On the Embedding Collapse When Scaling Up Recommendation Models

Xingzhuo Guo, Junwei Pan, Ximei Wang, Baixu Chen, Jie Jiang, Mingsheng Long

Recommendation Models Background

▶ Recommendation Models

- ▶ Predict users' action based on features of users/item based on a large amount of data.
- ▶ Embedding / Feature Interaction / Post Processing

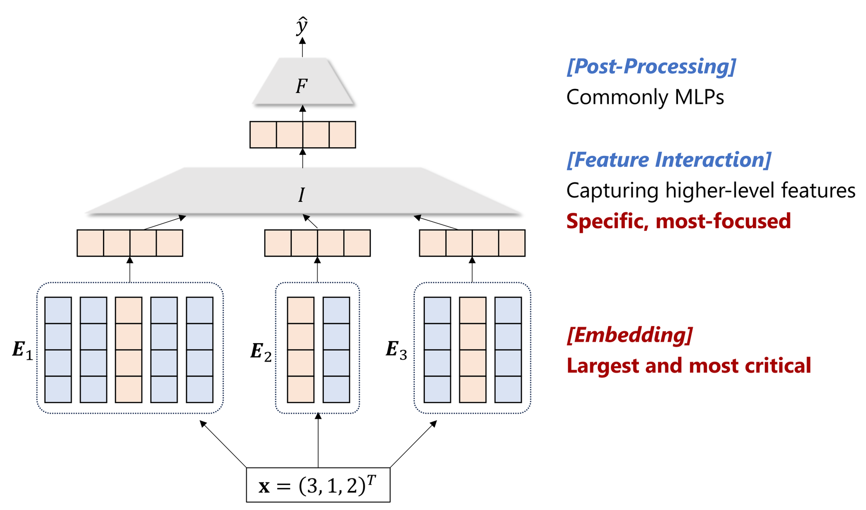
▶ Deficiency in Model Scalability

- ▶ Existing embedding sizes are too small.

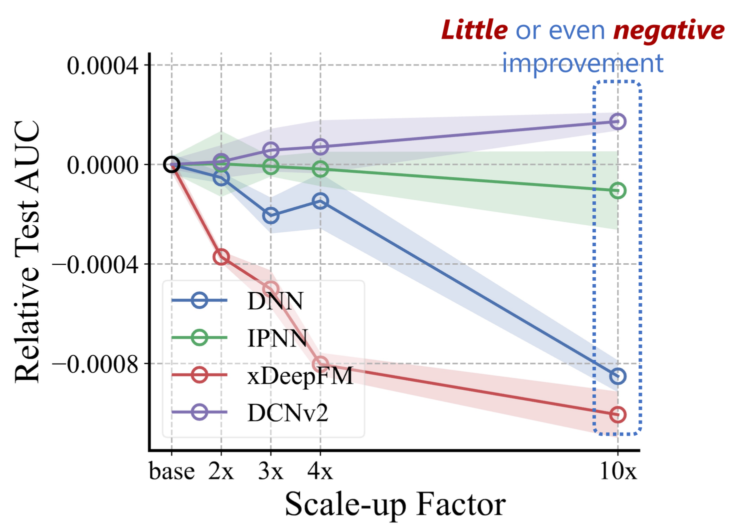
Recommendation Models	JL-Lemma	Embeddings in LLM
➤ $10^5 \sim 10^9$ of features ➤ Embedding size of $10 \sim 100$	➤ $K \geq 8\epsilon^{-2} \ln D$	➤ $\sim 10^5$ tokens ➤ $\sim 10^3$ embedding size

mismatch

- ▶ Scaling up recommendation models does not necessarily lead to performance gain.



Recommendation Model Architectures.



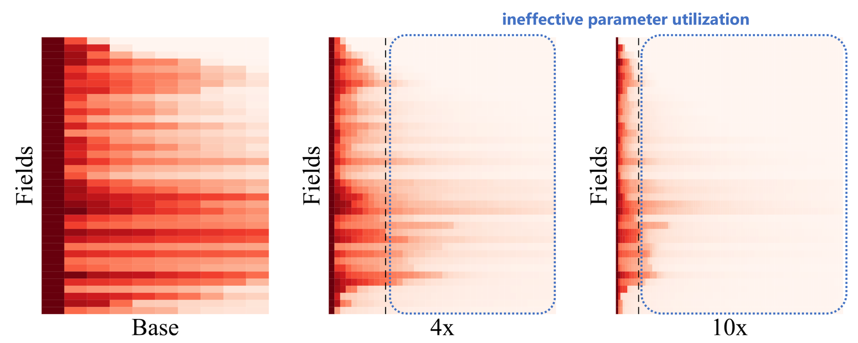
Scaling up Recommendation Models.

