L_1 Graph Based Sparse Model for Label De-noising

Xiaobin Chang
x.chang@qmul.ac.uk
Tao Xiang
t.xiang@qmul.ac.uk
Timothy M. Hospedales
t.hospedales@qmul.ac.uk

School of Electronic Engineering and
Computer Science
Queen Mary, University of London
London, E1 4NS
United Kingdom

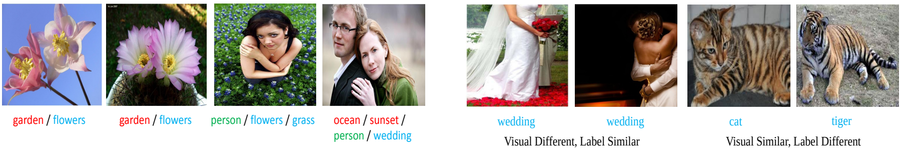
Abstract

The abundant images and user-provided tags available on social media websites provide an intriguing opportunity to scale vision problems beyond the limits imposed by manual dataset collection and annotation. However, exploiting user-tagged data in practice is challenging since it contains many noisy (incorrect and missing) labels. In this work, we propose a novel robust graph-based approach for label de-noising. Specifically, the proposed model is built upon (i) label smoothing via a visual similarity graph in a form of L_1 graph regulariser, which is more robust against visual outliers than the conventional L_2 regulariser, and (ii) explicitly modelling the label noise pattern, which helps to further improve de-noising performance. An efficient algorithm is formulated to optimise the proposed model, which contains multiple robust L_1 terms in its objective function and is thus non-trivial to optimise. We demonstrate our model's superior de-noising performance across the spectrum of problems from multi-class with label noise to real social media data with more complex multi-label structured label noise patterns.

1 Introduction

Constructing large manually annotated datasets [19] that conventionally required to scale up visual recognition can be prohibitively expensive. Therefore, the idea of exploiting the enormous amount of freely accessible visual data and associated tags on social media sites, such as Flickr, has attracted increasing attention [5, 8]. Nevertheless, the user-provided tags on social media sites can be extremely noisy, containing both incorrect and missing labels as illustrated in Fig. 1(a). Directly learning from noisy labels brings negative impacts on model performance [12]. Therefore, many studies focus on inferring more accurate and complete labels from the noisy ones through label de-noising [6, 12, 21, 28] (also called tag refinement [18, 37]) methods. Some of them consider either the incorrect labels [12, 22] or missing labels [3, 6, 27, 35] only. In this works, both types of label noise are considered.

In order to rectify incorrect and missing labels, various cues can be exploited. Label correlation is used to constrain predicted label sets with label co-occurence statistics [3, 29]. However, the required co-occurrence is typically estimated from noisy labels which limits its reliability. Visual appearance smoothness is another important cue for label de-noising. One strategy to exploit this is training classifiers based on appearance [1, 22, 24], but reliability is again limited as the classifier itself suffers from being trained with noisy labels. As an alternative mechanism to exploit visual appearance, visual similarity graph is widely used [6, 11, 27, 28, 29, 31]. It avoids incorporating noisy labels and directly exploits the intuition that labels should vary smoothly with visual appearance. As a result, visual similarity is more reliable than label correlation and visual appearance classifier cues for label de-noising. In practice, it is usually implemented through an optimisation with a Laplacian graph regularisation term [25, 36], to penalise label assignments that do not vary smoothly on the visual similarity graph.


garden / flowers
garden / flowers
person / flowers / grass
ocean / sunset / person / wedding
wedding
wedding
cat
tiger
Visual Different, Label Similar
Visual Similar, Label Different

(a) Noisy labels illustration (b) Visual outliers illustration

Figure 1: Illustrations of noisy labels (a) and visual outliers (b). Red indicates incorrect labels, green missing labels and blue correct labels.

Two challenges are identified in this work in effectively exploiting the visual similarity cue for label de-noising. The first challenge is the existence of visual outliers. As illustrated in Fig. 1(b), even when two images look very different, they can share the same label; while visually similar images can have different labels. The Laplacian graph regularisation term conventionally used to model the visual similarity cue is sensitive to such outliers since it is effectively an L_2 norm penalty. This limits its label de-noising efficacy. Inspired by the success of sparse learning for noisy problems in other vision tasks [33, 34], we propose a novel L_1 based visual similarity regulariser, which improves outlier robustness in visual similarity modelling and thus enhances label de-noising performance.

