ICML
International Conference
On Machine Learning

Forty-first International Conference on Machine Learning

Tencent 腾讯

On the Embedding Collapse When Scaling Up Recommendation Models

Xingzhuo Guo [1] **Junwei Pan** [2] **Ximei Wang** [2] **Baixu Chen** [1] **Jie Jiang** [2] **Mingsheng Long** [1]

Xingzhuo Guo

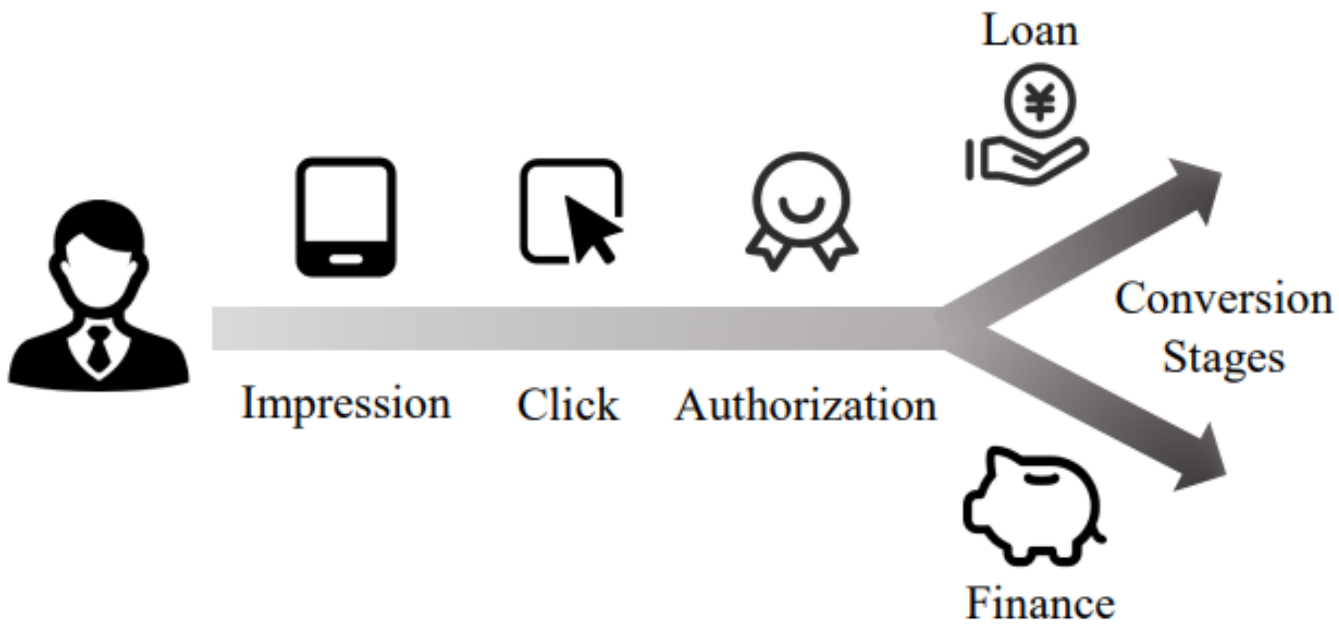
Junwei Pan

Ximei Wang

Baixu Chen

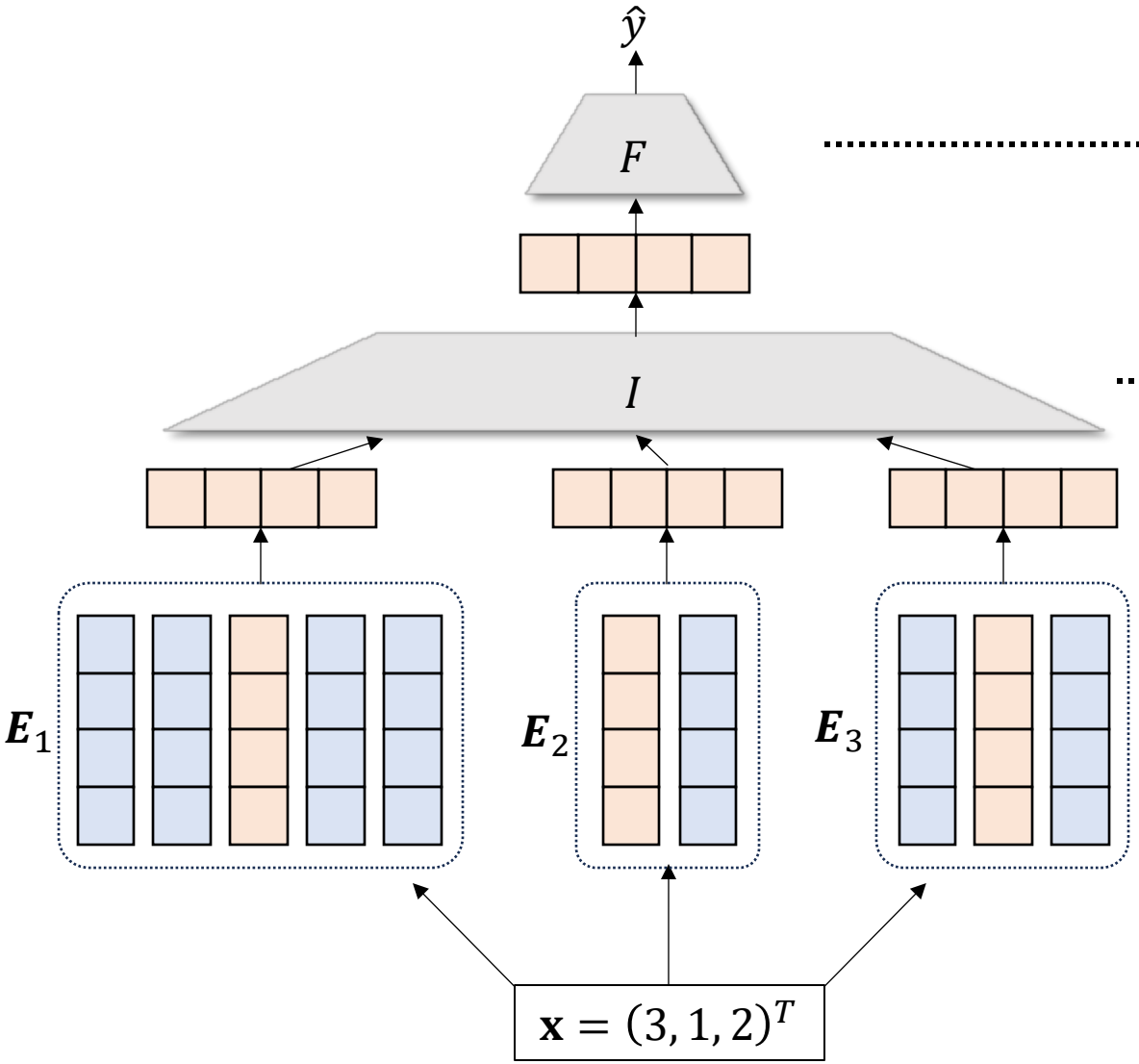
Mingsheng Long

Recommender Systems



Predict users' action based on ***features*** of users/item based on ***a large amount of data***

