Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 9, No. 2, 2021, 141-151.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Convolutional Neural Network Equipped with Attention Mechanism and Transfer Learning for Enhancing Performance of Sentiment Analysis

Hossein Sadr[1], Mir Mohsen Pedram [2*] and Mohammad Teshnehlab[3]

1. Department of Computer Engineering, Rasht Branch, Islamic Azad University, Rasht, Iran.
2. Department of Electrical and Computer Engineering Faculty of Engineering, Kharazmi University, Tehran, Iran.
3. Industrial Control Center of Excellence, Faculty of Electrical and Computer Engineering, K. N. Toosi University, Tehran, Iran.

Article Info

Article History:
Received 29 May 2020
Revised 11 November 2020
Accepted 21 January 2021

DOI:10.22044/jadm.2021.9618.2100

Keywords:
Sentiment Analysis, Deep Learning, Convolutional Neural Network, Attention Mechanism, Transfer Learning

*Corresponding author: *pedram@khu.ac.ir(M.M. Pedram).*

Abstract

With the rapid development of textual information on the web, sentiment analysis is changing to an essential analytic tool rather than an academic endeavor, and numerous studies have been carried out in recent years to address this issue. By the emergence of deep learning, deep neural networks have attracted a lot of attention and have become the mainstream in this field. Despite the remarkable success of deep learning models for sentiment analysis of text, they are in the early steps of development and their potential is yet to be fully explored. Convolutional neural network is one of the deep learning methods that has been surpassed for sentiment analysis but is confronted with some limitations. Firstly, convolutional neural network requires a large number of training data. Secondly, it assumes that all words in a sentence have an equal contribution to the polarity of a sentence. To fill these lacunas, a convolutional neural network equipped with the attention mechanism is proposed in this paper which not only takes advantage of the attention mechanism but also utilizes transfer learning to boost the performance of sentiment analysis. According to the empirical results, our proposed model achieved comparable or even better classification accuracy than the state-of-the-art methods.

1. Introduction

By the expeditious growth of information, finding and sharing information on the web about each particular topic is changing to an easy task. On the other hand, we are nowadays faced with a large number of textual data that is generated by users while they express their opinion about various entities. Considering the fact that analyzing such a large amount of unstructured data is not feasible, there have been great endeavors to propose an effective method to automatically collect and process them. The automatic process of text analysis with the aim of extracting subjective information existing in the text is known as sentiment analysis [1].

Sentiment analysis is known as a classification problem from the machine learning perspective which classifies textual data into positive or negative categories. Although machine learning based models obtained considerable results in this field, the emergence of deep learning was the birth of a revolution and attracted many researchers to use them instead of machine learning models to perform classification. It is worth mentioning that although deep learning methods have achieved significant progress in this regard [2, 3], their classification performance is still unsatisfactory and their potential is yet to be fully explored [2, 4].

Convolutional neural network is recognized as one of the most important deep learning models that has been extensively utilized for sentiment classification and obtained remarkable results. In spite of the significant efficiency of convolutional neural networks for that task sentiment analysis, they are still suffering from some difficulties [5, 6]. Firstly, convolutional neural networks can only

clarify the polarity of the document and are not able to provide a deep understanding of the text such as identifying the main word that has a remarkable influence on the polarity classification [4, 7]. In other words, despite human brains, they are not able to put on emphasis on the salient parts of a text which leads to a reduction in their effectiveness [8, 9]. Secondly, convolutional neural networks tend to be particularly data-hungry and require a large number of training data to accurately train the model as well as they need many parameters to be tuned preciously [10].

To fill these lacunas, convolutional neural network equipped with attention mechanism and transfer learning is proposed in this paper that tries to take advantage of both attention mechanism and transfer learning to improve the efficiency of sentiment classification. The reason behind using the attention mechanism is to simulate the attention mechanism found in human brains as well as emphasizing more important words and neglecting the less important ones [8, 11, 12]. The intuition behind using transfer learning is to overcome the large training data requirement of convolutional neural networks by training the model in large size datasets as a source domain and then transfer it to the target domain which can lead to performance enhancement [10]. Moreover, while a downside of convolutional neural networks is that they require practitioners to specify the exact model architecture to be used and set the accompanying hyper-parameters, we conducted a series of experiment to explore the sensitivity of the proposed model and achieve the optimal values of hyper-parameters to boost the classification performance

The rest of this paper is classified as follows. Review of literature is briefly discussed in Section 2. The proposed model and its details are mentioned in Section 3. Datasets, model configuration, training, and empirical results are completely described in Section 4. Section 5 includes conclusions and future research directions.

2. Literature Review

Machine learning models have been commonly utilized for sentiment analysis. However, with the emergence of deep learning, deep neural networks have attracted a lot of attention and become mainstream. While the focus of this paper is on presenting an attention-based deep learning method with transfer learning, studies in the field of deep learning, attention mechanism, and transfer learning focusing on sentiment analysis, are extensively explored in this section.

Deep learning based sentiment analysis. With the success of deep learning, a great variety of deep neural networks have been developed that are able to tackle the lack of interaction between the target entity and its context. To this end, various methods like Convolutional Neural Network (CNN), Recursive and Recurrent Neural Network (RNN), Recursive Auto Encode (RAE), and Deep Belief Network (DBN) have been extensively applied for sentiment classification [3].

In this regard, tree structure gated recurrent neural network was proposed by Kuta et al. [13] that was inspired by adaptation of gated recurrent unit to recursive model and tree structure LSTM. Long short term memory network that was integrated with several complex units was also used by Tai et al. [14] for sentiment analysis. They also conducted more experiments on two layers of bidirectional LSTM and achieved significant results.

