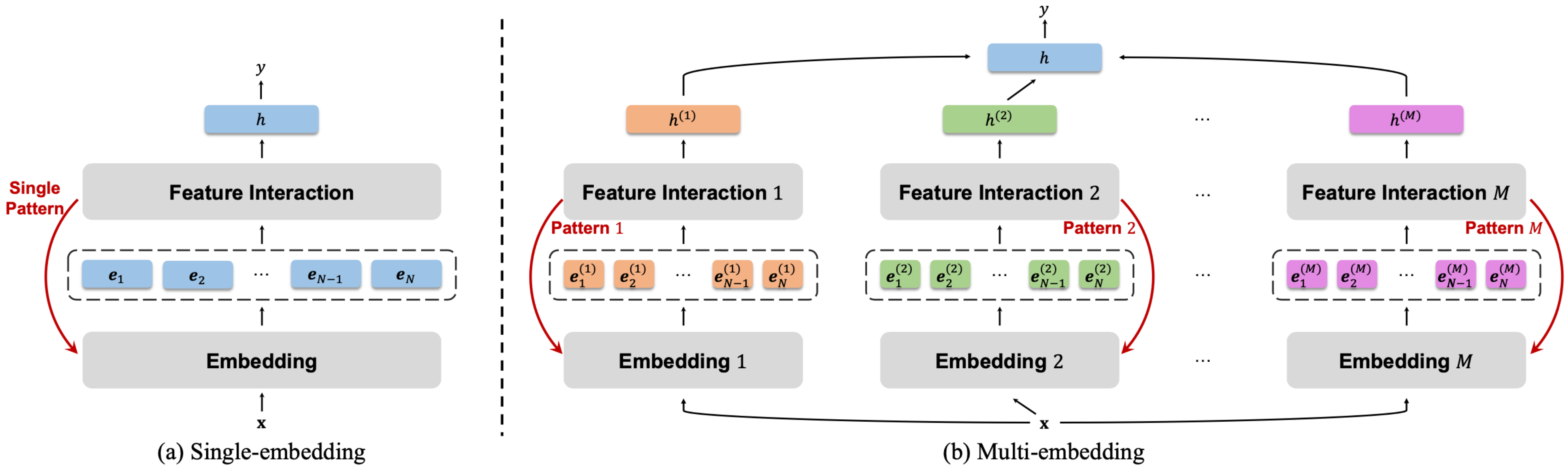
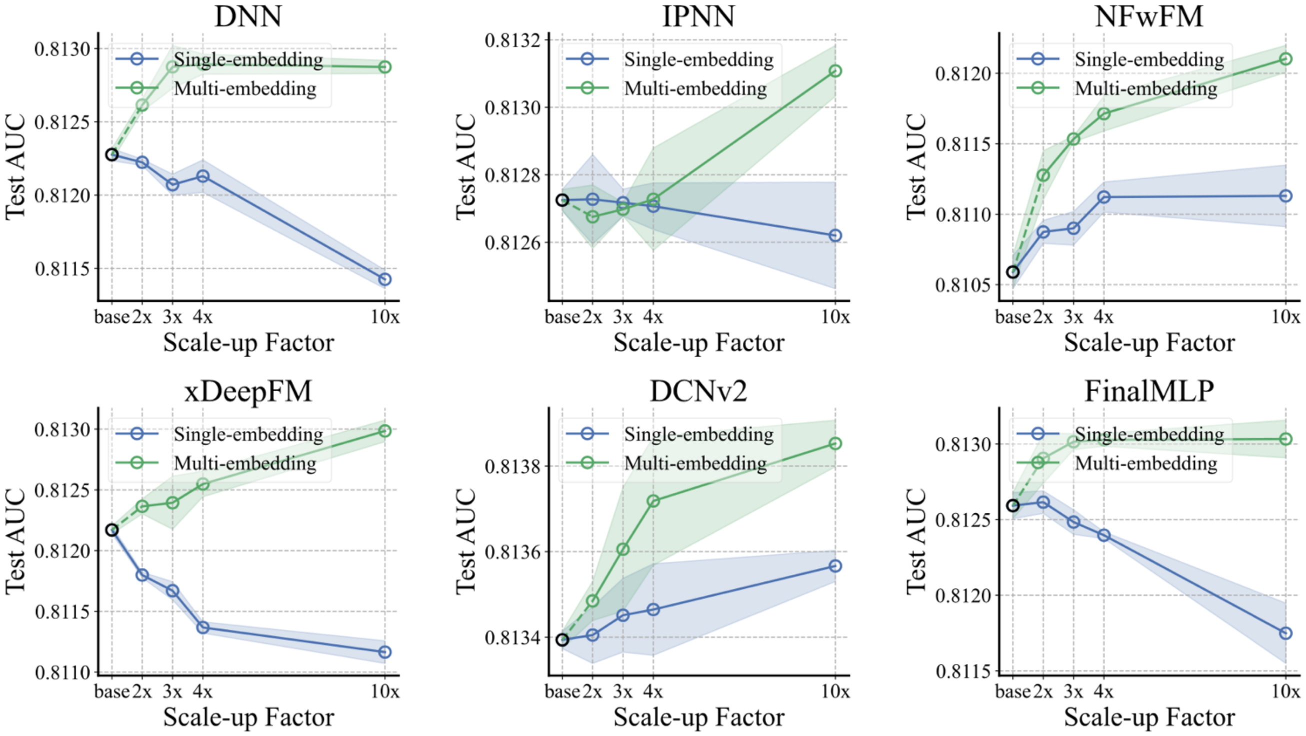


