wsdm 2021

THE 14TH ACM INTERNATIONAL CONFERENCE ON
WEB SEARCH AND DATA MINING

Multi-interactive Attention Network for Fine-grained Feature Learning in CTR Prediction

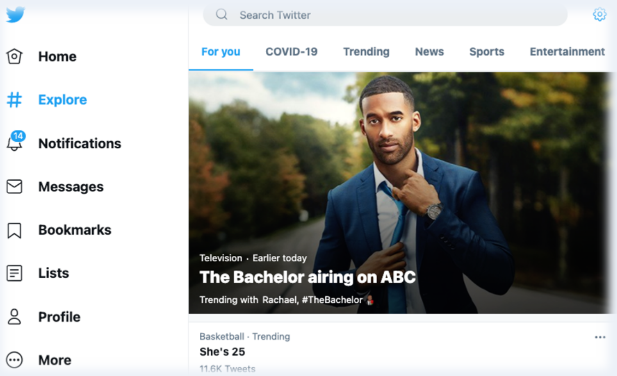
Kai Zhang[1], Hao Qian[3], Qing Cui[3], Qi Liu[1,2], Longfei Li[3], Jun Zhou[3], Jianhui Ma[1], Enhong Chen[1,2]

[1] Anhui Province Key Lab of Big Data Analysis and Application, School of Data Science

[2] School of Computer Science and Technology, University of Science and Technology of China

[3] Ant Group, Hangzhou, China

Consider some situations when you suffering online



Television · Earlier today

The Bachelor airing on ABC

Trending with Rachael, #TheBachelor

Basketball · Trending

She's 25

11.6K Tweets

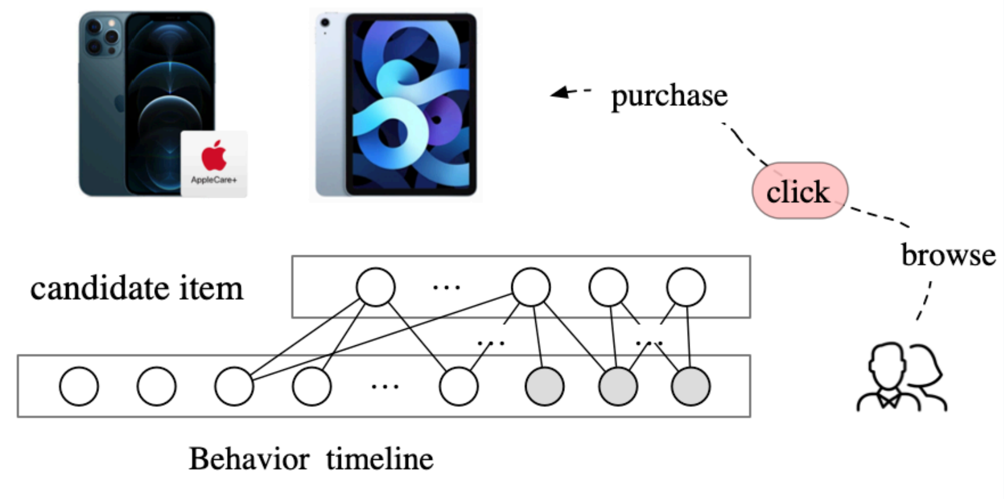
“How does this recommendation work?"

“How do they predict what we like to do in a certain scenario?"

Click-Though Rate (CTR) prediction

- CTR prediction, which aims to estimate the likelihood of a user clicking at an ad or an item.
- A high CTR score indicates that the candidate ads helpful and relevant.

- Data characteristics
  - categorical & value
  - high latitude
  - very sparse



Outline

1. Background
   a. Formal definition & previous methods
   b. Some existing problems
2. Our new : Fine-grained **feature learning**
3. Some motivating examples
4. Our new methods: Multi-interactive Attention Network
5. Experiments
6. Conclusion

Background – formal definition

To begin with, we first define the CTR prediction problem. It estimates the probability that a user clicks at candidate items based on the input feature representation.

General purpose:

$$\mathrm{CTR} = f(\,\cdot\,)$$

DNN & Sequential prediction function:

$$\mathrm{CTR} = f(\mathrm{candidate\_item}, history\_behavior, context, user\_profile)$$

