Computational Advertising: Recent Advances

Junwei Pan, Zhilin Zhang, Han Zhu

Jian Xu, Jie Jiang, Bo Zheng

Outline

- Part II, Prediction/Ranking
  - Perspectives
  - Feature Interaction
  - Sequential Models
  - Multi-Task and Multi-Domain Learning
  - Large Recommendation Models
  - LLM4Rec

- Part II, Prediction
  - **Perspectives**
  - Feature Interaction
  - Sequential Models
  - Multi-Task and Multi-Domain Learning
  - Large Recommendation Models
  - LLM4Rec

Perspectives and Off-the-shelf Tools

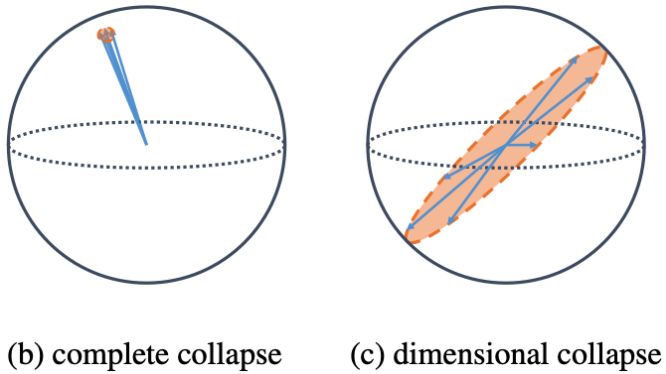
- Do the embeddings fully span the k-dimensional space?
  - **Dimensional Robustness/Collapse**
  - Tools: Singular Spectrum Analysis
- Does the model capture the correlation that we focus on?
  - **Discriminability**
  - Tools: Mutual Information Analysis
- Do the model disentangle factors in the data?
  - **Disentanglement, Redundancy Reduction**
  - Tools: Contradictory Preference Analysis, De-correlation Analysis

Perspective I: Dimensional Robustness

- Embeddings dominates the #params.
- How does the learned embedding span the k-dimensional space?
- What we want to avoid:

Dimensional Collapse

Definition: the embedding vectors end up spanning a **lower-dimensional subspace** instead of the entire available space

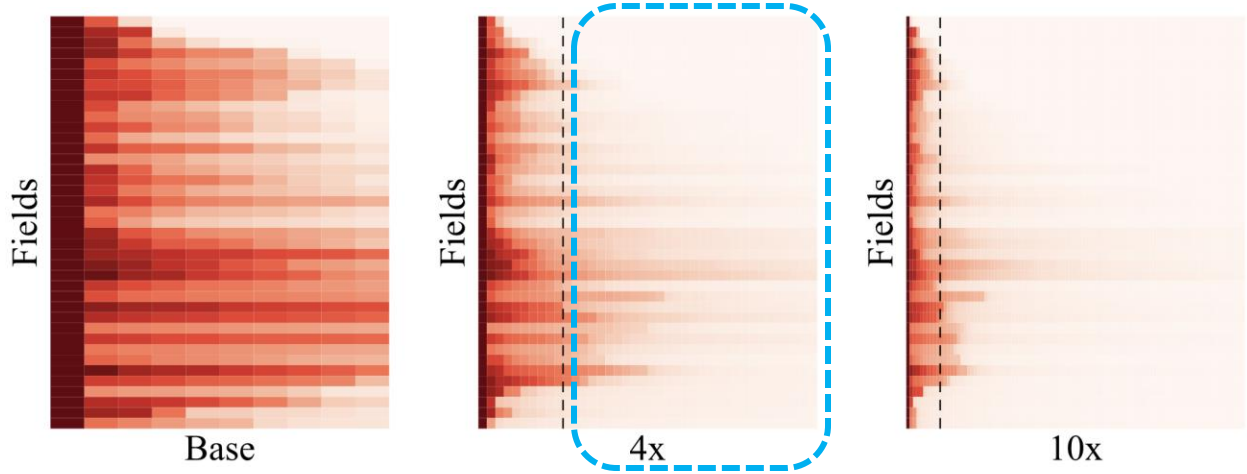


(b) complete collapse (c) dimensional collapse

Singular Spectrum Analysis

$$E = U \Sigma V^{\top}$$

E is the **embedding matrix**, it can be both sample-wise or feature-ID-wise. After **SVD**, The singular values Σ reflects the **scaling in each space dimension**.



Collapsed dimensions

Understanding Dimensional Collapse in Contrastive Self-Supervised Learning. ICLR, 2022.
On the Embedding Collapse When Scaling Up Recommendation Models. ICML 2024.

Perspective II: Discriminability

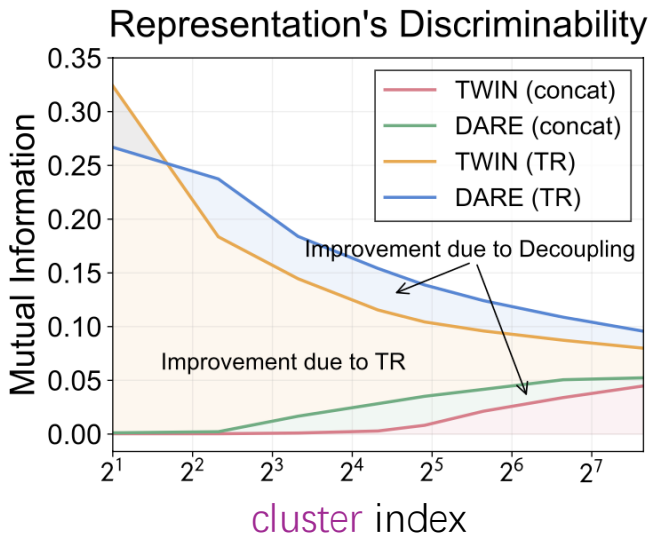
Whether the model capture **correlations** in hidden factors (e.g. layers) or explicit factors (e.g. features.)

Hidden factors

Discretized Mutual Information

$\mathrm{MI}(\mathrm{Discret}(h), Y)$

Mutual information between the discretized factors $\mathrm{Discret}(h)$, i.e., clusters, and the target Y

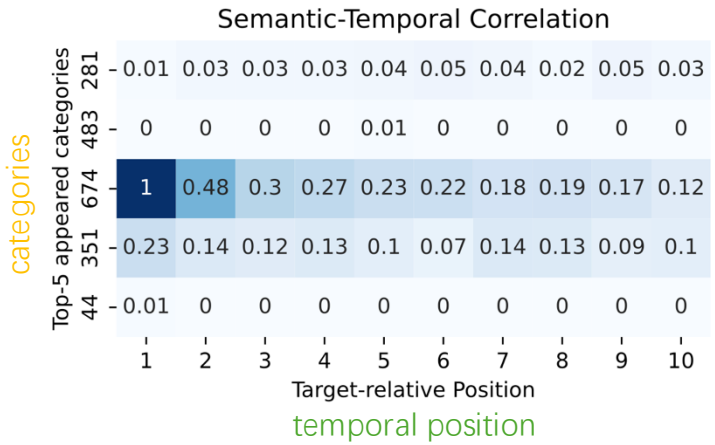


Explicit factors

Feature-wise Mutual Information

$\mathrm{MI}(X', Y)$

Mutual information between specified features, e.g., categories, or temporal position, and the target Y.



Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts. KDD 2018.
Temporal Interest Network for User Response Prediction. WWW 2024.

Tencent

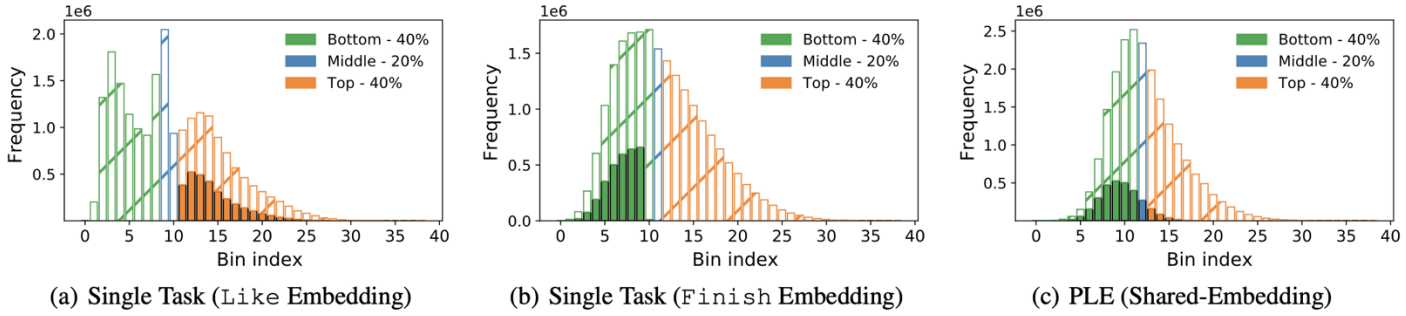
Perspective III: Disentanglement

Whether the model succeed in **decouple** **conflicting** or **redundant** interests?

For multi-task learning:

Contradictory Preference Analysis

- Pick up conflicting user-item pairs based on **single task** embeddings.
- Plot distance distribution of these user-item pairs using MTL embs.



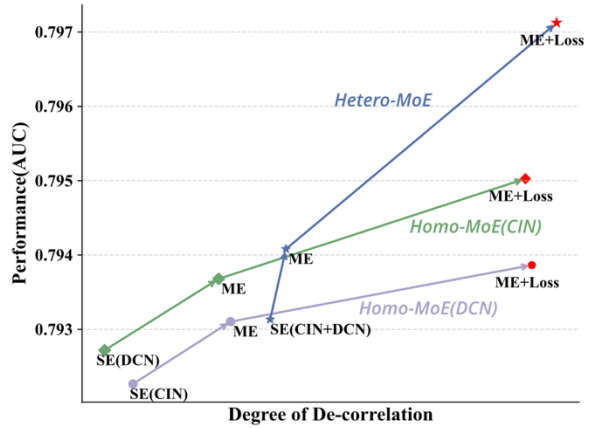
(a) Single Task (`Like` Embedding) (b) Single Task (`Finish` Embedding) (c) PLE (Shared-Embedding)

For MoE:

Inter-Expert De-Correlation

$$\sum_i \sum_j Corr(Ex_i, Ex_j)$$

The correlation between expert outputs Ex_i and Ex_j in the MoE architecture.



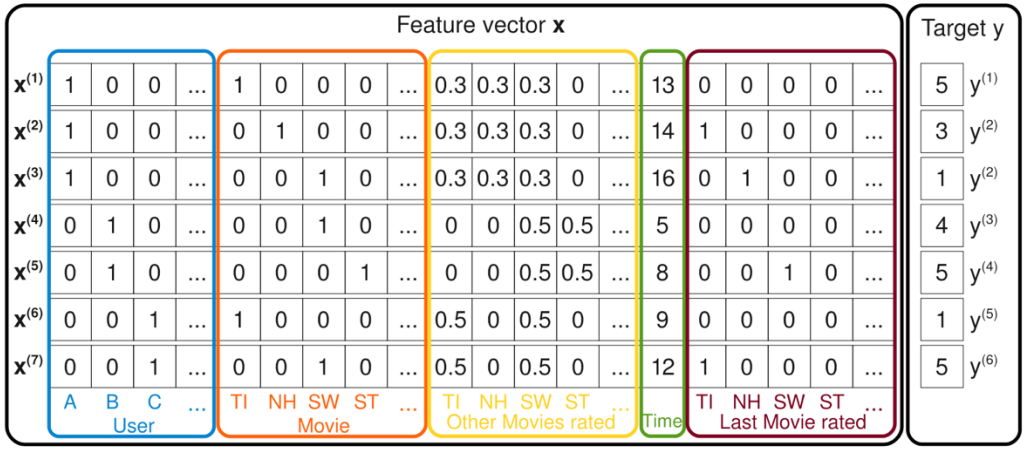
Expert De-Correlation Principle

Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts. KDD 2018.
Towards De-correlated Mixture-of-Experts for CTR Prediction. Under review.

Tencent

- Part II, Prediction
  - Perspectives
  - **Feature Interaction**
  - Sequential Models
  - Multi-Task and Multi-Domain Learning
  - Large Recommendation Models
  - LLM4Rec

FuxiCTR / BARs - Benchmark



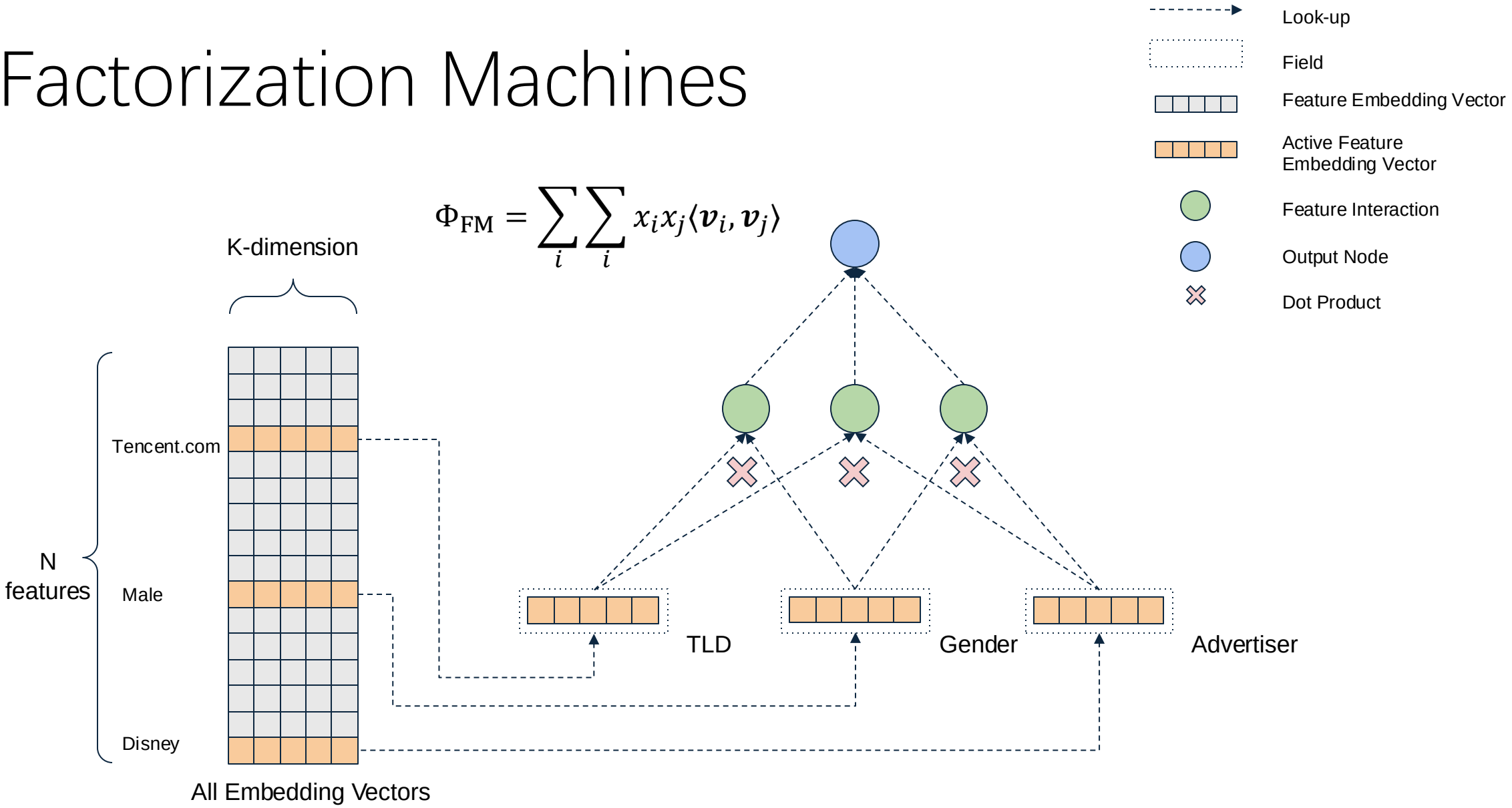
- Multi-field categorical data
- Tabular data

Year	Model	Criteo_x4		Avazu_x4		KKBox_x1	
		Logloss ($\times 10^{-2}$)	AUC(%)	Logloss ($\times 10^{-2}$)	AUC(%)	Logloss ($\times 10^{-2}$)	AUC(%)
2007	LR	45.68	79.34	38.15	77.75	57.46	76.78
2010	FM [77]	44.31	80.86	37.54	78.87	50.75	82.91
2015	CCPM [58]	44.15	81.04	37.45	78.92	50.13	83.72
2016	FFM [46]	44.07	81.13	37.20	79.31	49.74	83.76
2016	HOFM [20]	44.11	81.07	37.54	78.91	50.48	83.15
2016	PNN [75]	43.78	81.42	**37.12(4)**	**79.44(3)**	47.93	85.15
2016	DNN [27]	43.80	81.40	37.22	79.28	48.11	85.01
2016	Wide&Deep [25]	43.77	81.42	37.20	79.29	48.52	85.04
2016	DeepCrossing [84]	43.84	81.35	37.21	79.30	47.99	84.95
2017	NFM [39]	44.24	80.93	37.43	78.94	51.02	82.85
2017	AFM [112]	44.55	80.60	37.93	78.23	52.41	81.75
2017	DeepFM [37]	**43.76(3)**	**81.43(5)**	37.19	79.30	47.85	**85.31(4)**
2017	CrossNet [103]	44.56	80.60	37.79	78.40	52.83	81.16
2017	DCN [103]	**43.76(3)**	**81.44(4)**	37.19	79.31	**47.66(1)**	**85.31(4)**
2018	FwFM [70]	44.08	81.12	37.44	79.07	49.71	84.06
2018	CIN [55]	43.94	81.27	37.42	78.94	49.09	84.26
2018	xDeepFM [55]	**43.76(3)**	**81.43(5)**	37.18	79.33	**47.72(2)**	**85.35(2)**
2019	FiGNN [54]	43.83	81.38	37.36	79.15	48.96	84.72
2019	FiBiNET [43]	43.87	81.31	**37.05(2)**	**79.53(2)**	48.14	84.99
2019	AutoInt [90]	43.99	81.19	37.45	78.91	49.19	84.36
2019	AutoInt+ [90]	43.90	81.32	37.46	79.02	**47.73(3)**	**85.34(3)**
2019	HFM [98]	44.24	80.95	37.57	78.79	49.70	83.92
2019	HFM+ [98]	43.92	81.27	37.14	**79.44(3)**	**47.81(5)**	85.21
2019	FGCNN [57]	43.98	81.21	**37.11(3)**	**79.44(3)**	48.01	85.22
2020	LorentzFM [113]	44.34	80.83	37.56	78.85	51.88	82.02
2020	InterHAt [53]	44.14	81.04	37.49	78.82	48.63	84.59
2020	AFN [26]	44.02	81.15	37.40	79.07	49.10	84.26
2020	AFN+ [26]	43.84	81.38	37.26	79.29	48.42	84.89
2020	DeepIM [120]	**43.75(2)**	**81.46(2)**	**37.16(5)**	79.35	**47.75(4)**	**85.37(1)**
2020	ONN [118]	**43.72(1)**	**81.48(1)**	**36.83(1)**	**79.92(1)**	48.56	84.98
2021	FmFM [95]	43.97	81.24	37.47	79.01	49.46	84.07
2021	DCN-V2 [104]	**43.75(2)**	**81.45(3)**	37.19	79.31	47.87	**85.31(4)**

FuxiCTR: An Open Benchmark for Click-Through Rate Prediction. 2020.
BARS: Towards Open Benchmarking for Recommender Systems. SIGIR 2022.

Tencent

Factorization Machines

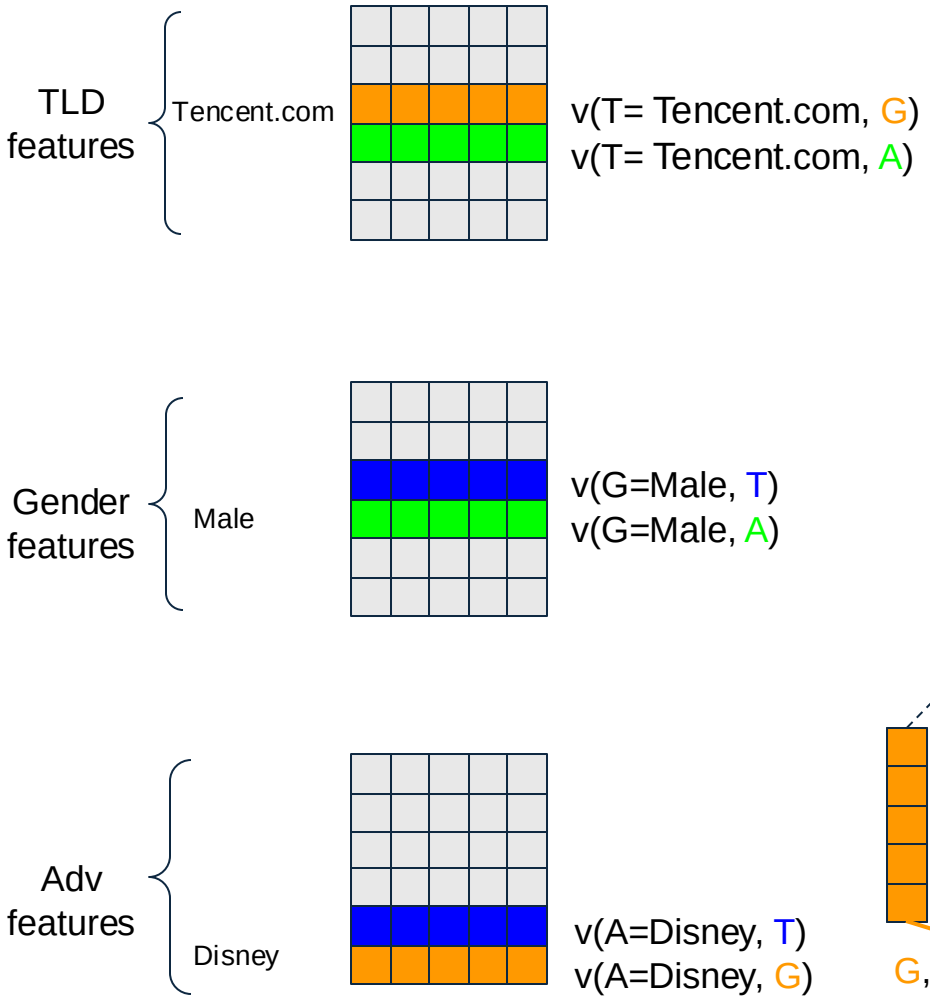


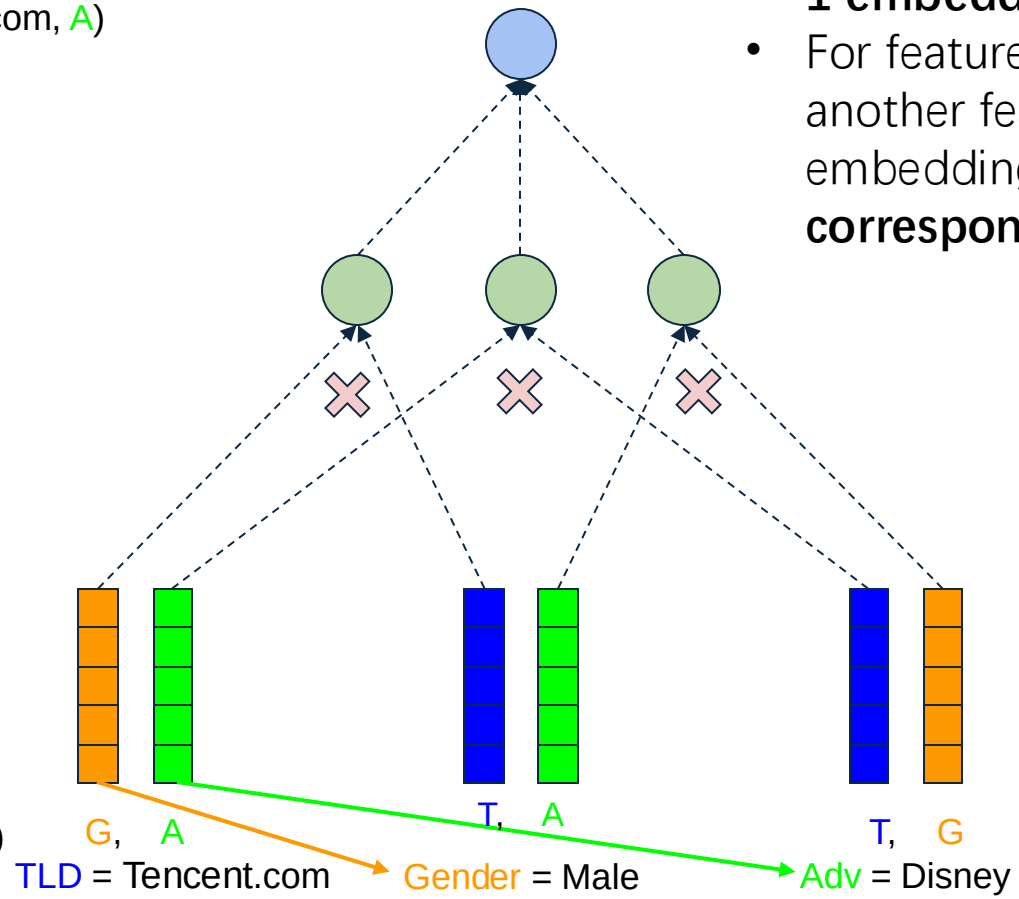
Factorization Machines. ICDM 2010.

FFM (Field-aware Factorization Machines)

$$\Phi_{\text{FFM}} = \sum_i \sum_i x_i x_j \langle \boldsymbol{v}_{i,F(j)}, \boldsymbol{v}_{j,F(i)} \rangle$$

- For a dataset with M fields, FFM learns **M-1 embeddings** for each feature.
- For feature i, when being interacted with another feature j, among the M-1 embeddings, it chooses **the one corresponding to the field of j**, i.e., $\boldsymbol{v}_{i,F(j)}$





All Embedding Vectors

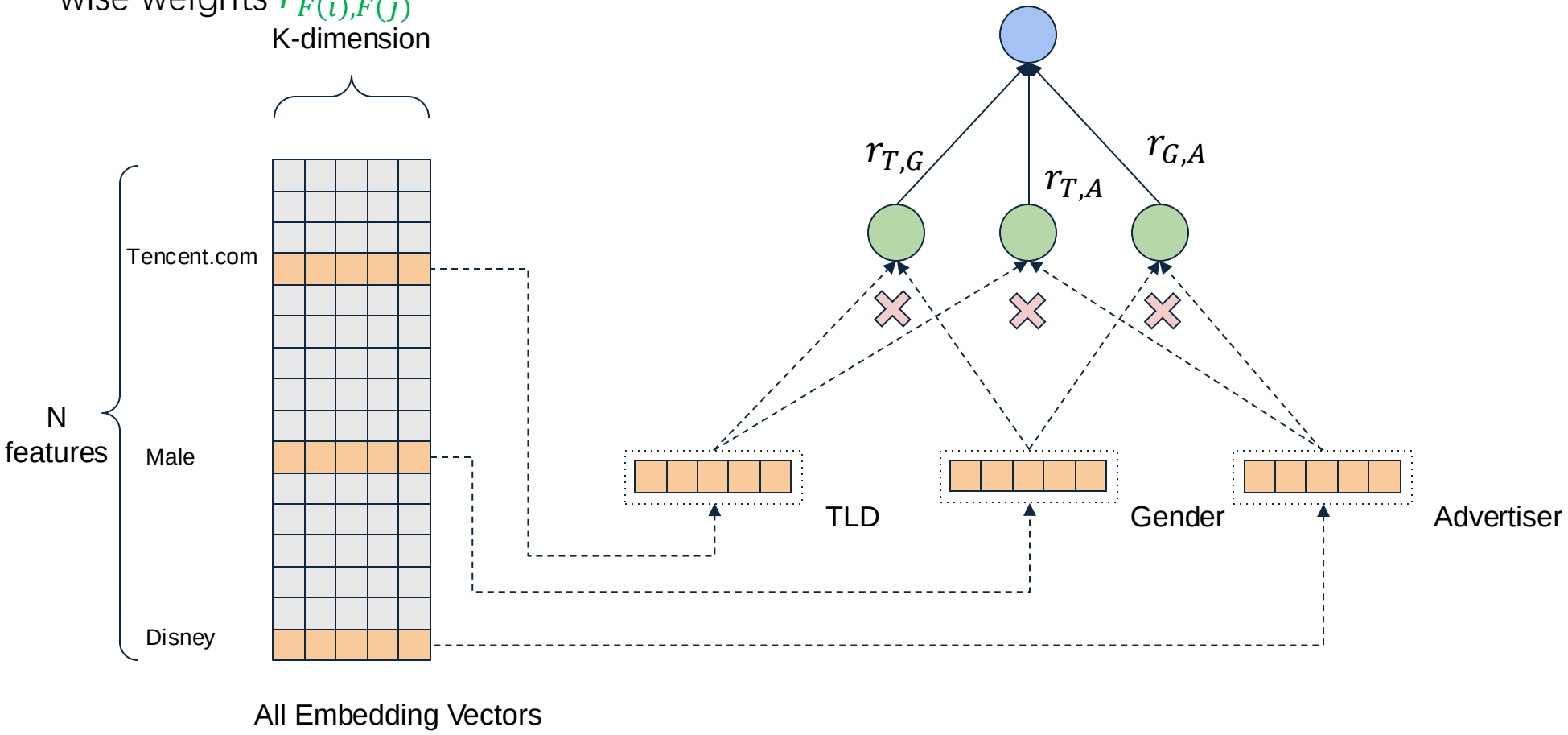
Field-aware Factorization Machines. RecSys 2016.

