Pattern Recognition 74 (2018) 15–24

Contents lists available at ScienceDirect

Pattern Recognition

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

Revisiting multiple instance neural networks

Xinggang Wang*, Yongluan Yan, Peng Tang, Xiang Bai*, Wenyu Liu

School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China

ARTICLE INFO

Article history:
Received 12 April 2017
Revised 24 July 2017
Accepted 23 August 2017
Available online 31 August 2017

Keywords:
Multiple instance learning
Neural networks
Deep learning
End-to-end learning

ABSTRACT

Of late, neural networks and Multiple Instance Learning (MIL) are both attractive topics in the research areas related to Artificial Intelligence. Deep neural networks have achieved great successes in supervised learning problems, and MIL as a typical weakly-supervised learning method is effective for many applications in computer vision, biometrics, natural language processing, and so on. In this article, we revisit Multiple Instance Neural Networks (MINNs) that the neural networks aim at solving the MIL problems. The MINNs perform MIL in an end-to-end manner, which take bags with a various number of instances as input and directly output the labels of bags. All of the parameters in a MINN can be optimized via back-propagation. Besides revisiting the old MINNs, we propose a new type of MINN to learn bag representations, which is different from the existing MINNs that focus on estimating instance label. In addition, recent tricks developed in deep learning have been studied in MINNs; we find *deep supervision* is effective for learning better bag representations. In the experiments, the proposed MINNs achieve state-of-the-art or competitive performance on several MIL benchmarks. Moreover, it is extremely fast for both testing and training, for example, it takes only 0.0003 s to predict a bag and a few seconds to train on MIL datasets on a moderate CPU.

1. Introduction

Multiple Instance Learning (MIL) was originally proposed for drug activity prediction [1]. Now it has been widely applied to many domains and is an important problem in machine learning. Many multimedia data have the Multiple Instance (MI) structure; for example, a text article contains multiple paragraphs, an image can be divided into multiple local regions, and a gene expression data contains multiple genes. MIL is useful to processing and understanding MI data.

Multiple instance learning is a kind of Weakly-Supervised Learning (WSL). Each sample is in the form of labeled bags, composed of a wide diversity of instances associated with input features. The aim of MIL, in a binary task, is to train a classifier to predict labels of testing bags, which is based on the assumption that a positive bag contains at least one positive instance, whereas a bag is negative if it is only constituted of negative instances. Thus, the crux of MIL is to deal with the ambiguity of the labels of the instances, especially in positive bags that have plenty of cases with different compositions.

There are many algorithms have been proposed to solve the MIL problem. According to the survey by Amores [2], MIL algorithms are in three folds: instance-space paradigm, bag-space paradigm, and embedded-space paradigm. Instance-space paradigm learns the instance classifier and performs bag classification by aggregating the responses of instance-level classifier. Bag-space paradigm exploits bag relations and treats bag as a whole; in particular, it calculates bag-to-bag distance/similarity; then the nearest neighbor or Bayesian classifier carries out bag classification based on the distances/similarities. Embedded-space paradigm embeds a bag into a vocabulary-based feature space to obtain a compact representation for the bag, for example, a vector representation; then classical classifiers can be applied to solve the bag classification problem.

Deep neural networks have been applied to solve many machine learning problems. For supervised learning, there are several kinds of neural networks. Deep Belief Networks (DBN) [3] use unsupervised pre-training and take a fixed length vector as input for feature learning, regression, and classification. Deep Convolutional Neural Networks (CNN) [4,5] take images as input and have dominated many computer vision problems. Deep Recurrent Neural Networks (RNN) [6] and Long Short Term Memory (LSTM) networks [7] take sequential data as input, such as text and speech, and are good at dealing with sequence prediction problems. Usually, training these deep networks requires a huge number of fully labeled data, that is, each training sample/instance needs a label.

* Corresponding author.
E-mail addresses: xgwang@hust.edu.cn (X. Wang), yongluanyan@hust.edu.cn (Y. Yan), pengtang@hust.edu.cn (P. Tang), xbai@hust.edu.cn, xiang.bai@gmail.com (X. Bai), liuwy@hust.edu.cn (W. Liu).

http://dx.doi.org/10.1016/j.patcog.2017.08.026
0031-3203/

However, in MIL, only bag-level labels are given. Meanwhile, MI data have a more complex structure which is a set of instances in various size. Also, MI data is different from the sequential data mentioned above, since there is no order information between instances. These problems make it difficult to deal with the MIL problem by conventional neural networks.

Before the raising of deep learning, some research studies were trying to solve the MIL problem using neural networks. In the year of 2000, Ramon and Raedt [8] firstly proposed a Multiple Instance Neural Network (MINN). The network estimates instance probabilities before the last layer and calculates bag probability using a convex max operator (i.e., log-sum-exp). The network was trained using back-propagation. Then, Zhang and Zhou [9] also proposed a multiple instance network that calculates bag probability by directly taking the max of instance probabilities.

A MINN takes a bag with multiple instances as input. Instance-level representation is gradually learned layer by layer guided by bag-level supervision. To inject the bag-level representation, there are two different network architectures. Following the naming style in a classical MIL study [10], we name the two networks as mi-Net and MI-Net, which aim at dealing with the MIL problem in instance-space paradigm and embedded-space paradigm [2], respectively. In mi-Net, there are instance classifiers in the each layer. We can obtain instance predictions for both training and testing bags, which is an appealing property in some applications. Different from MI-Net, there is no instance classifier. It directly builds a fixed-length vector as the bag representation and then learns bag classifier. Compared with mi-Net, MI-Net can obtain better bag classification accuracy. The previous studies are in the mi-Net category. We newly propose MI-Net in this article.