Background – previous methods

Shallow & deep methods

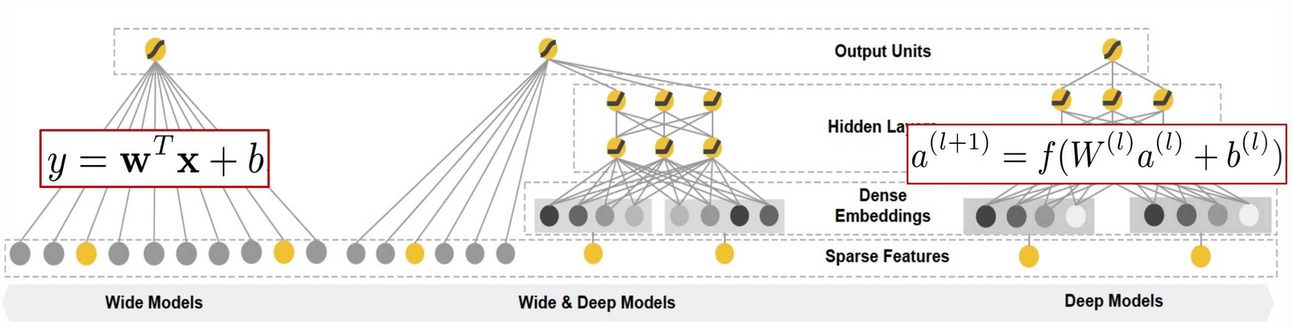
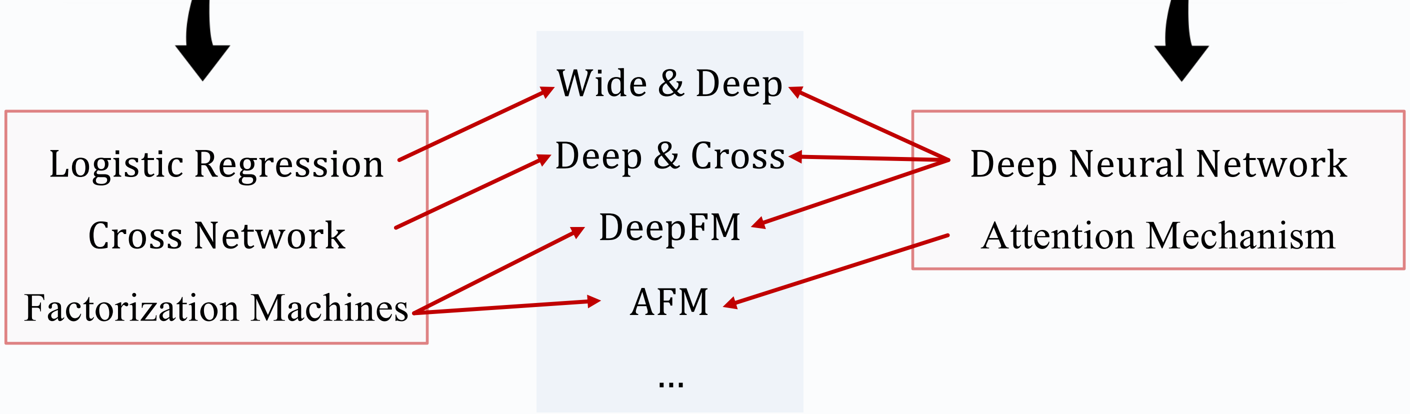
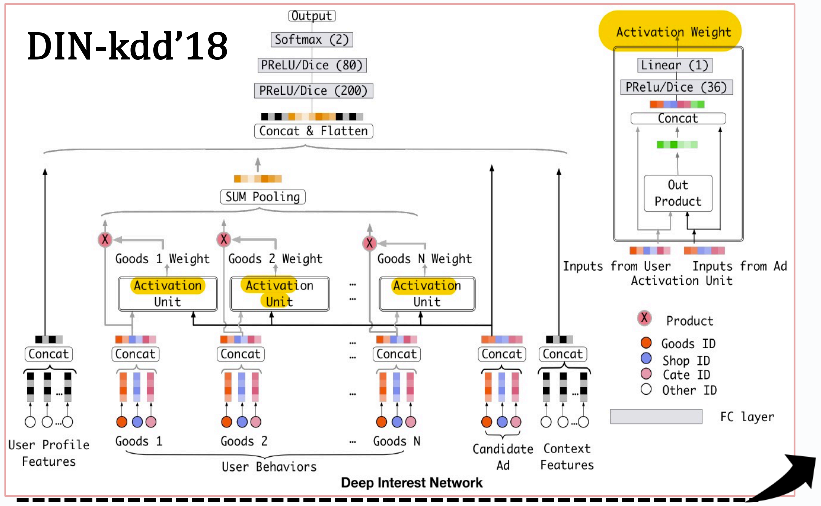


Figure 1: The spectrum of Wide & Deep models.



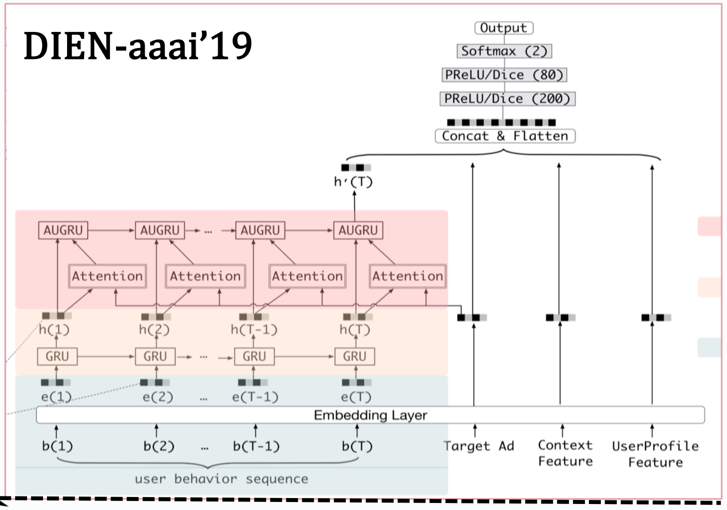
Background – previous methods

Sequential methods



DIN-kdd'18

Same features

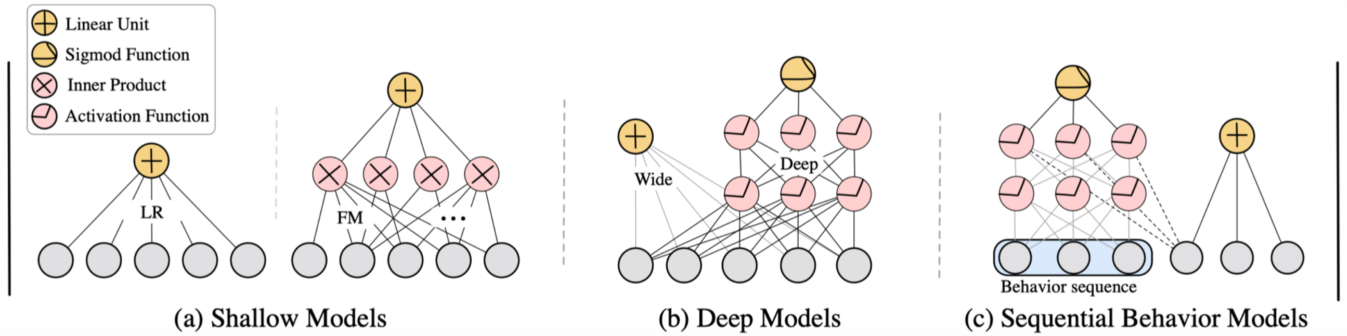


DIEN-aaai'19

1. Committed to fully exploring users' interests.
2. The attention mechanism is introduced to assign different weights to user behaviors, thus to capture user interest.

1. Assuming that users' interests change dynamically.
2. An AUGRU and attention mechanism are designed to capture the dynamic changes.

Background – existing problems



(a) Shallow Models (b) Deep Models (c) Sequential Behavior Models

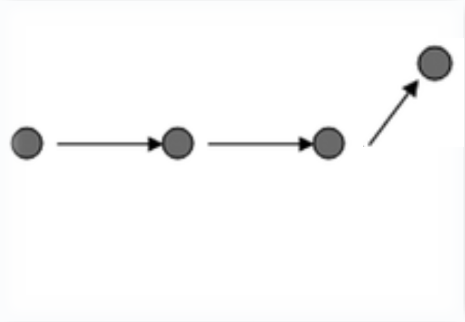
Shallow & deep methods

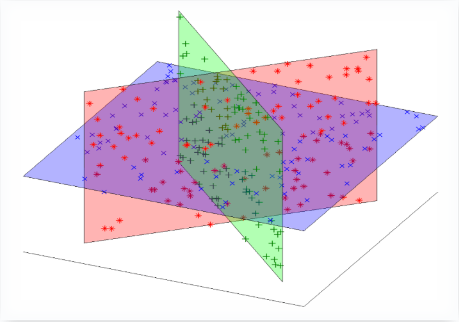
- Could not effectively learn the user’s interests and preferences from the user’s historical behaviors.