Tencent

FwFM (Field-weighted Factorization Machines)

FFM is field-aware by using field-specific embeddings within interaction. Another efficient way is to assign field-wise weights $r_{F(i),F(j)}$

$$\Phi_{\text{FwFM}} = \sum_i \sum_i x_i x_j \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle r_{F(i),F(j)}$$

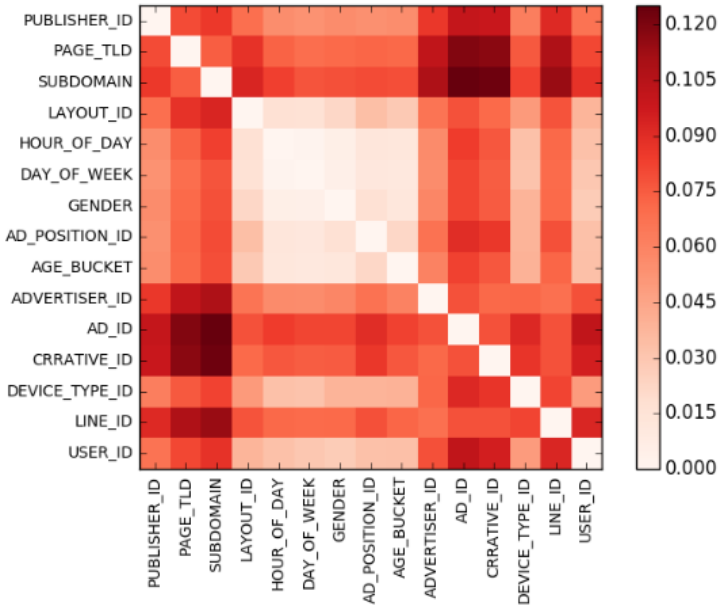


Field-weighted Factorization Machines for Click-Through Rate Prediction in Display Advertising. WWW 2018.

FwFM – A Discriminability Perspective

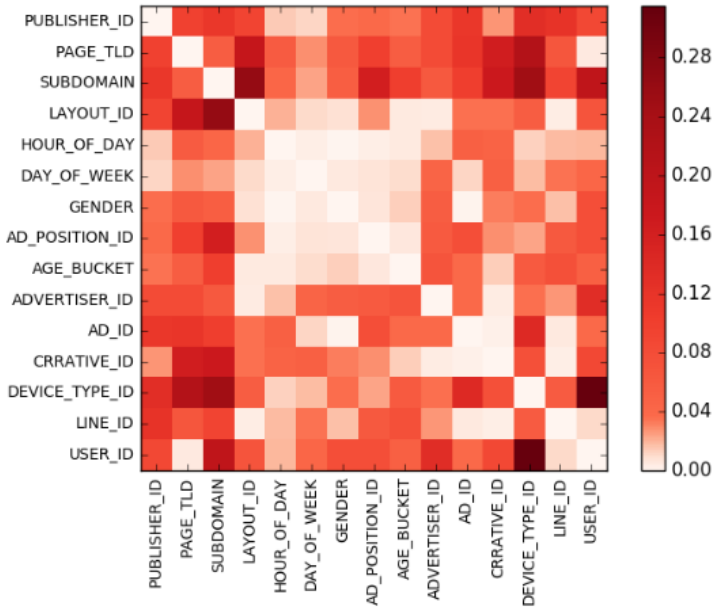
$$\Phi_{\text{FwFM}} = \sum_i \sum_i x_i x_j \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle r_{F(i),F(j)}$$

Ground-truth field-wise mutual information.



(a) Heat map of mutual informations between field pairs and labels.

Learned field-wise correlation.

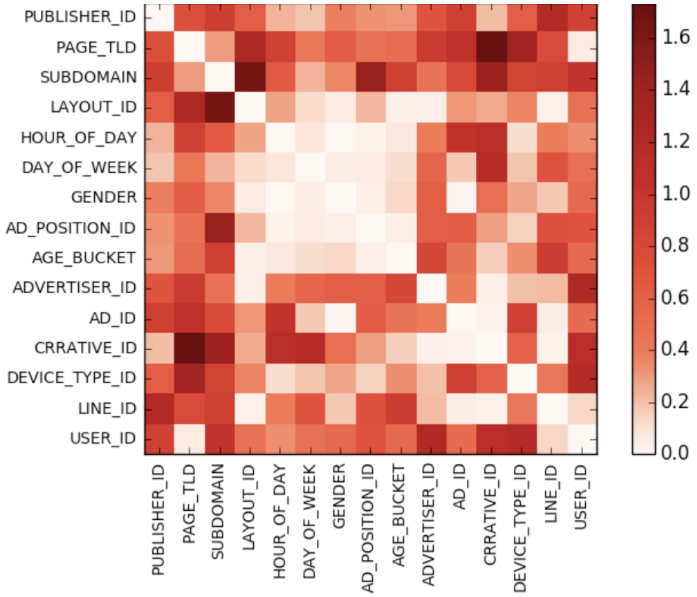


(d) Heat map of learned field interaction strengths of FwFMs. Pearson correlation coefficient with mutual informations: 0.4271.

$$\frac{\sum_{(i,j)\in(F_k,F_l)} I(i,j) \cdot \#(i,j)}{\sum_{(i,j)\in(F_k,F_l)} \#(i,j)}$$

$$I(i,j) = |\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle \cdot r_{F_k,F_l}|$$

Field-pair weights $|r_{F(i),F(j)}|$



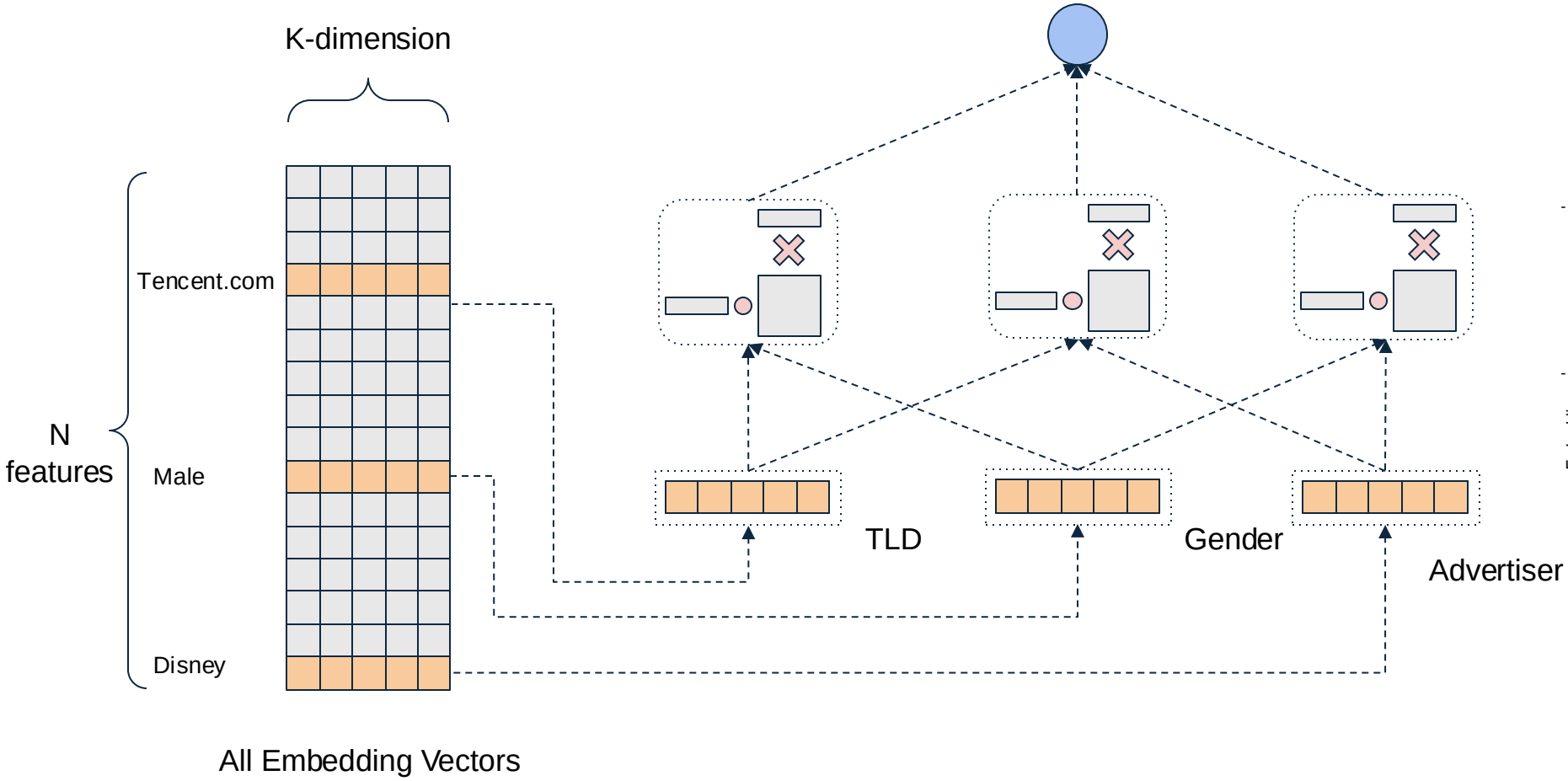
(a) Heat map of field interaction weight r_{F_k,F_l}. Its Pearson correlation coefficient with mutual information is 0.5554.

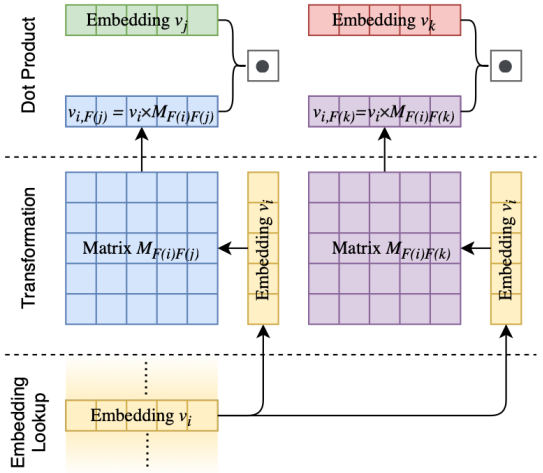
Field-weighted Factorization Machines for Click-Through Rate Prediction in Display Advertising. WWW 2018.

Tencent

FmFM (Field-matrixed Factorization Machines) $\Phi_{\mathrm{FmFM}} = \sum_i \sum_i x_i x_j \, (\boldsymbol{v}_i \, M_{F(i),F(j)}) \odot \boldsymbol{v}_j$

Involve field information with field-pair-wise projection matrix $M_{F(i),F(j)}$ within the interaction.





FmFM: Field-matrixed Factorization Machines for CTR Prediction. WWW 2021.

Tencent

Feature Interaction Evolution

-A Dimensional Collapse Perspective

With performance enhances from FM, FwFM to FmFM, the **projection matrix within interaction** also evolves, from identity matrix, scaled identity matrix to full matrix.

Identity matrix

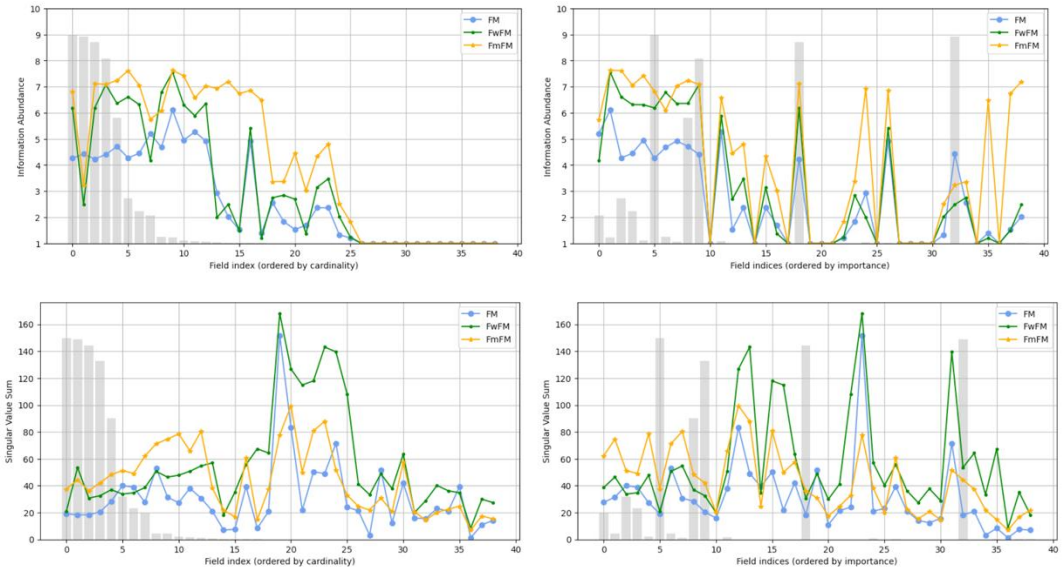
FM: $\langle v_i, v_j \rangle = \langle v_i I, v_j \rangle$

Scaled identity matrix

FwFM: $\langle v_i, v_j \rangle r_{F(i),F(j)} = \langle v_i I \cdot r_{F(i),F(j)}, v_j \rangle$

Full matrix

FmFM: $\langle v_i M_{F(i),F(j)}, v_j \rangle$



There is a trend of more dimensional robust embeddings from FM, FwFM to FmFM

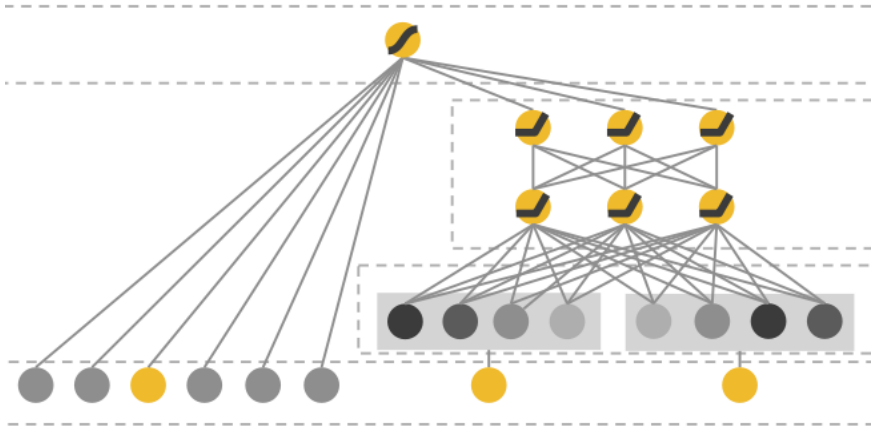
Among the 2nd-order interaction models, the more complicated the projection matrix (i.e., from identity, scaled identity, to full matrix) within the feature Interaction Function, the more robust (less collapsed) the learned embeddings regarding both information abundance and singular value sum.

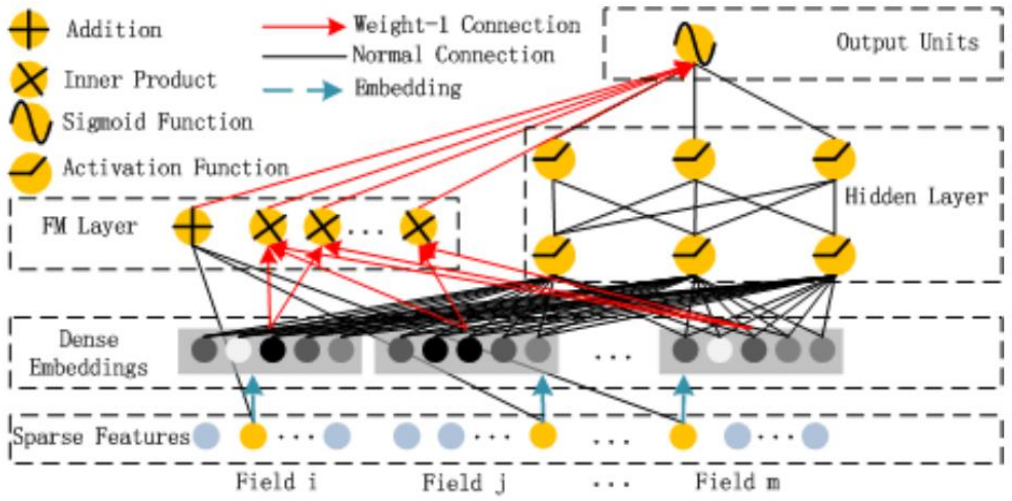
Towards Unifying Feature Interaction Models for Click-Through Rate Prediction. 2024.

Wide & Deep, DeepFM

Follow a Wide & Deep framework

- Wide: 2nd-order explicit interactions.
- Deep: implicit high-order interactions.





Wide & Deep

- Wide: Logistic Regression
- Deep: DNN

DeepFM

- Wide: FM
- Deep: DNN

Wide & deep learning for recommender systems. 2016.
DeepFM: A Factorization-Machine based Neural Network for CTR Prediction. 2017.

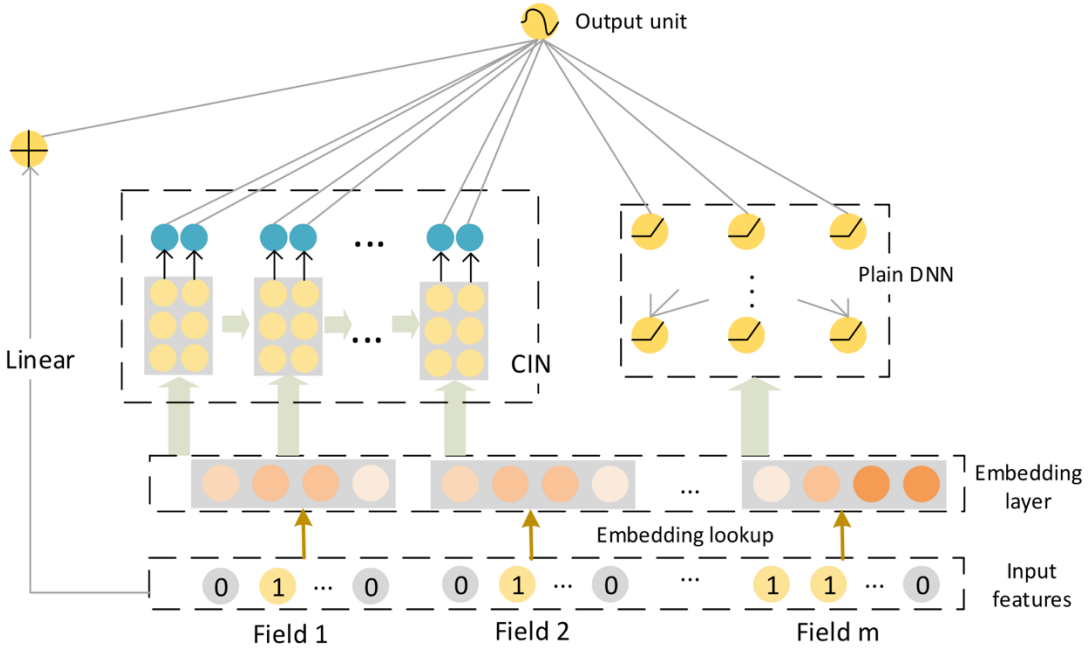
xDeepFM

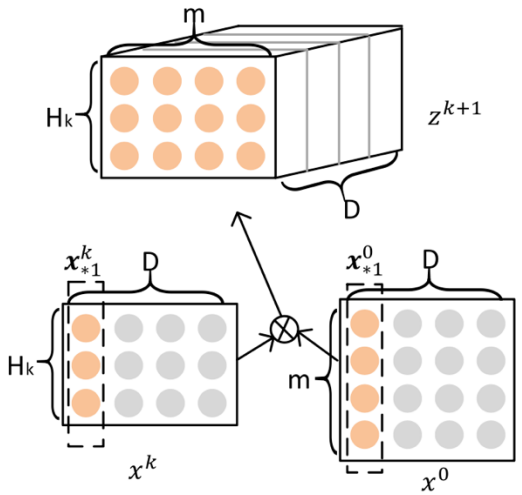
Follow a Explicit and Implicit framework

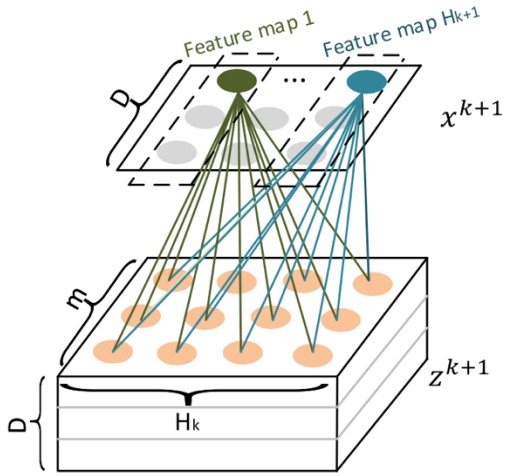
- Explicit: Compressed Interaction Network (CIN).
- Implicit: DNN.

$$\mathbf{x}_k = \mathbf{x}_0 \mathbf{x}_{k-1}^T \mathbf{w}_k + \mathbf{b}_k + \mathbf{x}_{k-1}$$

field-pair-wise weights, same as the weights $r_{F(i),F(j)}$ in FwFM





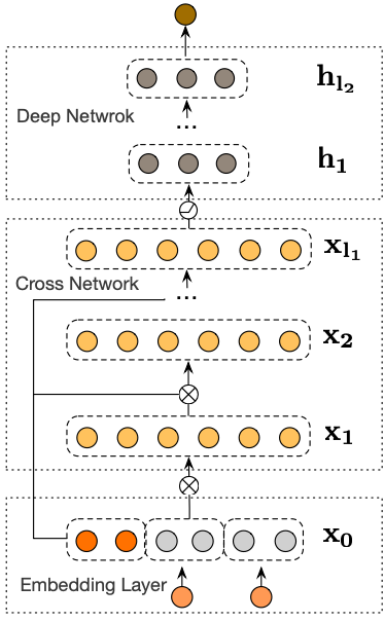


xDeepFM: Combining Explicit and Implicit Feature Interactions for Recommender Systems. KDD 2018.

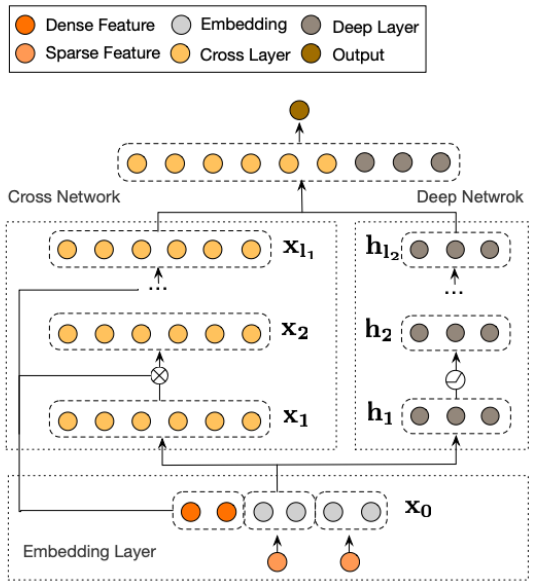
Tencent

DCN V2

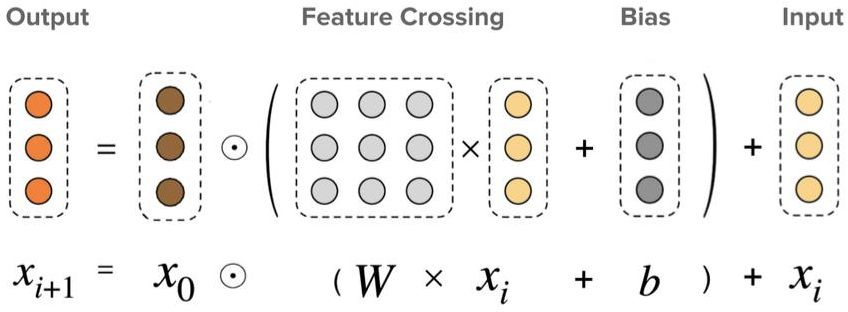
- Explicit: CrossNet
- Implicit: DNN.

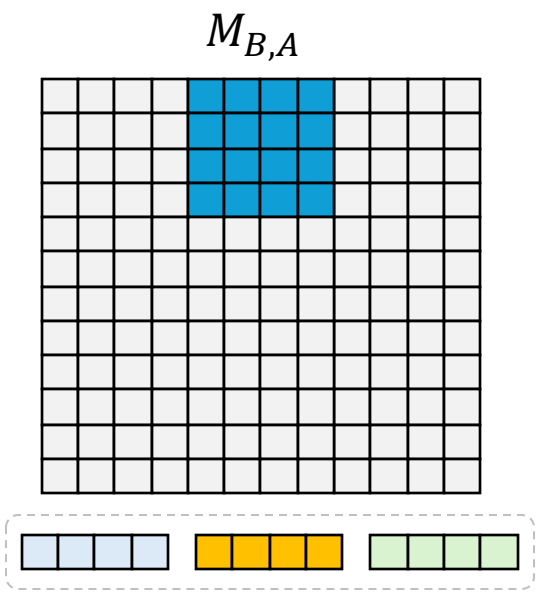


(a) Stacked

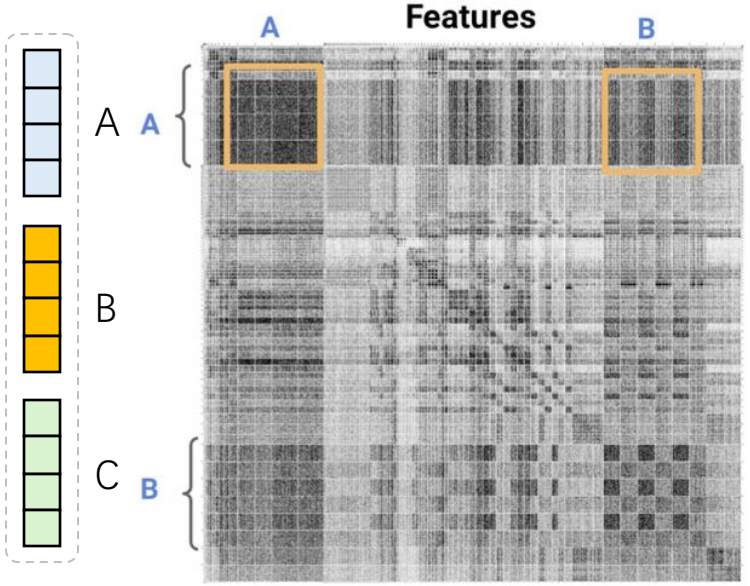


(b) Parallel





field-pair-wise matrix,
similar as the matrix in FmFM



F-norm of M_{F_k,F_l}

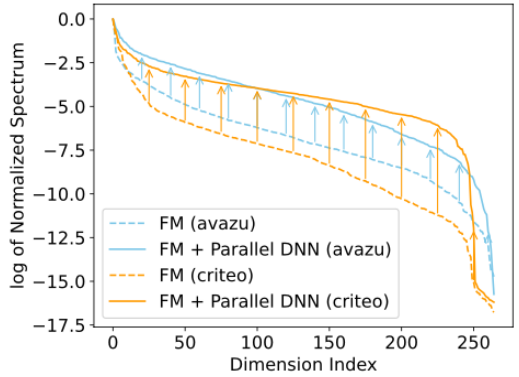
DCN v2: Improved Deep & Cross Network and Practical Lessons for Web-scale Learning to Rank Systems.