A key component in MINN is MIL Pooling Layer (MPL), which aggregates either instance probability distribution vectors or instance feature vectors into a bag probability/feature vector. It bridges MI data with conventional neural networks. As it must be differentiable, there are a few choices, such as max pooling, mean pooling, and log-sum-exp pooling. These pooling methods are compared and discussed in the experiments section. Besides MIL pooling layer, we use fully-connected layers with non-linear activations for instance feature learning. In MIL benchmarks, instance features are hand-crafted and raw data of instances are given. Even so, it is beneficial to do feature transformation guided by the bag-level supervision. Finally, for MI-Net, we use a fully-connected layer with only one neuron to match the predicted bag label with ground-truth in training.

Training neural networks using complex MI data is a challenging task. To learn good instance feature, we have tried to adopt various recent progresses of deep learning in MINN, such as dropout [11], Rectified Linear Unit (ReLU) [12], Deeply Supervised Nets (DSN) [13] and Residual Connections [14]. We find DSN is the most effective one because DSN can fuse the hierarchical features to make a better decision. Besides, residual connections are also helpful in MINNs.

To summarize, we revisit the problem of solving MIL using neural networks (MINNs), which are ignored in current MIL research community. Our experiments show that MINNs is very effective and efficient. Different from most MIL algorithms, MINNs optimize instance feature learning, bag feature learning, instance classification, and bag classification in a fully end-to-end manner via back-propagation. This article focuses on MINNs with comprehensive studies on MIL benchmarks. The main contributions of this article include two extremely fast and scalable methods for MIL, mi-Net, and MI-Net, and introducing deep supervision and residual connections for MIL.

The rest of this article is organized as follows. Section 2 briefly reviews previous studies on MIL. In Section 3, we propose end-to-end MIL networks. Our experimental results are presented on several MIL benchmarks in Section 4. Some discussions of experimental setups are presented in Section 5. Finally, in Section 6, we conclude the article with some future studies.

2. Related work

The previous MIL works based on neural networks were mainly proposed by Zhou et al. and Ramon et al. in [8,9,15,16]. and Raedt [8] introduced the use of a log-sum-exp function as the convex max to calculate bag probabilities from instance probabilities. Zhou and Zhang [9] changed to a different loss function and directly applied max function. Zhang and Zhou [15] improved multiple instance neural networks by feature selection using Diverse Density and PCA. Zhang and Zhou [16] showed that ensemble methods could be integrated with multiple instance neural networks. Subsequently, solving MIL using neural networks has been ignored in machine learning research. This article revisits this problem, proposes some new network structures, and investigates some of the recent neural network tricks. The idea of using neural networks for solving MIL problem has been studied in some computer vision studies, such as [17,18]. Wu et al. [17] proposed a deep MIL which uses max pooling to find positive instances/patches for image classification and annotation. Pinheheiro et al. [18] used log-sum-exp pooling in deep CNN for weakly supervised semantic segmentation. The studied mi-Net follows the path of these two works [17,18]. are applications of mi-Net. Thus, it is not necessary to compare to them in the experiments. In addition, in this article, we study the variants of mi-Net that utilize deep supervision and focus on more general MIL problems. Besides integrating MIL into deep neural networks, Wang et al. proposed a method to combine MIL with support vector machine using a relaxed MIL constraint [19] and applied this for object discovery. However, they pay more attention to vision applications (e.g., image classification, image annotation, and semantic segmentation, etc.), which are based on convolutional image features. Meanwhile, they always fine-tune neural network models pre-trained on other much larger datasets such as ImageNet [20]. Moreover, they only focus more on instance-space MIL. We focus on applying MINNs for more general MIL problems. Notice that for general MIL problems, there are no available large datasets for pre-training such as computer vision, which makes it harder to train MINNs efficiently. As we will show in experiments, training [8,9,17] like MINNs on such small scale MIL benchmarks directly cannot obtain satisfactory results. To solve this problem, We show many tricks to train our networks from the start on MIL benchmarks with limited training data, and have achieved many inspiring results. Meanwhile, we have investigated both mi-Net and MI-Net, and experiments have shown that MI-Net outperforms mi-Net in more cases.

Learning effective representation from (weakly-supervised) data, especially MIL, has received a lot of attention as it helps solve a range of real applications [21–25]. Till date, numerous MIL methods have been proposed to either develop effective MIL solvers or apply MIL to solve real application problems [26,27]. A comprehensive survey of MIL algorithms and applications can be found in [2]. Here, we focus on a brief review of the most recent MIL algorithms, especially the ones related to deep neural networks and feature learning. From the view of embedded-space paradigm for MIL, the most recent method is the scalable MIL algorithm, which solves MIL using Fisher Vector (FV) coding [28], called miFV [17]. miFV transforms instance feature into high-dimensional space using an unsupervised learned Gaussian Mixture Model (GMM) and FV coding. The proposed MI-Net learns instance feature using deep multiple instance supervision. In addition, MI-Net achieves better bag classification accuracy and is much faster than miFV.