A second key challenge is the existence of noise *patterns* in both incorrect and missing labels. As shown in Fig. 1(a), in the NUS-WIDE dataset [7], flower images are consistently mislabelled as 'garden', while users typically neglect to annotate 'person' in images with people. In practice, such patterned label noise occurs simultaneously with random label noise (e.g., the fourth image of Fig. 1(a)). However, few existing studies explicitly consider label noise patterns. Among the few exceptions, [9, 30] require a set of noise free (or less noisy) labels to estimate the noise pattern. However, this strong assumption undermines the initial scalability motivation of learning from labels in the wild. To estimate noise patterns purely with noisy labels, a further disambiguating cue is necessary. Visual appearance classifiers are used in conjunction with solely noisy labels to estimate noise patterns in [4, 16]. However, as mentioned earlier, this appearance cue is unreliable due to classifier training on noisy labels. Therefore, we instead use the visual similarity cue to provide the required disambiguating prior for noise pattern estimation. Specifically, we simultaneously model a novel L_1 based visual similarity graph (for visual outlier robustness), and learn a label noise pattern with a L_1 norm loss, which increases robustness to a variable numbers of noisy labels per image.

The main contributions of our proposed model, L_1 Graph based Sparse model with explicit noise Pattern modelling (L_1GSP), can be summarised as follows: (i) an L_1 based visual similarity graph regulariser is introduced to ensure labels vary smoothly with visual similarity, while being robust to visual outliers; (ii) we explicitly learn a transition matrix to model the label noise pattern along with our robust graph regulariser (iii) the resulting objective function has two L_1 norm terms for robustness to both visual outliers and multiple label errors. Optimisation of such an objective is non-trivial, so an efficient algorithm is formulated to solve it. Experimental analysis demonstrates our model's superior performance compared to the baseline methods at label de-noising, as well as the ultimate consequences of de-noising in terms of improving a trained classifier's performance on testing data.

2 Methodology

Our proposed L_1 Graph based Sparse model with explicit noise Pattern modelling (L_1GSP), takes images and associated noisy labels for label de-noising. L_1GSP consists of two key components, the robust L_1 visual similarity graph regulariser and the robust L_1 label regulariser with label noise pattern modelling, which are formulated in Sec. 2.1. The optimisation algorithm to solve the double L_1 objective is introduced in Sec. 2.2.

Notation We denote a set of N training samples $X = (x_1, ..., x_N)$ and the associated noisy labels $Y = (y_1, ..., y_N)^T$, where $x_i \in \mathcal{R}^D$ is a feature vector computed from the i^{th} image and $y_i \in \{0,1\}^C$. Thus there are C potential labels where $Y_{ij} = 1$ indicates that instance x_i has label $j \in C$. We also consider the multi-label setting, so more than one class can be assigned to an instance and y_i can be an all zero vector, meaning that x_i has no label. Label de-noising is to estimate the (unknown) set of cleaner labels $\hat{Y}$ given noisy labels Y and visual data X.

2.1 Similarity Graph and Noise Pattern Model Formulation

Visual Similarity Graph A visual similarity graph is typically used to express the prior belief that labels should vary smoothly with visual similarity. This is typically formalised through the graph Laplacian matrix. Specifically, we use a Gaussian kernel to compute a weight matrix $W = \{w_{ij}\}_{N\times N}$ from X, so that W_{ij} indicates the similarity between samples x_i and x_j. The normalised Laplacian matrix L is given by $L = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ where I is a $N \times N$ identity matrix and D is a diagonal matrix where $D_{ii} = \sum_j W_{ij}$.

A simple de-noising strategy to exploit the Laplacian graph as a regulariser [36] is:

$$\min_{\hat{Y}} tr(\hat{Y}^T L \hat{Y}) + \gamma ||\hat{Y} - Y||_F^2, \tag{1}$$

where $tr(\bullet)$ is the trace norm operator and γ controls the weights of the two terms. This optimisation problem aims to find the cleaner labels $\hat{Y}$ that are not only close to the observed noisy labels Y (second term) but also constrained by the visual similarity graph (first term). However, as explained earlier, the conventional Laplacian graph is not robust to visual outliers. This is because that despite its trace norm form in Eq. (1), it is derived from a L_2 norm term and thus sensitive to outlying training samples. Moreover, the label loss term $||\hat{Y} - Y||_F^2$ neither models label noise patterns nor is robust to noisy labels.

