Appendix for Multi-Faceted Hierarchical Multi-Task Learning for Recommender Systems

1 ABLATION STUDY

Table 1 demonstrates the ablation result of removing Multi-Faceted and removing both Multi-faceted and hierarchical. Comparing H-MTL 9-task model with MFH 9-task model when we remove Multi-Faceted design, significant performance decreases are observed for all tasks unanimously. Comparing flat 9-task model with MFH 9-task model when we remove both Multi-Faceted and hierarchical design, even more significant performance decreases are observed for all tasks unanimously. Table 2 shows the performance gap between flat 9-task model and H-MTL 9-task model, which is the performance decrease of the ablation of hierarchical design.

Table 1: MFH Ablation on Play Task MTL

New User Group			
Models	**PCR MSE**	**PFR AUC**	**PSR AUC**
MFH 9-task	.5138	.7813	.8003
H-MTL 9-task	$.5156_{(+0.35\%)}$	$.7801_{(-0.15\%)}$	$.7989_{(-0.17\%)}$
flat 9-task	$.5162_{(+0.47\%)}$	$.7799_{(-0.18\%)}$	$.7965_{(-0.47\%)}$
Low-Activity User Group			
Models	**PCR MSE**	**PFR AUC**	**PSR AUC**
MFH 9-task	.4776	.8006	.8198
H-MTL 9-task	$.4792_{(+0.34\%)}$	$.8001_{(-0.06\%)}$	$.8186_{(-0.15\%)}$
flat 9-task	$.4796_{(+0.42\%)}$	$.7990_{(-0.20\%)}$	$.8178_{(-0.24\%)}$
High-Activity User Group			
Models	**PCR MSE**	**PFR AUC**	**PSR AUC**
MFH 9-task	.4070	.8199	.8563
H-MTL 9-task	$.4073_{(+0.07\%)}$	$.8183_{(-0.20\%)}$	$.8549_{(-0.16\%)}$
flat 9-task	$.4079_{(+0.22\%)}$	$.8181_{(-0.22\%)}$	$.8536_{(-0.32\%)}$

Table 2: Hierarchical Ablation on Play Task MTL

New User Group			
Models	**PCR MSE**	**PFR AUC**	**PSR AUC**
H-MTL 9-task	.5156	.7801	.7989
flat 9-task	$.5162_{(+0.12\%)}$	$.7799_{(-0.03\%)}$	$.7965_{(-0.30\%)}$
Low-Activity User Group			
Models	**PCR MSE**	**PFR AUC**	**PSR AUC**
H-MTL 9-task	.4792	.8001	.8186
flat 9-task	$.4796_{(+0.08\%)}$	$.7990_{(-0.14\%)}$	$.8178_{(-0.10\%)}$
High-Activity User Group			
Models	**PCR MSE**	**PFR AUC**	**PSR AUC**
H-MTL 9-task	.4073	.8183	.8549
flat 9-task	$.4079_{(+0.15\%)}$	$.8181_{(-0.02\%)}$	$.8536_{(-0.15\%)}$

We also conduct another experiment ablating the heterogeneity of MFH. The results are shown in Table 3 where heter-abl stands for the ablation experiment where we remove all heterogeneous designs and adopt a task tower network of two-layer MLP of size (128,64) as the high-activity user group tasks' settings for all user groups, instead of heterogeneous task towers. Compared to the baseline model, the ablation of heterogeneity decreases the performance of the tasks to varying degrees as shown in Table 3.

Table 3: Heterogeneity Ablation on Play Task MTL

New User Group			
Models	**PCR MSE**	**PFR AUC**	**PSR AUC**
MFH 9-task	.4560	.8414	.8564
heter-abl 9-task	$.4562_{(+0.03\%)}$	$.8410_{(-0.05\%)}$	$.8560_{(-0.04\%)}$
Low-Activity User Group			
Models	**PCR MSE**	**PFR AUC**	**PSR AUC**
MFH 9-task	.4750	.8381	.8483
heter-abl 9-task	$.4760_{(+0.21\%)}$	$.8377_{(-0.05\%)}$	$.8480_{(-0.04\%)}$
High-Activity User Group			
Models	**PCR MSE**	**PFR AUC**	**PSR AUC**
MFH 9-task	.3942	.8766	.8822
heter-abl 9-task	$.3945_{(+0.05\%)}$	$.8766_{(-0.00\%)}$	$.8821_{(-0.01\%)}$

2 MITIGATION OF LOCAL OVERFITTING

Through multi-dimensional shared learning between the tasks, MFH maximizes the shared representation learning and mitigates the local overfitting phenomenon. Here we use the *Share Rate* task in the interactive task group as an example to show the mitigation of local overfitting with MFH. As in Fig. 1, we compared MFH 12-task MTL and the baseline flat 12-task MTL on training and testing errors on new users. *train_impr* denotes the relative improvement of training error from MFH 12-task to flat 12-task on Share Rate prediction, and *test_impr* denotes the relative improvement of testing error accordingly. As the training time proceeds shown in Fig. 1, the test error is reduced much more than the training error, which is about 2.5 times higher. Thus, compared to the flat MTL models, MFH mitigates the local overfitting phenomenon by reducing the gap between training and testing errors.

