通过从字符级、短语级和句子级别编码新闻，整合为三维矩阵在通过3DCNN匹配，组成细粒度的多层次的新闻推荐模型。

Fine-grained Interest Matching for Neural News Recommendation

Heyuan Wang[†], **Fangzhao Wu**[†], **Zheng Liu**[†], **and Xing Xie**[†]
[†] Microsoft Research Asia, Beijing 100080, China
heyuanww@163.com, wufangzhao@gmail.com,
{zhengliu, xingx}@microsoft.com

Abstract

Personalized news recommendation is a critical technology to improve users' online news reading experience. The core of news recommendation is accurate matching between user's interests and candidate news. The same user usually has diverse interests that are reflected in different news she has browsed. Meanwhile, important semantic features of news are implied in text segments of different granularities. Existing studies generally represent each user as a single vector and then match the candidate news vector, which may lose fine-grained information for recommendation. In this paper, we propose FIM, a Fine-grained Interest Matching method for neural news recommendation. Instead of aggregating user's all historical browsed news into a unified vector, we hierarchically construct multi-level representations for each news via stacked dilated convolutions. Then we perform fine-grained matching between segment pairs of each browsed news and the candidate news at each semantic level. High-order salient signals are then identified by resembling the hierarchy of image recognition for final click prediction. Extensive experiments on a real-world dataset from MSN news validate the effectiveness of our model on news recommendation.

1 Introduction

Recently, people's news reading habits have gradually shifted to digital content services. Many online news websites, such as Google News [1] and MSN News [2], aim to collect news from various sources and distribute them for users (Das et al., 2007; Lavie et al., 2010). However, the overwhelming number of newly-sprung news makes it difficult for users to find their interested content (Wu et al., 2019c). Therefore, personalized news recommendation becomes an important technology to alleviate information overload and improve users' online reading experience (IJntema et al., 2010).

Figure 1: Example of one user's reading behavior from MSN News. The user has various interests including NFL sports, pets and the issue about weight loss. The highlighted text segments are crucial semantic clues, and the arrows of different colors indicate the relevant matching pairs for candidate news recommendation.

The key to news recommendation lies in the accurate matching of user's interests and candidate news. The same user usually has diverse interests, which are reflected in different news she has browsed. Meanwhile, the important semantic features of news are implied in text segments of different granularities. Figure 1 illustrates the challenges with an example. As demonstrated, different historical browsed news can reveal user's interests about different topics or events. The first and second historical news are about pet dogs and the issue of weight loss respectively. Naturally, they provide critical clues to select the candidate news C_2 and C_3 which reveal relevant information. However, they are less informative to identify the candidate news C_1, which is about the competition of National Football League (NFL). Besides, the matched segment pairs across browsed news and candidate news lie in different granularities, such as the words "Dog's"-"puppy" and phrases "lost 245 pounds"-"Weight Loss". Moreover, different segments in news texts have different importance

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

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 836–845
July 5 - 10, 2020.

for selecting proper news candidates. For example, in the third historical browsed news D_3, "Philip Rivers" and "Chiefs" are more important than other words like "hilariously" and "after" for inferring that the user is a fan of NFL, since they refer to the famous quarterback and team of this sport.

Existing work, however, usually learns a single representation for each user by integrating all historical news that the user has browsed, then recommendations are performed by matching the final user vector and the candidate news vector (Okura et al., 2017; Wu et al., 2019e,b). For instance, Okura et al. (2017) encode news via denoising auto-encoders, and learn representations of users from their browsed news via a GRU network. Wu et al. (2019e) apply multi-head self-attentions to learn news representations, then learn user representations by modeling the relatedness between browsed news. Wu et al. (2019b) enhance personalized news and user representations by exploiting the embedding of user's ID to generate a query vector for attending to important words and news. Despite the improvements of these methods in news recommendation performance, they are limited in capturing fine-grained user-news matching signals, since user's various latent interests implied in distinct historical readings cannot match with the candidate news until the final step of click prediction.

In this paper, we propose a Fine-grained Interest Matching network (FIM), which is a new architecture for news recommendation that can tackle the above challenges. The advantages of FIM lie in two cores: the multi-level user/news representation and the fine-grained interest matching. Instead of representing each user as a single abstract vector, we employ hierarchical dilated convolutions in a unified module to construct multi-level representations of each news article based on the title and category annotations. By hierarchically stacking the dilated convolutions, the receptive input width at each layer grows exponentially, while the number of parameters increases only linearly. Meanwhile, the outputs of each layer are preserved as feature maps across different length of text segments, with no loss in coverage since any form of pooling or stride convolution is not applied. In this way, we can gradually obtain the semantic features of news from local correlation and long-term dependency at different granularities, including word, phrase, and sentence levels.