Following a similar line of research, an extensive set of experiments were performed by Kim et al. [15] on one-layer convolutional neural network. They trained their models on pre-trained vectors derived from Word2Vec embedding model. They also employed multi-channel representation and various filer sizes and achieved comparable results. Against modeling sentences at the word level, A character-level CNN for text classification was proposed by Zhang et al. [16] which presented remarkable improvement in classification accuracy. Moreover, Yin and Schutze presented a multichannel variable size CNN that employed combinations of various word embedding techniques as input [17]. Kalchbrenner et al. [18] proposed a dynamic CNN that utilized dynamic k-max pooling. While their model was able to handle input sentences of variable lengths, it could efficiently capture short and long-term dependencies. MV-RNN, a method in the family of recursive neural networks, was proposed by Socher et al. [19] that leveraged both matrix and vector aiming to represent words and phrases in the tree structure. Recursive Neural Tensor Network (RNTN) is another network in this field proposed by Socher et al. [20] where the tensor-based compositional matrix was used instead of matrix representation for all nodes in the tree structure. Sadr et. al [21] proposed a method that tried to take advantage of both convolutional and recursive neural networks for sentiment classification. In the following, they also utilized multi-view learning to make use of intermediate

features extracted from convolutional and recursive neural networks to perform classification and obtained considerable results [22].

In spite of the fact that deep neural networks have achieved significant results in the field of sentiment analysis, they are still confronting with some limitations and there is a strong need for progress in this area.

Attention-based Network for Sentiment Analysis. In spite of the fact that deep neural networks have achieved significant results in the field of sentiment analysis, their performance is still unsatisfactory [2, 3]. One of their general pitfalls is that they consider all words in the sentences equally and are not able to focus on salient parts of the text [7]. To fill this lacuna, the attention mechanism has been recently adopted in many tasks of natural language processing especially sentiment analysis due to its strength in providing an effective interpretation of the text. In fact, the attention mechanism was inspired by the visual attention mechanism found in humans which tries to focus on the more important part of the text rather than encoding the full sentence. In this regard, Yang et al. [7] modified RNN by adding weight that played the attention role for the aim of text classification. Wang et al. [23] also proposed an attention-based LSTM network that could focus on various parts of the sentences. A domain attention model for multi-domain sentiment classification was proposed by Yuag et.al [24]. Their proposed model used domain representation as attention to select the most appropriate domain-related features in each domain. It must be taken into consideration that despite promising results of applying attention mechanism on deep neural networks, only a few studies have been conducted in the field of sentiment analysis and its impact on the convolutional neural networks has been rarely explored.

Transfer Learning. Another challenge that deep neural networks are commonly confronted with refers to the lack of training data. In fact, deep neural networks require a large number of training data to be able to accurately train the model and as the number of data is increased, their performance is also enhanced [25]. The lack of available labeled training data has yielded to the emergence of a new concept known as transfer learning. Transfer learning is used when the training set is not large enough to efficiently train the model [26]. Therefore, the model is trained on a large dataset known as the source domain and then is transferred to the target domain which can significantly enhance the performance. Although transfer learning has been considerably employed for the task of image processing, its application in natural language processing, especially sentiment analysis, is still limited [10].

In this regard, Krizhevsky and Lee [27] presented the efficiency of transferring low-level neural layers in different tasks. In a similar study [28], the impact of transferring high-level layers in a deep neural network from the source dataset to a smaller size target dataset was investigated. However, it is worth mentioning that the effect of transfer learning for the task of sentiment analysis has been rarely explored [8].

Considering the conducted studies and the mentioned challenges, we decided to propose a convolutional neural network equipped with attention layer coupling with transfer learning that is able to make use of both attention mechanism and transfer learning. Despite previous studies, the proposed model applies an attention mechanism after the convolutional layer to extract informative words existing in the sentences. Then, the proposed model employs transfer learning that transfers knowledge from a source domain to various but related target domains aiming to enhance the performance.

3. Methodology

Convolutional neural networks are commonly known as good candidates for natural language processing. They can not only control the length of dependencies but also enables nearby input element to extract at lower layers while distant elements interact at higher layers. Therefore, hierarchical abstract representation of input text can be produced using multiple convolutional layers. In the following, the pooling layer is used to extract the most significant features. However, the pooling layer leads to the loss of local features that may contain valuable information. In this respect, the proposed model uses an attention layer before the pooling layer to specify the critical features as well as suppressing the effect of unimportant features and helping the pooling layer find the genuinely crucial features considering the context. Moreover, while increasing the number of training data generally leads to convolutional neural network performance enhancement, we also decided to explore the effect of transfer learning on the proposed method.

Briefly, this section describes the application of attention mechanism and transfer learning to construct the proposed method.

3.1. Convolutional Neural Network Equipped with Attention Layer

Our proposed model contains four layers where the attention layer is placed before the pooling layer to extract more significant information. In our proposed model, word vectors of input sentences are first extracted and then joined to form the initial input matrix. Thereafter, the convolutional filters are applied to the input matrix and feature maps are extracted. As the training is completed, feature maps extracted from similar filter sizes are merged and fed to the attention layer as a new matrix in the third layer. In the following, by extracting the informative words and assigning a higher weight to them using the attention mechanism and aggregating their representation to the previous features extracted by the convolutional neural network, new sentence vectors are formed. Finally, new word vectors are entered into a fully connected network and classification is performed. More detailed mathematical deduction about each layer is provided in the following.

3.1.1. Representation Layer

The convolutional neural network requires a sentence matrix as an input, where each row represents a word vector. If the dimensionality of the word vector is d and the length of a given sentence is s, the dimensionality of the sentence matrix would be $s \times d$ where padding is set to zero before the first word and after the last word in the sentence. Setting the padding to zero makes the number of times that each word is included in receptive filed during the convolution the same without considering the word position in the sentence. As a result, the sentence matrix is denoted by $A \in R^{s \times d}$ In this paper, various word embedding techniques including Word2Vec [29], Glove [30], and FastText [31] were employed to form the input matrix.

3.1.2. Convolutional Operation

Convolutional operation must be applied to the sentence matrix to produce the new features. According to the fact that the sequential structure of a sentence has an important effect in specifying its meaning, it is sensible to choose filter width equal to the dimensionality of word vectors (d). In this regard, only the height of filters (h), known as region size, can be varied.