▶ Question: What's behind the deficiency in recommendation model scalability?

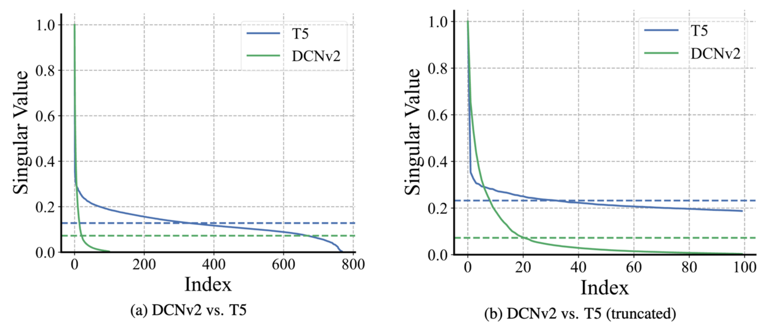
Embedding Collapse Phenomenon

▶ Observation of Embedding Collapse

- ▶ Many singular values tend to be small, embeddings tend to be low-rank.
- ▶ Compared with LLM, an intrinsic issue specific to recommendation models.



SVD on Embeddings.



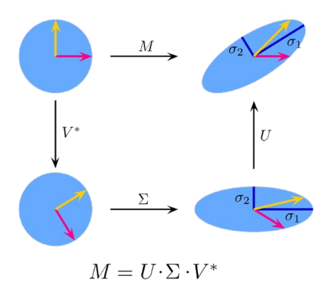
Comparision with LLMs.

▶ Analysis Tool of Embedding Collapse

$$E = U\Sigma V^T, \qquad \Sigma = \mathrm{diag}(\underbrace{\sigma_1, \sigma_2, \ldots, \sigma_k}_{\text{significances along spectra directions}}), \qquad \mathrm{rank}(E) = \|\sigma\|_0$$

Larger σ: carry more information ☺

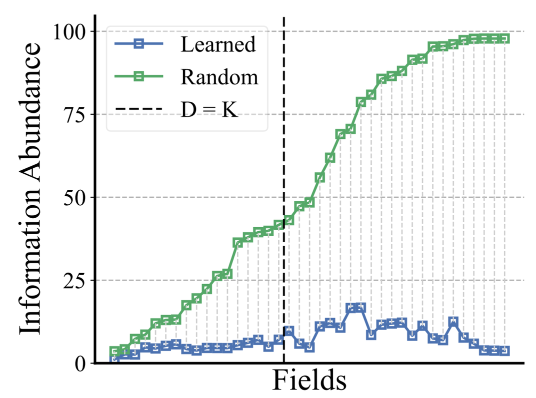
Smaller σ: more likely to be pruned ☹



Extend **rank** to ***information abundance***

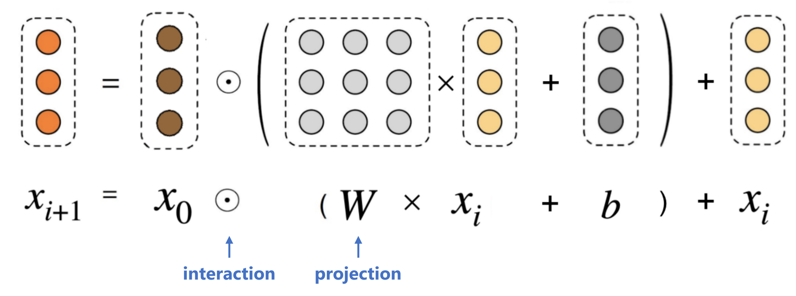
$$\mathrm{IA}(E) = \frac{\|\sigma\|_1}{\|\sigma\|_\infty}$$