Figure 7. Architectures of single-embedding (left) and multi-embedding (right) models.

- Instead of scaling up the model by enlarging the embedding dimensions, we scale by **introducing more embedding tables**: the **Multi-Embedding paradigm**.
- ***Less influenced*** by the interaction-collapse theory and ***mitigate embedding collapse*** while ***keeping*** the original interaction modules

KDD2024 BARCELONA, SPAIN

C2: Dimensional Collapse - Results



- *Can be applied to all existing CTR models*
- *Compared to Single-Embedding, Multi-Embedding paradigm, successfully achieves scale-up regarding #params*

KDD2024 BARCELONA, SPAIN

On the Embedding Collapse when Scaling up Recommendation Models. ICML 2024. Tencent Ads.

Outline

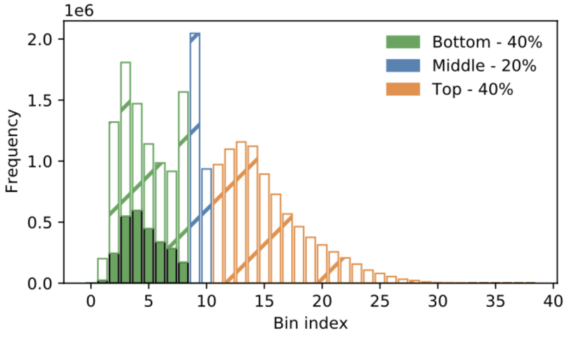
Three main challenges:

- **C1: Priors for Representation**
- **C2: Dimensional Collapse**
- **C3: Interest Entanglement**

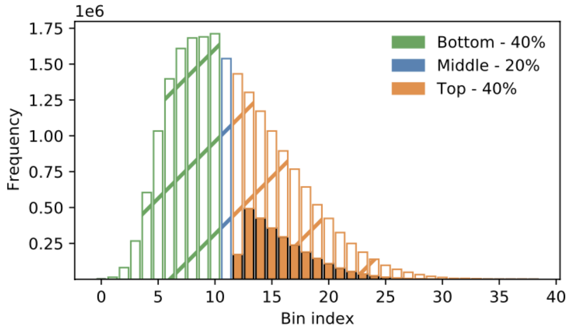
KDD2024
BARCELONA, SPAIN

C3: Interest Entanglement

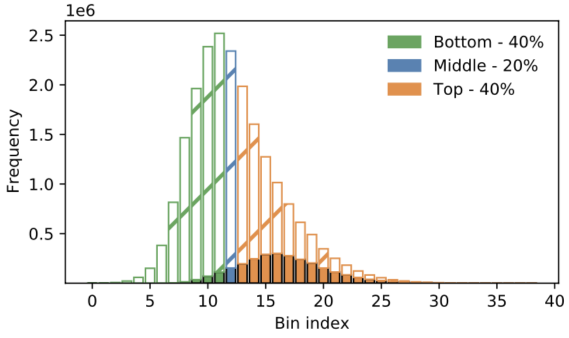
- Consider two tasks: like and finish.
- We select a set of (user, item) pairs with conflicting embedding distance (interest) according two individual tasks' embeddings.



