Supplementary Document: Robust and Discriminative Self-Taught Learning

Hua Wang HUAWANGCS@GMAIL.COM
Colorado School of Mines, 1610 Illinois Street, Golden, Colorado 80401 USA

Feiping Nie FEIPINGNIE@GMAIL.COM
Heng Huang HENG@UTA.EDU
The University of Texas at Arlington, 500 UTA Boulevard, Arlington, Texas 76019 USA

This document includes the following supplementary materials to the main text of the paper.

- Section 1: experimental data descriptions.
- Section 2: detailed algorithm derivation.
- Section 3: study of robustness by using $\ell_{2,1}$-norm loss.
- Section 4: additional single-label classification results.

1. Data description

In Section 4 of the main text of this paper, we experiment with the following three multi-label image data sets, which are broadly used computer vision studies.

MSRC-v2[1] data set is an extension of MSRC-v1 data set, which has 591 images annotated by 22 classes at pixel level. In our experiments, we use the image level annotation.

TRECVID 2005 data set[2] contains 61901 subshots labeled with 39 concepts according to LSCOM-Lite annotations (Naphade et al., 2005). Following (Wang et al., 2009b), we randomly sample the data such that each concept (label) has at least 100 video key frames.

NUS-WIDE-Object data set[3] was created by Lab for Media Search in National University of Singapore. The data set includes 269,648 images and the associated tags from Flickr, with a total number of 5,018 unique tags. For the Object subset, we randomly select images from the data set, such that at least 200 images are selected for each class.

[1]http://research.microsoft.com/en-us/projects/objectclassrecognition

[2]http://www-nlpir.nist.gov/projects/trecvid/

[3]http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

For the auxiliary image data set and the three target data sets described above, following (Gehler & Nowozin, 2009), we extract SIFT descriptors for the experimental images, which are computed on a regular grid on the image with a spacing of 10 pixels and for the four different radii $r = 4, 8, 12, 16$. The descriptors are subsequently quantized into a vocabulary of 300 visual words that is generated by k-means clustering.

2. Detailed derivation of the algorithm

Due to the non-smoothness of the $\ell_{2,1}$-norm function, the objective $J_{\text{RD-STL}}$ in Eq. (9) is highly non-smooth as it involves $K+2$ terms of $\ell_{2,1}$-norm. Thus, minimizing $J_{\text{RD-STL}}$ is hard in general by existing algorithms. In this section, we derive an efficient algorithm to solve it.

Because $J_{\text{RD-STL}}$ has two variables, *i.e.*, $\mathbf{D}$ and $\mathbf{A}$, we alternately optimize them.

First, when $\mathbf{D}$ is fixed, the objective $J_{\text{RD-STL}}$ in Eq. (10) can be decoupled to the following problems for each $k\,(0 \le k \le K)$:

$$\min\ J(\mathbf{A}_k) = \left\|(\mathbf{X}_k - \mathbf{D}\mathbf{A}_k)^T\right\|_{2,1} + \lambda \left\|\mathbf{A}_k\right\|_{2,1}\ . \quad (1)$$

Because the function $f(\mathbf{M}) = \|\mathbf{M}\|_{2,1}$ is non-smooth and not differentiable when $\mathbf{M} = 0$, as mentioned earlier in footnote 1 in the main text of the paper, following (Gorodnitsky & Rao, 1997), we introduce a small perturbation $\zeta > 0$ to replace $\|\mathbf{M}\|_{2,1}$ by $\sum_i \sqrt{\|\mathbf{m}^i\|_2^2 + \zeta}$. Apparently, $\sum_i \sqrt{\|\mathbf{m}^i\|_2^2 + \zeta}$ reduces to $\|\mathbf{M}\|_{2,1}$ and the perturbed objective reduces to our original objective in Eq. (10), when $\zeta \to 0$. In the sequel of this document, we implicit apply this replacement for all $\|\cdot\|_{2,1}$.