Considering $A \in R^{s \times d}$ s a sentence matrix, convolution filter $H \in R^{h \times d}$ is applied to A to produce its sub-matrix as a new feature $A[i:j]$. As the convolution operation is applied repeatedly on the matrix of A, $O \in R^{s-h+1 \times d}$ as the output sequence is achieved (Eq. 1).

$$O_i = w.A[i:i+h-1] \quad (1)$$

Here, $i = 1,..., s-h+1$ and $\cdot$ is the dot product between two matrices of the convolution filter and input submatrix. Bias term $b \in R$ and an activation function are also added to each O_i. Finally, feature maps $C \in R^{s-h+1}$ are generated (Eq. 2).

$$C_i = f(O_i + b) \quad (2)$$

3.1.3. Attention Mechanism

Considering the fact that all words in a sentence do not contribute equally to represent the meaning of a sentence and pooling layer in convolutional neural network yields to loss of local features, there is a need for an attention mechanism to emphasize on such words that have more impact on the meaning of the sentences considering the context and interaction of words.

To perform the attention mechanism on the convolutional layer, feature maps that are extracted from the same filter size are aggregated to form a new matrix. Suppose that in the convolution layer, M different region sizes are considered and for each of them m different filters are employed. Therefore, after applying $H_{ij} \in R^{h_i \times d}$ filters on sentence matrix A where $i = 1,2,..., M$ and $j = 1,2,..., m$, $M \times m$ feature map is obtained. By concatenating feature maps extracted from the same filter size, a new sentence matrix $X_i \in R^{n \times m}$ (Eq. 3) is obtained. Where n is the number of words and each element of this matrix represents the feature extracted from the input using filters of the same size.

$$X_i = \begin{bmatrix} x_{1,1} & \text{K} & x_{1.m} \\ \text{M} & \text{O} & \text{M} \\ x_{n-c_i+1,1} & \Lambda & x_{n-c_i+1,m} \end{bmatrix} \quad (3)$$

The objective of the attention mechanism is to assign a specific weight to each row for extracting informative parts of the sentence. For this aim, firstly, the new word matrix X_i is fed through a single layer perceptron using $w \in R^{m \times d}$ and $U_i \in R^{n-h_i+1 \times d}$ as a hidden representation of X_i is obtained (Eq. 4).

$$U_i = \tanh(X_i W + b) \quad (4)$$

In the following, the importance of each word is measured as the similarity of U_i with a context vector $u \in R^{d \times 1}$ and to achieve the normalized importance weight $a_i \in R^{n-h_i+1 \times 1}$, Softmax function

is used (Eq. 5). Notably, the context vector u can be considered as a high-level representation to specify the informative words and it is like the mechanism that is used in the memory networks [32, 33].

$$a_i = soft\max(U_i u) \quad (5)$$

Notably, u is set to zero at the beginning to consider the same weight for various rows in the matrix of X_i and it is learned along the training process. After that $\hat{X}_i$ (a new representation of X_i is computed by multiplying each element of a_i to its corresponding row in X_i matrix ($o$is the element-wise product) (Eq. 6).

$$\overline{X}_i = a_i \, o X_i \quad (6)$$

Generally, $\hat{X}_i$ is a new representation of X_i while the attention mechanism is applied to it in order to specify the informative words.

The overall process of the attention layer is schematically depicted in Figure 1. As it can be clearly seen, after merging feature maps extracted from the same filter sizes, X_i matrix is created. Then, by applying a single layer perceptron, a new representation of X_i, known as U_i is created. In the following, the normalized importance weight a_i, indicating the importance of each word, is computed as the similarity between U_i and content vector u which is a hyper-parameter and is tuned during the training process. Finally, $\hat{X}_i$ a new representation of X_i, is achieved by multiplying each element of a_i to its corresponding row in X_i. Generally, applying the attention mechanism leads to informative words extraction by assigning more
weight to them.

3.1.4. Pooling and Classification

While various feature maps according to different filter sizes are generated, a pooling function is required to induce fixed size vectors. Various strategies, such as average pooling, minimum pooling, and maximum pooling can be used for this aim. The idea behind the pooling layer is to capture the most important feature from each feature map and reduce dimensionality. Generated features from the pooling layer from each filter are concatenated into a feature vector o_i .The feature vector is then passed to a fully connected Softmax layer to specify the final classification.

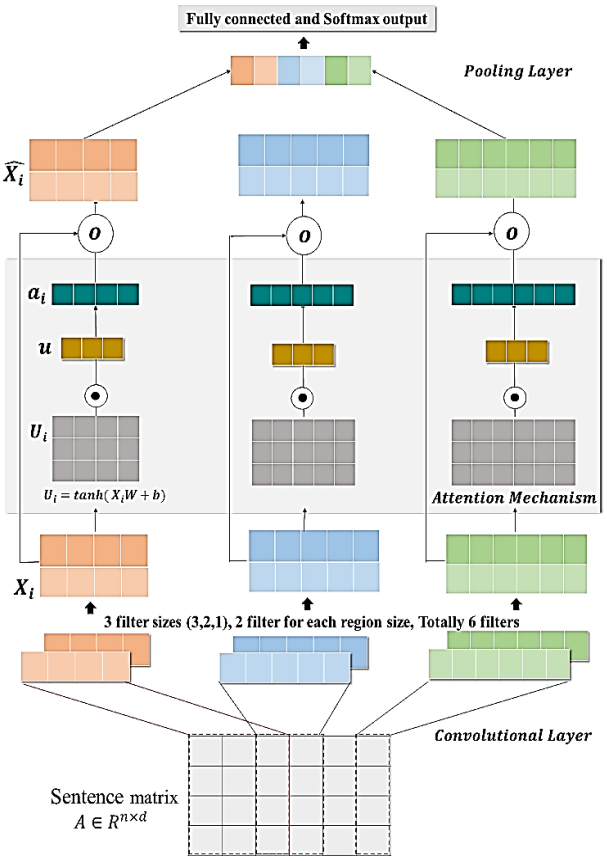

Figure 1. Overall process of convolutional neural network with attention layer

In other words, Softmax determines the probability distribution over all sentiment categories and is calculated as follows (Eq. 7).

$$P_i = \frac{\exp(o_i)}{\sum_{j=1}^{c} \exp(o_i)} \quad (7)$$

To clarify the difference between the real sentiment distribution $\hat{P}_i(C)$ and the distribution achieved from the model $P_i(C)$, cross-entropy as the loss function is employed (Eq. 8).

$$Loss = -\sum_{s \in T} \sum_{i=1}^{v} \hat{P}_i(C) \log(P_i(C)) \quad (8)$$

Where T is the training set and V is the sentiment categories. Stochastic Gradient Descent (SGD) is also used for end to end training of the model.