Tencent

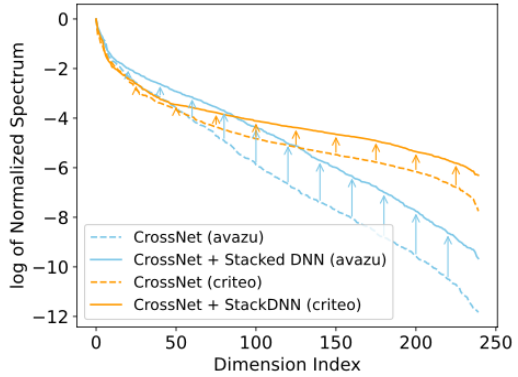
Role of DNN

- Why hybrid models work better?
- Does DNN really learn implicit interaction?
- Rendle proves that **DNN is hard to learn dot products**.

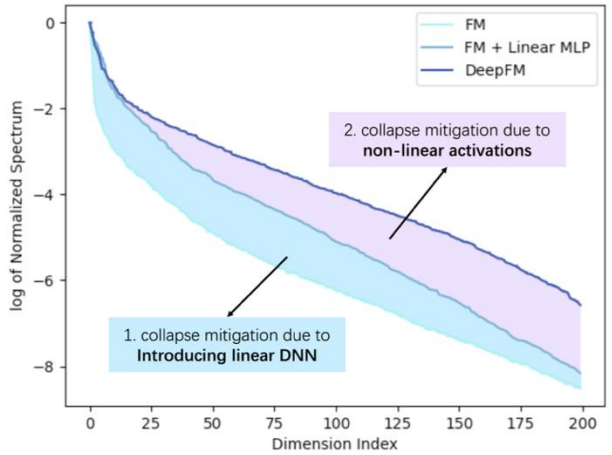
The **parallel and stacked DNN** mitigates the dimensional collapse of embeddings in recommendation models.



(a) Parallel DNN



(b) Stacked DNN



Both **linear DNN** and **non-linear activations** can mitigate **dimensional collapse**.

Neural Collaborative Filtering vs. Matrix Factorization Revisited. RecSys 2020.
Understanding DNNs in Feature Interaction Models: A Dimensional Collapse Perspective. Under review.

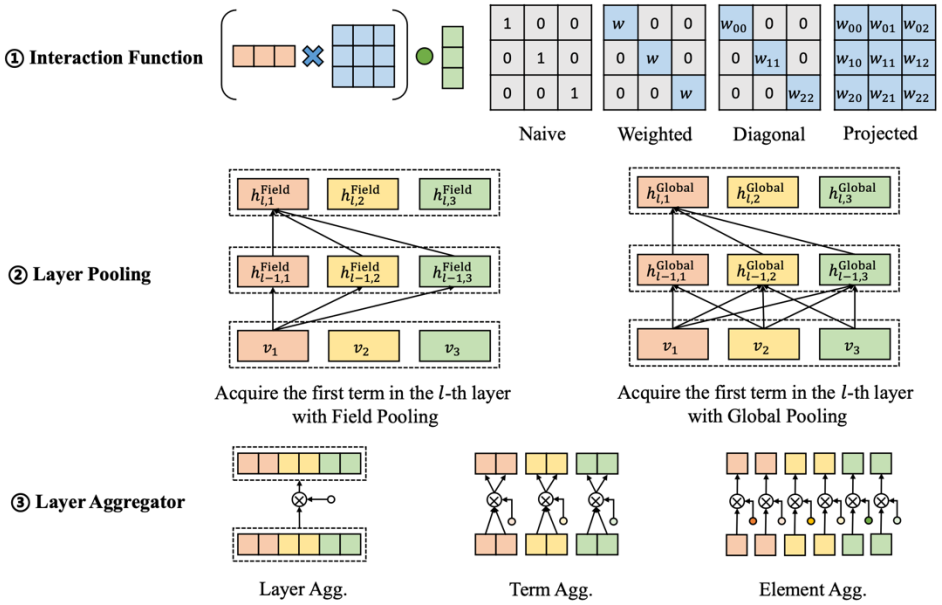
Tencent

IPA: Unified Framework

A unified framework with three components can summarize most existing feature interaction models:

- Feature interaction function: $\langle v_i I, v_j \rangle$, $\langle v_i I \cdot r_{F(i),F(j)}, v_j \rangle$, $\langle v_i\, D(\{\sigma_i\}), v_j \rangle$, $\langle v_i M_{F(i),F(j)}, v_j \rangle$
- Layer pooling: **Field Pooling**, **Global Pooling**
- Layer aggregation: **direct/layer/term/element**-aggregation

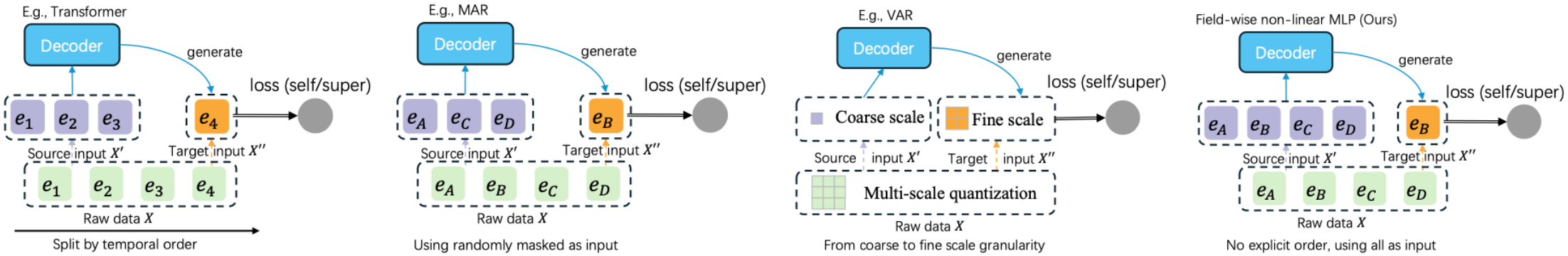
Model	Interaction Function	Layer Pooling	Layer Aggregator	Criteo			Avazu		
				L	AUC	Logloss	O	AUC	Logloss
FM	naive	field/global	direct	2	0.8009(2e-4)	0.4507(4e-4)	2	0.7758(1e-4)	0.3821(2e-4)
DeepFM	naive	field/global	direct	2	0.8122(1e-4)	0.4399(2e-4)	2	0.7899(5e-4)	0.3741(5e-4)
HOFM	naive	field/global	direct	4	0.8040(3e-4)	0.4479(3e-4)	5	0.7781(8e-4)	0.3807(6e-4)
FwFM	weighted	field/global	direct	2	0.8095(2e-4)	0.4423(3e-4)	2	0.7854(4e-4)	0.3768(4e-4)
xDeepFM	weighted	global	term	4	0.8119(2e-4)	0.4401(2e-4)	5	0.7897(7e-4)	0.3741(8e-4)
FvFM	diagonal	field/global	direct	2	0.8103(2e-4)	0.4415(5e-3)	2	0.7870(3e-4)	0.3754(4e-4)
FmFM	projected	field/global	direct	2	0.8115(3e-4)	0.4403(3e-4)	2	0.7882(5e-4)	0.3750(3e-4)
FiBiNet	projected	field/global	direct	2	0.8113(2e-4)	0.4405(2e-4)	2	0.7907(4e-4)	0.3738(5e-4)
DCN V2	projected	field'	direct	4	0.8137(3e-4)	0.4384(4e-4)	5	**0.7917(1e-4)**	**0.3729(4e-4)**
WFL	weighted	field	layer	4	0.8124(2e-4)	0.4394(2e-4)	5	0.7891(3e-4)	0.3746(5e-4)
DFL	diagonal	field	layer	4	0.8123(1e-4)	0.4395(2e-4)	5	0.7903(9e-4)	0.3740(7e-4)
PFL	projected	field	layer	4	**0.8138(3e-4)**	**0.4381(4e-4)**	5	0.7916(4e-4)	0.3731(4e-4)
PFT	projected	field	term	4	**0.8138(3e-4)**	**0.4381(3e-4)**	5	0.7904(4e-4)	0.3738(6e-4)
PFE	projected	field	element	4	**0.8138(3e-4)**	**0.4381(4e-4)**	5	0.7907(2e-4)	0.3735(3e-4)
PFD	projected	field	direct	4	0.8131(4e-4)	0.4388(4e-4)	5	0.7912(5e-4)	0.3732(4e-4)



Towards Unifying Feature Interaction Models for Click-Through Rate Prediction. Under review.

Feature Generation

- All existing feature interaction models are under the **discriminative paradigm** that models P(Y|X).
- We aim to model P(X). What's P(X) in feature interaction?
- **Predict each feature embedding** based on a **decoder network** over the whole input



(a) Sequential generation (b) Masked generation (c) Coarse-to-fine generation (d) Feature generation

Supervised Feature Generation framework for CTR models, shifting from the **discriminative "feature interaction"** paradigm to the **generative "feature generation"** paradigm.

From Feature Interaction to Feature Generation: A Generative Paradigm of CTR Prediction Models. Under review

Feature Interaction v.s. Feature Generation

$$\Phi_{\text{FM}}^{\text{DIS}} = \sum_i \sum_i x_i x_j (\boldsymbol{v}_i \odot \boldsymbol{v}_j)$$

interaction

$$\Phi_{\text{FM}}^{\text{GEN}} = \sum_i \sum_i x_i x_j (\boldsymbol{\sigma}([\boldsymbol{v}] \cdot W_{F(i)}) \odot \boldsymbol{v}_j)$$

decoder

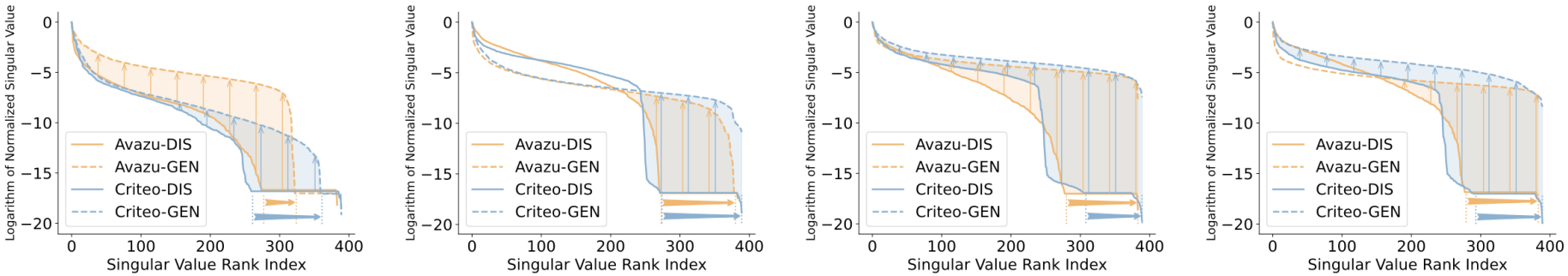
generation

Model	Discriminative	Generative
FM	$\sum_{i,j=1}^{N} \boldsymbol{v}_j \odot \boldsymbol{v}_i$	$\sum_{i,j=1}^{N} \boldsymbol{\sigma}([\boldsymbol{v}] \cdot W_{F(j)}) \odot \boldsymbol{v}_i$
FmFM	$\sum_{i,j=1}^{N} \boldsymbol{v}_j \odot [\boldsymbol{v}_i \cdot M_{F(i)\to F(j)}]$	$\sum_{i,j=1}^{N} \boldsymbol{\sigma}([\boldsymbol{v}] \cdot W_{F(j)}) \odot [\boldsymbol{v}_i \cdot M_{F(i)\to F(j)}]$
CrossNet V2	$\sum_{l=1}^{L} \sum_{i,j=1}^{N} \boldsymbol{v}_j^0 \odot (\boldsymbol{v}_i^l \cdot M_{F(i)\to F(j)}^l)$	$\sum_{l=1}^{L} \sum_{i,j=1}^{N} \boldsymbol{\sigma}([\boldsymbol{v}]^l \cdot W_{F(j)}^l) \odot (\boldsymbol{v}_i^l \cdot M_{F(i)\to F(j)}^l)$
DeepFM	$\sum_{i,j=1}^{N} \boldsymbol{v}_j \odot \boldsymbol{v}_i + \text{DNN}([\boldsymbol{v}])$	$\sum_{i,j=1}^{N} \boldsymbol{\sigma}([\boldsymbol{v}] \cdot W_{F(j)}) \odot \boldsymbol{v}_i$
xDeepFM	$\sum_{l=1}^{L} \sum_{i,j=1}^{N} \text{Conv}^l(\boldsymbol{v}_j^0 \odot \boldsymbol{v}_i^l) + \text{DNN}([\boldsymbol{v}])$	$\sum_{l=1}^{L} \sum_{i,j=1}^{N} \text{Conv}^l(\boldsymbol{\sigma}([\boldsymbol{v}]^l \cdot W_{F(j)}^l) \odot \boldsymbol{v}_i^l) + \text{DNN}([\boldsymbol{v}])$
IPNN	$\text{DNN}([[\boldsymbol{v}], \sum_{i,j=1}^{N} \boldsymbol{v}_j \odot \boldsymbol{v}_i])$	$\text{DNN}([[\boldsymbol{v}], \sum_{i,j=1}^{N} \boldsymbol{\sigma}([\boldsymbol{v}] \cdot W_{F(j)}) \odot \boldsymbol{v}_i])$
DCN V2	$\sum_{l=1}^{L} \sum_{i,j=1}^{N} \boldsymbol{v}_j^0 \odot (\boldsymbol{v}_i^l \cdot M_{F(i)\to F(j)}^l) + \text{DNN}([\boldsymbol{v}])$	$\sum_{l=1}^{L} \sum_{i,j=1}^{N} \boldsymbol{\sigma}([\boldsymbol{v}]^l \cdot W_{F(j)}^l) \odot (\boldsymbol{v}_i^l \cdot M_{F(i)\to F(j)}^l) + \text{DNN}([\boldsymbol{v}])$

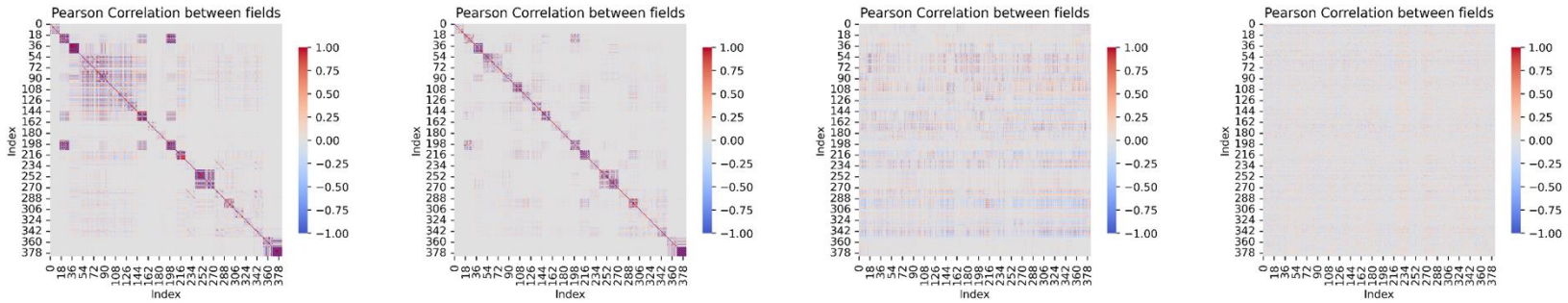
From Feature Interaction to Feature Generation: A Generative Paradigm of CTR Prediction Models. Under review.

New Perspectives: Dimensional Collapse and Redundancy Reduction.

- **Similar work** treat it as gating/mask: FibiNet, MaskNet, Final, FinalMLP, PEPNet.
- We explain it's effectiveness from the following perspectives:
  - Dimensional Collapse: Singular Spectrum
  - Redundancy Reduction: De-Correlation between Fields



(a) FM (b) DeepFM (c) CrossNet (d) DCN V2

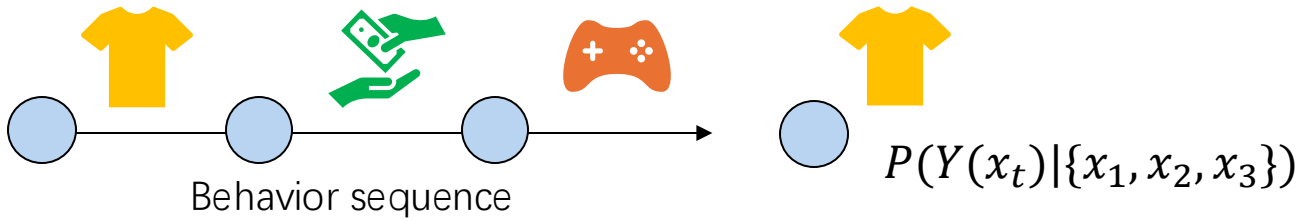


(a) FM (DIS) (b) DeepFM (DIS) (c) DCN V2 (DIS) (d) DCN V2 (GEN)

From Feature Interaction to Feature Generation: A Generative Paradigm of CTR Prediction Models. Under review.

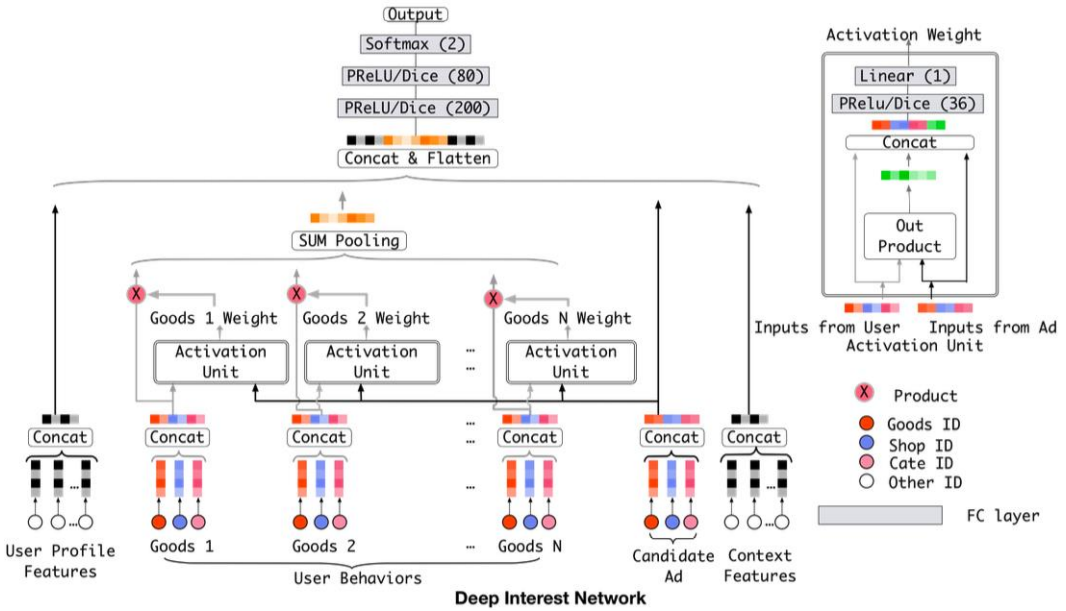
- Part II, Prediction
  - Perspectives
  - Feature Interaction
  - **Sequential Models**
  - Multi-Task and Multi-Domain Learning
  - Large Recommendation Models
  - LLM4Rec

DIN



In prediction, the most import sequential correlation is the behavior-target correlation.
DIN: **Target Attention**, captures the correlation between each **behavior** and the **target**.

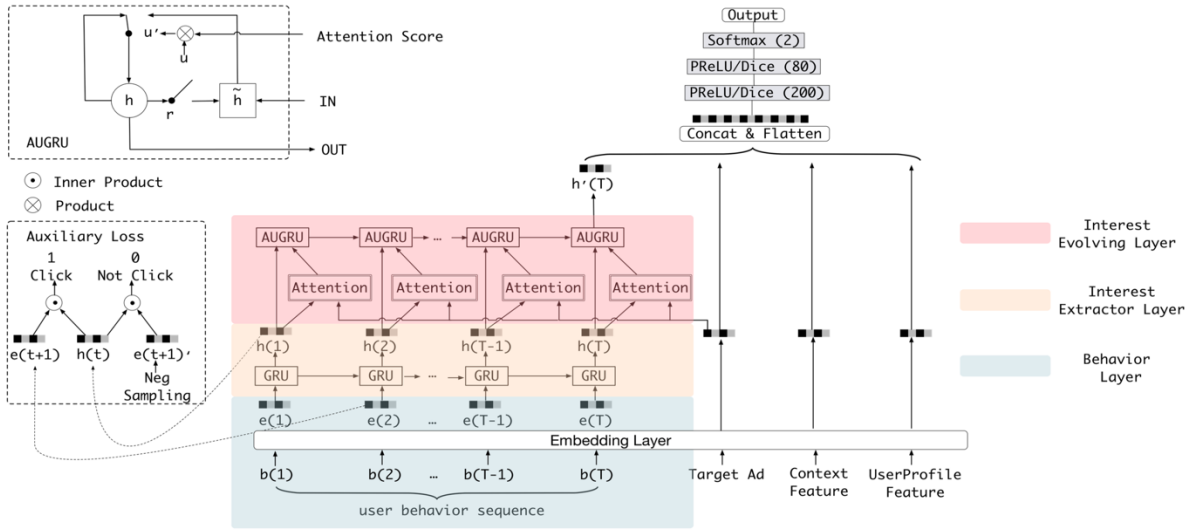
$$\alpha(e_i, v_t) \cdot e_i$$

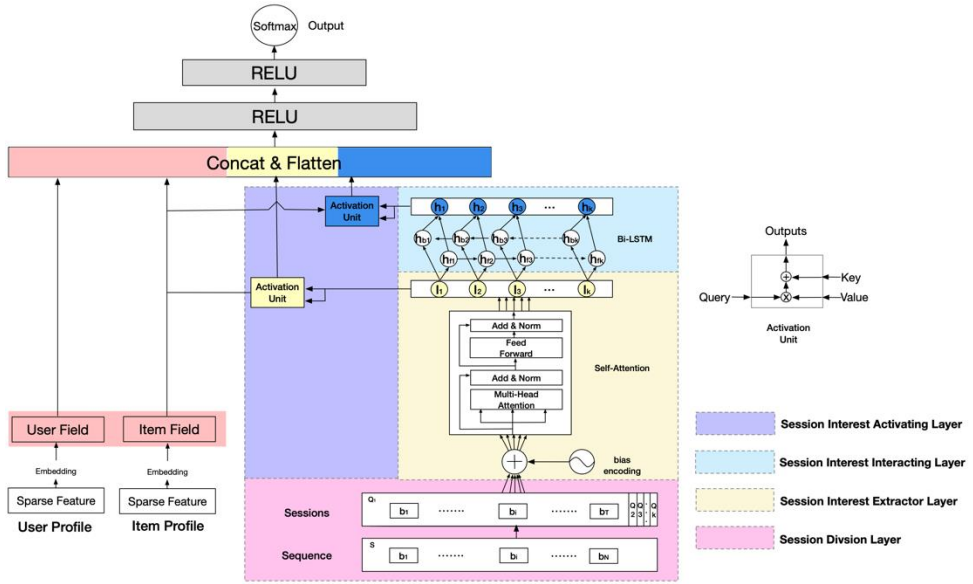


Deep Interest Network for Click-Through Rate Prediction. KDD 2018.

DIEN, DSIN

DIEN: **GRU** to capture the temporal correlations
DSIN: **LSTM** to capture correlations within and between session.





Deep interest evolution network for click-through rate prediction. AAAI 2019.
Deep Session Interest Network for Click-Through Rate Prediction. 2019.

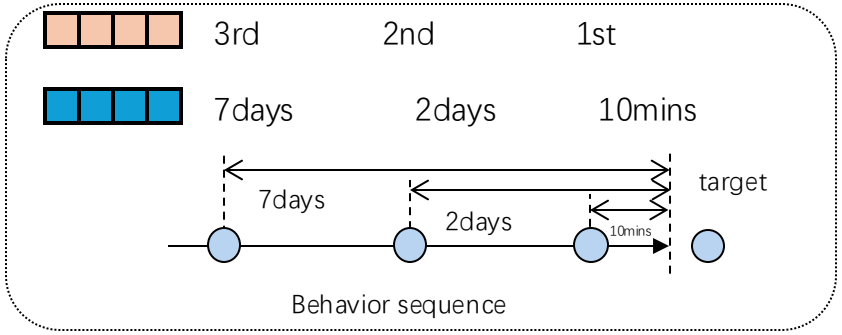
Tencent

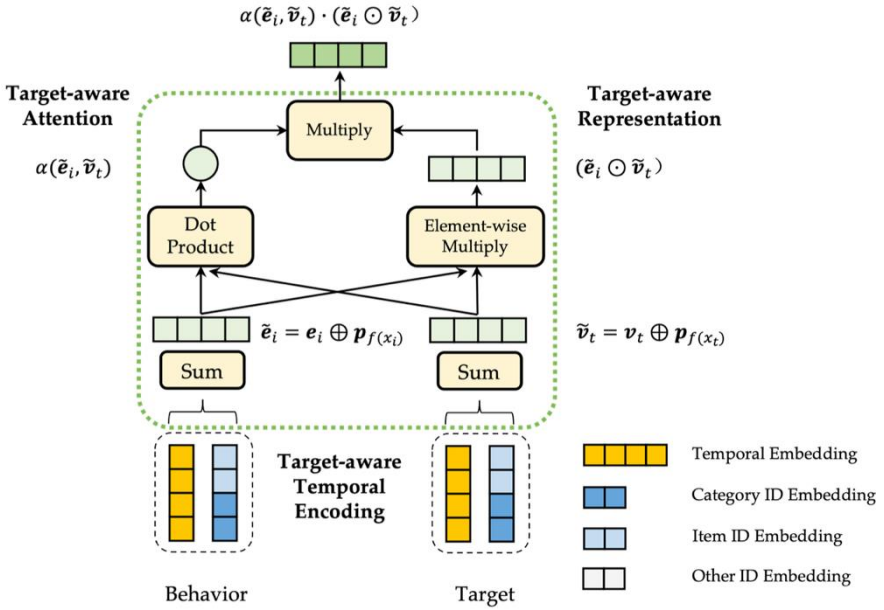
TIN (Temporal Interest Network)

$$\sum_{X_i \in \mathcal{H}} \underbrace{\alpha(\tilde{\boldsymbol{e}}_i, \tilde{\boldsymbol{v}}_t)}_{\text{TA}} \cdot \underbrace{(\tilde{\boldsymbol{e}}_i \odot \tilde{\boldsymbol{v}}_t)}_{\text{TR}}$$

A **Transformer variant** for sequential ranking
Three critical components to capture the temporal-semantic correlation

- **Target-aware Temporal Encoding (TTE)** to capture the temporal correlations: target-relative positions, target-relative time intervals
- **Target Attention (TA)**
- **Target-aware representation (TR)** to enhance the discriminability.