Taking the derivative of J with respect to $\mathbf{A}_k$, and

setting the derivative as zero, we have:

$$\frac{\partial J}{\partial \mathbf{A}_k} = -2\mathbf{D}^T\mathbf{X}_k\mathbf{U}_k + 2\mathbf{D}^T\mathbf{D}\mathbf{A}_k\mathbf{U}_k + 2\lambda\mathbf{V}_k\mathbf{A}_k = 0 \ , \tag{2}$$

where $\mathbf{U}_k$ is a diagonal matrix with the i-th diagonal element as $1/\left(2\left\|(\mathbf{X}_k)_i - \mathbf{D}\left(\mathbf{A}_k\right)_i\right\|_2\right)$ and $\mathbf{V}_k$ is a diagonal matrix with the i-th diagonal element as $1/\left(2\left\|(\mathbf{A}_k)^i\right\|_2\right)$. Here $(\mathbf{X}_k)_i$ is the i-th column of $\mathbf{X}_k$, $(\mathbf{A}_k)_i$ is the i-th column of $\mathbf{A}_k$ and $(\mathbf{A}_k)^i$ is the i-th row of $\mathbf{A}_k$. Because $\mathbf{U}_k$ is a diagonal matrix, the equation in Eq. (2) can be decoupled into n_k subproblems for each column of $\mathbf{A}_k$, *i.e.*, $(\mathbf{A}_k)_i$ $(1 \le i \le n_k)$, as following:

$$(\mathbf{U}_k)_{ii}\,\mathbf{D}^T(\mathbf{X}_k)_i = \left((\mathbf{U}_k)_{ii}\,\mathbf{D}^T\mathbf{D} + \lambda\mathbf{V}_k\right)(\mathbf{A}_k)_i \ , \tag{3}$$

which is a linear equation and can be efficiently solved. Here $(\mathbf{U}_k)_{ii}$ is the i-th diagonal element of $\mathbf{U}_k$. Upon solution, $\widetilde{\mathbf{A}}$ can be constructed from the resulted $\mathbf{A}_k$ $(1 \le k \le K)$.

Note that, both $\mathbf{U}_k$ and $\mathbf{V}_k$ are dependent on $\mathbf{A}_k$ and $\mathbf{D}$. Therefore they are unknown variables and can be seen as two latent variables of the objective $J_{\text{RD-STL}}$, which can be solved under the same alternative optimization framework.

Second, when fixing $\widetilde{\mathbf{A}}$, we need to solve the following problem:

$$\min_{\mathbf{D}\in\mathcal{C}} \left\|(\widetilde{\mathbf{X}} - \mathbf{D}\widetilde{\mathbf{A}})^T\right\|_{2,1} \ . \tag{4}$$

Instead of solving Eq. (4), under the framework of alternating optimization, we solve the following problem:

$$\min_{\mathbf{D}\in\mathcal{C}} \left\|(\hat{\mathbf{X}} - \mathbf{D}\hat{\mathbf{A}})^T\right\|_{\text{F}}^2 \ , \tag{5}$$

where $\hat{\mathbf{X}} = \widetilde{\mathbf{X}}\,(\mathbf{U})^{\frac{1}{2}}$ and $\hat{\mathbf{A}} = \widetilde{\mathbf{A}}\,(\mathbf{U})^{\frac{1}{2}}$. Here, $\mathbf{U}$ is a diagonal matrix, whose i-th diagonal element is $u_{ii} = 1/\left\|\tilde{\mathbf{x}}_i - \mathbf{D}\tilde{\mathbf{a}}_i\right\|_2$, *i.e.*, $\mathbf{U} = \text{diag}\left(\text{diag}\left(\mathbf{U}_1\right), \ldots, \text{diag}\left(\mathbf{U}_K\right)\right)$. Eq. (5) can be solved following (Lee et al., 2007), which solves the Lagrangian dual problem and much faster than standard QCQP solver.

Finally, upon the solved $\mathbf{A}_k$ and $\mathbf{D}$, $\mathbf{U}_k$ (thereby $\mathbf{U}$) and $\mathbf{V}_k$ are updated by their respective definitions.

The whole procedures to optimize $J_{\text{RD-STL}}$ Eq. (10) is summarized in Algorithm 1. The convergence of the algorithm is guaranteed by Theorem 1 in the main text of the paper, which has been rigorously proved. Moreover, when the objective value in the iterations remains unchanged, due to step 3 of Algorithm 1, the K.K.T. condition in Eq. (2) or Eq. (3) is satisfied, which indicates that the objective function reaches the optimal value[4].

