

Discriminative Semi-supervised Learning Based on Visual Concept-Like Features

Fang Liu and Xiaofeng Wu^(✉)

Department of Electronic Engineering, Fudan University,
Shanghai 200433, China
xiaofengwu@fudan.edu.cn

Abstract. A discriminative semi-supervised learning method based on visual concept-like high-level features is proposed in this paper. Previous semi-supervised learning methods usually use unlabeled data to augment the training set or regularize the decision boundary of classifiers. The classification results rely on the precision on unlabeled data using supervised classifiers trained with limited labeled samples. When a small number of labeled samples are provided, these methods are likely to get bad results. Differently, the proposed method directly uses the distribution information of all available data in the feature space to learn a new representation which is achieved by computing the similarities of a chosen image and some discriminative data exemplars sampled from the feature space. A semi-supervised distance metric learning method by learning a projection matrix under the equivalence constraints of similar pairs and dissimilar pairs is introduced to measure these similarities, and a pseudo-mahalanobis distance is thus obtained to represent the similarities between data samples instead of Euclidean distance. Experiments showed the effectiveness of this learned distance. The new representation can be fed into standard classifiers for image classification task. The training data of our system can either be original image data or handcrafted features or image features learned by deep architectures. Therefore, the proposed method can be applied in both feature extraction and feature enhancement. In the semi-supervised classification task on eight standard datasets, the proposed method achieves improved performance over many of the previous existing methods.

Keywords: Semi-supervised image classification · Discriminative feature learning · Metric learning

1 Introduction

Recent years have witnessed great progress in image classification with less or no annotations, since annotating data by human efforts is both time-consuming and expensive while unlabeled data is numerous and easy to achieve. Semi-supervised methods try to reveal the information carried by unlabeled data to improve performance. A large number of methods regularize the decision boundary by forcing it to pass through the region with lower density of unlabeled data. Another widely-used scheme is the self-training scheme [1]. It first annotates the unlabeled data by training a supervised classifier on labeled data. Then the training set is augmented by adding the

most confident unlabeled data with their predicted labels. The system iterates between training models and augmenting the training set until some termination condition is reached. However, self-training relies on the predicted labels on unlabeled data for training a new classifier. It can probably make an error when a small number of labeled data is available. Differently, some ensemble algorithms assign a pseudo-label to unlabeled data, and then sample them for training a new classifier. They iterate to construct the ensemble classifier under the restriction of a cost function. The precision relies on the prediction of pseudo-labels using the constructed classifier at each iteration. In contrast, the proposed method compares the learned distance between data samples to obtain pseudo-labels. Then ensemble supervised classifiers are trained and used to extract a new visual concept-like representation of the input data.

Different object classes carry many discriminative visual concept-like features that can help us distinguish the classes, such as colors, shapes and textures. We call them visual concepts for simplicity in this paper. These concepts are lower-dimensional compared with the features or images. A new category can be learned by comparing the new object with the existing categories from the perspective of visual concepts. For instance, a volleyball has similar shape to a basketball but has different texture and color. This learning procedure is called learning by comparison. It is a part of Eleanor Rosch’s prototype theory [2] which states that an object’s class is determined by its similarities to prototypes that represent object classes. This theory has been used successfully in transfer learning, where labeled data from different classes are available.

Since the visual concepts exist in images regardless of labels and can provide discriminative information, we consider to learn these visual concepts from all available data. In particular, we aim to learn some data samples that contain several typical visual concepts. The typical concepts are called concept exemplars such as “spherical”, “red” and “brown”. And a group of data samples containing several similar concept exemplars are named a “subset”. For instance, the class of apples can be viewed as a subset of “red” and “smooth” at least. These subsets are sampled from all available data in an unsupervised way based on the assumption that neighboring samples in the feature space share similar concepts exemplars. To be discriminative, samples of the same subset should be close to each other and samples from different subsets should be as far as possible. In other words, the subsets are expected to be inter-distinct and intra-compact. We combine several subsets to form a cluster. Discriminative information can be learned by concatenated the similarities between the chosen image and these subsets. However, a cluster is formed in one sampling trial and can be noisy because the concepts we learned in one trial are limited, so a rich set of clusters is necessary for our learning procedure to cover enough concepts.

