MA-BERT: Learning Representation by Incorporating Multi-Attribute Knowledge in Transformers

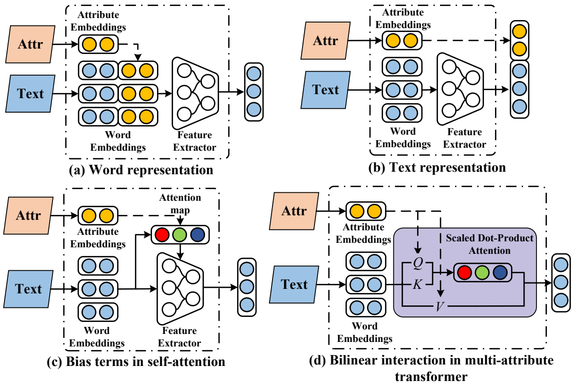
You Zhang[†], **Jin Wang**[†, 1], **Liang-Chih Yu**[‡, 2] and **Xuejie Zhang**[†, 3]
[†]School of Information Science and Engineering, Yunnan University, Yunnan, P.R. China
[‡]Department of Information Management, Yuan Ze University, Taiwan
`Contact:{wangjin`[1]`, xjzhang`[3]`}@ynu.edu.cn, lcyu@saturn.yzu.edu.tw`[2]

Abstract

Incorporating attribute information such as user and product features into deep neural networks has been shown to be useful in sentiment analysis. Previous works typically accomplished this in two ways: concatenating multiple attributes to word/text representation or treating them as a bias to adjust attention distribution. To leverage the advantages of both methods, this paper proposes a multi-attribute BERT (MA-BERT) to incorporate external attribute knowledge. The proposed method has two advantages. First, it applies multi-attribute transformer (MA-Transformer) encoders to incorporate multiple attributes into both input representation and attention distribution. Second, the MA-Transformer is implemented as a universal layer and stacked on a BERT-based model such that it can be initialized from a pre-trained checkpoint and fine-tuned for the downstream applications without extra pre-training costs. Experiments on three benchmark datasets show that the proposed method outperformed pre-trained BERT models and other methods incorporating external attribute knowledge.

1 Introduction

To learn a distributed text representation for sentiment classification (Pang and Lee, 2008; Liu, 2012), conventional deep neural networks, such as convolutional neural networks (CNN) (Kim, 2014) and long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997), and common integration technics, such as self-attention mechanisms (Vaswani et al., 2017; Chaudhari et al., 2019) and dynamic routing algorithms (Gong et al., 2018; Sabour et al., 2017), are usually applied to compose the vectors of constituent words. To further enhance the performance, pre-trained models (PTMs), such as BERT (Devlin et al., 2019), ALBERT (Lan et al., 2019), RoBERTa (Liu et al., 2019), and XLM-RoBERTa (Conneau et al., 2019) can be fine-tuned and transferred for sentiment analysis tasks. Practically, PTMs were first fed a large amount of unannotated data, and trained using a masked language model or next sentence prediction to learn the usage of various words and how the language is written in general. Then, the models are transferred to another task to be fed another smaller task-specific dataset.



(a) Word representation

(b) Text representation

(c) Bias terms in self-attention

(d) Bilinear interaction in multi-attribute transformer

Figure 1: Different strategies to incorporate external attribute knowledge into deep neural networks.

The abovementioned methods only use features from plain texts. Incorporating attribute information such as users and products can improve sentiment analysis task performance. Previous works typically incorporated such external knowledge by concatenating these attributes into word and text representations (Tang et al., 2015), as shown in Figs. 1(a) and (b). Such methods are often introduced in shallow models to attach attribute information to modify the representation of either words or texts. However, this may lack interaction between attributes and the text since it equally aligns words to attribute features, thus the model is unable to emphasize important tokens. Several works have used attribute features as a bias term in self-attention mechanisms to model meaningful rela-

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 2338–2343
August 1–6, 2021. ©2021 Association for Computational Linguistics

tions between words and attributes (Wu et al., 2018; Chen et al., 2016b; Dong et al., 2017; Dou, 2017), as shown in Fig. 1(c). By using the $softmax$ function for normalization to calculate the attention score, the incorporated attribute features only impact the allocation of the attention weights. As a result, the representation of input words has not been updated, and the information of these attributes will be lost. For example, depending on individual preferences for *chili*, readers may focus on reviews talking about *spicy*, but only those who like *chili* would consider such review recommendations useful. However, current self-attention models that learn text representations by adjusting the weights of *spicy* may still produce the same word representation of *spicy* for different persons, leading to confusion in distinguishing people who like *chili* or not.

