MTRec: Multi-Task Learning over BERT for News Recommendation

Qiwei Bi[1*], **Jian Li**[2*], **Lifeng Shang**[2], **Xin Jiang**[2],
Qun Liu[2], **Hanfang Yang**[3,1†]
[1]School of Statistics, Renmin University of China
[2]Huawei Noah's Ark Lab
[3]Center for Applied Statistics, Renmin University of China
{bqw, hyang}@ruc.edu.cn
{lijian703, shang.lifeng, jiang.xin, qun.liu}@huawei.com

Abstract

Existing news recommendation methods usually learn news representations solely based on news titles. To sufficiently utilize other fields of news information such as category and entities, some methods treat each field as an additional feature and combine different feature vectors with attentive pooling. With the adoption of large pre-trained models like BERT in news recommendation, the above way to incorporate multi-field information may encounter challenges: the shallow feature encoding to compress the category and entity information is not compatible with the deep BERT encoding. In this paper, we propose a multi-task learning framework to incorporate the multi-field information into BERT, which improves its news encoding capability. Besides, we modify the gradients of different tasks based on their gradient conflicts, which further boosts the model performance. Extensive experiments on the MIND news recommendation benchmark show the effectiveness of our approach.

1 Introduction

Online News platforms such as Google News and MSN News have become a prevalent way for users to access news information (Das et al., 2007). To alleviate information overload and improve the reading experience, personalized news recommendation has become an essential part of these platforms (Liu et al., 2010; Phelan et al., 2011).

Traditional Recommendation models focus on modeling feature interactions (Rendle, 2012; Cheng et al., 2016; Guo et al., 2017; Wang et al., 2017). Accurate modeling of news and users is critical for news representation. Previous neural methods usually learn news representation vectors solely based on news titles and then learn a user representation by aggregating the previously browsed news via sequential or attentive models (Okura et al., 2017; An et al., 2019; Wu et al., 2019d). Though effective, these methods only utilize the title information and neglect other valuable news information such as categories and entities, which we call multi-field information. To fully utilize this information, as shown in Figure 1, existing methods usually transform each field of information (e.g., title, category, and entities) into a feature vector and combine different representations via attentive multi-field learning (Wu et al., 2019a, 2021a).

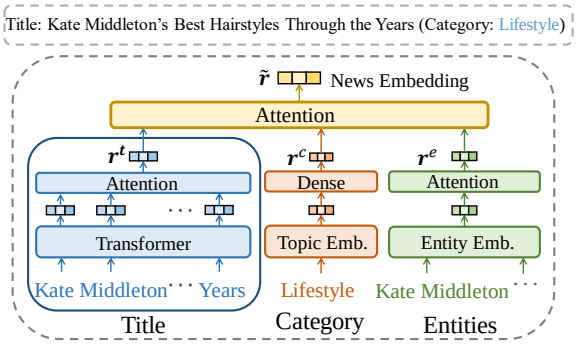

Figure 1: Traditional way to incorporate multi-field news information with attentive multi-field learning.

With the widespread use of large pre-trained language models, news recommendations start to adopt BERT (Devlin et al., 2019) as the cornerstone to encode news contents (e.g., encoding title as the blue box in Figure 1). However, when employing the above attentive way to combine other fields of information, we may encounter challenges: the shallow feature encoding to compress the category and entity information is *not compatible* with the deep BERT encoding for the title. Consequently, the ineffective adaption of multi-field information arises when we employ large pre-trained models in news recommendation.

In this paper, we propose a novel multi-task learning (Collobert and Weston, 2008; Stickland and Murray, 2019; Li et al., 2019) framework over BERT for news recommendation, named *MTRec*,

* Equal contribution. This work was done when Qiwei Bi was an intern at Huawei Noah's Ark Lab.
† Corresponding author

Findings of the Association for Computational Linguistics: ACL 2022, pages 2663 - 2669

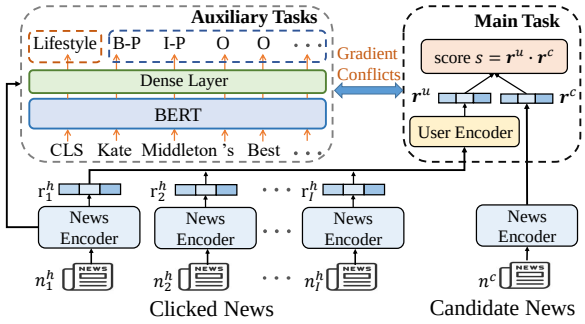

Figure 2: The overall framework of *MTRec*. We employ BERT as the news encoder and additive attention as the user encoder. In addition to the main task of news recommendation, we design two auxiliary tasks (i.e., category classification and NER) to further incorporate the category and entity information.