As stated above, the learning-by-comparison procedure has a critical demand on similarity measurement. Euclidean distance is widely used to represent the similarity. However, it is not enough when features have high dimensionality. We introduce a semi-supervised distance metric learning method which learns a pseudo-mahalanobis distance by learning a projection matrix under the equivalence constraints. The pseudo-mahalanobis distance can measure the similarity in a better way, which means similar samples are closer and dissimilar samples are as far as possible.

The rest of this paper is organized as follows. Section 2 briefly reviews related work in the field of both metric learning and high-level feature learning. Section 3

introduces the proposed method in details, followed by experiments for performance evaluation in Sect. 4. Section 5 concludes the paper.

2 Related Work

Our work is generally relevant to image classification with metric learning under equivalence constraints and high-level feature learning based on Eleanor Rosch’s prototype theory. In this section, some current related works in these two research fields are simply reviewed, and the similarities and differences among our method and these works are discussed.

2.1 Metric Learning Under Equivalence Constraints

There are many widely-used metric learning methods using equivalence constraints. Relevant components analysis [3] (RCA) learns an embedding which allocates larger weights to the most relevant dimensions of the features and lower weights to less relevant ones. But it does not incorporate dissimilarity constraints, and is limited in the original input space to learn linear transformations. Discriminative components analysis [4] (DCA) incorporates dissimilarity constraints into RCA and Kernel RCA. The semi-supervised discriminative common vector method [5] (SS-DCV) introduced in our method is similar in spirit to DCA, but it overcomes a serious shortcoming of DCA – the criterion of maximizing the classical LDA (Linear Discriminant Analysis) function does not have a unique solution while the dimensionality of the sample space is much larger than the number of similar sample pairs, which leads to miss the optimal projection direction. SS-DCV method projects the data onto the subspace orthogonal to the linear span of the difference vectors of similar sample pairs first, in which similar pairs have identical projections. Then it learns a linear embedding that maximizes the scatter of the dissimilar sample pairs. This corresponds to a pseudo-metric characterized by a positive semi-definite matrix in the original input space. The integrate derivation of SS-DCV can be found in [5].

2.2 High-Level Feature Learning Using Eleanor Rosch’s Prototype Theory

A method that is closely related to ours is Ensemble Projection [6]. It is also based on the Eleanor Rosch’s prototype theory. It samples an ensemble of prototype sets that present different classes. In return, an ensemble of diverse projection functions are learned based on these prototype sets. The prototype is similar to our visual concept exemplar. But their prototypes indicate the classes directly and our visual concept exemplar is expected to be a subset of several visual concepts shared among different object classes. Projection values of an individual data sample through these projection functions are stacked together to form a new feature representation. However, different from our approach, Ensemble Projection is purely unsupervised and do not leverage label information for a specific task. In particular, Ensemble Projection samples the prototypes in the original feature space. In our method, visual concept exemplars are

learned in a lower-dimensional subspace. In the projected space, similar pairs have identical projections which indicate that they have similar subsets of visual concepts. In the experiments on different datasets for image classification, we will compare our algorithm with the Ensemble Projection method.

3 Discriminative Semi-supervised Learning Based on Visual Concept-Like Features

Following the notations that are widely used in semi-supervised feature learning, the input of our method includes labeled data $\mathbf{x}_{1:L}, y_{1:L}$ and unlabeled data $\mathbf{x}_{L+1:L+U}$, where \mathbf{x}_i denotes the feature vector of image i , $y_i \in \{1, \dots, C\}$ indicates its label, and C is the number of classes, L is the length of labeled data, U is the length of unlabeled data. A new image representation \mathbf{f} is learned using both the unlabeled data and the labeled data.