Sequential methods

- No feature is divided into finer granularity for interactive learning of features.

Background – existing problems





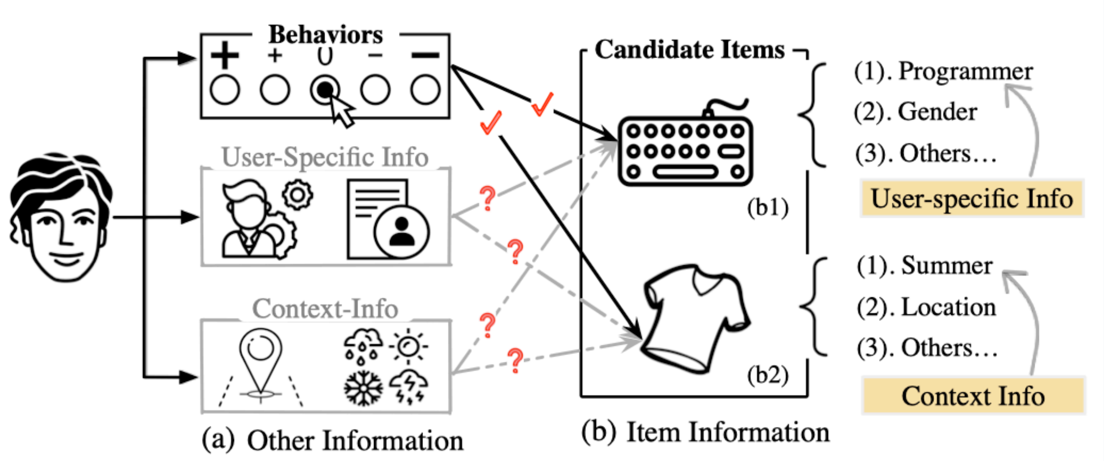
In real prediction scenario, still suffer from some limitations：

1. Most methods utilize attention on the behaviors, which may mislead the CTR prediction because users often click on new products that are irrelevant to any historical behaviors.
2. There are numerous users that have operations (i.e., behaviors) a long time ago, but turn relatively inactive in recent times.
3. Multiple representations of user’s historical behaviors in different feature subspaces are largely ignored.

Outline

1. Background
   a. Formal definition & previous methods
   b. Some existing problems
2. **Our new solution: Fine-grained feature learning**
3. Some motivating examples
4. Our new methods: Multi-interactive Attention Network
5. Experiments
6. Conclusion

Our new solution – fine-grained feature learning



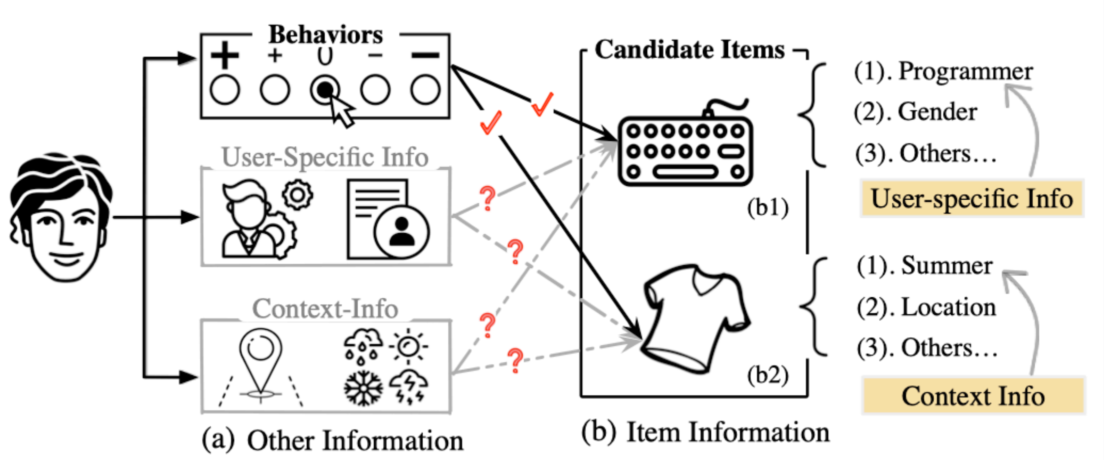
e.g., age, gender, and occupations

e.g., weather, city and location

Exploring the Fine-grained attributes：

1. There exists a large amount of user-specific and context information.
2. This fine-grained information provides various clues to infer the user’s current state.

