Artificial Intelligence in Medicine xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

What matters in a transferable neural network model for relation classification in the biomedical domain?

Sunil Kumar Sahu, Ashish Anand*

Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, India

ARTICLE INFO

Article history:
Received 9 August 2017
Received in revised form 13 February 2018
Accepted 29 March 2018

Keywords:
Transfer learning
Recurrent neural network
LSTM network
Relation classification

ABSTRACT

A lack of sufficient labeled data often limits the applicability of advanced machine learning algorithms to real life problems. However, the efficient use of *transfer learning* (TL) has been shown to be very useful across domains. TL make use of valuable knowledge learned in one task (*source task*), where sufficient data is available, in order to improve performance on the task of interest (*target task*). In the biomedical and clinical domain, a lack of sufficient training data means that machine learning models cannot be fully exploited. In this work, we present two unified recurrent neural models leading to three transfer learning frameworks for relation classification tasks. We systematically investigate the effectiveness of the proposed frameworks in transferring knowledge from a source task to a target task when the characteristics of the source data vary, such as similarity or relatedness between the source and target tasks, and the size of training data for the source task. Our empirical results show that the proposed frameworks, in general, improve the model performance. However, these improvements do depend on characteristics of source and target tasks. This dependence then finally determine the choice of a particular TL framework.

1. Introduction

Recurrent neural networks (RNN) and their variants, such as long short term memory (LSTM) network, have shown to be effective models for many natural language processing tasks [1–6]. However, the requirement of huge gold standard labeled datasets for training makes it difficult to apply them to tasks for which few such resources exist, which is often the case in the biomedical domain. In the biomedical domain, obtaining labeled data is not only time consuming and costly, but also requires domain knowledge. Transfer learning (TL) has been used successfully in such scenarios across multiple domains. The aim of transfer learning is to apply the knowledge gained while training a model for a Task-A (*Source Task*), where we have sufficient gold standard labeled data, to a different Task-B (*Target Task*) where we do not have enough training data [7]. In literature, various TL frameworks have been proposed [7–9]. With the recent surge in applications of TL using neural network based models in computer vision and image processing [9,10] as well as in NLP [8,11,12], this work explores TL frameworks using neural models for relation classification in the biomedical domain.

Applying TL in a sequential framework on a source and a target task is a standard approach. We refer to this approach as *sequential TL*. Furthermore, if there exists a bijection mapping between the label sets of the source and target tasks, then the entire model trained on the source task can be transferred to the target task. Otherwise, only a partial model can be utilized. In NLP, transfer of feature representations is the most common form of partial model transfer. Instead of performing the training in a sequential manner, an alternative method is to train the model on both source and target data simultaneously [12]. This is very similar to *multi-task learning* [13]. This way of simultaneous training can be carried out in multiple ways. These options make it possible to deign several variants of the TL framework.

In addition the options of using training data in different ways, using partial or complete model transfer, and the presence or absence of bijection mapping between two label sets, other aspects such as *selection of the source task*, *its size* and *relatedness or similarity to the target task* determine the selection of the most relevant TL model. Intuitively, it is preferable for the source task to be as similar as possible to the target task. For example, if the target task concerns the binary classification of drug–drug interaction (DDI) mentioned

* Corresponding author.
E-mail address: anand.ashish@iitg.ernet.in (A. Anand).

https://doi.org/10.1016/j.artmed.2018.03.006
0933-3657/

in social media text or in doctors' notes, then a potentially suitable source task would be the binary classification of DDIs mentioned in research articles. Here, the difference lies in the nature of texts appearing in the two corpora. In the case of doctors' notes, the text is likely to be short and precise compared to the research articles. In other words, the feature spaces representing data for source and target tasks differ from each other, although the two label sets are same. On the other hand, it is also possible that there does not exist any bijection between the labels of the source and target tasks. An example of such a scenario would be if the target task required multi-class classification of DDIs, while the source task only concerned binary classification.

According to the various possible scenarios introduced in the above discussion, we present two LSTM based models and three different corresponding TL frameworks in this study. Our motivation is to systematically explore various TL frameworks for the task of relation classification in the biomedical domain and try to empirically analyze the results that we obtain. Our contribution can be summarized as follows:

- We present and evaluate three TL framework variants based on LSTM models for different relation classification tasks in biomedical and clinical text.
- We analyze the impact of relatedness (implicit or explicit) between the source and target tasks on the effectiveness of TL framework.
- We explore how the size of the training data corresponding to source task impacts on the effectiveness of TL frameworks.