Furthermore, to avoid information loss, FIM matches the text segments of the candidate news and each historical news browsed by the user at each semantic granularity. In practice, for each pair of news, the model constructs a segment-segment similarity matrix from word-level to sentence-level based on the hierarchical news representations. By this means, user's reading interests implied in the browsing history can be recognized under the supervision of candidate news, and carried into matching with minimal loss, so as to provide sufficient clues about the content relevance for recommending proper news. Afterwards, we merge the multiple matching matrices of each news pair at each granularity into a 3D image, whose channels indicate the relevant degrees of different kinds of user-news matching patterns. By resembling the CNN-based hierarchy of image recognition, higher-order salient signals are identified to predict the probability of the user clicking the candidate news.

We conducted extensive experiments on a real-world dataset collected from MSN news. Experimental results validate that our approach can effectively improve the performance of news recommendation compared with the state-of-the-art methods.

2 Related Works

With the explosive growth of digital news, building personalized news recommender systems has drawn more attentions in both natural language processing and data mining fields (Phelan et al., 2011; Zheng et al., 2018; Wu et al., 2019a). Conventional news recommendation methods focus on utilizing manual feature engineering to build news and user representations for matching (Phelan et al., 2009; Li et al., 2010; Liu et al., 2010; Son et al., 2013; Li et al., 2014; Bansal et al., 2015). For example, Liu et al. (2010) used topic categories and interest features generated by a Bayesian model to build news and user representations. Son et al. (2013) extracted topic and location features from Wikipedia pages to build news representations for location-based news recommendation.

In recent years, deep learning based models have achieved better performance than traditional methods for news recommendation, due to their capabilities of distilling implicit semantic features in news content (Okura et al., 2017; Wang et al., 2018; An et al., 2019; Wu et al., 2019e,d). For example, Okura et al. (2017) learned news representations via denoising auto-encoders, then used recurrent neural networks to aggregate historical browsed

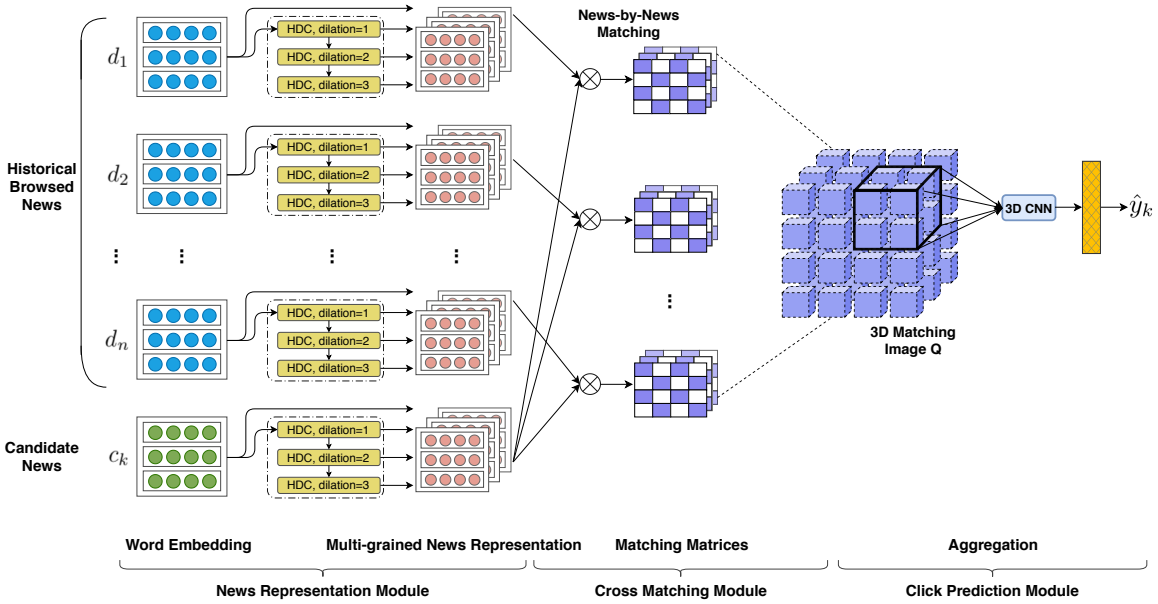


Figure 2: Architecture of our FIM model. *HDC* (*hierarchical dilated convolution*) is the news encoder.

news to learn user representations. Wang et al. (2018) enhanced the representation of news by exploiting the embeddings of extracted entities in a knowledge graph as a separate channel of the CNN input. Wu et al. (2019e) leveraged multi-head self-attentions to construct news representations based on the interactions between words, and constructed user representations based on the relatedness between news. An et al. (2019) proposed to learn long-term user preferences from the embeddings of their IDs, and learn short-term user interests from their recently browsed news via GRU network. (Wu et al., 2019a) proposed an attentive multi-view learning model to learn unified news representations from titles, bodies and topic categories by regarding them as different views of news. Different from these existing methods, in FIM, the representations of user's multiple browsed news are not fused into an abstract user vector before matching with the candidate news. Instead, we perform matching between each pair of segments in the news texts from multiple semantic levels. Therefore, more fine-grained information can be distilled for the final recommendation.