3.2. Convolutional Neural Network Equipped with Attention Layer and Transfer Learning

There is no doubt that deep learning is super-efficient for many tasks, however, it is not a silver bullet that can solve hidden problems particularly when the data is limited [2]. In fact, deep learning is not an ideal solution while sufficient data is not available and as the amount of training data is increased, deep learning becomes more efficient. Convolutional neural network, as a representative

of deep learning methods, is also really data-hungry and its performance is highly dependent on the size of the available training data. To overcome the problem of lack of data, transfer learning has emerged as a popular learning framework. Although transfer learning is suitable for a wide range of applications, its influence on natural language processing especially sentiment analysis has been rarely explored [10]. The focus of transfer learning is on storing learned knowledge from the source domain and then applying it to the different but related domain. Considering that $D_s = \{(x_{s1}, y_{s1}),(x_{s2}, y_{s2}),...,(x_{sn}, y_{sn})\}$ and $D_t = \{(x_{t1}, y_{t1}),(x_{t2}, y_{t2}),...,(x_{tn}, y_{tn})\}$ refer to the source and target domain respectively while T_s and T_t respectively show the source and target task. Moreover, $x_{si} \in X_s$ and $x_{ti} \in X_t$ respectively refer to the i-th data in the source and the target domain while $y_{si} \in Y_s$ and $y_{ti} \in Y_t$ show the i-th label of the source and target domain respectively. Therefore, transfer learning is used to enhance T_t of the target prediction function f_t at D_t by employing knowledge in D_s for T_s. The utilized transfer learning process for training the proposed model is depicted in Figure 2.

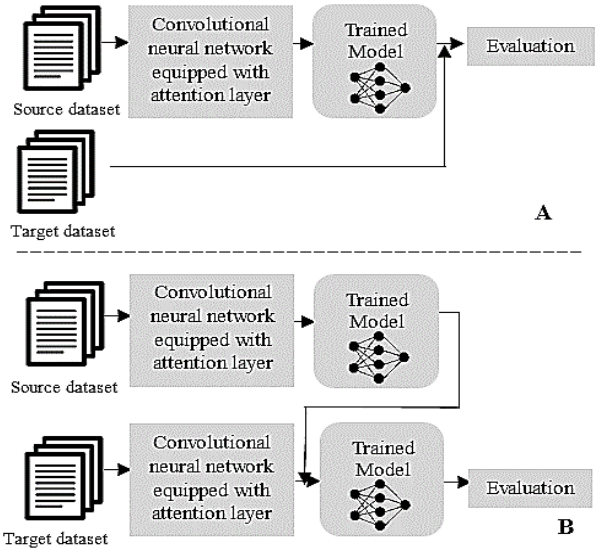

Figure 2. Learning process using transfer learning without (A) and with (B) training the model on the target domain.

As it is clear, the proposed model is firstly trained on the source domain based on the process flow that is shown in Figure 2, and then the trained model is transferred to the target domain. Transferring the trained model also includes two different processes. In the first one, the proposed model is not trained on the target domain. It means that the trained model is only tested on the target domain (Figure 2 A) while in the second process, the model is trained on the target domain to incrementally learn and update its knowledge (Figure 2 B). It means that in the second process the model is trained on both source and target domains and can use their combination to increase the classification performance. Finally, this new trained model is evaluated on the test set of the target domain.

4. Experiments

The experiments of this paper are divided into two sections. The first section investigates the effect of the proposed attention mechanism on the performance of convolutional neural networks. The second section explores the influence of transfer learning on the classification performance in the task of sentiment analysis as well as analyzing the sensitivity of the proposed model to parameters on different datasets.

4.1. Dataset

To provide a comprehensive investigation of the effectiveness of the proposed model, standard datasets for the aim of sentiment classification were used in our experiments. While the focus of this paper was also on exploring the impact of transfer learning, six different datasets were used in our experiments that were divided into two groups of source domain (D_S) and target domain (D_t). While transfer learning is generally used while the target data is not large enough for accurately training the model, small size datasets were considered as the target domain while large size datasets were assumed as the source domain in our experiments. Statistics of the used datasets are presented in Table 1. As it is clear, the source datasets have larger values of metrics compared to the target datasets that are smaller in size. More details about the used datasets are also provided in the following.

Table 1. Summary statistics of the datasets used.

	Dataset	C	L	S	V
D_S	AMZ-5	5	84	3650000	1057296
	AMZ-2	2	82	3000000	1112820
	YELP	2	141	560000	246735
	IMDB	2	257	25000	81321
D_T	SST-5	5	18	11855	17836
	SST-2	2	19	9613	16185

*D_S: Source domain, D_T: Target Domain, C: Number of classes, L: Average sentence length, S: Number of sentences, V: Vocabulary size

- Amazon Review (AMZ): This dataset contains reviews about Amazon products collected by

Zhang et al. [16]. This dataset has two class (AMZ-2) and five class (AMZ-5) versions.

- Yelp polarity review (YELP): This is a large size dataset that contains business reviews that are classified into two classes [16].
- IMDB: This dataset contains multi-sentence reviews about movies that are classified into two class [34].
- Stanford Sentiment Treebank (SST): This is the dataset that is commonly used for sentiment classification which also contains two classes (SST-2) and five classes (SST-5) versions. In fact, this is the extended version of the MR dataset [35], which also contains train/dev/test sets and fine-grained labels.