Deep Recommendation Models



[Post-Processing]

Commonly MLPs

[Feature Interaction]

Capturing higher-level features

Specific, most-focused

[Embedding]

Largest and most critical

Deficiency in Model Scalability: Part I

Embeddings should be ***large enough*** to be more informative for large recommendation models

Existing embedding sizes are too small

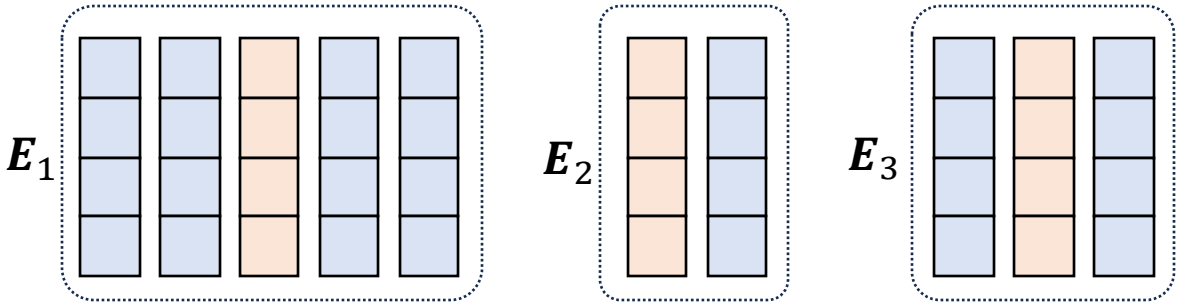
- $10^5 \sim 10^9$ of features
- Embedding size of $10 \sim 100$

JL-Lemma

- $K \geq 8\epsilon^{-2} \ln D$
- Mismatch

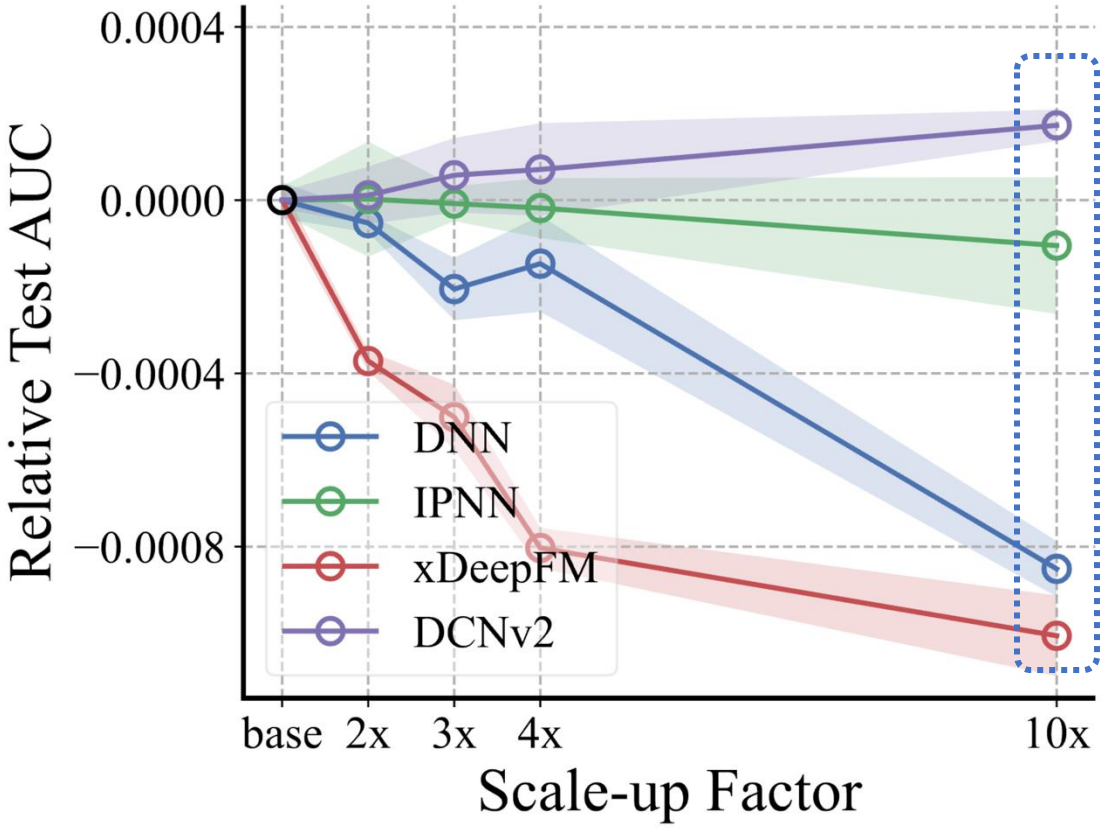
Embeddings in LLM

- $\sim 10^5$ tokens
- $\sim 10^3$ embedding size



Deficiency in Model Scalability: Part II

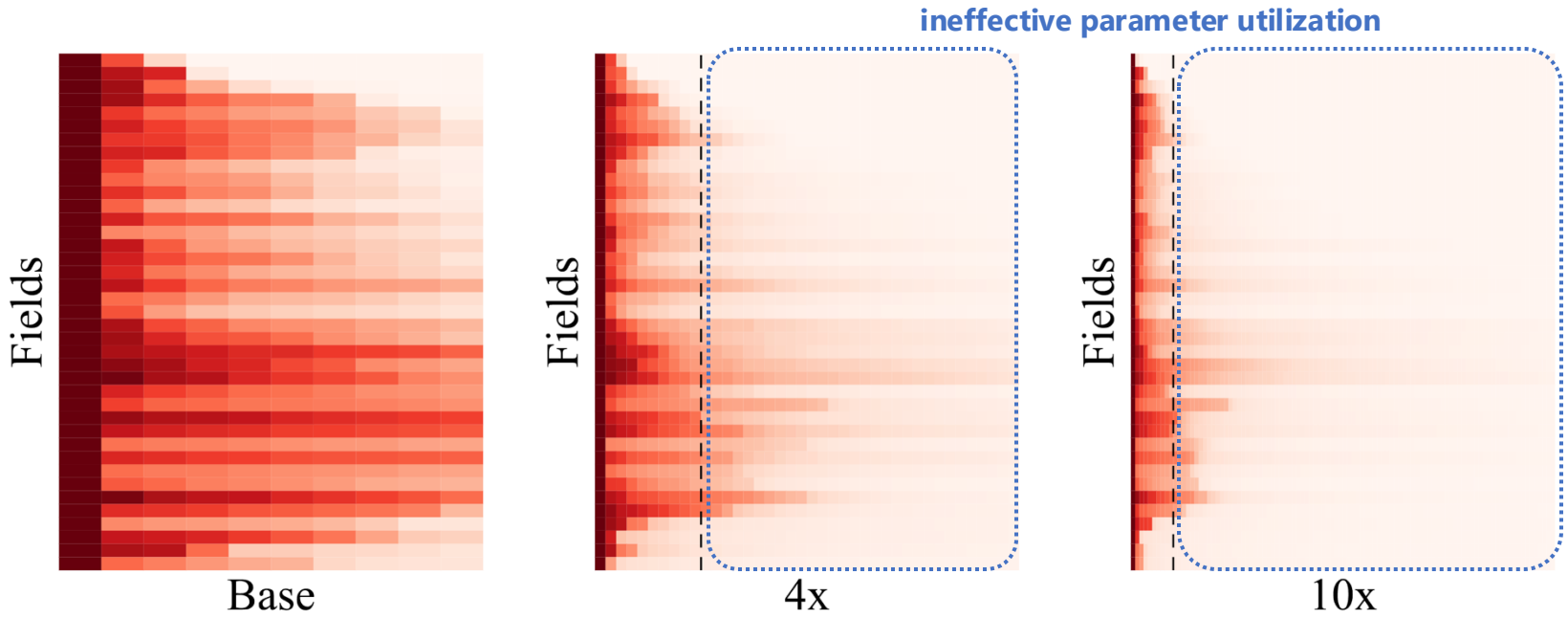
Scaling up recommendation models does ***not*** lead to performance gain



