Dimension Relation Modeling for Click-Through Rate Prediction

Zihao Zhao,Zhiwei Fang,Yong Li,Changping Peng,Yongjun Bao,Weipeng Yan
{zhaozihao3,fangzhiwei2,liyong5,pengchangping,baoyongjun,paul.yan}@jd.com
Business Growth BU, JD

ABSTRACT

Embedding mechanism plays an important role in Click-Through-Rate (CTR) prediction. Essentially, it tries to learn a new feature space with some learned latent properties as the basis, and maps the high dimensional and categorical raw data to dense, rich and expressive representations, i.e., the embedding features. Current researches usually focus on learning the interactions through operations on the whole embedding features without considering the relations among the learned latent properties. In this paper, we find it has clear positive effects on CTR prediction to model such relations and propose a novel Dimension Relation Module (DRM) to capture them through dimension recalibration. We show that DRM can improve the performance of existing models consistently and the improvements are more obvious when the embedding dimension is higher. We further boost **F**ield-wise and **E**lement-wise embedding methods with our **D**RM and name this new model FED network. Extensive experiments demonstrate that FED is very powerful in CTR prediction task and achieves new state-of-the-art results on Criteo, Avazu and JD.com datasets.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; **Computational advertising**.

KEYWORDS

Neural Networks, Deep Learning, Recommendation

ACM Reference Format:
Zihao Zhao,Zhiwei Fang,Yong Li,Changping Peng,Yongjun Bao,Weipeng Yan. 2020. Dimension Relation Modeling for Click-Through Rate Prediction. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3340531.3412108

1 INTRODUCTION

Click-Through-Rate (CTR) prediction plays a primary role in industry advertising system. It reflects the probability of an ad to be clicked on and influences the rank of items in advertising system. Thus, the accuracy of CTR prediction usually makes a direct effect on final revenue.

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3412108

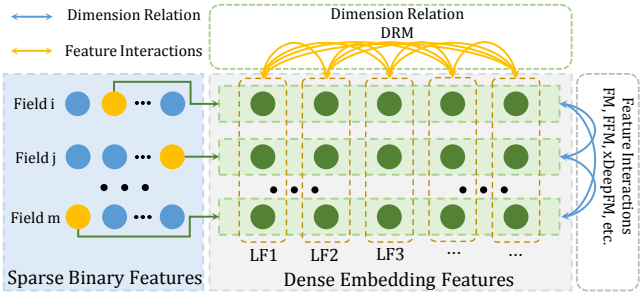

Figure 1: Feature Interaction and Dimension Relation. The former learns the relations among raw categorical fields while the latter models relations among the learned latent dense fields.

One challenge in CTR prediction is that the input variables are mostly discrete and categorical, leading to a large and sparse feature space. A common practice is to transform the data into high-dimensional sparse binary feature via one hot encoding and apply linear models such as logistic regression to make predictions. Although linear model is fast and simple, it's hard to capture high-order interactions under the linear assumption. Thus, embedding method is proposed to project each sparse and discrete source feature into a dense embedding feature with a fixed length. With the dense representations, more complicate models can be applied to learn potential patterns of feature interactions without much effort in hand-craft feature design. Meanwhile, embedding based methods have the ability to improve model generalization and explore different level (low-order or high-order) feature interactions thus making it the base of many models in CTR prediction, [1, 3–6, 8, 10].

Now, let's review the embedding mechanism from the view of feature space transformation. In the raw feature space, a sample is represented by its responses on a series of fields, such as *Location*, *Gender*, *Time* or *Click*. Namely, the basis of raw feature space is a set of fields. Since the responses on these fields are usually discrete and categorical, the raw features (in one hot format) are also discrete, sparse and high-dimensional. The embedding mechanism assumes that each categorical field can be mapped into a new feature space whose basis is a series of some *latent fields*. These latent fields are continuous, unknown but learnable. Usually, the raw fields are not independent, thus it's effective to learn information from their interactions, such as low-order interactions [4, 8] or high-order interactions [3, 5], as shown in Fig.1. Similarly, in the embedding feature space, the learned latent fields are also in high probability dependent since there is no constraint to force the basis of the space to be orthogonal. However, current researches only model

the interactions among the raw fields without deep research on the relations among latent fields in embedding feature space.

In this paper, we aim to explicitly model such relations to improve CTR prediction. Specifically, we propose a novel module based on dimension recalibration and self-attention mechanism [2, 9] to learn the relations among latent fields. Since each latent field corresponds to a specific dimension in the embedding feature space, we name our module **D**imension **R**elation **M**odule (DRM). Our experimental results show that the proposed DRM module can catch extra useful information for CTR prediction and boost the performance of existing state-of-the-art methods such as [1, 3], especially when the size of embedding dimension is high.