L_1 Visual Similarity Graph To define a more robust visual similarity regulariser, we propose a novel L_1 visual similarity graph regulariser. Since the Laplacian matrix L is symmetric, it can be reformulated as the decomposition: $L = V\Sigma V^T = (\Sigma^{\frac{1}{2}} V^T)^T \Sigma^{\frac{1}{2}} V^T = S^T S$, where $S = \Sigma^{\frac{1}{2}} V^T$, Σ is a diagonal matrix, Σ_{ii} is the eigenvalue of L and each column in V is the eigenvector of L. This leads to a re-expression of the trace norm as:

$$tr(\hat{Y}^T L \hat{Y}) = tr(\hat{Y}^T S^T S \hat{Y}) = ||S\hat{Y}||_F^2 \tag{2}$$

Our robust L_1 visual similarity graph regulariser is defined by transforming the Frobenius norm $||\bullet||_F^2$ in Eq. (2) into $||\bullet||_1$ (where $||K||_1 = \sum_{ij}|k_{ij}|$), thus leading to:

$$||S\hat{Y}||_F^2 \Rightarrow ||S\hat{Y}||_1 \tag{3}$$

L_1 Label Regulariser with Noise Pattern The robust L_1 visual similarity graph regularisation term improves robustness to visual outliers, but the conventional label loss term ($||\hat{Y}-Y||_F^2$) can only deal with moderate random noise, and it is still vulnerable to strong and patterned label noise. To address this, we first introduce a $C \times C$ transition matrix Q to encode the label-noise pattern. Second, inspired by the success of L_1 losses in strong noisy data problems [26, 34], we introduce an L_1 label regulariser leading to the loss $||\hat{Y}-YQ||_1$.

Objective Function of L_1GSP Computing Q is trivial if both ground-truth and noisy labels are known. However, we aim to avoid this assumption, as it undermines the motivation for learning with noisy labels. To this end, the proposed L_1 visual similarity graph is used to provide the required disambiguating prior for learning a noise pattern purely from the noisy labels. Jointly estimating the cleaner labels $\hat{Y}$ and the noise pattern Q leads to the optimisation problem:

$$\min_{\hat{Y},Q} ||S\hat{Y}||_1 + \gamma||\hat{Y}-YQ||_1 + \frac{\beta}{2}||Q||_F^2. \tag{4}$$

where the two weighting factors γ and β control the relative strengths of the graph regulariser, the label loss, and a regulariser on Q to prevent overfitting.

2.2 Optimisation

Our label de-noising framework L_1GSP exploits both robust visual similarity regularisation and robust label regulariser with noise pattern modelling. However, the optimisation of Eq. (4) is non-trivial because the two L_1 norm terms make it significantly harder [32] than the more common case of a single L_1 norm. An alternating optimisation procedure is first formulated in order to simplify the joint optimisation of $\hat{Y}$ and Q. Specifically, we optimise $\hat{Y}$ and Q iteratively by first breaking Eq. (4) into the following alternating objectives:

$$\hat{Y}^* = \arg\min_{\hat{Y}} ||S\hat{Y}||_1 + \gamma||\hat{Y}-YQ^*||_1, \tag{5}$$

$$Q^* = \arg\min_{Q} \gamma||\hat{Y}^*-YQ||_1 + \frac{\beta}{2}||Q||_F^2, \tag{6}$$

where Q^* is an identity matrix initially. We next explain how to solve each of these in turns.

Robust Graph Solution We first focus on solving Eq. (5), which contains the proposed L_1 visual similarity graph, for cleaner labels $\hat{Y}$. Here Q^* is a constant from initial condition or the solution of Eq. (6) and Y is the given noisy labels, which is also a constant. Therefore, we denote a new constant $Z \equiv YQ^*$ in Eq. (5) for simplification. Inspired by [20], we introduce an intermediate variable $F \in R^{N\times C}$, which leads to a new F-norm term,

$$\min_{\hat{Y},F} \frac{1}{2}||\hat{Y}-F||_F^2 + \lambda||SF||_1 + \gamma||\hat{Y}-Z||_1 \tag{7}$$

Eq. (7) can then be solved by alternating optimisation for F and $\hat{Y}$,

$$F^* = \arg\min_{F} \frac{1}{2}||F-\hat{Y}^*||_F^2 + \lambda||SF||_1, \tag{8}$$

$$\hat{Y}^* = \arg\min_{\hat{Y}} \frac{1}{2}||\hat{Y}-F^*||_F^2 + \gamma||\hat{Y}-Z||_1, \tag{9}$$

where $\hat{Y}^* = Z$ initially. Each step in this alternating optimisation procedure is now simpler since only one L_1 norm term is present. In particular, the sub-problem Eq. (9) has a closed-form solution [2]: $\hat{Y}^* = soft\_thr(F^*, Z, \gamma)$, where $soft\_thr(\cdot, \cdot, \gamma)$ is a piecewise soft-thresholding function. We define $z = soft\_thr(x, y, \gamma)$ as:

$$z = \begin{cases} z_1 = max(x-\gamma, y), f_1 \leq f_2 \\ z_2 = max(0, min(x+\gamma, y)), f_1 > f_2 \end{cases},$$