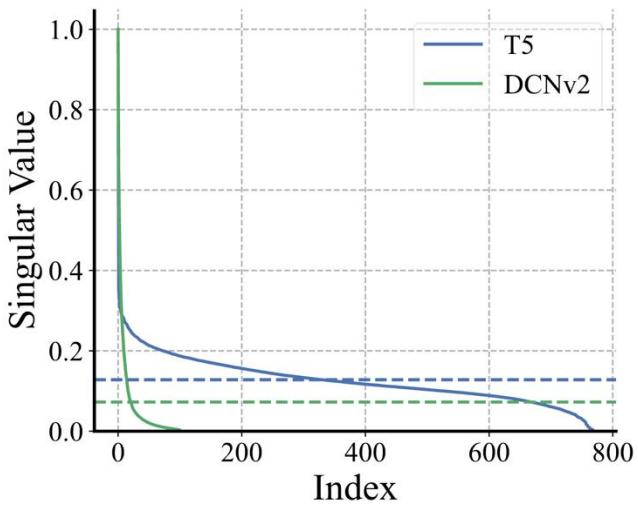
Little or even ***negative*** improvement

Embedding Collapse Phenomenon

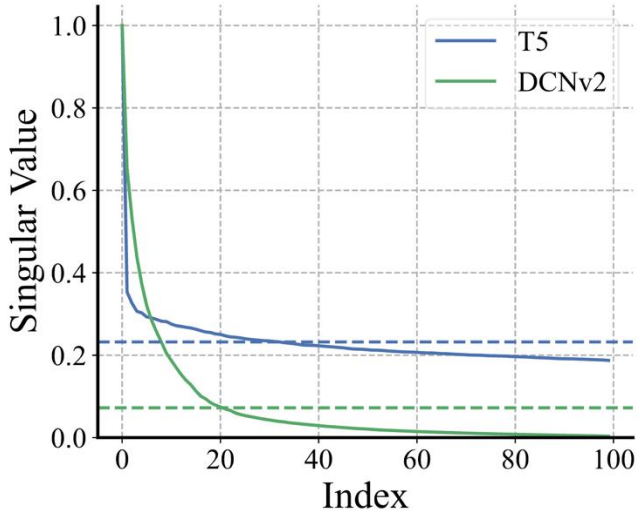
Many singular values tend to be ***small***, embeddings tend to be ***low-rank***



Comparison with LLMs



(a) DCNv2 vs. T5



(b) DCNv2 vs. T5 (truncated)

An ***intrinsic*** issue ***specific*** to recommendation models

SVD and Information Abundance

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

significances along spectra directions

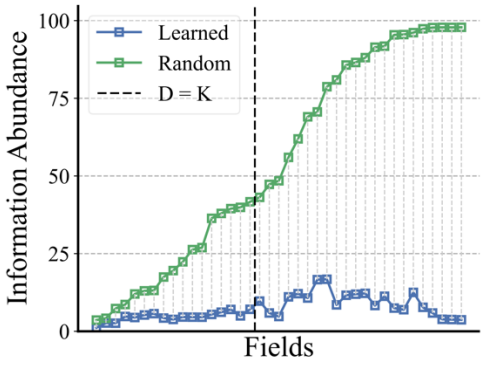
Larger σ: carry more information ☺

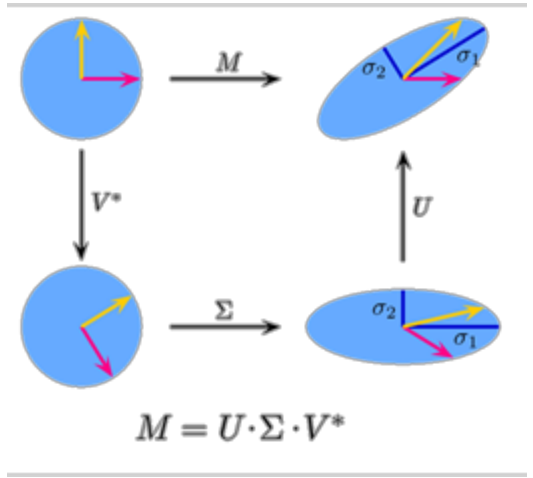
Smaller σ: more likely to be pruned ☹

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

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

Embedding Collapse: low IA



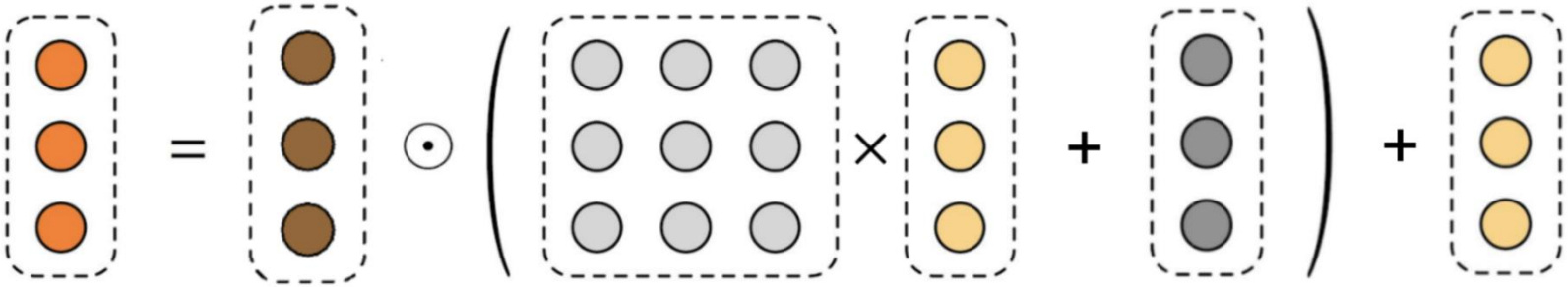


Feature Field ID	Emb Dim	Feat. N in Field	Feature Field ID	Emb Dim	Feat. N in Field
Field #01	3	62	Field #21	8	633
Field #02	8	113	Field #22	2	3
Field #03	5	125	Field #23	13	46,329
Field #04	7	50	Field #24	14	5,228
Field #05	9	223	Field #25	8	243,452
Field #06	8	147	Field #26	13	3,176
Field #07	6	99	Field #27	4	26
Field #08	5	78	Field #28	14	11,744
Field #09	8	103	Field #29	10	225,320
Field #10	3	8	Field #30	6	10
Field #11	5	31	Field #31	14	4,726
Field #12	3	56	Field #32	12	2,056
Field #13	6	81	Field #33	2	3
Field #14	8	1,457	Field #34	9	238,638
Field #15	12	555	Field #35	4	16
Field #16	2	245,195	Field #36	6	15
Field #17	11	166,164	Field #37	12	67,854
Field #18	5	305	Field #38	7	87
Field #19	4	18	Field #39	11	50,940
Field #20	14	12,054			

