This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Learning Multi-Instance Enriched Image Representations via Non-Greedy Ratio Maximization of the ℓ_1-Norm Distances

Kai Liu[†], Hua Wang[†]*, Feiping Nie[‡], Hao Zhang[†]
[†]Department of Computer Science, Colorado School of Mines, Golden, Colorado 80401, U.S.A.
[‡]School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, P. R. China
kaliu@mines.edu, huawangcs@gmail.com, feipingnie@gmail.com, hzhang@mines.edu

Abstract

Multi-instance learning (MIL) has demonstrated its usefulness in many real-world image applications in recent years. However, two critical challenges prevent one from effectively using MIL in practice. First, existing MIL methods routinely model the predictive targets using the instances of input images, but rarely utilize an input image as a whole. As a result, the useful information conveyed by the holistic representation of an input image could be potentially lost. Second, the varied numbers of the instances of the input images in a data set make it infeasible to use traditional learning models that can only deal with single-vector inputs. To tackle these two challenges, in this paper we propose a novel image representation learning method that can integrate the local patches (the instances) of an input image (the bag) and its holistic representation into one single-vector representation. Our new method first learns a projection to preserve both global and local consistencies of the instances of an input image. It then projects the holistic representation of the same image into the learned subspace for information enrichment. Taking into account the content and characterization variations in natural scenes and photos, we develop an objective that maximizes the ratio of the summations of a number of ℓ_1-norm distances, which is difficult to solve in general. To solve our objective, we derive a new efficient non-greedy iterative algorithm and rigorously prove its convergence. Promising results in extensive experiments have demonstrated improved performances of our new method that validate its effectiveness.

1. Introduction

Learning images representations plays an important role in many real-world applications due to the overwhelming amount of images and videos nowadays brought by modern technologies. Recently, image representation techniques using semi-local, or patch-based, features, such as SIFT and geometric blur, have demonstrated some of the best performance in image retrieval and object recognition applications. These algorithms choose a set of patches in an image, and for each patch compute a fixed-length feature vector. This gives a set of vectors per image, where the size of the set can vary from image to image. Armed with these patch-based features, image categorization and retrieval are recently formulated as a *multi-instance learning (MIL)* problem with improved retrieval, indexing and annotation performances [19, 3, 40, 30, 28, 31]. Under the framework of MIL, an image is viewed as a *bag*, which contains a number of *instances* corresponding to the patches in the image. For example, in the image in Figure 1 there exist a total of four patches (surrounded by the yellow bounding boxes) that represent a set of four different objects, including a car, a bicycle, and two persons (one person is riding the bicycle and the other one is standing aside). The image thereby is a bag and each of the four patches, represented as a vector, in the image is consider as an instance. If any of these instances is related to a semantic concept, the entire image will be annotated with the corresponding semantic label.

Despite a number of successes in applying MIL in image learning, there exist two critical challenges that prevent one from effectively using available visual information in an image as much as possible.

First, most, if not all, existing MIL methods only use the instances of an input image to model the semantic concepts, but not the entire image. For example, when a MIL method [19, 3, 40, 30, 28, 31, 24, 35, 38] is used to study the image in Figure 1, only the four instances are associated with some semantic concepts. However, these four patches only occupy a small percent of the area of the entire image, while the remained areas of the image that are outside of the yellow bounding boxes are completely dis-

*Corresponding author. This work was partially supported by NSF-IIS 1423591 and NSF-IIS 1652943.

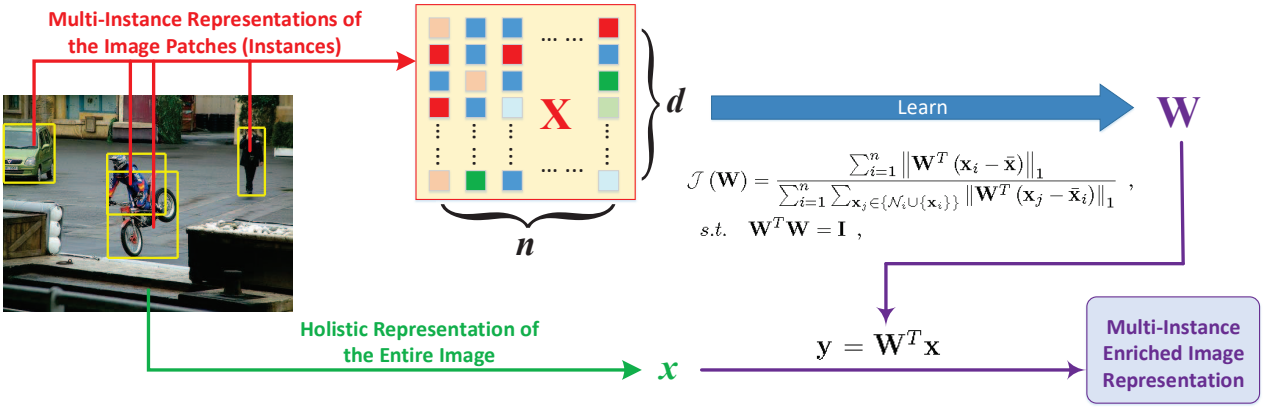

Figure 1. Illustration of the proposed method to learn a multi-instance enriched image representation in the ***single-vector*** format.

carded, which, though, could potentially convey valuable semantic information. For instance, a large area of this picture is "ground", which is closely correlated to "automobiles" and "bicycles" and could be used to improve the image categorization accuracy [25, 26, 27]. Indeed, recent studies have already demonstrated that holistic image representations based upon global features are a necessity to decode scene contents [10, 37]. Therefore, it is desirable to learn an image representation that is able to capture both instance-wise and holistic information of an input image.

Second, in MIL an image is represented as a set of vectors and the numbers of the vectors in the images of a data set are different in general. Although the *multiple-vector representation* could describe the image details with better granularity, varied data sizes make it infeasible to use traditional machine learning models that can only deal with data represented by single-vectors, *i.e.*, one vector per data sample. Therefore, it would be beneficial to learn a ***single-vector representation*** for an image that can integrate the information from both its instances and its entire context.

To address the above two challenges in multi-instance image learning, in this paper we present a novel image representation learning method. It first learns a projection from the instances of an input image. Then it projects its holistic image representation into the learned subspace. A schematic illustration of our new method is shown in Figure 1. Through these procedures, the learned image representation simultaneously captures the information from both multi-instance image patches and the holistic summarization of the entire image. In the proposed objective to learn the projection from the instances of an input image, we aim to preserve both global and local consistencies of the instances in the projected subspace, which leads to an optimization problem that maximizes the ratio of matrix traces. By further recognizing the variations of the content and visual characterization in natural scenes and photos, we further develop the proposed objective by replacing the squared ℓ_2-norm distances by the ℓ_1-norm distances in our formulation, such that the robustness of the learned image representations against outlying samples and features is promoted [2, 4, 9, 15, 16, 36, 20, 34].

