ABSTRACT

Image tagging is a well known challenge in image processing. It is typically addressed through multi-instance multi-label (MIML) classification methodologies. Convolutional Neural Networks (CNNs) possess great potential to perform well on MIML tasks, since multi-level convolution and max pooling coincide with the multi-instance setting and the sharing of hidden representation may benefit multi-label modeling. However, CNNs usually require a large amount of carefully labeled data for training, which is hard to obtain in many real applications. In this paper, we propose a new approach for transferring pre-trained deep networks such as VGG16 on Imagenet to small MIML tasks. We extract features from each group of the network layers and apply multiple binary classifiers to them for multi-label prediction. Moreover, we adopt an $L_1$-norm regularized Logistic Regression ($L_1$LR) to find the most effective features for learning the multi-label classifiers. The experiment results on two most-widely used and relatively small benchmark MIML image datasets demonstrate that the proposed approach can substantially outperform the state-of-the-art algorithms, in terms of all popular performance metrics.

Index Terms—CNN, Multi-instance, Multi-label, Small dataset, Transfer Learning

1. INTRODUCTION AND RELATED WORK

Abstracting meaningful visual information from images into semantics is one of the popular research areas in image processing, namely, feature extraction. A fundamental but unsolved problem is how to automatically generate text interpretation or description of images, an ultimate goal of image understanding. Publicly accessible images are usually posted for transmitting information, so each of them is likely worth many (if not a thousand) words. If some of these words are treated as labels, then the prediction of them becomes a multi-instance learning problem. Meanwhile, each image can be viewed as a bag of local regions. If labels are assigned to a whole image but not its specific regions, i.e. an image (bag) label is positive if any of its regions (instances) has this positive label, the prediction of each image (bag) label is also naturally a multi-instance learning problem. Therefore, the so-called multi-instance multi-label (MIML) problem [1] is in fact very common in real life and of great importance for bridging images and text in multimedia researches. Figure 1 illustrates a general setting for automatic image tagging (annotation) in the MIML perspective\textsuperscript{1}. Recently, Convolutional Neural Networks (CNNs) have been successfully applied to numerous computer vision tasks, such as face recognition [2], image categorization [3], scene recognition [4], and semantic segmentation [5]. These tasks are still essentially a multi-class classification problem, where only a single label is assigned to a concerned input, not yet a MIML problem.

However, here we would like to promote CNNs further that CNNs are in fact suitable for solving MIML problems. Firstly, traditional algorithms solve multi-instance problems by extracting a bag of instances from each image and then applying multi-instance classifiers. The convolutional layers of CNNs slide through an image (the bag) and create the candidate instances, and the max-pooling layers select the most representative instances inside the bag. Secondly, traditional algorithms solve multi-label problems by explicitly or implicitly building the relationship between the labels. In CNNs, through utilizing a deep hierarchical structure, different levels of representation of all the labels have already been embedded in the network, and the relationship of the labels can be explored via checking the sharing of the neurons, for exam-

\textsuperscript{1}In some related work multi-instance learning also concerns instance-level labeling, but in this paper we focus on bag-level labeling.
2. TRANSFER LEARNING FOR MIML

To use a representative and concrete exemplar model to illustrate the transferring of CNNs, in this paper we choose the VGG16 net [6] trained on the multi-class classification task of Imagenet. VGG16 is a widely-used simple yet powerful CNN structure, enjoying good performance on Imagenet and many successful transferring records. Imagenet has been proved to be a great source for transferring as it has a wide coverage (totally 1000) of common object categories and contains great variations of them in its 1.3 million training images.

VGG16 has 13 convolutional layers and 3 Fully-Connected (FC) layers. The 13 convolutional layers form 5 groups (2-2-3-3-3) and at the end of each group one max-pooling layer is utilized. During the training process, VGG16 learns filters/kernels, which would slide through each image at multiple scales and generate nearly all possible instances (patches) inside the bag (image). Max pooling, which would add position robustness, selects the most representative instances inside the bag in fact. Therefore, the features learned via CNNs are potentially suitable for multi-instance tasks.

Meanwhile, during the training process with Imagenet, a larger number of labels have been simultaneously employed to learn a representation network. The hierarchical structure of the network builds different levels of representation of all the labels. Later layer filters, which are semantic detectors [7, 8], are in fact closely related to the semantic description space of the labels [5, 9]. Earlier layer filters would be more about the low level descriptions of the labels [10].

With the pre-trained network, different levels of semantic relationship have already been captured by the network via the sharing of neurons at different layers. Therefore, the features learned via CNNs are also suitable for multi-label tasks on small datasets. One widely adopted approach to transfer learning is fine-tuning. However, an effective fine-tuning usually also requires a large number of training instances [11] or pixel-level strongly supervised information [5]. Hence, for small MIML image tagging tasks, we introduce the following transfer learning method.