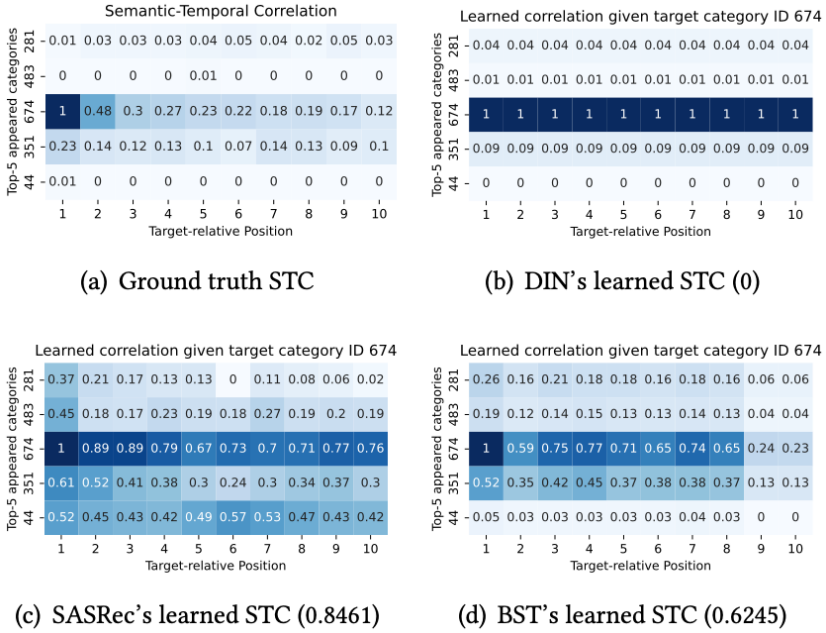


Temporal Interest Network for User Response Prediction. WWW 2024.

TIN – Temporal-Semantic Correlation Analysis

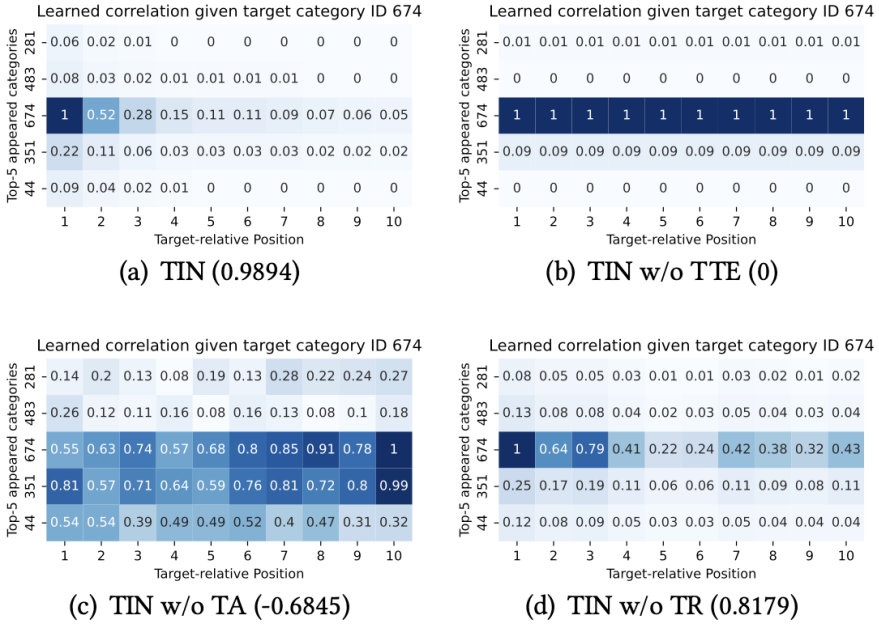
$$\sum_{X_i \in \mathcal{H}} \underbrace{\alpha(\tilde{\boldsymbol{e}}_i, \tilde{\boldsymbol{v}}_t)}_{\text{TA}} \cdot \underbrace{(\tilde{\boldsymbol{e}}_i \odot \tilde{\boldsymbol{v}}_t)}_{\text{TR}}$$

Mutual Information at category and position level



categories

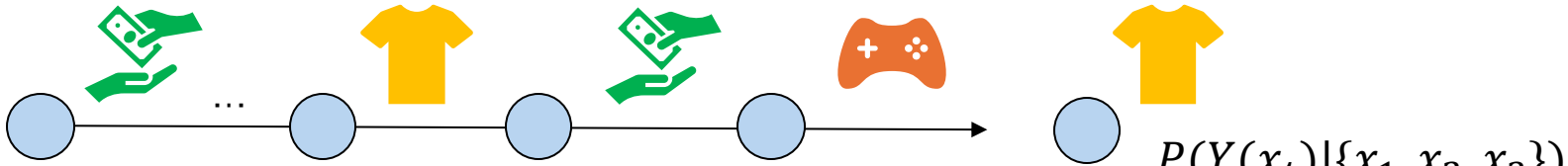
temporal position



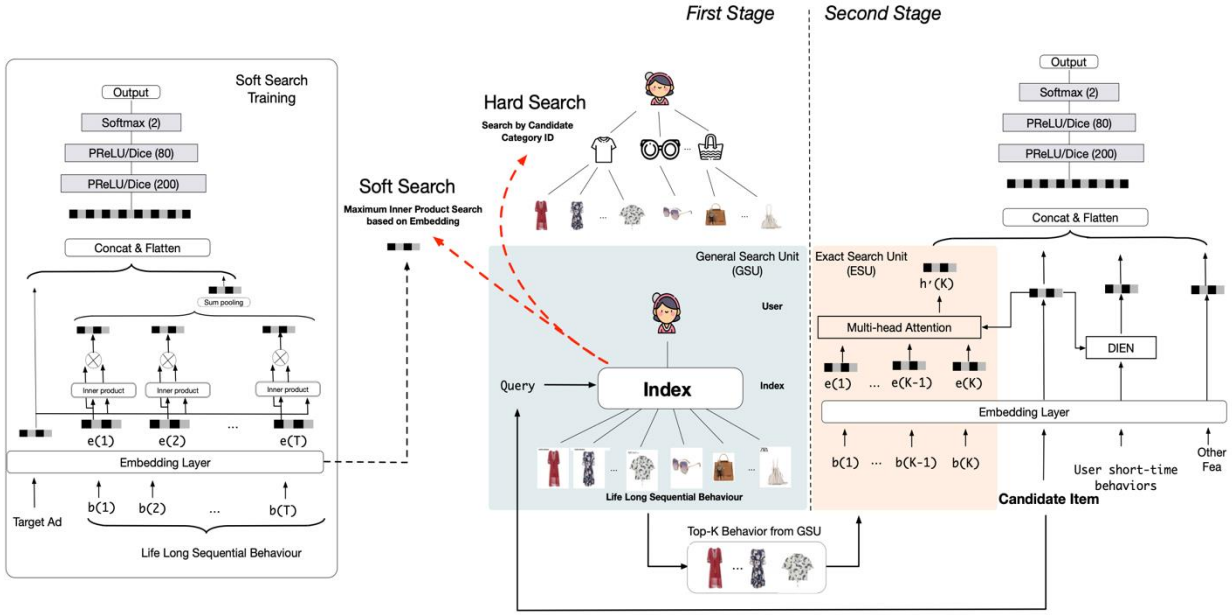
Every component is critical. Ablating any of them makes the model unable to capture the temporal-semantic correlation.

Temporal Interest Network for User Response Prediction. WWW 2024.

SIM



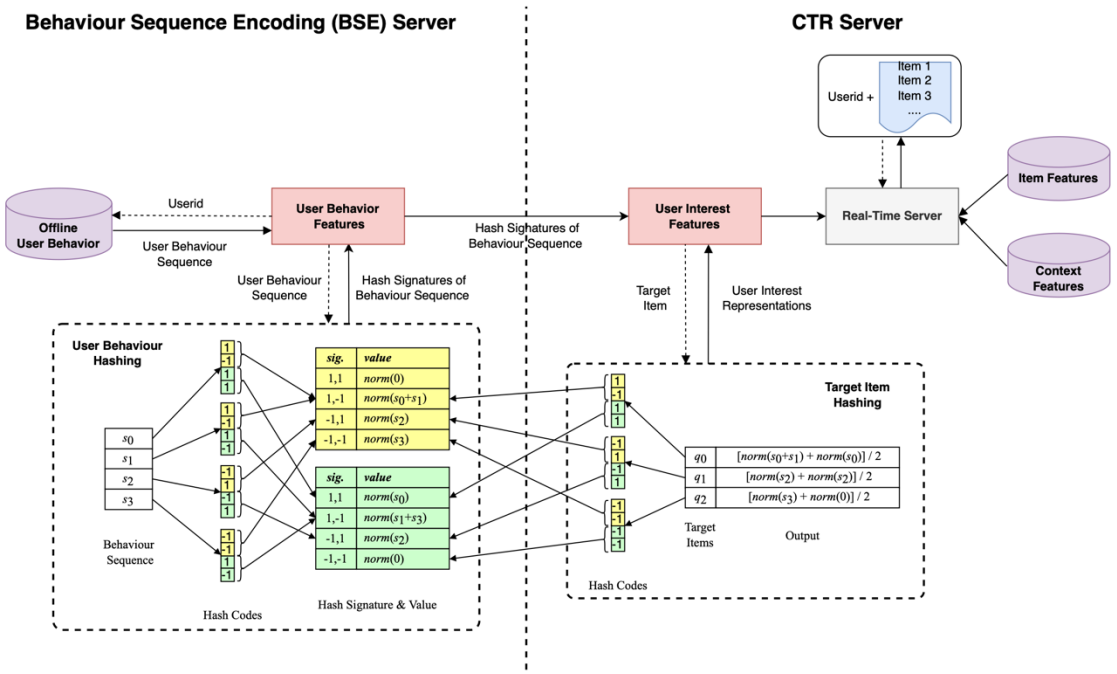
- User's behavior sequence with length of 1K ~ 1M, making **target-attention** infeasible
- Two-stage design for efficient user interest modeling
  - Stage 1: General Search Unit (**GSU**): retrieve the most informative behaviors
    - **Hard Search**: Same category,.
    - **Soft Search**: Attention score.
  - Stage 2: Exact Search Unit (**ESU**): model user interest on the retrieved behaviors, e.g., DIN.

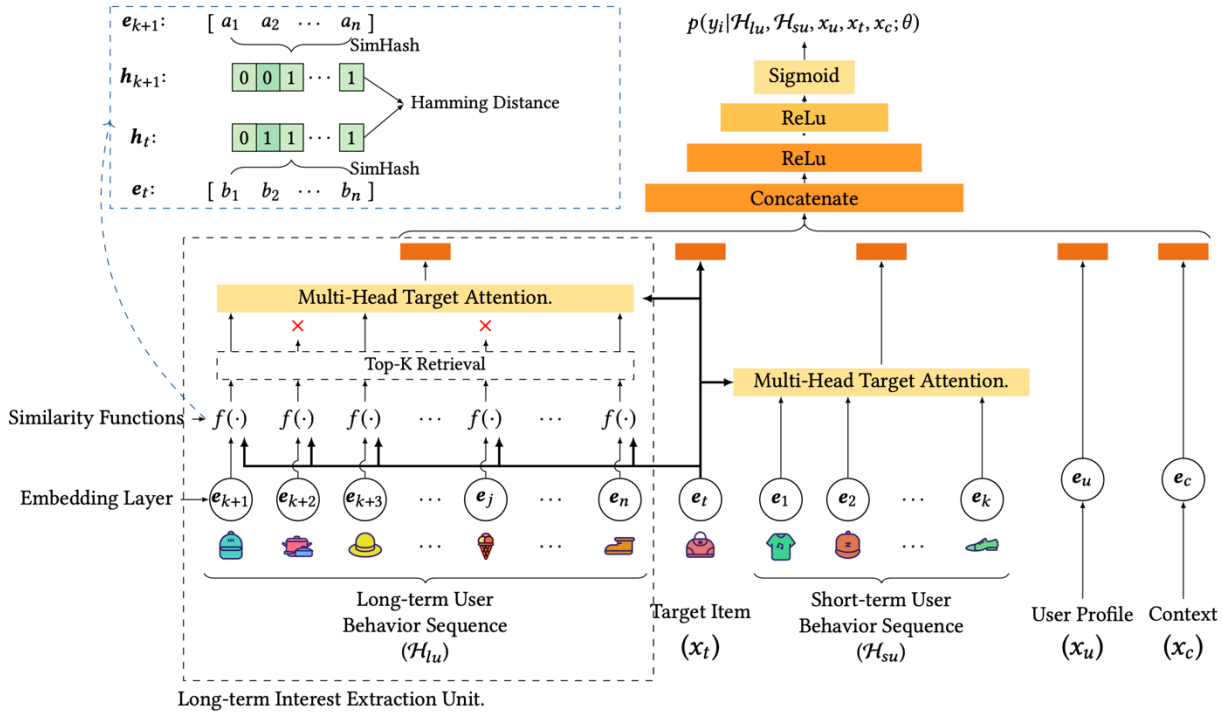


Search-based User Interest Modeling with Lifelong Sequential Behavior Data for Click-Through Rate Prediction. CIKM 2020.

SDIM, ETA

The **pre-trained** embedding in GSU of SIM is **outdated**
Solution: **end-to-end training**, search by exact match or Hamming distance based on the **LSH**





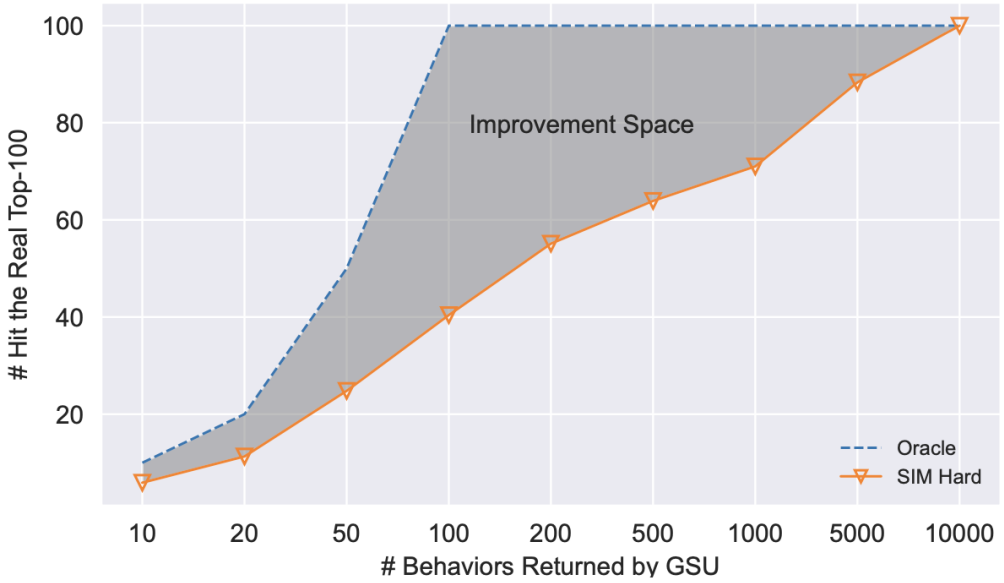
Sampling is all you need on modeling long-term user behaviors for CTR prediction. CIKM 2022.
ETA: End-to-End User Behavior Retrieval in Click-Through Rate Prediction Model. 2021.

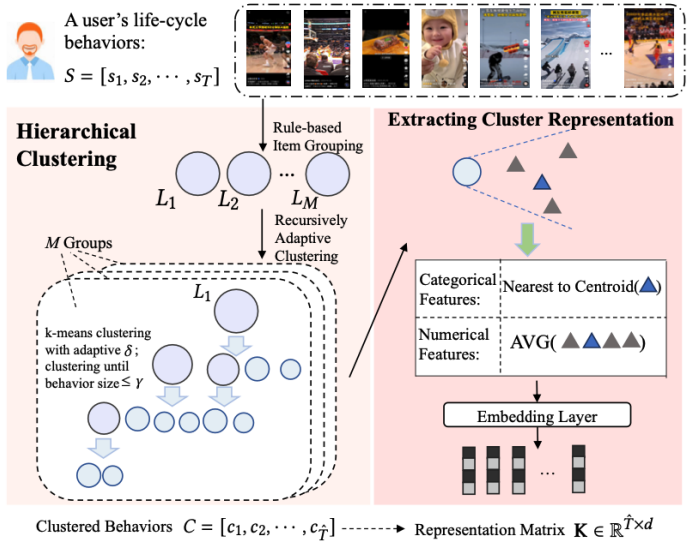
Tencent

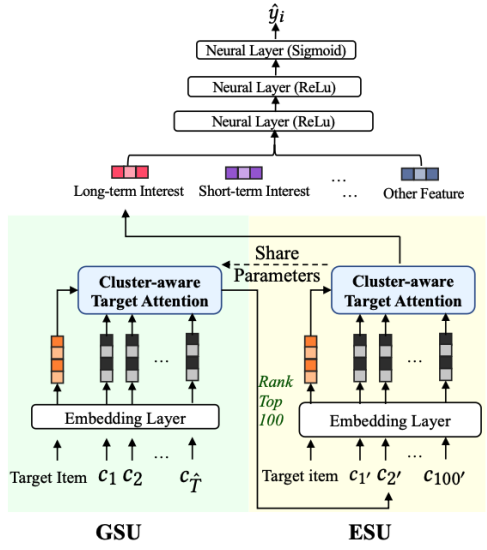
TWIN, TWIN-V2

Challenge: The relevance metric in GSU is **coarse** and **inconsistent** with ESU.
Solution: **identical** target-behavior relevance metrics in GSU as the Target-Attention in ESU.

Divide and Conquer with a hierarchical clustering method for efficient search.







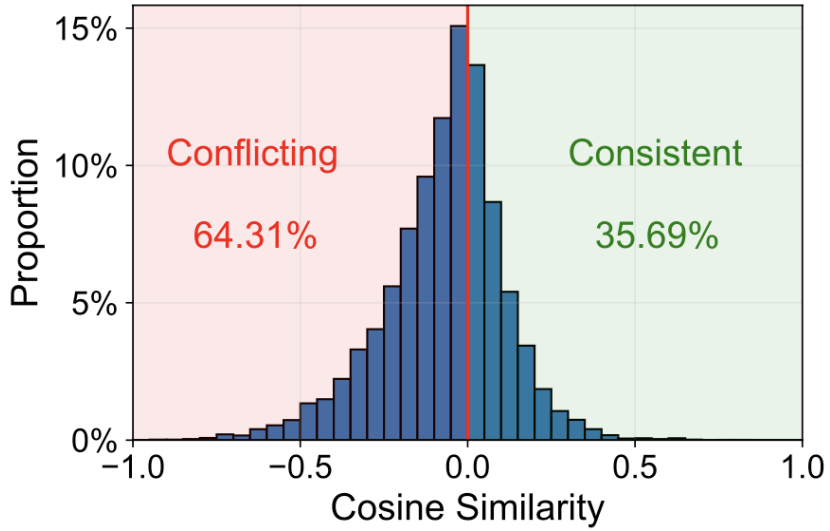
TWIN: TWo-stage Interest Network for Lifelong User Behavior Modeling in CTR Prediction at Kuaishou. KDD 2023.
TWIN V2: Scaling Ultra-Long User Behavior Sequence Modeling for Enhanced CTR Prediction at Kuaishou. CIKM 2024.

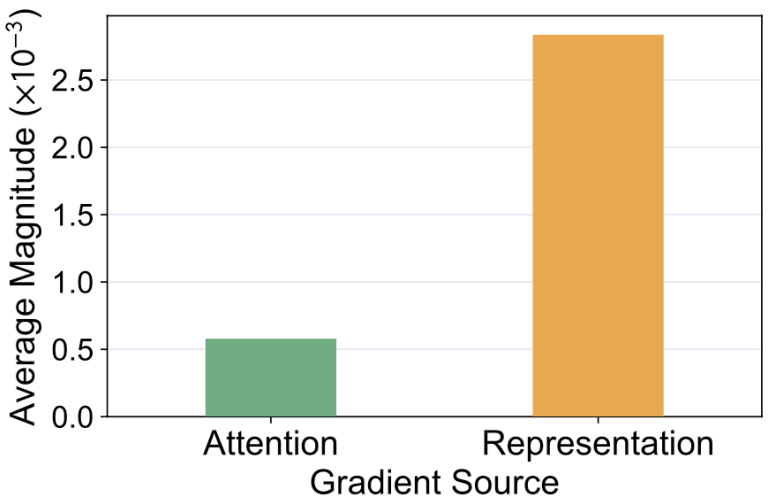
Tencent

DARE (Decoupled Attention and Representation Embeddings)

$$\sum_{X_i \in \mathcal{H}} \underbrace{\alpha(\tilde{\boldsymbol{e}}_i, \tilde{\boldsymbol{v}}_t)}_{\text{TA}} \cdot \underbrace{(\tilde{\boldsymbol{e}}_i \odot \tilde{\boldsymbol{v}}_t)}_{\text{TR}}$$

- TWIN and TWIN-V2 use the attention score for retrieval.
- There are **conflicts** between the target **attention** and target **representation**.
- The gradients are **dominated** by the target **representation**.
- Gradient domination deteriorates the attention learning and hence **harm the search quality**.

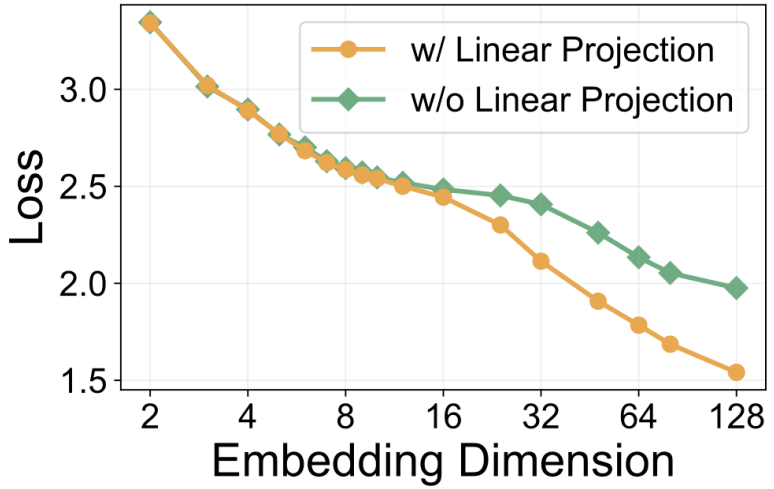




Long-Sequence Recommendation Models need Decoupled Embeddings. ICLR 2025.

DARE

- Straightforward approach: projection matrix for attention and representation.
- Unlike NLP, the Q, K, U, V **projection matrix** in RecSys is much smaller, making it **unable to decouple** attention and representation.



Projection matrix only works with large embedding dimension.

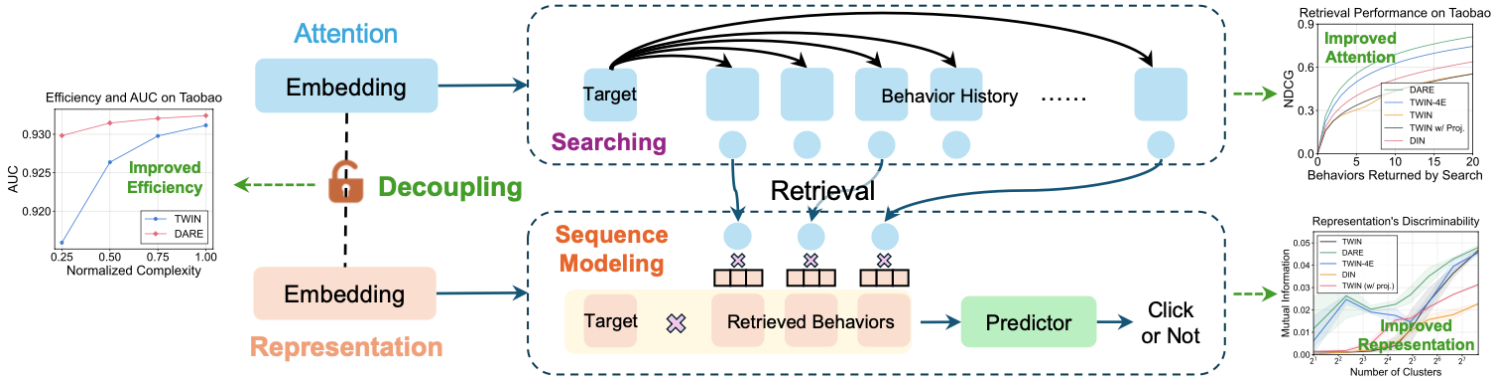
NLP

#tokens: 128,000
#params: 16,777,216 (4096 x 4096)

Ads Recommendation

#tokens: 1,000,000,000
#params: 16,384 (128 x 128)

- Solution: **decouple the embeddings** for attention and representation.



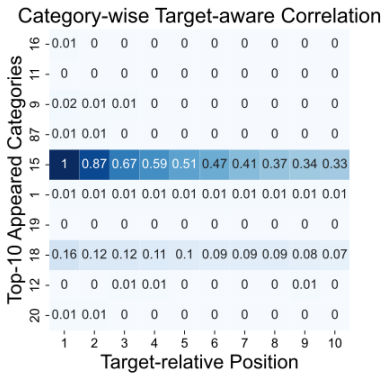
Long-Sequence Recommendation Models need Decoupled Embeddings. ICLR 2025.

Tencent

DARE – Discriminability Analysis

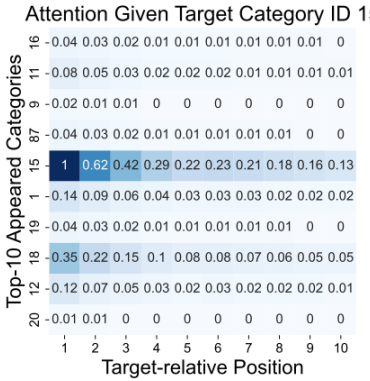
Embedding decoupling resolves the conflicts between attention and representation, leading to:

- **Better correlation** modeling in attention
- **More discriminative** representation



(a) GT mutual information

(b) TWIN learned correlation



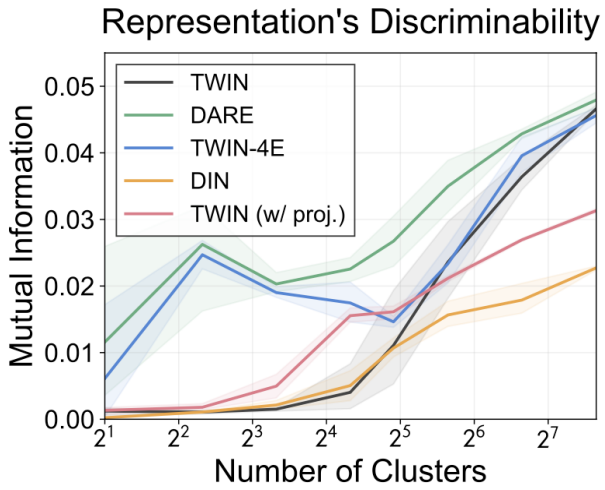
(c) DARE learned correlation

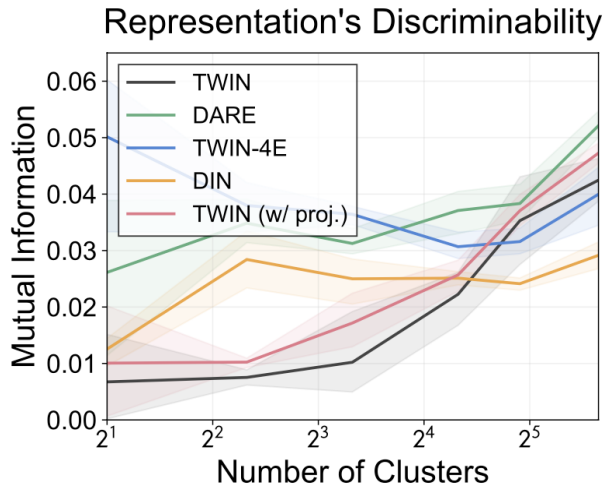
Temporal-Semantic Correlation

Discretized Mutual Information

$$\mathrm{MI}(\mathrm{Discret}(h), Y)$$

Mutual information between the discretized factors $\mathrm{Discret}(h)$, i.e., clusters, and the target Y



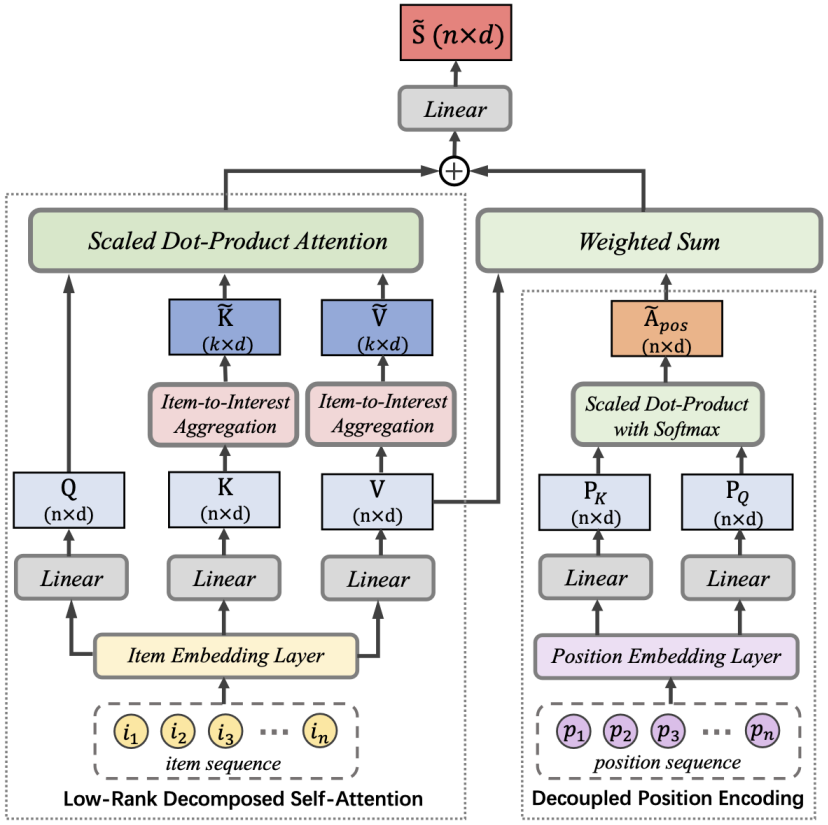


Representation **Discriminability**

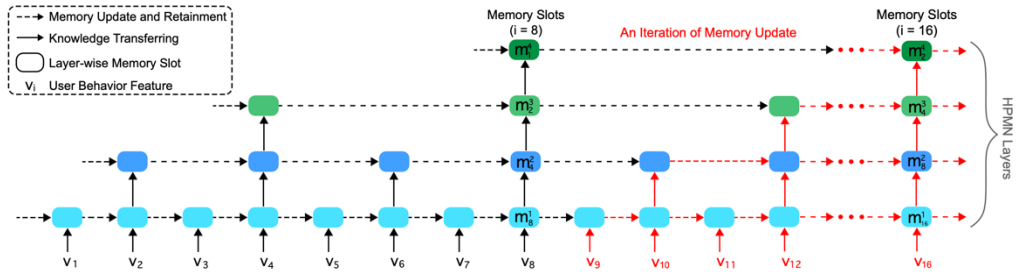
Long-Sequence Recommendation Models need Decoupled Embeddings. ICLR 2025.