3.1 Projection onto a Subspace Under Equivalence Constraints

Given a small amount of labeled data, we aim to learn a class-discriminative subspace and achieve a better judgement of similarity using SS-DCV [5].

Let $\mathbf{x}_i \in \mathbb{R}^d, i = 1, \dots, N$ denote the samples of the training set. A set of equivalence constraints in the form of similar and dissimilar pairs are given and we aim to learn a pseudo-mahalanobis distance of the form

$$d_A(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_A = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{A} (\mathbf{x}_i - \mathbf{x}_j)}, \quad (1)$$

where $\mathbf{A} \geq 0$ is a symmetric positive semi-definite matrix reflecting the underlying relationships between different dimensions of the feature. If $q = \text{Rank}(\mathbf{A}) \leq d$, \mathbf{A} can be written in the form $\mathbf{A} = \mathbf{W}\mathbf{W}^T$ where \mathbf{W} is a full-rank rectangular matrix of size $d \times q$, so that

$$\|\mathbf{x}_i - \mathbf{x}_j\|_A^2 = \|\mathbf{W}^T \mathbf{x}_i - \mathbf{W}^T \mathbf{x}_j\|^2 \quad (2)$$

i.e. the pseudo-mahalanobis distance between samples are equivalent to Euclidean distances on their linear projections by \mathbf{W}^T .

\mathbf{X}_S and \mathbf{X}_D denote the matrixes whose columns are the difference vectors of the given similar and dissimilar pairs

$$\mathbf{X}_S = [\mathbf{x}_{s1,1} - \mathbf{x}_{s1,2}, \mathbf{x}_{s2,1} - \mathbf{x}_{s2,2}, \dots, \mathbf{x}_{sn,1} - \mathbf{x}_{sn,2}], \quad (3)$$

$$\mathbf{X}_D = [\mathbf{x}_{d1,1} - \mathbf{x}_{d1,2}, \mathbf{x}_{d2,1} - \mathbf{x}_{d2,2}, \dots, \mathbf{x}_{dm,1} - \mathbf{x}_{dm,2}], \quad (4)$$

where $\mathbf{x}_{si,1}$ and $\mathbf{x}_{si,2}$ respectively represent the first and second samples of the i -th similar sample pair; $\mathbf{x}_{di,1}$ and $\mathbf{x}_{di,2}$ respectively represent the first and second samples of the i -th dissimilar sample pair.

The full method of semi-supervised discriminative common vector can be presented as follows:

1. Compute \mathbf{X}_s and its orthonormal basis matrix \mathbf{U} .
2. Project the dissimilar sample pairs to the null space of \mathbf{X}_s using $\tilde{\mathbf{X}}_D = (\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{X}_D$.
3. Compute $\tilde{\mathbf{S}}_D = \tilde{\mathbf{X}}_D\tilde{\mathbf{X}}_D^T$ and its leading eigenvectors \mathbf{W} , then output the final distance metric $\mathbf{A} = \mathbf{W}\mathbf{W}^T$.

3.2 Visual Concept Clusters Learning and Feature Extraction

In the projected lower-dimensional subspace, we learn S cluster sets $\mathcal{C}_{1:S}$. Each cluster \mathcal{C}_s contains $e_{1:m \times n}, l_{1:m \times n}$, where e_i denotes the i -th exemplar and $l_i \in \{1, \dots, n\}$ denotes the pseudo label. There are n subsets of exemplars in a cluster and the pseudo label indicates which subsets the exemplar belongs to. In each subset, we have m exemplars so the total number of exemplars in a cluster is $m \times n$.

Firstly, a group of feature samples that are far apart from each other are chosen as the skeleton of the subsets. Then their k -nearest neighbors are found to enrich the subsets. The pseudo label is shared in the same subset. The method of learning the cluster set can be summarized as follows:

1. Randomly choose n feature samples as a skeleton for t times and choose the furthest one.
2. Calculate m nearest neighbors of every feature sample in the skeleton.
3. Repeat step 1 to step 2 for S times to get the cluster sets $\mathcal{C}_{1:S}$.

