Spatial Co-Training for Semi-Supervised Image Classification

Yi Hong
YHong@walmartlabs.com

@WalmartLabs
San Bruno, CA 94066, USA

Co-training [1] is a famous learning algorithm used when there are small amounts of labeled data and large amounts of unlabeled data, but it has limited applications in image classification due to the unavailability of two independent and sufficient representations of the images. In this paper, we propose a novel co-training algorithm in which these two independent and sufficient representations are automatically learned from the data. We call it as the spatial co-training algorithm (SCT). The main idea of SCT is to divide an image into two subregions and consider each of them as an independent representation. In SCT, the division of the image is firstly learned by an EM style algorithm on small amounts of labeled images, and finally relearned by a co-training style algorithm on many confident unlabeled images; while the classification of the image is performed jointly with the division of the image.

1 Problem definition

Our proposed algorithm is based on the bag-of-words model that represents each image as the histogram of its local image patches. In particular, the bag-of-words model performs to: (1) extract a collection of local descriptors such as SIFT [2] from the images; (2) quantize them as indexes; (3) and represent each image as the histogram of indexes of its local image patches. A lot of visual codebook learning algorithms have been proposed so far, and in this paper we employ the k-means clustering algorithm because of its simplicity and wide applications.

Given a set of labeled images $\{I_1,...,I_l\}$, we extract local sift descriptors densely from each image and express them as a matrix F:

$$F=\{f_{ij}|i=1,...,w;j=1,...,v\}, \tag{1}$$

where f_{ij} is the sift descriptor of the local image patch (i,j), w and v are the height and width of the matrix of the SIFT descriptor. To simplify our notations, we assume that all images have the same size and thus have the same values of w and v of the matrix F. Then we apply the k-means clustering algorithm to quantize these local sift descriptors into indexes and rearrange indexes from one image as a matrix C:

$$C=\{c_{ij}|i=1,...,w;j=1,...,v\}, \tag{2}$$

where $c_{ij}\in\{1,...,K\}$ is the visual code of the local image patch (i,j) and K is the codebook size. As illustrated in the introduction section, the basic idea of the proposed algorithm is to partition an image into two subregions and consider each of them as an independent representation. Now suppose that I_1^s and I_2^s are a partition of the image I that corresponds to a partition C_1^s and C_2^s of all visual codes in C:

$$C_1^s \bigcap C_2^s=\emptyset; \quad C_1^s \bigcup C_2^s=C, \tag{3}$$

we calculate the histograms of visual codes in both I_1^s and I_2^s as:

$$h_{1,k}=\frac{\sum_{(i,j)\in C_1^s}\delta(c_{ij},k)}{\sum_{(i,j)\in C_1^s}\mathbf{1}}; \quad h_{2,k}=\frac{\sum_{(i,j)\in C_2^s}\delta(c_{ij},k)}{\sum_{(i,j)\in C_2^s}\mathbf{1}}, \tag{4}$$

after normalized for $k=1,...,K$; and $\delta(a,b)=1$ if $a=b$, and $\delta(a,b)=0$ if $a\neq b$. According to (4), each division (C_1^s,C_2^s) of the image will lead to a representation pair (h_1,h_2) of the image:

$$(C_1^s,C_2^s)\Rightarrow(h_1,h_2), \tag{5}$$

where both h_1 and h_2 are the histograms with K bins. Consider the overall number of divisions of each image is around wv that is too large to be processed, in this paper we restrict the division of the image as a vertical line. In this case, the partition of C at the position d is $(C_{1:w,1:d},C_{1:w,d+1:v})$, and we simplify it as $(C_{1:d},C_{d+1:v})$ together with their histogram representation pair as $(h_{1:d},h_{d+1:v})$. Therefore, the candidate pool of all possible representation pairs is:

$$H=\{(h_{1:d},h_{d+1:v})|\ d=1,...,v-1\}. \tag{6}$$

Based on the above notations, the proposed algorithm in this paper aims to learn good divisions $\{d_1,...,d_\ell,d_{\ell+1},...,d_{\ell+u}\}$ of both labeled and unlabeled images such that the co-training algorithm using $\{(h_{i,1:d_i},h_{i,d_i+1:v})|i=1,...,\ell;\ell+1,...,\ell+u\}$ as inputs can succeed. Two points should be mentioned about the above definition: (1) different images may have different good values of d; (2) the above restriction of the division of the image may not be optimal, but our experimental results indicate that it works quite well.

2 Solutions

Fig.1 and Fig.2 are frameworks of our proposed algorithm.

```
• Input
    – A set $\{I_1,...,I_\gamma\}$ of labeled positive training images
    – A set $\{I_{\gamma+1},...,I_\ell\}$ of labeled negative training images
• Process
  (1) Set initial divisions $\{d_1,...,d_\gamma\}$ of the images
  (2) Loop:
      (a) $(svm^L,svm^R)$ ←Fix $(d_1,...,d_\gamma)$ and train two one-class SVMs
      (b) $(d_1,...,d_\gamma)$ ←Fix $(svm^L,svm^R)$ and update divisions of the images
    End loop
  (3) Determine divisions $\{d_{\gamma+1},...,d_\ell\}$ of negative training images
```

Figure 1: The EM style algorithm to learning divisions of labeled images.

```
• Input
    – A set $L$ of labeled training images
    – A set $U$ of unlabeled images
• Process
  (1) Learn divisions $D=\{d_1,...,d_\ell\}$ of labeled images in $L$ by EM style algorithm
  (2) Create a pool $U'$ of examples by choosing $u'$ examples at random from $U$
  (3) Loop for $k$ iterations
      (a) Learn divisions $\{d'_1,...,d'_{u'}\}$ of unlabeled images in $U'$
      (b) Train a classifier $F_1$ on $\{h_{1,1:d_1},...,h_{\ell,1:d_\ell}\}$
      (c) Train a classifier $F_2$ on $\{h_{1,d_1+1:v},...,h_{\ell,d_\ell+1:v}\}$
      (d) Label $p$ positive and $n$ negative on which $F_1$ is most confident
      (e) Label $p$ positive and $n$ negative on which $F_2$ is most confident
      (f) Add these self-labeled images to $L$ and their divisions to $D$
      (g) Update $(svm^L,svm^R)$ by training two one-class SVMs on augmented $D$
      (h) $\ell=\ell+2p+2n$
      (i) Randomly choose $2p+2n$ examples from $U$ to replenish $U'$
    End loop
  (4) Train a single view classifier $F$ on the augmented training dataset $L$
  (5) Label all the rest of unlabeled images by $F$
```

Figure 2: The spatial co-training algorithm (SCT).

[1] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. *COLT*, 1998.

[2] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.