Semantic Gated Network for Efficient News Representation

Xuxiao Bu
Xi'an Jiaotong University
Xi'an, Shaanxi, China
bo951024@stu.xjtu.edu.cn

Bingfeng Li
Department of PCG, Tencent
Beijing, China
bingfengli@tencent.com

Yaxiong Wang
Xi'an Jiaotong University
Xi'an, Shaanxi, China
wangyx15@stu.xjtu.edu.cn

Jihua Zhu
Xi'an Jiaotong University
Xi'an, Shaanxi, China
zhujh@mail.xjtu.edu.cn

Xueming Qian
Xi'an Jiaotong University
Xi'an, Shaanxi, China
qianxm@mail.xjtu.edu.cn

Marco Zhao
Department of PCG, Tencent
Beijing, China
marcozhao@tencent.com

ABSTRACT

Learning an efficient news representation is a fundamental yet important problem for many tasks. Most existing news-relevant methods only take the textual information while abandoning the visual clues from the illustrations. We argue that the textual title and tags together with the visual illustrations form the main force of a piece of news and are more efficient to express the news content. In this paper, we develop a novel framework, namely **S**emantic **G**ated **N**etwork (SGN), to integrate the news title, tags and visual illustrations to obtain an efficient joint textual-visual feature for the news, by which we can directly measure the relevance between two pieces of news. Particularly, we first harvest the tag embeddings by the proposed self-supervised classification model. Besides, news title is fed into a sentence encoder pretrained by two semantically relevant news to learn efficient contextualized word vectors. Then the feature of the news title is extracted based on the learned vectors and we combine it with features of tags to obtain textual feature. Finally, we design a novel mechanism named semantic gate to adaptively fuse the textual feature and the image feature. Extensive experiments on benchmark dataset demonstrate the effectiveness of our approach.

CCS CONCEPTS

• **Information systems → Document representation**.

KEYWORDS

Self-supervised Classification Model, Contextualized Word Vector, Multimodal Fusion, Semantic Gate

ACM Reference Format:
Xuxiao Bu, Bingfeng Li, Yaxiong Wang, Jihua Zhu, Xueming Qian, and Marco Zhao. 2020. Semantic Gated Network for Efficient News Representation. In *Proceedings of the 2020 International Conference on Multimedia Retrieval (ICMR '20), June 8–11, 2020, Dublin, Ireland.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3372278.3390719

1 INTRODUCTION

With the development of the World Wide Web, the common manner of news reading has been moving from the traditional news paper and TV to online news web, such as Google News and Tencent News. Meanwhile, massive articles on the internet can be overwhelming to users. To help users find interesting news and alleviate information overload problem, extensive efforts has been dedicated to news recommendation and retrieval.

Many existing news-relevant tasks design their models based on news body [9, 16, 18]. However, news body is not concise enough and may include irrelevant information, the advertisements for example. Meanwhile, news illustrations contain plenty of information and have auxiliary effect in news representation. For example,if two news possess similar illustrations, they will be semantically relevant with higher probability. Instead of only extracting the news feature from expatiatory body, our work proposes a new framework to get an efficient news representation from the news title, tags and illustrations. To obtain a robust integral feature for news, there are several key problems which need to be tackled.

The first difficulty is that the similarity between textual features does not accurately reflect relevance among news. For instance, two news articles might share a majority of words and their news representation are similar, yet their actual topic could be very different [8].Traditional methods such as one-hot and tf-idf can not solve this problem because they represent sentences according to words' occurrence and frequency. Word vectors in NLP such as NNLM [3] and word2vec [14] also seem not to help because there exists polysemy problem in these methods. Fortunately, contextualized word vectors can alleviate this problem, because this kind of vector contains its context information in the entire input sentence.

Contextualized word vectors are always realized through pretrained model and its classical methods includes CoVe [13] and ELMO [15]. Sentence or document encoders which produce contextual token representations are always pretrained from unlabeled text and then fine-tuned for downstream tasks. In this paper, we propose a new contextualized word vector based on a Bi-directional Long-Short Term Memory (BLSTM) [6] encoder. We first pretrain the encoder from an attentional seq2seq model through two semantically relevant news. Then we can obtain word vector and corresponding state in the bidirectional time directions which compose contextualized word vector of each word. Finally, we combine these contextualized word vectors into an embedded matrix to get the feature of news title.