Tencent

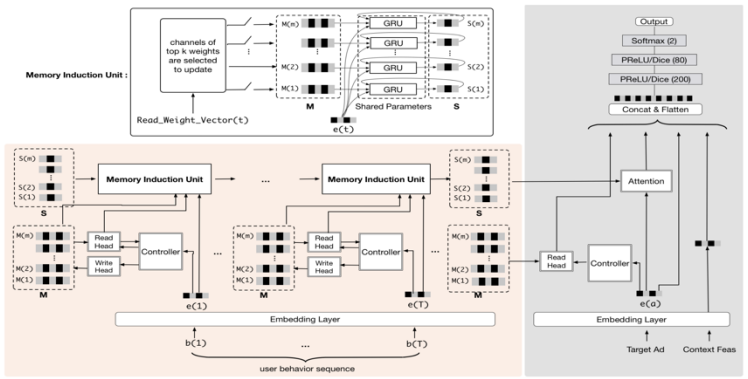
Compression - Memory Networks, Low-rank Decomposition



LightSANs



HPMN



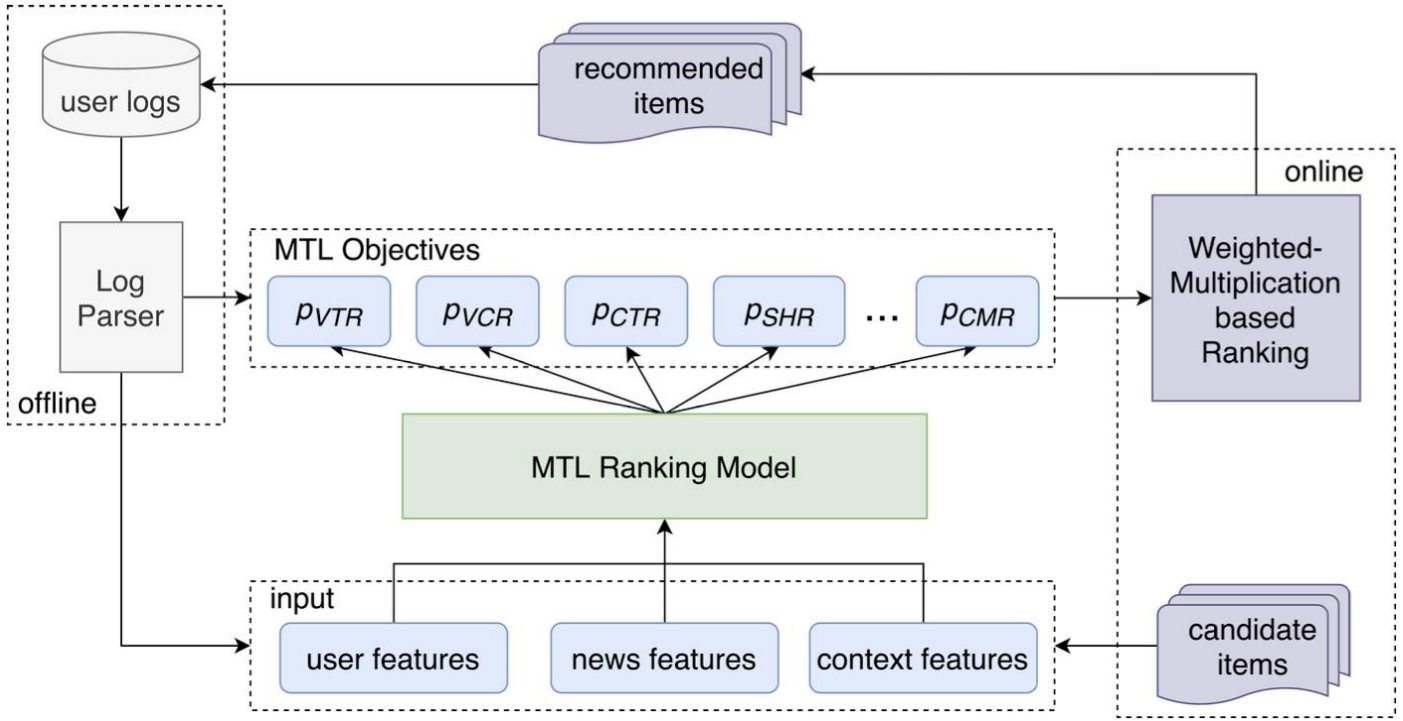
MIMN

Lighter and Better: Low-Rank Decomposed Self-Attention Networks for Next-Item Recommendation. 2021.
Lifelong Sequential Modeling with Personalized Memorization for User Response Prediction. SIGIR 2019.
Practice on Long Sequential User Behavior Modeling for Click-Through Rate Prediction. KDD 2019.

- Part II, Prediction
  - Perspectives
  - Feature Interaction
  - Sequential Models
  - **Multi-Task and Multi-Domain Learning**
  - Large Recommendation Models
  - LLM4Rec

Multi-task/domain Learning

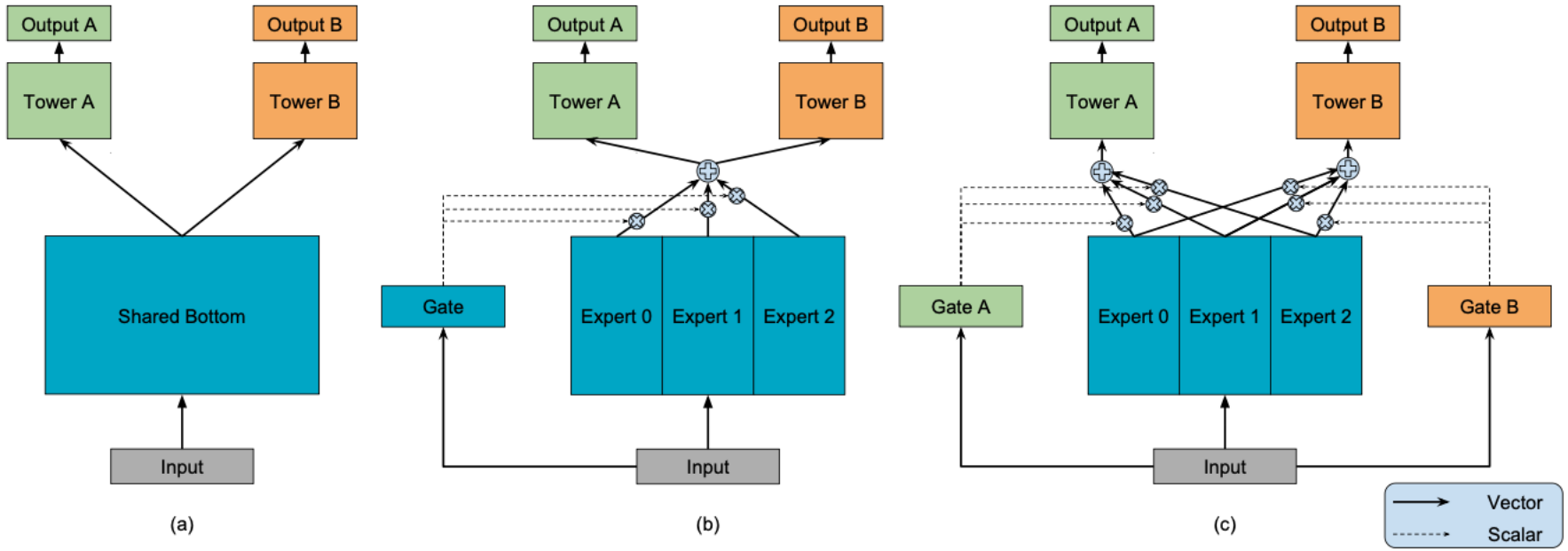
- Multi-task Learning aim to utilize the shared/transferable knowledge from other tasks to enhance every task.
- Challenges: **negative transfer**, **seesaw phenomenon**.



Progressive Layered Extraction (PLE): A Novel Multi-Task Learning (MTL) Model for Personalized Recommendations. RecSys, 2020.

MMoE (Multi-Gate MoE)

- An art of balance between **shared** and **specific** components



Shared-Bottom
One **shared** experts for all tasks.

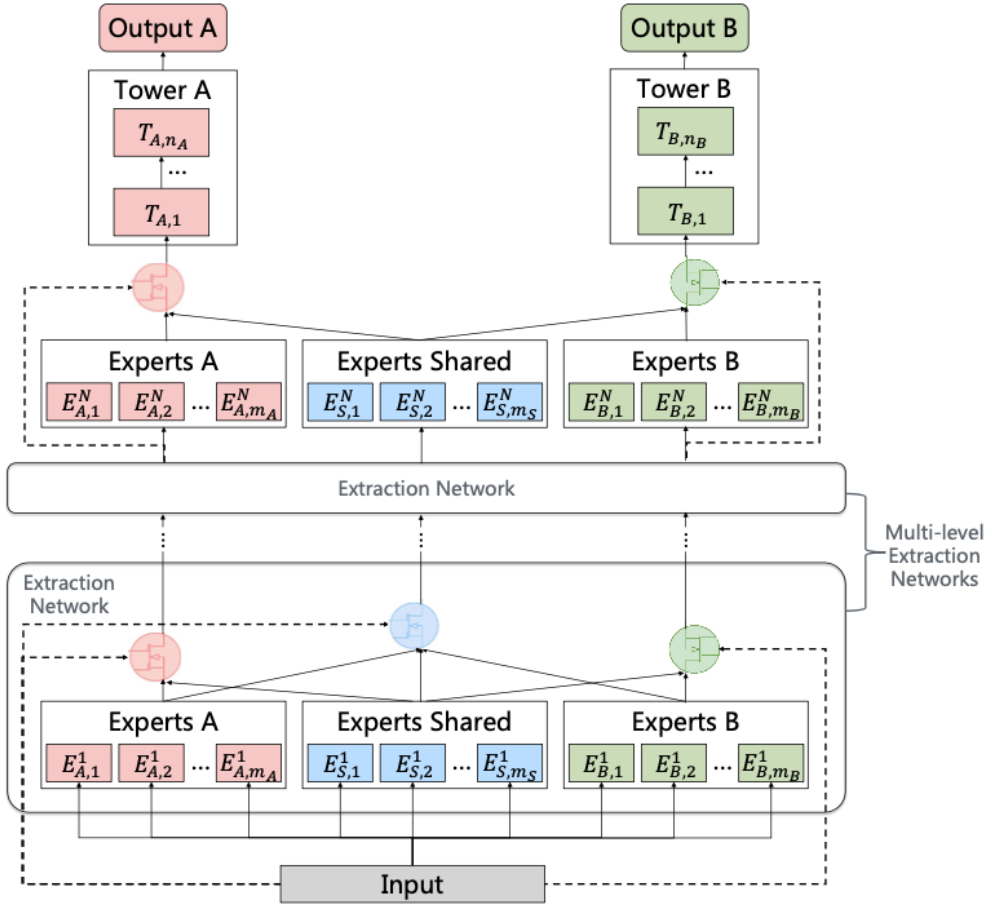
MoE
one **shared** gating
multiple **shared** experts,

MMoE
task-specific gating
multiple **shared** experts,

Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts. KDD 2018.

Tencent

PLE (Progressive Layered Extraction)

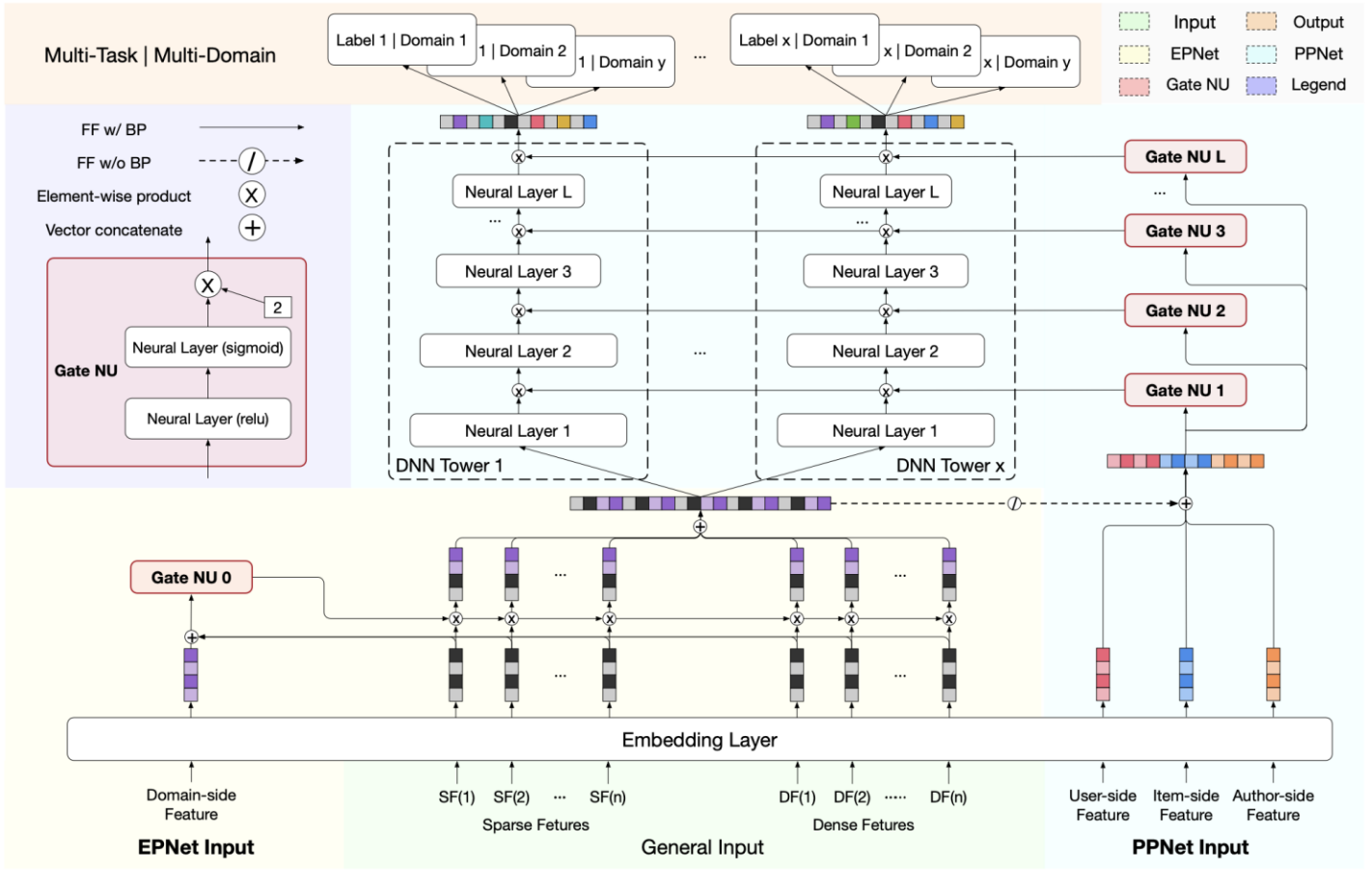


task-specific gating
shared and **task-specific** experts

Progressive Layered Extraction (PLE): A Novel Multi-Task Learning (MTL) Model for Personalized Recommendations. RecSys, 2020.

Tencent

PEPNet (Parameter and Embedding Personalized Networks)



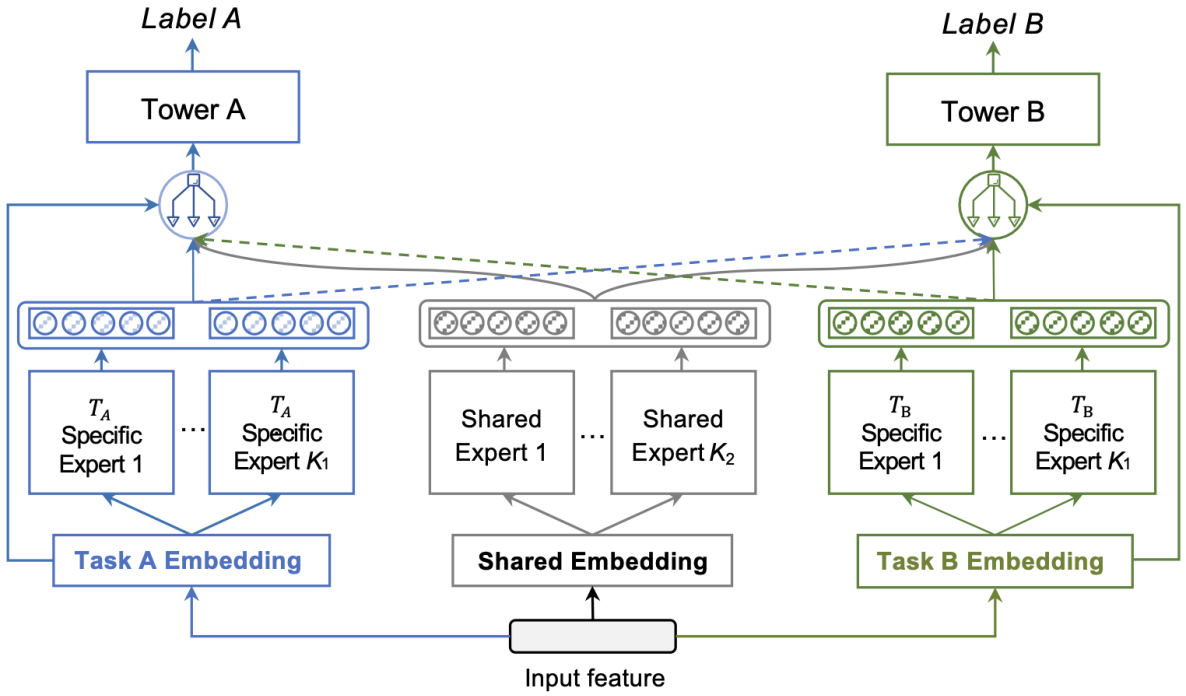
task-specific gating
task-specific embedding gating (experts)

PEPNet: Parameter and Embedding Personalized Network for Infusing with Personalized Prior Information. 2023.

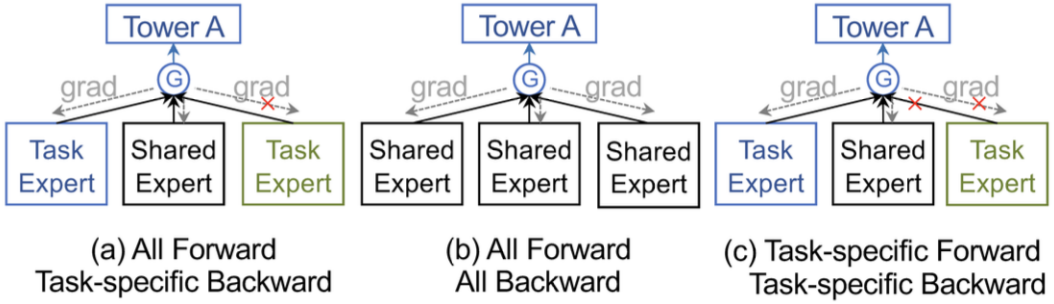
Tencent

STEM (Shared and Task-specific EMbeddings)

Learn shared and **task-specific embeddings**.



task-specific gating
shared and **task-specific** experts
shared and **task-specific** embeddings



All Forward **Task-specific Backward** Gating

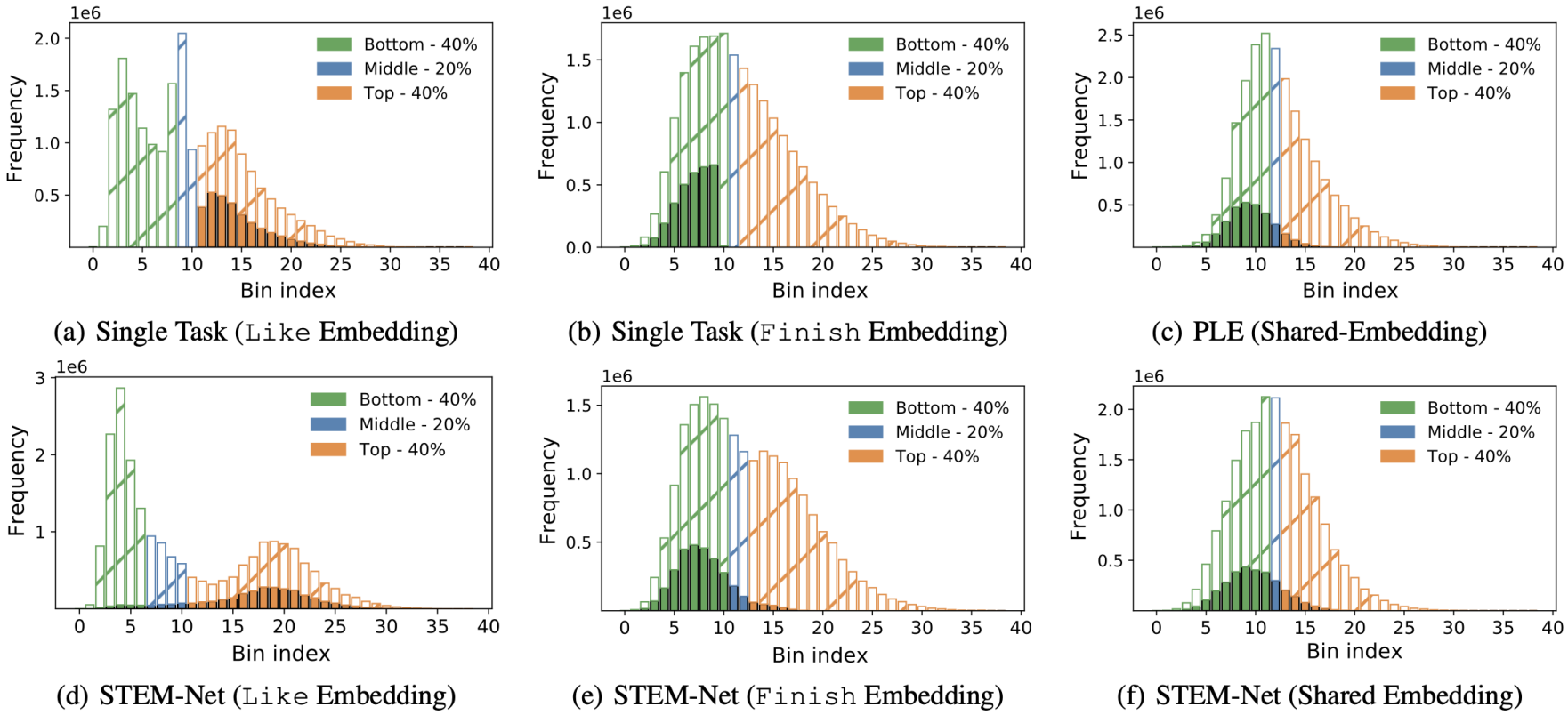
STEM: Unleashing the Power of Embeddings for Multi-Task Recommendation. AAAI 2024.

Tencent

STEM – contradictory preference analysis

Pick up user-item pairs that have contradictory distances in task like and finish.

- E.g., user-item pairs that have **large** distance in the Like Single Model embedding, while have **small** distance in Finish Single Model Embedding

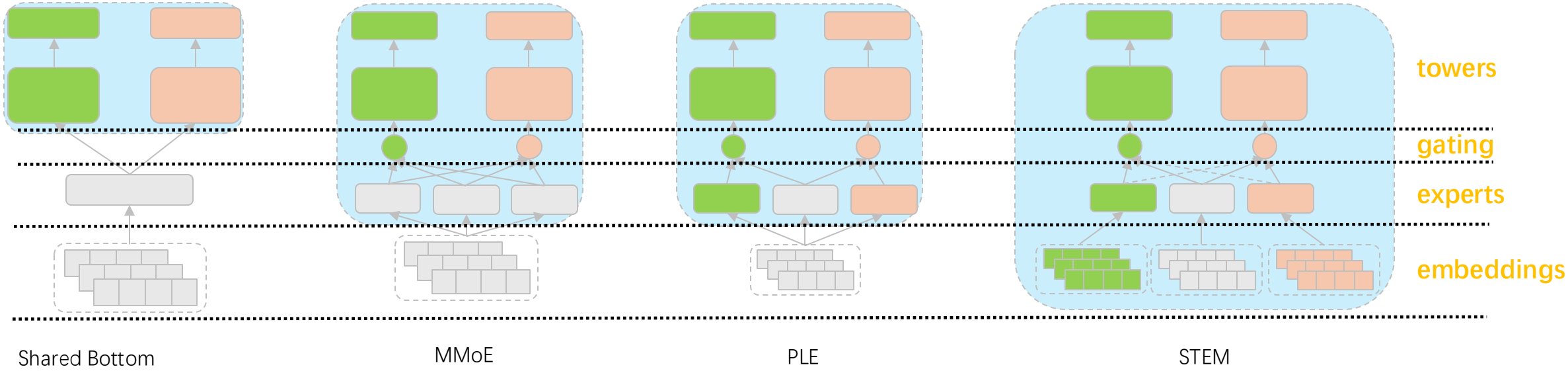


(a) Single Task (`Like` Embedding) (b) Single Task (`Finish` Embedding) (c) PLE (Shared-Embedding)

(d) STEM-Net (`Like` Embedding) (e) STEM-Net (`Finish` Embedding) (f) STEM-Net (Shared Embedding)

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

MTL Models Evolution

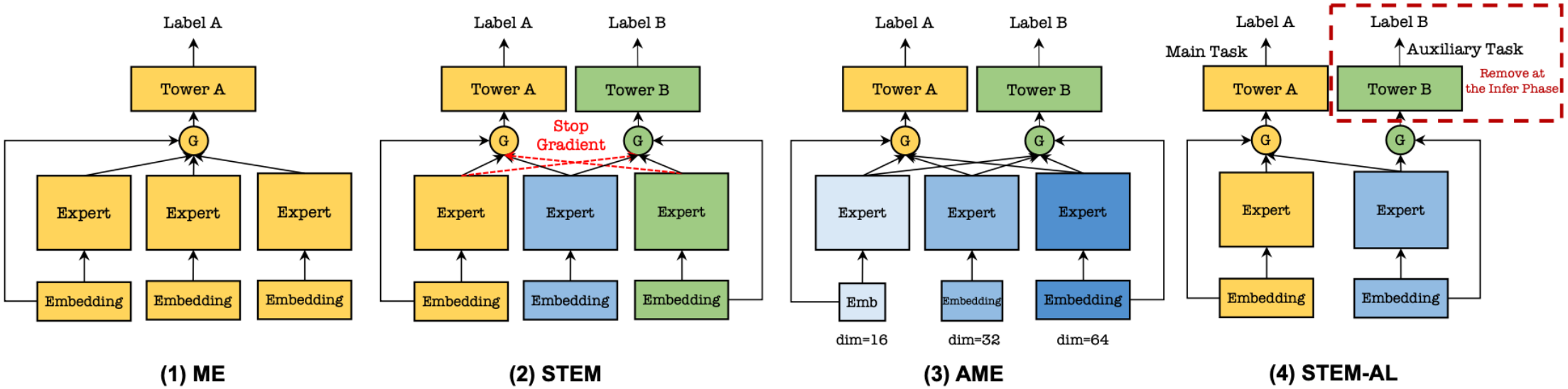


A trend to decouple more bottom components, i.e., from gating (MMoE) to experts (PLE, PEPNet) to embeddings (STEM)

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

AME (Asymmetric Multi-Embedding)

- STEM requires to learn task-specific embedding, which is **infeasible** with many task, e.g., more than **hundreds** in Tencent. We may group tasks, but how?
- AME: Learn multiple **shared** yet **asymmetric** embeddings and tasks, with different embedding dimensions **to break the homogeneity**.
- For auxiliary learning, try STEM-AL.



