Neural News Recommendation with Multi-Head Self-Attention

Chuhan Wu[1], **Fangzhao Wu**[2], **Suyu Ge**[1], **Tao Qi**[1], **Yongfeng Huang**[1],**and Xing Xie**[2]
[1]Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[2]Microsoft Research Asia, Beijing 100080, China
{wu-ch19, gsy17, qit16, yfhuang}@mails.tsinghua.edu.cn,
{fangzwu, xing.xie}@microsoft.com

Abstract

News recommendation can help users find interested news and alleviate information overload. Precisely modeling news and users is critical for news recommendation, and capturing the contexts of words and news is important to learn news and user representations. In this paper, we propose a neural news recommendation approach with multi-head self-attention (NRMS). The core of our approach is a news encoder and a user encoder. In the news encoder, we use multi-head self-attentions to learn news representations from news titles by modeling the interactions between words. In the user encoder, we learn representations of users from their browsed news and use multi-head self-attention to capture the relatedness between the news. Besides, we apply additive attention to learn more informative news and user representations by selecting important words and news. Experiments on a real-world dataset validate the effectiveness and efficiency of our approach.

1 Introduction

Online news platforms such as Google News[1] and MSN News[2] have attracted many users to read news online (Das et al., 2007). Massive news articles are generated everyday and it is impossible for users to read all news to find their interested content (Phelan et al., 2011). Thus, personalized news recommendation is very important for online news platforms to target user interests and alleviate information overload (IJntema et al., 2010).

Learning accurate representations of news and users are two core tasks in news recommendation (Okura et al., 2017). Several deep learning based methods have been proposed for these tasks (**?**Kumar et al., 2017; Khattar et al., 2018; Wu et al., 2019b,c,a; Zhu et al., 2019; An et al., 2019). For example, Okura et al. (2017) proposed to learn news representations from news bodies via auto-encoders, and learn representations of users from their browsed news via GRU. However, GRU is quite time-consuming, and their method cannot capture the contexts of words. Wang et al. (2018) proposed to learn news representations from news titles via a knowledge-aware convolutional neural network (CNN), and learn representations of users based on the similarities between candidate news and their browsed news. However, CNN cannot capture the long-distance contexts of words, and their method cannot model the relatedness between browsed news.

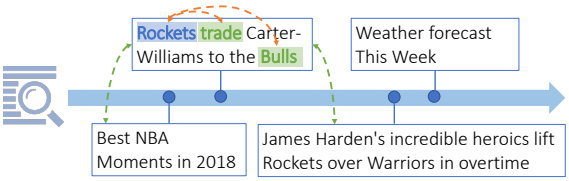


Figure 1: Several news browsed by an example user. Orange and green dashed lines represent the interactions between words and news respectively.

Our work is motivated by several observations. First, the interactions between words in news title are important for understanding the news. For example, in Fig. 1, the word "Rockets" has strong relatedness with "Bulls". Besides, a word may interact with multiple words, e.g., "Rockets" also has semantic interactions with "trade". Second, different news articles browsed by the same user may also have relatedness. For example, in Fig. 1 the second news is related to the first and the third news. Third, different words may have different importance in representing news. In Fig. 1, the word "NBA" is more informative than "2018". Besides, different news articles browsed by the same user may also have different importance in repre-

[1]https://news.google.com/
[2]https://www.msn.com/en-us/news

senting this user. For example, the first three news articles are more informative than the last one.

In this paper, we propose a neural news recommendation approach with multi-head self-attention (NRMS). The core of our approach is a news encoder and a user encoder. In the news encoder, we learn news representations from news titles by using multi-head self-attention to model the interactions between words. In the user encoder, we learn representations of users from their browsing by using multi-head self-attention to capture their relatedness. Besides, we apply additive attentions to both news and user encoders to select important words and news to learn more informative news and user representations. Extensive experiments on a real-world dataset show that our approach can effectively and efficiently improve the performance of news recommendation.

2 Our Approach

Our NRMS approach for news recommendation is shown in Fig. 2. It contains three modules, i.e., *news encoder*, *user encoder* and *click predictor*.

2.1 News Encoder

The *news encoder* module is used to learn news representations from news titles. It contains three layers. The first one is word embedding, which is used to convert a news title from a sequence of words into a sequence of low-dimensional embedding vectors. Denote a news title with M words as $[w_1, w_2, ..., w_M]$. Through this layer it is converted into a vector sequence $[\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_M]$.