2. Model architectures

In this section, we firstly explain a generic LSTM model architecture for relation classification tasks. Then we explain three ways of using this architecture for transferring knowledge from source tasks to target tasks. We assume that a relation exists between two entities, referred to as *target entities*, whose positions within the sentence are known.

The generic neural network architecture for the relation classification task consist of the following layers: *word level feature layer*, *embedding layer*, *sentence level feature extraction layer*, *fully connected and softmax layers*. We define features for all words in the *word level feature layer*, which also includes some features relative to the two targeted entities. In the *embedding layer* every feature gets mapped to a vector representation through a corresponding embedding matrix. Raw features are combined with the entire sentence and a fixed length feature representation is obtained in the *sentence level feature extraction layer*. Although a convolution neural network (CNN) or other variants of a recurrent neural network can be used in this layer, we use a bidirectional LSTM because of its relatively better ability to take into account discontinuous features. The *fully connected and softmax layer* map the sentence level feature vectors to a class probability. In summary, the input for these models would be a sentence containing the two targeted entities and the output would be a probability distribution over each possible relation class between them.

2.1. BLSTM-RE

Suppose $w_1, w_2, \ldots, w_m$ is a sentence of length m. Two targeted entities e_1 and e_2 correspond to some words (or phrases) w_i and w_j respectively. In this work, we use the word and its position from both targeted entities as features in the word level feature layer. Positional features are important for relation extraction, because they inform the model about the targeted entities [14,15]. The output of *embedding layer* would be a sequence of vectors $x_1, x_2, \ldots, x_m$ where $x_i \in \mathbb{R}^{(d_1+d_2+d_3)}$ is the concatenation of word and position vectors. d_1,d_2 and d_3 are the embedding lengths of the word, the position from the first entity and the position from the second entity respectively. We use a bidirectional LSTM with *max* pooling in the *sentence level feature extraction layer*. This layer is responsible for obtaining an optimal fixed length feature vector from entire sentence. The basic architecture of *BLSTM-RE* model is shown in Fig. 1a. We omit the mathematical equations as there are no modifications made to the standard bi-directional LSTM model [2].

2.2. T-BLSTM-Mixed

T-BLSTM-Mixed is a specific way to use the *BLSTM-RE* model in a transfer learning framework. In this case, the instances from both source and target tasks are fed into the same *BLSTM-RE* model. While training, we pick one batch of data from the source or target in a random order with equal probability. Since the training happens simultaneously for both the source and target datasets, we can say that the model will learn features which are applicable to both datasets. It is quite obvious that this model is only applicable for those cases in which bijection mapping between labels of the source and target tasks exists.

2.3. T-BLSTM-Seq

The convergence of neural network based models depends on the initialization of model parameters. Several studies [16,17,14] have shown that initializing parameters with values from other supervised or unsupervised pre-trained models often improves the model convergence. In this framework of transfer, we firstly train our model with the source tasks dataset and use the learned parameters to initialize the model parameters for training a separate model to carry out the target task. We call this framework *T-BLSTM-Seq*. *T-BLSTM-Seq* can be applicable for the transfer of both the *same label set* and *disparate label sets* transfer. We transfer the entire set of network parameters if there exists a bijection mapping between the source and target label sets, otherwise, we only share model parameters up to the second last layer of the network. The left out last layer is randomly initialized.

2.4. T-BLSTM-Multi

We propose another transfer learning framework, called *T-BLSTM-Multi*, using the same backbone of *BLSTM-RE* model. As shown in Fig. 1b, this model has two *fully connected and softmax* layers, one for the source task and other is for the target task. The other layers of the models are shared for the two tasks. While training, the parameters of the shared block are updated with training instances from both source and target data and the *fully connected* layer is updated only with its corresponding task data. A batch of instances are picked in a similar manner to *T-BLSTM-Mixed*. This method of training is also called *multi-task learning* but in that case, the focus is on the performance of both the source and target tasks. The *T-BLSTM-Multi* model is also applicable for both *disparate label set* as well as *same label set* transfer.

2.5. Training and implementation

Pre-trained word vectors are used to initialize the word embeddings and random vectors are used for other feature embeddings. We apply GloVe [18] on Pubmed corpus [19] to obtain word vectors. The dimensions of word and position embeddings are set to 100 and 10, respectively. Adam optimization [20] is used for training all models. All parameters, i.e., word embeddings, position embeddings, and network parameters are updated during training. We fixed the batch size to 100 for all the experiments. In the case of

Please cite this article in press as: Sahu SK, Anand A. What matters in a transferable neural network model for relation classification in the biomedical domain? Artif Intell Med (2018), https://doi.org/10.1016/j.artmed.2018.03.006