Despite its clear motivations to integrate the information of an input image from both its local instances and its holistic representation, the proposed objective ends up to be an optimization problem that simultaneously maximizes and minimizes the summations of a number of ℓ_1-norm distances, which is difficult to solve in general. To solve this challenging optimization problem, we derived an efficient iterative algorithm with theoretically guaranteed convergence. It is worth noting that, different from many previous works, our new solution algorithm is a ***non-greedy*** algorithm, such that it has better chance to find the global optima. To the best of our knowledge, our new algorithm solves the general optimization problem that maximizes the ratio of the summations of the ℓ_1-norm distances in a non-greedy way for the first time in literature, which can find many applications to improve a number of machine learning models.

Finally, we performed extensive experiments on three benchmark multi-instance image data sets, the promising experimental results have demonstrated the effectiveness of our new method in image learning applications.

2. Learning multi-instance enriched image representations in the single-vector format

In this section, first we formalize the representation learning problem for images with semantic patches, where we introduce the notations used in this paper. Then we gradually develop the proposed objective to learn a single-vector representation of an input image that captures both global and local consistencies of the image patches and integrates the holistic information conveyed by the entire image.

2.1. Notations and problem formalization

Throughout this paper, we write matrices as bold uppercase letters and vectors as bold lowercase letters. The trace

of the matrix $\mathbf{M} = [m_{ij}]$ is defined as $\mathbf{tr}(\mathbf{M}) = \sum_i m_{ii}$, and the ℓ_1-norm of $\mathbf{M}$ is defined as $\|\mathbf{M}\|_1 = \sum_i \sum_j |m_{ij}|$. The ℓ_1-norm of a vector $\mathbf{v}$ is defined as $\|\mathbf{v}\|_1 = \sum_i |v_i|$ and the ℓ_2-norm of $\mathbf{v}$ is defined as $\|\mathbf{v}\|_2 = \sqrt{\sum_i v_i^2}$.

In image retrieval and annotation tasks, we study a set of images and every image contains a collection of semantically meaningful patches. For a given image, we represent it as $\mathcal{X} = \{\mathbf{x}, \mathbf{X}\}$, where $\mathbf{x} \in \Re^d$ denotes the holistic representation of the entire image and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \Re^{d \times n}$ denotes a collection of n semantic patches, respectively. Here $\mathbf{x}_i \in \Re^d$ (for $i = 1, 2, \ldots, n$) represents a patch of the input image, which can be illustrated as a yellow box in the image in Figure 1. Under this framework, every image is considered as a bag of instances (patches). In general, the numbers of the semantic patches in the images of a data are different from one to another. In order to tackle the two critical challenges analyzed before in Section 1, different from existing MIL studies that model the associations between the instances of input images and the predictive targets directly, in this paper we aim to learn from an input image $\mathcal{X}$ a ***single-vector representation*** of $\mathbf{y} = f(\mathcal{X})$ that captures the information in both local patches and the entire image. Because the new representations of the images in a data set are of the same length, they can be readily used by traditional learning models in various image learning tasks. In the following, we use *instance* and *image patches* interchangeably when there is no risk of ambiguity.

2.2. Our objective

In this subsection, we will develop the proposed objective to learn a new single-vector representation for an input image from its holistic representation and its semantic instances. When we integrate the holistic representation and the semantic instances of the input image, we aim to preserve both global and local consistencies among the semantic instances in the learned projected subspace.

Learning with global consistency via PCA. With recent advances in digital imaging techniques, one can easily have a camera with very high resolution. As a result, the derived visual descriptors from a raw picture are usually of high dimensionality. When the image dimensionality grows, most image retrieval and annotation methods will fail due to "the curse of dimensionality" [8] and intractable computational costs. Thus, learning a lower-dimensional image representation while maintaining the original geometrical structures of the input image is valuable for practical use. To achieve this goal, principal component analysis (PCA) [14] is the right tool to preserve as much information as possible by learning a projection $\mathbf{W} \in \Re^{d \times r}$ (usually $r \ll d$) from the semantic instances $\mathbf{X}$ of the input image $\mathcal{X}$, which maps $\mathbf{x}_i$ in the high d-dimensional space into a vector $\mathbf{y}_i$ in a lower r-dimensional space by computing $\mathbf{y}_i = \mathbf{W}^T \mathbf{x}_i$, such that the overall variance of the input data in the projected space $\Re^r$ is maximized. Formally, let the global mean vector of the input data $\mathbf{X}$ as $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$, PCA seeks the projection $\mathbf{W}$ by maximizing the following objective:

$$\mathcal{J}_{\text{Global}}(\mathbf{W}) = \mathbf{tr}\left(\mathbf{W}^T \mathbf{S}_G \mathbf{W}\right) = \sum_{i=1}^n \left\|\mathbf{W}^T (\mathbf{x}_i - \bar{\mathbf{x}})\right\|_2^2 \ , \quad s.t. \quad \mathbf{W}^T \mathbf{W} = \mathbf{I} \ , \tag{1}$$

where $\mathbf{S}_G = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$ is the covariance matrix of $\mathbf{X}$ and the constant factor $\frac{1}{n}$ is omitted for brevity. Because $\mathcal{J}_{\text{Global}}$ maximizes the global variance of the instances of the input image in the projected subspace, the projected instances via the learned projection $\mathbf{W}$ are globally consistent in terms of information preservation.

Learning with local consistency via neighborhood variances. Besides taking advantage of the global consistency of the semantic instances of the input image, we further take into account the local geometric structures of these semantic instances in the projected subspace and consider their local consistency. Ideally, in the learned subspace the instances with similar semantic labels should be close to each other, while those with different semantic labels should be far away from each other. In other words, in contrast to maximizing the projected global variance, we also want to minimize the local variance around every instance in the projected subspace. Mathematically, denoting the K-nearest neighbors of $\mathbf{x}_i$ as $\mathcal{N}_i$ and the local mean vector of $\mathbf{x}_i$ as $\bar{\mathbf{x}}_i = \frac{1}{K+1} \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} \mathbf{x}_j$, we can achieve the overall local consistency by minimizing the following objective:

$$\mathcal{J}_{\text{Local}}(\mathbf{W}) = \mathbf{tr}\left(\mathbf{W}^T \mathbf{S}_L \mathbf{W}\right) \quad s.t.\ \mathbf{W}^T \mathbf{W} = \mathbf{I} \ , \tag{2}$$

where, following our previous work [33], we define:

$$\mathbf{S}_L = \sum_{i=1}^n \mathbf{S}_{Li};\ \mathbf{S}_{Li} = \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} (\mathbf{x}_j - \bar{\mathbf{x}}_i)(\mathbf{x}_j - \bar{\mathbf{x}}_i)^T \ .$$

Obviously, $\mathbf{S}_{Li}$ computes the local covariance matrix of the data points around $\mathbf{x}_i$. Thus minimizing $\mathbf{tr}\left(\mathbf{W}^T \mathbf{S}_{Li} \mathbf{W}\right)$ ensures the local consistency around $\mathbf{x}_i$ and minimizing $\mathcal{J}_{\text{Local}}$ in Eq. (2) ensures the overall local consistency around all the instances in a bag. Here, again the constant factor $\frac{1}{K+1}$ is omitted for brevity.