to effectively incorporate the multi-field information. Specifically, we use BERT to encode the news title as news embedding, and design two auxiliary tasks on top of BERT, i.e., category classification and named entity recognition (NER). The two auxiliary tasks are trained together with the main news recommendation task. We believe such a multi-task way can help BERT better capture the news semantics. To further improve the model performance, we adopt the recently proposed *gradient surgery* technique (Yu et al., 2020) which eliminates the gradient conflicts among different tasks during the multi-task training. While Zhang et al. (2021) study homogeneous multi-field news information including titles, abstracts and bodies, we study titles, categories and entities, which are heterogeneous thus can provide valuable information from different perspectives.

Finally, we find that combining the proposed multi-task learning and traditional attentive multi-field learning can further boost the performance of our model. Extensive experiments on the real-world MIND (Wu et al., 2020) news recommendation dataset show that *MTRec* can effectively improve the accuracy of news recommendation.

2 Method

Given I historical clicked news of a user $N^h = [n_1^h, n_2^h, ..., n_I^h]$ and a set of candidate news $N^c = [n_1^c, n_2^c, ..., n_J^c]$. Our goal is to calculate the user interest score s_j of each candidate news according to the historical behavior of the user, then the candidate news with the highest interest score is recommended to the user. For each news, we have its title text T, category label p^c, and entity set $\mathcal{E}$.

2.1 News Recommendation Framework

As shown in Figure 2, there are three main components in news recommendation framework, i.e., a *news encoder*, a *user encoder*, and a *click predictor*.

News Encoder For each news n, we encode its title with pre-trained BRET (Devlin et al., 2019). Specifically, we feed the tokenized text T into the BERT model and adopt the embedding of [CLS] token as the news representation $\mathbf{r}$. We denote the encoded vectors of historical clicked news N^h and candidate news N^c as $\mathbf{R}^h = [\mathbf{r}_1^h, \mathbf{r}_2^h, ..., \mathbf{r}_I^h]$ and $\mathbf{R}^c = [\mathbf{r}_1^c, \mathbf{r}_2^c, ..., \mathbf{r}_J^c]$, respectively.

User Encoder To gain a user representation from the representations of historical clicked news, existing methods usually employ sequential (An et al., 2019) or attentive models (Wu et al., 2019d; Li et al., 2018). In this paper, we adopt additive attention as the user encoder to compress the historical information $\mathbf{R}^h$. The user representation $\mathbf{r}^u$ is then denoted as:

$$\mathbf{r}^u = \sum_{i=1}^{I} a_i^u \mathbf{r}_i^h, \quad a_i^u = \text{softmax}(\mathbf{q}^u \cdot \tanh(\mathbf{W}^u \mathbf{r}_i^h)), \tag{1}$$

where $\mathbf{q}^u$ and $\mathbf{W}^u$ are trainable parameters.

Click Predictor For each candidate news, we obtain its interest score s_j by matching the candidate news vector $\mathbf{r}_j^c$ and the user representation $\mathbf{r}^u$ via dot product:

$$s_j = \mathbf{r}_j^c \cdot \mathbf{r}^u. \tag{2}$$

Loss Function Following previous work (Huang et al., 2013; Wu et al., 2019d), we employ the NCE loss to train the main ranking model. Then the main task loss $\mathcal{L}_{Main}$ is the negative log-likelihood of all positive samples in the training dataset $\mathcal{D}$:

$$\mathcal{L}_{Main} = -\sum_{i=1}^{|\mathcal{D}|} \log \frac{\exp(s_i^+)}{\exp(s_i^+) + \sum_{j=1}^{L} \exp(s_i^j)}, \tag{3}$$

where s^+ denotes the interest scores of positive news, L indicates the number of negative news.