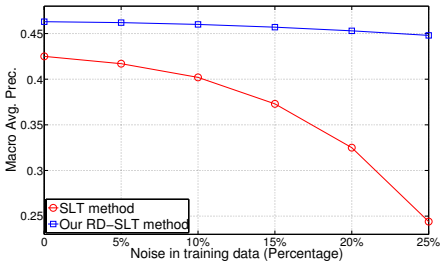

Figure 1. Comparison of the robustness of the proposed RD-STL method and STL method against noises in labeled images on TRECVID 2007 data set. Obviously, our method is impacted much less by the noises.

3. Study of the robustness against noises in labeled images

Although we introduce robustness in the proposed RD-STL method mainly addressing the outlier samples in unlabeled images due to their randomness and high heterogeneity, our model is also robust against noises in training images because of the $\ell_{2,1}$ loss function used in J_{RD} in Eq. (9). We evaluate this by manually introducing noises in the labeled images on TRECVID 2007 data set. Specifically, for each class we randomly pick up a certain percentage of labeled images and set their labels to be incorrect, to emulate noises. When we vary the amount of the imposed noises, we examine the classification performances of the proposed method, which are shown in Figure 1. We also report the classification accuracies of SLT (Raina et al., 2007) method at every corresponding noise condition for comparison. From Figure 1 we can see that the classification performance of our method does not degrades much when the amount of noises increases, whereas the classification performance of SLT method drops significantly with the increase of imposed noises. These results are consistent with the mathematical formulations of these two methods in that SLT method uses squared ℓ_2 loss function while the reconstruction errors in proposed RD-STL method are not squared. Therefore, the proposed method is more robust against noises in training data, which adds to its practical value.

[4]We rigorously proved that the objective function value is non-increasing during iterations. For most machine learning problems, such proofs provide some guarantee of the algorithm. In general, proving the solution converge to a fixed point using Cauchy theorem is much harder, and such proofs exits for only very limited cases.

4. Improved single-label image classification

Besides evaluating the proposed methods in multi-label image data sets as in the main text of the paper, we also evaluate our new approach on the three following single-label image data sets.

Caltech-101 data set[5] contains 8677 images of objects, belonging to 101 categories. Following (Dueck & Frey, 2007), besides using the full Caltech-101 data set, we also use the following two subsets of the data, which leads to two different classification tasks: the 7-class subset includes Faces, Motorbikes, Dolla-Bill, Garfield, Snoopy, Stop-Sign, Windsor-Chair, and has 441 images in total; and the 20-class subset includes Faces, Leopards, Motorbikes, Binocular, Brain, Camera, Car-Side, Dollar-Bill, Ferry, Garfield, Hedgehog, Pagoda, Rhino, Snoopy, Stapler, Stop-Sign, Water-Lilly, Windsor-Chair, Wrench, Yin-Yang, and has 1230 images. contains 8677 images of objects, belonging to 101 categories. Following (Dueck & Frey, 2007), we use two subsets of the data: a 7-class subset with 441 images and a 20-class subset with 1230 images.

MSRC-v1 data set[6] contains 240 images with 9 classes. Following (Lee & Grauman, 2009), we refine the data set to get 7 classes including tree, building, airplane, cow, face, car, bicycle, and each refined class has 30 images. Compared to the Caltech-101 data set, MSRC-v1 data set has more clutter and variability in the objects appearances. **MSRC-v1** data set contains 240 images with 9 classes. Following (Lee & Grauman, 2009), we refine the data set to get 7 classes, each of which has 30 images.

Following (Vedaldi & Fulkerson, 2008), we extract DSIFT features for both labeled and unlabeled images. Following (Wang et al., 2009a), we resize the images to 256×256 and extract features with grid size of 5 pixels. As a result, 2601 DSIFT features of are exacted for every image. The experimental results using DSIFT image descriptors under the same settings as Section 4 are reported in Table 1. Besides the same observations reported in the main text, we can see that our method is able to achieve state-of-the-art classification performance, which adds to its practical value.