Embedding Collapse: low IA

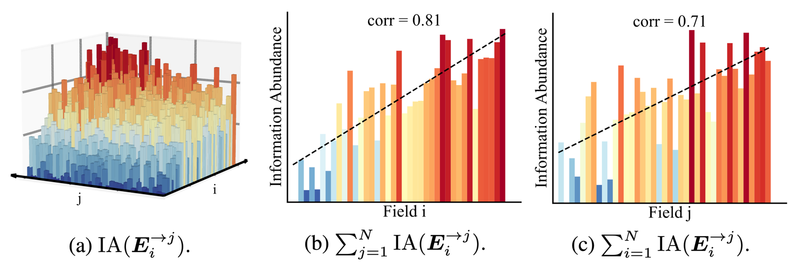


Interaction-Collapse Theory

▶ Empirical Analysis on DCNv2



DCNv2: *pre-projection* before *feature interaction*
Sub-embedding: $E_i^{\to j} = E_i W_{i \to j}^\top$



(a) $\mathrm{IA}(E_i^{\to j})$. (b) $\sum_{j=1}^N \mathrm{IA}(E_i^{\to j})$. (c) $\sum_{i=1}^N \mathrm{IA}(E_i^{\to j})$.

$\mathrm{IA}(E_i^{\to j})$ is *proportional* to both $\mathrm{IA}(E_i)$ *and* $\mathrm{IA}(E_j)$
Even if $E_i^{\to j}$ are simple projections of *the same* E_i

▶ Theoretical Analysis on FM

$$\frac{\partial \mathcal{L}}{\partial e_1} = \frac{1}{B}\sum_{b=1}^{B} \frac{\partial \ell^{(b)}}{\partial h^{(b)}} \cdot \frac{\partial h^{(b)}}{\partial e_1} = \frac{1}{B}\sum_{b=1}^{B} \frac{\partial \ell^{(b)}}{\partial h^{(b)}} \cdot \sum_{i=2}^{N} E_i^\top \mathbf{1}_{x_i^{(b)}}$$

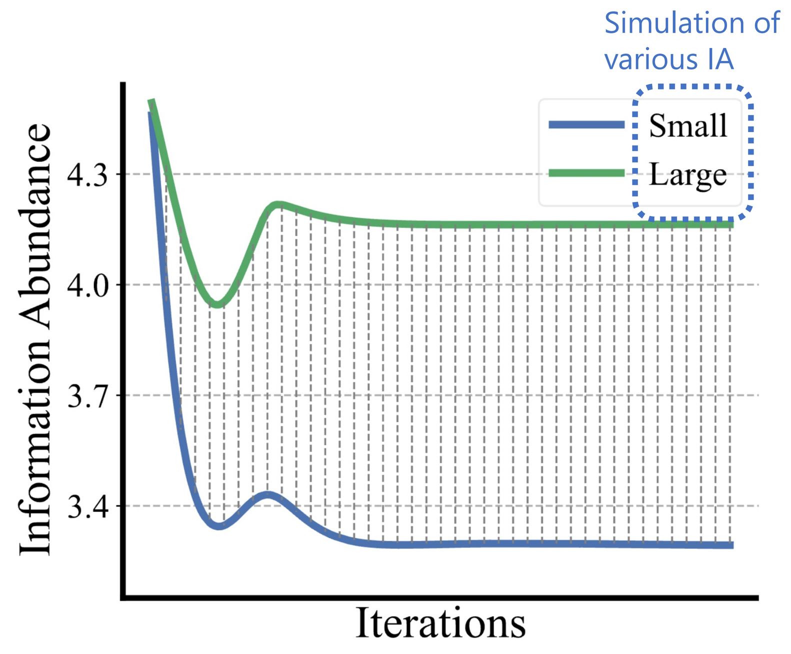
$$= \frac{1}{B}\sum_{b=1}^{B} \frac{\partial \ell^{(b)}}{\partial h^{(b)}} \cdot \sum_{i=2}^{N}\sum_{k=1}^{K} \sigma_{i,k} v_{i,k} u_{i,k}^\top \mathbf{1}_{x_i^{(b)}}$$

$$= \sum_{i=2}^{N}\sum_{k=1}^{K} \left(\frac{1}{B}\sum_{b=1}^{B} \frac{\partial \ell^{(b)}}{\partial h^{(b)}} u_{i,k}^\top \mathbf{1}_{x_i^{(b)}} \right) \sigma_{i,k} v_{i,k}$$

$$= \sum_{i=2}^{N}\sum_{k=1}^{K} \alpha_{i,k}\sigma_{i,k} v_{i,k} = \sum_{i=2}^{N} \theta_i,$$

$$\text{where} \quad \theta_i = \sum_{k=1}^{K} \alpha_{i,k} \boxed{\sigma_{i,k} v_{i,k}}$$