After the cluster set is created, we train logistic regression on each cluster. The classification scores of a chosen feature are then concatenated to form the new representation. Classification scores indicate the similarities between the chosen feature and the visual concepts exemplars we created, so they can represent the correlation of the current feature and the shared visual concepts.

4 Experiments

The primary performance evaluation experiments are executed for semi-supervised image classification on eight standard datasets as follows:

1. Texture-25 [7]: texture images divided into 25 categories, with 40 images per class.
2. Caltech-101 [8]: 8677 images from 101 object classes, with 31 to 800 samples per class.
3. STL-10 [9]: 100000 unlabeled images and 13000 labeled images from 10 object classes with 500 training images and 800 test images per class.
4. Scene-15 [10]: 4485 scene images of indoor and outdoor environments divided into 15 classes, with 200 to 400 samples per class.
5. Indoor-67 [11]: 15620 images from 67 indoor classes, with at least 100 images per class.

6. Event-8 [12]: 1574 images from 8 sports event categories.
7. Building-25 [13]: 4794 images from 25 architectural styles, such as American craftsman, Baroque, and Gothic.
8. LandUse-21 [14]: 2100 satellite images divided into 21 classes, with 100 samples per class.

The inputs of our algorithm are CNN features with the dimensionality of 4096 obtained from an off-the-shelf CNN [15] pre-trained on the ImageNet dataset. For comparison, we use the same datasets and CNN features used by [6].

4.1 Experiment Settings

Three baselines are used in our primary performance evaluation experiments: k-nearest neighbor (k-NN) algorithm, Logistic Regression (LR) and support vector machines (SVMs) with radial basis function (RBF) kernels for semi-supervised classification. The original convolutional neural network (CNN) features, the features learned by the Ensemble Projection method and features learned by our algorithm were fed into these classifiers to evaluate the method. For different datasets, the parameters in the experiments were fixed to the following values: $S = 100$, $m = 6$, $n = 30$, and $t = 50$. Different numbers of training images per class were tested. In keeping with most existing systems for semi-supervised classification [16–19], we evaluate the method in the transductive manner, where we take the training and test samples as a whole, and randomly choose labeled samples from the whole dataset to learn and infer labels of other samples whose labels are held back as the unlabeled samples. The reported results are the average performance over 5 runs with random labeled-unlabeled splits.

4.2 Classification Results

Table 1 lists the precision of the methods using 5 labeled training examples per class. Three kinds of classifiers are used: k-NN, Logistic Regression and SVMs with RBF kernel. They worked with three feature inputs: the original CNN features, features learned by ensemble projection (indicated by “+EP”) and features learned by our visual concept method (indicated by “+VC”). The best performance for each dataset is indicated in bold, and the second best is in bold italic. It is easy to observe that

Table 1. Precision (%) of image classification on the eight datasets, with 5 labeled training examples per class

Methods	Scene-15	LandUse-21	Texture-25	Building-25	Event-8	Caltech-101	Indoor-67	STL-10
<i>k</i> -NN	61.27	69.04	81.50	30.57	75.51	69.75	20.86	54.16
<i>k</i> -NN+EP	74.86	75.73	83.97	34.64	86.35	70.49	24.64	64.75
<i>k</i> -NN+VC	76.03	79.94	89.51	35.82	86.67	61.86	26.75	65.71
LR	73.07	77.70	86.96	36.61	84.71	81.38	31.02	65.66
LR+EP	80.06	80.56	87.52	40.36	89.47	79.28	34.50	74.00
LR+VC	80.78	84.12	91.72	41.42	90.01	80.21	36.30	75.29
SVMs	72.54	73.78	82.00	38.75	83.12	76.92	32.05	65.62
SVMs+EP	79.66	77.23	83.10	39.38	87.31	74.77	31.76	70.99
SVMs+VC	81.39	81.91	90.35	40.34	88.39	68.21	34.26	73.57

classifiers consistently have better performance when working with our features. Logistic regression performs best among these three classifiers. Working with our features, logistic regression gets the highest precision on six of the eight datasets followed by SVMs, which get higher precision than k-NN.