2.2 Multi-Field Information

Besides the contents of news (e.g., titles), there is also other valuable information available in news recommendation, for example, category labels and entity annotations, which we call multi-field information. To fully utilize the multi-field information, existing methods usually treat them as additional input features (Wu et al., 2019a, 2021a). As the example in Figure 1, each field of information (i.e., title, category, and entities) is firstly transformed into

vectors via embedding lookup and attention mechanisms. Then the representations $\mathcal{R} = \{\mathbf{r}^t, \mathbf{r}^c, \mathbf{r}^e\}$ for title, category and entities are combined as the final news representation $\widetilde{\mathbf{r}}$ via attentive multi-field learning[1]:

$$\widetilde{\mathbf{r}} = \sum_{\mathbf{r}_i \in \mathcal{R}} w_i \mathbf{r}_i, \quad w_i = \text{softmax}(\mathbf{q}^r \cdot \tanh(\mathbf{W}^r \mathbf{r}_i)), \tag{4}$$

where $\mathbf{q}^r$ and $\mathbf{W}^r$ are trainable parameters.

Though effective with traditional text encoding, attentive multi-field learning may not work well with deep BERT encoding. Since the shallow feature encoding to compress the category and entity information may not be in the same feature space with the deep BERT encoding, directly combining them together may cause *incompatibility* problem thus ineffective use of multi-field information.

2.3 Multi-Task Learning

To effectively utilize the multi-field information with the BERT news encoder, we propose to employ multi-task learning with two auxiliary tasks on top of BERT: category classification and named entity recognition, as illustrated in Figure 2.

Category Classification To incorporate the news category information, we design a classification task on top of BERT, which uses the [CLS] embedding to predict the category distribution of news n_i:

$$\hat{\mathbf{p}}_i^c = \text{softmax}(\mathbf{W}^c \mathbf{r}_i + \mathbf{b}^c), \tag{5}$$

where $\mathbf{b}^c$ and $\mathbf{W}^c$ are trainable parameters. Then the loss function of category classification task is:

$$\mathcal{L}_{\text{Category}} = -\frac{1}{I} \sum_{i=1}^{I} \sum_{k=1}^{K^c} p_{i,k}^c \log(\hat{p}_{i,k}^c)), \tag{6}$$

where K^c is the number of categories.

Named Entity Recognition We also design a NER task (Lample et al., 2016) on top of BERT, so that the model can recognize important entities in the title thus better matching interested news. Specifically, we locate the given entities in the news title according to exact match and use "B" to indicate the beginning word of an entity, "I" to indicate the internal words. The other non-entity words in the title are denoted as "O". Then a tag prediction task is performed based on the BERT output embeddings:

$$\hat{\mathbf{p}}_{t_i}^n = \text{softmax}(\mathbf{W}^n \mathbf{r}^{t_i} + \mathbf{b}^n), \tag{7}$$

[1]Concatenation is another option but generally performs worse than attentive pooling.

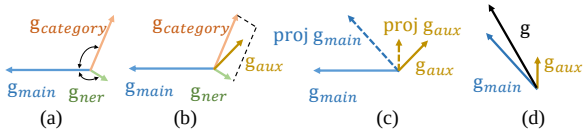

Figure 3: Illustration of the Gradient Surgery (GS).

where $\mathbf{r}^{t_i}$ is the output embedding of i-th token, $\mathbf{b}^n$ and $\mathbf{W}^n$ are trainable parameters. The loss function of the NER task is thus formulated as:

$$\mathcal{L}_{\text{NER}} = -\frac{1}{I} \sum_{i=1}^{I} \sum_{l=1}^{l_i} \sum_{k=1}^{K^n} p_{l,k}^n \log(\hat{p}_{l,k}^n)), \tag{8}$$

where K^n is the number of all NER tags, l_i is the title length of i-th news.

We optimize the loss function of the main task, category classification, and NER task simultaneously, which derives the final loss function:

$$\mathcal{L}_{\text{MTRec}} = \mathcal{L}_{\text{Main}} + \mathcal{L}_{\text{Category}} + \mathcal{L}_{\text{NER}}. \tag{9}$$

Multi-Task Learning with Gradient Surgery Yu et al. (2020) find that multi-task learning is not always beneficial, since there may exist *gradient conflicts* among different tasks. The problem means that the gradient directions of different tasks form an angle larger than 90° thus harm each other, as shown in Fig. 3(a). To alleviate this issue, Yu et al. (2020) propose a technique called Gradient Surgery (GS) that projects the gradient of the i-th task $\mathbf{g}_i$ onto the normal plane of another conflicting task's gradient $\mathbf{g}_j$:

$$\mathbf{g}_i = \mathbf{g}_i - \frac{(\mathbf{g}_j \cdot \mathbf{g}_i)}{\|\mathbf{g}_j\|^2} \cdot \mathbf{g}_j. \tag{10}$$