Finding I: Interaction-Collapse Theory

DCNv2: ***pre-projection*** before ***feature interaction***

Sub-embedding: $\boldsymbol{E}_i^{\rightarrow j} = \boldsymbol{E}_i \boldsymbol{W}_{i\rightarrow j}^{\top}$

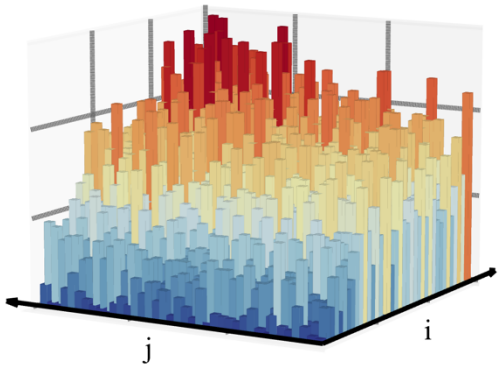


$$x_{i+1} = x_0 \odot (W \times x_i + b) + x_i$$

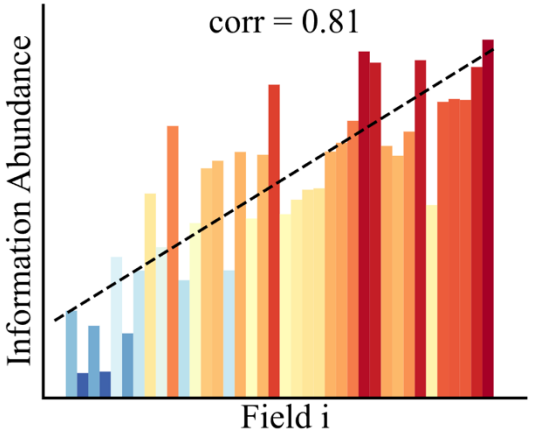
interaction projection

If $\boldsymbol{W}_{i\rightarrow j}^{\top}$ preserves singular values, $\mathrm{IA}(\boldsymbol{E}_i^{\rightarrow j}) = \mathrm{IA}(\boldsymbol{E}_i)$

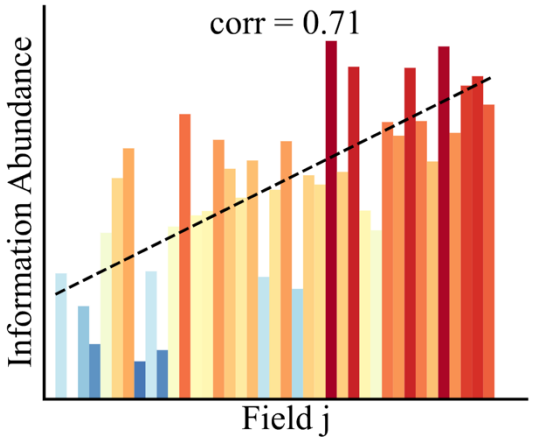
Finding I: Interaction-Collapse Theory



(a) $\mathrm{IA}(\boldsymbol{E}_i^{\rightarrow j})$.



(b) $\sum_{j=1}^{N} \mathrm{IA}(\boldsymbol{E}_i^{\rightarrow j})$.



(c) $\sum_{i=1}^{N} \mathrm{IA}(\boldsymbol{E}_i^{\rightarrow j})$.

$\mathrm{IA}(\boldsymbol{E}_i^{\rightarrow j})$ is ***proportional*** to both $\mathrm{IA}(\boldsymbol{E}_i)$ ***and*** $\mathrm{IA}(\boldsymbol{E}_j)$

Even if $\boldsymbol{E}_i^{\rightarrow j}$ are simple projections of *the same* $\boldsymbol{E}_i$

Finding I: Interaction-Collapse Theory

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{e}_1} = \frac{1}{B}\sum_{b=1}^{B}\frac{\partial \ell^{(b)}}{\partial h^{(b)}}\cdot\frac{\partial h^{(b)}}{\partial \boldsymbol{e}_1} = \frac{1}{B}\sum_{b=1}^{B}\frac{\partial \ell^{(b)}}{\partial h^{(b)}}\cdot\sum_{i=2}^{N}\boldsymbol{E}_i^{\top}\mathbf{1}_{x_i^{(b)}}$$

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

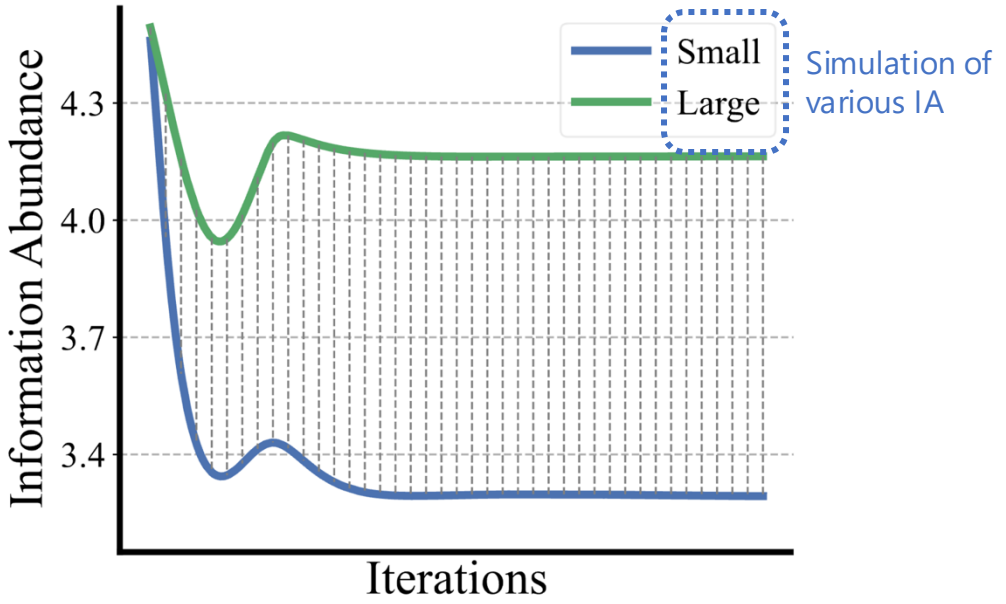
$$= \sum_{i=2}^{N}\sum_{k=1}^{K}\left(\frac{1}{B}\sum_{b=1}^{B}\frac{\partial \ell^{(b)}}{\partial h^{(b)}}\boldsymbol{u}_{i,k}^{\top}\mathbf{1}_{x_i^{(b)}}\right)\sigma_{i,k}\boldsymbol{v}_{i,k}$$

$$= \sum_{i=2}^{N}\sum_{k=1}^{K}\alpha_{i,k}\sigma_{i,k}\boldsymbol{v}_{i,k} = \sum_{i=2}^{N}\boldsymbol{\theta}_i,$$

where $\boldsymbol{\theta}_i = \sum_{k=1}^{K}\alpha_{i,k}\sigma_{i,k}\boldsymbol{v}_{i,k}$