Some motivating examples

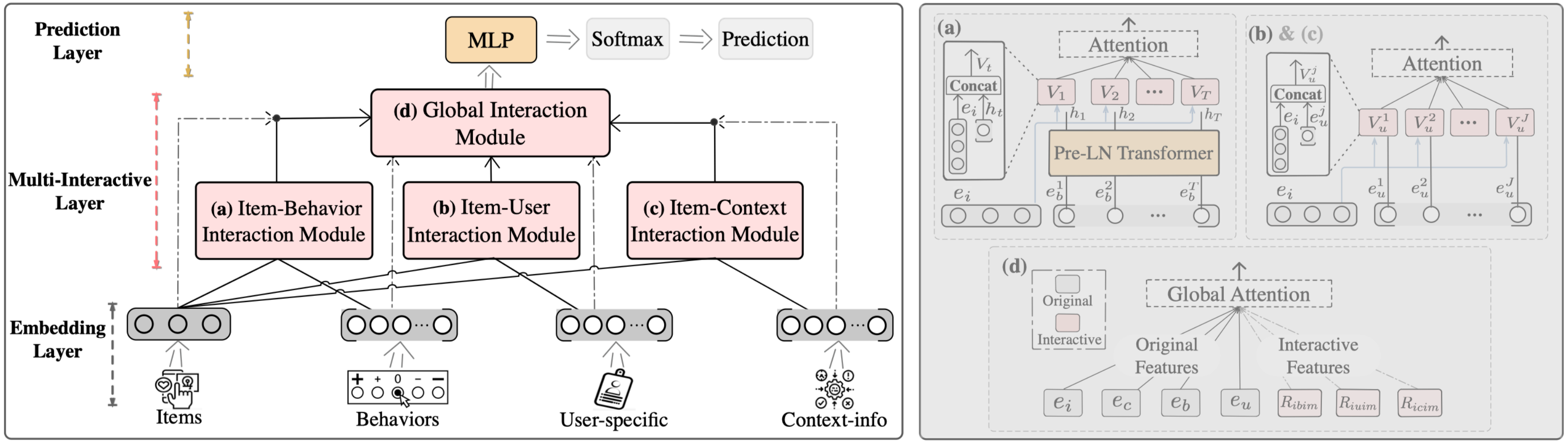


1. In the above Figure b1, the candidate item "mechanical keyboard" may be more relevant to users' current occupation "programmer" which is hard to represent in the historical behaviors. (user's different preference)
2. In the above Figure b2, the "T-shirt" may be activated as a user's behavior representation in "summer", rather than "winter". (different semantic subspace)

Outline

1. Background
    a. Formal definition & previous methods
    b. Some existing problems
2. Our new solution: Fine-grained feature learning
3. Some motivating examples
4. **Our new methods: Multi-interactive Attention Network**
5. Experiments
6. Conclusion

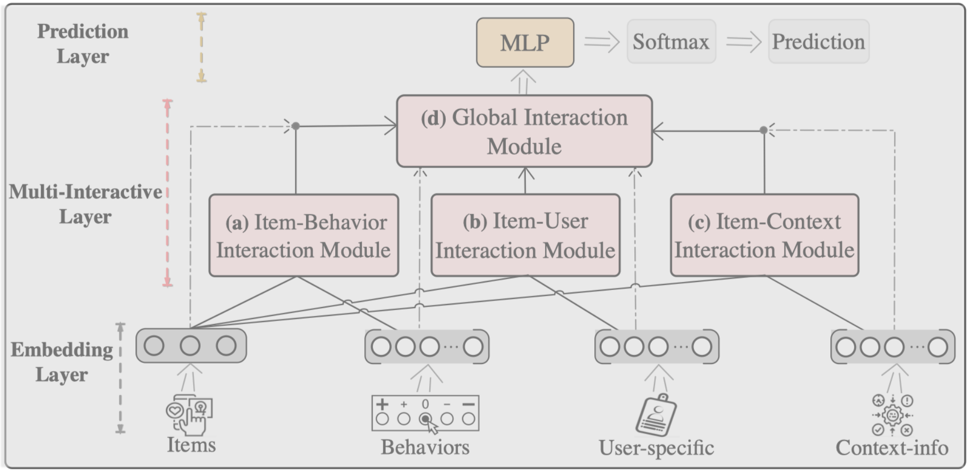
Our new method – multi-interactive attention network

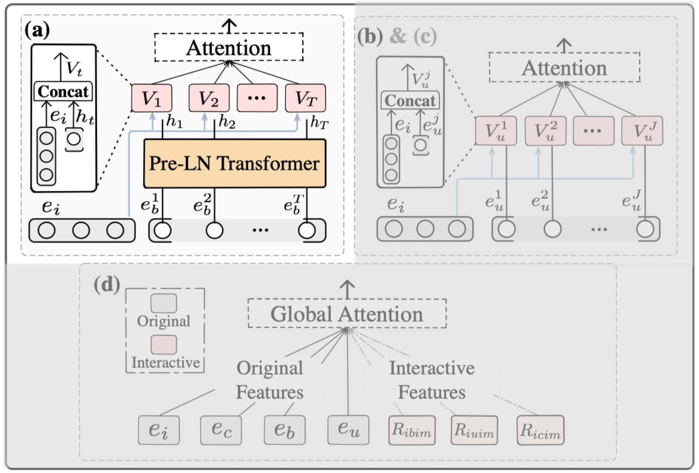


Overall architecture of MIAN

1. a common y used Embedding Layer (feature embedding);
2. a novel Multi-interactive Layer, i.e., contains (a), (b), (c) and (d);
3. and a general Prediction Layer (MLP + Softmax).

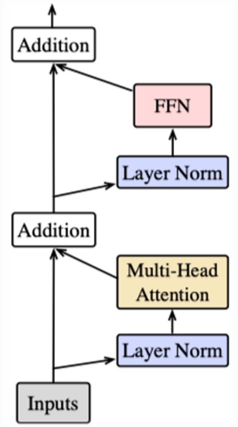
MIAN – item-behaviors interaction module (IBIM)


Prediction Layer
MLP
Softmax
Prediction
(d) Global Interaction Module
Multi-Interactive Layer
(a) Item-Behavior Interaction Module
(b) Item-User Interaction Module
(c) Item-Context Interaction Module
Embedding Layer
Items
Behaviors
User-specific
Context-info