[5]http://www.vision.caltech.edu/Image_Datasets/Caltech101/

[6]http://research.microsoft.com/en-us/projects/objectclassrecognition/

5. Image Classification with Irrelevant Unlabeled Data

One of the main advantage of self-taught learning lies in that it is able to leverage the large amount of inexpensive unlabeled data. Besides evaluating the performance of the proposed methods when unlabeled data are relevant to the labeled data as in Section 4 of the main text of the paper and Section 4 in this supplementary document, though from different distributions, we also evaluate our new method when the unlabeled data are irrelevant. To emulate this by generating random vectors as unlabeled data, and the classification performance on the four single-label image data sets are reported in the second row of Table 2. As a baseline, we also report the performance of our method when it does not utilize the unlabeled, which is equivalent to solve the following problem:

$$J_{\mathrm{RD}}(\mathbf{D},\mathbf{A}) = \left\| \left(\tilde{\mathbf{X}} - \mathbf{D}\tilde{\mathbf{A}} \right)^T \right\|_{2,1} + \lambda \sum_{k=1}^{K} \|\mathbf{A}_k\|_{2,1} \quad , \quad (6)$$
$$\underset{\mathbf{D}\in\mathcal{C}}{}$$

where $\tilde{\mathbf{X}} = [\mathbf{X}_1, \ldots, \mathbf{X}_K]$ and $\tilde{\mathbf{A}} = [\mathbf{A}_1, \ldots, \mathbf{A}_K]$. The results when not using unlabeled data are shown in the first row of Table 2. From the results we can see that, when using irrelevant unlabeled data, the performance of our method is about same as the that when not using any unlabeled data. This is because our new method is able to automatically identify relevant dictionary bases by the structured sparse regularization introduced in Eqs.8 9 to learn an adaptive dictionary, such that irrelevant, or even harmful data, will not be able to impact the classification performance of our new self-taught learning method.

References

Dueck, D. and Frey, B.J. Non-metric affinity propagation for unsupervised image categorization. In *ICCV*, 2007.

Gehler, P. and Nowozin, S. On feature combination for multiclass object classification. In *ICCV*, 2009.

Gorodnitsky, I.F. and Rao, B.D. Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm. *IEEE Transactions on Signal Processing*, 45(3):600–616, 1997.

Lee, H., Battle, A., Raina, R., and Ng, A.Y. Efficient sparse coding algorithms. In *NIPS*, 2007.

Lee, Y.J. and Grauman, K. Foreground focus: Unsupervised learning from partially matching images. *IJCV*, 85(2):143–166, 2009. ISSN 0920-5691.

Table 1. Classification accuracies of the compared methods in the three single-label image classification tasks.

Methods	Caltech-101 (7 class)	Caltech-101 (20 class)	Caltech (full)	MSRC
SVM	0.661	0.570	0.537	0.792
TSVM	0.674	0.567	0.536	0.801
GF	0.541	0.466	0.431	0.698
KTW	0.649	0.545	0.513	0.799
STC	0.621	0.504	0.488	0.767
STL	0.672	0.586	0.576	0.820
J-STL	0.716	0.607	0.601	0.826
R-STL	0.874	0.796	0.759	0.860
RA-STL	0.884	0.810	0.753	0.867
D-STL	0.835	0.764	0.741	0.837
RD-STL	**0.921**	**0.842**	**0.784**	**0.891**

Table 2. Classification accuracies of when using irrelevant unlabeled data.

Methods	Caltech-101 (7 class)	Caltech-101 (20 class)	Caltech (full)	MSRC
RD-STL (no unlabeled data)	0.813	0.754	0.732	0.821
RD-STL (random unlabeled data)	0.821	0.752	0.730	0.824

Naphade, M., Kennedy, L., Kender, JR, Chang, SF, Smith, JR, Over, P., and Hauptmann, A. LSCOM-lite: A light scale concept ontology for multimedia understanding for TRECVID 2005. Technical report, Technical report, IBM Research Tech. Report, RC23612 (W0505-104), 2005.

Raina, R., Battle, A., Lee, H., Packer, B., and Ng, A.Y. Self-taught learning: Transfer learning from unlabeled data. In *ICML*, 2007.

Vedaldi, A. and Fulkerson, B. VLFeat: An open and portable library of computer vision algorithms, 2008.

Wang, C., Blei, D., and Li, F.F. Simultaneous image classification and annotation. In *CVPR*, 2009a.

Wang, H., Huang, H., and Ding, C. Image annotation using multi-label correlated Green's function. In *ICCV*, 2009b.