Our objective to integrate the global and local consistencies of the semantic instances. Armed with the objectives that can capture the global and local consistencies of the semantic instances of an input image separately, we can develop a combined objective to capture both of them simultaneously. Among several possible ways to combine the two objectives in Eqs. (1-2), we can formulate our new objective using the trace ratio of matrices [12], which maximizes the

following objective:

$$\mathcal{J}_{\ell_2}(\mathbf{W}) = \frac{\mathbf{tr}\left(\mathbf{W}^T\mathbf{S}_G\mathbf{W}\right)}{\mathbf{tr}\left(\mathbf{W}^T\mathbf{S}_L\mathbf{W}\right)} = \frac{\sum_{i=1}^{n}\left\|\mathbf{W}^T\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)\right\|_2^2}{\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\{\mathcal{N}_i\cup\{\mathbf{x}_i\}\}}\left\|\mathbf{W}^T\left(\mathbf{x}_j-\bar{\mathbf{x}}_i\right)\right\|_2^2} \ , \quad s.t. \quad \mathbf{W}^T\mathbf{W}=\mathbf{I} \ . \tag{3}$$

A critical problem of $\mathcal{J}_{\ell_2}$ in Eq. (3) lies in that it computes the ratio of the summations of a number of squared ℓ_2-norm distances, which are notoriously known to be sensitive to both outlying sample and outlying features [4, 34]. Many images from natural scenes and photos often have clustered objects. This is particularly true when there exist a crowd of people in a picture, where each individual people may not characterize the the semantic category of "person" appropriately and many instances have to be considered as outlying samples. Similarly, due to cropped objects and (partially) shaded objects in pictures, such as the car in the image in Figure 1, outlying features also inevitably exist in real image data sets. Following many previous works [2, 4, 9, 15, 16, 36, 29, 32, 21, 34], to deal with the feature and content variances in natural images, we propose to learn the projection by maximizing the following objective:

$$\mathcal{J}(\mathbf{W}) = \frac{\sum_{i=1}^{n}\left\|\mathbf{W}^T\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)\right\|_1}{\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\{\mathcal{N}_i\cup\{\mathbf{x}_i\}\}}\left\|\mathbf{W}^T\left(\mathbf{x}_j-\bar{\mathbf{x}}_i\right)\right\|_1} \ , \quad s.t. \quad \mathbf{W}^T\mathbf{W}=\mathbf{I} \ , \tag{4}$$

in which we compute the summations of the ℓ_1-norm distances, because the ℓ_1-norm distance can promote the robustness against both outlier samples and outlier features.

Upon solving the optimization problem in Eq. (4), the learned $\mathbf{W}$ not only preserves the global variance of the semantic instances of an input image, but also rewards the local geometric structures of the semantic instances, which thereby is both globally and locally consistent in the learned subspace. Then we enrich the holistic representation $\mathbf{x}$ of the input image $\mathcal{X}$ by computing $\mathbf{y} = \mathbf{W}^T\mathbf{x}$, which is a fixed-length single-vector representation and can be readily used by any traditional single-instance machine learning models. This indeed is the main contribution of this paper.

3. An efficient solution algorithm

Our new objective in Eq. (4) maximizes the ratio of the summations of a number of the ℓ_1-norm distances, which is obviously not smooth and thus difficult to solve in general. To solve the general problem that maximizes the ratio of the summations of the ℓ_1-norm distances, such as our objective in Eq. (4), in this section we will derive an efficient iterative algorithm that is non-greedy. We will also prove the convergence of our new solution algorithm.

3.1. Solving a general ratio maximization problem

We first generalize the objectives in Eq. (3) and Eq. (4) into the following general optimization problem and then derive its solution algorithm:

$$v_{\text{opt}} = \underset{v\in\Omega}{\operatorname{argmax}} \frac{h(v)}{m(v)} \ , \quad \forall v \in \Omega \quad \begin{cases} C_2 \geq m(v) \geq C_1 > 0 \ , \\ C_4 \geq h(v) \geq C_3 > 0 \ . \end{cases} \tag{5}$$

where Ω is the feasible domain.

Motivated by our previous works [34, 33, 24], we propose a simple, yet efficient, iterative framework in Algorithm 1 to solve the objective in Eq. (5), whose convergence is rigorously guaranteed by Theorems 1.

Algorithm 1: Algorithm to solve Eq. (5).

1. Randomly initialize $v^0 \in \Omega$ and set $k = 1$.
while *not converge* **do**
  2. Calculate $\lambda^k = \frac{h(v^{k-1})}{m(v^{k-1})}$.
  3. Find a $v^k \in \Omega$ satisfying $h(v^k) - \lambda^k m(v^k) > h(v^{k-1}) - \lambda^k m(v^{k-1}) = 0$.
  4. $k = k + 1$.
Output: v.

Theorem 1. *In Algorithm 1, for each iteration we have (1)* $\frac{h(v^k)}{m(v^k)} \geq \frac{h(v^{k-1})}{m(v^{k-1})}$*; and (2)* $\forall\delta$*, there exists a* $\hat{k}$ *such that* $\forall k > \hat{k}$ $\frac{h(v^k)}{m(v^k)} - \frac{h(v^{k-1})}{m(v^{k-1})} < \delta$.

Proof. In Algorithm 1, from step 3, we have $h(v^k) - \lambda^k m(v^k) > 0$. Because $\forall v \in \Omega \ \ m(v) > 0$, we can get $\frac{h(v^k)}{m(v^k)} > \lambda^k = \frac{h(v^{k-1})}{m(v^{k-1})}$, which completes the proof of the first statement of Theorem 1.

Suppose that for the k-th iteration, there exists a c^k such that $h(v^k) - \lambda^k m(v^k) = c^k > 0$. We have:

$$\frac{h(v^k)}{m(v^k)} = \frac{h(v^{k-1})}{m(v^{k-1})} + \frac{c_k}{m(v^k)} \ , \tag{6}$$

by which we can derive:

$$\frac{h(v^k)}{m(v^k)} = \frac{h(v^0)}{m(v^0)} + \sum_{i=1}^{k}\frac{c^i}{m(v^i)} \ . \tag{7}$$

From Eq. (7), we can derive:

$$\frac{h(v^0)}{m(v^0)} + \frac{1}{C_2}\sum_{i=1}^{k}c^i \leq \frac{h(v^k)}{m(v^k)} \leq \frac{h(v^0)}{m(v^0)} + \frac{1}{C_1}\sum_{i=1}^{k}c^i \ . \tag{8}$$

Suppose that there exist a positive constant C such that $\lim_{k\to\infty}\sum_{i=1}^{k}c^i = C$. If this is not true, we have