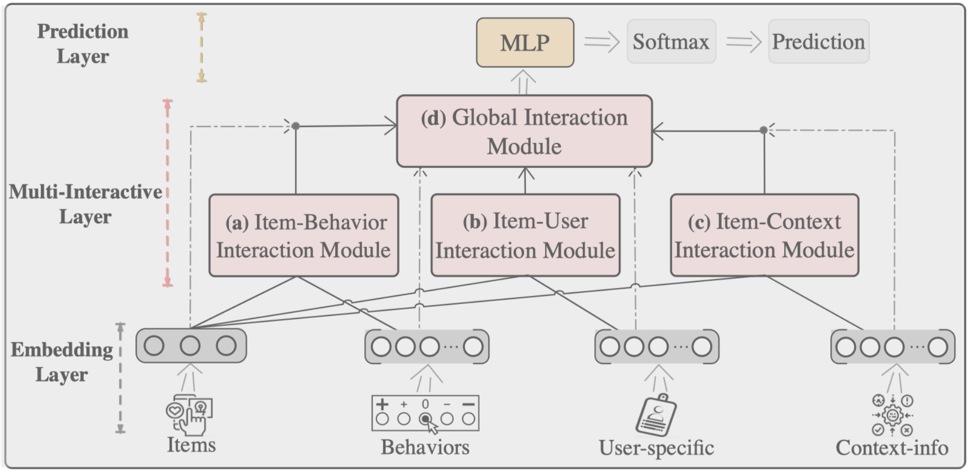


IBIM module's architecture

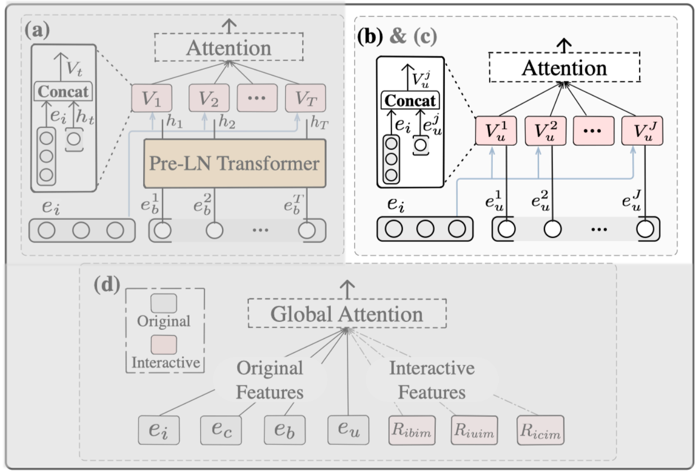
1. Just like DIN or DIEN, mining the users' behavior preference;
2. Pre-LN Transformer + Attention;
3. The output of IBIM is represented as R_{ibim}.



MIAN – item- context/user interaction module (ICIM / IUIM)



Prediction Layer
MLP
Softmax
Prediction
Multi-Interactive Layer
(d) Global Interaction Module
(a) Item-Behavior Interaction Module
(b) Item-User Interaction Module
(c) Item-Context Interaction Module
Embedding Layer
Items
Behaviors
User-specific
Context-info



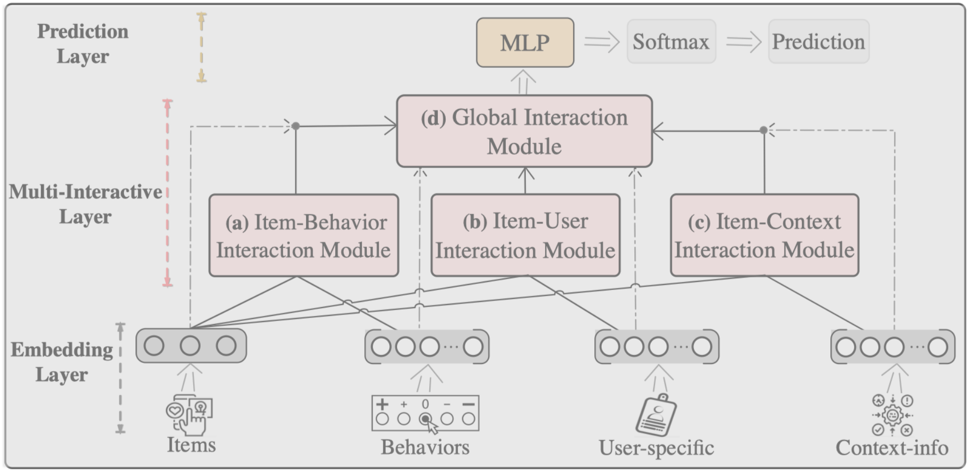
ICIM & IUIM modules' architecture (very similar)

1. Concatenate each fine-grained vector (context / user-specific) with candidate item embedding;
2. Calculate the attention score;

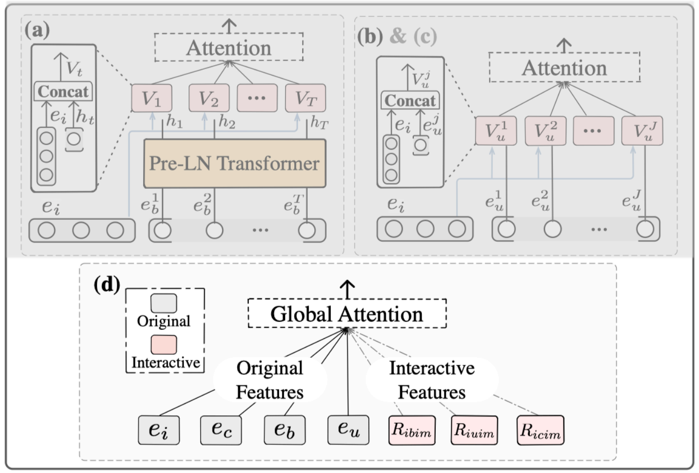
$$R_{icim} = \sum_{k=1}^{K} \frac{exp(tanh(V_c^k \cdot W_k + \widehat{b}_k))}{\sum_{k=1}^{K} exp(tanh(V_c^k \cdot W_k + \widehat{b}_k))} V_c^k$$

3. The output of ICIM/IUIM is represented as R_{icim} and R_{iuim}.

MIAN – global interaction module (GIM)



Prediction Layer
MLP
Softmax
Prediction
Multi-Interactive Layer
(d) Global Interaction Module
(a) Item-Behavior Interaction Module
(b) Item-User Interaction Module
(c) Item-Context Interaction Module
Embedding Layer
Items
Behaviors
User-specific
Context-info



1. Connect all feature vectors as the input to global attention;

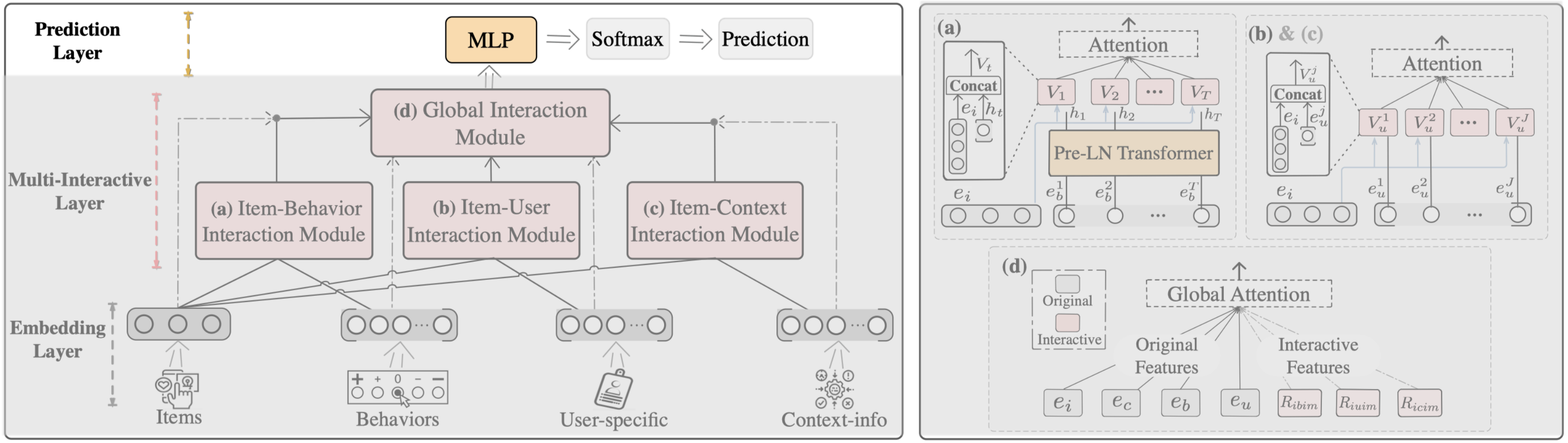
   (e_b、 R_{ibim}、 e_i、 R_{iuim}、 e_u、 R_{icim}、 e_c)