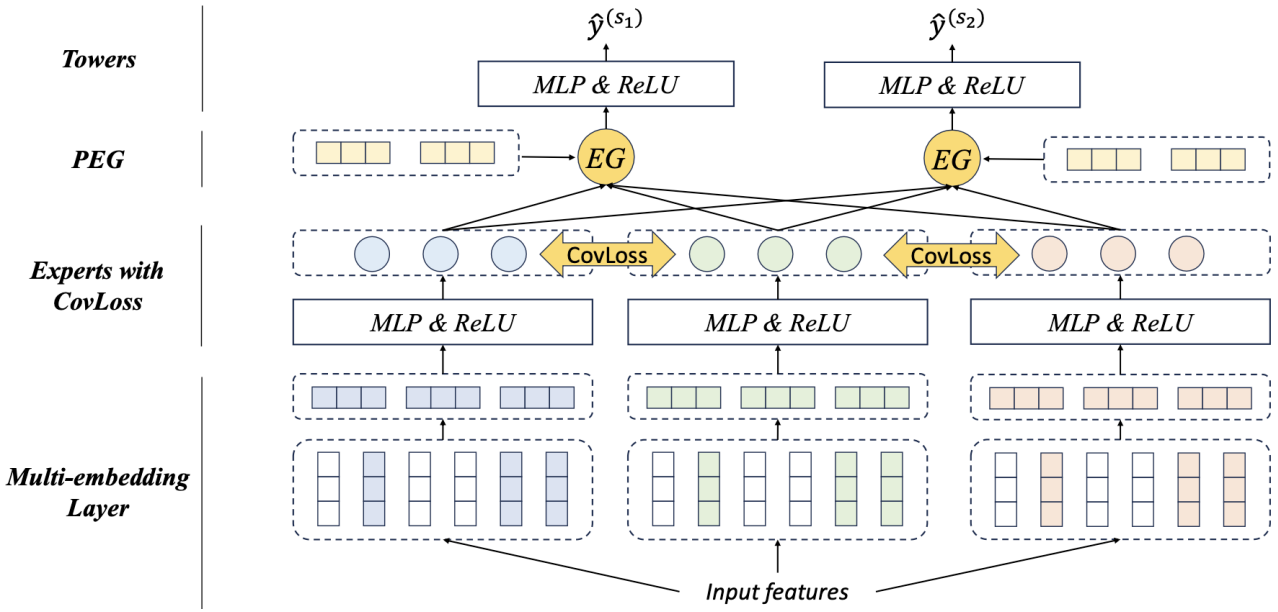
Ads Recommendation in a Collapsed and Entangled World. KDD 2024.

Tencent

Crocodile (Cross-experts Covariance Loss for Disentangled Learning)

- Different embedding dimensions (AME) or heterogeneous expert modules (DHEN) **constrains** the architecture design.
- A more general approach: **Cross Expert Covariance** loss, to de-correlate the experts directly.

$$\mathcal{L}_{Cov} = \frac{1}{d^2} \sum_{p,q \in M \times M, p \geq q} ||[\mathbf{O}^{(p)} - \overline{\mathbf{O}}^{(p)}]^T [\mathbf{O}^{(q)} - \overline{\mathbf{O}}^{(q)}]||_1$$

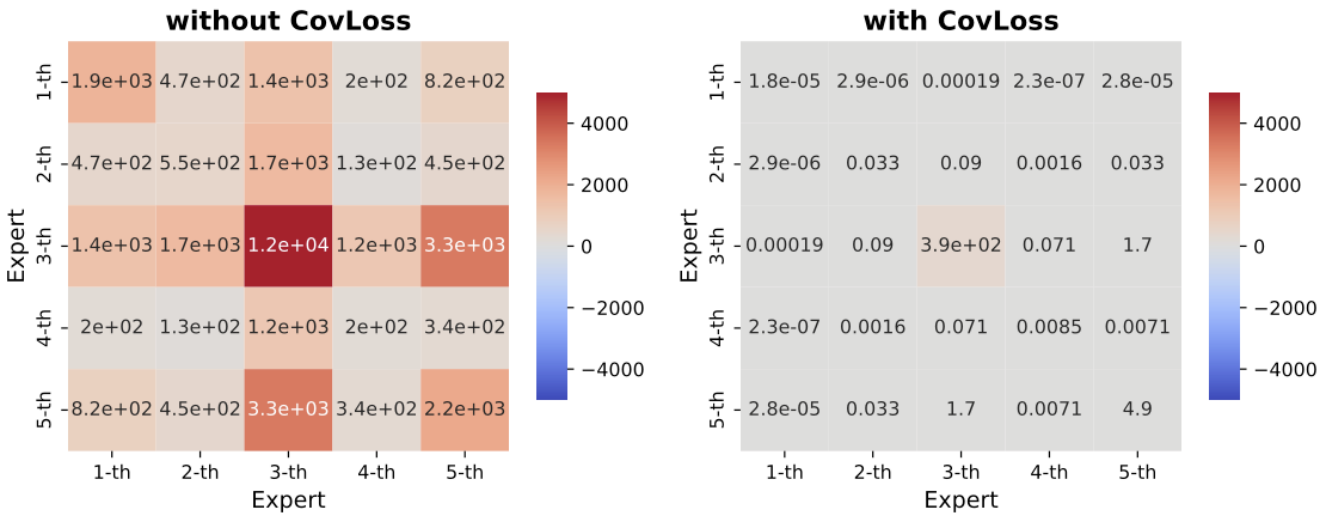


Crocodile: Cross Experts Covariance for Disentangled Learning in Multi-Domain Recommendation. Under review.

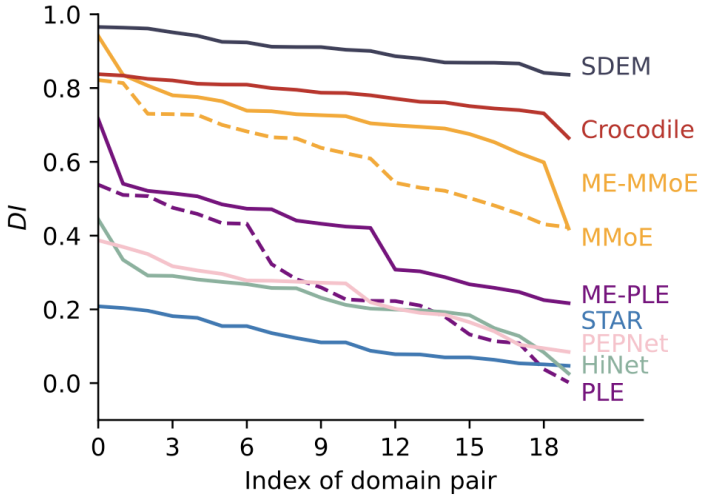
Tencent

Crocodile - Disentanglement

The CovLoss significantly **reduces** the inter-expert correlations.



Better Diversity-Index based on the contradictory preference analysis.



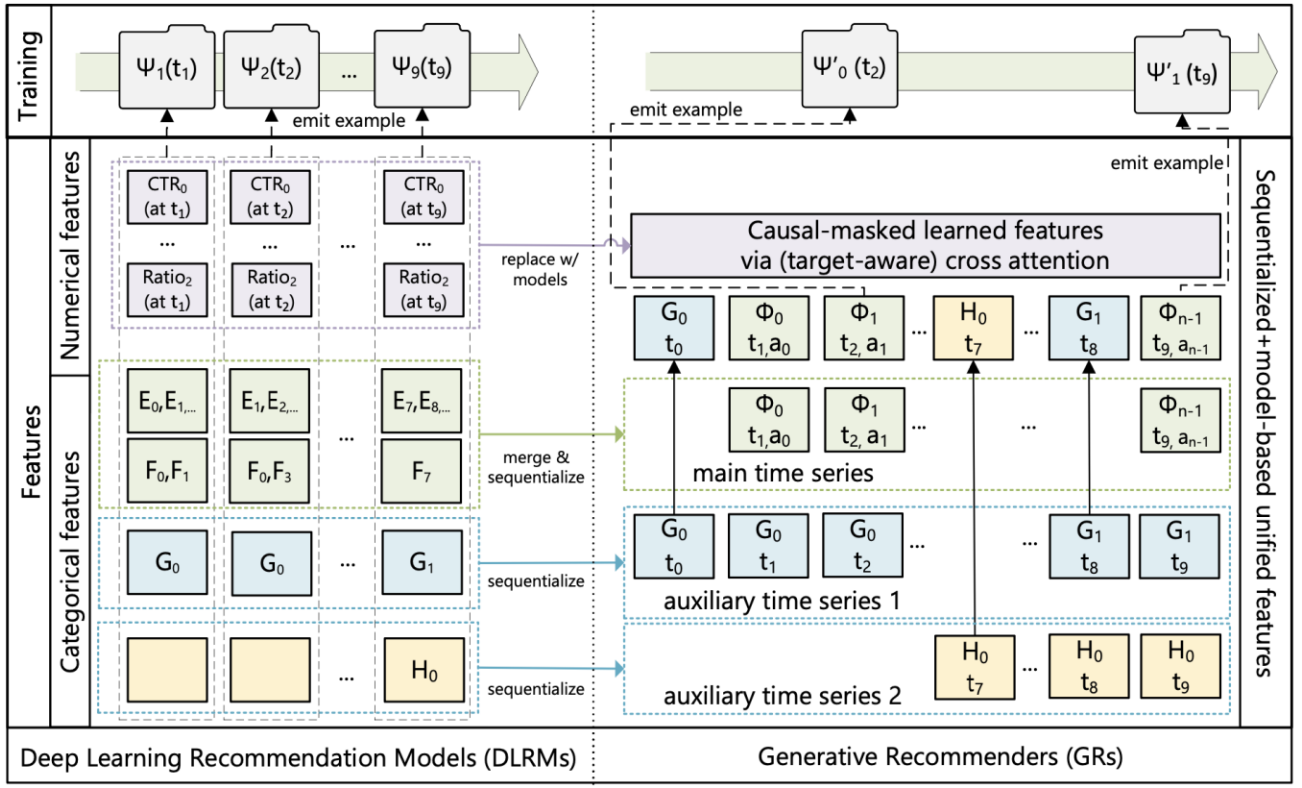
Crocodile: Cross Experts Covariance for Disentangled Learning in Multi-Domain Recommendation. Under review.

Tencent

- Part II, Prediction
  - Perspectives
  - Feature Interaction
  - Sequential Models
  - Multi-Task and Multi-Domain Learning
  - **Large Recommendation Models**
  - LLM4Rec

HSTU

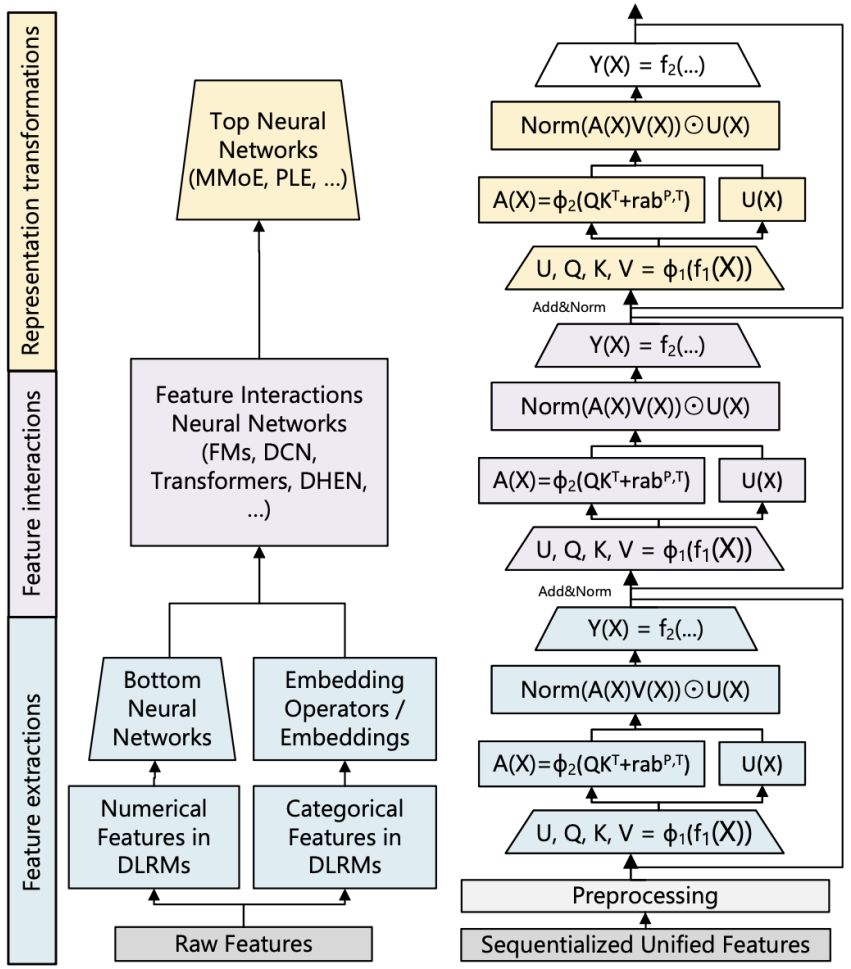
- **Scaling laws** in LLM.
- Formulate recommendation as sequential transduction tasks within a **generative** modeling framework.
- Build a universal sequence to unify categorical and numerical features.



Actions Speak Louder than Words: Trillion-Parameter Sequential Transducers for Generative Recommendations. ICML 2024.

Tencent

HSTU



$$U(X), V(X), Q(X), K(X) = \text{Split}(\phi_1(f_1(X))) \quad (1)$$

$$A(X)V(X) = \phi_2\left(Q(X)K(X)^T + \text{rab}^{p,t}\right)V(X) \quad (2)$$

$$Y(X) = f_2\left(\text{Norm}\left(A(X)V(X)\right) \odot U(X)\right) \quad (3)$$

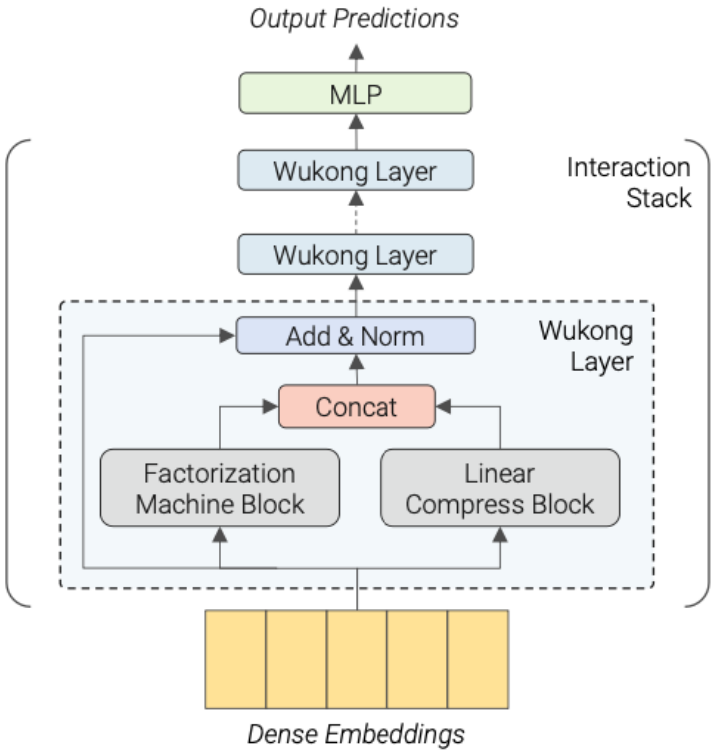
Can be interpreted as a variant of **SwiGLU**, with a bilinear term $V(X) \odot U(X)$, which is similar with the **target-aware representation** in TIN.

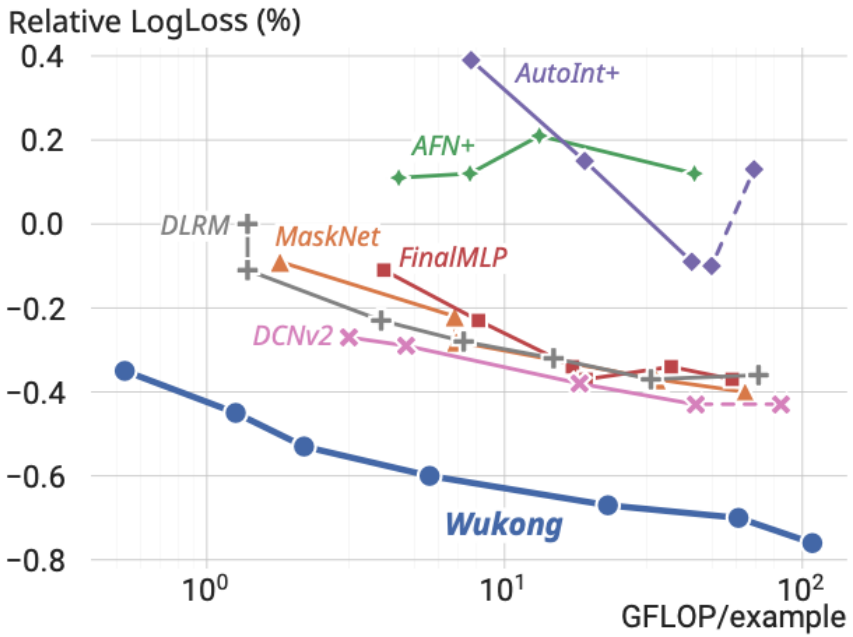
$$\sum_{X_i \in \mathcal{H}} \underbrace{\alpha(\tilde{\boldsymbol{e}}_i, \tilde{\boldsymbol{v}}_t)}_{\text{TA}} \cdot \underbrace{(\tilde{\boldsymbol{e}}_i \odot \tilde{\boldsymbol{v}}_t)}_{\boxed{\text{TR}}}$$

Actions Speak Louder than Words: Trillion-Parameter Sequential Transducers for Generative Recommendations. ICML 2024.

Wukong

Each layer in the Interaction Stack consists of a **Factorization Machine Block** and a **Linear Compression Block**.





- The linear compression block helps **mitigate** the **dimensional collapse**.
- **Scale up** (HSTU, Wukong) v.s. **scale-out**

Wukong: Towards a Scaling Law for Large-Scale Recommendation. ICML 2024.

Tencent

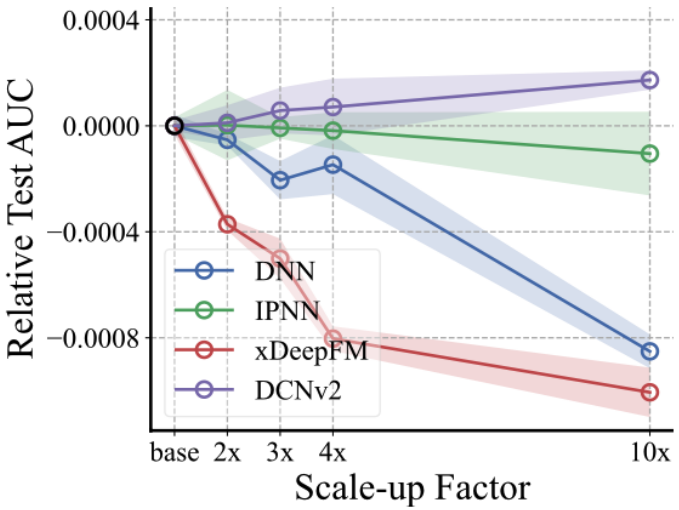
Dimensional Collapse

- Embeddings **dominates** the parameters in RecSys. A straightforward way to scale is to **enlarge embedding dimensions** K.
- However, the performance **deteriorates** when we increase K.
- Through singular spectrum analysis, we found severe **dimensional collapse**.

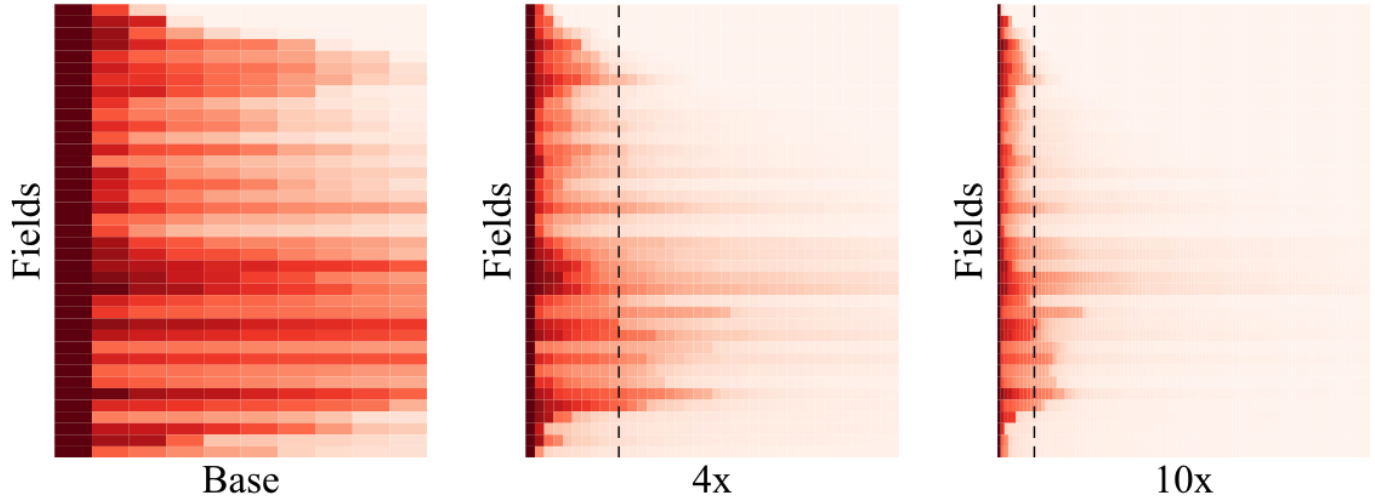
Information Abundance

$$IA = \frac{|\Sigma|_1}{|\Sigma|_\infty}$$

Sum of all singular values divided by the maximum one



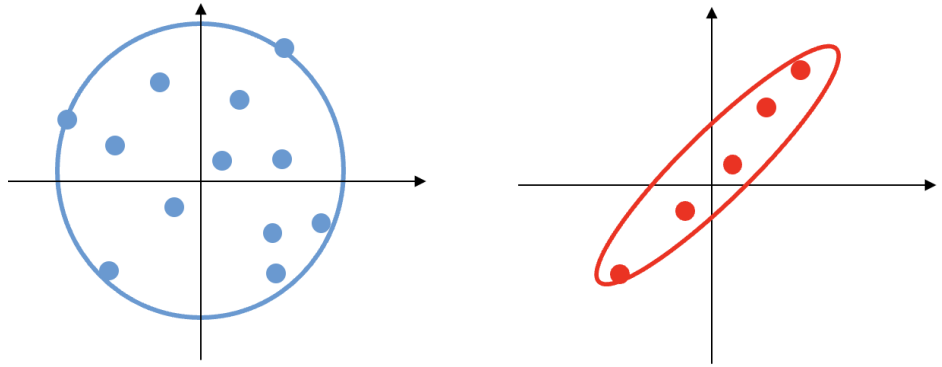
(a) Performance when scaling up recommendation models



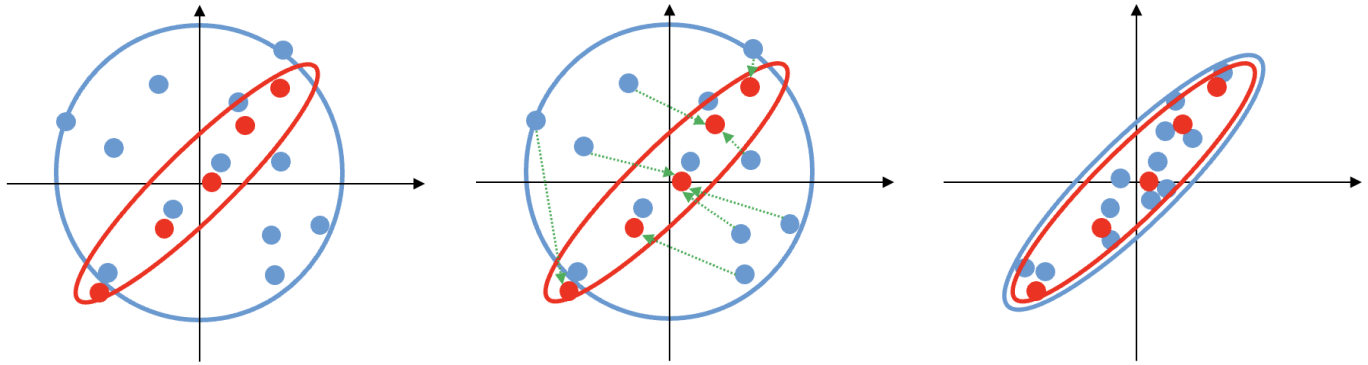
(b) Singular values of DCNv2 under different model size, with the dashed lines corresponding to the base size.

On the Embedding Collapse When Scaling Up Recommendation Models. ICML 2024.

Interaction Collapse Theory



High-cardinality features (item ID) Low-cardinality features(Gender)

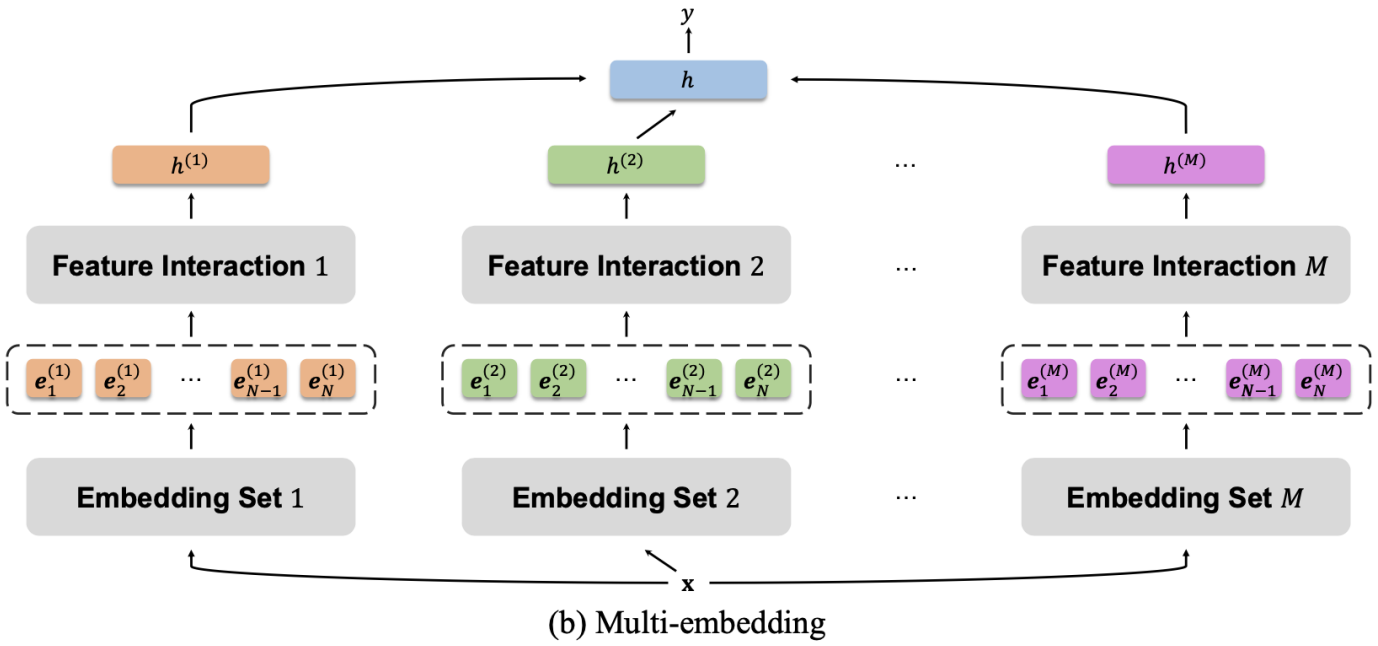


In feature interaction of recommendation models, fields with low- information-abundance (low cardinality) embeddings constrain the information abundance of other fields, resulting in collapsed embedding matrices.