$\lim_{k\to\infty}\sum_{i=1}^{k} c^i = \infty$, by which, together with Eq. (8), we can derive $\lim_{k\to\infty}\sum_{i=1}^{k}\frac{h(v^k)}{m(v^k)} = \infty$. This, however, contradicts the fact that $\frac{h(v^k)}{m(v^k)}$ is bounded as defined in Eq. (5), which means that $\lim_{k\to\infty}\sum_{i=1}^{k} c^i = C$ holds. Thus, we have $\lim_{k\to\infty} c^k = 0$, *i.e.*, $\lim_{k\to\infty}\frac{c^k}{m(v^k)} = 0$, which indicates that $\forall\delta > 0$, there must exist a $\hat{k}$ such that:

$$\forall k > \hat{k} \quad \frac{c^k}{m(v^k)} < \delta \ , \tag{9}$$

by which and Eq. (6), we have:

$$\forall k > \hat{k} \quad \frac{h(v^k)}{m(v^k)} - \frac{h(v^{k-1})}{m(v^{k-1})} < \delta \ , \tag{10}$$

which indicates that Algorithm 1 converges to a local optimum and completes the proof of the second statement of Theorem 1. □

3.2. Our algorithm to solve the objective in Eq. (4)

To solve our objective in Eq. (4), according to Step 3 in Algorithm 1, we need find a solution that satisfy the constraint of $\mathbf{W}^T\mathbf{W} = \mathbf{I}$ and the following inequality:

$$F(\mathbf{W}) = H(\mathbf{W}) - \lambda^k M(\mathbf{W}) > 0 \ , \tag{11}$$

where λ^k is computed by

$$\lambda^k = \frac{\sum_{i=1}^{n}\left\|(\mathbf{W}^{k-1})^T(\mathbf{x}_i - \bar{\mathbf{x}})\right\|_1}{\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\{\mathcal{N}_i\cup\{\mathbf{x}_i\}\}}\left\|(\mathbf{W}^{k-1})^T(\mathbf{x}_j - \bar{\mathbf{x}}_i))\right\|_1} \ , \tag{12}$$

and $\mathbf{W}^{k-1}$ denotes the projection matrix in the $(k-1)$-th iteration, which is already known in the k-th iteration. Here, for notation brevity, we define:

$$H(\mathbf{W}) = \sum_{i=1}^{n}\left\|\mathbf{W}^T(\mathbf{x}_i - \bar{\mathbf{x}})\right\|_1 \ , \tag{13}$$

$$M(\mathbf{W}) = \sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\{\mathcal{N}_i\cup\{\mathbf{x}_i\}\}}\left\|\mathbf{W}^T(\mathbf{x}_j - \bar{\mathbf{x}}_i)\right\|_1 \ . \tag{14}$$

Now we need solve the problem in Eq. (11), for which we first introduce the following two lemmas.

Lemma 1. *[18, Theorem 1] For any vector $\boldsymbol{\xi} = [\xi_1, \cdots, \xi_m]^T \in \Re^m$, we have $\|\boldsymbol{\xi}\|_1 = \max_{\boldsymbol{\eta}\in\Re^m}(\mathrm{sign}(\boldsymbol{\eta}))^T\boldsymbol{\xi}$, where the maximum value is attained if and only if $\boldsymbol{\eta} = a\times\boldsymbol{\xi}$, where $a > 0$ is a scalar.*

Lemma 2. *[11, Lemma 3.1] For any vector $\boldsymbol{\xi} = [\xi_1, \cdots, \xi_m]^T \in \Re^m$, we have $\|\boldsymbol{\xi}\|_1 = \min_{\boldsymbol{\eta}\in\Re^m_+}\frac{1}{2}\sum_{i=1}^{m}\frac{\xi_i^2}{\eta_i} + \frac{1}{2}\|\boldsymbol{\eta}\|_1$, where the minimum value is attained if and only if $\eta_j = |\xi_j|, j \in \{1, 2, \cdots, m\}$.*

First, motivated by Lemma 1 and Lemma 2, we construct the following objective:

$$L(\mathbf{W}, \mathbf{W}^{k-1}) = K(\mathbf{W}) - \lambda^k N(\mathbf{W}) \ , \tag{15}$$

where $K(\mathbf{W})$ and $N(\mathbf{W})$ are defined as:

$$K(\mathbf{W}) = \sum_{g=1}^{r}\mathbf{w}_g^T\mathbf{B}\,\mathrm{sign}\left(\mathbf{B}^T\mathbf{w}_g^{k-1}\right) \ , \tag{16}$$

$$N(\mathbf{W}) = \frac{1}{2}\sum_{g=1}^{r}\mathbf{w}_g^T\mathbf{A}_g\mathbf{w}_g + \left(\mathbf{w}_g^{k-1}\right)^T\mathbf{A}_g\mathbf{w}_g^{k-1} \ . \tag{17}$$

Here $\mathbf{w}_g$ and $\mathbf{w}_g^{k-1}$ denote the g-th column of matrices $\mathbf{W}$ and $\mathbf{W}^{k-1}$, respectively; $\mathbf{B}$ and $\mathbf{A}_g$ for $g = 1, 2, \cdots, r$ are defined as follows:

$$\mathbf{B} = [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}, \bar{\mathbf{x}}_2 - \bar{\mathbf{x}}, \cdots, \bar{\mathbf{x}}_n - \bar{\mathbf{x}}] \ , \tag{18}$$

$$\mathbf{A}_g = \sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\{\mathcal{N}_i\cup\{\mathbf{x}_i\}\}}\frac{(\mathbf{x}_j - \bar{\mathbf{x}}_i)(\mathbf{x}_j - \bar{\mathbf{x}}_i)^T}{\left|\left(\mathbf{w}_g^{k-1}\right)^T(\mathbf{x}_j - \bar{\mathbf{x}}_i)\right|} \ , \tag{19}$$

and $\mathrm{sign}(x)$ is the sign function.

Then, using the definition of $L(\mathbf{W}, \mathbf{W}^{k-1})$ in Eq. (15) and Lemmas 1—2, we can prove the following theorem.

Theorem 2. *For any $\mathbf{W} \in \Re^{d\times r}$, we have*

$$L(\mathbf{W}, \mathbf{W}^{k-1}) \le F(\mathbf{W}) \ . \tag{20}$$

The equality holds on if and only if $\mathbf{W} = \mathbf{W}^{k-1}$.