ICMR '20, June 8–11, 2020, Dublin, Ireland

ACM ISBN 978-1-4503-7087-5/20/06…$15.00
https://doi.org/10.1145/3372278.3390719

*There exist a few methods which can effectively integrate the feature of news illustrations into news representation.*Although image and text both contain rich information, they reside in heterogeneous modalities and this brings about great difficulties to the feature fusion of them. Simply concatenating them together can not relieve challenges in multimodal fusion. The challenges can be caused by these factors: 1) signals in heterogeneous space might not be temporally aligned; 2) it is difficult to exploit supplementary information, which may carry redundant information and noisy information; 3) each modality might exhibit different types and different levels of noise [2]. Consequently, we design a novel feature fusion mechanism named semantic gate to fuse these two kind of features. The mechanism can extract useful information from visual feature by calculating relationship between visual feature and textual feature.

With the proposed framework SGN, we can obtain an efficient news representation that encodes the feature from the news title, tags and illustrations, thus many news-relevant tasks can be conducted accordingly. The contributions of this paper can be summarized as follows:

- We propose a new contextualized word vector based on a sentence encoder and we pretrain the encoder from an attentional seq2seq model through two semantically relative news. Then we take these vectors as the input of textual feature extractor to get the textual feature of news title.
- We design a self-supervised classification model to build the semantic consistence between news tags. Then we extract the feature of news tags based on embeddings trained by the classification model and combine it with feature of news title to get textual feature.
- We propose a novel multimodal feature fusion mechanism named semantic gate which can combine textual feature and visual feature effectively.

2 METHODOLOGY

In this section, we will introduce our framework named SGN for semantical news representation. Given news n_i and n_j from news pool $\mathcal{N} = \{n_1, n_2, ..., n_l\}$, where each news n_i is associated with three data cells, i.e., title T_i, tag set A_i, and illustration I_i, the target of our framework is to learn an efficient representation from the textual title, tags and the visual image. Then we can directly calculate the relevance score s_{ij} between news n_i and news n_j through their representation. The overview of SGN is shown in Fig. 1.

2.1 Pre-training Language Representations

In this subsection, we will introduce the pre-training language representations including news tag and title. The designed self-supervised classification model for tag embedding is presented in subsection 2.1.1 and subsection 2.1.2 introduces the model for contextual word vector.

2.1.1 *Self-supervised Classification Model.*

For tags $A_i = \{a_i^1, ..., a_i^l\}$ and tags $A_j = \{a_j^1, ..., a_j^l\}$ in two semantically relevant news n_i and news n_j, integrating the relations among these co-occurring tags would benefit the news representation. Consequently, building semantic consistence among tags through their embeddings becomes the key problem. We design a self-supervised classification model to predict tags $A_j = \{a_j^1, ..., a_j^l\}$ in news n_j according to tags $A_i = \{a_i^1, ..., a_i^l\}$ in its semantically relevant news n_i.

To simplify the symbol, we still use the a_i^k to indicate tag's one-hot representation. First, we average the embeddings of all tags in A_i as follows

$$h_a^0 = \frac{\sum_{k=1}^{l} V_a \cdot a_i^k}{l}, \tag{1}$$

where $V_a \in \mathbb{R}^{k_a \times |\mathcal{V}|}$ are the mapping matrices for the embedding of tags and $\mathcal{V}$ is the vocabulary of all tags. For better performance, we can add more layers of non-linear transformations into our model and h_a^0 is its initial input

$$h_a^{i+1} = \varphi(W_a^{i+1} \cdot h_a^i + b_a^{i+1}), \tag{2}$$

where $W_a^{i+1} \in \mathbb{R}^{(k_a \times k_a)}$ is the mapping matrix for the variables in the hidden layers, φ is the tanh function and i is the index of a hidden layer. Let h_a donate the final output of the non-linear transformations. Finally, we add a fully-connected layer to map h_a into a tag vocabulary-size vector o_a, which can be described as follows

$$o_a = \varsigma(W_a \cdot h_a + b_a), \tag{3}$$

where $W_a \in \mathbb{R}^{|\mathcal{V}| \times k_a}$, $b_a \in \mathbb{R}^{|\mathcal{V}|}$ and ς is the softmax function.

For convenience, we use the Negative Log-Likelihood (NLL) as the loss function. When the training done, we can obtain the embedding of each tag through the mapping matrix V_a, and they can reflect co-occurrence, namely semantic consistence, among news tags.