Gradients are correlated with spectra

Theoretical Analysis



Simulation of various IA

Toy Experiments

Finding I: Interaction-Collapse Theory

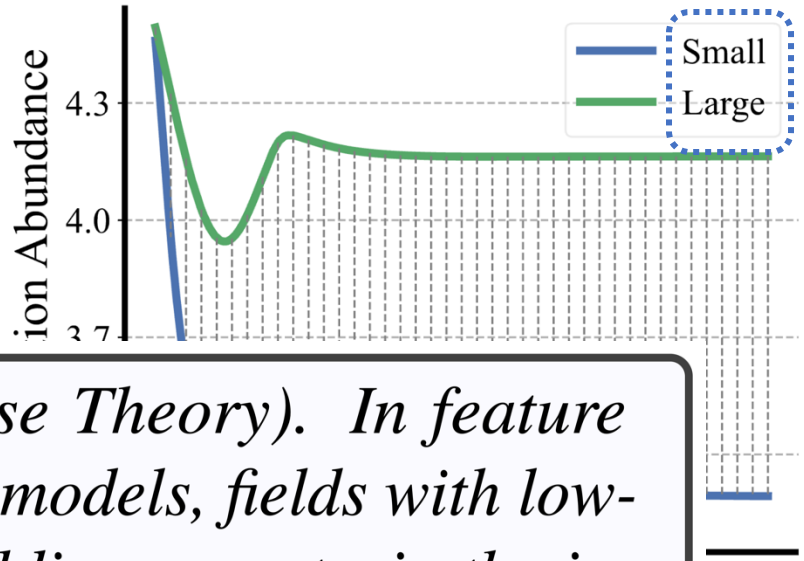
$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{e}_1} = \frac{1}{B}\sum_{b=1}^{B}\frac{\partial \ell^{(b)}}{\partial h^{(b)}}\cdot\frac{\partial h^{(b)}}{\partial \boldsymbol{e}_1} = \frac{1}{B}\sum_{b=1}^{B}\frac{\partial \ell^{(b)}}{\partial h^{(b)}}\cdot\sum_{i=2}^{N}\boldsymbol{E}_i^{\top}\mathbf{1}_{x_i^{(b)}}$$

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

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

$$= \sum_{i=2}^{N}$$

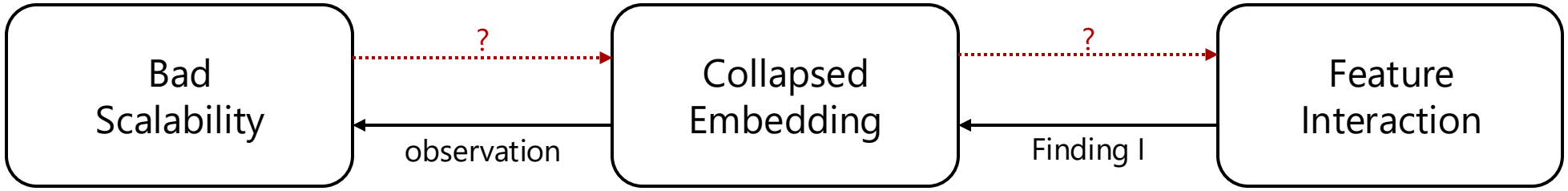
where



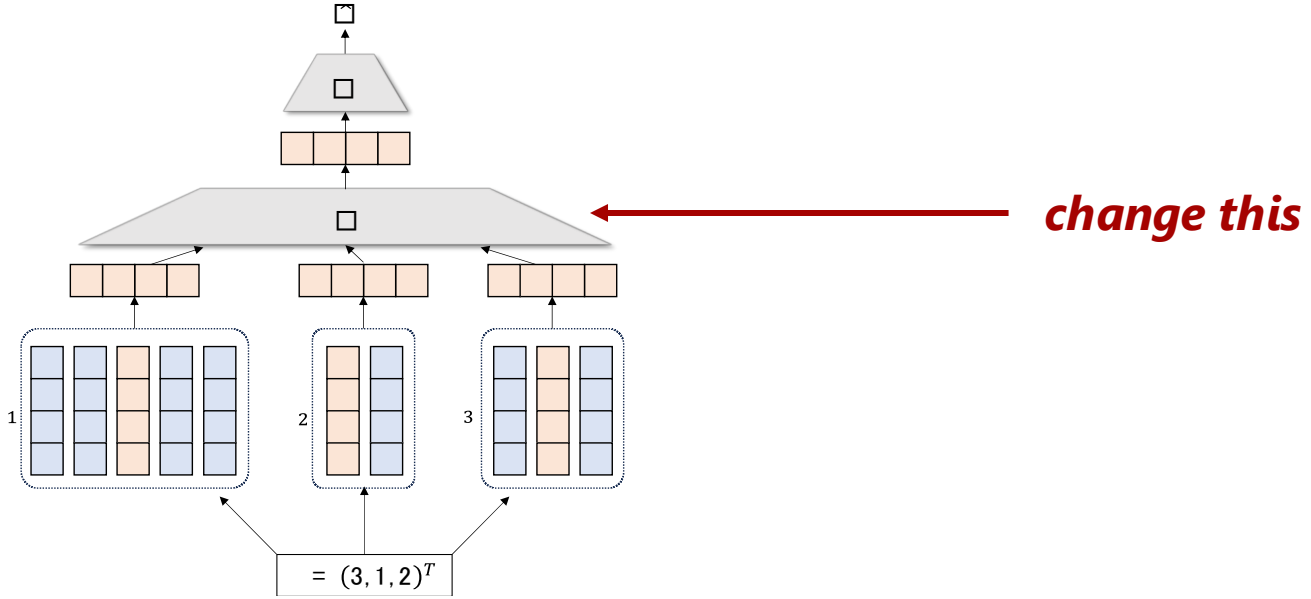
Simulation of various IA

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

Finding II: Necessity for Interaction



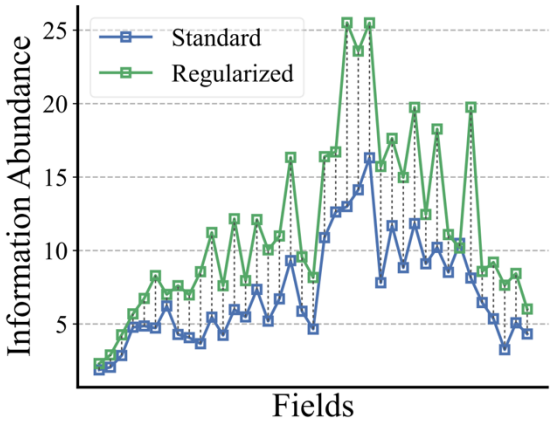
Can ***suppressing*** the feature interaction to ***mitigate collapse*** lead to ***model scalability***?