The second layer is a word-level multi-head self-attention network (Vaswani et al., 2017; Wu et al., 2018). The interactions between words are important for learning news representations. For example, in the news title "Rockets Ends 2018 with a Win", the interaction between "Rockets" and "Win" is useful for understanding this news, and such long-distance interactions usually cannot be captured by CNN. In addition, a word may interact with multiple words in the same news. For instance, in above example the word "Rockets" has interactions with both "Ends" and "Win". Thus, we propose to use multi-head self-attention to learn contextual representations of words by capturing their interactions. The representation of the i_{th} word learned by the k_{th} attention head is

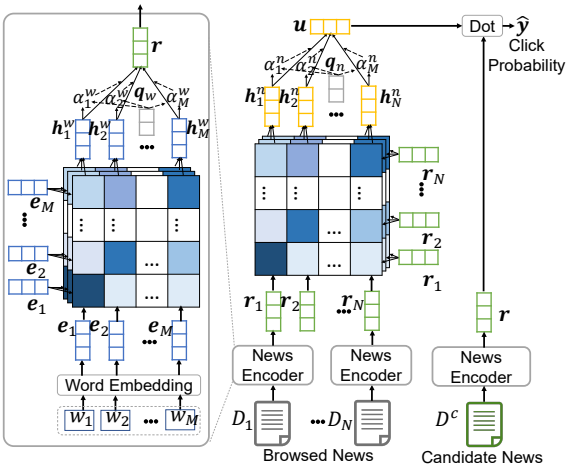


Figure 2: The framework of our *NRMS* approach.

computed as:

$$\alpha_{i,j}^k = \frac{\exp(\mathbf{e}_i^T \mathbf{Q}_k^w \mathbf{e}_j)}{\sum_{m=1}^{M} \exp(\mathbf{e}_i^T \mathbf{Q}_k^w \mathbf{e}_m)}, \quad (1)$$

$$\mathbf{h}_{i,k}^w = \mathbf{V}_k^w (\sum_{j=1}^{M} \alpha_{i,j}^k \mathbf{e}_j), \quad (2)$$

where $\mathbf{Q}_k^w$ and $\mathbf{V}_k^w$ are the projection parameters in the k_{th} self-attention head, and $\alpha_{i,j}^k$ indicates the relative importance of the interaction between the i_{th} and j_{th} words. The multi-head representation $\mathbf{h}_i^w$ of the i_{th} word is the concatenation of the representations produced by h separate self-attention heads, i.e., $\mathbf{h}_i^w = [\mathbf{h}_{i,1}^w; \mathbf{h}_{i,2}^w; ...; \mathbf{h}_{i,h}^w]$.

The third layer is an additive word attention network. Different words in the same news may have different importance in representing this news. For example, in the second news of Fig. 1, the word "NFL" is more informative than "Today" for understanding this news. Thus, we propose to use attention mechanism to select important words in news titles for learning more informative news representations. The attention weight α_i^w of the i-th word in a news title is computed as:

$$a_i^w = \mathbf{q}_w^T \tanh(\mathbf{V}_w \times \mathbf{h}_i^w + \mathbf{v}_w), \quad (3)$$

$$\alpha_i^w = \frac{\exp(a_i^w)}{\sum_{j=1}^{M} \exp(a_j^w)}, \quad (4)$$

where $\mathbf{V}_w$ and $\mathbf{v}_w$ are projection parameters, and $\mathbf{q}_w$ is the query vector. The final representation of a news is the weighted summation of the contextual word representations, formulated as:

$$\mathbf{r} = \sum_{i=1}^{M} \alpha_i^w \mathbf{h}_i^w. \quad (5)$$

2.2 User Encoder

The *user encoder* module is used to learn the representations of users from their browsed news. It contains two layers. The first one is a news-level multi-head self-attention network. Usually, news articles browsed by the same user may have some relatedness. For example, in Fig. 1, the first two news articles are related. In addition, a news article may interact with multiple news articles browsed by the same user. Thus, we propose to apply multi-head self-attention to enhance the representations of news by capturing their interactions. The representation of the i_{th} news learned by the k_{th} attention head is formulated as follows:

$$\beta_{i,j}^k = \frac{\exp(\mathbf{r}_i^T \mathbf{Q}_k^n \mathbf{r}_j)}{\sum_{m=1}^M \exp(\mathbf{r}_i^T \mathbf{Q}_k^n \mathbf{r}_m)}, \tag{6}$$

$$\mathbf{h}_{i,k}^n = \mathbf{V}_k^n (\sum_{j=1}^M \beta_{i,j}^k \mathbf{r}_j), \tag{7}$$

where $\mathbf{Q}_k^n$ and $\mathbf{V}_k^n$ are parameters of the k_{th} news self-attention head, and $\beta_{i,j}^k$ represents the relative importance of the interaction between the j_{th} and the k_{th} news. The multi-head representation of the i_{th} news is the concatenation of the representations output by h separate self-attention heads, i.e., $\mathbf{h}_i^n = [\mathbf{h}_{i,1}^n; \mathbf{h}_{i,2}^n; ...; \mathbf{h}_{i,h}^n]$.