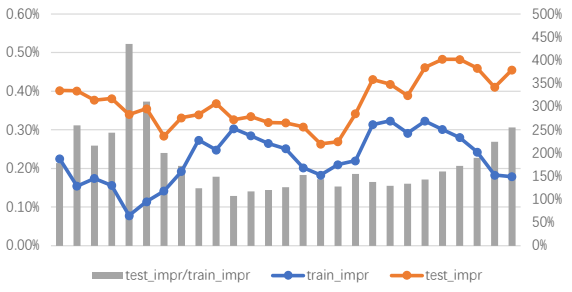

Figure 1: Mitigation of Local Overfitting on Share Rate in Interactive Tasks MTL

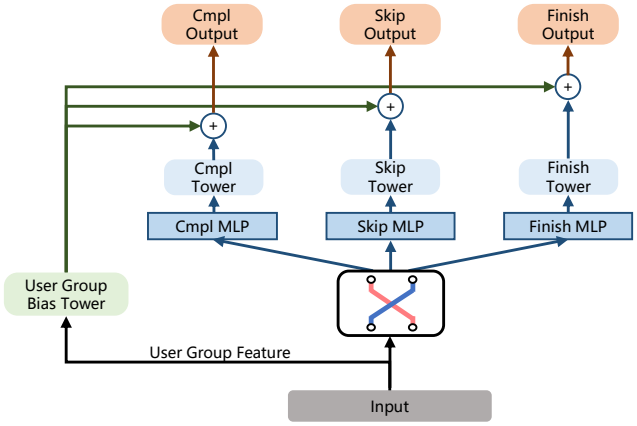

Figure 2: The Biasnet Model

3 COMPARISON WITH BIASNET

Biasnet [1, 2] is a natural alternative to address the user group differences with a side shallow bias tower that outputs a bias logit to be combined with the main tower. Viewing the neural network as a network learning a nonlinear mapping from the input feature space to the labels, the MTL approach provides a more general mapping than the biasnet approach as the task specific towers provide greater flexibility in fitting the label. Biasnet uses a stronger induction bias which may be beneficial for cases where the bias difference can be modeled with a few layers of simple neural mappings and the biased feature value is continuous. Thus, the MTL approach is a more general framework, which can actually adopt the biasnet structure as part of the MTL network. Offline evaluation is also conducted to compare MFH's performance with biasnet. As shown Fig. 2, we replace the user group facet with a bias tower that takes the user group feature as its input, and the bias output is added to the logits of task towers. The bias tower is composed of two-layer MLP of size (128,64) with RELU activation.

Table 4: Performance of MFH *vs.* Biasnet on Play Tasks. The improvements of *MFH* over *Biasnet* are shown in brackets.

New User Group			
Models	**PCR MSE**	**PFR AUC**	**PSR AUC**
Biasnet 9-task	.4457	.7767	.7914
MFH 9-task	**.4423**$_{(\textbf{-0.77\%})}$	**.7773**$_{(\textbf{+0.08\%})}$	**.7926**$_{(\textbf{+0.15\%})}$
Low-Activity User Group			
Models	**PCR MSE**	**PFR AUC**	**PSR AUC**
Biasnet 9-task	.4327	**.8019**	.8041
MFH 9-task	**.4276**$_{(\textbf{-1.19\%})}$	.8011$_{(-0.10\%)}$	**.8051**$_{(\textbf{+0.12\%})}$
High-Activity User Group			
Models	**PCR MSE**	**PFR AUC**	**PSR AUC**
Biasnet 9-task	.3829	**.8201**	.8497
MFH 9-task	**.3799**$_{(\textbf{-0.79\%})}$	.8191$_{(-0.12\%)}$	**.8504**$_{(\textbf{+0.08\%})}$

As Table 4 shows, the 2 facet MFH 9-task MTL outperforms Biasnet on the *Cmpl* task's PCR (Predicted Completion Ratio) MSE loss for all user groups, also outperforms Biasnet in all tasks for new users, and slightly underperforms Biasnet on PFR (Predicted Finish Rate) and PSR (Predicted Skip Rate) on partial user groups. As the PCR task and the new user group are more important, overall MFH performs better than biasnet.

4 IMPLEMENTATION DETAILS OF THE EXPERIMENTS

In this subsection, additional experiment details are supplemented for reproducibility. As introduced earlier, we tune some hyper-parameters such as the number of experts, number of layers in the MLP for the experimented models. In the play task group MTL, some additional tuned results for MFH 9-task model are: for the level 0 switcher, a simple shared-bottom switcher with a single layer MLP of size 512 is adopted; for the level 1 switchers, PLE is adopted, 2 experts for each of the three customized branches and 2 shared experts, every expert is implemented with a single layer MLP, the first level PLE expert uses a single-layer MLP of size 256 and the second level expert uses a single-layer MLP of size 128; for the level 2 switchers, CGC is adopted, 1 expert for each of the three customized branches and 1 shared expert, every expert is implemented with a single layer MLP of size 128; finally all the MLPs in Fig. 3 are set to be zero layer as the model complexity is already high.

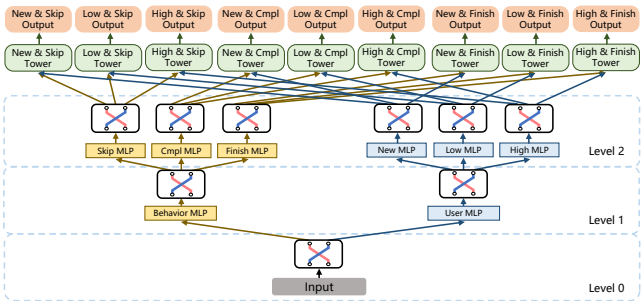

Figure 3: Multi-Faceted Hierarchical MTL Model(MFH)

REFERENCES

[1] John Moore, Joel Pfeiffer, Kai Wei, Rishabh Iyer, Denis Charles, Ran Gilad-Bachrach, Levi Boyles, and Eren Manavoglu. Modeling and simultaneously removing bias via adversarial neural networks. *arXiv preprint arXiv:1804.06909*, 2018.

[2] Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi, and Ed Chi. Recommending what video to watch next: a multitask ranking system. In *Proceedings of the 13th ACM RecSys*, pages 43–51, 2019.