To address the above problems, this study proposes a multi-attribute BERT (MA-BERT) model which applies multi-attribute transformer (MA-Transformer) encoders to incorporate external attribute knowledge. Different from being incorporated into the attention mechanism as bias terms, multiple attributes can be injected into both attention maps and input token representations using bilinear interaction, as shown in Fig. 1(d). In addition, the MA-Transformer is implemented as a universal layer and stacked on a BERT-based model such that it can be initialized from a pre-training checkpoint and fine-tuned for downstream tasks without extra pre-training costs. Experiments are conducted on three benchmark datasets (IMDB, Yelp-2013, and Yelp-2014) for sentiment polarity classification. The results show that the proposed MA-BERT model outperformed pre-trained BERT models and other methods incorporating external attribute knowledge.

The remainder of this paper is organized as follows. Section 2 provides a detailed description of the proposed methods. The empirical experiments are reported with analysis in Section 3. Conclusions are finally drawn in Section 4.

2 Multi-Attribute BERT Model

Fig. 2 shows an overview of the MA-BERT model. It mainly consists of two parts, including a BERT-based PTM model and several MA-Transformer encoders as extra layers stacked on BERT. Both components are described in detail below.

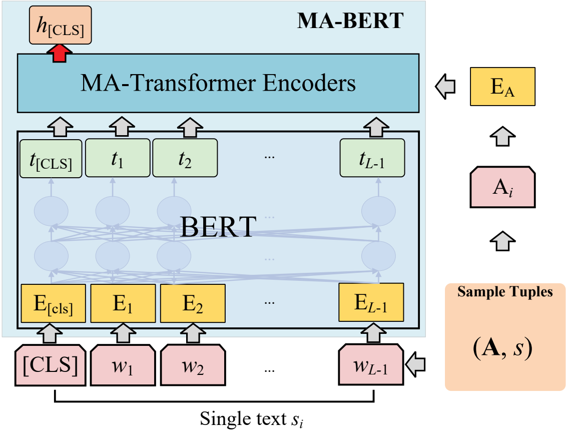


Figure 2: Overall architecture of the MA-BERT model.

2.1 BERT Encoder

By applying a word piece tokenizer (Wu et al., 2016), the input text can be denoted as a sequence of tokens, i.e., $s = \{w_0, w_1, w_2, \ldots, w_{L-1}\}$, where L is the length of the text and $w_0 =$ [CLS] is a special classification token. Moreover, its corresponding attributes are denoted as $\mathbf{A} = \{a_1, a_2, \ldots, a_M\}$, where M is the number of attributes in the text. Thus, the i-th input sample can be denoted as a tuple, i.e., $(\mathbf{A}_i, s_i)$.

To learn the hidden representation, the pre-trained language model BERT (Devlin et al., 2019) was used, achieving impressive performance for various natural language processing (NLP) tasks. We then fed the token sequence into the BERT model to obtain the representation, denoted as,

$$T = [t_0, ..., t_{L-1}] = f_{\text{BERT}}([w_0, ..., w_{L-1}]; \theta_{\text{BERT}}) \quad (1)$$

where $T \in \mathbb{R}^{L \times d_t}$ is the output representation of all tokens; θ_{BERT} is the trainable parameters of BERT, which is initialized from a pretrained checkpoint and then fine-tuned during the model training; d_t=768 is the dimensionality of the output representation.

According to Wu et al. (2018) and Wang et al. (2017), all the attributes are mapped to attribute embeddings $E_{\text{A}} = [E_{\text{A},1}, E_{\text{A},2}, \ldots, E_{\text{A},M}] \in \mathbb{R}^{M \times d_E}$, which are randomly initialized and updated in the following training phase.