2. Captures the implicit relationship between the original features and the generated interaction features.

$$\begin{aligned} r_g &= [e_b;\ R_{ibim};\ e_i;\ R_{iuim};\ e_u;\ R_{icim};\ e_c] \\ &= [r_1;\ r_2;\ r_3;\ r_4;\ r_5;\ r_6;\ r_7]. \end{aligned}$$

$$R_g = \sum_{l=1}^{L} \frac{exp(tanh(r_l \cdot W_l\ +\ \hat{b_l}))}{\sum_{l'=1}^{n} exp(tanh(r_{l'} \cdot W_{l'}\ +\ \hat{b}_{1'}))}\ r_l\ .$$

MIAN – DNN prediction layer



1. Multi-layer perceptron network;
2. Make final CTR predictions through Softmax function.

$$\begin{aligned}
R_1 &= Relu(W_1 R_g + b_1), \\
R_2 &= Relu(W_2 R_1 + b_2), \\
\cdots \quad & \cdots \quad \cdots \\
R_h &= Relu(W_h R_{h-1} + b_h),
\end{aligned}$$

$$\hat{y} = softmax(W_q R_h + b_q).$$

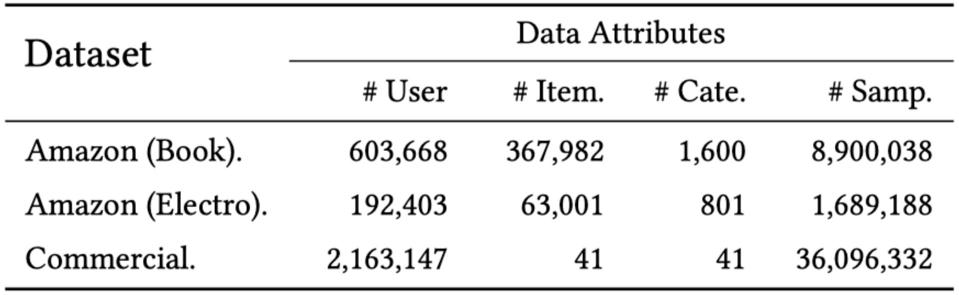
Outline

1. Background
   a. Formal definition & previous methods
   b. Some existing problems
2. Our new solution: Fine-grained feature learning
3. Some motivating examples
4. Our new methods: Multi-interactive Attention Network
5. **Experiments**
6. Conclusion

Experiments

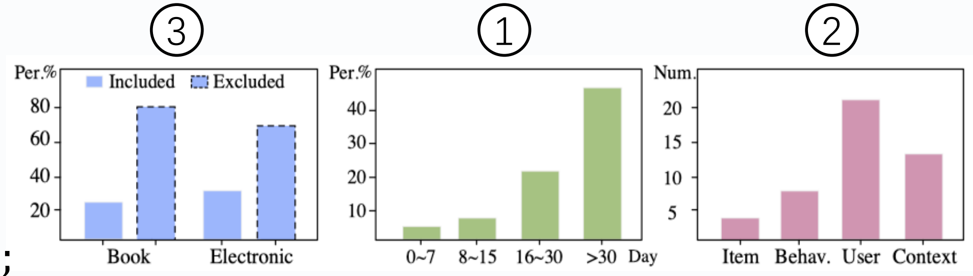
Dataset

1. Amazon: Book and Electronics
    - Two subset collected from Amazon.com；
2. Commercial: Alipay recommendation system;



Dataset	Data Attributes			
	# User	# Item.	# Cate.	# Samp.
Amazon (Book).	603,668	367,982	1,600	8,900,038
Amazon (Electro).	192,403	63,001	801	1,689,188
Commercial.	2,163,147	41	41	36,096,332

Data analysis



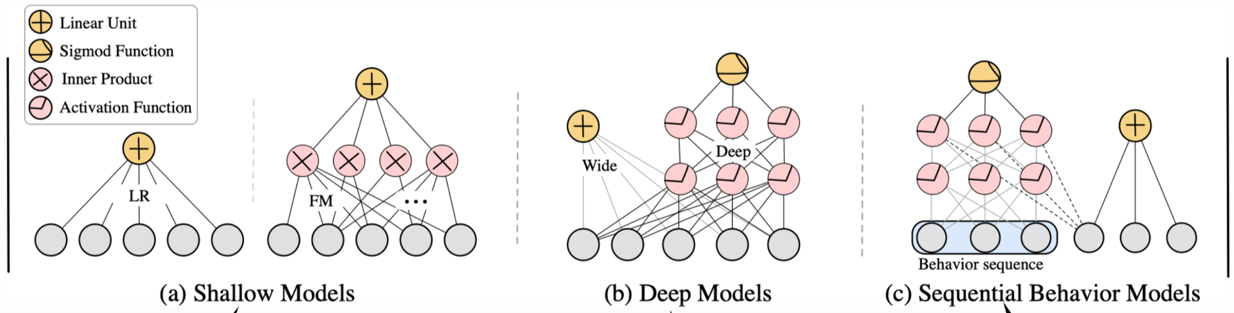
① 50% of user behavior occurred 30 days ago;

② Exist a large amount of fine-grained information;

③ There are a considerable fraction of candidate items that are irrelevant (i.e., not appeared) to any historical behaviors in Amazon datasets;

Experiments

Baseline methods



(a) Shallow Models (b) Deep Models (c) Sequential Behavior Models

- **LR:** a widely used linear transformation baseline;
- **Wide&Deep:** jointly trains a linear model and a deep MLP model;