The main contributions of this paper are concluded as follows:

- We find it's useful for CTR prediction to model the relations of *latent fields* in embedding space and propose a novel module named DRM to capture such relations. DRM is easily incorporated into existing methods to learn more powerful embedding representations, leading to better performance, especially when the dimension of embedding space is high.
- We design FED network based on DRM, which takes the advantage of field-wise and element-wise modeling simultaneously to learn interactions explicitly and implicitly.
- We conduct extensive experiments on several datasets like Criteo, Avazu and JD.com datasets. Experimental results demonstrate the superiority over the state-of-art models.

2 METHOD

2.1 Feature Embedding

In advertising systems, the input samples are sparse, of huge dimension and containing response values on a series of categorical fields. The field-aware one-hot encoding is a widely used form in current related works to present such data.

Embedding mechanism maps the one-hot vector of each field to a d-dimensional dense vector, i.e., the embedding feature. Let $\mathbf{v}_i = [v_{i,1}, v_{i,2}, ..., v_{i,d}]^T \in \mathbb{R}^d$ denote the embedding feature for i^{th} field, then after embedding, an input sample is described as:

$$\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_i, ..., \mathbf{v}_m]^T \tag{1}$$

where $\mathbf{V} \in \mathbb{R}^{m \times d}$ and m is the number of fields in source features.

In embedding feature space, each dimension stands for one of the d learned *latent fields*. Then the representation of i^{th} dimension is $\mathbf{u}_i = [v_{1,i}, v_{2,i}, ..., v_{m,i}]^T \in \mathbb{R}^m$. Let $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_d]^T$, the relation between $\mathbf{U}$ and $\mathbf{V}$ is obvious:

$$\mathbf{U} = \mathbf{V}^T \in \mathbb{R}^{d \times m} \tag{2}$$

From Eq.2, we have some insight into embedding feature $\mathbf{v}_i$ and dimension feature $\mathbf{u}_i$. For a specific $\mathbf{v}_i$, it recodes the response values of the corresponding field on all *latent fields*, while a $\mathbf{u}_i$ reflects different fields' behaviors on a single *latent fields*. Unlike the raw fields, it's hard to assign a clear and unique semantic *nature word* such as `Location` or `Gender` to each *latent field*, while $\mathbf{u}_i$ provides another observation for each of them.

2.2 Dimension Relation Module

We propose a module named **D**imension **R**elation **M**odule (DRM) based on attention mechanism. DRM helps to learn sample-level enhanced embedding vectors, which selectively emphasizes specific dimensions and suppress less useful ones for the given task. Specifically, self attention is preferred to determine the importance and relations among dimensions, then we elaborate the process to adaptively aggregate dimension information.

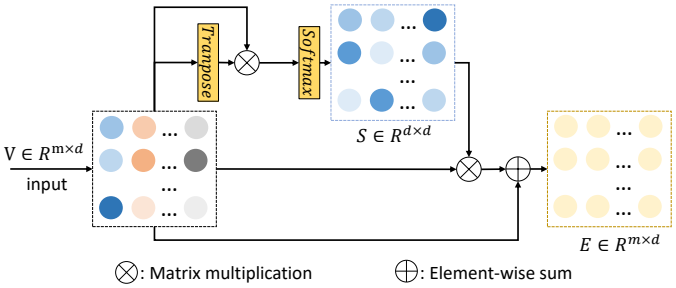

Figure 2: The overview architecture of DRM. We omit the transform matrix for simplicity.

The DRM module is shown in Fig.2. Given the dimension feature matrix $\mathbf{U} \in \mathbb{R}^{d \times m}$, the dimension attention map matrix $\mathbf{S} \in \mathbb{R}^{d \times d}$ can be achieved by:

$$\mathbf{S} = softmax((\mathbf{U}\mathbf{W}_\theta)(\mathbf{U}\mathbf{W}_\phi)^T) \tag{3}$$

where $\mathbf{W}_\theta, \mathbf{W}_\phi \in \mathbb{R}^{m \times m}$ is the transform matrix. Each entry $\mathbf{S}_{ji}$ in $\mathbf{S}$ measures the i^{th} dimension's impact on j^{th} dimension. Then embedding vectors can be enhanced by performing the attention map matrix to the origin input as follows:

$$\mathbf{E} = (\mathbf{W}_\delta \mathbf{V})\mathbf{S} \tag{4}$$

where $\mathbf{W}_\delta \in \mathbb{R}^{m \times m}$ is the transform matrix to original input. For better optimization, the residual connection is added to the attention embedding vectors like Resnet, and the final form of DRM is as follows:

$$\mathbf{E} = (\mathbf{W}_\delta \mathbf{V})\mathbf{S} + \mathbf{V} \tag{5}$$

We can see that the output $\mathbf{E}$ at each dimension is a weighted sum of the vectors of all dimension and original embeddings, where the weight is determined by the similarity of dimensions. Thus, if a dimension is important, it would generate more impact on $\mathbf{E}$ and finally contribute more to the result.

2.3 FED Network

The framework of FED-net is shown in Fig.3, the origin embedding $\mathbf{V}$ is first passed through DRM, which helps to learn more powerful embedding representation by explicitly modeling relations among dimensions. FED-net models the complex interactions among features from multiple levels, including the field-wise and element-wise. The predicted value $\hat{y}$ for a given sample is as follows:

$$\hat{y} = \sigma([\mathbf{x}_{dim}; \mathbf{x}_{field}; \mathbf{x}_{element}]^T \mathbf{w}_{logistic}) \tag{6}$$

where $\sigma(x) = 1/(1+exp(-x))$ is the sigmoid function, $\mathbf{x}_{dim} \in \mathbb{R}^{dm}$ is the vectorized output of DRM module, $\mathbf{x}_{field} \in \mathbb{R}^{dm}$ is the vectorized output of field-wise module, $\mathbf{x}_{element} \in \mathbb{R}^l$ is the output of element-wise module and $\mathbf{w}_{logistic} \in \mathbb{R}^{2md+l}$ is the weight for the logistic layer.

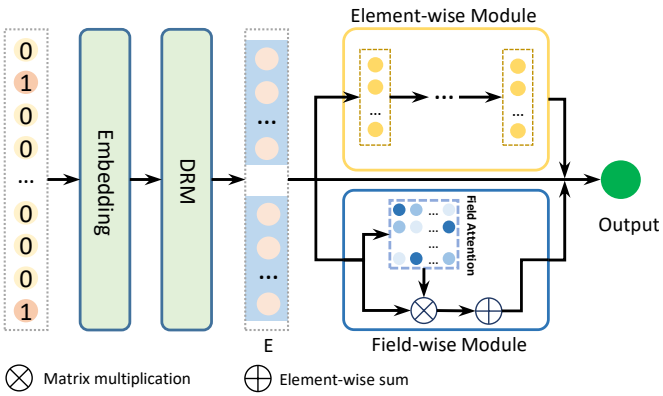

Figure 3: The overview architecture of FED-net.

Then the loss function is defined as:

$$loss = -\frac{1}{N}\sum_{i=1}^{N} y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i) \tag{7}$$

where y_i is the ground truth label for the given sample i, and N is the size of training dataset.

2.3.1 Field-wise Module. Field information is an explicit description for a given sample. Different samples are supposed to pay attention to different fields and its interactions for the given task [11]. We propose to utilize attention mechanism to model field-wise feature interactions.

In the field-wise module, we first perform matrix multiplication between the output embedding E of DRM and its transpose E^T. Then a softmax function is applied to calculate the field attention map $H \in \mathbb{R}^{m \times m}$. H_{ji} is computed by:

$$\mathbf{H} = softmax((\mathbf{E}\mathbf{W}_{\theta^1}) \cdot (\mathbf{E}\mathbf{W}_{\phi^1})^T) \tag{8}$$

where $\mathbf{W}_{\theta^1}, \mathbf{W}_{\phi^1} \in \mathbb{R}^{d \times d}$ is the tranform matrix, and $\mathbf{H}_{ji}$ measures the i^{th} field's impact on j^{th} field. Thereafter, the field attention map $\mathbf{H}$ applies to the embedding vectors $\mathbf{E}$ to enhance field information, and identity term is added. Then, the output $\mathbf{F}$ of field-wise module is:

$$\mathbf{F} = \mathbf{H}(\mathbf{E}\mathbf{W}_{\delta^1}) + \mathbf{E} \tag{9}$$

where $\mathbf{W}_{\delta^1} \in \mathbb{R}^{d \times d}$ is the transform matrix to original input.

With such structure of layer, each field feature will be updated into a high-order feature which takes the field interactions into account. Therefore, we can model certain order combinatorial features by stacking multiple layers. Finally, the output $\mathbf{F}$ of final layer is vectorized into $\mathbf{x}_{field}$.