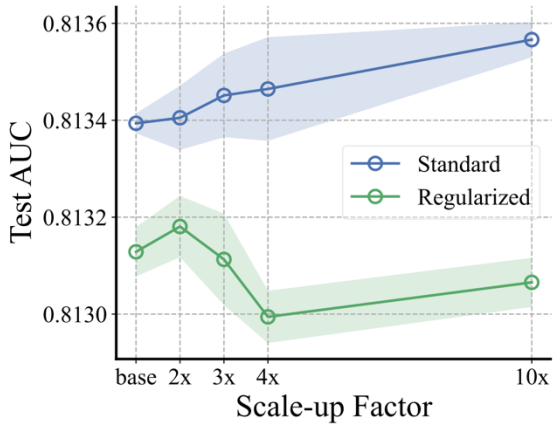
Finding II: Necessity for Interaction

Regularization to force it

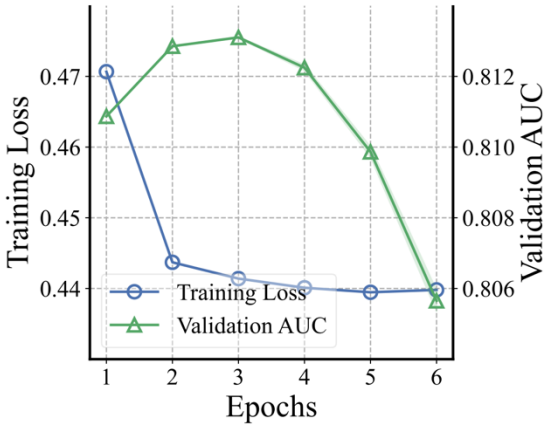
If $\boldsymbol{W}_{i\to j}^{\top}$ preserves singular values, $\mathrm{IA}\left(\boldsymbol{E}_i^{\to j}\right) = \mathrm{IA}(\boldsymbol{E}_i)$



(a) IA w/ 10x size.



(b) Test AUC w.r.t. size.



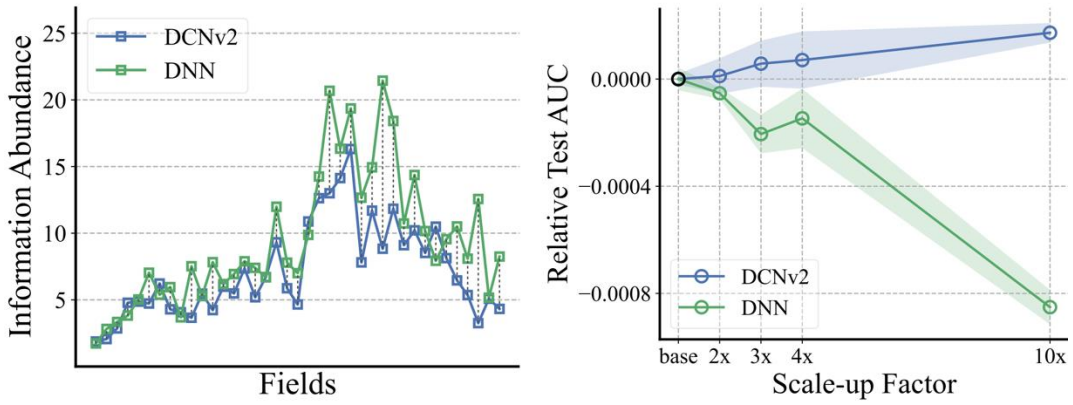
(c) Training curve.

Less collapse ***but*** overfitting and bad scalability

Finding II: Necessity for Interaction

DCNv2 to DNN

Replacing **explicit interaction** with **implicit interaction**

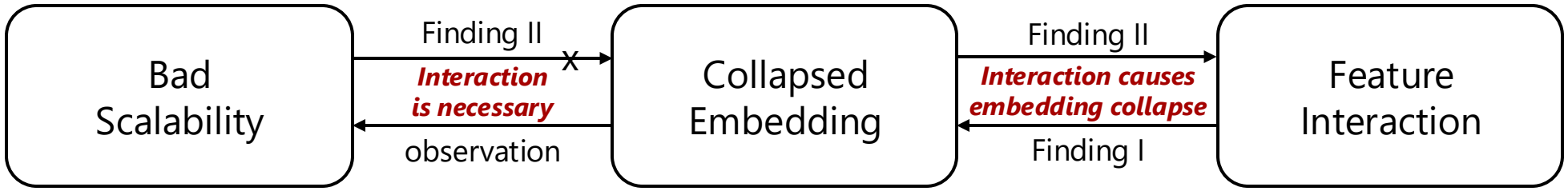


(a) IA w/ 10x size. (b) Test AUC w.r.t. size.

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

Finding II: Necessity for Interaction

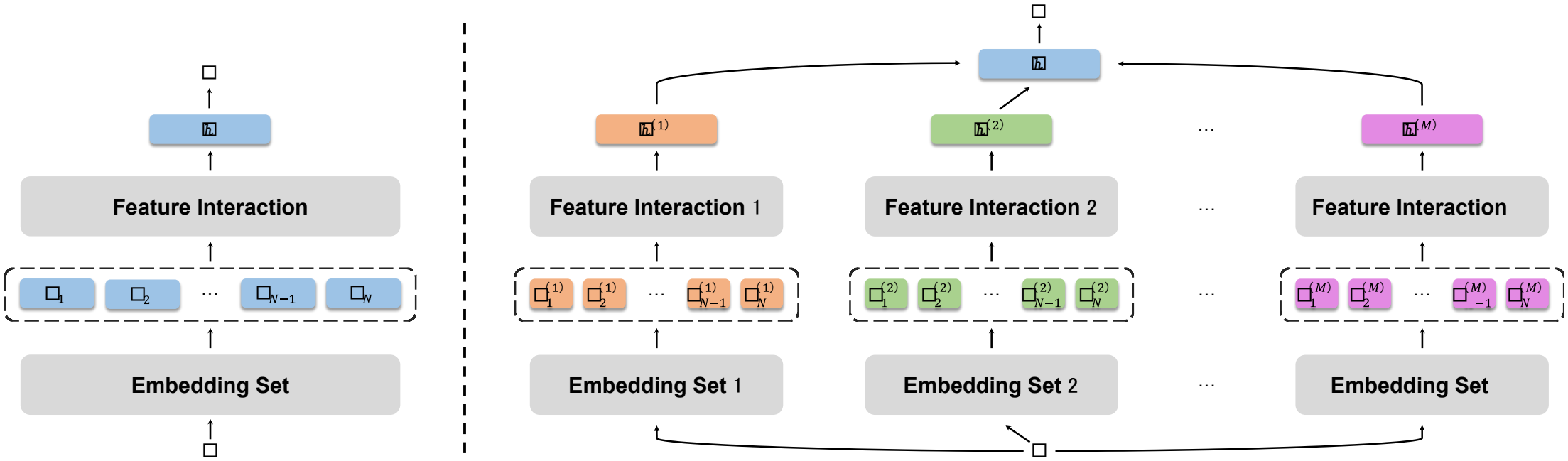
Finding 2. A less-collapsed model with feature interaction suppressed improperly is insufficient for scalability due to overfitting concern.