Gradients are correlated with spectra



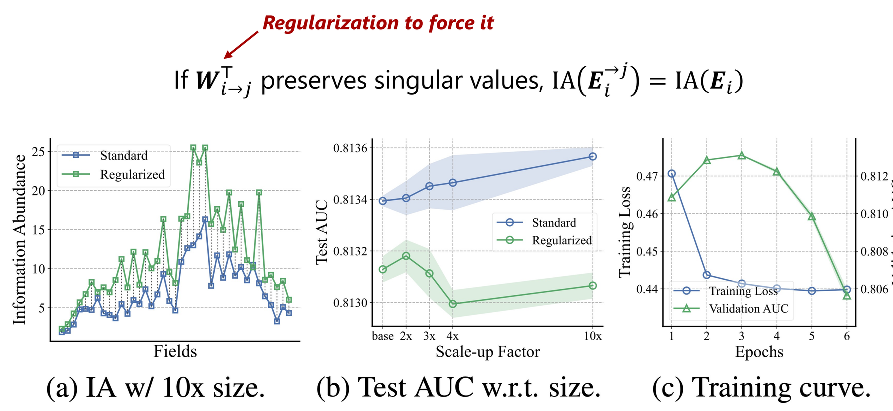
Embedding Collapse is an Optimization Issue. Toy Experiments on FM.

▶ 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.

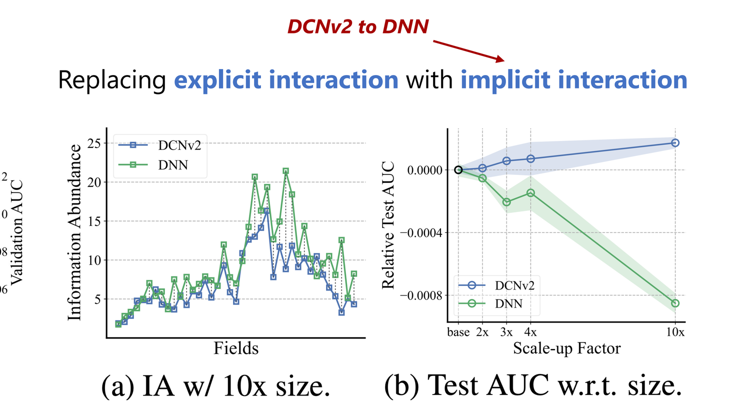
Necessity of Interaction

▶ Restricting or Replacing Interaction that Causes Collapse



(a) IA w/ 10x size. (b) Test AUC w.r.t. size. (c) Training curve.

Less collapse **but** overfitting and bad scalability



(a) IA w/ 10x size. (b) Test AUC w.r.t. size.

Less collapse **but** negative improvement and bad scalability

▶ Necessity of Interaction

A less-collapsed model with feature interaction suppressed improperly is insufficient for scalability due to overfitting concern.

▶ Overall Connections

Bad Scalability	Finding II: *Interaction is necessary* / observation	Collapsed Embedding	Finding II: *Interaction causes embedding collapse* / Finding I	Feature Interaction