Proof. First, according to Lemma 1 we can compute:

$$\begin{aligned} H(\mathbf{W}) &= \sum_{i=1}^{n}\left\|\mathbf{W}^T(\mathbf{x}_i - \bar{\mathbf{x}})\right\|_1 \\ &= \sum_{i=1}^{n}\sum_{g=1}^{r}\left\|\mathbf{w}_g^T(\mathbf{x}_i - \bar{\mathbf{x}})\right\|_1 \\ &\ge \sum_{g=1}^{r}\sum_{i=1}^{n}\mathrm{sign}\left[(\mathbf{w}_g^{k-1})^T(\mathbf{x}_i - \bar{\mathbf{x}})\right]\left[\mathbf{w}_g^T(\mathbf{x}_i - \bar{\mathbf{x}})\right] \\ &= \sum_{g=1}^{r}\mathbf{w}_g^T\mathbf{B}\,\mathrm{sign}\left(\mathbf{B}^T\mathbf{w}_g^{k-1}\right) = K(\mathbf{W}) \ . \end{aligned} \tag{21}$$

Then, according to Lemma 2 we have:

$$\begin{aligned} &\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\{\mathcal{N}_i\cup\{\mathbf{x}_i\}\}}\left\{\frac{1}{2}\frac{\boldsymbol{\xi}^T(\mathbf{x}_j - \bar{\mathbf{x}}_i)(\mathbf{x}_j - \bar{\mathbf{x}}_i)^T\boldsymbol{\xi}}{\boldsymbol{\xi}^T(\mathbf{x}_j - \bar{\mathbf{x}}_i)} + \frac{1}{2}\left\|\boldsymbol{\xi}^T(\mathbf{x}_j - \bar{\mathbf{x}}_i)\right\|_1\right\} \\ &\le \\ &\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\{\mathcal{N}_i\cup\{\mathbf{x}_i\}\}}\left\{\frac{1}{2}\frac{\boldsymbol{\xi}^T(\mathbf{x}_j - \bar{\mathbf{x}}_i)(\mathbf{x}_j - \bar{\mathbf{x}}_i)^T\boldsymbol{\xi}}{\boldsymbol{\eta}^T(\mathbf{x}_j - \bar{\mathbf{x}}_i)} + \frac{1}{2}\left\|\boldsymbol{\eta}^T(\mathbf{x}_j - \bar{\mathbf{x}}_i)\right\|_1\right\} \ , \end{aligned} \tag{22}$$

which indicates that:

$$M(\mathbf{W}) = \sum_{i=1}^{n} \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} \left\| \mathbf{W}^T (\mathbf{x}_j - \bar{\mathbf{x}}_i) \right\|_1 \tag{23}$$

$$= \sum_{g=1}^{r} \sum_{i=1}^{n} \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} \left\{ \frac{1}{2} \frac{\mathbf{w}_g^T (\mathbf{x}_j - \bar{\mathbf{x}}_i) (\mathbf{x}_j - \bar{\mathbf{x}}_i)^T \mathbf{w}_g}{\mathbf{w}_g^T (\mathbf{x}_j - \bar{\mathbf{x}}_i)} + \frac{1}{2} \left\| \mathbf{w}_g^T (\mathbf{x}_j - \bar{\mathbf{x}}_i) \right\|_1 \right\}$$

$$\leq \sum_{g=1}^{r} \sum_{i=1}^{n} \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} \left\{ \frac{1}{2} \frac{\mathbf{w}_g^T (\mathbf{x}_j - \bar{\mathbf{x}}_i) (\mathbf{x}_j - \bar{\mathbf{x}}_i)^T \mathbf{w}_g}{\left(\mathbf{w}_g^{k-1}\right)^T (\mathbf{x}_j - \bar{\mathbf{x}}_i)} + \frac{1}{2} \left\| \left(\mathbf{w}_g^{k-1}\right)^T (\mathbf{x}_j - \bar{\mathbf{x}}_i) \right\|_1 \right\}$$

$$= \frac{1}{2} \sum_{g=1}^{r} \mathbf{w}_g^T \mathbf{A}_g \mathbf{w}_g + \left(\mathbf{w}_g^{k-1}\right)^T \mathbf{A}_g \mathbf{w}_g^{k-1} = N(\mathbf{W}) \ .$$

Combining Eq. (21) and Eq. (23), we can derive:

$$\begin{aligned} L(\mathbf{W}, \mathbf{W}^{k-1}) &= K(\mathbf{W}) - \lambda^k N(\mathbf{W}) \\ &\leq H(\mathbf{W}) - \lambda^k M(\mathbf{W}) = F(\mathbf{W}) \ . \end{aligned} \tag{24}$$

According to Lemma 1 and Lemma 2, it is easy to verify that equality holds in Eq. (21) and Eq. (23) if and only if $\mathbf{W} = \mathbf{W}^{k-1}$. Thus, equality holds in Eq. (24) if and only if $\mathbf{W} = \mathbf{W}^{k-1}$. This completes the proof of Theorem 2. □

Now we continue to solve our objective. Let $\mathbf{W} = \mathbf{W}^{k-1}$, by substituting it into the objective, we have:

$$L(\mathbf{W}^{k-1}, \mathbf{W}^{k-1}) = F(\mathbf{W}^{k-1}) = 0 \ . \tag{25}$$

In the k-th iteration in solving the objective in Eq. (4), $\mathbf{W}^\star$ satisfies:

$$L(\mathbf{W}^\star, \mathbf{W}^{k-1}) \geq L(\mathbf{W}^{k-1}, \mathbf{W}^{k-1}) = 0 \ . \tag{26}$$

Then, we have:

$$\begin{aligned} F(\mathbf{W}^\star) &\geq L(\mathbf{W}^\star, \mathbf{W}^{k-1}) \\ &\geq L(\mathbf{W}^{k-1}, \mathbf{W}^{k-1}) = F(\mathbf{W}^{k-1}) = 0 \ . \end{aligned} \tag{27}$$

Theorem 2 and Eq. (27) indicate that the solution of the objective function in Eq. (11) can be transformed to solve the objective function $L(\mathbf{W}, \mathbf{W}^{k-1}) \geq 0$, which can be easily solved by the projected subgradient method with Armijo line search [23]. The subgradient of $L(\mathbf{W}, \mathbf{W}^{k-1})$ at $\mathbf{W}$ is computed as:

$$\begin{aligned} \partial L(\mathbf{W}, \mathbf{W}^{k-1}) = \mathbf{B} \operatorname{sign}\left(\mathbf{B}^T \mathbf{W}^{k-1}\right) \\ - \lambda^k \left[\mathbf{A}_1 \mathbf{w}_1, \mathbf{A}_2 \mathbf{w}_2, \cdots, \mathbf{A}_r \mathbf{w}_r\right] \ . \end{aligned} \tag{28}$$

Note that, for any matrix $\mathbf{W}$ the operator $P(\mathbf{W}) = \mathbf{W}\left(\mathbf{W}^T \mathbf{W}\right)^{-\frac{1}{2}}$ can project it onto an orthogonal cone.

Algorithm 2: Algorithm to maximize $F(\mathbf{W})$.

Input: $\mathbf{W}^{k-1}$ and Armijo parameter $0 < \beta < 1$.
1. Calculate λ^k by Eq. (12) the subgradient $\mathbf{G}^{k-1} = \partial L(\mathbf{W}^{k-1}, \mathbf{W}^{k-1})$ by Eq. (28) and set $m = 1$.
while *not* $F(\mathbf{W}^k) > F(\mathbf{W}^{k-1}) = 0$ **do**
  2. Calculate $\mathbf{W}^k = P(\mathbf{W}^{k-1} + \beta^m \mathbf{G}^{k-1})$.
  3. Calculate $F(\mathbf{W}^k)$ by Eq. (11).
  4. $m = m + 1$.
Output: $\mathbf{W}^k$.