Multi-Embedding

Scale up ***#embedding sets*** instead of embedding dim

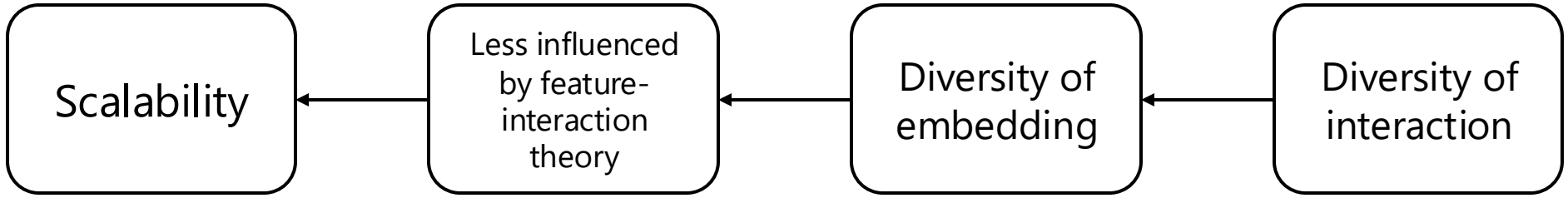
Each embedding set owns its ***specific interaction layers***

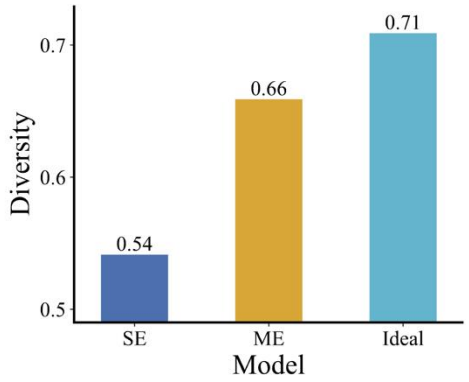


(a) Single-embedding (b) Multi-embedding

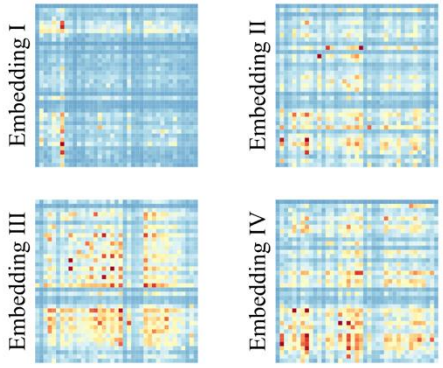
Less influenced by the interaction-collapse theory and ***mitigate embedding collapse*** while ***keeping*** the original interaction modules

How Multi-Embedding works?

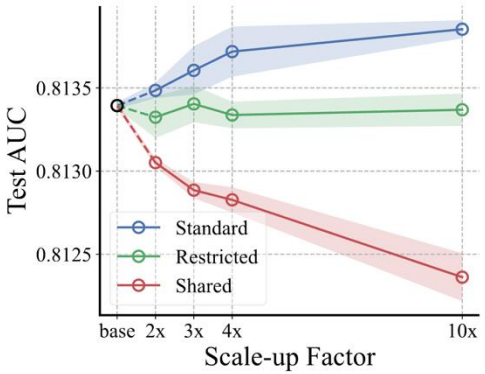




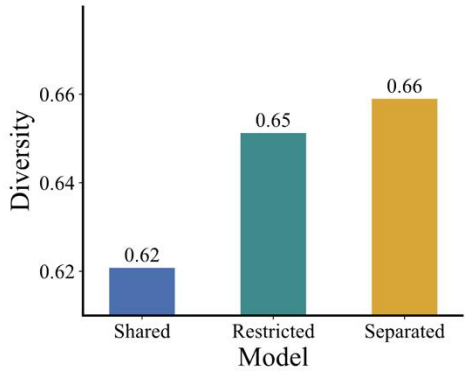
(a) Diversity of ME & SE.



(b) $\|\boldsymbol{W}_{i\to j}^{(m)}\|_{\mathrm{F}}$.

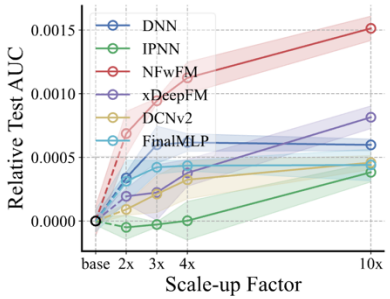


(c) Scaling up ME variants.

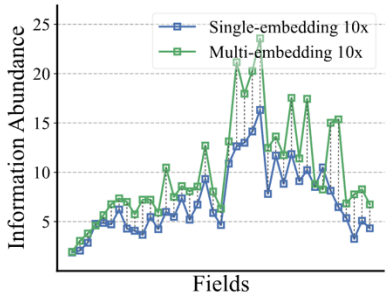


(d) Diversity of ME variants.

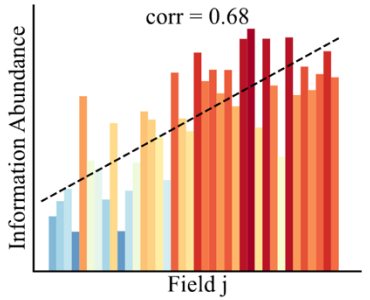
Experiment Results



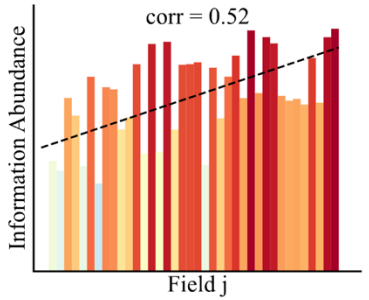
(a) Multi-embedding on Criteo.



(b) $\mathrm{IA}(\boldsymbol{E}_i)$ on DCNv2.



