IEICE TRANS. INF. & SYST., VOL.E103–D, NO.10 OCTOBER 2020

PAPER

Completion of Missing Labels for Multi-Label Annotation by a Unified Graph Laplacian Regularization

Jonathan MOJOO[†a)], **Yu ZHAO**[†], **Muthu Subash KAVITHA**[†], ***Nonmembers***, **Junichi MIYAO**[†], ***Member***, ***and*** **Takio KURITA**[†], ***Fellow***

SUMMARY The task of image annotation is becoming enormously important for efficient image retrieval from the web and other large databases. However, huge semantic information and complex dependency of labels on an image make the task challenging. Hence determining the semantic similarity between multiple labels on an image is useful to understand any incomplete label assignment for image retrieval. This work proposes a novel method to solve the problem of multi-label image annotation by unifying two different types of Laplacian regularization terms in deep convolutional neural network (CNN) for robust annotation performance. The unified Laplacian regularization model is implemented to address the missing labels efficiently by generating the contextual similarity between labels both internally and externally through their semantic similarities, which is the main contribution of this study. Specifically, we generate similarity matrices between labels internally by using Hayashi's quantification method-type III and externally by using the word2vec method. The generated similarity matrices from the two different methods are then combined as a Laplacian regularization term, which is used as the new objective function of the deep CNN. The Regularization term implemented in this study is able to address the multi-label annotation problem, enabling a more effectively trained neural network. Experimental results on public benchmark datasets reveal that the proposed unified regularization model with deep CNN produces significantly better results than the baseline CNN without regularization and other state-of-the-art methods for predicting missing labels.

key words: *multi-label image annotation, regularization, missing labels*

1. Introduction

Recently, multi-label image annotation has achieved great progress in different domains such as multi-object recognition [21], [22], scene recognition [20], facial action detection [19], and medical diagnostic prediction [17], [18]. In an image annotation task, there is usually more than one label to annotate an image with, and these labels provide significant information such as location, features and events. However, it is difficult to prepare complete labels to ensure correct predictions for multi-label learning which makes training of neural networks with missing labels a challenging problem in automated image annotation.

Several recently developed techniques based on deep neural networks have produced acceptable annotation performances by utilizing the label relationships. The relationships between labels are derived using various strategies such as generating a tree-structured graph [30], using a structured inference neural network [29] and co-occurrence dependency [13]. However, these methods only implement internal relations for multi-label learning. To better understand ambiguous examples and minimize false predictions, multi-label image annotation requires not only internal relations between labels but also strong relations that are external to the ground truth [29].

Fig. 1 Example image demonstrating the consistency of predicted label (**arctic**) with the given labels (bear, snow, polar, tundra)

Incorporating both internal and external label relations in training a deep CNN could capture diverse contextual relations and specifically infer the remaining missing labels for a given set of labels for a particular image. For example, in Fig. 1, the missing label "arctic" can be correctly recovered, which is meaningful and directly related to the given label subset as well as to the visual content of the image. In this study, we propose a regularization term that combines strong dependencies derived from both internal and external similarities between labels. Regularization enables superior and effective fine-tuning of the network layers, thus improving its performance. This could be understood from earlier studies showing that incorporating semantic regularization in a deep network improves accuracy and convergence speed [25], [27], [28].

The contributions of this study can be summarized as follows: (1) A novel contextual regularization for CNN models is proposed for improved image annotation, which differs from conventional approaches, by introducing a unified internal and external graph Laplacian regularization term in the objective function of the CNN; (2) a strong contextual similarity between labels both internally and externally is generated by using Hayashi's quantification method-type III and the word2vec method, respectively; (3) extensive evaluation on three different datasets is performed to confirm whether a unified internal-external label-relation

Manuscript received December 9, 2019.
Manuscript revised May 3, 2020.
Manuscript publicized July 3, 2020.
[†]The authors are with Department of Information Engineering, Graduate School of Advanced Science and Engineering, Hiroshima University, Higashihiroshima-shi, 739–8511 Japan.
a) E-mail: jonathanmojoo@yahoo.com
DOI: 10.1587/transinf.2019EDP7318

regularization graph derived from co-occurrence data could produce better performance than individual regularization for an image annotation deep CNN.

2. Related Works

The problem of multi-label learning for data with missing-labels has been studied in a number of contexts. In this section we review related works on multi-label learning with a focus on label relations followed by a discussion of regularization in neural networks.