2.3.2 Element-wise Module. The element-wise module is a feed-forward DNN network. The output embedding vectors of DRM are vectorized into a one-dimension vector, and fed into the hidden layers of DNN network. Specifically, for the l^{th} hidden layer, computation is performed as:

$$\mathbf{h}_{l+1} = f(\mathbf{W}_l \mathbf{h}_l + \mathbf{b}_l) \tag{10}$$

where $\mathbf{h}_l$ is the input of the l^{th} hidden layer, and $\mathbf{h}_{l+1}$ is the output of the l^{th} hidden layer. $\mathbf{W}_l$ is the weight matrix and $\mathbf{b}_l$ is bias at l^{th} layer, and f is the activation function which is chosen as rectified linear units (ReLU). The output of last layer is vectorized to $\mathbf{x}_{element}$.

Table 1: DRM boosting DNN and Field Attention Network on Criteo dataset.

Model	AUC	Logloss
DNN	0.8073	0.4439
DNN+DRM	**0.8084**	**0.4429**
Field Attention	0.8060	0.4453
Field Attention + DRM	**0.8073**	**0.4441**

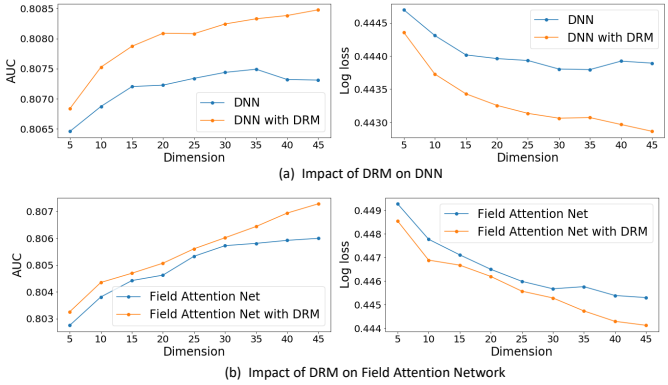

Figure 4: Impact of DRM on Criteo dataset while dimension number increases.

3 EXPERIMENT

3.1 Datasets

We conduct our experiments on two public datasets Criteo, Avazu and one commercial dataset JD.com. Criteo and Avazu are both popular industry benchmarking datasets for predicting ads CTR. For thees two datasets, data is randomly split into three parts, where 80% samples are used for training, and 10% for validation and the 10% for testing.

JD.com Dataset: The JD.com dataset covers a period of 9 days, where the first 8 days are used for training, and the last day is for validation and testing. Finally there are 82 million samples for training, 5 million for validation and 5 million for testing. Each sample has 20 categorical features including user information, item information and context information (city, request time and so on).

3.2 Dimension Relation Module

In this section, we validate the effectiveness of DRM in boosting element-wise model (DNN) and field-wise model (Field Attention Network). Furthermore, we perform experiments with different embedding sizes, and experimental results show that improvement of DRM is consistent, and DRM achieves more remarkable improvement when embedding size is high.

Firstly, we show experiments of DRM with DNN network and Field Attention Network on the Criteo dataset in Table 1. We find that both architectures get improvements in AUC and decreases

Table 2: Ablation experiments about different modules.

Model	AUC	Logloss
Only Element-wise Module	0.8073	0.4439
Only Field-wise Module	0.8060	0.4421
Element-wise & Field-wise	0.8094	0.4453
FED	**0.8113**	**0.4403**

Table 3: Performances on three Datasets.

	Criteo		Avazu		JD.com	
Model	AUC	Logloss	AUC	Logloss	AUC	Logloss
LR	0.7948	0.4553	0.7713	0.3849	0.6495	0.1494
FM	0.8025	0.4496	0.7804	0.3805	0.6609	0.1460
W&D [1]	0.8076	0.4436	0.7869	0.3755	0.6685	0.1432
AFM [11]	0.8038	0.4478	0.7817	0.3792	0.6618	0.1457
DeepFM [3]	0.8091	0.4423	0.7878	0.3753	0.6708	0.1372
DCN [10]	0.8093	0.4420	0.7875	0.3759	0.6702	0.1361
PNN [7]	0.8094	0.4414	0.7879	0.3752	0.6705	0.1360
xDeepFM [5]	0.8096	0.4412	0.7877	0.3753	0.6710	0.1356
FGCNN [6]	0.8093	0.4422	0.7878	0.3752	0.6709	0.1361
FED (Ours)	**0.8113**	**0.4403**	**0.7889**	**0.3748**	**0.6735**	**0.1341**

in LogLoss, which indicates that DRM can learn extra information from dimension relations for CTR prediction. It should be explained that increase at 10^{-3} level in Criteo dataset is already clear compared with recent works such as xDeepFM [5] and DCN [10].