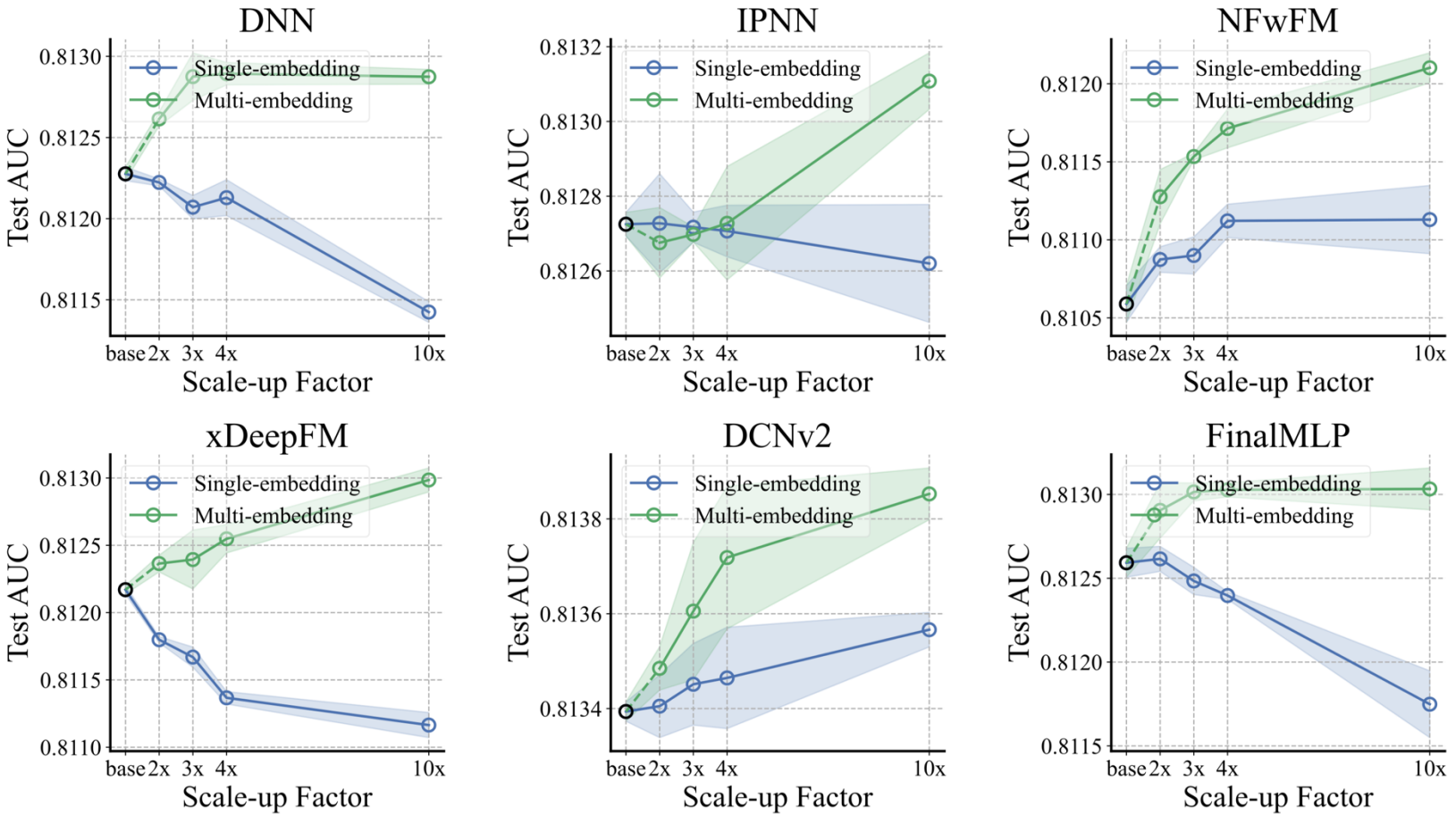
(a) $\sum_{i=1}^{N} \mathrm{IA}(\boldsymbol{E}_i^{\to j})$, SE.



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

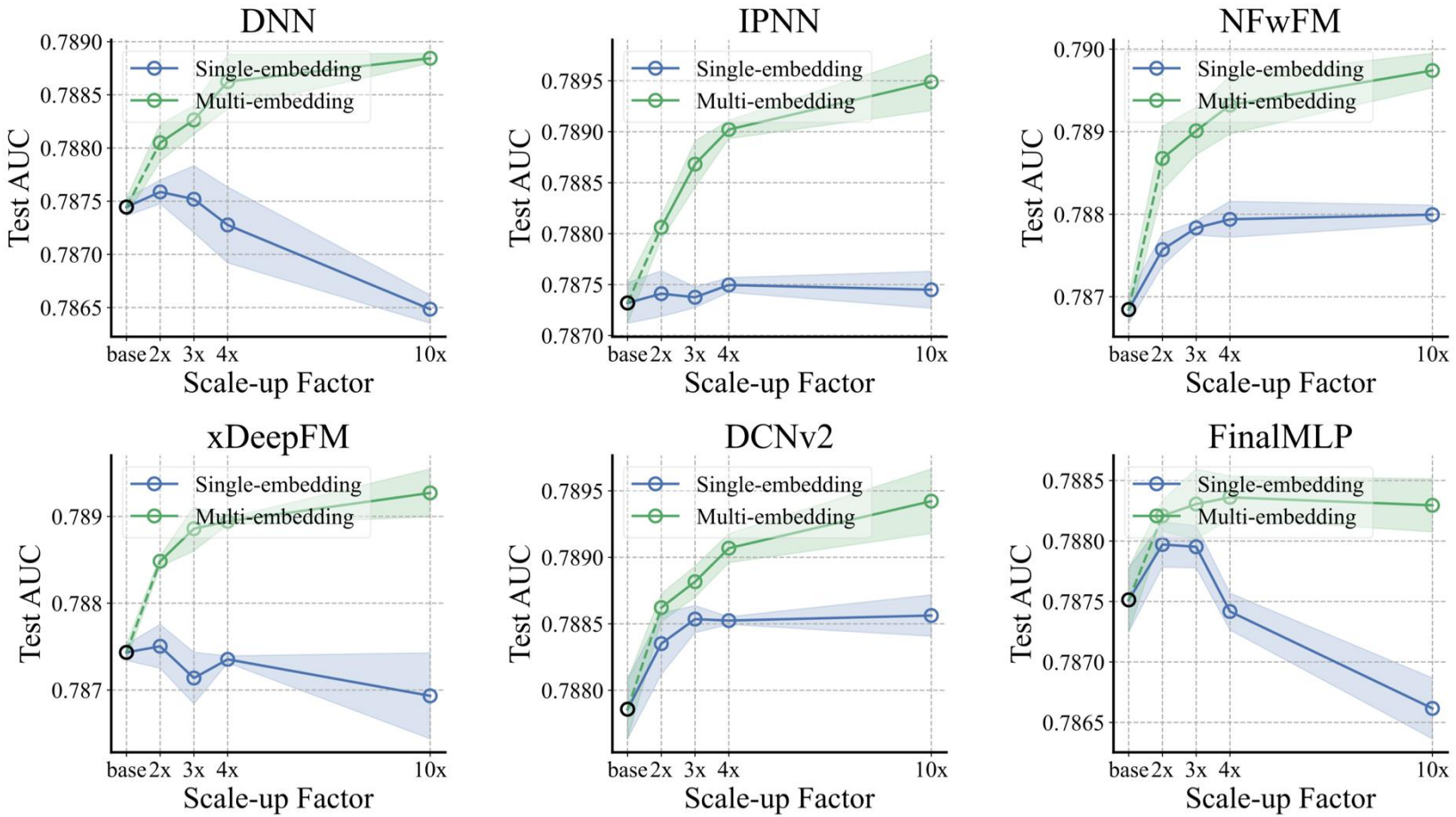
Model		Criteo					Avazu				
		base	2x	3x	4x	10x	base	2x	3x	4x	10x
DNN	SE	0.81228	0.81222	0.81207	0.81213	0.81142	0.78744	0.78759	0.78752	0.78728	0.78648
	ME		0.81261	**0.81288**	**0.81289**	**0.81287**		0.78805	0.78826	0.78862	**0.78884**
IPNN	SE	0.81272	0.81273	0.81272	0.81271	0.81262	0.78732	0.78741	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**

Experiment Results



Results on Criteo

Experiment Results



Results on Avazu

Tencent 腾讯

Thank You!

Xingzhuo Guo

gxz23@mails.tsinghua.edu.cn

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