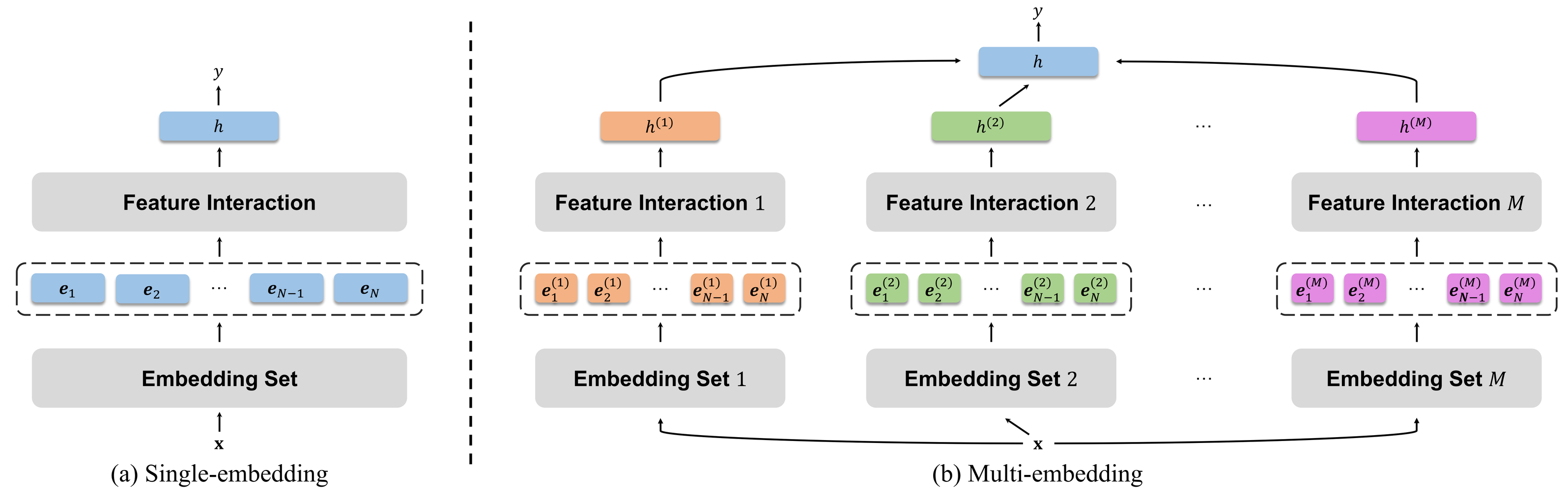
Multi-Embedding Design

▶ Principle for Scalable Model Design

- ▶ Capable of less-collapsed embeddings.
- ▶ Be within the existing feature interaction framework instead of removing interaction.

▶ Multi-Embedding Models are Scalable Recommendation Models.

Scale up ***#embedding sets*** instead of embedding dim
Each embedding set owns its ***specific interaction layers***

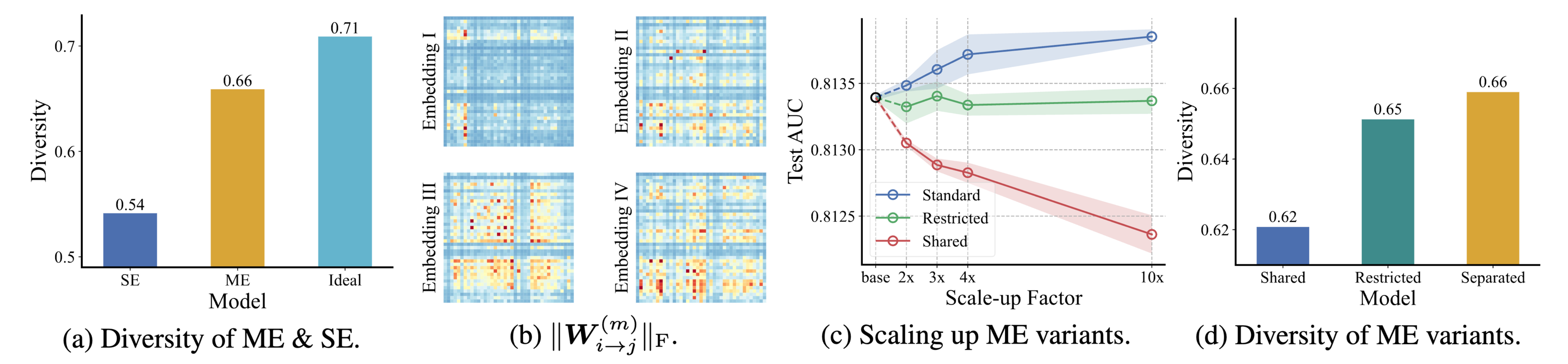


(a) Single-embedding (b) Multi-embedding

Less influenced by the interaction-collapse theory and ***mitigate embedding collapse*** while ***keeping*** the original interaction modules

▶ How Multi-Embedding Works

Scalability ← Less influenced by feature-interaction theory ← Diversity of embedding ← Diversity of interaction