The second layer is an additive news attention network. Different news may have different informativeness in representing users. For example, in Fig. 1 the first news is more informative than the fourth news in modeling user interest, since the latter one is usually browsed by massive users. Thus, we propose to apply the additive attention mechanism to select important news to learn more informative user representations. The attention weight of the i_{th} news is computed as:

$$a_i^n = \mathbf{q}_n^T \tanh(\mathbf{V}_n \times \mathbf{h}_i^n + \mathbf{v}_n), \tag{8}$$

$$\alpha_i^n = \frac{\exp(a_i^n)}{\sum_{j=1}^N \exp(a_j^n)}, \tag{9}$$

where $\mathbf{V}_n$, $\mathbf{v}_n$ and $\mathbf{q}_n$ are parameters in the attention network, and N is the number of the browsed news. The final user representation is the weighted summation of the representations of the news browsed by this user, which is formulated as:

$$\mathbf{u} = \sum_{i=1}^N \alpha_i^n \mathbf{h}_i^n. \tag{10}$$

2.3 Click Predictor

The *click predictor* module is used to predict the probability of a user clicking a candidate news. Denote the representation of a candidate news D^c as $\mathbf{r}^c$. Following (Okura et al., 2017), the click probability score $\hat{y}$ is computed by the inner product of the user representation vector and the news representation vector, i.e., $\hat{y} = \mathbf{u}^T \mathbf{r}^c$. We also explored other kinds of scoring methods such as perception, but dot product shows the best performance and efficiency.

2.4 Model Training

Motivated by (Huang et al., 2013), we use negative sampling techniques for model training. For each news browsed by a user (regarded as a positive sample), we randomly sample K news which are shown in the same impression but not clicked by the user (regarded as negative samples). We shuffle the orders of these news to avoid possible positional biases. Denote the click probability score of the positive and the K negative news as $\hat{y}^+$ and $[\hat{y}_1^-, \hat{y}_2^-, ..., \hat{y}_K^-]$ respectively. These scores are normalized by the softmax function to compute the posterior click probability of a positive sample as follows:

$$p_i = \frac{\exp(\hat{y}_i^+)}{\exp(\hat{y}_i^+) + \sum_{j=1}^K \exp(\hat{y}_{i,j}^-)}. \tag{11}$$

We re-formulate the news click probability prediction problem as a pseudo $(K+1)$-way classification task, and the loss function for model training is the negative log-likelihood of all positive samples $\mathcal{S}$, which is formulated as follows:

$$\mathcal{L} = -\sum_{i \in \mathcal{S}} \log(p_i). \tag{12}$$

3 Experiments

3.1 Datasets and Experimental Settings

We conducted experiments on a real-world news recommendation dataset collected from MSN News[3] logs in one month (Dec. 13, 2018 to Jan. 12, 2019). The detailed statistics are shown in Table 1. The logs in the last week were used for test, and the rest were used for training. We randomly sampled 10% of training data for validation.

[3]https://www.msn.com/en-us/news

# users	10,000	avg. # words per title	11.29
# news	42,255	# positive samples	489,644
# impressions	445,230	# negative samples	6,651,940

Table 1: Statistics of our dataset.

In our experiments, the word embeddings are 300-dimensional and initialized by the Glove embedding (Pennington et al., 2014). The self-attention networks have 16 heads, and the output of each head is 16-dimensional. The dimension of the additive attention query vectors is 200. Following (Wu et al., 2019b), the negative sampling ratio K is 4. Adam (Kingma and Ba, 2014) is used for model optimization. We apply 20% dropout to the word embeddings to mitigate overfitting. The batch size is 64. These hyperparameters are tuned on validation set. We conducted experiments on a machine with Xeon E5-2620 v4 CPUs and a GTX1080Ti GPU. We independently repeated each experiment 10 times and reported average results in terms of AUC, MRR, nDCG@5 and nDCG@10.