where $f_1 = (z_1 - x)^2 + 2\gamma|z_1 - y|$ and $f_2 = (z_2 - x)^2 + 2\gamma|z_2 - y|$. The sub-problem Eq. (8) is not tractable when the data N is large since S is a $N \times N$ matrix. This can be addressed by taking a small fraction m of the Laplacian graph L's eigenvectors. In particular, we can significantly reduce the dimension of F by decomposing it to $F = V_m A$, where $A = \{a_{ij}\}_{m \times C}$ collects the reconstruction coefficients and $V_m \in R^{N \times m}$ contains the eigenvectors with the smallest eigenvalues of L. Thus the sub-problem Eq. (8) becomes:

$$\begin{aligned} &\arg\min_A \frac{1}{2}||V_m A - \hat{Y}^*||_F^2 + \lambda||SV_m A||_1 \\ =&\arg\min_A \sum_{j=1}^{C} (\frac{1}{2}||V_m A_{.j} - \hat{Y}^*_{.j}||_2^2 + \lambda \sum_{i=1}^{m} \Sigma_{ii}^{\frac{1}{2}} |a_{ij}|), \end{aligned} \tag{10}$$

where $\hat{Y}^*_{.j}$ and $A_{.j}$ denote the j^{th} column of $\hat{Y}^*$ and A, respectively. The orthogonality of V is exploited here to simplify the term $||SV_m A||_1$. The L_1 optimisation problem in the final line of Eq. (10) can be solved by existing solvers and we use $L_1 General$ [23].

Transition Matrix Solution Next we solve Eq. (6) for estimating the transition matrix Q. Note that $\hat{Y}^*$ is the solution from Eq. (5) and thus a constant in Eq. (6). By introducing an intermediate variable $E = \hat{Y}^* - YQ$ in Eq. (6), we get the new objective: $\min_{Q,E} \gamma||E||_1 + \frac{\beta}{2}||Q||_F^2 + \frac{1}{2}||E - \hat{Y}^* + YQ||_F^2$, which can in turn be solved by alternately optimising the following two simple objectives,

$$E^* = \arg\min_E \gamma||E||_1 + \frac{1}{2}||E - \hat{Y}^* + YQ^*||_F^2, \tag{11}$$

$$Q^* = \arg\min_Q \frac{\beta}{2}||Q||_F^2 + \frac{1}{2}||E^* - \hat{Y}^* + YQ||_F^2. \tag{12}$$

3 Experiments

We validate our proposed de-noising method on multi-class data with synthetic noise, where the strengths and patterns of label noise can be experimentally controlled, as well as on two multi-label datasets with user-provided tags reflecting real label noise distributions.

3.1 Datasets

Datasets We evaluate our model on MNIST [17], Pascal VOC 2007 [10] and NUS-WIDE [7]. The **MNIST** dataset contains 60K training and 10K testing samples of 10 classes. We use raw pixels as image features. Synthetic label noise is simulated by flipping specified proportions of ground-truth labels. Flips are generated either (i) with uniform random label noise so every wrong label is equiprobable, (ii) based on a synthetic pattern illustrated in Fig. 3(a) (patterned label noise), or (iii) a hybrid of random and patterned label noise. In

Pascal VOC 2007, we take the Flickr tags [14] associated with the original source Pascal images as noisy labels. Preprocessing based on standard natural language processing procedures is performed to keep only the noisy tags that are identical or synonyms of the 20 object classes. This results in 1,521 training samples with noisy labels and 3,490 training samples without any label – this reflects the fact that the user provided tags are extremely sparse. The standard test set (4,952 images) is used for testing on the 20 classes. In **NUS-WIDE**, we download the raw images still available on Flickr, filtering out the low quality (too small) images. Following [13], we also discard images for which all labels are absent. This leaves about 100,000 images, which we randomly partition into 60K for training and 40K testing on the 81 ground-truth concepts. For both Pascal and NUS-WIDE, we extract 4096-dimension CNN feature vectors with the pre-trained Caffe Reference network [15].

3.2 Settings

Parameter Settings The proposed model, L_1GSP, has three free parameters: γ and β in Eq. (4), and λ in Eq. (7). However, since for label de-noising the ground-truth labels are unavailable, we cannot tune the model parameters by cross validation. In our experiments, we fix the free parameters $\lambda = 0.05$ and $\beta = 1.0$ across different datasets, and set $\gamma = 0.1, 0.1$ and 0.001 for MNIST, NUS-WIDE and Pascal, respectively. γ for Pascal is smaller than the others because a large proportion of its training labels are totally missing – smaller γ means the de-noising relies less on the given noisy labels. The model's sensitivity to the parameters is discussed in Sec. 3.4.

