Deep Multiple Instance Learning for Image Classification and Auto-Annotation

Jiajun Wu[1], Yinan Yu[2,3], Chang Huang[2], Kai Yu[2]
[1]Massachusetts Institute of Technology [2]Institute of Deep Learning, Baidu [3]Tsinghua University

The recent development in learning deep representations has demonstrated its wide applications in traditional vision tasks. However, there has been little investigation on how we could build up a deep learning framework in a weakly supervised setting. In this paper, we attempt to model deep learning in a weakly supervised learning (multiple instance learning) framework for the tasks of image classification and annotation. In our setting, each image follows a dual multi-instance assumption, where its object proposals and possible text annotations can be regarded as two instance sets. We thus design effective systems to exploit the MIL property with deep learning strategies from the two ends; we also try to jointly learn the relationship between object and annotation proposals.

Dual MIL assumption: Recent methods can generate a number of object proposals with very high recalls. We may therefore assume that objects lies in at least one of the proposals, *i.e.* it becomes natural to treat the object proposals of each image as a positive bag in multiple instance learning.

From a different angle, there have been many techniques for collecting keywords from the web for a given image. These keywords alone are often too noisy for tasks like image classification. However, it is justifiable to assume that there must be at least one relevant keyword within a number of most confident keywords. This again corresponds to the multiple instance assumption. These findings encourage us to design a multi-instance learning scheme to jointly learn about visual objects and verbal keywords.

DMIL with regions: Considering the recent advances achieved by deep learning, we are interested in using deep convolutional neural network for learning visual representation with MIL. The structure is inspired by [3], and contains five convolutional layers, a pooling layer, and three fully connected layers. Given one training sample x, the network extracts layer-wise representations from the first convolutional layer to the output of the last fully connected layer $fc_8 \in \mathbb{R}^m$, which can be viewed as high level features of the input image. Followed by a softmax layer, fc_8 is transformed into a probability distribution $\mathbf{p} \in \mathbb{R}^m$ for objects of m categories, and cross entropy is used to measure the prediction loss of the network. The gradients of the deep convolutional neural network is calculated via back-propagation.

For MIL, we denote $\{\mathbf{x_j}|j=1,2,\ldots,n\}$ as a bag of n instances and $t=\{t_i|t_i\in\{0,1\},i=1,\ldots,m\}$ as the label of the bag; a multiple instance convolutional neural network extracts representations of the bag: $h=\{h_{ij}\}\in R^{m\times n}$, in which each column is the representation of an instance. The aggregated representation of the bag for MIL is $\hat{h_i}=f(h_{i1},h_{i2},\ldots,h_{in})$, where function f can be $\max_j(h_{ij})$, $\text{avg}_j(h_{ij})$, or $\log\left[1+\sum_j\exp(h_{ij})\right]$, among others. The distribution of visual categories of the bag and the loss L are therefore

$$p_i=\frac{\exp(\hat{h_i})}{\sum_i\exp(\hat{h_i})} \quad \text{and} \quad L=-\sum_i t_i\log(p_i). \tag{1}$$

We employ stochastic gradient descent to minimize the loss function of the DMIL. For the $\max(\cdot)$ layer, back propagation [4] provides

$$\frac{\partial L}{\partial \hat{h_i}}=p_i-t_i \quad \text{and} \quad \frac{\partial \hat{h_i}}{\partial h_{ij}}=\begin{cases}1, & h_{ij}=\hat{h_i}\\ 0, & \text{else}\end{cases}. \tag{2}$$

For the task of image classification, we first employ existing methods to generate object proposals within each image; we then apply the deep multiple instance learning framework to perform image classification.

DMIL with keywords: Images on the web are connected with rich documents, which may describe images in more detail. For each image, we first use Baidu image search to find a set of most similar images from the web. We then crawl the surrounding documents of each retrieved image. The nouns which appear in the surrounding documents are considered as the keywords of the image. The keywords extracted from webpages are highly noisy. However, some words provide more informative descriptions than the category labels do (for instance, F-22 Raptor vs aeroplane).

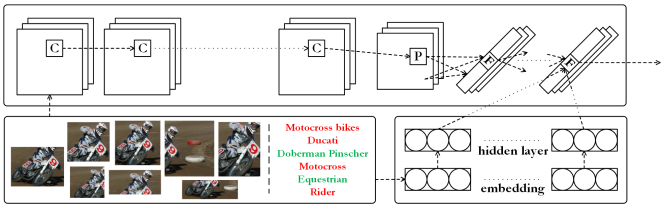

Figure 1: Our deep multiple instance learning framework for jointly learning from image regions and keywords.

Keywords gathered from the web, as a type of "noisy input", fit the multi-instance assumption well. We use another deep formulation to predict image category from keywords, which contains one input layer, one hidden layer, and one output layer with softmax. Instead of using original word indices as input, a 128-dimensional word-to-vector feature is used to relieve the computational burden.

Joint learning: Object proposals and keywords are two sets of instances satisfying the multiple instance assumption; a cross combination of the regions and the words leads to the possibility that we can label regions with proper words. We build a joint deep multi-instance learning architecture to learn the object proposals and keywords simultaneously.

Specifically, we combine the outputs of image and text understanding systems in the final fully connected layer, as illustrated in Figure 1. This can be viewed as a straightforward generalization of the aggregate equation with

$$\hat{h_i}=f\begin{pmatrix}h_{i11} & h_{i12} & \ldots & h_{i1n}\\ h_{i21} & h_{i22} & \ldots & h_{i2n}\\ \vdots & \vdots & \ddots & \vdots\\ h_{im1} & h_{im2} & \ldots & h_{imn}\end{pmatrix}, \tag{3}$$

where m is the number of keywords and n is the number of patches.

Experiments: Table 1 shows the results of our deep multiple instance learning (DMIL) system for image classification on PASCAL VOC 07 and MIT Indoor. DMIL outperforms previous efforts. Both region proposals and keyword proposals contribute to the overall performance, although regions proposals play a more central role.

Methods	mAP
GHM	64.7
AGS	71.1
NUS	70.5
CNN-SVM [5]	73.0
DMIL (region)	74.7
DMIL (keyword)	61.8
DMIL (joint)	75.5

Methods	mAcc
Object Bank	37.6
RBow	37.9
BoP	46.1
miSVM	46.4
D-Parts	51.4
MLrep	64.0
CNN-SVM [5]	57.7
DMIL (region)	60.0
DMIL (keyword)	48.3
DMIL (joint)	61.2

Table 1: Classification results on VOC 07 (left) and MIT Indoor (right).

[1] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Mid-level visual element discovery as discriminative mode seeking. In *NIPS*, 2013.

[2] Mayank Juneja, Andrea Vedaldi, CV Jawahar, and Andrew Zisserman. Blocks that shout: Distinctive parts for scene classification. In *CVPR*, 2013.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[4] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[5] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *ICLR*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.