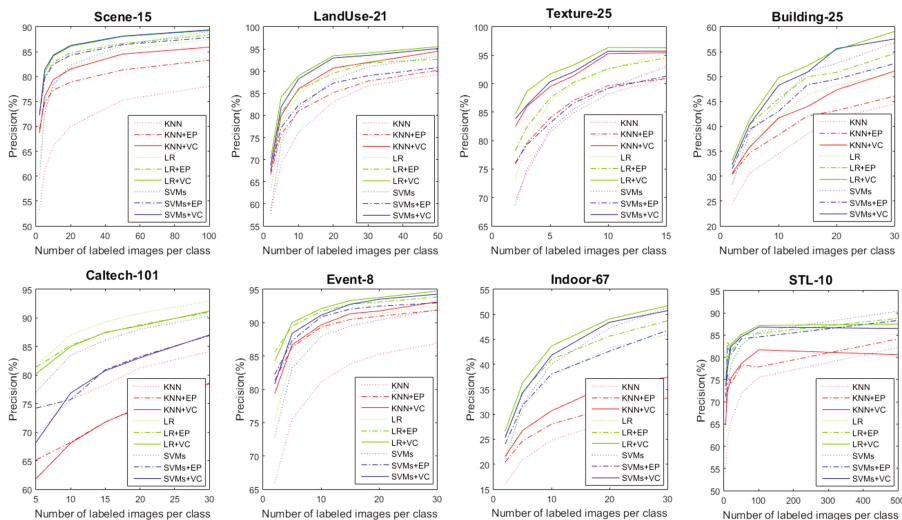


Fig. 1. Precision (%) of image classification on the eight datasets using different number of labeled images per class. Classifiers and features used are as same as Table 1.

Figure 1 shows the classification results on the eight datasets using these three kinds of classifiers and three feature inputs when different number of labeled training examples are provided. Numbers of labeled images per class are chosen according to the different structures of the datasets: different number of classes and different number of images per class. Results achieved by the same kind of classifier are shown in the same color and results generated by three classifiers working with same features are shown with the same line type. The figure shows the advantages of our features over the original CNN features and the features learned by Ensemble Projection across different datasets and classifiers. For example, when given 5 labeled samples per class, we obtain a 4.2% improvement over the Ensemble Projection on the Texture-25 dataset using LR and 3.56% on the LandUse-21 dataset.

The enhancement of precision is attributed to our discriminative common vector learning for optimal similarity measure metrics, and projecting all available data to the lower-dimensional subspace. This projection applies the supervisory information carried by the labeled sample pairs and learns a lower-dimensional visual concept in the subspace. When given a small number of labeled data, our method can be used as an effective feature enhancing method.

To clarify the influence and contributions of the parameters in our method on various image datasets, we tested a variety of values. In details, the number of clusters

and the number of subsets per cluster and the number of exemplars per subset are adjusted. It shows that learning more exemplars or more clusters can slightly improve the accuracy. But in a large range of values, results are not sensitive to the parameters. Values fixed in the experiments can achieve a promising result. The proposed method is also efficient because the training of logistic regression is efficient.

5 Conclusion

A semi-supervised high-level feature learning method that aims to approach the lower-dimensional visual concepts in our cognitive process is proposed. By extracting similar and dissimilar sample pairs from labeled data and projecting features onto a lower-dimensional subspace under the equivalence constraints, we leveraged the discriminative information carried by labeled data and obtained reduced dimensional features which are closer to the visual concepts. A rich set of concept exemplar subsets are learned in the subspace. They not only included the difference between image classes, but also carried the similarities among all available data in terms of visual concepts. Images are classified and linked to these subsets. The classification scores are stacked to form the new representation. Experiments conducted on eight standard datasets show the effectiveness of our method.