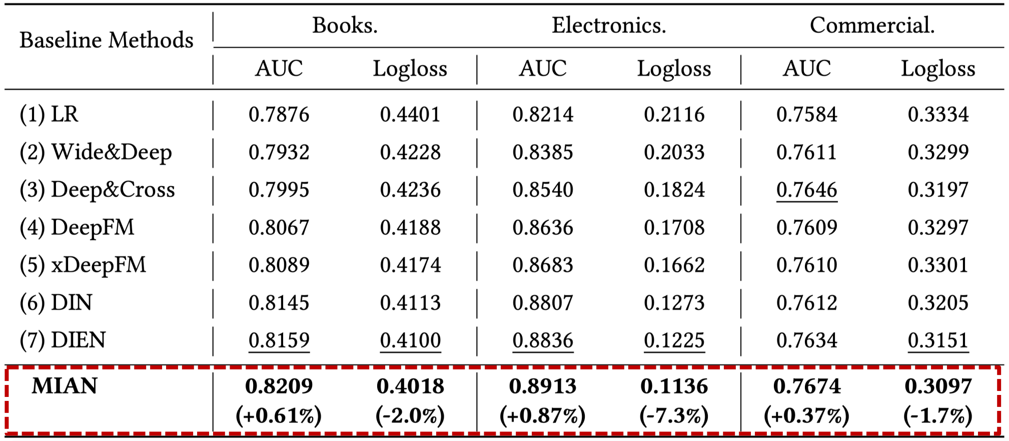
- **Deep&Cross:** to handle and learn cross high-order features;
- **DeepFM:** combines the explicit high-order interaction with MLP and traditional FM;
- **xDeepFM:** compress interaction network to enumerate and compress all feature interactions;

- **DIN:** exploits users' historical behaviors through the attention mechanism;
- **DIEN:** integrates GRUs with attention mechanism to capture users' involved interests;

Experiments

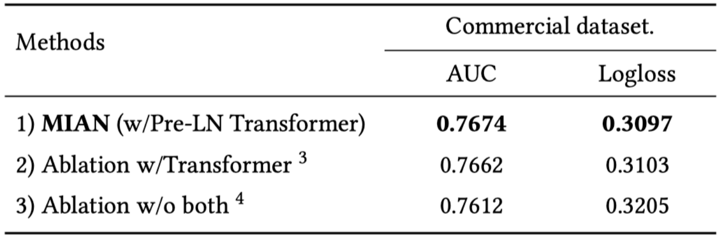
Overall performance

1. DIN and DIEN perform better than the other baselines (shallow/deep -based);
2. MIAN always has the best performance on all three datasets over all the metrics;

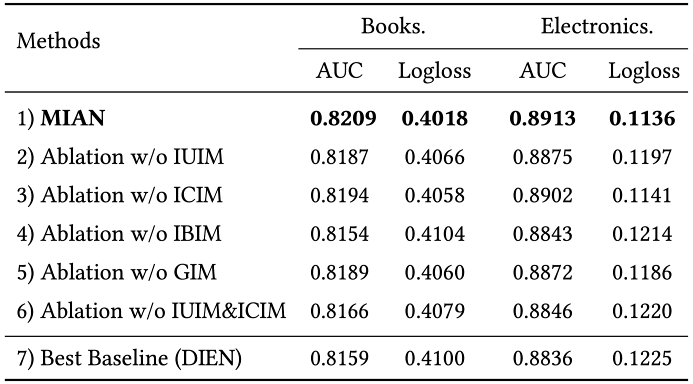


Baseline Methods	Books.		Electronics.		Commercial.	
	AUC	Logloss	AUC	Logloss	AUC	Logloss
(1) LR	0.7876	0.4401	0.8214	0.2116	0.7584	0.3334
(2) Wide&Deep	0.7932	0.4228	0.8385	0.2033	0.7611	0.3299
(3) Deep&Cross	0.7995	0.4236	0.8540	0.1824	0.7646	0.3197
(4) DeepFM	0.8067	0.4188	0.8636	0.1708	0.7609	0.3297
(5) xDeepFM	0.8089	0.4174	0.8683	0.1662	0.7610	0.3301
(6) DIN	0.8145	0.4113	0.8807	0.1273	0.7612	0.3205
(7) DIEN	0.8159	0.4100	0.8836	0.1225	0.7634	0.3151
MIAN	**0.8209** **(+0.61%)**	**0.4018** **(-2.0%)**	**0.8913** **(+0.87%)**	**0.1136** **(-7.3%)**	**0.7674** **(+0.37%)**	**0.3097** **(-1.7%)**

Ablation studies clearly demonstrate the effectiveness of each component (i.e., IBIM, ICIM, IUIM, GIM and Pre-Transformer).

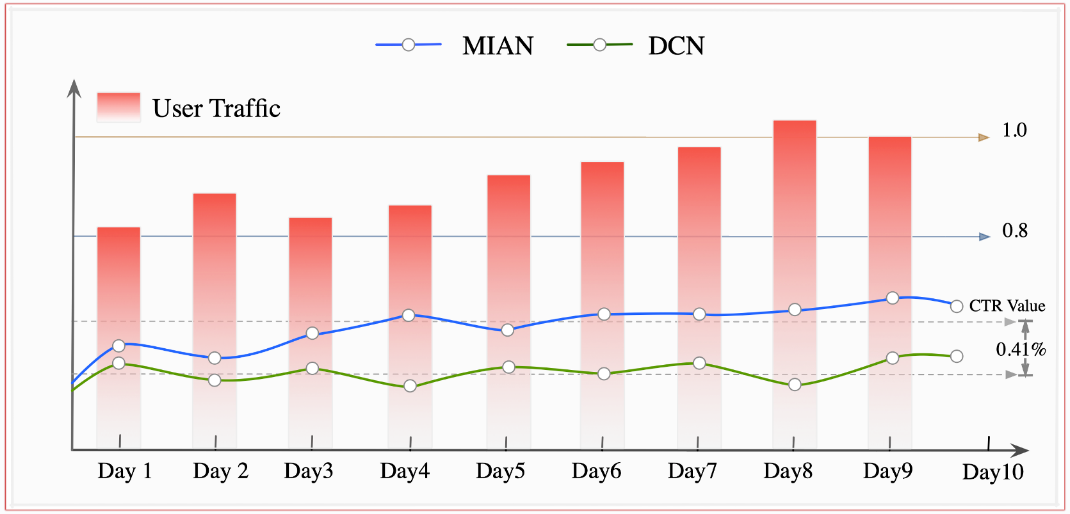


Methods	Commercial dataset.	
	AUC	Logloss
1) **MIAN** (w/Pre-LN Transformer)	**0.7674**	**0.3097**
2) Ablation w/Transformer [3]	0.7662	0.3103
3) Ablation w/o both [4]	0.7612	0.3205