Classifier Training Once the de-noising process produces cleaner labels $\hat{Y}$, we can use them to train a classifier. However, the cleaner labels are not necessarily $\{0,1\}$ vectors. Therefore, in each de-noised sample, we take the τ classes with largest scores as 1s and the others as 0s. For MNIST we use $\tau = 1$ (single label). For multi-label Pascal and NUS-WIDE, τ is set as $2,3$ respectively based on the average number of labels per training image. For MNIST, we train the LeNet [17] CNN on the de-noised labels. For NUS-WIDE we fine-tune the Caffe Reference Net [15]. For the smaller Pascal07, we employ SVM instead of CNN.

Evaluation Metrics MNIST: we evaluate de-noising performance by the remaining training label error rate, and testing performance by accuracy. Pascal VOC 07: we use mean Average Precision (mAP) metric to evaluate both de-noising and testing performance. NUS-WIDE: we follow [18] in using the complementary metrics of per-class mAP(mAPc) and per-image mAP (mAPi), for de-noising and testing.

Competitors Three label de-noising methods are used for comparison: L_2 Visual similarity Graph model (L_2VG) follows the method in [36] as given in Eq. (1). L_2 Visual similarity Graph and Label similarity Graph model (L_2VGLG) [27] combines the visual similarity cue and label correlation cue into the following optimisation problem:

$$\min_{\hat{Y}} \frac{\lambda_X}{2} tr(\hat{Y}^T L_X \hat{Y}) + \frac{\lambda_C}{2} tr(\hat{Y} L_C \hat{Y}^T) + \gamma ||\hat{Y} - Y||_F^2, \tag{13}$$

where both visual similarity L_X and label correlation L_C are exploited as graph regularisers. Moreover, a Robust PCA (RPCA) based de-noising method [37] is also used for comparison. RPCA uses a L_1 error term between cleaner and noisy labels (similar to our L_1 label regularisation term) and imposes a low-rank constraint on the de-noised labels together with the conventional visual similarity (Laplacian) graph and label correlation graph. L_2VGLG and RPCA are only used for multi-label problems (NUS-WIDE and Pascal VOC 2007) since they rely on label correlation.

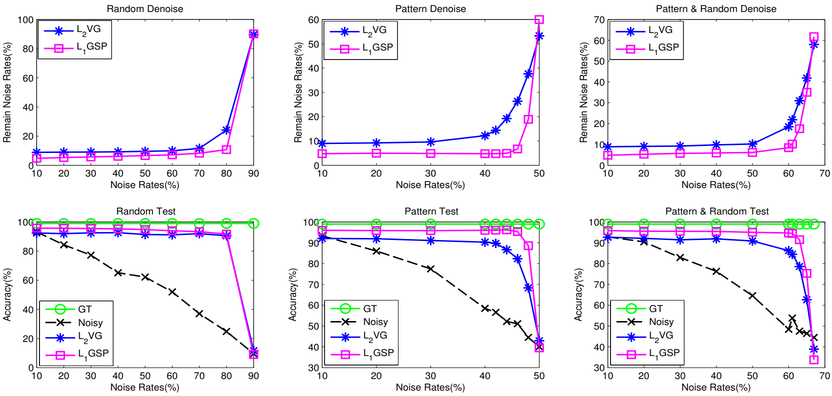


Figure 2: De-noising (first row: remaining noise rate) and Testing (second row: accuracy) performance on MNIST with varying label noise types and levels . Best viewed in color.

3.3 Results

MNIST The de-noising performance is shown in the first row of Fig. 2 for varying noise strengths and types. Our proposed L_1GSP (magenta) achieves consistently lower remaining noise rates than L_2VG at each noise level. The testing performance (Fig. 2, second row) reveals that the relative performance above carries over to testing when a classifier is trained on the de-noised labels. Directly learning from the noisy labels (black) results in poor performance, and the two compared de-noising models push the results much closer to the upper bound (green) of ground-truth labels with the proposed L_1GSP clearly better. In MNIST, the visual similarity is relatively reliable (less outliers) compared to Pascal and NUS-WIDE. The explanation for the performance margin of L_1GSP over L_2VG thus primarily lies in the L_1 label regulariser with noise pattern modelling in our method. More illustrations on label noise patterns and the learned transition matrices are in Sec. 3.4.

Pascal VOC 2007 This dataset is challenging for label de-noising due to the numerous visual outliers compared to MNIST, and the extremely noisy and sparse labels. De-noising results are summarised in Table 1. Our proposed model (L_1GSP) outperforms all the baseline methods (L_2VG, L_2VGLG and RPCA). This suggests that our proposed L_1 visual similarity regulariser is more robust to visual outliers than the L_2 norm term used by the baselines.