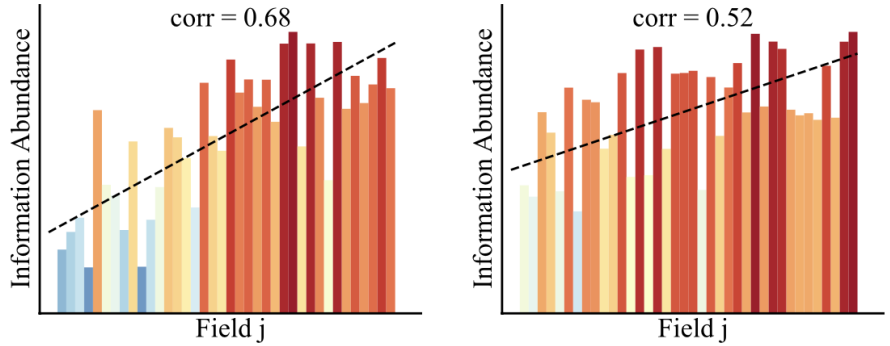
On the Embedding Collapse When Scaling Up Recommendation Models. ICML 2024.

Multi-Embedding Paradigm

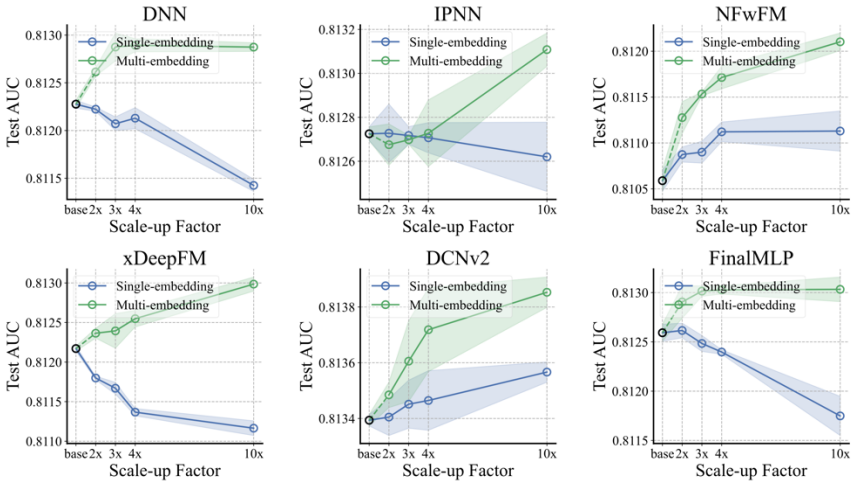
$$\sum_i g_i f_i(\Phi(x, E_i))$$



(b) Multi-embedding



(a) $\sum_{i=1}^{N} \mathrm{IA}(\boldsymbol{E}_i^{\to j})$, SE. (b) $\sum_{i=1}^{N} \mathrm{IA}(\boldsymbol{E}_i^{\to j})$, ME.



- Both ME and FFM learns **multiple embeddings** for each feature.
- Different motivation than STEM and Crocodile for Multi-task learning.

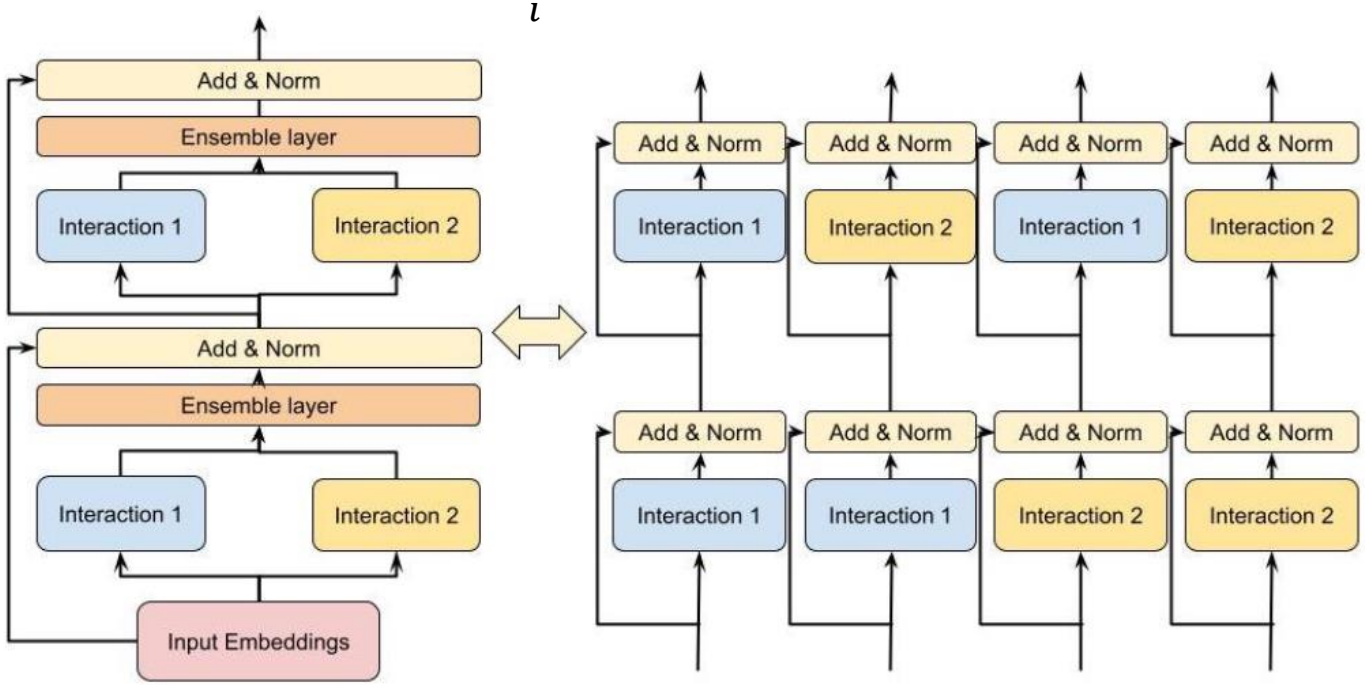
On the Embedding Collapse When Scaling Up Recommendation Models. ICML 2024.

Tencent

DHEN (Deep and Hierarchical Ensemble Network)

Heterogeneous interaction experts.

$$\sum_i g_i f_i(\Phi_i(x, E))$$

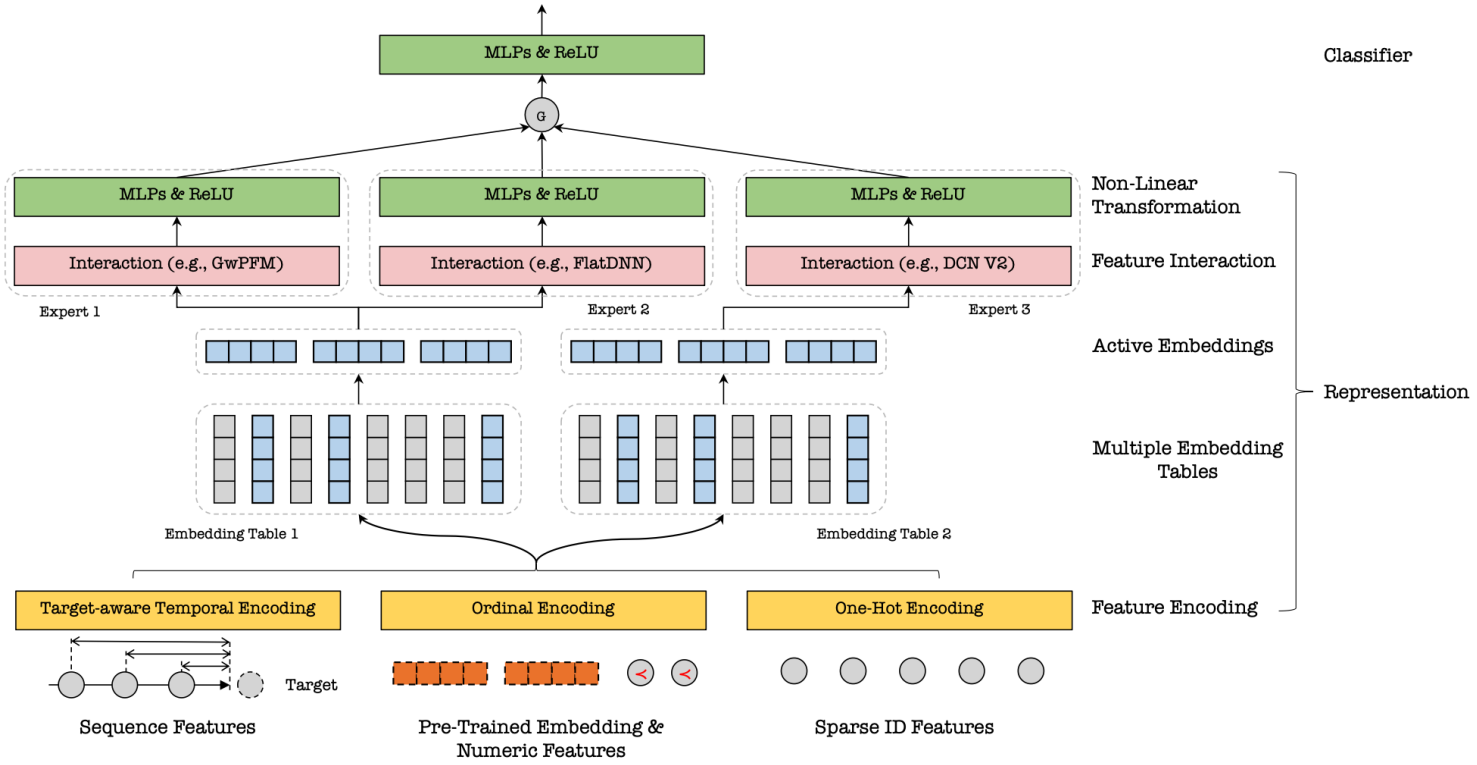


DHEN: A Deep and Hierarchical Ensemble Network for Large-Scale Click-Through Rate Prediction. 2022.

Hetero-ME (Heterogeneous Multi-Embedding)

Heterogeneous interaction experts, with **multiple independent embeddings**.

$$\sum_i g_i f_i(\Phi_i(x, E_i))$$



Ads Recommendation in a Collapsed and Entangled World. KDD 2024.

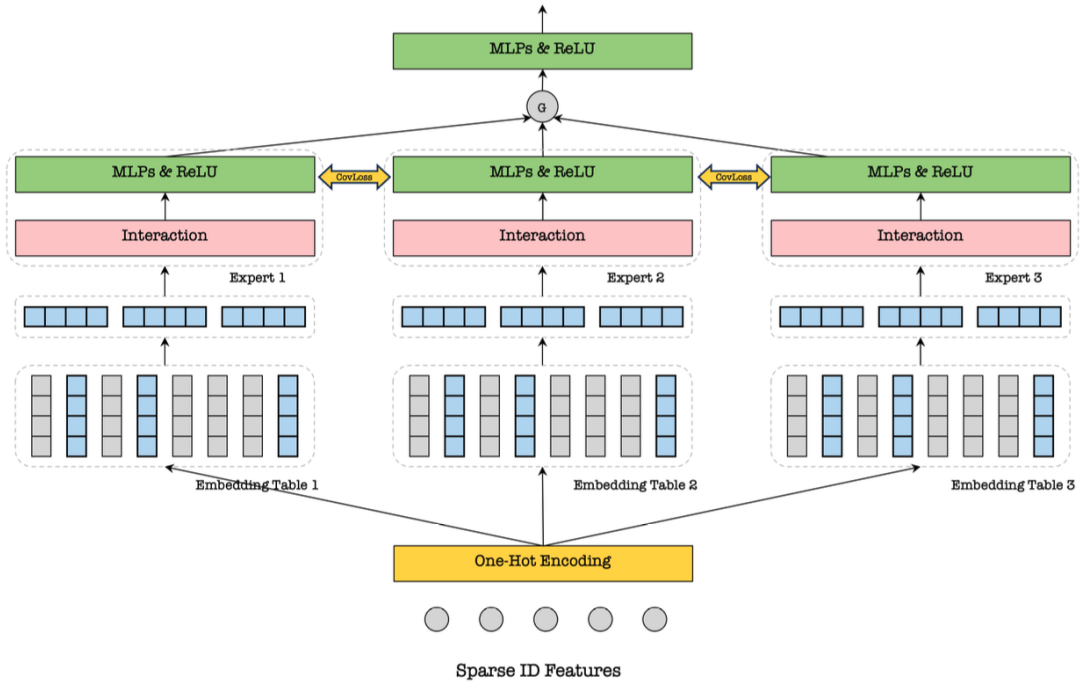
Tencent

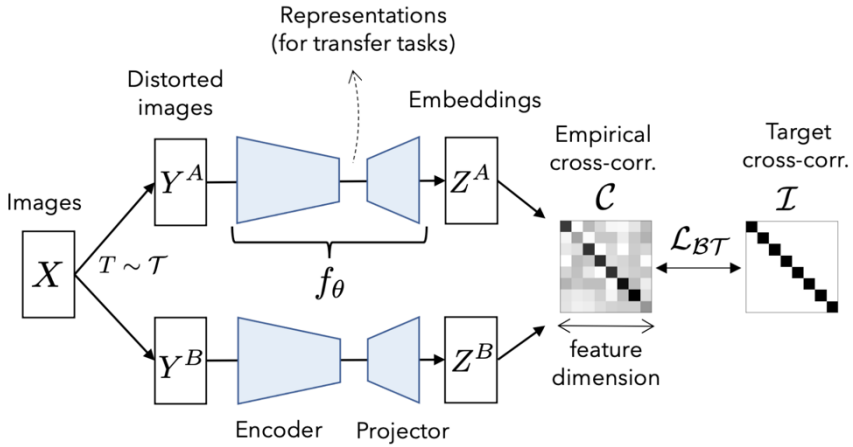
D-MoE (De-Correlated MoE)

Besides adopting heterogeneous architectures (DHEN), or different embedding dimensions, we can directly de-correlate the experts by an **Inter-Expert De-Correlation Loss**

$$\mathrm{L_{BCE}}\left(\sum_i g_i f_i\left(\Phi_i(x, E_i)\right), y\right) + \sum_i \sum_j L_{\mathrm{Corr}}(\Phi_i(x, E_i), \Phi_j(x, E_j))$$

$$\mathcal{L}_{Corr} = \frac{1}{d^2} \sum_{\substack{m_1, m_2 \in \{1,\ldots,M\} \\ m_1 < m_2}} \left\| \left[\frac{\mathbf{O}^{(m_1)} - \overline{\mathbf{O}}^{(m_1)}}{\sigma(\mathbf{O}^{(m_1)})} \right]^T \left[\frac{\mathbf{O}^{(m_2)} - \overline{\mathbf{O}}^{(m_2)}}{\sigma(\mathbf{O}^{(m_2)})} \right] \right\|_2$$



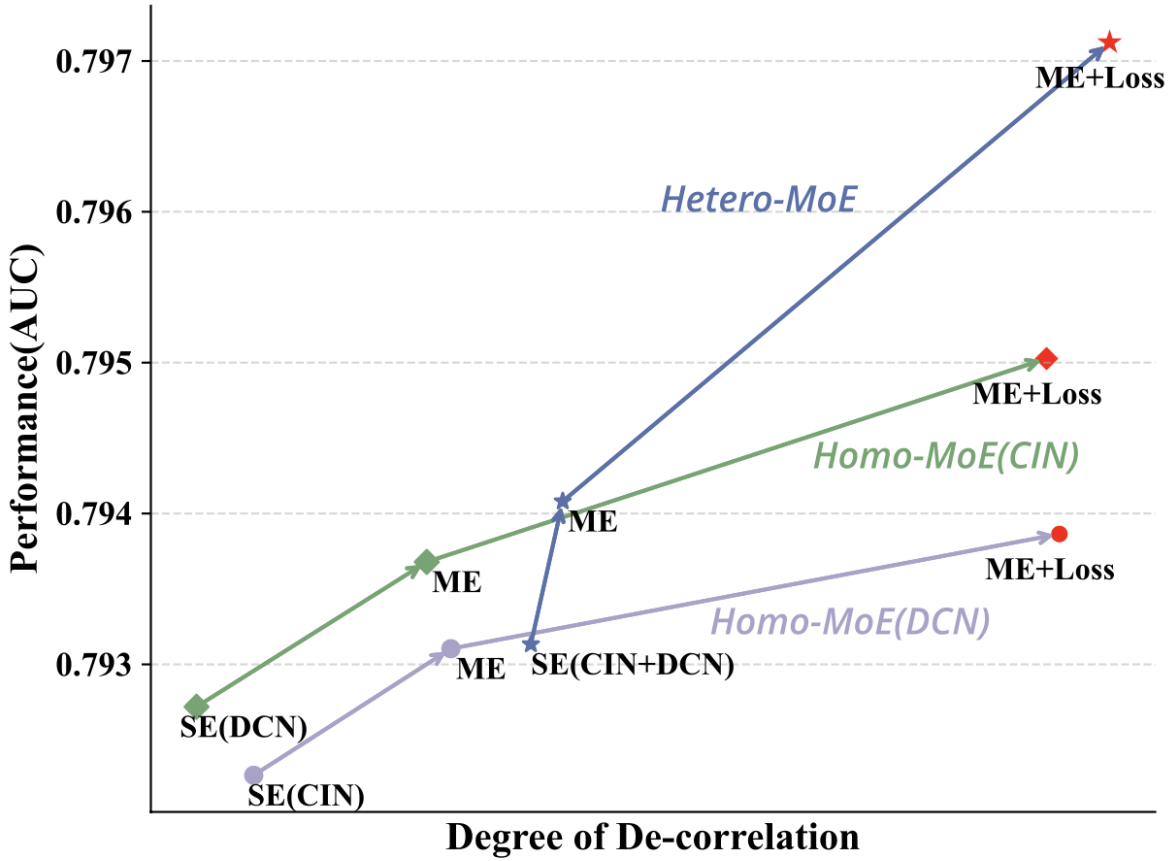


Barlow-Twins

Towards De-correlated Mixture-of-Experts for CTR Prediction. Under review.

D-MoE - Redundancy Reduction

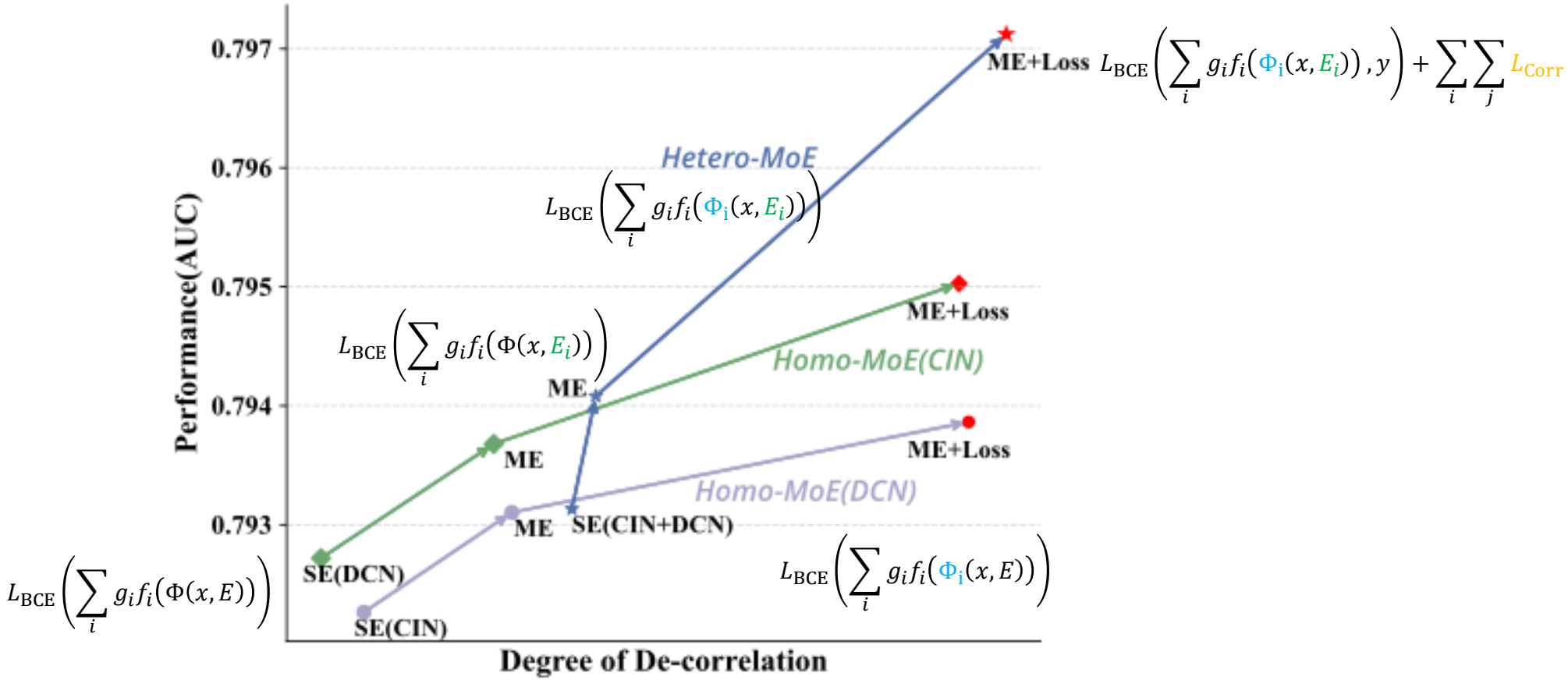
Applying the **Inter-Expert De-Correlation Loss** on ME-MoE or ME-Hetero-MoE can further reduce the correlation between behaviors.



Towards De-correlated Mixture-of-Experts for CTR Prediction. Under review.

Expert De-Correlation Principle

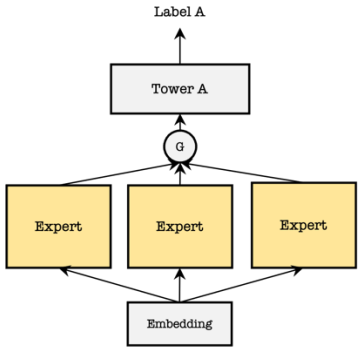
By combining the Heterogeneous MoE, Multi-Embedding, Inter-Expert De-correlation, we can progressively de-correlate the experts, while achieving performance lift.



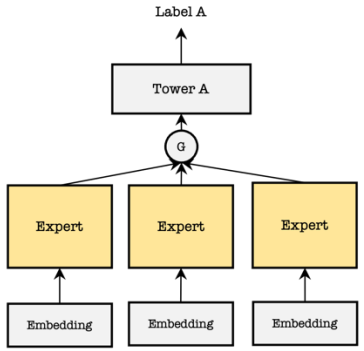
Towards De-correlated Mixture-of-Experts for CTR Prediction. Under review.

Tencent

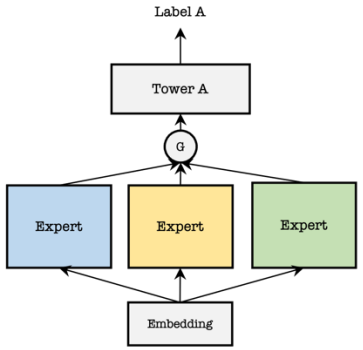
MoE Model Zoo



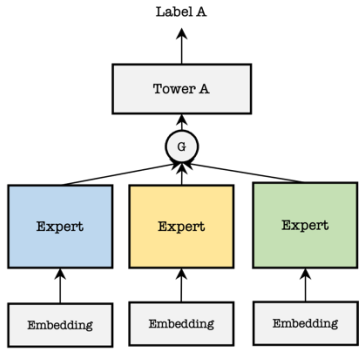
MoE



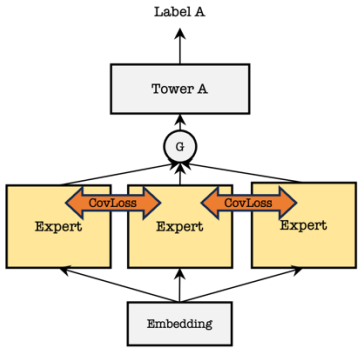
ME-MoE



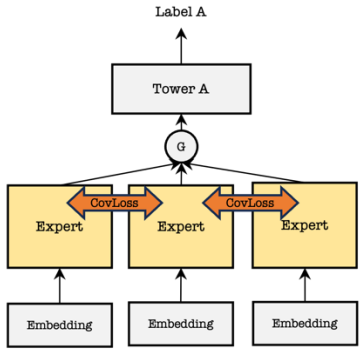
Hetero-MoE



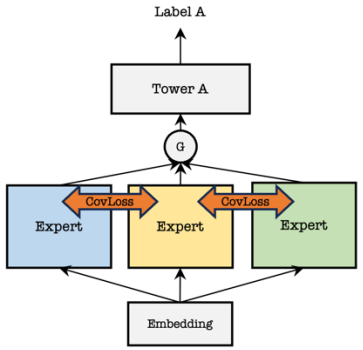
ME-Hetero-MoE



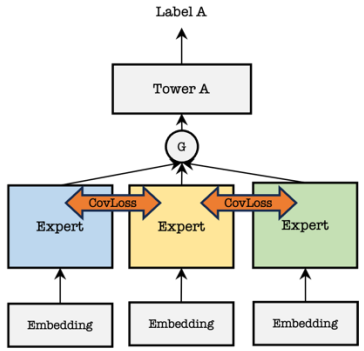
D-MoE



D-ME-MoE

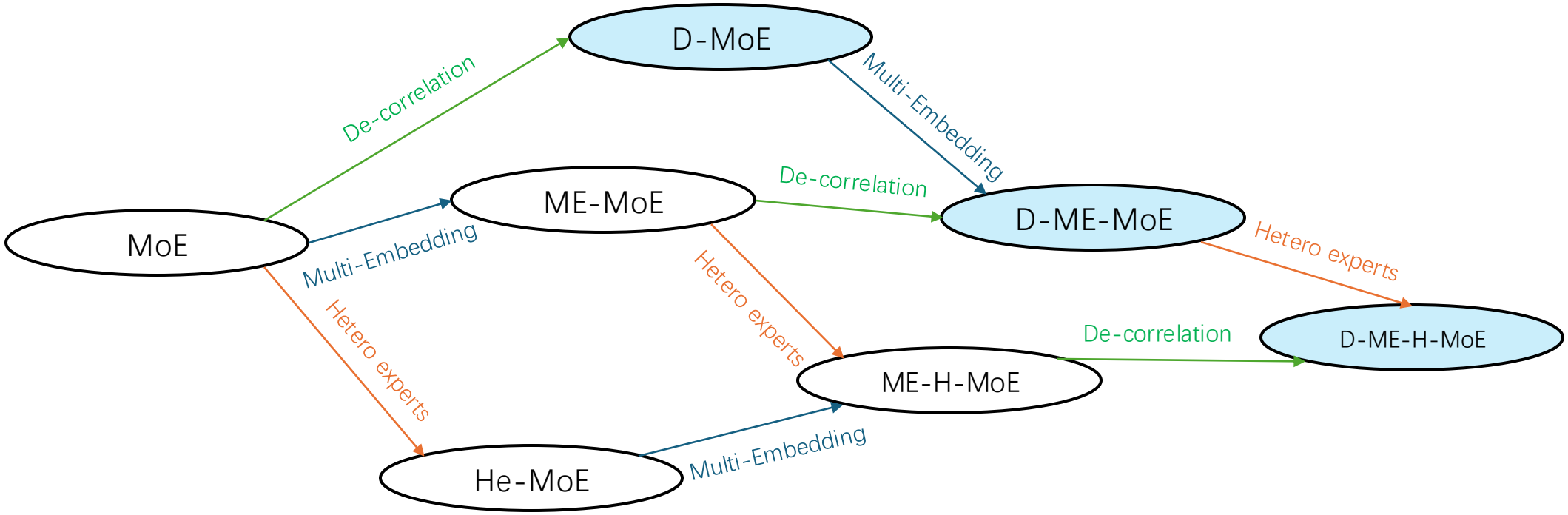


D-Hetero-MoE



D-ME-Hetero-MoE

MoE Evolution Map



- Part II, Prediction
  - Perspectives
  - Feature Interaction
  - Sequential Models
  - Multi-Task and Multi-Domain Learning
  - Large Recommendation Models
  - **LLM4Rec**

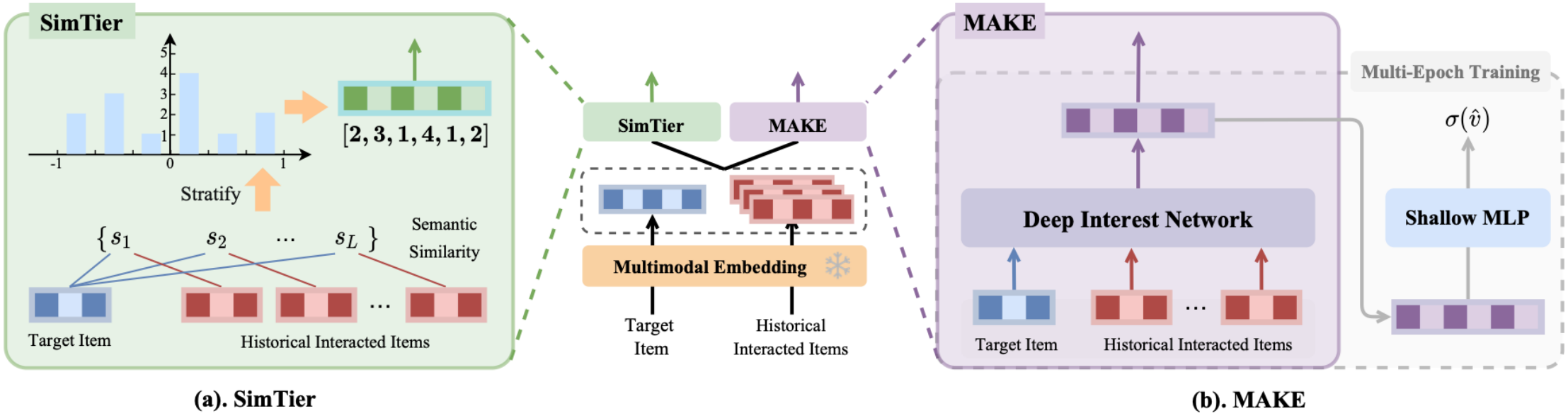
LLM4Rec

How to transfer the knowledge from LLM to RecSys?