4.2. Results of Applying Attention Mechanism

The details of the experiments that were conducted to explore the effect of the proposed attention mechanism on the convolutional neural network and the obtained results are presented in this section.

4.2.1. Model Configuration

Pre-processing is considered as one of the most important components of training. In this regard, the documents were firstly split into sentences and Stanford core NLP was applied for the aim of tokenization. Extracted tokens were then used to obtain word embedding vectors using unsupervised Word2Vec [29], Glove [30], and FastText [31] models. It is worth mentioning that target datasets were only used in this set of experiments. For training word vectors, the dimension of word vectors was considered as 200, and window size was set to 3. Word vectors were updated based on a learning rate of 0.1. We used ADADELTA with a learning rate of 0.01 as an update rule for stochastic gradient descent and the mini-batch was 25. Filter size and the number of filters were considered as the hyperparameters. Their values for training the proposed model are presented in Table 2.

Table 2. Hyper-parameters configuration

Hyper-parameters	Value
Filter region size	3,4,5
Number of filters	128
Dropout rate	0.5
Activation function	ReLU
Batch size	25

As it is illustrated, it is found that filter size (3,4,5) and 128 filters led to better results on both datasets. Furthermore, the convolutional layer was regularized with a dropout rate of 0.5. ReLU (Rectified Linear Unit) was also used in our experiments as the activation function. 60 epochs were also used for training the model. Noteworthy, five different variations of the proposed model were used in our experiments as follows:

- *CNN+Attention-Rand:* It employs random initialized vectors as the input.
- *CNN+Attention-Static*: It employs pre-trained word vectors obtained from Word2Vec as the input. It is worth mention that weights are not updated along the training process.
- *CNN+Attention-Non-Static*: It employs pre-trained word vectors achieved from Word2Vec as the input. It is notable that weights are updated along the training process.
- *CNN+Attention-2channels*: It employs a combination of pre-trained word vectors obtained from Word2Vec and random initialized vectors as the input.
- *CNN+Attention-4channels*: It employs a combination of pre-trained word vectors obtained from Word2Vec, Glove, and FastText as the input. It is notable that weights are also updated along the training process.

4.2.2. Results

To create a baseline and conduct a fair comparison among the proposed model and another state-of-the-art, we only trained our proposed model on the target domains without considering transfer learning. The accuracy comparison of the proposed model against other existing models is provided in Table 3.

Table 3. Classification accuracy of the proposed model on target datasets

Model	Accuracy (%)	
	SST-2	**SST-5**
NB[20]	81.8	41
BiNB[20]	83.1	41.9
SVM [36]	79.4	40.7
WordVec-AVE [36]	80.1	32.7
CNN-1 layer [18]	77.1	37.4
CNN-non static[15]	87.2	48
CNN-multichannel[15]	88.1	47.4
DCNN [18]	86.8	48.5
LSTM[14]	85.2	46.2
Bi-LSTM[14]	87.5	49.1
Tree-LSTM[14]	88.0	51.0
Tree-GRU[37]	88.6	50.5
Tree-GRU+ attention[37]	89.0	51.0
LSTM+RNN attention[23]	86.1	48.0
RecRNN[20]	82.4	43.2
RNTN[20]	85.4	45.7
MVRNN[19]	82.9	44.4
CNN+Attention-Rand	88.61	49.76
CNN+Attention-Static	89.95	50.06
CNN+Attention-Non-Static	90.57	51.31
CNN+Attention-2channel	91.02	**52.18**
CNN+Attention-4channel	**91.59**	52.08

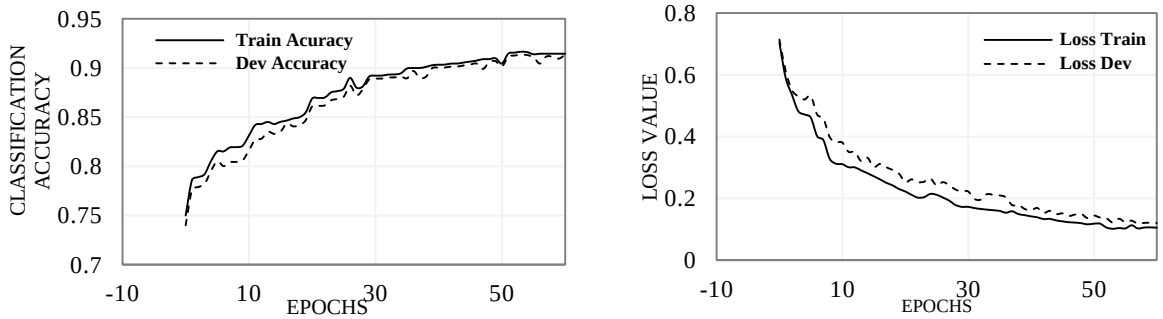

Figure 3. CNN-Attention- 4channel evaluation on SST2 dataset. The left figure plots the classification accuracy for training and validation samples and the right figure plots the cross-entropy loss for training and validation samples at different epochs.

As it is clear, most of the variations of the proposed model have a slight superior performance compared to existing models and therefore it can be considered as a good option for being used in transfer learning. It is obvious that among all variations of the proposed model, CNN+Attention-Rand has the lowest classification performance on both target datasets which can be due to the employment of random initialized vectors as input. Better performance of other variations can also be attributed to the employment of pre-trained vectors that can solve the semantic sparsity problem. Moreover, considering the fact that CNN+Attention-Static has the lowest classification accuracy besides CNN+Attention-Rand, it is obvious that updating word vectors along the training process could yield to obtain a higher performance without considering if the word vectors were previously trained or not. Finally, CNN+Attention-4channel has the highest classification accuracy of 91.59% on SST-2 and CNN+Attention-2channel has the highest classification accuracy of 52.18% on SST-5 dataset.