3 Our Approach

3.1 Problem Definition

The news recommendation problem can be formulated as follows. Given a user u, the set of historical news she has browsed at the online news platform is formulated as $s_u = \{d_1, \ldots, d_n\}$. For a news candidate c_i, a binary label $y_i \in \{0, 1\}$ is adopted to indicate whether u will click c_i in latter impressions. The aim is to build a prediction model $g(\cdot, \cdot)$. For each pair of user and candidate news (u, c), we can predict the probability that u would like to click c using the function $g : s_u, c \rightarrow \hat{y}$. Recommendations are performed based on the ranking of candidate news according to their click scores.

3.2 Model Overview

We present a Fine-grained Interest Matching network (FIM) to model $g(\cdot, \cdot)$. The architecture of FIM is illustrated in Figure 2, which contains three major components, i.e., a news representation module to construct hierarchical semantic features for news text segments, a cross interaction module to exploit and aggregate matching information from each pair of news at each level of granularity, and a prediction module to calculate the probability that the user will click the candidate news. Next, we introduce each component in detail.

3.2.1 News Representation Module

We design a *hierarchical dilated convolution (HDC)* encoder to learn representations of news from multiple semantic views. Besides titles that can reflect the central information of news, at many digital platforms such as MSN, news articles are usually labeled with a category annotation (e.g., "sports", "entertainment") and a subcategory annotation (e.g., "football_nba", "movies_celebrity") to help indicate news topics and target users' in-

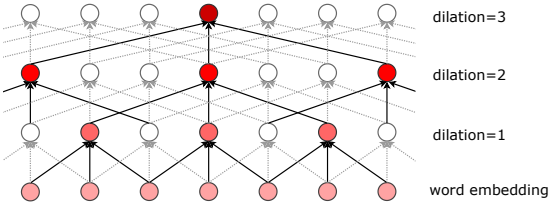


Figure 3: Hierarchical Dilated Convolution (HDC).

terests. HDC encodes each news by connecting its title, category and subcategory annotations into a sequence of words as input. Given the word sequence $d = [x_1, \ldots, x_N]$, where N is the sequence length, the model first looks up an embedding table to transform d into a matrix $\mathbf{d}^0 = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, where $\mathbf{x}_j \in \mathbb{R}^d$ is a d-dimensional word embedding. Then hierarchical dilated convolution layers are applied to capture multi-grained semantic features in news texts. Different from standard convolution that convolves a contiguous subsequence of the input at each step, dilated convolution (Yu and Koltun, 2016) has a wider receptive field by skipping over δ input elements at a time, where δ is the dilation rate. For a context of $\mathbf{x}_j$ and a convolution kernel $\mathbf{W}$ of size $2w+1$, the dilated convolution operation is:

$$F(x_t) = ReLU(\mathbf{W} \bigoplus_{k=0}^{w} \mathbf{x}_{j \pm k\delta} + \mathbf{b}), \quad (1)$$

where $\bigoplus$ is the vector concatenation, $\mathbf{b}$ is the bias and ReLU (Nair and Hinton, 2010) is the nonlinear activation function. As shown in Figure 3, the darker output of each convolution layer is a weighted combination of the lighter regular spaced inputs in the previous layer. We start with $\delta = 1$ (equals to standard convolution) for the first layer to ensure that no element of the input sequence is excluded. Afterwards, by hierarchically stacking the dilated convolutions with wider dilation rates, the length of convolved text segments expands exponentially, and the semantic features of different n-grams can be covered using only a few layers and a modest number of parameters.