Methods	Books.		Electronics.	
	AUC	Logloss	AUC	Logloss
1) **MIAN**	**0.8209**	**0.4018**	**0.8913**	**0.1136**
2) Ablation w/o IUIM	0.8187	0.4066	0.8875	0.1197
3) Ablation w/o ICIM	0.8194	0.4058	0.8902	0.1141
4) Ablation w/o IBIM	0.8154	0.4104	0.8843	0.1214
5) Ablation w/o GIM	0.8189	0.4060	0.8872	0.1186
6) Ablation w/o IUIM&ICIM	0.8166	0.4079	0.8846	0.1220
7) Best Baseline (DIEN)	0.8159	0.4100	0.8836	0.1225

Experiments

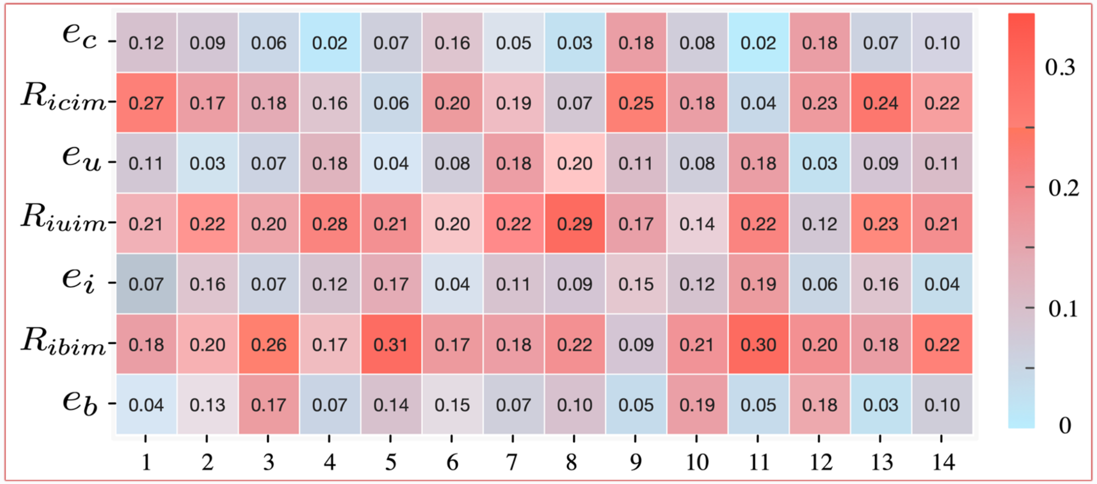


Online A/B testing

1. We conducted an online A/B testing in the production environment of Alipay for 10 days;
2. MIAN model brings a 0.41% gain in CTR while a 0.27% drop in cost which contributes a considerable business revenue growth;
3. Applied in multiple scenarios of Alipay;

Experiments

Visualization



	1	2	3	4	5	6	7	8	9	10	11	12	13	14
e_c	0.12	0.09	0.06	0.02	0.07	0.16	0.05	0.03	0.18	0.08	0.02	0.18	0.07	0.10
R_{icim}	0.27	0.17	0.18	0.16	0.06	0.20	0.19	0.07	0.25	0.18	0.04	0.23	0.24	0.22
e_u	0.11	0.03	0.07	0.18	0.04	0.08	0.18	0.20	0.11	0.08	0.18	0.03	0.09	0.11
R_{iuim}	0.21	0.22	0.20	0.28	0.21	0.20	0.22	0.29	0.17	0.14	0.22	0.12	0.23	0.21
e_i	0.07	0.16	0.07	0.12	0.17	0.04	0.11	0.09	0.15	0.12	0.19	0.06	0.16	0.04
R_{ibim}	0.18	0.20	0.26	0.17	0.31	0.17	0.18	0.22	0.09	0.21	0.30	0.20	0.18	0.22
e_b	0.04	0.13	0.17	0.07	0.14	0.15	0.07	0.10	0.05	0.19	0.05	0.18	0.03	0.10

1. We randomly select 14 cases from the Amazon dataset and visualize the attention weights in the global attention module;
2. The attention weights of interactive features “R_{ibim} ”,“ R_{iuim} ” and “R_{icim} ” are much larger than the original features, i.e., e_i , e_b , e_u , e_c .
3. The visualization results not only demonstrate the importance of the fine-grained feature learning but also indicate that MIAN is able to learn deeply interactive associations.

Outline

1. Background
    a. Formal definition & previous methods
    b. Some existing problems
2. Our new solution: Fine-grained feature learning
3. Some motivating examples
4. Our new methods: Multi-interactive Attention Network
5. Experiments
6. **Conclusion**

Conclusion

1. We point out some existing problems in the real CTR scenario and propose to study the problems via multiple fine-grained feature learning, that, to our knowledge, has not been explicitly and fully modeled by previous CTR methods.
2. We design a novel MIAN model, which contains a multi-interactive layer for fine-grained feature interactive learning, and a Transformer-based module to extract multiple meaning of user behavior in different feature subspaces.
3. Offline&Online experiments as well as the ablation studies illustrate the effectiveness and interpretability of each module, which may bring insights for future work.

References

[1] Steffen Rendle. 2010. Factorization machines. In 2010 IEEE ICDM'10, 995–1000.

[2] Heng-Tze Cheng, Levent Koc, et al. 2016. Wide & deep learning for recommender systems. In Proceedings of the 1st workshop on deep learning for recommender systems. ACM, 7–10.

[3] Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. 2017. Deep & cross network for ad click predictions. In Proceedings of the ADKDD'17. ACM, 12.

[4] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. arXiv preprint arXiv:1703.04247 (2017).

[5] Guorui Zhou, Xiaoqiang Zhu, et al. 2018. Deep interest network for click-through rate prediction. In Proceedings of the 24th ACM SIGKDD'18, 1059–1068.

[6] Guorui Zhou, Na Mou, and et al. 2019. Deep interest evolution network for click-through rate prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33. 5941–5948.

Thanks !