Feature Decomposition for Reducing Negative Transfer: A Novel Multi-Task Learning Method for Recommender System (Student Abstract)

Jie Zhou, Qian Yu*, Chuan Luo, Jing Zhang

School of Software, Beihang University, China
{zhoujiee, qianyu, chuanluo, zhang_jing}@buaa.edu.cn

Abstract

We propose a novel multi-task learning method termed Feature Decomposition Network (FDN). The key idea of the proposed FDN is to reduce the phenomenon of feature redundancy by explicitly decomposing features into task-specific features and task-shared features with carefully designed constraints. Experimental results show that our proposed FDN can outperform the state-of-the-art (SOTA) methods by a noticeable margin on *Ali-CCP*.

Introduction

The recommendation system (RS) is an effective tool to help users to handle *information overload*. It has been widely used in many commercial fields, such as advertising computing, social networks, and e-commerce(Wen et al. 2019). To improve the user experience, the idea of using multi-task learning (MTL) to satisfy user requirements from multiple aspects has become increasingly popular in the RS community. Many deep learning (DL) based MTL models (Ma et al. 2018; Tang et al. 2020) have been proposed for RS. However, there is no explicit constraint to force these experts to extract features as required. As a result, the features captured by different experts may still be a mixture of the task-specific feature and the task-shared feature (as shown in Fig. 1-(a)), which is called *feature redundancy* phenomenon in this paper and will degrade the effectiveness of the model in handling the issue of negative transfer. In the ideal case (as shown in Fig. 1-(b)), feature spaces should be more pure.

This paper proposes a novel MTL method for recommendation system, named **F**eature **D**ecomposition **N**etwork (FDN). The key idea is to reduce features redundancy by decomposing with *explicit* constraints. Specifically, we introduce a **DeC**omposition **P**air (DCP) to capture task-shared and task-specific features, respectively. We conducted experiments on a public datasets (i.e., Ali-CCP dataset) to demonstrate the effectiveness of our proposed FDN.

Methodology

Preliminaries As shown in Fig. 2, FDN consists of multiple newly-proposed decomposition pairs (DCPs). For each task, there are M DCP modules. Each DCP module has a pair of experts, i.e. a task-specific expert and a task-shared expert. Task-specific expert, which is denoted by f^p, captures features that are specific for each task. Task-shared expert is denoted by f^s, which extracts features that are potentially shared by all tasks. After extracting the decomposed features, the model then fuses these features and feeds the features into the prediction layer of each task. The process can be represented as follows:

$$f_k^{MTL}(X) = \sigma(g_k(f^p(X), f^s(X))) \tag{1}$$

where $g_k(.)$ is a fusion function for task k and σ is an activation function.

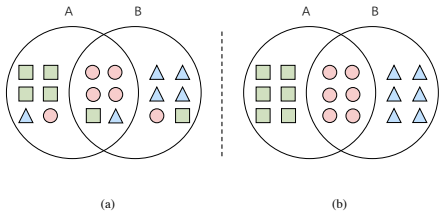

Figure 1: The green rectangle indicates the task-specific features of Task A, the blue triangle indicates the task-specific features of Task B, and the pink circle indicates task-shared features.

DeComposition Pair As explained before, the negative transfer is caused by feature redundancy. To address this problem, our idea is to reduce such redundancy by *explicitly* decomposing features into task-specific and task-shared features. Such design will make the model recombine features based on the requirements of each task. Three constraints are introduced into the DCP module.

Orthogonal Constraint: To decompose the original features into task-specific features and shared features as much as possible, we add the orthogonal constraint between the extracted features of the two experts. The orthogonal constraint is represented as follows:

$$L_{orth} = \sum_{k=1}^{K} \sum_{m=1}^{M} \left\| (f_m^s)^\top f_m^p \right\|_F^2 \tag{2}$$

where M denotes the number of modules of task k, f_m^s denotes the feature extracted by the m-th task-shared expert

*Corresponding author: Qian Yu.