2.1.2 *Contextualized Word Vector.*

Inspired by McCann et al. [13], who use a deep LSTM encoder from an attentional seq2seq model [1] trained for MT to contextualize word vectors, we adapt a similar model to build the relationship between titles T_i and T_j in two semantically relevant news n_i and n_j. This is because titles in two semantically relevant news are with similar semantics but different expression ways, which is similar to MT. During pretraining, we take title T_i as the input of encoder and title T_j as the target sequence of decoder. Attention mechanism can identify important information in title T_i for the generation of title T_j and this will benefit news representation. After pretraining the attentional seq2seq model, we transfer what is learned by the encoder to news representation by treating the outputs of the encoder as contextualized word vectors.

First, we can utilize the BLSTM encoder from our attentional seq2seq model to obtain all of the layer representations for each word in input sentence. For word w_i in news title, we can get its word vector $\tilde{w_i}$, its corresponding state $\overrightarrow{h_i}$ in the positive time direction and $\overleftarrow{h_i}$ in the negative time direction. Then we can combine these three vectors as the contextualized word vector of word w_i as follows

$$CoVe(w_i) = [\tilde{w_i}; \overrightarrow{h_i}; \overleftarrow{h_i}]. \tag{4}$$

Our contextualized word vector can represent the meaning of word according to its context, so it can relieve the polysemy problem and get actual topic of articles. Finally, we can combine all

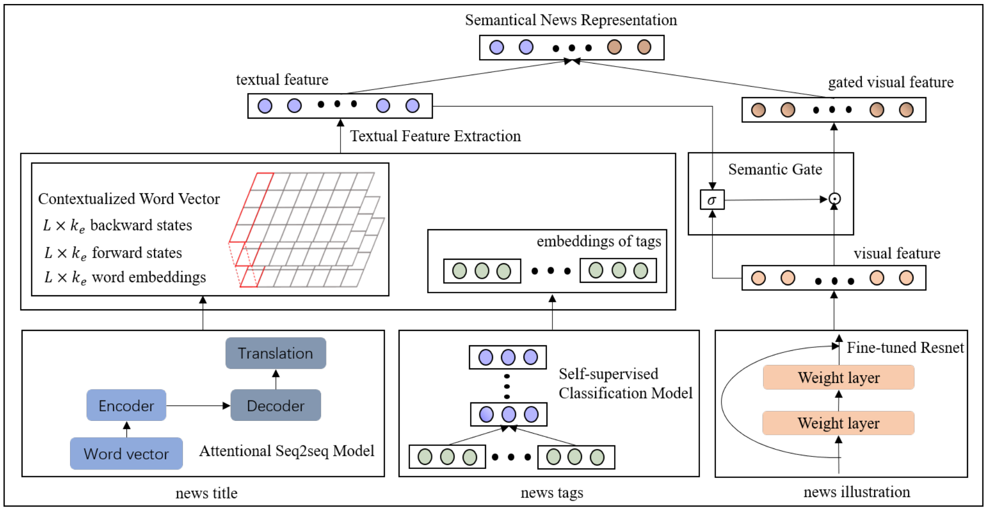

Figure 1: Our proposed framework SGN for efficient news representation.

contextualized word vectors of the input sentence into an embedded matrix $V \in \mathbb{R}^{L \times k_e \times 3}$. L is the length of news title. k_e is the embedding size of words and the hidden size of BLSTM.

2.2 Feature Extraction and Fusion

In this section, feature extraction is presented in subsection 2.2.1. Then, semantic gate is introduced in subsection 2.2.2.

2.2.1 Feature Extraction.

We feed the embedded matrix V into a convolutional layer which consists of k_c neurons. Then, the feature corresponding to each neuron is computed using a max-pooling operation. The output is the concatenation of the output from k_c neurons, denoted by

$$O_t = [o_1, o_2, ..., o_{k_c}]. \tag{5}$$

We treat O_t as the feature of news title. Meanwhile, we can obtain the feature O_a of news tags as we do in self-supervised classification model to obtain h_a. We concatenate O_a and O_t as $[O_t, O_a]$ directly. Then the output $[O_t, O_a]$ is passed to a fully connected layer

$$f_t = \varphi(W_t \cdot [O_t, O_a] + b_t), \tag{6}$$

where weight matrix $W_t \in \mathbb{R}^{k_t \times (k_c + k_a)}$, bias $b_t \in \mathbb{R}^{k_t}$ and φ is activation function softsign φ [17].We treat f_t as the textual feature.