3.2 Performance Evaluation

We evaluate the performance of our approach by comparing it with several baseline methods, including: (1) *LibFM* (Rendle, 2012), a matrix factorization based recommendation method; (2) *DSSM* (Huang et al., 2013), deep structured semantic model; (3) *Wide&Deep* (Cheng et al., 2016), a popular neural recommendation method; (4) *DeepFM* (Guo et al., 2017), another popular neural recommendation method; (5) *DFM* (Lian et al., 2018), deep fusion model for news recommendation; (6) *DKN* (Wang et al., 2018), deep knowledge-aware network for news recommendation; (7) *Conv3D* (Khattar et al., 2018), a neural news recommendation method with 3-D CNNs to learn user representations; (8) *GRU* (Okura et al., 2017), a neural news recommendation method using GRU to learn user representations; (9) *NRMS*, our approach. In methods (1) and (3-5), we use one-hot encoded user ID, news ID and the TF-IDF features extracted from news titles as the model input. In methods (6-9), we all use news titles for fair comparison. The results of these methods are summarized in Table 2.

We have several observations from Table 2. First, neural recommendation methods such as *DSSM* and *NRMS* outperform traditional recommendation methods such as *LibFM* on news recommendation. This may be because neural networks can learn better representations of news and users than matrix factorization methods. Thus, it may be more appropriate to learn news and user representations via neural networks rather than craft them manually. Second, among the deep learning based methods, the methods which exploit the relatedness between news (e.g., *Conv3D*, *GRU* and *NRMS*) can outperform other methods. This may be because the news browsed by the same user usually have relatedness, and capturing the news relatedness is useful for understanding these news and modeling user interests. Third, our approach performs better than all baseline methods. This is because our approach can capture the interactions between both words and news via multi-head self-attention to enhance representation learning of news and users. Besides, our approach employs additive attention to select important words and news for learning informative news and user representations. These results validate the effectiveness of our approach.

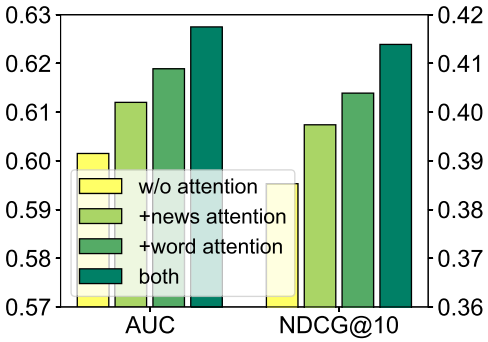
Methods	AUC	MRR	nDCG@5	nDCG@10
LibFM	0.5661	0.2414	0.2689	0.3552
DSSM	0.5949	0.2675	0.2881	0.3800
Wide&Deep	0.5812	0.2546	0.2765	0.3674
DeepFM	0.5830	0.2570	0.2802	0.3707
DFM	0.5861	0.2609	0.2844	0.3742
DKN	0.6032	0.2744	0.2967	0.3873
Conv3D	0.6051	0.2765	0.2987	0.3904
GRU	0.6102	0.2811	0.3035	0.3952
NRMS*	**0.6275**	**0.2985**	**0.3217**	**0.4139**

Table 2: The results of different methods. *The improvement is significant at $p < 0.01$.

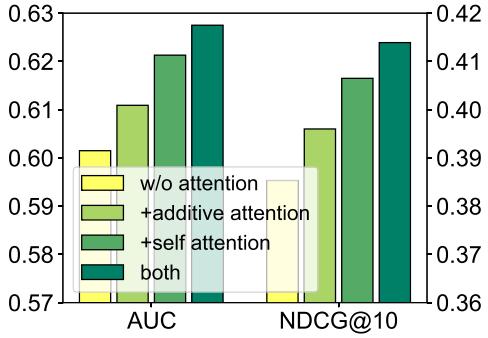
We also conducted experiments to compare the time efficiency of our approach with several popular news recommendation methods. The results are shown in Table 3. From the results, we find our approach has a smaller parameter size and a lower time complexity in learning news and user representations than existing news recommendation methods. In addition, different from *DKN*, our approach does not need to memorize the news browsing histories of users when computing the click probability scores. In addition, since our approach can be further accelerated by computing the hidden representations of different attention heads in parallel, our approach is more suitable for being deployed in large-scale news recommendation scenarios. These results validate the efficiency of our approach.