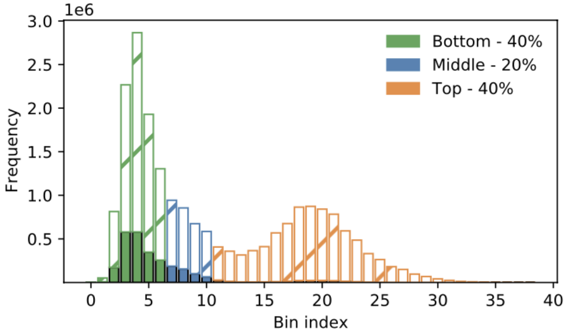
(a) Embeddings of Single Task Like



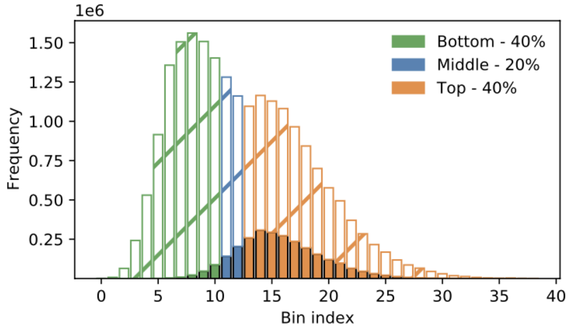
(b) Embeddings of Single Task Finish



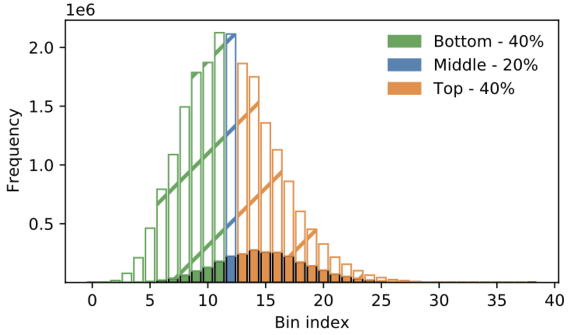
(c) Shared-Embeddings in PLE



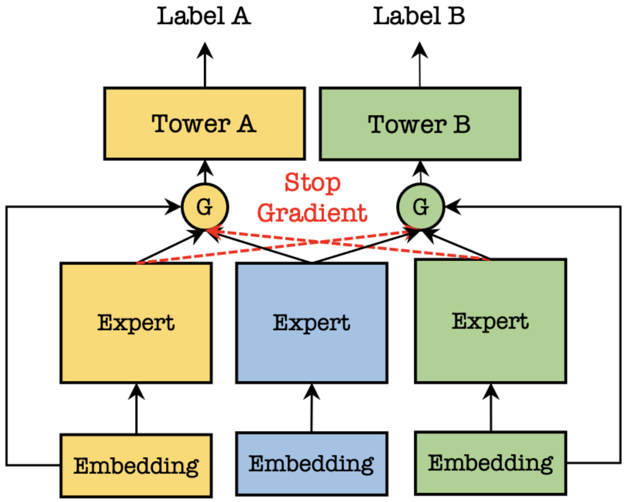
(d) Like-Specific Embeddings in STEM



(e) Finish-Specific Embeddings in STEM



(f) Shared Embeddings in STEM

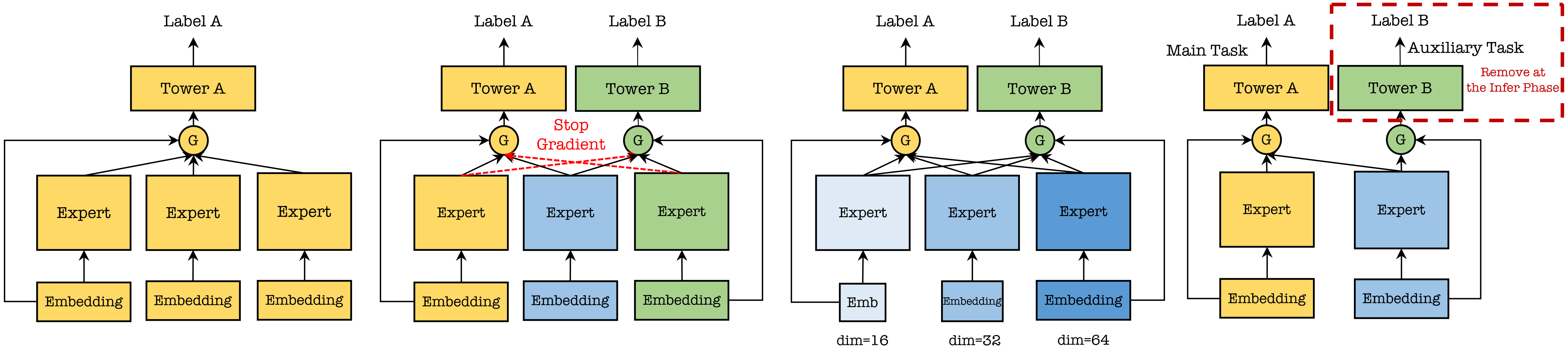


(2) STEM

Shared-and Task-specific Embedding paradigm.

STEM: Unleashing the Power of Embedding for Multi-Task Recommendation. AAAI 2024.

C3: Interest Entanglement, Architecture



(1) ME

For Single-Task Learning
Tackle dimensional collapse

(2) STEM **(3) AME**

For **Multi-Task Learning**
Tackle interest entanglement

(4) STEM-AL

For **Auxiliary Learning**
Tackle interest entanglement

Multi-Task or Multi-Domain Learning

- STEM: One embedding table for each task: too many embedding tables!
- AME (Asymmetric Multiple-Embedding): Fixed number of embeddings, with different dimension to disentangle them.
- Crocodile: Fixed number of embeddings, with an explicit disentangle loss.

Auxiliary Learning

- STEM-AL: One embedding for the main task, and one shared embedding for all tasks.

KDD2024 BARCELONA, SPAIN

STEM: Unleashing the Power of Embedding for Multi-Task Recommendation. AAAI 2024.