ResNet-101 [7] is one practical and classical model for image feature extraction. We fine-tune it with our dataset according to news category. After fine-tuning, the feature $O_r \in \mathbb{R}^{k_r}$ of an illustration can be obtained from the second last layer. The output O_r is then passed to a fully connected layer

$$f_r = \varphi(W_i \cdot O_r + b_i), \tag{7}$$

where weight matrix $W_i \in \mathbb{R}^{k_i \times k_r}$, bias $b_i \in \mathbb{R}^{k_i}$ and φ is activation function softsign. We treat f_r as the visual feature.

2.2.2 Semantic Gate.

Textual feature and visual feature reside in heterogeneous modalities and this brings about great challenges to the feature fusion of them. Just concatenating them together will introduce a lot of useless information. Motivated by the forget gate in LSTM, we design a novel mechanism named semantic gate for the fusion of them. Useless information can be discarded and complementary information can be distilled in visual feature through semantic gate.

Firstly, we concatenate these two vectors, namely f_t and f_r together as $[f_t, f_r]$. The concatenated result is then passed to a fully connected layer as follows

$$M = \varphi(W_m \cdot [f_t, f_r] + b_m), \tag{8}$$

where weight matrix $W_m \in \mathbb{R}^{k_i \times (k_t + k_i)}$, bias $b_m \in \mathbb{R}^{k_i}$ and φ is a nonlinear activation function sigmoid.

We can see that the dimension of M is the same as f_r, because each value in M means whether value of current position in f_r is noisy or complementary to semantical news representation. Then we multiply the corresponding position elements of M and f_r, namely calculating Hadamard product between M and f_r, and concatenate the gated visual feature with f_t. The final semantical news representation f which fuse textual feature and visual feature can be represented as

$$f = [f_t, M \odot f_r]. \tag{9}$$

2.3 Loss Function

After obtaining feature for news n_i and news n_j, namely f_i and f_j, we can utilize cosine distance to calculate the relevance score $s_{i,j}$. During training process, we employ a hinge-based triplet ranking loss with margin α

$$\mathcal{L}_t = max(0, s_{i,neg} - s_{i,pos} + \alpha) + \lambda_n \parallel \theta \parallel_2^2, \tag{10}$$

where $s_{i,neg}$ is the relevance score between news n_i and its irrelevant news n_{neg} and $s_{i,pos}$ is the opposite. θ is the set of parameters and λ_n is the weight of $\parallel \theta \parallel_2^2$. The hard negative samples are sampled within a mini-batch by the sampling strategy in [11].

Table 1: Performance comparison on efficient news representation based on news tags, title and illustration

Method	Pre	MRR	MAP@5	MAP@10	MAP@20	MNDCG@5	MNDCG@10	MNDCG@20
Word2vec	0.954	0.506	0.700	0.614	0.487	0.663	0.684	0.740
TextCNN	0.965	0.426	0.635	0.567	0.463	0.604	0.642	0.715
BLSTM+TextCNN	0.967	0.426	0.638	0.570	0.465	0.608	0.646	0.719
BERT+TextCNN	0.978	0.539	0.750	0.660	0.525	0.711	0.742	0.804
SGN(w/o illustration)	0.983	0.604	0.802	0.714	0.566	0.770	0.799	0.848
Feature Concatenation	0.974	0.537	0.755	0.670	0.534	0.721	0.751	0.808
SGN (OURS)	**0.985**	**0.636**	**0.822**	**0.733**	**0.579**	**0.790**	**0.817**	**0.865**

3 EXPERIMENTAL ANALYSIS

In this section, experiments is performed on one commercial dataset from Tencent News to demonstrate the effectiveness of SGN. We obtain semantically relevant news pairs as follows. First, we calculate euclidean distance between news pairs based on word2vec which will be discussed in subsection 3.1 and select pairs whose euclidean distance are less than 0.65. Then we choose pairs whose titles have at least one same word. Finally, we discard pairs whose similarity is higher than 75 percent through Locality Sensitive Hashing(LSH). Following these steps, a dataset with 5,930,779 news pairs and 1,083,014 news is built. We employ Precision, Mean Reciprocal Rank (MRR), Average Precision (AP) and NDCG to quantitatively evaluate the performance of the proposed SGN.

3.1 Baselines

To evaluate the performance of our framework SGN, we compare it with the following methods.

Word2vec: Word2vec [14] can output embeddings of words. We train it on news articles to get embeddings of news tags and average the embeddings of news tags to get news representation.