Multi-Attribute Attention. To incorporate multiple attributes into the MA-Transformer, we introduce multi-attribute attention (MAA), which is expressed as,

$$Y = \text{MAA}(T, E_{\text{A}}) = [U_1, \ldots, U_M]\, W_o \quad (2)$$

$$U_m = \text{Att}(T, E_{\text{A},m}) = softmax\left(\frac{Q_m {K_m}^\top}{\sqrt{d}}\right) V_m \quad (3)$$

where U_m is the attention from m-th attribute; $W_o \in R^{(M \cdot d) \times d_t}$ is the output linear projection and d denotes the dimensionality of Q, K and V; Q, K and V are matrices that package the queries, keys and values, which are defined as,

$$Q_m = T \cdot W_{q,m} \odot E_{\mathrm{A},m} \tag{4}$$

$$K_m = T \cdot W_{k,m} \odot E_{\mathrm{A},m} \tag{5}$$

$$V_m = T \cdot W_{v,m} \odot E_{\mathrm{A},m} \tag{6}$$

where Q_m, K_m and $V_m \in \mathbb{R}^{L \times d_E}$ are bilinear transformations (Huang et al., 2019) applied on the input representation T and attribute representation $E_{\mathrm{A,m}}$. $W_{q,m}$, $W_{k,m}$ and $W_{v,m} \in \mathbb{R}^{d_t \times d_E}$ are weight matrices for query, key and value projections, and $\cdot$ and $\odot$ respectively denote the inner and the Hadamard product.

Similar to Vaswani et al. (2017), we also introduced multi-head mechanism for MA-Transformer, denoted as,

$$U_m = \bigoplus_{k=1}^{K} \mathrm{Att}(T, E^k_{\mathrm{A},m}) \in \mathbb{R}^{L \times (K \cdot d_E)} \tag{7}$$

where K is the number of heads for each attribute and $\oplus$ denotes the concatenation operator; $E^k_{\mathrm{A},m} \in \mathbb{R}^{d_E}$ is the m-th attribute representation in the k-th head, and its dimensionality should be ensured that $d_E = d_t/K$. Given that different heads can capture different relation types along with text representations, different parameters are considered for different heads.

2.2 MA-Transformer

Taking the representation of both text T and attribute $\mathbf{A}$ as input, an MA-Transformer encoder then processes the same as a standard transformer encoder (Vaswani et al., 2017) to generate $Y \in \mathbb{R}^{L \times d_t}$. Then, Y is connected by a normalization layer and a residual layer from the input representation T. The intermediate output is then passed to a two-layered feed-forward network with a rectified linear unit (ReLU) activate function. Similarly, residual and normalization layers are connected to generate the final output which is taken as the input for the next encoder.

By stacking several MA-Transformer encoders on the BERT model, the MA-BERT model generates a review representation $h_{[\mathrm{CLS}]}$ consistent with the special token [CLS]. Then, a classifier comprised of a linear projection and a $softmax$ activation (with the dimension identical to the number of classes) is used for classification.

3 Comparative Experiments

Datasets. Following the experimental settings used in Tang et al. (2015), the proposed MA-BERT model is evaluated using three benchmark datasets [1] (IMDB, Yelp-2013, and Yelp-2014). The evaluation metrics include accuracy (Acc.) and root mean squared error ($RMSE$). Higher Acc. and lower $RMSE$ scores indicate higher performance.

Implementation Details. The baseline methods can be divided into three groups. The first group includes the methods without user and product information such as **CNN** (Kim, 2014), **BiLSTM** (Hochreiter and Schmidhuber, 1997), neural sentiment classification (**NSC**) (Chen et al., 2016a) and its variant with a local attention mechanism (**NSC+LA**). For the BERT-based methods, the uncased-base-**BERT** model consisting of 12 layers of transformer encoders was implemented for comparison. **ToBERT** (Pappagari et al., 2019) was trained non-end2end using a word-to-segment strategy in a two-stage way.