- Distance Transfer
- Embedding Alignment
- Semantic IDs

SimTier

- Calculate the **similarity score** between multi-modal representation of behaviors and the target.
- **Histogram** of the scores in a N pre-defined buckets.
- Use the N-dimension vector as a new representation.



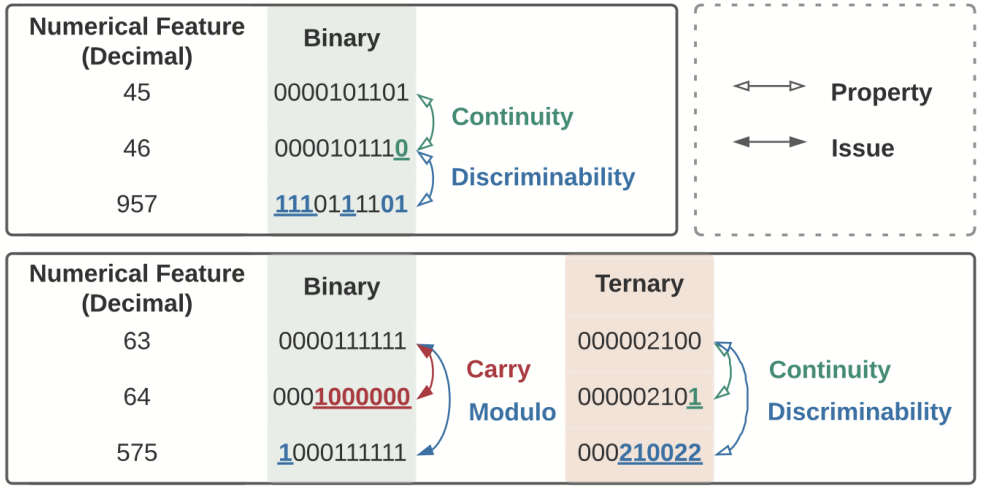
(a). SimTier (b). MAKE

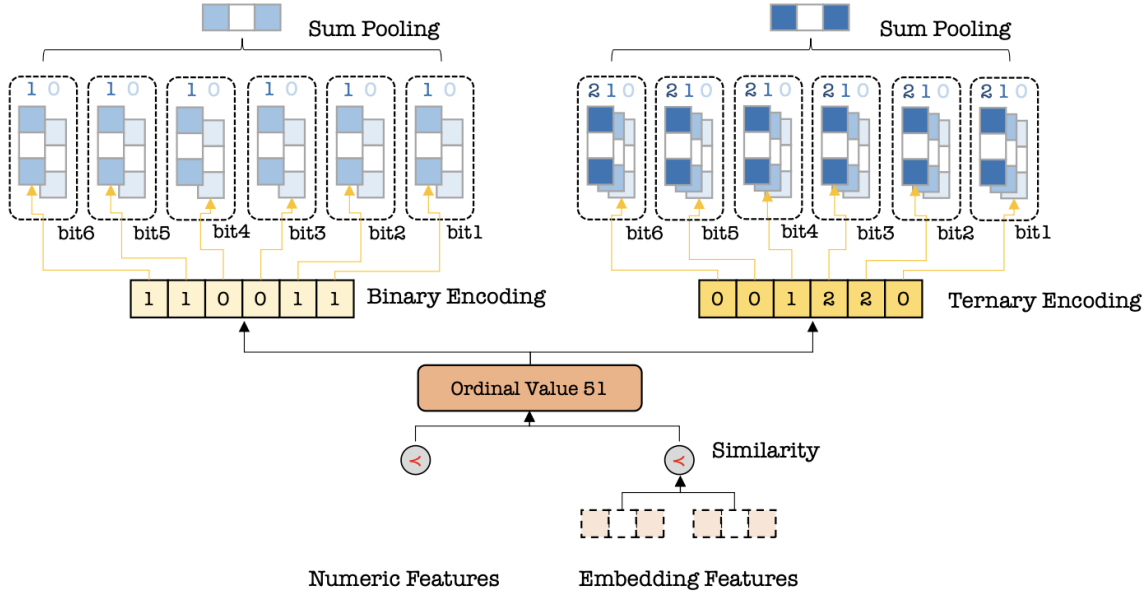
MNSE (Multiple Numeric Systems Encoding)

- Calculate the **distance** between two features based on the source embedding.
- Encoding the distance with **n-ary**
- Train the encoded embeddings in the target task, together with other embeddings in the target space.

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

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





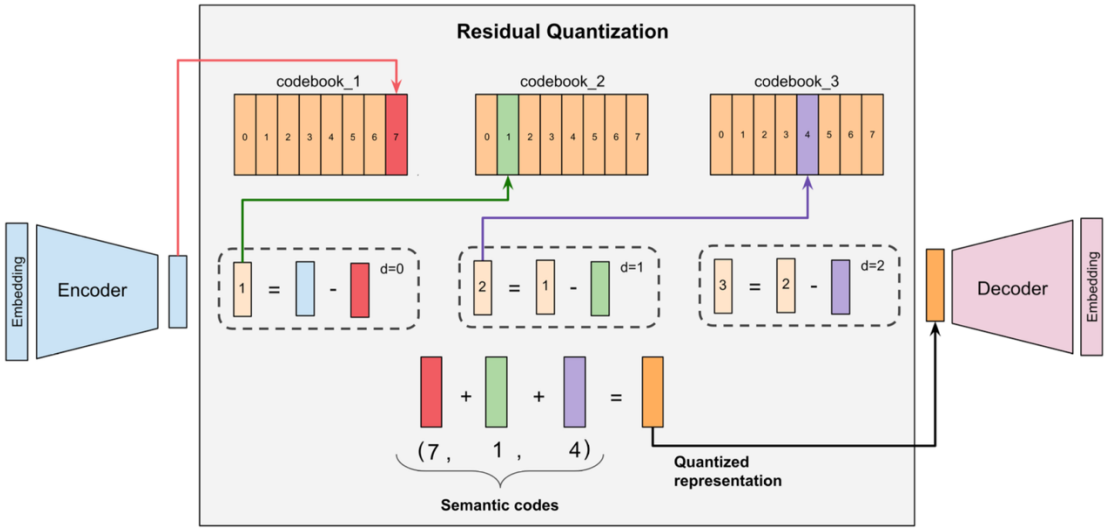
Key properties of n-ary: continuity, discriminability

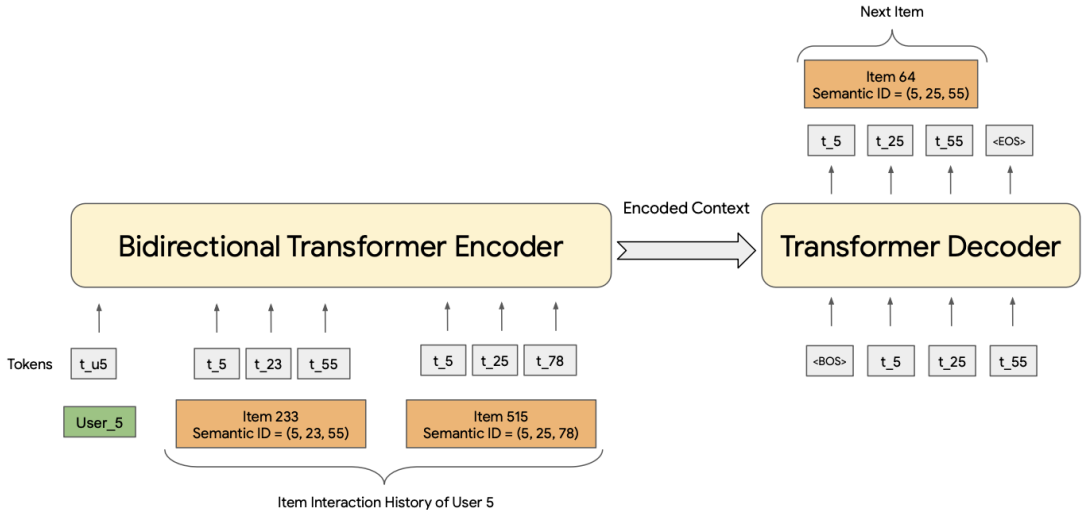
Ads Recommendation in a Collapsed and Entangled World. KDD 2024.
Numerical Feature Representation with Hybrid N-ary Encoding, Bo Chen. CIKM 2022.

Tencent

Tiger

- Use **RQ-VAE** on the LLM representation to get **semantic ID**s.
- Use semantic IDs in the downstream recommendation models.





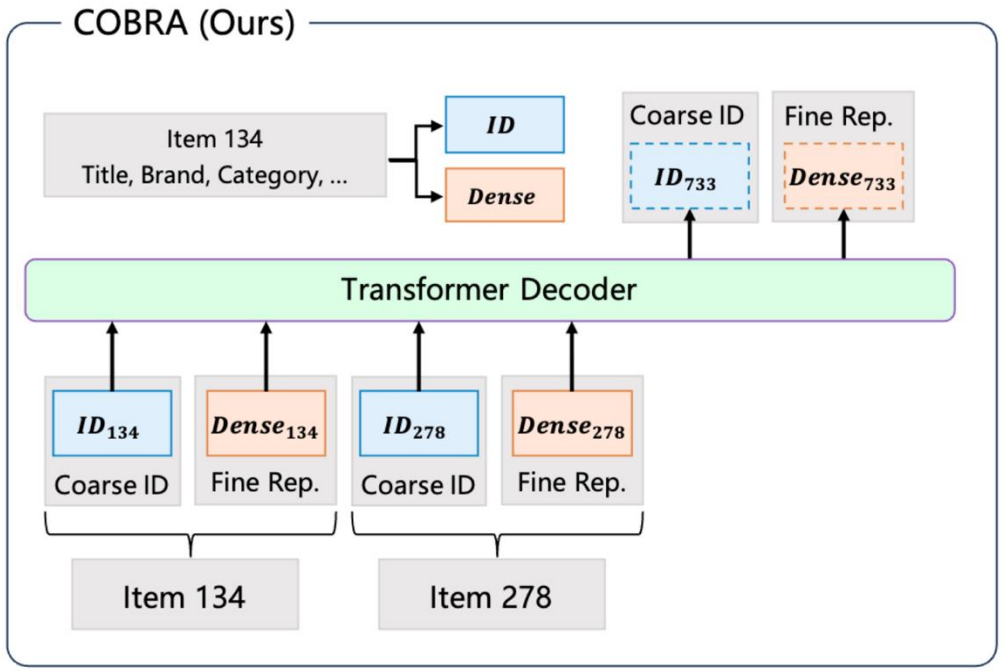
Why it works?

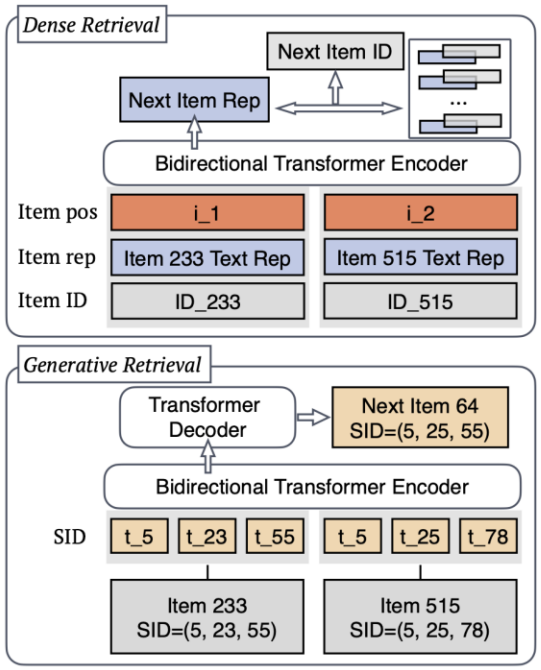
- **RQ-VAE** and **Semantic IDs** to capture the **source** space structures.
- **Semantic ID Embeddings** to align with the **target** space.

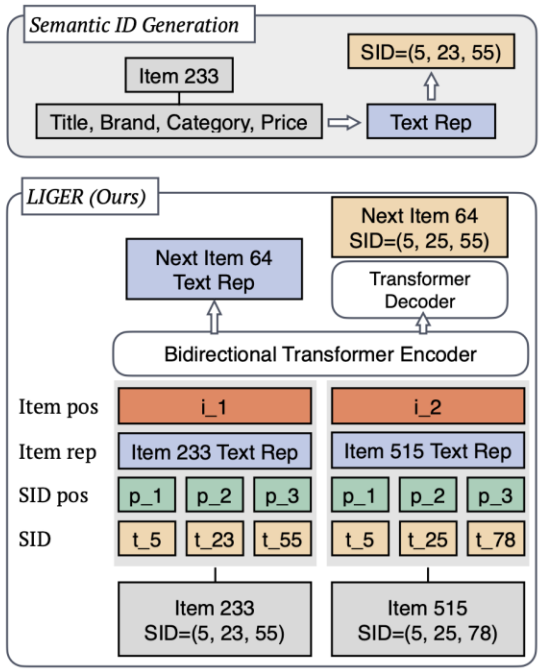
Recommender Systems with Generative Retrieval. NIPS 2023.

COBRA, LIGER

- Utilize both **sparse** (Semantic IDs through **RQ-VAE**) and **dense** representations to extract knowledge from the LLM.





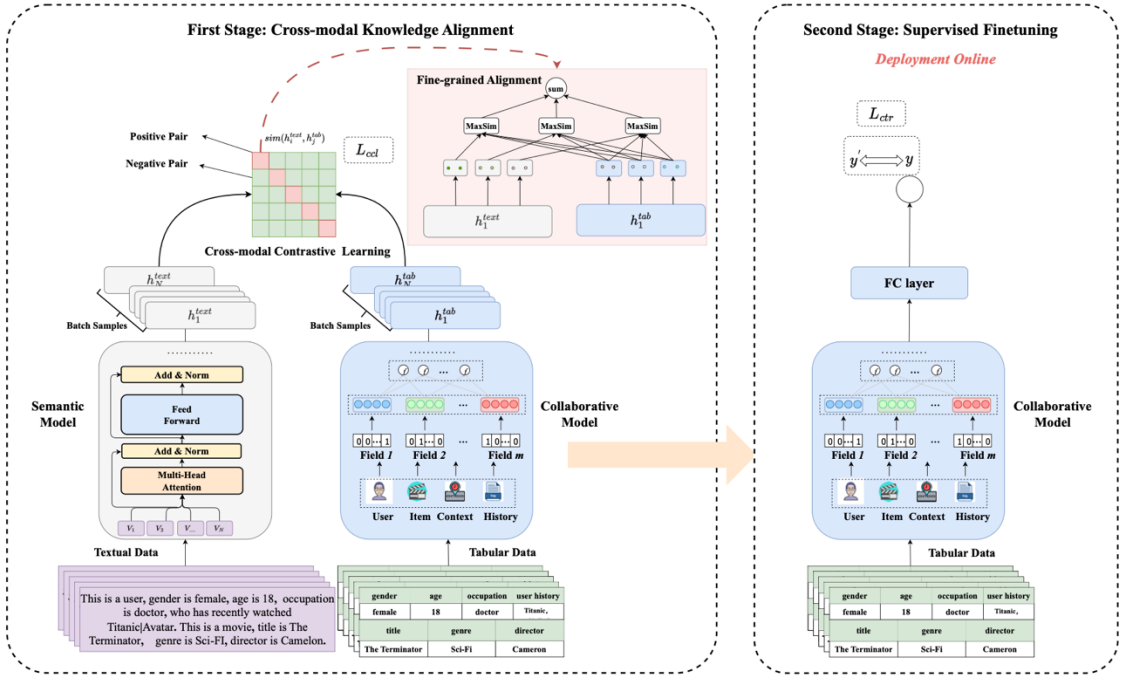


Sparse Meets Dense: Unified Generative Recommendations with Cascaded Sparse-Dense Representations. 2025.
Unifying Generative and Dense Retrieval for Sequential Recommendation. 2025.

CTRL

Taxtual-to-Tabular **contrastive loss**

$$\mathcal{L}^{textual2tabular} = -\frac{1}{N}\sum_{k=1}^{N} log\frac{exp(sim(\mathbf{h}_k^{text}, \mathbf{h}_k^{tab})/\tau)}{\sum_{j=1}^{N} exp(sim(\mathbf{h}_k^{text}, \mathbf{h}_j^{tab})/\tau)},$$



CTRL: Connect Collaborative and Language Model for CTR Prediction.

PAD (Pre-train, Align and Disentangle)

Adopt **MK-MMD** as the alignment loss to capture all information about the distribution.
Combine the alignment with the BCE loss to avoid **catastrophic forgetting**.

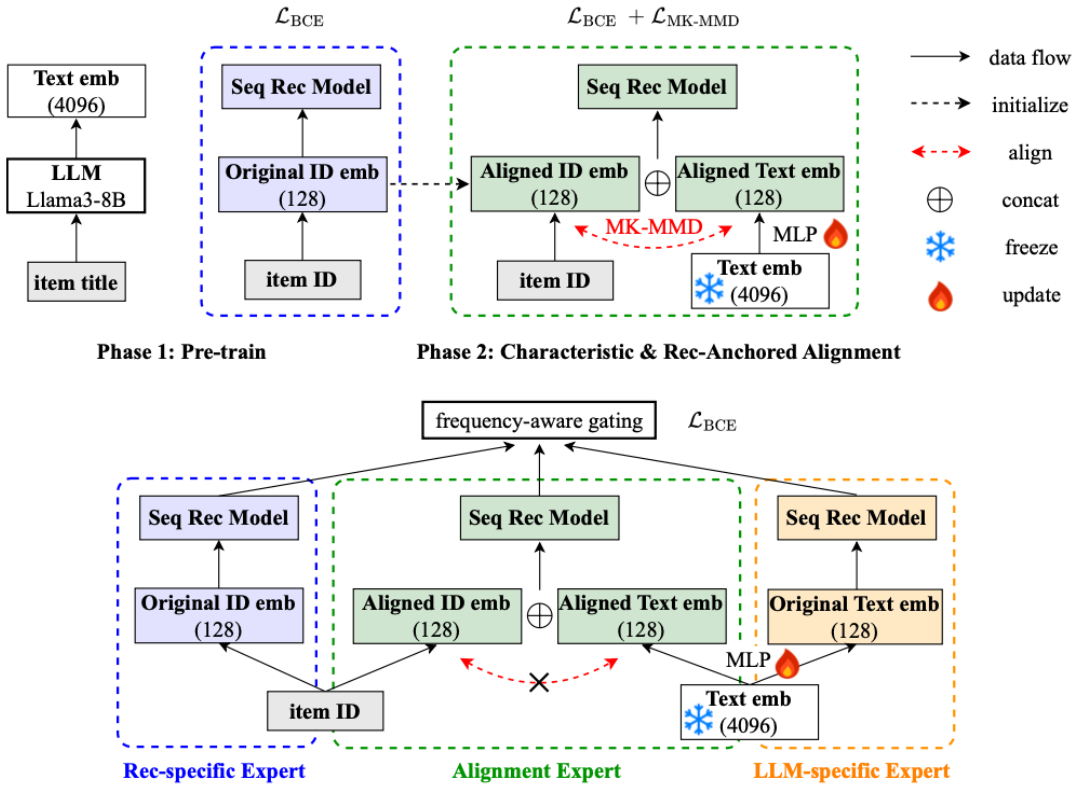
Avoid **catastrophic forgetting** **Space Alignment**

$$\mathcal{L} = \mathcal{L}_{\text{REC}} + \gamma \cdot \mathcal{L}_{\text{MK-MMD}}$$

$$\mathcal{L}_{\text{MK-MMD}} = D_k^2 \left(f_{\text{MLP}}(\text{SG}(\mathcal{D}_{\text{text}}), \boldsymbol{w}), \mathcal{D}_{\text{rec}} \right)$$

$$\mathcal{L}_{\text{REC}} = \frac{1}{n} \sum_{i=1}^{n} \text{BCE} \left(f_\theta \left(\{\mathbf{h}_i^s\}, \{\mathbf{h}_i^c\}, \mathbf{x}_i^s, \mathbf{x}_i^c \right), y_i \right)$$

$$\text{MK-MMD}^2(X_s, X_t) = \left\| \frac{1}{n} \sum_{i=1}^{n} \phi_k(x_s^i) - \frac{1}{m} \sum_{j=1}^{m} \phi_k(x_t^j) \right\|_{\mathcal{H}_k}^2$$



Pre-train, Align and Disentangle: Empowering Sequential Recommendation with Large Language Models. SIGIR 2025.

Tencent

Summary

- Perspectives
- Feature Interaction Models: Dimensional Collapse
- Sequential Models: Discriminability
- Multi-task/domain Learning: Disentanglement
- Large Recommendation Models
- LLM4Rec

References, I

- Factorization Machines. ICDM 2010.
- Field-aware Factorization Machines. RecSys 2016.
- **Field-weighted Factorization Machines for Click-Through Rate Prediction in Display Advertising. WWW 2018.**
- **FmFM: Field-matrixed Factorization Machines for CTR Prediction. WWW 2021.**
- **Towards Unifying Feature Interaction Models for Click-Through Rate Prediction. 2024.**
- Wide & Deep Learning for Recommender Systems. 2016.
- DeepFM: A Factorization-Machine based Neural Network for CTR Prediction. 2017.
- xDeepFM: Combining Explicit and Implicit Feature Interactions for Recommender Systems. KDD 2018.
- DCN v2- Improved Deep & Cross Network and Practical Lessons for Web-scale Learning to Rank Systems.
- Neural Collaborative Filtering vs. Matrix Factorization Revisited. RecSys 2020.
- **Understanding DNNs in Feature Interaction Models - A Dimensional Collapse Perspective. 2025.**
- **From Feature Interaction to Feature Generation: A Generative Paradigm of CTR Prediction Models. Under review.**
- **Deep Interest Network for Click-Through Rate Prediction. KDD 2018.**
- **Deep interest evolution network for click-through rate prediction. AAAI 2019.**

References, II

- **Deep Session Interest Network for Click-Through Rate Prediction. 2019.**
- **Temporal Interest Network for User Response Prediction. WWW 2024.**
- Search-based User Interest Modeling with Lifelong Sequential Behavior Data for Click-Through Rate Prediction. CIKM 2020.
- Sampling is all you need on modeling long-term user behaviors for CTR prediction. CIKM 2022.
- ETA: End-to-End User Behavior Retrieval in Click-Through Rate Prediction Model. 2021.
- TWIN: TWo-stage Interest Network for Lifelong User Behavior Modeling in CTR Prediction at Kuaishou.
- TWIN V2: Scaling Ultra-Long User Behavior Sequence Modeling for Enhanced CTR Prediction at Kuaishou. CIKM 2024.
- **Long-Sequence Recommendation Models need Decoupled Embeddings. ICLR 2025.**
- Lighter and Better: Low-Rank Decomposed Self-Attention Networks for Next-Item Recommendation. 2021.
- **Lifelong Sequential Modeling with Personalized Memorization for User Response Prediction. SIGIR 2019.**
- **Practice on Long Sequential User Behavior Modeling for Click-Through Rate Prediction. KDD 2019.**
- Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts. KDD 2018.
- Progressive Layered Extraction (PLE): A Novel Multi-Task Learning (MTL) Model for Personalized Recommendations. RecSys, 2020.
- PEPNet: Parameter and Embedding Personalized Network for Infusing with Personalized Prior Information. 2023.

References, III

- **STEM: Unleashing the Power of Embeddings for Multi-Task Recommendation. AAAI 2024.**
- **Ads Recommendation in a Collapsed and Entangled World. KDD 2024.**
- **Crocodile - Cross Experts Covariance for Disentangled Learning in Multi-Domain Recommendation. Under review.**
- Actions Speak Louder than Words: Trillion-Parameter Sequential Transducers for Generative Recommendations. ICML 2024.
- Wukong: Towards a Scaling Law for Large-Scale Recommendation. ICML 2024.
- **On the Embedding Collapse When Scaling Up Recommendation Models. ICML 2024.**
- DHEN: A Deep and Hierarchical Ensemble Network for Large-Scale Click-Through Rate Prediction. 2022.
- **Towards De-correlated Mixture-of-Experts for CTR Prediction. Under review.**
- **Enhancing Taobao Display Advertising with Multimodal Representations: Challenges, Approaches and Insights.**
- Recommender Systems with Generative Retrieval. NIPS 2023.
- Sparse Meets Dense: Unified Generative Recommendations with Cascaded Sparse-Dense Representations. 2025.
- CTRL: Connect Collaborative and Language Model for CTR Prediction.
- **Pre-train, Align and Disentangle: Empowering Sequential Recommendation with Large Language Models. SIGIR 2025.**

Thanks!

Q & A