TextCNN: TextCNN [10] is the state-of-the art model for text classification. We concatenate news tags and title as its input and change its loss function to hinge-based triplet ranking loss.

BLSTM + TextCNN: BLSTM is first integrated into text classification in [12]. We concatenate news tags and title as its input. Then we utilize it in the same way with our feature extraction of news title but without parameter initialization with the attentional seq2seq model.

BERT + TextCNN: BERT [5] can output vector of each word in a sentence according to its context. We input these vectors into a TextCNN to get the final textual feature.

Feature Concatenation: Feature concatenation [4] is the simplest way for feature fusion, and it concatenates feature from multimodal. We take it as the comparative method to prove the effectiveness of semantic gate.

3.2 Results

Table 1 shows the comparison results of our framework SGN and baseline methods on semantical news representation. Based on the results, the following conclusions can be drawn.

Firstly, comparing with both TextCNN and BLSTM+TextCNN, word2vec works better. This proves that it is reliable to select news pairs. Meanwhile, comparing with word2vec, SGN(w/o illustration)(our framework without news illustration) is 14.6 percent higher in MNDCG@20. This illustrations it is effective in textual feature extraction.

Secondly, contextualized word vectors based models (SGN(w/o illustration), BLSTM+TextCNN and BERT+TextCNN) perform better than TextCNN not utilizing contextual information. This is because models incorporating valuable information from context can capture the multi-faceted nature of words from news titles, and thereby build a fine-grained semantical representation for news. Consequently, they can relieve polysemous problem in specific context. This is significant for many problems in the field of NLP.

Thirdly, models based on pretraining (BERT+TextCNN, SGN(w/o illustration)) are more effective than these models without pretraining (BLSTM+TextCNN). BERT is pretrained by a large text corpus (like Wikipedia) and the BLSTM in our framework is pretrained by two semantically relevant news through the attentional seq2seq model. Consequently, word vectors from BERT can reflect knowledge from other corpus and word vectors obtained by BLSTM can extract more crucial information in news representation.

Finally, as shown in last two lines in Table 1, SGN which designs a novel feature fusion mechanism named semantic gate works better than model just concatenating two multimodal feature together. This is because news illustration contains useful information, and useless even noise information at the same time. Concatenating it with textual information will make news representation even worse. Our designed semantic gate can distill helpful information, and filter out information which not only is no use but also disturbs the effect. Meanwhile, we can find out that the assistant of news illustration coupling with designed semantic gate structure makes the performances surpass all the competing methods by a large margin. This reveals that the news illustration is helpful for news representation and the overall designed network is efficient.

4 CONCLUSIONS

In this paper, we propose a framework SGN for efficient news representation. It can effectively extract information from news tags, news titles and news illustration and combine them together. As for future work, we will explore the potential advantages of knowledge graph into news representation. This is worth studying because there are plenty of entities in new title and news article. Combining information from knowledge graph will bring useful information and benefit news representation.

5 ACKNOWLEDGMENTS

This work was supported in part by the NSFC under Grant 61772407.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2018), 423–443.

[3] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research* 3, Feb (2003), 1137–1155.

[4] Xingyue Chen, Yunhong Wang, and Qingjie Liu. 2017. Visual and textual sentiment analysis using deep fusion convolutional neural networks. In *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 1557–1561.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A Smith. 2015. Transition-based dependency parsing with stack long short-term memory. *arXiv preprint arXiv:1505.08075* (2015).

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[8] Kevin Joseph and Hui Jiang. 2019. Content based News Recommendation via Shortest Entity Distance over Knowledge Graphs. In *Companion Proceedings of The 2019 World Wide Web Conference*. ACM, 690–699.

[9] Dhruv Khattar, Vaibhav Kumar, Vasudeva Varma, and Manish Gupta. 2018. Weave&Rec: A Word Embedding based 3-D Convolutional Network for News Recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 1855–1858.

[10] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[11] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 201–216.

[12] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101* (2016).

[13] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*. 6294–6305.

[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[15] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* (2018).

[16] Gabriele Sottocornola, Panagiotis Symeonidis, and Markus Zanker. 2018. Session-based News Recommendations. In *Companion Proceedings of the The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 1395–1399.

[17] Joseph Turian, James Bergstra, and Yoshua Bengio. 2009. Quadratic features and deep architectures for chunking. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. Association for Computational Linguistics, 245–248.

[18] Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. DKN: Deep knowledge-aware network for news recommendation. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 1835–1844.