2.1 Multi-Label Learning with Label Relations

Co-occurrence distributions learned from the internal label space have been used to compensate for missing labels [13]. For example, the multi-label local correlation approach encodes the local influence of label correlations using the feature representation of each instance [3]. Other researchers have proposed parametric models, which combine pair-wise correlations of class labels to solve the multi-label learning problem [6]. Several works on multi-label learning have suggested that external knowledge of label relations could improve label prediction [4], [5]. For example, [12] generated the co-occurrence of pairs of labels using external knowledge for multi-label annotation. Furthermore, to adequately address the problem of missing labels, an integrated framework can be used to learn the complex correlations between labels for multi-label classification with missing labels [2]. Lee et al. [7] proposed that label relations observed in the external space can be used to identify multiple unseen class labels for each input instance for performing multi-label classification.

The above-mentioned methods independently handled internal or external label relations in learning algorithms, while our proposed model incorporates co-occurrence distributions of both internal and external label relations. Note that inter-relations between different labels have already been exploited in recent techniques by using a mixed graph to encode a network of label dependencies [16], a unified correlative multi-label method to classify the labels [24], and quadratic energy function graphs for constructing complete labels [26]. However, these algorithms are too complex and are not appropriate for large datasets.

2.2 Regularization Techniques in a Neural Network

The introduction of a label relation graph in the regularization term of a deep neural network model enables more efficient training and avoids over-fitting which in turn leads to better performance. Pengfei et al. [1] used semantic information to regularize the combination of two different neural network layers. Similarly, Yan et al. [28] implemented an attribute induced semantic regularization to tune the middle embedding layer. Mojoo et al. [13] combined the original objective function of a neural network with the graph Laplacian regularization term based on the internal co-occurrence dependency of the labels. Several works have focused on the loss function of a neural network to solve the multi-label learning with missing labels (MLML) problem with large-scale labels. Wu et al. [15] proposed a sub-modular objective function to handle the problem of large numbers of negative labels. Another study [14] exploited the structure of a specific loss function for the annotation problem. Inspired by these research techniques, we measure the effectiveness of a unified label-relation regularization graph in training a deep CNN for multi-label image annotation.

In our proposed method, we combine two different matrices. One matrix is the graph Laplacian matrix of all the Word2Vec similarities between labels in the dataset [11]. The Word2Vec similarity is calculated from a model trained on the Wikipedia dumped data [9]. The second matrix is the graph Laplacian matrix with all the co-occurrence-based label similarities calculated by Hayashi's quantification method-type III [10]. The combination of these two matrices is used as a regularization term added to the neural network's original objective function. Thus, the weights of the graph Laplacian matrix are calculated by using the similarities between the vectors obtained from both the internal and external label spaces. Hence, the regularization term in the proposed objective function introduces correlation information between each pair of labels in the training process and increases the co-occurrence probability of labels with high observed co-occurrence frequency.

3. Unified Approach to Image Annotation

The framework of the proposed multi-label image annotation model is presented in Fig. 2. We first present an overview of the label contextual similarity measurement by Hayashi's quantification method-type III and the Word2Vec word-embedding method. The proposed unified approach for multi-label image annotation is later described in Sect. 4.

3.1 Hayashi's Quantification Method-Type III

Hayashi's quantification method-type III (HQ-III) is applied in the understanding of categorical data, including cross-tabulation or contingency tables [10]. It is used to calculate the vector representations of each row and column by utilizing the information of co-occurrences. Suppose the cross-tabulation that records the frequency of co-occurrence is represented by $\boldsymbol{T} = [t_{ij}]$, where

$i = 1, 2, \cdots, M$ and $j = 1, 2, \cdots, N$. Using this table, we can extract the vector representations $\boldsymbol{q}_i$ and $\boldsymbol{q}_j$ of the i^{th} row and the j^{th} column respectively, by applying HQ-III. Note that the distance between these vectors becomes small if the pattern of responses in the cross-tabulation is identical. In this study, HQ-III is used to calculate the distance between the vectors of each label, which can explain the internal similarity between each pair of labels.