Though GS is effective to some degree, our task is a little different from the ordinary multi-task learning as Yu et al. (2020): we aim to use auxiliary tasks to boost the main task performance rather than treating them equally. Therefore, it would be beneficial to apply fewer gradient modifications to the main task. To this end, we slightly *revise* the original GS by firstly merging the gradients of auxiliary tasks, then adopt factor λ to scale them (Fig. 3(b)):

$$\mathbf{g}_{aux} = \lambda(\mathbf{g}_{category} + \mathbf{g}_{ner}), \tag{11}$$

where λ is empirically set to 0.3. Then we apply GS between the gradients of the main task and the merged auxiliary task (Fig. 3(c)) and derive the final gradient $\mathbf{g}$ (Fig. 3(d)).

MIND-small				
Methods	AUC	MRR	nDCG@5	nDCG@10
NAML	66.12	31.53	34.88	41.09
LSTUR	65.87	30.78	33.95	40.15
NRMS	65.63	30.96	34.13	40.52
HieRec	67.95	32.87	36.36	42.53
BERT (baseline)	68.26	32.52	35.89	42.33
LSTUR+BERT	68.28	32.58	35.99	42.32
NRMS+BERT	68.60	32.97	36.55	42.78
BERT+AMF	68.96	33.42	37.10	43.27
MTRec	69.43	33.79	37.64	43.74
MTRec+AMF	**69.51**	**34.06**	**38.05**	**44.03**

Table 1: Performance of different methods. *MTRec* is our proposed multi-task method and "AMF" denotes attentive multi-field learning.

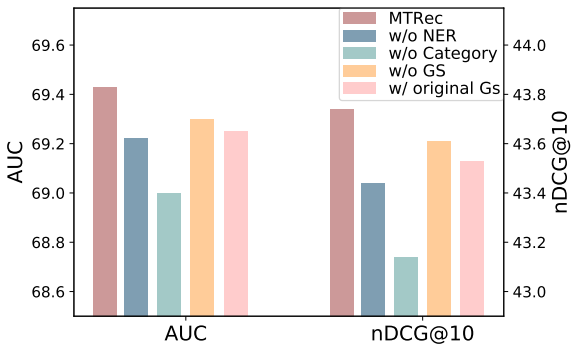

Figure 4: Ablation study to show the effectiveness of auxiliary tasks and gradient surgery (GS).

3 Experiment

3.1 Dataset and Settings

We evaluate our approach on a real-world news recommendation dataset MIND (Wu et al., 2020), and we use the small version for quick experiments. Following previous work (Wu et al., 2019b; Qi et al., 2021), we utilize users' most recent 50 clicked news as historical behavior and each positive news is paired with 4 negative news. More details about the settings are in the Appendix A.

We compare our approach against several competitive baselines including NAML (Wu et al., 2019a), LSTUR (An et al., 2019), NRMS (Wu et al., 2019d), HieRec (Qi et al., 2021). While the above methods all adopt shallow text encodings, we also employ BERT as the news encoder, implementing a BERT baseline. Further, we reproduce two best-performing BERT-based methods (Wu et al., 2021b), denoted as LSTUR+BERT and NRMS+BERT. We also combine attentive multi-field learning to incorporate the multi-field information with the BERT baseline and MTRec, denoted as BERT+AMF and MTRec+AMF respectively.

3.2 Results

The main experimental results are listed in Table 1, from which we have the following observations. Firstly, the news recommendation system clearly performs better when BERT is utilized as the news encoder. For example, LSTUR+BERT and NRMS+BERT, for which we only replace the news encoder with BERT in LSTUR and NRMS, surpass their shallow versions significantly. Secondly, BERT+AMF performs better than the BERT baseline, which proves the value of the multi-field information. Different users prefer different categories and entities of news and this information is beneficial for the system to make personalized recommendations. Thirdly, MTRec performs significantly better than BERT+AMF, indicating the effectiveness of the multi-task learning strategy. It's worth noting that the attentive multi-field learning applies Glove (Pennington et al., 2014) and TransE (Bordes et al., 2013) embeddings to vectorize the information of categories and entities respectively. We claim that these feature encodings may not be in the same feature space as the deep BERT encoding, thus causing the insufficient use of multi-field information in BERT+AMF. Finally, MTRec+AMF achieves the best results. Ruder (2017) proposes that multi-task learning can be regarded as a kind of regularization. Thus, we deduce that the attentive multi-field learning, which augments the news representation directly, is not in conflict with the multi-task learning in MTRec.