Moreover, to prevent vanishing or exploding of gradients, we apply layer normalization (Ba et al., 2016) at the end of each convolution layer. Since there may be irrelevant information introduced to semantic units at a long distance, we practically design the multi-level dilation rates based on the performance in validation. The output of each stacked layer l is preserved as feature maps of the news text at a specific level of granularity, formulated as $\mathbf{d}^l = [\mathbf{x}_j^l]_{j=1}^N \in \mathbb{R}^{N \times f_s}$, where f_s is the number of filters for each layer. Suppose there are L layers stacked, the multi-grained news representations can be defined as $[\mathbf{d}^0, \mathbf{d}^1, \ldots, \mathbf{d}^L]$. By this means, HDC gradually harvests lexical and semantic features from word and phrase levels with small dilation rates, and captures long dependences from sentence level with larger dilation rates. Meanwhile, the computational path is greatly shortened, and the negative effects of information loss caused by down-sampling methods such as max-pooling can be reduced. Our news encoder is superior to the recurrent units in parallel ability and the entirely attention-based approach in reducing token-pair memory consumptions.

3.2.2 Cross Interaction Module

Given representations of the k-th browsed news $[\mathbf{d}_k^l]_{l=0}^L$ and the candidate news $[\mathbf{c}^l]_{l=0}^L$, a segment-segment matching matrix is constructed for each granularity, i.e., $\mathbf{M}_{k,c}^l \in \mathbb{R}^{N_{d_k} \times N_c}$, where $l \in \{0, L\}$ is the semantic level, N_{d_k} and N_c are the length of the news d_k and c. The (i,j)-th element of $\mathbf{M}_{k,c}^l$ is calculated by scaled dot product as:

$$\mathbf{M}_{k,c}^l[i,j] = \frac{\mathbf{d}_k^l[i] \cdot \mathbf{c}^l[j]^T}{\sqrt{f_s}}, \quad (2)$$

indicating the relevance between the i-th segment in d_k and the j-th segment in c according to the l-th representation type. The $L+1$ matching matrices for the news pair $<d_k, c>$ can be viewed as different feature channels of their matching information.

To summarize the information of user's entire reading sequence, FIM fuses all interaction matrices across each browsed news and the candidate news into a 3D matching image $\mathbf{Q}$, formulated as:

$$\mathbf{Q} = \{Q_{k,i,j}\}_{n \times N_{d_k} \times N_c}, \quad (3)$$

where n denotes the total number of browsed news in user history, and each pixel $Q_{k,i,j}$ is defined as:

$$Q_{k,i,j} = [\mathbf{M}_{k,c}^l[i,j]]_{l=0}^L. \quad (4)$$

Specifically, each pixel is a concatenated vector with $L+1$ channels, indicating the matching degrees between a certain segment pair of the news content at different levels of granularity.

As user's click behaviors may be driven by personalized interests or temporary demands and events, different historical browsed news has different usefulness and representativeness for matching

and recommending the proper candidate news. Inspired by Zhou et al. (2018) in the issue of dialogue system, we resemble the compositional hierarchy of image recognition, and employ a layered 3D convolution & max-pooling neural network to identify the salient matching signals from the whole image. The 3D convolution is the extension of typical 2D convolution, whose filters and strides are 3D cubes. Formally, the higher-order pixel at (k, i, j) on the z-th feature map of the t-th layer is computed as:

$$\mathbf{Q}_{k,i,j}^{(t,z)} = ELU\left(\sum_{z'}\sum_{w=0}^{W_t-1}\sum_{h=0}^{H_t-1}\sum_{r=0}^{R_t-1}\mathbf{K}_{w,h,r}^{(t,z)} \cdot \mathbf{Q}_{k+w,i+h,j+r}^{(t-1,z')} + \mathbf{b}^{(t)}\right), \tag{5}$$

where z' denotes each feature map of the previous layer, $\mathbf{K}^{(t,z)} \in \mathbb{R}^{W_t \times H_t \times R_t}$ is a 3D convolution kernel with the size of $W_t \times H_t \times R_t$, and $\mathbf{b}^{(t)}$ is the bias for the t-th layer. A max pooling operation is then adopted to extract salient signals as follows:

$$\widehat{\mathbf{Q}_{k,i,j}^{(t,z)}} = max\left(\mathbf{Q}_{[k:k+P_w^{(t,z)}-1],[i:i+P_h^{(t,z)}-1],[j:j+P_r^{(t,z)}-1]}^{(t,z)}\right), \tag{6}$$

where $P_w^{(t,z)}$, $P_h^{(t,z)}$ and $P_r^{(t,z)}$ are sizes of 3D max-pooling. Outputs of the final layer are concatenated as the integrated matching vector between the user and the candidate news, denoted as $\mathbf{s}_{u,c} \in \mathbb{R}^v$.