The testing results when noisy/de-noised labels are used to train a classifier are also given in Table 1. It is noteworthy that for testing, baseline models L_2VG and L_2VGLG give even worse results than training directly on the noisy labels (NL). This is largely due to the particularly poor performance on a few specific classes. In particular, the less robust L_2 similarity graph term induces a strong imbalance in the estimated labels for certain classes. For example, bottle and sofa are estimated as having 17 and 47 samples after de-noising, compared to the true 248 and 295 samples. Meanwhile, dog and horse are over-estimated as having 1408 and 2840 samples.

		GT	NL	L_2VG	L_2VGLG	RPCA	L_1GSP
De-noising	mAP	-	-	52.21	55.01	56.39	**60.09**
Testing	mAP	71.98	42.34	40.33	41.10	53.54	**58.66**

Table 1: Pascal VOC 2007 de-noising performance and testing performance (mAP, %). GT for Ground-truth; NL for Noisy Labels.

NUS-WIDE The de-noising and testing results are shown in Table 2. The proposed L_1GSP achieves significantly better results on both de-noising and testing than the baselines L_2VG and L_2VGLG. RPCA achieves similar mAPi performance to our model but significantly lower in mAP metric. It suggests that RPCA's de-noising performance on NUS-WIDE is imbalanced among classes and some classes are sacrificed to boost the others. On the mAPi testing metric, the L_2VG and L_2VGLG baseline methods make little improvement (less than 2%) on the Noisy Label lower bound, but our proposed L_1GSP improves the baseline by a good margin of about 10%. However, on the mAP metric, the improvements made by all de-noising methods are not very significant. This is because NUS-WIDE is a very imbalanced dataset. Some classes (e.g. map, earthquake) have very few (only dozens of) samples. Learning from such extremely rare classes is particularly hard, but these rare classes are penalised equally under the mAP metric and limit the average performance.

	De-noising		Testing	
	mAPc	mAPi	mAPc	mAPi
GT	-	-	47.76	74.31
NL	-	-	30.07	47.88
L_2VG	52.39	57.45	33.81	48.52
L_2VGLG	53.02	59.68	34.69	49.45
RPCA	48.89	64.10	31.20	54.21
L_1GSP	**58.46**	**66.98**	**35.70**	**57.84**

Table 2: De-noising (left) and testing (right) performance (mAP, %) on NUS-WIDE. GT for Ground-truth; NL for Noisy Labels.

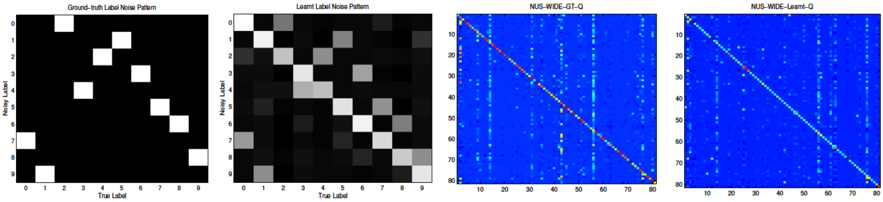
3.4 Further Analysis

Label Noise Patterns In our model the label noise pattern is represented by the transition matrix Q and iteratively learned along with cleaner labels $\hat{Y}$ without relying on ground-truth labels. The intrinsic noise patterns and the learned transition matrices are illustrated in Fig. 3. Fig. 3(a) shows the noise pattern used to synthesise patterned label noise in MNIST, e.g. a fixed portion of labels '8' are consistently flipped to label '6' as the patterned incorrect labels. The learned transition matrix Q (condition: 64% of hybrid label noise including 32% randomly incorrect and 32% with the given noise pattern) is shown in Fig. 3(b). Here we can see that the learned matrix includes: (i) a bright diagonal representing unchanged labels, (ii) a bright pattern of off diagonal elements matching the true label noise pattern in Fig. 3(a), and (iii) some weak background elements corresponding to random label noise. The true and learned transition matrix Q for NUS-WIDE are shown in Fig. 3(c) and Fig. 3(d), respectively. The true transition matrix Q is computed by $\min_{Q} ||\bar{Y} - YQ||_F^2 + ||Q||_F^2$ using ground-truth labels $\bar{Y}$. We see that our model estimates a very similar Q despite not having access to the ground-truth labels.