This guarantees the orthogonality constraint of the projection matrix, *i.e.* $\left(\mathbf{W}^k\right)^T \left(\mathbf{W}^k\right) = \mathbf{I}$. Algorithm 2 summarizes the algorithm to maximize $F(\mathbf{W})$ in Eq. (11).

Finally, based on Algorithm 2, we can derive a simple yet efficient iterative algorithm as summarized in Algorithm 3 to solve ratio maximization problem for the ℓ_1-norm distances, *i.e.*, our objective in Eq. (4).

Algorithm 3: Algorithm for non-greedy ratio maximization of the ℓ_1-norm distances.

1. Randomly initialize $\mathbf{W}^0$ satisfying $\left(\mathbf{W}^0\right)^T \mathbf{W}^0 = \mathbf{I}$ and set $k = 1$.
while *not converge* **do**
  2. Calculate λ^k by Eq. (12).
  3. Find a $\mathbf{W}^k$ satisfying $F(\mathbf{W}^k) > F(\mathbf{W}^{k-1}) = 0$ by Algorithm 2.
  4. $k = k + 1$.
Output: $\mathbf{W}$.

3.3. Convergence analysis of our algorithm

Theorem 3. *If $\mathbf{W}^k$ is the solution of the objective function in Eq. (11) and satisfies $\left(\mathbf{W}^k\right)^T \left(\mathbf{W}^k\right) = \mathbf{I}$, then we have $\mathcal{J}(\mathbf{W}^k) \geq \mathcal{J}(\mathbf{W}^{k-1})$.*

Proof. Since $\mathbf{W}^k$ is the solution of the objective function in Eq. (11), we have

$$\begin{aligned} F(\mathbf{W}^k) &= \sum_{i=1}^{n} \left\| \left(\mathbf{W}^k\right)^T (\mathbf{x}_i - \bar{\mathbf{x}}) \right\|_1 \\ &\quad - \lambda^k \sum_{i=1}^{n} \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} \left\| \left(\mathbf{W}^k\right)^T (\mathbf{x}_j - \bar{\mathbf{x}}_i)) \right\|_1 \\ &\geq 0 \ , \end{aligned} \tag{29}$$

from which we can easily derive:

$$\begin{aligned} \mathcal{J}(\mathbf{W}^k) &= \frac{\sum_{i=1}^{n} \left\| \left(\mathbf{W}^k\right)^T (\mathbf{x}_i - \bar{\mathbf{x}}) \right\|_1}{\sum_{i=1}^{n} \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} \left\| \left(\mathbf{W}^k\right)^T (\mathbf{x}_j - \bar{\mathbf{x}}_i)) \right\|_1} \\ &\geq \lambda^k \ . \end{aligned} \tag{30}$$

Now by substituting Eq. (12) into Eq. (30), we have

$$\begin{aligned}
&\mathcal{J}(\mathbf{W}^k) \\
&= \frac{\sum_{i=1}^{n} \left\| \left(\mathbf{W}^k\right)^T (\mathbf{x}_i - \bar{\mathbf{x}}) \right\|_1}{\sum_{i=1}^{n} \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} \left\| \left(\mathbf{W}^k\right)^T (\mathbf{x}_j - \bar{\mathbf{x}}_i)) \right\|_1} \\
&\geq \frac{\sum_{i=1}^{n} \left\| (\mathbf{W}^{k-1})^T (\mathbf{x}_i - \bar{\mathbf{x}}) \right\|_1}{\sum_{i=1}^{n} \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} \left\| (\mathbf{W}^{k-1})^T (\mathbf{x}_j - \bar{\mathbf{x}}_i)) \right\|_1} \\
&= \mathcal{J}(\mathbf{W}^{k-1}) \ , \qquad (31)
\end{aligned}$$

which completes the proof of Theorem 3. □

Theorem 4. *The objective in Eq.* (4) *is upper bounded.*

Proof. First, using Cauchy-Schwarz inequality we have the following for the numerator of our objective in Eq. (4):

$$\begin{aligned}
&\sum_{i=1}^{n} \left\| \mathbf{W}^T (\mathbf{x}_i - \bar{\mathbf{x}}) \right\|_1 = \sum_{i=1}^{n} \sum_{j=1}^{r} \left\| \mathbf{w}_j^T (\mathbf{x}_i - \bar{\mathbf{x}}) \right\|_1 \qquad (32) \\
&\leq \sum_{i=1}^{n} \sum_{j=1}^{r} \left\| \mathbf{w}_j^T \right\|_2 \left\| (\mathbf{x}_i - \bar{\mathbf{x}}) \right\|_2 = \sum_{i=1}^{n} r \left\| (\mathbf{x}_i - \bar{\mathbf{x}}) \right\|_2 \ .
\end{aligned}$$

Obviously, given an input data set, $\sum_{i=1}^{n} r \left\| (\mathbf{x}_i - \bar{\mathbf{x}}) \right\|_2$ is a constant, which indicates that the numerator of our objective in Eq. (4) is upper bounded for a given data set.

Second, it can be verified that $\sqrt{\sum_{i=1}^{n} v_i^2} \leq \sum_{i=1}^{n} |v_i|$, *i.e.*, $\forall \mathbf{v} \in \Re^n \ \left\| \mathbf{v} \right\|_2 \leq \left\| \mathbf{v} \right\|_1$, by which we can derive the following for the denominator of our objective in Eq. (4):

$$\begin{aligned}
&\sum_{i=1}^{n} \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} \left\| \mathbf{W}^T (\mathbf{x}_j - \bar{\mathbf{x}}_i) \right\|_1 \\
&\geq \sum_{i=1}^{n} \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} \sqrt{\left\| \mathbf{W}^T (\mathbf{x}_j - \bar{\mathbf{x}}_i) \right\|_2^2} \\
&\geq \sqrt{\sum_{i=1}^{n} \sum_{\mathbf{x}_j \in \{\mathcal{N}_i \cup \{\mathbf{x}_i\}\}} \left\| \mathbf{W}^T (\mathbf{x}_j - \bar{\mathbf{x}}_i) \right\|_2^2} \qquad (33) \\
&= \sqrt{\mathbf{tr}\left(\mathbf{W}^T \mathbf{S}_L \mathbf{W}\right)} \geq \sqrt{\sum_{i=1}^{r} \lambda_i} \ ,
\end{aligned}$$

where $\lambda_i \ (i = 1, \ldots, r)$, ordered by $\lambda_1 \leq \cdots \leq \lambda_r$, are the eigenvalues of $\mathbf{S}_L$. The last inequality in Eq. (33) is obtained by the Ky Fan's inequality [7], which states that $\mathbf{tr}\left(\mathbf{W}^T \mathbf{S}_L \mathbf{W}\right) \geq \sum_{i=1}^{r} \lambda_i$. Again, given an input data set, $\mathbf{S}_L$ is an constant matrix thereby $\sum_{i=1}^{r} \lambda_i$ is a constant. Thus the denominator of our objective in Eq. (4) is lower bounded.