3.2.3 Click Prediction Module

In the recommendation scenario studied in this paper, recommendations are made based on ranking the candidate news articles according to their probabilities of being clicked by a user in an impression. Given the integrated matching vector $\mathbf{s}_{u,c}$ of a user and candidate news pair, the final click probability is calculated as:

$$\hat{y}_{u,c} = \mathbf{W}_o^T \mathbf{s}_{u,c} + b_o\,, \tag{7}$$

where $\mathbf{W}_o$ and b_o are learned parameters.

Motivated by (Huang et al., 2013b) and (Wu et al., 2019e), we leverage the negative sampling technique for model training. For each news browsed by a user (regarded as a positive sample), we randomly sample K news which are showcased in the same impression but not clicked by the user as negative samples. Besides, the orders of these news are shuffled to avoid positional biases. FIM jointly predicts the click probability scores of the positive news and the K negative news during training. By this means, the news click prediction problem is reformulated as a $(K+1)$-way classification task. The loss function is designed to minimize the summation of negative log-likelihood of all positive samples, which is defined as:

$$-\sum_{i=1}^{S}\log\frac{\exp(\hat{y}_{u_i,c_i}^{+})}{\exp(\hat{y}_{u_i,c_i}^{+}) + \sum_{k=1}^{K}\exp(\hat{y}_{u_i,c_{i,k}}^{-})}\,, \tag{8}$$

where S is the number of positive training samples, and $c_{i,k}$ is the k-th negative sample in the same impression with the i-th positive sample.

4 Experiments

4.1 Dataset and Experimental Settings

We conducted experiments on the Microsoft News dataset used in (Wu et al., 2019b)[3], which was built from the user click logs of Microsoft News[4]. The detailed statistics are shown in Table 1. Logs in the last week were used for test, and the rest for model training. Besides, we randomly sampled 10% of logs in the training data for validation.

In our experiments, the word embeddings are 300-dimensional and initialized using pre-trained Glove embedding vectors (Pennington et al., 2014). Due to the limitation of GPU memory, the maximum length of the concatenated word sequence of news title and category is set to 20, and at most 50 browsed news are kept for representing the user's recently reading behaviors. We tested stacking 1-5 HDC layers with different dilation rates. The reported results utilize [1-2-3] hierarchy (dilation rate for each convolution layer) as it gains the best performance on the validation set. The window size and number of convolution filters for news representation are 3 and 150 respectively. For the cross interaction module, we use two-layered composition to distill higher-order salient features of the 3D matching image, and the number and window size of 3D convolution filters are 32-[3,3,3] for the first layer and 16-[3,3,3] for the second layer, with [1,1,1] stride. The followed max-pooling size is [3,3,3] with [3,3,3] stride. Meanwhile, the negative sampling ratio K is set to 4. Adam (Kingma and Ba, 2014) is used as the optimizer, the mini-batch size is 100, and the initial learning rate is 1e-3.

Following the settings of state-of-the-art methods (Okura et al., 2017; Wu et al., 2019e), we use popular ranking metrics to evaluate the performance of each model, including AUC (Area

[3] A large-scale public version of Microsoft News dataset for news recommendation can be found at https://msnews.github.io

[4] https://microsoftnews.msn.com

# users	10,000	# topic categories	14
# news	42,255	# subtopic categories	284
# impressions	445,230	# positive samples	489,644
avg. # words per title	11.29	# negative samples	6,651,940

Table 1: Statistics of the dataset.

Under the ROC Curve) (Bradley, 1997), MRR (Mean Reciprocal Rank) (Voorhees et al., 1999), and NDCG (Normalized Discounted Cumulative Gain) (Järvelin and Kekäläinen, 2002). We independently repeated each experiment for 10 times and reported the average performance.

4.2 Comparison Methods

We compare FIM with the following methods:

Manual Feature-based Methods: Traditional recommendation methods which rely on manual feature engineering to build news and user representations, including (1) *LibFM* (Rendle, 2012), a feature-based matrix factorization model that is widely used in recommendations. We extract TF-IDF features from users' browsed news and candidate news, and concatenate them as the input for *LibFM*; (2) *DSSM* (Huang et al., 2013a), a deep structured semantic model with word hashing via character trigram and multiple dense layers. All browsed news are merged into a long document as the query; (3) *Wide & Deep* (Cheng et al., 2016), a popular recommendation method that combines a wide channel for linear transformations and a deep channel with multiple dense layers. The same features with *LibFM* are used for both channels; (4) *DeepFM* (Guo et al., 2017), combining factorization machines and deep neural networks with the same features as *LibFM*.