Since DRM learns the relations of dimensions in embedding features space, its performance is affected by the number of dimensions, i.e., the embedding size. The experiments are summarized in Fig.4. From Fig.4, we have the following observations when dimension increases: 1) the performance of DNN increases very marginally and even becomes worse if dimension is higher than 35; 2) DNN with DRM keeps increases all the time; 3) the margin between the two curves keeps growing. Similar phenomena can also be found in Field Attention Network without/with DRM in Fig.4. This indicates that DRM can boost different architectures and improve their learning capacity. The reason is that higher dimensions can enhance the expression of embedding features but will introduce noise and cause hard optimization, while our DRM alleviates such problem by applying enhancement and suppression operations on embedding features with the learned dimension relations.

3.3 FED Network

3.3.1 Ablation Experiments. In FED network, there are three core modules, i.e., DRM, element-wise module and field-wise module, to contribute to the final prediction. In order to analyze their effects, we conduct an ablation study which is summarized in Tab.2. From the results, we can have the following observations. Firstly, element-wise module outperforms field-wise module, which means that element-wise transformation still plays an essential role in CTR prediction although it is simple. Secondly, when further using DRM, the FED network achieves the best performance, which indicates that the dimension relation has positive effectives and the three modules complement each other in FED network.

3.3.2 Comparison to the State-of-the-art. The performance of different models is listed in Tab.3. On Criteo dataset, we observe that LR is far worse than all the rest models, which demonstrates that embedding-based models are essential for measuring sparse features. W&D, DeepFM, DCN, PNN, xDeepFM, FGCNN and FED are significantly better than FM, which directly reflects that deep learning is important for boosting the accuracy. Besides, as a useful practice in CTR prediction, incorporating hybrid components is used nearly in all the deep learning based models. Our proposed FED achieves the best performance with a clear margin, which verifies that learning the dimension relations in embedding space is another effective way to enhance CTR prediction.

To further ensure the generality of above conclusion, we conduct the comparisons on another two datasets: Avazu and JD.com, and the results are shown in Tab.3. We can see that, in all three datasets, our FED outperforms other models, which again demonstrates that FED can integrate different level information to improve the accuracy of CTR prediction.

4 CONCLUSION

In this paper, we propose a novel module to model dimension relations named DRM. DRM helps to learn sample-level enhanced embedding vectors, which selectively emphasizes specific features and suppress less useful ones for the given task. DRM can be easily incorporated into the existing methods to boost their performance. Furthermore, we propose a unified model FED-net based on DRM, which models field-wise network and element-wise network jointly. Extensive experiments on three real-world datasets demonstrate the effectiveness of FED-net.

REFERENCES

[1] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*. ACM, 7–10.

[2] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. 2019. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3146–3154.

[3] 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).

[4] Yuchin Juan, Yong Zhuang, Wei-Sheng Chin, and Chih-Jen Lin. 2016. Field-aware factorization machines for CTR prediction. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 43–50.

[5] Jianxun Lian, Xiaohuan Zhou, Fuzheng Zhang, Zhongxia Chen, Xing Xie, and Guangzhong Sun. 2018. xDeepFM: Combining explicit and implicit feature interactions for recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 1754–1763.

[6] Bin Liu, Ruiming Tang, Yingzhi Chen, Jinkai Yu, Huifeng Guo, and Yuzhou Zhang. 2019. Feature generation by convolutional neural network for click-through rate prediction. In *The World Wide Web Conference*. 1119–1129.

[7] Yanru Qu, Bohui Fang, Weinan Zhang, Ruiming Tang, Minzhe Niu, Huifeng Guo, Yong Yu, and Xiuqiang He. 2018. Product-Based Neural Networks for User Response Prediction over Multi-Field Categorical Data. *ACM Transactions on Information Systems (TOIS)* 37, 1 (2018), 5.

[8] Steffen Rendle. 2010. Factorization machines. In *2010 IEEE International Conference on Data Mining*. IEEE, 995–1000.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[10] 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.

[11] Jun Xiao, Hao Ye, Xiangnan He, Hanwang Zhang, Fei Wu, and Tat-Seng Chua. 2017. Attentional factorization machines: Learning the weight of feature interactions via attention networks. *arXiv preprint arXiv:1708.04617* (2017).