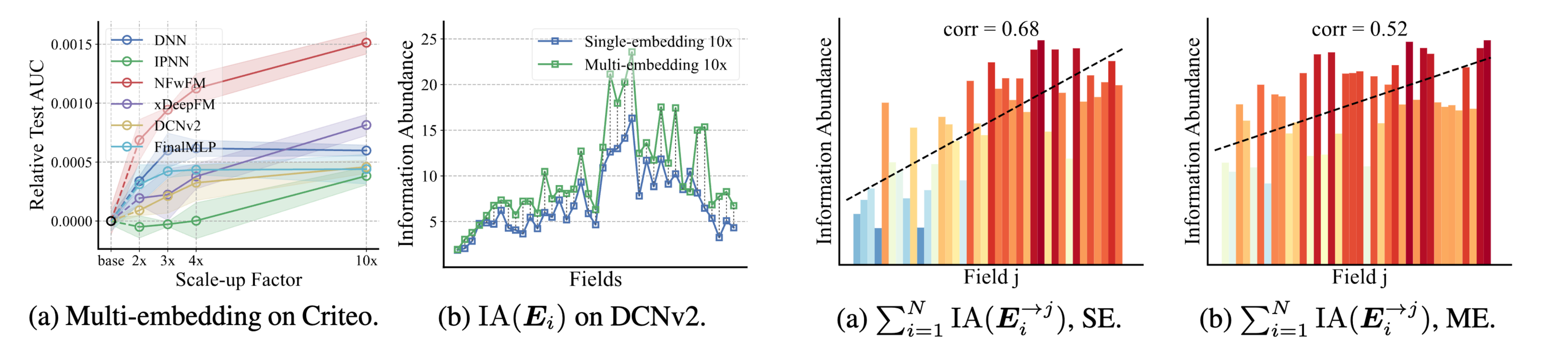


(a) Diversity of ME & SE. (b) $\|W_{i\to j}^{(m)}\|_F$. (c) Scaling up ME variants. (d) Diversity of ME variants.

Experimental Results

▶ Significant and Consistent Scalability

- ▶ Multi-embedding achieve success on 2 benchmark datasets and 6 baseline models.



(a) Multi-embedding on Criteo. (b) $\mathrm{IA}(E_i)$ on DCNv2. (a) $\sum_{i=1}^N \mathrm{IA}(E_i^{\to j})$, SE. (b) $\sum_{i=1}^N \mathrm{IA}(E_i^{\to j})$, ME.

Model		Criteo base	Criteo 2x	Criteo 3x	Criteo 4x	Criteo 10x	Avazu base	Avazu 2x	Avazu 3x	Avazu 4x	Avazu 10x
DNN	SE	0.81228	0.81222	0.81207	0.81213	0.81142	0.78744	0.78759	0.78752	0.78728	0.78648
	ME		0.81261	**0.81288**	**0.81289**	**0.81287**		0.78805	0.78826	0.78862	**0.78884**
IPNN	SE	0.81272	0.81273	0.81272	0.81271	0.81262	0.78732	0.78741	0.78738	0.78750	0.78745
	ME		0.81268	0.81270	0.81273	**0.81311**		0.78806	0.78868	0.78902	**0.78949**
NFwFM	SE	0.81059	0.81087	0.81090	0.81112	0.81113	0.78684	0.78757	0.78783	0.78794	0.78799
	ME		0.81128	0.81153	0.81171	**0.81210**		0.78868	0.78901	0.78932	**0.78974**
xDeepFM	SE	0.81217	0.81180	0.81167	0.81137	0.81116	0.78743	0.78750	0.78714	0.78735	0.78693
	ME		0.81236	0.81239	0.81255	**0.81299**		0.78848	0.78886	0.78894	**0.78927**
DCNv2	SE	0.81339	0.81341	0.81345	0.81346	0.81357	0.78786	0.78835	0.78854	0.78852	0.78856
	ME		0.81348	0.81361	**0.81382**	**0.81385**		0.78862	0.78882	0.78907	**0.78942**
FinalMLP	SE	0.81259	0.81262	0.81248	0.81240	0.81175	0.78751	0.78797	0.78795	0.78742	0.78662
	ME		0.81290	**0.81302**	**0.81303**	**0.81303**		0.78821	**0.78831**	**0.78836**	**0.78830**

School of Software, BNRist, Tsinghua University; Tencent Inc

Code: `https://github.com/thuml/Multi-Embedding`

Mail: gxz23@mails.tsinghua.edu.cn