Neural Recommendation Methods: Neural networks specially designed for news recommendation, including (1) *DFM* (Lian et al., 2018), a deep fusion model combining dense layers with different depths and using attention mechanism to select important features; (2) *DKN* (Wang et al., 2018), incorporating entity information in knowledge graphs with Kim CNN (Kim, 2014) to learn news representations and using news-level attention network to learn user representations; (3) *GRU* (Okura et al., 2017), using auto-encoders to represent news and a GRU network to represent users; (4) *NRMS* (Wu et al., 2019e), leveraging multi-head self-attentions for news and user representation learning; (5) *Hi-Fi Ark* (Liu et al., 2019), summarizing user history into highly compact and complementary vectors as archives, and learning candidate-dependent user representation via attentive aggregation of such archives; (6) *NPA* (Wu et al., 2019b), using personalized attention with user ID's embedding as the query vector to select important words and news.

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
GRU	0.6102	0.2811	0.3035	0.3952
NRMS	<u>0.6275</u>	0.2985	0.3217	0.4139
Hi-Fi Ark	0.6027	0.3162	0.3335	0.4204
NPA	0.6243	<u>0.3321</u>	<u>0.3535</u>	<u>0.4380</u>
FIM	**0.6359***	**0.3354***	**0.3582***	**0.4436***
FIM$_{first}$	0.6258	0.3266	0.3484	0.4348
FIM$_{last}$	0.6319	0.3323	0.3549	0.4407

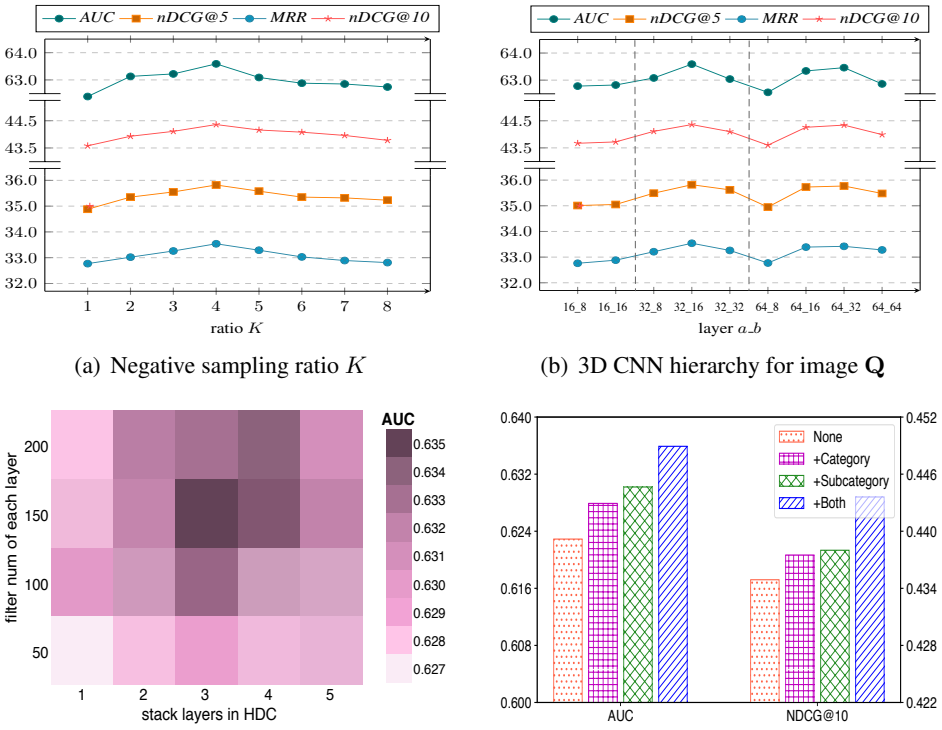
Table 2: The performance of different methods on news recommendation. The best and second best results are highlighted in boldface and underlined respectively. *The improvement over all baseline methods is significant at p-value < 0.05.

Ablation Variants: To verify the effects of multi-grained representation and sequential matching, we further setup two comparing ablation models, i.e., (1) *FIM*$_{first}$: a variant in which we use feature maps of the first news representation layer for matching and recommendation. In this scenario, the HDC module degenerates into a one-layer standard CNN encoder. (2) *FIM*$_{last}$: a variant using the outputs of the last layer in HDC (namely, the L-th embedding type) to represent each news for matching. Due to the hierarchical representation architecture, higher-level features synthesize information from lower-level features, and can model more complex lexical and semantic clues.

4.3 Experimental Results

Table 2 shows the results of our model and all comparative methods. Several observations can be made. First, neural news recommendation methods (e.g., *GRU*, *NRMS*, *Hi-Fi Ark*, *NPA*) are generally better than traditional methods (e.g., *LibFM*, *DeepFM*) that are based on manual feature engineering. The reason might be that handcrafted features are usually not optimal, and deep neural networks take the advantages of extracting implicit semantic features and modeling latent relationships between user and news representations.