Method	# Parameters	News Encoding Time	User Encoding Time
DKN	681 K	46.4 s	10.1 min
Conv3D	575 K	29.8 min	16.7 min
GRU	541 K	59.8 s	148.0 min
NRMS	530 K	39.7 s	6.7 min

Table 3: The number of parameters in *NRMS* and baseline methods, and their time in encoding 1 million news and 1 million users. *Word embeddings are excluded.



(a) Attention mechanisms at different levels.



(b) Attention mechanisms of different kinds.

Figure 3: Effectiveness of different attention networks.

3.3 Effectiveness of Attention Mechanism

Next we explore the effectiveness of attentions in our approach. First, we verify the word- and news-level attentions. The results are shown in Fig. 3(a). We find the word-level attention is very useful. This is because modeling the interactions between words and selecting important words can help learn informative news representations. Besides, the news-level attention is also useful. This is because capturing the relatedness of news and selecting important news can benefit the learning of user representations. Moreover, combining both word- and news-level attentions can further improve the performance of our approach.

We also study the influence of additive and self-attentions on our approach. The results are shown in Fig. 3(b). From these results, we find the self-attentions are very useful. This is because the interactions between words and news are important for understanding news and modeling users. In addition, the additive attentions are also helpful. This is because different words and news may usually have different importance in representing news and users. Thus, selecting important words and news can help learn more informative news and user representations. Combining both additive and self-attention can further improve our approach. Thus, these results validate the effectiveness of the attention mechanism in our approach.

4 Conclusion and Future Work

In this paper we propose a neural news recommendation approach with multi-head self-attention. The core of our approach is a news encoder and a user encoder. In both encoders we apply multi-head self-attentions to learn contextual word and news representations by modeling the interactions between words and news. In addition, we use additive attentions to select important words and news to learn more informative news and user representations. Extensive experiments validate the effectiveness and efficiency of our approach.

In our future work, we will try to improve our approach in the following potential directions. First, in our framework we do not consider the positional information of words and news, but they may be useful for learning more accurate news and user representations. We will explore position encoding techniques to incorporate the word position and the time-stamps of news clicks to further enhance our approach. Second, we will explore how to effectively incorporate multiple kinds of news information in our framework, especially long sequences such as news body, which may challenge the efficiency of typical self-attention networks.

Acknowledgments

The authors would like to thank Microsoft News for providing technical support and data in the experiments, and Jiun-Hung Chen (Microsoft News) and Ying Qiao (Microsoft News) for their support and discussions. This work was supported by the National Key Research and Development Program of China under Grant number 2018YFC1604002, the National Natural Science Foundation of China under Grant numbers U1836204, U1705261, U1636113, U1536201, and U1536207.

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.

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 *DLRS*, pages 7–10.

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

Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. Deepfm: a factorization-machine based neural network for ctr prediction. In *AAAI*, pages 1725–1731. AAAI Press.

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. ACM.

Wouter IJntema, Frank Goossen, Flavius Frasincar, and Frederik Hogenboom. 2010. Ontology-based news recommendation. In *Proceedings of the 2010 EDBT/ICDT Workshops*, page 16.

Dhruv Khattar, Vaibhav Kumar, Vasudeva Varma, and Manish Gupta. 2018. Weave&rec: A word embedding based 3-d convolutional network for news recommendation. In *CIKM*, pages 1855–1858.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Vaibhav Kumar, Dhruv Khattar, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep neural architecture for news recommendation. In *CLEF (Working Notes)*.

Jianxun Lian, Fuzheng Zhang, Xing Xie, and Guangzhong Sun. 2018. Towards better representation learning for personalized news recommendation: a multi-channel deep fusion approach. In *IJCAI*, pages 3805–3811.

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

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. Springer.

Steffen Rendle. 2012. Factorization machines with libfm. *TIST*, 3(3):57.

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 *NIPS*, pages 5998–6008.

Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. Dkn: Deep knowledge-aware network for news recommendation. In *WWW*, pages 1835–1844.

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. ACM.

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, Junxin Liu, Sixing Wu, Yongfeng Huang, and Xing Xie. 2018. Detecting tweets mentioning drug name and adverse drug reaction with hierarchical tweet representation and multi-head self-attention. In *SMM4H*, pages 34–37.

Qiannan Zhu, Xiaofei Zhou, Zeliang Song, Jianlong Tan, and Li Guo. 2019. Dan: Deep attention neural network for news recommendation. In *AAAI*, volume 33, pages 5973–5980.