According to the previous papers such as [12, 13, 14, 15],
features from all the layers are useful for the final classification. More precisely, they extract features from all layers and then conduct 4-quadrant max-pooling for each filter of the convolution layers. Similarly, we use 4-quadrant max-pooling to extract features from the pooling layer after each group of convolution layers, as illustrated in Figure 2. The second, instead of the last, FC layer is adopted, because the second, instead of the last, FC layer is adopted, because the CNN features extracted and employed may be related to only a small subset of the Imagenet classes.

Therefore, many of the pre-trained CNN filters may not be useful for the final classification. More precisely, they extract features from all layers and then conduct 4-quadrant max-pooling for each filter of the convolution layers. Similarly, we use 4-quadrant max-pooling to extract features from the pooling layer after each group of convolution layers, as illustrated in Figure 2. The second, instead of the last, FC layer is adopted, because the CNN features extracted and employed may be related to only a small subset of the Imagenet classes.

3. EXPERIMENTS

3.1. Setting

We use VGG+L1LR (VGG+LR) to represent the transferring algorithm adopting the features extracted from VGG16 and then using L1LR (LR) to do multi-label classification. VGG+L1LR and VGG+LR are compared with other six state-of-the-art MIML methods, namely MIMLfast [17], DBA [18], KISAR [19], MIMLkNN [20], MIMLSVM [21] and RankLSIM [22].

The experiments are conducted on two popular and relatively small benchmark datasets: MSRC v2 [23], a subset of the Microsoft Research Cambridge (MSRC) image dataset with 591 images and 23 labels (on average 5 labels per image), and Scene [21], a natural scene dataset with 2000 images and 5 different labels (on average 1.2 labels per image).

The VGG features are extracted via Caffe [24]; LR and L1LR are implemented via Liblinear [25]. The shrinkage parameter of L1LR is fixed as default of 1. The performance of those state-of-the-art algorithms are based on [17], which adopts cross validation to select suitable values for the parameters of the algorithms. For each dataset, 2/3 of the instances are randomly selected to train the classifiers and the remaining 1/3 instances are taken as the test instances. We repeat the random data partition for 30 times, and report the average results over the 30 repetition.

The performance evaluation of multi-label algorithms is more complex than that of multi-class ones. We choose five
frequently adopted metrics [1] for evaluation as follows.

- Average Precision (a.p.): the average fraction of labels ranked above a particular label \( y \in Y \) which are actually in \( Y \);
- Coverage (co.): how far we need, on average, to go down the list of labels in order to cover all the proper labels of the instance\(^2\);
- Hamming Loss (h.l.): the percentage of instance-label pairs which are misclassified;
- One Error (o.e.): the average number of the top-ranked label predicted but not a label of the instance;
- Ranking Loss (r.l.): the average fraction of label pairs that are reversely ordered for the instance.

### 3.2. Results

Firstly, as illustrated in Table 1, it is clearly that both VGG+L\(_1\)LR and VGG+LR outperform the 6 state-of-the-art algorithms remarkably, which proves the proposed VGG-based transfer learning framework is suitable for MIML problems and could perform well even with a small number of training instances (394 on the dataset of MSRC and 1333 on the dataset of Scene). In terms of average precision, which is the higher the better, the VGG-based algorithms perform at least 0.16 and 0.15 better on the two datasets respectively. In terms of the other four metrics, namely coverage, hamming loss, one error and ranking loss, the VGG-based algorithms are at least 0.098, 0.033, 0.146, 0.062 smaller than the state-of-the-art algorithms on the dataset of MSRC and 0.111, 0.098, 0.237, 0.130 smaller on the dataset of Scene. The experiment results on these two datasets indicate that the proposed algorithms, which are based on transferring VGG, enjoy at least 21.8\%, 41.1\%, 33\%, 49.4\% and 57.4\% performance improvement on the five metrics, respectively.

Secondly, after comparing VGG+L\(_1\)LR against VGG+LR, the sparsity regularization also has shown noticeable positive effect. On the dataset of MSRC, we can see 0.082(9.6\%), 0.038(27.1\%), 0.018(35.29\%), 0.089(59.7\%) and 0.027(60\%) improvement on the five metrics, respectively. Similarly, 0.022(2.4\%), 0.014(14.6\%), 0.020(22.2\%), 0.032(28.1\%) and 0.027(30.9\%) improvement is observed on the dataset of Scene. The results verify the effectiveness of sparsity regularization.

Finally, the performance of the algorithms with even smaller numbers of training instances is illustrated via Figure 3. We can see that adding L\(_1\) regularization always produces better performance when the number of training instances is relatively small. On the dataset of MSRC, when comparing against the state-of-the-art algorithms, we can observe that, with only 200 training instances, VGG+L\(_1\)LR can perform better than the state-of-the-art algorithms with 394 training instances, in terms of all five metrics. Even more striking, on the dataset of Scene, it is clearly that with only 30 training instances, L\(_1\)LR could get better performance than the state-of-the-art algorithms with 1333 training instances. These results verify that the proposal also works with a relatively small number of training instances.

### 4. CONCLUSION

In this paper, for multi-instance multi-label problems with small datasets, we propose a CNN-based transfer learning framework with sparsity regularization. The proposal has achieved substantial improvement in classification performance when compared with six state-of-the-art algorithms.
5. REFERENCES