The two bounds in Eq. (32) and Eq. (33) together indicate that our objective in Eq. (4) is upper bounded. □

Theorem 3 indicates that our proposed Algorithm 3 monotonically increase the objective function value in each iteration. Theorem 4 indicates that the objective function is upper bounded, which, together with Theorem 3, indicates that Algorithm 3 converges to a local optimum.

4. Experiments

In this section, we experimentally evaluate the proposed image representation method in an automatic image annotation task, where we use the following three multi-instance image data sets: the PASCAL VOC 2010 data set [6], the Corel5K data set [5], and the Scene data set [40]. We perform our evaluations using standard 5-fold cross-validation and report the average performances over the 5 trials.

The proposed image representation learning method has two parameters, the number of neighborhoods K of an instance and the dimensionality r of the projected subspace. In our experiments, the performance of the proposed method is very stable with respect to these two parameters in considerably large value ranges. Empirically, in all our experiments we select $K = \min\{3, n\}$ where n is the number of instances in an image bag and $r = d/10$ where d is the dimensionality of the instance vectors.

Experimental settings. We first compare our method to two baseline classification methods including support vector machine (SVM) method and the transductive support vector machine (TSVM) [13] method. The former is the most broadly used supervised classification method in statistical learning, while the latter is an extension of the former one and is a semi-supervised classification method. Because both of these two methods are designed for single-instance data, they are not able to deal with data with representations of varied sizes. Therefore, we train and classify images using the holistic representations of the experimental images. Specifically, for each class we train a one-*vs*.-others classier using the images in the training data set, and classify the images in the test data set. Gaussian kernel is used in the both methods, *i.e.*, $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\beta \left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2\right)$, where β and the regularization box parameter C are fine tuned by searching the grid of $\left\{10^{-5}, \ldots, 10^{-1}, 1, 10, \ldots, 10^{5}\right\}$ via an internal 5-fold cross-validation using the training data of each of the 5 trails. The both methods are implemented using SVM^{light} software package [1].

We also compare our method against two very recent MIL methods including the miGraph [39] method and the MIMLSVM+ [17] method. Because miGraph method is a single-label classification method, one-*vs*.-others strategy is used to conduct classification, one class at a time. We implement these two methods using the codes published by the respective authors. Because the both methods are multi-instance classification methods, we perform classification using the semantic instances of the input images.

Table 1. Comparison of the performances (mean ± std) of the compared methods in the image annotation tasks.

	Method	Hamming loss ↓	One-error ↓	Coverage ↓	Rank loss ↓	Average precision ↑
PASCAL	SVM	0.183 ± 0.016	0.336 ± 0.018	1.025 ± 0.014	0.186 ± 0.015	0.476 ± 0.022
	TSVM	0.180 ± 0.015	0.331 ± 0.016	1.022 ± 0.012	0.183 ± 0.016	0.478 ± 0.025
	miGraph	0.173 ± 0.011	0.306 ± 0.018	1.013 ± 0.018	0.178 ± 0.013	0.483 ± 0.023
	MIMLSVM+	0.176 ± 0.014	0.323 ± 0.024	0.999 ± 0.015	0.177 ± 0.010	0.485 ± 0.022
	Our method (3NN)	0.165 ± 0.009	0.289 ± 0.011	0.975 ± 0.010	0.151 ± 0.002	0.481 ± 0.013
	Our method (SVM)	$\mathbf{0.155 \pm 0.014}$	$\mathbf{0.272 \pm 0.011}$	$\mathbf{0.962 \pm 0.016}$	$\mathbf{0.139 \pm 0.012}$	$\mathbf{0.507 \pm 0.015}$
Corel5K	SVM	0.283 ± 0.011	0.584 ± 0.011	5.972 ± 0.011	0.291 ± 0.011	0.465 ± 0.012
	TSVM	0.276 ± 0.005	0.579 ± 0.012	5.993 ± 0.052	0.291 ± 0.006	0.476 ± 0.015
	miGraph	0.246 ± 0.015	0.571 ± 0.009	5.510 ± 0.013	0.233 ± 0.011	0.545 ± 0.013
	MIMLSVM+	0.238 ± 0.004	0.568 ± 0.013	5.104 ± 0.009	0.241 ± 0.015	0.559 ± 0.018
	Our method (3NN)	0.211 ± 0.011	0.526 ± 0.013	$\mathbf{4.611 \pm 0.021}$	0.216 ± 0.012	$\mathbf{0.611 \pm 0.016}$
	Our method (SVM)	$\mathbf{0.204 \pm 0.015}$	$\mathbf{0.507 \pm 0.009}$	4.751 ± 0.021	$\mathbf{0.207 \pm 0.008}$	0.604 ± 0.010
Scene	SVM	0.228 ± 0.011	0.374 ± 0.011	1.041 ± 0.018	0.209 ± 0.011	0.695 ± 0.021
	TSVM	0.231 ± 0.005	0.381 ± 0.011	1.078 ± 0.009	0.211 ± 0.012	0.701 ± 0.018
	miGraph	0.221 ± 0.012	0.384 ± 0.012	1.071 ± 0.014	0.241 ± 0.022	0.715 ± 0.033
	MIMLSVM+	0.215 ± 0.011	0.370 ± 0.012	1.015 ± 0.003	0.238 ± 0.012	0.709 ± 0.022
	Our method (3NN)	0.187 ± 0.011	$\mathbf{0.350 \pm 0.017}$	0.995 ± 0.012	0.182 ± 0.014	$\mathbf{0.795 \pm 0.026}$
	Our method (SVM)	$\mathbf{0.175 \pm 0.004}$	0.355 ± 0.016	$\mathbf{0.980 \pm 0.026}$	$\mathbf{0.174 \pm 0.007}$	0.794 ± 0.011

For our method, once the multi-instance enriched representations of the input images are learned, they can be directly fed into any traditional single-instance classifiers. Thus we evaluate our new image representation learning method using two most broadly used classifiers: the K-nearest neighbour (KNN) classifier and the SVM. In our experiments, we select $K = 3$ in KNN classifiers and use the same settings as detailed above the SVM classifiers.

Experimental results. Because the three experimental image data sets are all multi-label data sets, we evaluate the classification performances of the compared methods using five broadly used multi-label evaluation metrics as in Table 1, where "↓" indicates "the smaller is the better", while "↑" indicates "the bigger is the better". We refer readers to [22] for detailed definitions of these evaluation metrics.