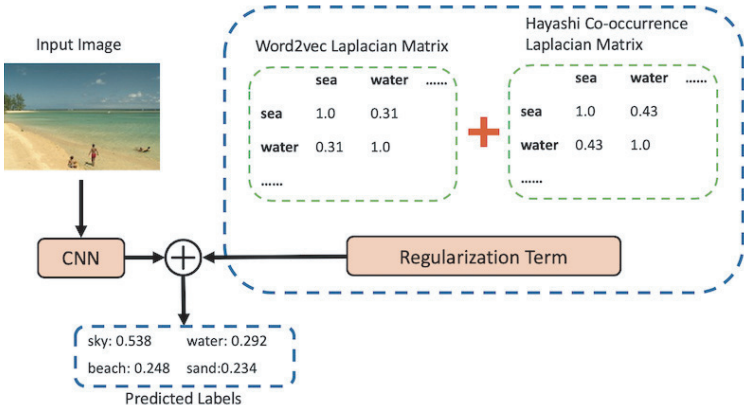

Fig. 2 Illustration of the proposed approach

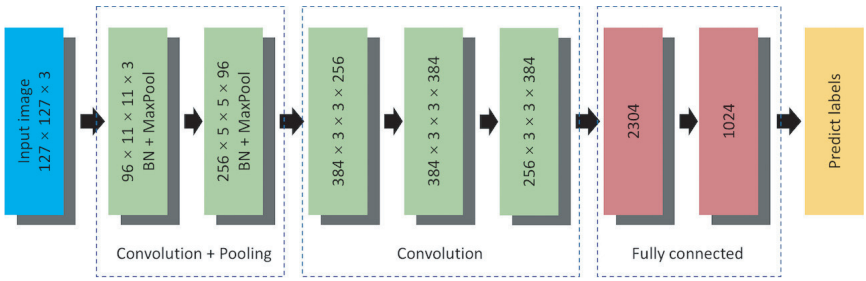

Fig. 3 CNN architecture

3.2 Word2Vec Method

The Word2Vec method was developed by Mikolov et al. [11] in 2013. The model is trained on a corpus of text (like Wikipedia or Google News) and then outputs the vector representations of all words in the text. Unlike the previous method, where a neural network learns the expression vectors of words, Word2Vec employs the Skip-gram model, which reduces the calculation of dense matrix multiplications. In our proposed method, we use the Word2Vec model to obtain the vector representations of labels in a dataset and then calculate the similarity between each pair of labels.

4. Proposed Approach

4.1 Training with Missing Labels

Let $\{(\boldsymbol{x}_i, \boldsymbol{t}_i)\}_{i=1}^{M} = \{\boldsymbol{X}, \boldsymbol{T}\}$ be the set of training samples with missing labels, where

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \cdots \boldsymbol{x}_i \cdots & \boldsymbol{x}_M \end{bmatrix}^T \tag{1}$$

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{t}_1 & \cdots \boldsymbol{t}_i \cdots & \boldsymbol{t}_M \end{bmatrix}^T \tag{2}$$

Here, $\boldsymbol{x}_i$ is the i^{th} image used as an input to the CNN and $\boldsymbol{t}_i = \begin{bmatrix} t_{i1} & \cdots & t_{iN} \end{bmatrix}$ represents the binary vector representation of the labels for the i^{th} image, where $t_{ij} = 1$ when the j^{th} label assigned to the i^{th} image; otherwise, $t_{ij} = 0$. M and N are the number of samples and labels, respectively.

Let $\boldsymbol{y} = f(\boldsymbol{x}; \boldsymbol{\theta})$ be the function of the CNN model for the input image $\boldsymbol{x}$ where $\boldsymbol{\theta}$ represents the parameters of the CNN. AlexNet [23], shown in Fig. 3, is used as the baseline CNN architecture. The network has five convolutional layers and two fully connected layers. The acronym "BN" in the figure denotes batch normalization.

To estimate the posterior probability of each label, a sigmoid activation function is after the neurons in the output layer and the sum of the sigmoid cross entropies of each label is used as the loss function. After training the parameters $\boldsymbol{\theta}$, we can estimate the labels $\hat{\boldsymbol{y}}$ by feeding the test input image $\hat{\boldsymbol{x}}$ into the trained CNN model

$$\hat{\boldsymbol{y}} = f(\hat{\boldsymbol{x}}; \boldsymbol{\theta}) \tag{3}$$

The value of each component in the estimated vector $\hat{\boldsymbol{y}}$ is the probability of the corresponding label.

4.2 Internal Distributions of Label Similarity

The internal similarity distributions of each label in the dataset is obtained by applying the HQ-III on the frequency Table $\boldsymbol{T}$. Let $\boldsymbol{q}_j, (j = 1, \ldots, N)$ be the HQ-III vector representation of each label. We calculate the co-occurrence distance between a pair of labels using the vectors $\boldsymbol{q}_i$ and $\boldsymbol{q}_j$ in their internal label space as

$$d_{ij}^{h} = \|\boldsymbol{q}_i - \boldsymbol{q}_j\|^2 \tag{4}$$

The similarity using internal co-occurrence distributions is defined as

$$s_{ij}^{h} = \exp(-\delta \times d_{ij}^{h}) \tag{5}$$

where δ controls the influence of the distance.