The second group includes existing methods incorporating user and product information such as NSC with user (U) and product (P) information incorporated into an attention (A) mechanism (**NSC+UPA**), user product neural network (**UPNN**) (Tang et al., 2015), hierarchical model with separate user attention and product attention (**HUAPA**) (Wu et al., 2018), and the chunk-wise importance matrix model (**CHIM**) (Amplayo, 2019).

The third group includes a set of BERT-based methods incorporating user and product information using different strategies, presented in Figs. 1(a)-(c). In detail, an uncased-base-BERT model first extracted fixed feature vectors from texts. Then, the **BERT Concat (word)** model incorporates attribute features into each word vector and stacks another 6 transformer encoders as the feature extractor. Similarly, the **BERT Concat (text)** incorporates attribute features into the representation of the special token [CLS] for the classification. Finally, the **BERT Attention (bias)** applied 6 more MA-Transformers which only inject attributes into Q and K to calculate attention score instead of V in Eq. (6).

The proposed **MA-BERT** models applied 6 MA-Transformer encoders to incorporate user and product attributes, and stacking over the BERT model.

[1] http://ir.hit.edu.cn/~dytang/paper/acl2015/dataset.7z

Models	IMDB		Yelp-2013		Yelp-2014	
	Acc. (%)	$RMSE$	Acc. (%)	$RMSE$	Acc. (%)	$RMSE$
without user and product information						
CNN (UPNN w/o UP)	40.5	1.629	57.7	0.812	58.5	0.808
BiLSTM	43.3	1.494	58.4	0.764	59.2	0.733
NSC	44.3	1.465	62.7	0.701	63.7	0.686
NSC+LA	48.7	1.381	63.1	0.706	63.0	0.715
BERT	51.8	1.191	67.7	0.627	67.2	0.630
ToBERT	50.8	1.194	66.7	0.626	66.9	0.620
with user and product information						
UPNN	43.5	1.602	59.6	0.803	60.8	0.764
NSC+UPA	53.3	1.281	65.0	0.692	66.7	0.654
HUAPA	55.0	1.185	68.3	0.628	68.6	0.626
CHIM$_{embedding}$	56.4	1.161	67.8	0.646	69.2	0.629
BERT Concat (word)	56.8	1.106	69.9	0.602	70.9	0.582
BERT Concat (text)	54.6	1.168	68.5	0.616	71.0	0.590
BERT Attention (bias)	52.5	1.177	68.0	0.635	67.6	0.617
MA-BERT	**57.3**	**1.042**	**70.3**	**0.588**	**71.4**	**0.573**

Table 1: Comparative results of different methods for sentiment classification. The **boldface** figures indicate the best results among all methods and underscored figures represent the best performance for each group of methods. All results are averaged over five runs.

Each attribute is initialized in a uniform distribution $U \sim (-0.25, 0.25)$ with the dimension of 768 (d_t) and head number of 12 (K). Thus, the dimension of each head (d_E) is set to 64. All other hyper-parameters in MA-Transformer encoders are identical with BERT-transformer encoders due to their isomorphic structure. For all models, the AdamW (Loshchilov and Hutter, 2017) optimizer was used with a base learning rate of 2e-5 in a warmup linear schedule. Early stopping (Prechelt, 1998) strategy with a patience of 3 epochs was also applied to avoid overfitting. The code for this paper is available at: `https://github.com/yoyo-yun/MA-Bert`.

Comparative Results and Discussion. Table 1 shows the comparative results of different methods for sentiment ordinal classification. For models without user and product attributes, BiLSTM outperforms CNN (UPNN w/o UP), due to its ability to encode text. Furthermore, both NSC and NSC+LA outperformed BiLSTM mainly because of its hierarchical structure.

Incorporating the user and product attributes improved performance. For example, UPNN achieved a better result than its variant without user and product attributes, i.e., CNN (UPNN w/o UP). In addition, both NSC+UPA and HUAPA introduced the user and product information as a bias to guide the hierarchical attention, and thus outperformed NSC and NSC+LA.