The average classification performances (mean ± standard deviation) of the compared methods over the 5 trials of the experiments are reported in Table 1, from which we can see a number of interesting observations as following. First, the proposed method is consistently better than the other four competing methods, sometimes very significantly. Second, the MIL methods are generally better than the two baseline classification methods that only use the holistic image representations. This observation is reasonable in that the two baseline methods are both single-instance classification methods, which only use the holistic image representations. As a result, the important structural information contained in image patches with semantic meanings are not exploited, which leads to inferior performance. Last, but not least, the SVMs using the raw holistic image representations perform drastically worse than those using the learned image representations by our new method, *i.e.*, the holistic image representation with multi-instance enrichments. This observation firmly confirms that our proposed method can improve the image representations in terms of image annotation. To summarize, the experimental results in Table 1 clearly demonstrate the effectiveness of the proposed methods in multi-instance multi-label image classification.

5. Conclusions

In this paper, we have presented a novel image representation learning method that is able integrates the information conveyed by both local image patches and the holistic representation of the entire image. Our new method first learns a projection to preserve both global and local consistencies of the instances of the input image in a projected subspace, then it projects the holistic representation of the entire image into the learned subspace for information enrichment. Taking into account the content and characterization variations in pictures for nature scenes and photos, we developed an objective that simultaneously maximizes and minimizes the summations of a number of ℓ_1-norm distances, which is difficult to solve in general. Thus, we derived an efficient iterative solution algorithm that is non-greedy and theoretically proved to converge. Our new method has been validated in extensive experiments to simulate the real-world applications.

References

[1] SVMLight: Support Vector Machine.

[2] A. Baccini, P. Besse, and A. de Faguerolles. A l1-norm pca and heuristic approach. *International Conference on Ordinal and Symbolic Data Analysis*, pages 359–368, 1996.

[3] Y. Chen and J. Wang. Image categorization by learning and reasoning with regions. *JMLR*, 5:913–939, 2004.

[4] C. Ding, D. Zhou, X. He, and H. Zha. R1-pca: rotational invariant l1-norm principal component analysis for robust subspace factorization. In *ICML*, pages 281–288, 2006.

[5] P. Duygulu, K. Barnard, J. F. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results.

[7] K. Fan. On a theorem of weyl concerning eigenvalues of linear transformations ii. *Proceedings of the National Academy of Sciences*, 36(1):31–35, 1950.

[8] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic press, 2013.

[9] J. Gao. Robust l1 principal component analysis and its bayesian variational inference. *Neural Computation*, 20:555–572, 2008.

[10] F. Han, X. Yang, Y. Deng, M. Rentschler, D. Yang, and H. Zhang. SRAL: Shared representative appearance learning for long-term visual place recognition. *IEEE Robotics and Automation Letters*, 2(2):1172–1179, 2017.

[11] R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. In *International Conference on Artificial Intelligence and Statistics*, 2010.

[12] Y. Jia, F. Nie, and C. Zhang. Trace ratio problem revisited. *IEEE Transactions on Neural Networks*, 20(4):729–735, 2009.

[13] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, pages 200–209.

[14] I. T. Jolliffe. Principal component analysis and factor analysis. In *Principal component analysis*, pages 115–128. Springer, 1986.

[15] Q. Ke and T. Kanade. Robust l1 norm factorization in the presence of outliers and missing data by alternative convex programming. In *CVPR*, pages 592–599, 2004.

[16] N. Kwak. Principal component analysis based on l1-norm maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:1672–1680, 2008.

[17] Y. Li, S. Ji, S. Kumar, J. Ye, and Z. Zhou. Drosophila Gene Expression Pattern Annotation through Multi-Instance Multi-Label Learning. *ACM/IEEE TCBB*, 2011.

[18] Y. Liu, Q. Gao, S. Miao, X. Gao, F. Nie, and Y. Li. A non-greedy algorithm for l1-norm lda. *IEEE Transactions on Image Processing*, 26(2):684–695, 2017.

[19] O. Maron and A. L. Ratan. Multiple-instance learning for natural scene classification. In *ICML*, 1998.

[20] F. Nie, H. Huang, C. Ding, D. Luo, and H. Wang. Robust principal component analysis with non-greedy ℓ_1-norm maximization. In *IJCAI*, 2011.

[21] F. Nie, H. Wang, H. Huang, and C. H. Ding. Adaptive loss minimization for semi-supervised elastic embedding. In *IJCAI*, pages 1565–1571, 2013.

[22] R. Schapire and Y. Singer. BoosTexter: A boosting-based system for text categorization. *Machine learning*, 39(2):135–168, 2000.

[23] W. Sun and Y.-X. Yuan. *Optimization theory and methods: nonlinear programming*, volume 1. Springer Science & Business Media, 2006.

[24] H. Wang, C. Deng, H. Zhang, X. Gao, and H. Huang. Drosophila gene expression pattern annotations via multi-instance biological relevance learning. In *AAAI*, pages 1324–1330, 2016.

[25] H. Wang, H. Huang, and C. Ding. Image annotation using multi-label correlated green's function. In *CVPR*, pages 2029–2034, 2009.

[26] H. Wang, H. Huang, and C. Ding. Multi-label feature transform for image classifications. *Computer Vision–ECCV 2010*, pages 793–806, 2010.

[27] H. Wang, H. Huang, and C. Ding. Image annotation using bi-relational graph of images and semantic labels. In *CVPR*, pages 793–800. IEEE, 2011.

[28] H. Wang, H. Huang, F. Kamangar, F. Nie, and C. H. Ding. Maximum margin multi-instance learning. In *NIPS*, volume 1, page 3, 2011.

[29] H. Wang, F. Nie, W. Cai, and H. Huang. Semi-supervised robust dictionary learning via efficient $\ell_{2,1}$-norms minimization. In *ICCV*, 2013.

[30] H. Wang, F. Nie, and H. Huang. Learning instance specific distance for multi-instance classification. In *AAAI*, 2011.

[31] H. Wang, F. Nie, and H. Huang. Robust and discriminative distance for multi-instance learning. In *CVPR*, 2012.

[32] H. Wang, F. Nie, and H. Huang. Robust and discriminative self-taught learning. In *International Conference on Machine Learning*, pages 298–306, 2013.

[33] H. Wang, F. Nie, and H. Huang. Globally and locally consistent unsupervised projection. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[34] H. Wang, F. Nie, and H. Huang. Robust distance metric learning via simultaneous ℓ_1-norm minimization and maximization. In *ICML*, pages 1836–1844, 2014.

[35] X.-S. Wei and Z.-H. Zhou. An empirical study on image bag generators for multi-instance learning. *Machine Learning*, 105(2):155–198, 2016.

[36] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma. Robust principal component analysis: Exact recovery of corrupted. *NIPS*, page 116, 2009.

[37] J. Wu and J. M. Rehg. CENTRIST: A visual descriptor for scene categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1489–1501, 2011.

[38] J.-S. Wu, S.-J. Huang, and Z.-H. Zhou. Genome-wide protein function prediction through multi-instance multi-label learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11(5):891–902, 2014.

[39] Z. Zhou, Y. Sun, and Y. Li. Multi-instance learning by treating instances as non-I.I.D. samples. In *ICML*, 2009.

[40] Z.-H. Zhou and M.-L. Zhang. Multi-instance multi-label learning with application to scene classification. *NIPS*, 19:1609, 2007.