Second, our model FIM consistently outperforms other baselines in terms of all metrics, including the state-of-the-art deep learning based mod-



(a) Negative sampling ratio K

(b) 3D CNN hierarchy for image $\mathbf{Q}$

(c) Hierarchy of HDC news encoder

(d) Incorporating two-level category annotations

Figure 4: Performances w.r.t. different hyper-parameters and input information.

els. This validates the advantage of the pair-wise multi-level matching architecture in synthetically detecting fine-grained matching information from news segment pairs to predict the probability of a user clicking a candidate news.

Third, both FIM$_{first}$ and FIM$_{last}$ show a decrease of performance compared to FIM. The latter is better than the former, indicating the effectiveness of constructing higher-level representations on the basis of low levels via the hierarchical mechanism of *HDC*. Besides, compared with DKN that utilizes knowledge-enhanced CNNs to learn news representations, FIM$_{first}$ has a better performance, illustrating the advantage of pair-wise matching fashion. Another notable thing is that while FIM$_{last}$ underperforms FIM, it can outperform all other competitors on all metrics. However, the benefit of interacting news pairs at multi-grained semantic levels is still significant.

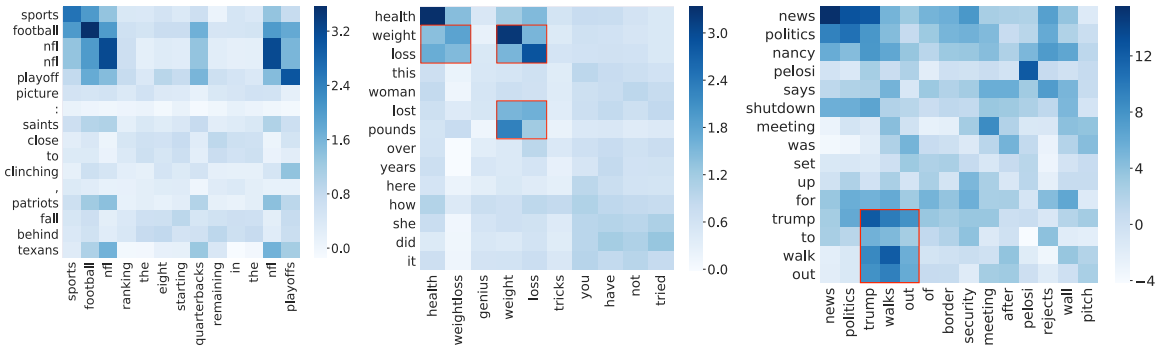
5 Analysis

In this section, we further investigate the impacts of different parameters and inputs on the model performance, and discuss the contribution of multi-grained representation and matching architecture.

5.1 Quantity & Input Analysis

We first study how FIM perfroms with different negative sampling ratio K. Figure 4(a) shows the experimental results. We can find that the performance consistently improves when K is lower than 5, then begins to decline. The possible reason is that with a too small K, the useful information exploited from negative samples is limited. However, when too many negative samples are incorporated, they may become dominant and the imbalance of training data will be increased. Thus it is more difficult for the model to precisely recognize the positive samples, which will also affect the recommendation performance. Overall, the optimal setting of K is moderate (e.g., $K = 4$).

We then explore the influence of the 3D convolution & max-pooling neural network for processing the matching image $\mathbf{Q}$. Comparing results are illustrated in Figure 4(b), where the CNN hierarchy a_b means that the number of filters for the first layer and the second layer are set to a and b, separately. As shown, given the filter number a for the first layer, the performance first increases with a larger filter number b for the second layer, since more high-order information can be extracted. Then the performance begins to decrease, possibly because



(a) $\mathbf{M}_1$ (b) $\mathbf{M}_2$ (c) $\mathbf{M}_3$

Figure 5: Matching matrices visualization, darker area means larger value.

more noisy patterns are introduced to the model (e.g., the group of [32_8, 32_16, 32_32]). Besides, a similar trend exists in the hierarchies with the same value b and different value a (e.g., the group of [16_8, 32_8, 64_8]). We conduct other experiments by changing the window size in [2,3,4,5] and the number of convolution layers in [1,2,3]. Results show that the optimal hierarchy is two-layered CNNs, with 32×[3,3,3] filters for the first layer and 16×[3,3,3] filters for the second layer.