In order to provide a better analysis of the performance of the proposed model, the plot of classification accuracy and loss values per epoch over training and validation sets for CNN+Attention-4channel on SST2 dataset are illustrated in Figure. 3. As depicted, by enhancing the number of the epoch, classification accuracy is increased while the loss value is decreased. Moreover, the plot of training loss and validation loss both decreased to a point of stability, and validation loss had a small gap with the training loss. Therefore, it can be concluded that the model is finely tuned and is a good fit.

4.3. Effect of Applying Transfer Learning

Considering the fact that the lack of enough training data is known as one of the prominent challenges that deep learning is confronted with, we decided to explore the effect of transfer learning on the performance of the proposed model. The details of the experiments that were conducted to investigate the efficiency of transfer learning and the obtained results are presented in this section.

4.3.1. Model Configuration

Generally, one of the downsides of the convolutional neural networks refers to their free number of hyper-parameters which require practitioners to determine the exact model architecture. While the hyper-parameters' values have a significant influence on the efficiency of deep neural networks, we decided to optimize the proposed model hyper-parameters on the source domains and then apply the optimal parameters on the target domains while transfer learning is applied. Therefore, we used the previously mentioned configuration as a baseline and tried to carry out extensive sets of experiments to obtain their optimal values on a source domain to observe their effect on transfer learning. Notably, 10-fold cross-validation where 10% of the training data was randomly selected as a test set was performed and each experiment was repeated for 5 times and the average results are reported.

- To explore the influence of the *filter size*, different numbers of filter sizes were used in our experiments, while the other parameters were kept constant. As it is depicted in Table 4, different filter size has a remarkable effect on the efficiency of the model and the greatest accuracies were achieved while the multiple filter size was set as (4, 5, 6) in all source datasets.

Table 4. The influence of filter on the performance of the proposed model based on different source domains.

Filter size	Accuracy (%)			
	AMZ-2	AMZ-5	YELP	IMDB
(3,4,5)	92.44	57.31	93.46	86.41
(4,5,6)	**92.98**	**57.65**	**94.65**	**88.63**
(6,7,8)	91.8	56.85	93.61	87.63
(8,9,10)	90.78	55.85	93.08	86.34
(9,10,11)	90.45	55.32	92.35	85.41
(14,15,16)	90.30	55.24	92.58	85.92
(3,4,5,6)	91.10	56.07	92.18	84.18
(6,7,8,9)	90.17	55.14	93.45	85.02

- To consider the influence of the *number of filters*, other configurations were kept constant, and only the number of filters in each filter region was changed. According to the empirical results (Table 5), the highest classification accuracy was achieved when the number of filters was set to 300 on AMZ-2 and AMZ-5 datasets. On YELP and IMDB datasets the highest values were obtained when the number of filters was respectively set to 256 and 128.

Table 5. Influence of number of filter on the performance of the proposed model based on different source domains.

Number of filters	Accuracy (%)			
	AMZ-2	**AMZ-5**	**YELP**	**IMDB**
128	91.5	56.32	93.25	**88.94**
256	91.78	56.54	**94.68**	87.63
300	**92.83**	**57.74**	94.21	86.45
512	92.44	57.31	90.15	85.12
450	90.25	55.34	91.12	86.08

- To explore the influence of *dropout* as a technique that is generally used for the aim of regularization, various dropout rates in the range of 0.1-0.9 were employed in our experiments. According to the empirical results (Table 6), the highest accuracy was achieved while the dropout rate was about 0.6 on AMZ-2 and AMZ-3 datasets and about 0.5 on YELP and IMDb datasets.

Table 6. Influence of dropout rate on the performance of the proposed model based on different source domains.

Dropout rate	Accuracy (%)			
	AMZ-2	AMZ-5	**YELP**	**IMDB**
0.1	91.54	56.73	92.87	87.86
0.2	91.40	56.32	93.94	88.01
0.3	91.93	56.57	93.89	87.54
0.4	91.8	56.23	94.02	88.21
0.5	92.44	57.31	**94.83**	**88.83**
0.6	**92.89**	**57.68**	94.73	88.15
0.7	91.48	56.41	94.23	87.35
0.8	92.08	57.12	93.84	87.26
0.9	92.28	57.43	93.97	86.18

- In order to investigate the influence of *activation function*, different activation functions, such as ReLU, Tanh, SoftPlus, and linear were used in our experiments. According to the empirical results (Table 7), the ReLU function outperformed the other activation function on all source datasets.

Table 7. Influence of activation function on the performance of the proposed model based on different source domains.

Activation function	Accuracy (%)			
	AMZ-2	**AMZ-5**	**YELP**	**IMDB**
Tanh	91.45	57.12	92.05	87.63
Softplus	80.25	40.43	93.17	88.01
ReLU	**92.4**	**57.31**	**94.93**	**88.94**
Linear	90.31	56.25	91.35	85.42

4.3.2. Results

While the performance of deep neural networks is highly dependent on the number of data, and increasing the number of training data has a significant effect on their performance, we used transfer learning in our experiment and trained our model on the source domains first, and then transferred it to target domains. Two different learning processes were used in our experiments (Figure 2). In one of them, the proposed model was directly used for sentiment classification in the target domain and in the other one, the model is incrementally trained on the target domain using the optimal values that were previously discussed. To make a comparison between these two learning processes, the accuracy of the proposed model with and without incremental learning in the different source and target domains is presented in Table 8.

Table 8. Transfer learning performance comparison with and without incremental learning.

$D_S \rightarrow D_T$	Description	Accuracy (%)
AMZ-2→SST-2	Transfer learning without incremental learning	86.3
	Transfer learning with incremental learning	92.53
AMZ-2→SST-5	Transfer learning without incremental learning	49.77
	Transfer learning with incremental learning	53.93

It must be mentioned that CNN-Attention-4channel is used in this set of experiments. As it is clear, the accuracy of the model is lower when it is directly used for classification which indicates that the knowledge obtained in the source domain is not enough to be applied in the target domain.