Our method can be used in feature extraction or feature enhancement. For a specific semi-supervised classification task, it is easy to achieve the CNN features by fine-tuning based on a pre-trained model. Then our method can be applied to reveal the information carried by the labeled data and enhance the feature using both labeled and unlabeled data to improve the classification results. To prepare more comprehensive equivalence constraints and apply the proposed method to practical tasks would be our future work.

References

1. Rosenberg, C., Hebert, M., Schneiderman, H.: Semi-supervised self-training of object detection models. In: IEEE Workshops on Application of Computer Vision, vol. 1, pp. 29–36. IEEE Computer Society (2005)
2. Rosch, E.: Principles of categorization. In: Concepts: Core Readings, pp. 189–206 (1999)
3. Shental, N., Hertz, T., Weinshall, D., Pavel, M.: Adjustment learning and relevant component analysis. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2353, pp. 776–790. Springer, Heidelberg (2002). doi:[10.1007/3-540-47979-1_52](https://doi.org/10.1007/3-540-47979-1_52)
4. Hoi, S.C., Liu, W., Lyu, M.R., Ma, W.-Y.: Learning distance metrics with contextual constraints for image retrieval. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2072–2078. IEEE (2006)
5. Cevikalp, H.: Semi-supervised discriminative common vector method for computer vision applications. *Neurocomputing* **129**, 289–297 (2014)
6. Dai, D., Van Gool, L.: Unsupervised high-level feature learning by ensemble projection for semi-supervised image classification and image clustering. arXiv preprint [arXiv:1602.00955](https://arxiv.org/abs/1602.00955) (2016)

7. Lazebnik, S., Schmid, C., Ponce, J.: A sparse texture representation using local affine regions. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 1265–1278 (2005)
8. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.* **106**, 59–70 (2007)
9. Coates, A., Lee, H., Ng, A.Y.: An analysis of single-layer networks in unsupervised feature learning. In: *AISTATS 2011, Ann Arbor*, vol. 1001, no. 2 (2010)
10. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2169–2178. IEEE (2006)
11. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, pp. 413–420. IEEE (2009)
12. Li, L.-J., Fei-Fei, L.: What, where and who? Classifying events by scene and object recognition. In: *IEEE 11th International Conference on Computer Vision, ICCV 2007*, pp. 1–8. IEEE (2007)
13. Xu, Z., Tao, D., Zhang, Ya., Wu, J., Tsoi, A.C.: Architectural style classification using multinomial latent logistic regression. In: *Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS*, vol. 8689, pp. 600–615. Springer, Cham (2014). doi:[10.1007/978-3-319-10590-1_39](https://doi.org/10.1007/978-3-319-10590-1_39)
14. Yang, Y., Newsam, S.: Bag-of-visual-words and spatial extensions for land-use classification. In: *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 270–279. ACM (2010)
15. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. *arXiv preprint [arXiv:1405.3531](https://arxiv.org/abs/1405.3531)* (2014)
16. Ebert, S., Larlus, D., Schiele, B.: Extracting structures in image collections for object recognition. In: *Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS*, vol. 6311, pp. 720–733. Springer, Heidelberg (2010). doi:[10.1007/978-3-642-15549-9_52](https://doi.org/10.1007/978-3-642-15549-9_52)
17. Fergus, R., Weiss, Y., Torralba, A.: Semi-supervised learning in gigantic image collections. In: *Advances in Neural Information Processing Systems*, pp. 522–530 (2009)
18. Liu, W., He, J., Chang, S.-F.: Large graph construction for scalable semi-supervised learning. In: *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, pp. 679–686 (2010)
19. Pitelis, N., Russell, C., Agapito, L.: Semi-supervised learning using an unsupervised atlas. In: *Calders, T., Esposito, F., Hüllermeier, E., Meo, R. (eds.) ECML PKDD 2014. LNCS*, vol. 8725, pp. 565–580. Springer, Heidelberg (2014). doi:[10.1007/978-3-662-44851-9_36](https://doi.org/10.1007/978-3-662-44851-9_36)