We further compare different combinations of the number of dilated convolution filters and stacked layers in the HDC news representation module. Figure 4(c) demonstrates the results, where darker areas represent larger values. We observe a consistent trend over settings with different number of filters at each layer, i.e., there is a significant improvement during the first few stacked layers, and then the performance decreases a lot when the depth grows to 5. The results indicate that depth of representation layers indeed matters in terms of matching and recommendation accuracy. The optimal setting of the number of stacked layers and convolution filters is 3 and 150 respectively. We think the reason might be that in this scenario, the perceived field of dilated convolution filters at each layer ranges among [3-7-13] (with dilation rates as [1-2-3]), which is sufficient for modeling multi-grained n-gram features through hierarchical composition of local interactions, compared to the average length of news word sequences.

We also investigate the effectiveness of incorporating two-level category annotations of news as inputs. The results are shown in Figure 4(d). We can find that incorporating either categories or subcategories can benefit the performance of our model. This is interpretable since category annotations are helpful to reveal user's interested aspects more explicitly. In addition, enhancing news representations with subcategories is better than with categories. This is probably because compared to the general category labels, subcategories can provide more concrete and detailed information to indicate the core topic of news content. Overall, jointly incorporating the two-level category annotations can achieve the best performance.

5.2 Visualization

In this subsection, we further study the effectiveness of constructing hierarchical news representations and performing multi-grained interest matching. Figure 5 gives visualizations of the multi-grained matching matrices (defined as formula 2) between historical browsed news and candidate news for a user, where $\mathbf{M}_l$ denotes a matching matrix of a news pair at the l-th representation level. We observe that the important matching information captured by the 1st-level matching matrix is mainly lexical relevance. For example, the words "football", "nfl", "playoff", "playoffs" and "quarterbacks" are more correlated and assigned higher matching values in $\mathbf{M}_1$, which may due to their similar co-occurrence information encoded in word embeddings. Differently, higher-level matching matrices have the ability to identify more sophisticated semantic structures and latent long-term dependencies. From Figure 5(b), the interactive areas between the segments "weight loss" in the candidate news and "lost pounds" in the browsed news significantly gain larger matching scores among the 2-nd level semantic representations. In the matching matrix $\mathbf{M}_3$ in Figure 5(c), the subsequences about "trump walks out" are distinguished, since the expressions have correlated meanings. Mean-

while, the results also indicate that our model has the ability to identify important segments of a sentence and ignore the parts with less information, which is helpful to capture user's interested topics or events more accurately.

6 Conclusion and Future Work

In this paper, we propose a new architecture for neural news recommendation based on multi-grained representation and matching. Different from previous work that first integrates user's reading history into a single representation vector and then matches the candidate news representation, our model can capture more fine-grained interest matching signals by performing interactions between each pair of news at multi-level semantic granularities. Extensive experiments on a real-world dataset collected from MSN news show that our model significantly outperforms the state-of-the-art methods. In the future, we will do more tests and surveys on the improvement of business objectives such as user experience, user engagement and service revenue.

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.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Trapit Bansal, Mrinal Das, and Chiranjib Bhattacharyya. 2015. Content driven user profiling for comment-worthy recommendations of news and blog articles. In *RecSys*, pages 195–202.

Andrew P Bradley. 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159.

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.

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

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

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

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

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*, pages 1746–1751.

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

Talia Lavie, Michal Sela, Ilit Oppenheim, Ohad Inbar, and Joachim Meyer. 2010. User attitudes towards news content personalization. *International journal of human-computer studies*, 68(8):483–495.

Lei Li, Li Zheng, Fan Yang, and Tao Li. 2014. Modeling and broadening temporal user interest in personalized news recommendation. *Expert Systems with Applications*, 41(7):3168–3177.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *WWW*, pages 661–670.

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.

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

Zheng Liu, Yu Xing, Fangzhao Wu, Mingxiao An, and Xing Xie. 2019. Hi-fi ark: Deep user representation via high-fidelity archive network. In *IJCAI*, pages 3059–3065.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML*, pages 807–814.

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 D. 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.

Owen Phelan, Kevin McCarthy, and Barry Smyth. 2009. Using twitter to recommend real-time topical news. In *RecSys*, pages 385–388.

Steffen Rendle. 2012. Factorization machines with libfm. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):57.

Jeong-Woo Son, A Kim, Seong-Bae Park, et al. 2013. A location-based news article recommendation with explicit localized semantic analysis. In *SIGIR*, pages 293–302.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

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.

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, Mingxiao An, Tao Qi, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019d. Neural news recommendation with heterogeneous user behavior. In *EMNLP*, pages 4873–4882.

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

Fisher Yu and Vladlen Koltun. 2016. Multi-scale context aggregation by dilated convolutions. In *ICLR*.

Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. 2018. Drn: A deep reinforcement learning framework for news recommendation. In *WWW*, pages 167–176.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *ACL*, pages 1118–1127.