On the other hand, when the model is incrementally learned in the target domain, the accuracy is significantly improved. To this end, we decided to use transfer learning with incremental learning in our experiments.

To provide more analysis of the impact of transfer learning on sentiment classification, the results of employing transfer learning with incremental learning on all variations of source and target domains are presented in Table 9. As it is clear, employing transfer learning has generally enhanced the overall classification performance. In fact, it can be stated that employing transfer learning remarkably enhanced the performance of the proposed model which can be due to the large size of the source domain and richer embedding which helps the model to learn contextual information better. Presented results of this set of experiments are also attributed to CNN-Attention-4channel.

Table 9. Accuracy (%) of the proposed model with transfer learning on different variations of source and target domains.

D_T	→ SST-2				→SST-5			
D_S	AMZ-2	AMZ-5	YELP	IMDB	AMZ-2	AMZ-5	YELP	IMDB
CNN+Attention -4channel	92.53	**92.88**	91.68	91.82	**53.93**	53.26	52.17	52.83

As can be seen, various source domains yield different classification accuracy on target domains. In this respect, it can be said that choosing the most compatible source dataset for transferring knowledge to the target dataset is a primary challenge in transfer learning. According to the fact that IMDB is semantically more similar to SST-2 compared to AMZ and YELP datasets, we supposed that using IMDB as a source domain may lead to an enhancement in classification performance. However, the obtained results were completely otherwise, and AMZ-5→SST-2 obtained the highest classification performance which can be due to its larger size that can provide richer embedding. The same result was also obtained for AMZ-2→ SST-5 although the source and target domains did not have an equal number of classes.

Although YELP was also a large size dataset, using it as a source dataset led to lower accuracy compared to IMDB which could be due to the fact that the words in the target dataset were not available in the source dataset. In fact, YELP dataset contains business reviews and its vocabulary domain is different from the assumed source datasets. In summary, besides semantics, size and out of vocabulary metrics of the source dataset have also a great impact on the classification accuracy.

5. Conclusion

The contribution of this paper is two-fold. Firstly, an attention-based convolutional neural network for the aim of sentiment classification was proposed that used an attention mechanism before the pooling layer to focus on the salient part of the sentences and consider the context to determine the polarity of sentences. Secondly, considering the fact that the lack of enough training data is known as one of the major challenges that deep neural networks are confronted with, we explored the effects of transfer learning and the sensitivity of the proposed model parameters to overcome this issue.

According to the empirical results, the proposed attention-based convolution neural network significantly outperformed other existing models. Moreover, employing transfer learning greatly improved the classification accuracy.

In fact, the proposed model obtained an accuracy of 91.59% and 52.18% on SST-2 and SST-5 datasets respectively without transfer learning. On the other hand, by applying transfer learning the classification accuracy increased and the obtained accuracies were about 92.88% and 53.93% on SST-2 and SST-5 datasets respectively. Based on the observations, it can also be said that the size and semantic of the source domains have also a great impact on the efficiency of transfer learning.

Following a similar line of research, the proposed model using transfer learning can be performed in another target domain or for other natural language processing tasks. Applying transfer learning to other deep neural networks is also worth exploring.

References

[1] S. A. Salloum, R. Khan, and K. Shaalan, "A survey of semantic analysis approaches," in *Joint European-US Workshop on Applications of Invariance in Computer Vision*, 2020: Springer, pp. 61-70.

[2] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: a review," *Artificial Intelligence Review*, Vol. 53, No. 6, pp. 4335-4385, 2020.

[3] M. I. Prabha and G. U. Srikanth, "Survey of sentiment analysis using deep learning techniques," in *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, 2019: IEEE, pp. 1-9.

[4] O. Habimana, Y. Li, R. Li, X. Gu, and G. Yu, "Sentiment analysis using deep learning approaches: an overview," *Science China Information Sciences*, Vol. 63, No. 1, pp. 1-36, 2020.

[5] X. Xie, S. Ge, F. Hu, M. Xie, and N. Jiang, "An improved algorithm for sentiment analysis based on maximum entropy," *Soft Computing*, Vol. 23, No. 2, pp. 599-611, 2019.

[6] H. Sadr, M. N. Soleimandarabi, M. Pedram, and M. Teshnelab, "Unified Topic-Based Semantic Models: A Study in Computing the Semantic Relatedness of Geographic Terms," in *2019 5th International Conference on Web Research (ICWR)*, 2019: IEEE, pp. 134-140.

[7] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1480-1489.

[8] Z. Zhang, Y. Zou, and C. Gan, "Textual sentiment analysis via three different attention convolutional neural networks and cross-modality consistent regression," *Neurocomputing*, Vol. 275, pp. 1407-1415, 2018.

[9] H. Sadr, M. M. Pedram, and M. Teshnelab, "Improving the Performance of Text Sentiment Analysis using Deep

Convolutional Neural Network Integrated with Hierarchical Attention Layer," *International Journal of Information and Communication Technology Research*, Vol. 11, No. 3, pp. 57-67, 2019.

[10] R. Liu, Y. Shi, C. Ji, and M. Jia, "A survey of sentiment analysis based on transfer learning," *IEEE Access*, Vol. 7, pp. 85401-85412, 2019.

[11] H. Sadr and M. Nazari Solimandarabi, "Presentation of an efficient automatic short answer grading model based on combination of pseudo relevance feedback and semantic relatedness measures," *Journal of Advances in Computer Research*, Vol. 10, No. 2, pp. 1-10, 2019.

[12] H. Sadr, M. Nazari, M. M. Pedram, and M. Teshnehlab, "Exploring the Efficiency of Topic-Based Models in Computing Semantic Relatedness of Geographic Terms," *International Journal of Web Research*, Vol. 2, No. 2, pp. 23-35, 2019.

[13] M. Kuta, M. Morawiec, and J. Kitowski, "Sentiment Analysis with Tree-Structured Gated Recurrent Units," *Springer International Publishing AG 2017*

[14] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," *arXiv preprint arXiv:1503.00075*, 2015.