Contributions of Model Components The proposed L_1GSP has two L_1 regularisation terms: visual outlier robust L_1 similarity graph and L_1 label regulariser with label noise pattern modelling. In order to evaluate the contributions of these components for de-noising, we compare the proposed model (L_1GSP) with the following variants:

$\mathbf{L_1}$GS: Incorporate robust L_1 visual similarity term and robust L_1 label loss, but no noise pattern modelling: $\min_{\hat{Y}} ||S\hat{Y}||_1 + \gamma||\hat{Y} - Y||_1$;

$\mathbf{L_1}$VG$\mathbf{L_2}$L: L_1 Visual similarity Graph with L_2 Label regulariser: $\min_{\hat{Y}} ||S\hat{Y}||_1 + \gamma||\hat{Y} - Y||_F^2$;



(a) (b) (c) (d)

Figure 3: Ground-truth label noise pattern and learned transition matrices. (a) Ground-truth noise pattern used for noise synthesis in MNIST; (b) Transition matrix Q learned from 64% hybrid label noise by our model; (c) NUS-WIDE ground-truth transition matrix; (d) Learned NUS-WIDE transition matrix Q by our model.

L_2VGL_1L: L_2 Visual similarity Graph with L_1 Label regulariser: $\min_{\hat{Y}} ||S\hat{Y}||_F^2 + \gamma||\hat{Y} - Y||_1$;

L_2VG: L_2 Visual similarity Graph with L_2 Label regulariser as in Eq. (1).

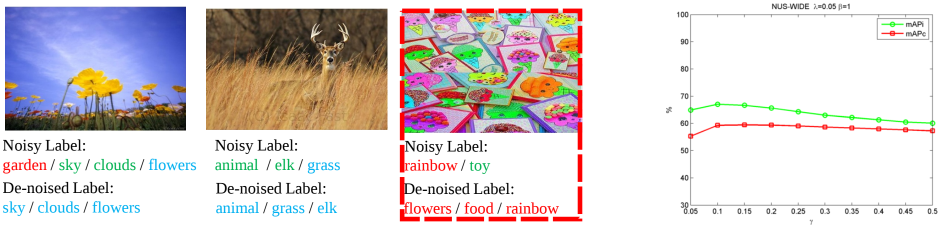
We use Pascal VOC 2007 to compare these variants, with results summarised in Table 3. Conclusions can be drawn as follows: (i) Having the L_1 norm for visual similarity graph makes the biggest contribution (the performance margin between L_1VGL_2L and L_2VG is 4.21%) to model performance. This is because visual similarity is an important but noisy cue for label de-noising and the proposed L_1 visual similarity graph regularisation term is more robust to the negative impacts of visual outliers than the L_2/Frobenius norm term; (ii) Comparing L_1GSP with L_1GS, we find that explicitly modelling the noise pattern helps to further boost the de-noising performance by 2.59%; (iii) L_1 label loss also contributes to the performance boosts because of its robustness to label noise; (iv) Our proposed L_1GSP model combines the two L_1 regularisation terms with explicit label noise pattern modelling. As a result, it achieves the best performance among the variants.

	L_2VG	L_2VGL_1L	L_1VGL_2L	L_1GS	L_1GSP
mAP	52.21	53.69	56.42	57.50	**60.09**

Table 3: Component contributions evaluated on Pascal VOC 2007 (mAP, %).

Qualitative Results Qualitative results of label de-noising are shown in Fig. 4(a). The first example shows that incorrect labels can be eliminated from the top ranking predictions of our de-noising model. The patterned incorrect label ‘garden’ does not appear in our de-noised labels and more relevant ones, such as ‘sky’ and ‘clouds’, show up instead. The effectiveness of the proposed model to recover missing labels is illustrated in the second image of Fig. 4(a), where the missing labels ‘elk’ and ‘animal’ appear in the top ranks of the de-noised labels. The last image of Fig. 4(a) shows a failure case using our model, which is mainly due to the unconventional appearance of toys. The predicted label ‘food’ may corresponds to the toy ice creams in the image, while labels ‘flowers’ and ‘rainbow’ are given may due to the colourful image content.

Parameter Robustness We found that our model is insensitive to λ and β and thus fix them as $\lambda = 0.05$ and $\beta = 1.0$ across the three different datasets we used. The sensitivity of our model’s de-noising performance with respect to γ is illustrated in Fig. 4(b). The results suggest that the impacts of different γs are small.



(a) Qualitative Results

(b) Parameter Robustness

Figure 4: (a) Illustrations of label de-noising results on NUS-WIDE dataset (top 3 scoring of the de-noised labels by our model are shown). Red indicates the incorrect labels, green for missing labels and blue for correct labels. Failure case in red dashed line; (b) Illustration of the effect of γ on our L_1GSP de-noising model with NUS-WIDE dataset.

4 Conclusion

We have provided a step towards the sought-after capability of learning from noisy labels in social media data by introducing a novel robust label de-noising model and formulating an efficient algorithm to solve it. The proposed model is based on a visual-outlier robust L_1 visual similarity graph regularisation term, and estimating the label noise pattern along with the visual similarity constraint to further improve label de-noising performance. In future work, we aim to integrate appearance modelling and feature representation learning into our de-noising model.