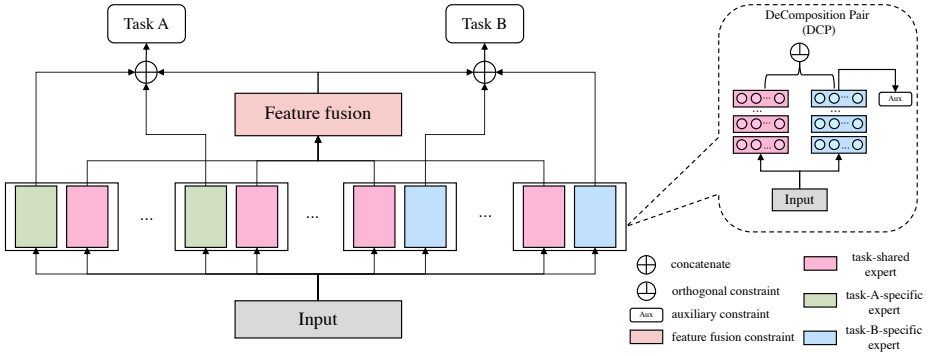

Figure 2: The architecture of the proposed Feature Decomposition Network (FDN): The gray rectangle box represents the original input, each DeComposition Pair (DCP) consists of two expert networks, including a task-shared expert network (pink box) and a task-specific expert network (green or blue box). The detail of the DCPs is shown in the dashed box. An auxiliary task head is added for each DCP to promote feature decomposition.

of tash k, and f_m^p is from the m-th task-specific expert of tash-k. $\|\cdot\|_F^2$ denotes squared Frobenius norm.

Auxiliary Task Constraint: The orthogonal constraint can not necessarily decompose the features into task-shared and task-specific. Therefore, to capture task-specific features, we introduce an auxiliary task to each task-specific expert as a regularizer for the prediction of a specific task (as shown in Fig. 2(Right) rounded rectangular box marked with "aux" text). It can be represented as follows:

$$L_{aux} = \sum_{k=1}^{K} \sum_{m=1}^{M} L_{k,m}\left(\hat{y}_m^k, y^k\right) \tag{3}$$

$$\hat{y}_m^k = \sigma(f_m^{k,p}(X)) \tag{4}$$

where $L_{k,m}(\cdot)$ is the k-th auxiliary task loss, and $\hat{y}_m^k$ denotes the prediction result of using the extracted feature of the m-th task-specific expert.

Task-shared Feature Fusion Constraint: The fusion of task-shared experts in DCPs is equivalent to a constraint, which will induce the model to further learn task-shared features. Once obtaining the decomposed features, the model fuses these features by concatenation ($g_k(.)$ in Eq. 1), as shown in Fig. 2.

Total Loss and Implementation Details: The final loss function of FDN is defined as $L = L_{task} + L_{orth} + L_{aux}$.

Experiments on Ali-CCP Dataset

To demonstrate the effectiveness of the newly proposed FDN, we conducted experiments on "Alibaba Click and Conversion Prediction" (Ali-CCP). The optimization tasks are including CTR and CVR respectively. We compare FDN with the SOTA models, Multi-gate Mixture-of-Experts (MMoE), and Progressive Layered Extraction (PLE).

The experimental results are shown in Table 1. FDN outperforms other models. We also provide the performance of single-task models for comparison. From the results, we can have the following observations: (1) our proposed model achieves the best performance on both tasks; (2) The performance of the MTL model, MMoE, is lower than single-task models, indicating the hard-sharing architecture can limit the model's expressivity; (3) Our model is efficient, and it has the least number of parameters.

Model	CTR/AUC	CVR/AUC	#params
Single-Task	0.6189	0.6248	-
MMoE	0.6175	0.6239	5.75×10^9
PLE	0.6241	0.6308	7.03×10^9
FDN (ours)	**0.6252**	**0.6783**	5.11×10^9

Table 1: Quantitative Results on Ali-CPP Dataset

Conclusion

In this paper, we propose a novel multi-task learning model, termed FDN. The core of FDN is a new feature decomposition module with three constraints. We conduct experiments on *Ali-CCP* and FDN outperforms the SOTA models.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62002012, No. 62202025 and No. 62006012).

References

Ma, J.; Zhao, Z.; Yi, X.; Chen, J.; Hong, L.; and Chi, E. H. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1930–1939.

Tang, H.; Liu, J.; Zhao, M.; and Gong, X. 2020. Progressive Layered Extraction (PLE): A Novel Multi-Task Learning (MTL) Model for Personalized Recommendations. In *Fourteenth ACM Conference on Recommender Systems*, 269–278.

Wen, H.; Zhang, J.; Lin, Q.; Yang, K.; and Huang, P. 2019. Multi-level deep cascade trees for conversion rate prediction in recommendation system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 338–345.