[15] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[16] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Advances in neural information processing systems*, 2015, pp. 649-657.

[17] W. Yin, H. Schütze, B. Xiang, and B. Zhou, "Abcnn: Attention-based convolutional neural network for modeling sentence pairs," *arXiv preprint arXiv:1512.05193*, 2015.

[18] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," *arXiv preprint arXiv:1404.2188*, 2014.

[19] R. Socher, B. Huval, C. D. Manning, and A. Y. Ng, "Semantic Compositionality through Recursive Matrix-Vector Spaces," *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics*., 2012.

[20] R. Socher *et al*., "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," in EMNLP, 2013.

[21] H. Sadr, M. M. Pedram, and M. Teshnehlab, "A Robust Sentiment Analysis Method based on Sequential Combination of Convolutional and Recursive Neural Networks," *Neural Processing Letters*, pp. 1-17, 2019.

[22] H. Sadr, M. M. Pedram, and M. Teshnehlab, "Multi-View Deep Network: A Deep Model based on Learning Features From Heterogeneous Neural Networks for Sentiment Analysis," *IEEE Access*, Vol. 8, pp. 86984-86997, 2020.

[23] Y. Wang, M. Huang, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, pp. 606-615.

[24] Z. Yuan, S. Wu, F. Wu, J. Liu, and Y. Huang, "Domain attention model for multi-domain sentiment classification," *Knowledge-Based Systems*, Vol. 155, pp. 1-10, 2018.

[25] T. Semwal, P. Yenigalla, G. Mathur, and S. B. Nair, "A practitioners' guide to transfer learning for text classification using convolutional neural networks," in *Proceedings of the 2018 SIAM International Conference on Data Mining*, 2018: SIAM, pp. 513-521.

[26] F. Zhuang et al., "A comprehensive survey on transfer learning," *arXiv preprint arXiv:1911.02685*, 2019.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[28] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[29] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[30] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532-1543.

[31] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information,"

[32] S. Sukhbaatar, J. Weston, and R. Fergus, "End-to-end memory networks," in *Advances in neural information processing systems*, 2015, pp. 2440-2448.

[33] A. Kumar et al., "Ask me anything: Dynamic memory networks for natural language processing," in *International conference on machine learning*, 2016, pp. 1378-1387.

[34] A. Maas, R. E. Daly, P. T. P. am, D. Huang, A. Y. Ng, and C. Potts, "Learning Word Vectors for Sentiment Analysis," *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics*, 2011.

[35] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proceedings of the 43rd annual meeting on association for computational linguistics*, 2005: Association for Computational Linguistics, pp. 115-124.

[36] C. Du and L. Huang, "Sentiment Classification Via Recurrent Convolutional Neural Networks," *DEStech Transactions on Computer Science and Engineering*, No. cii, 2017.

[37] F. Kokkinos and A. Potamianos, "Structural attention neural networks for improved sentiment analysis," *arXiv preprint arXiv:1701.01811*, 2017.

شبکه عصبی پیچشی مجهز به سازوکار توجه و یادگیری انتقالی برای افزایش دقت تجزیه و تحلیل احساسات

حسین صدر[۱]، میرمحسن پدرام[۲،*] و محمد تشنه‌لب[۳]

[۱] گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد رشت، گیلان، ایران.

[۲] گروه مهندسی برق و کامپیوتر، دانشکده فنی و مهندسی، دانشگاه خوارزمی، تهران، ایران.

[۳] دانشکده مهندسی برق، گروه سیستم‌ها و کنترل، دانشگاه صنعتی خواجه نصیر طوسی، تهران، ایران.

ارسال ۲۰۲۰/۰۵/۲۹؛ بازنگری ۲۰۲۰/۱۱/۱۱؛ پذیرش ۲۰۲۱/۰۱/۲۱

چکیده:

با توسعه سریع داده‌های متنی در سطح وب، تجزیه و تحلیل احساسات از یک موضوع تحقیقاتی به یک ابزار ضروری تحلیل تبدیل شده و در سال‌های اخیر مطالعات زیادی پیرامون این موضوع صورت گرفته است. با ظهور یادگیری ژرف، شبکه‌های عصبی ژرف نیز توجه بسیاری از محققان را به خود جلب کرده و به جریان اصلی در این زمینه تبدیل شده اند. علی‌رغم موفقیت چشم‌گیر مدل‌های یادگیری ژرف برای تجزیه و تحلیل احساسات از متن، همچنان در ابتدای مسیر پیشرفت قرار دارند و پتانسیل آن‌ها هنوز به‌طور کامل بررسی نشده است. در این راستا، شبکه عصبی پیچشی یکی از روش‌های یادگیری ژرف است که علی‌رغم به دست آوردن دقت قابل‌توجه در این حوزه همچنان با محدودیت‌هایی مواجه می‌باشد. اولاً، شبکه‌های عصبی پیچشی به تعداد زیادی داده برای آموزش نیاز دارند. ثانیاً، این شبکه‌ها با توجه به اینکه از فیلترهای محلی برای استخراج ویژگی‌ها استفاده می‌کنند، سهم یکسانی برای تمام کلمات موجود در متن هنگام تعیین قطبیت یک جمله در نظر گرفته و تأثیر کلمات کلیدی را نادیده می‌گیرند. برای غلبه بر این چالش‌ها، در این مقاله یک شبکه عصبی پیچشی مجهز به سازوکار توجه و یادگیری انتقالی معرفی شده است که به‌طور هم‌زمان از سازوکار توجه و یادگیری انتقالی برای بهبود عملکرد تجزیه و تحلیل احساسات از متن بهره می‌برد. با توجه به نتایج به دست آمده از آزمایش‌ها، مدل پیشنهادی از دقت بالاتری نسبت به سایر روش‌های پیشین برخوردار است.

کلمات کلیدی: تجزیه و تحلیل احساسات، یادگیری ژرف، شبکه عصبی پیچشی، ساز وکار توجه، یادگیری انتقالی.