References

[1] Wenjuan An and Mangui Liang. Fuzzy support vector machine based on within-class scatter for classification problems with outliers or noises. *Neurocomputing*, 2013.

[2] Francis Bach, Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, et al. Convex optimization with sparsity-inducing norms. *Optimization for Machine Learning*, 2011.

[3] Wei Bi and James T Kwok. Multilabel classification with label correlations and missing labels. In *AAAI*, 2014.

[4] Jakramate Bootkrajang and Ata Kabán. Label-noise robust logistic regression and its applications. In *MLKDD*. 2012.

[5] Xinlei Chen and Abhinav Gupta. Webly supervised learning of convolutional networks. In *ICCV*, 2015.

[6] Zheng Chen, Minmin Chen, Kilian Q Weinberger, and Weixiong Zhang. Marginalized denoising for link prediction and multi-label learning. In *AAAI*, 2015.

[7] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao. Zheng. NUS-WIDE: A real-world web image database from national university of singapore. In *CIVR*, 2009.

[8] Santosh Divvala, Ali Farhadi, and Carlos Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*, 2014.

[9] Yunyan Duan and Ou Wu. Learning with auxiliary less-noisy labels. *TNNLS*, 2016.

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[11] Zheyun Feng, Songhe Feng, Rong Jin, and Anil K Jain. Image tag completion by noisy matrix recovery. In *ECCV*, 2014.

[12] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *TNNLS*, 2014.

[13] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894*, 2013.

[14] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Multimodal semi-supervised learning for image classification. In *CVPR*, 2010.

[15] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[16] Neil D Lawrence and Bernhard Schölkopf. Estimating a kernel fisher discriminant in the presence of label noise. In *ICML*, 2001.

[17] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

[18] Xirong Li, Tiberio Uricchio, Lamberto Ballan, Marco Bertini, Cees GM Snoek, and Alberto Del Bimbo. Socializing the semantic gap: A comparative survey on image tag assignment, refinement and retrieval. *arXiv preprint arXiv:1503.08248*, 2015.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*. 2014.

[20] Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In *ICML*, 2010.

[21] Zhongang Qi, Ming Yang, Zhongfei Mark Zhang, and Zhengyou Zhang. Mining partially annotated images. In *SIGKDD*, 2011.

[22] Zhongang Qi, Ming Yang, Zhongfei Mark Zhang, and Zhengyou Zhang. Mining noisy tagging from multi-label space. In *CIKM*, 2012.

[23] Mark Schmidt, Glenn Fung, and Rómer Rosales. Fast optimization methods for l1 regularization: A comparative study and two new approaches. In *ECML*. 2007.

[24] Guillaume Stempfel and Liva Ralaivola. Learning svms from sloppily labeled data. In *ICANN*. 2009.

[25] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 2007.

[26] John Wright, Allen Y Yang, Arvind Ganesh, Shankar S Sastry, and Yi Ma. Robust face recognition via sparse representation. *TPAMI*, 2009.

[27] Baoyuan Wu, Zhilei Liu, Shangfei Wang, Bao-Gang Hu, and Qiang Ji. Multi-label learning with missing labels. In *ICPR*, 2014.

[28] Baoyuan Wu, Siwei Lyu, and Bernard Ghanem. Ml-mg: Multi-label learning with missing labels using a mixed graph. In *ICCV*, 2015.

[29] Baoyuan Wu, Siwei Lyu, and Bernard Ghanem. Constrained submodular minimization for missing labels and class imbalance in multi-label learning. In *AAAI*, 2016.

[30] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, 2015.

[31] Wenxuan Xie, Zhiwu Lu, Yuxin Peng, and Jianguo Xiao. Graph-based multimodal semi-supervised image classification. *Neurocomputing*, 2014.

[32] Junfeng Yang and Yin Zhang. Alternating direction algorithms for l_1-problems in compressive sensing. *SIAM journal on scientific computing*, 2011.

[33] Lei Zhang, Meng Yang, and Xiangchu Feng. Sparse representation or collaborative representation: Which helps face recognition? In *ICCV*, 2011.

[34] Tianzhu Zhang, Bernard Ghanem, Si Liu, and Narendra Ahuja. Low-rank sparse learning for robust visual tracking. In *ECCV*. 2012.

[35] Feipeng Zhao and Yuhong Guo. Semi-supervised multi-label learning with incomplete labels. In *ICAI*, 2015.

[36] Dengyong Zhou and Bernhard Schölkopf. A regularization framework for learning from graph data. In *ICML workshop*, 2004.

[37] Guangyu Zhu, Shuicheng Yan, and Yi Ma. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *ICM*, 2010.