Crocodile: Cross Experts Covariance for Disentangled Learning in Multi-Domain Recommendation. arxiv 2024.

Overview

- Outline
  - Priors of Representations
    - Sequence Features: Temporal Interest Module — WWW 2024
    - Numeric Features: Multiple Numeral Systems Encoding
    - Embedding Feature: Similarity Encoding Embedding
  - Dimensional Collapse
    - Multi-Embedding Paradigm — ICML 2024
    - GwPFM, a variant of FFM
    - More robust interaction functions, e.g. FmFM, DCN V2
  - Interest Entanglement
    - Shared-and Task-specific EMbeddings (STEM) — AAAI 2024
    - Asymmetric Multi-Embeddings (AME)
    - STEM for Auxiliary Learning
  - Model Training
    - Auxiliary Ranking Loss — KDD 2024
    - Repeated Exposure: an IPW approach
    - Real-time Online Learning
    - Exploration and Exploitation
  - Analysis Tools
    - Feature Correlation Analysis
    - Singular Spectrum Analysis
    - Interest Entanglement Analysis

Q & A

Recommender

Reference papers by our team:

- Temporal Interest Network for User Response Prediction. WWW 2024.
- On the Embedding Collapse when Scaling up Recommendation Models. ICML 2024.
- STEM: Unleashing the Power of Embeddings for Multi-task Recommendation. AAAI 2024.
- Understanding the Ranking Loss for Recommendation with Sparse User Feedback. KDD 2024.

QR of Github repo for analysis

KDD2024
BARCELONA, SPAIN