3.3 Ablation Study

Auxiliary Tasks Firstly, we drop the category classification and NER tasks respectively to explore their impacts on the system. As shown in Figure 4, the model performances decrease to varying degrees when only introducing a single auxiliary task. But their performances are still better than the BERT baseline, which proves that both auxiliary tasks contribute additional information to BERT. Wu et al. (2019c) only utilizes the title and category, which is denoted as w/o NER in Figure 4. Note that the performance drops the most when we remove the category classification task, possibly due to that categories are document-level labels and contain richer information than entities.

Gradient Surgery Further, we remove the Gradient Surgery technique in MTRec. As shown in Figure 4, the model performance drops greatly, which verifies the benefits to alleviate the gradient con-

flicts among different tasks. When we apply the original Gradient Surgery as Yu et al. (2020) in MTRec, the performances even get worse. The reason is that we aim to use auxiliary tasks to boost the main task performance rather than treating them equally, which is different from the ordinary multi-task learning. We also record and plot the gradient cosine similarity between the main and merged auxiliary task during training in the Appendix B.

4 Conclusion

We propose a novel multi-task learning framework over BERT for news recommendation, named MTRec, to effectively incorporate the multi-field information. We also modify the Gradient Surgery technique to reduce gradient conflicts and further improve the model performance. Finally, we find that combining multi-task learning with traditional attentive multi-field learning achieves the best results. Extensive experiments on the MIND dataset show the effectiveness of our approach. In the future, we will also combine MTRec with more advanced user modeling methods (Li et al., 2022).

5 Acknowledgements

This work was supported by National Key R&D Program of China (Grant No. 2018YFC2000302). We thank the anonymous reviewers for their insightful comments.

References

Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural news recommendation with long-and short-term user representations. In *ACL*, pages 336–345.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NIPS*, volume 26.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. 2016. Wide & deep learning for recommender systems. In *(DLRS@RecSys*, pages 7–10.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*, pages 160–167.

Abhinandan S Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. 2007. Google news personalization: scalable online collaborative filtering. In *WWW*, pages 271–280.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. Deepfm: a factorization-machine based neural network for ctr prediction. In *IJCAI*.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *CIKM*, pages 2333–2338.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *NAACL*, pages 260–270.

Jian Li, Zhaopeng Tu, Baosong Yang, Michael R Lyu, and Tong Zhang. 2018. Multi-head attention with disagreement regularization. In *EMNLP*, pages 2897–2903.

Jian Li, Baosong Yang, Zi-Yi Dou, Xing Wang, Michael R Lyu, and Zhaopeng Tu. 2019. Information aggregation for multi-head attention with routing-by-agreement. In *NAACL*, pages 3566–3575.

Jian Li, Jieming Zhu, Qiwei Bi, Guohao Cai, Lifeng Shang, Zhenhua Dong, Xin Jiang, and Qun Liu. 2022. Miner: Multi-interest matching network for news recommendation. In *Findings of ACL*.

Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. 2010. Personalized news recommendation based on click behavior. In *IUI*, pages 31–40.

Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. 2017. Embedding-based news recommendation for millions of users. In *KDD*, pages 1933–1942.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Owen Phelan, Kevin McCarthy, Mike Bennett, and Barry Smyth. 2011. Terms of a feather: Content-based news recommendation and discovery using twitter. In *ECIR*, pages 448–459.

Tao Qi, Fangzhao Wu, Chuhan Wu, Peiru Yang, Yang Yu, Xing Xie, and Yongfeng Huang. 2021. Hierec: Hierarchical user interest modeling for personalized news recommendation. In *ACL*, pages 5446–5456.

Steffen Rendle. 2012. Factorization machines with libfm. *ACM TIST*, 3(3):1–22.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Asa Cooper Stickland and Iain Murray. 2019. BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning. In *ICML*, pages 5986–5995.

Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. 2017. Deep & cross network for ad click predictions. In *ADKDD*, pages 12:1–12:7.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019a. Neural news recommendation with attentive multi-view learning. In *IJCAI*, pages 3863–3869.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019b. Npa: Neural news recommendation with personalized attention. In *KDD*, pages 2576–2584.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Yongfeng Huang, and Xing Xie. 2019c. Neural news recommendation with topic-aware news representation. In *ACL*, pages 1154–1159.

Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019d. Neural news recommendation with multi-head self-attention. In *EMNLP*, pages 6390–6395.

Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. 2021a. User-as-graph: User modeling with heterogeneous graph pooling for news recommendation. In *IJCAI*.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021b. Empowering news recommendation with pre-trained language models. In *SIGIR*, page 1652–1656.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. 2020. Mind: A large-scale dataset for news recommendation. In *ACL*, pages 3597–3606.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. In *NeurIPS*.

Qi Zhang, Qinglin Jia, Chuyuan Wang, Jingjie Li, Zhaowei Wang, and Xiuqiang He. 2021. Amm: Attentive multi-field matching for news recommendation. In *SIGIR*, pages 1588–1592.

A Dataset and Settings

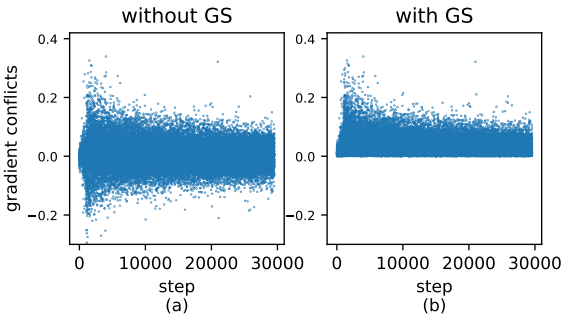

Figure 5: The fluctuation of cosine similarity for the main task and the merged auxiliary task. 'GS' indicated the gradient surgery.

Dataset We evaluate our approach on a real-world news recommendation dataset MIND (Wu et al., 2020), which is collected from the user behavior logs of Microsoft News. There are two versions of the dataset, namely MIND-large and MIND-small. The MIND-large contains more than 15 million impression logs generated by 1 million users, from which the MIND-small randomly samples 50,000 users. An impression log records the clicked and non-clicked news that are displayed to a user at a specific time and his historical news click behaviors before this impression. Besides, MIND contains off-the-shelf category labels and a set of entities of each news.

Settings Following previous work (Wu et al., 2019b; Qi et al., 2021), we utilize users' most recent 50 clicked news as historical behavior. We use *bert-base-uncased* pre-trained model as the news encoders. Only news title is used as the model input in this paper and the maximum length is set to 20. The dimension of the query vector in the additive attention is set as 200. Following previous work (Wu et al., 2019b; Qi et al., 2021), we apply Glove (Pennington et al., 2014) and TransE (Bordes et al., 2013) embeddings to vectorize the information of categories and entities respectively. The total number of news categories is 19 and 22 entity classes are identified in this paper. The embeddings dimension of the entities and categories are 100, and both are finetuned during model training. For the embedding of categories and entities, we also apply a dense layer to align the feature dimensions with the corresponding title encodings. The negative sampling rate L is set to 4 during training, i.e., each positive news is paired with 4 negative

news. The learning rate is set to $2e^{-5}$ and linearly decayed with 10% warmup steps. We employ Adam (Kingma and Ba, 2015) as the optimization algorithm. As previous work (Wu et al., 2020), we employ four ranking metrics, i.e., AUC, MRR, nDCG@5, and nDCG@10, for evaluation.

B Gradient Conflicts

As shown in the Figure 5, we record and plot the gradient cosine similarity between the main and merged auxiliary task $\frac{\mathbf{g}_{main} \cdot \mathbf{g}_{aux}}{\|\mathbf{g}_{main}\| \|\mathbf{g}_{aux}\|}$ in each step. It's easy to find that there are often conflicts (negative points) between the main task and the merged auxiliary task before applying the gradient modification (Fig. 5(a)). Contrastively, our method eliminate these conflicts (Fig. 5(b)). There is no doubt that it is great internal consumption for optimization if the gradient directions among different tasks are opposite. Without alleviating the gradient conflicts, the model cannot balance multiple tasks well. In this case, the multi-filed auxiliary tasks are even harmful to the performance of the recommendation system.