The proposed MA-BERT achieved the best performance on all datasets. Compared with baselines without user and product attributes, the MA-BERT can leverage implicit attribute features to boost performance. MA-BERT outperformed methods already incorporating user and product attributes (i.e., NSC+UPA, HUAPA and CHIM$_{embedding}$) because the proposed model can incorporate attribute knowledge to both the attention map and input representation.

The BERT and ToBERT models achieved improvement on all datasets against the conventional models, due to the large knowledge migration from pre-training. Unfortunately, a lack of implicit extra features resulted in performance lower than that of the proposed MA-BERT model. MA-BERT also outperformed BERT Concat (word), BERT Concat (text) and BERT Attention (bias), indicating that the proposed MA-Transformer architecture can improve existing incorporation strategies. Furthermore, the proposed MA-BERT could be initialized from the pre-trained checkpoint of BERT, thus making full use of the parameter settings without bringing additional costs for pre-training.

4 Conclusion

This paper proposes a MA-BERT model capable of incorporating multiple attributes into BERT-based PTMs for learning attribute-specific text representation. Different from existing attention models, the MA-Transformer can inject external knowledge to both attention maps and the input representation.Additionally, the proposed model could be extended to other tasks by using the MA-Transformer encoder as a universal layer and stacking it on a BERT-based model. Future work will attempt to incorporate such or similar multiple attributes into PTMs in the pre-training phases.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61966038 and 61762091, and by the Ministry of Science and Technology, Taiwan, ROC, under Grant No. MOST107-2628-E-155-002-MY3. The authors would like to thank the anonymous reviewers for their constructive comments.

References

Reinald Kim Amplayo. 2019. Rethinking Attribute Representation and Injection for Sentiment Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5601–5612, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sneha Chaudhari, Gungor Polatkan, Rohan Ramanath, and Varun Mithal. 2019. An Attentive Survey of Attention Models. *arXiv preprint arXiv:1904.02874*.

Huimin Chen, Maosong Sun, Cunchao Tu, Yankai Lin, and Zhiyuan Liu. 2016a. Neural Sentiment Classification with User and Product Attention. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2016)*, pages 1650–1659.

Tao Chen, Ruifeng Xu, Yulan He, Yunqing Xia, and Xuan Wang. 2016b. Learning User and Product Distributed Representations Using a Sequence Model for Sentiment Analysis. *IEEE Computational Intelligence Magazine*, 11(3):34–44.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv preprint arXiv:1911.02116*.

J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, pages 4171–4186.

Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. 2017. Learning to generate product reviews from attributes. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2017)*, volume 1, pages 623–632.

Zi-Yi Dou. 2017. Capturing User and Product Information for Document Level Sentiment Analysis with Deep Memory Network. In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP-2017)*, pages 521–526.

Jingjing Gong, Xipeng Qiu, Shaojing Wang, and Xuanjing Huang. 2018. Information Aggregation via Dynamic Routing for Sequence Encoding. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2742–2752.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Tongwen Huang, Zhiqi Zhang, and Junlin Zhang. 2019. Fibinet: Combining feature importance and bilinear feature interaction for click-through rate prediction. In *13th ACM Conference on Recommender Systems (RecSys-2019)*, pages 169–177.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP-2014)*, pages 1746–1751, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv preprint arXiv:1909.11942*.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101*.

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis: Foundations and Trends in Information Retrieval. *Foundations and Trends® in Information Retrieval*, 2(1-2):1–135.

Raghavendra Pappagari, Piotr Żelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical Transformers for Long Document Classification. *arXiv preprint arXiv:1910.10781*.

Lutz Prechelt. 1998. Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks*, 11(4):761–767.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic Routing Between Capsules. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS-2017)*, pages 3859–3869.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Learning Semantic Representations of Users and Products for Document Level Sentiment Classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-2015)*, pages 1014–1023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in neural information processing systems(nips-2017)*, pages 5598–6008.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *the Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI-2017)*, pages 3316–3322.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144*.

Zhen Wu, Xin Yu Dai, Cunyan Yin, Shujian Huang, and Jiajun Chen. 2018. Improving review representations with user attention and product attention for sentiment classification. In *Proceedings of The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 5989–5996, New Orleans, Louisiana, USA.