4.3 External Distributions of Label Similarity

The external similarity of each label in the dataset is derived using the Word2Vec method. Let $\boldsymbol{v}_j$ be the vector of the j-th ($j = 1, \ldots, N$) label. The word2vec distance between a pair of labels $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ is defined as

$$s_{ij}^{w} = \exp(-\epsilon \times d_{ij}^{w}) \tag{6}$$

where ϵ controls the influence of the Word2Vec distance.

4.4 Regularization Graph with Unified Co-Occurrence Distributions

To control the similarities s_{ij} between estimated labels, we introduce the graph Laplacian regularization term, defined as

$$G = \frac{1}{2}\sum_{i,j}^{N}(\hat{y_i} - \hat{y_j})^2 s_{ij} = \hat{\boldsymbol{y}}^T \boldsymbol{L}\hat{\boldsymbol{y}} \tag{7}$$

where $\hat{\boldsymbol{y}} = \begin{bmatrix}\hat{y_1} & \cdots & \hat{y_N}\end{bmatrix}$ and $(0 \le \hat{y_i} \le 1])$ is the binary vector representation of the estimated labels. $\boldsymbol{L}$ is the Laplacian matrix defined as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S}$, where $\boldsymbol{D} = \left[\sum_j^N s_{ij}\right]$ and $\boldsymbol{S} = \left[s_{ij}\right]$.

Since we have two similarities s_{ij}^{h} and s_{ij}^{w} representing both internal and external similarity, we can define two Laplacian regularization terms. The regularization term for the internal co-occurrence similarity graph is defined as

$$G_h = \frac{1}{2}\sum_{i,j}^{N}(\hat{y_i} - \hat{y_j})^2 s_{ij}^{h} = \hat{\boldsymbol{y}}^T \boldsymbol{L_h}\hat{\boldsymbol{y}} \tag{8}$$

Similarly, the regularization term for the external co-occurrence similarity graph is defined as

$$G_w = \frac{1}{2}\sum_{i,j}^{N}(\hat{y_i} - \hat{y_j})^2 s_{ij}^{w} = \hat{\boldsymbol{y}}^T \boldsymbol{L_w}\hat{\boldsymbol{y}} \tag{9}$$

where $\boldsymbol{L_h}$ and $\boldsymbol{L_w}$ are the co-occurrence dependency Laplacian matrices from the internal and external label spaces, respectively. For M training samples, we can define the average graph Laplacian regularization terms as

$$D_h = \frac{1}{M}\sum_{l=1}^{M}\left(\frac{1}{2}\sum_{i,j}^{N} s_{ij}^{h}(\hat{y_{li}} - \hat{y_{lj}})^2\right) = \sum_{l=1}^{M}\hat{\boldsymbol{y}_l}^T \boldsymbol{L}_h\hat{\boldsymbol{y}_l} \tag{10}$$

$$D_w = \frac{1}{M}\sum_{l=1}^{M}\left(\frac{1}{2}\sum_{i,j}^{N} s_{ij}^{w}(\hat{y_{li}} - \hat{y_{lj}})^2\right) = \sum_{l=1}^{M}\hat{\boldsymbol{y}_l}^T \boldsymbol{L}_w\hat{\boldsymbol{y}_l} \tag{11}$$

The values of D_h and D_w become small if the estimated labels are similar for a given pair of labels with similar vector representations implying high co-occurrence or similar meaning. Finally we combine the graph Laplacian regularization term from internal and external co-occurrence similarities as

$$D = \alpha D_h + (1-\alpha)D_w = \sum_{l}^{M}\hat{\boldsymbol{y}}_l^T(\alpha\boldsymbol{L}_h + (1-\alpha)\boldsymbol{L}_w)\hat{\boldsymbol{y}}_l \tag{12}$$

where α controls the contribution of each regularization term.

4.5 Training with a Combined Objective Function

The original objective function for standard multi-label annotation over all the training samples is given by

$$E = \sum_{l=1}^{M}\sum_{k=1}^{N}\{-t_{lk}\log(\hat{y_{lk}}) - (1-t_{lk})\log(1-\hat{y_{lk}})\}. \tag{13}$$

Here, the objective function is modified by combining the original objective function E with the graph Laplacian regularization term D, which gives

$$Q = E + \beta D. \tag{14}$$

The parameter β controls the effect of the regularization term.

5. Experiments

To validate the proposed approach, we use three standard image datasets and compare the results to a baseline CNN without regularization, a CNN using internal similarity alone with HQ-III, and a CNN using external similarity alone by the Word2Vec method. In this section, we describe the datasets, model parameter settings, and experimental results.

5.1 Datasets

We use three benchmark image annotation datasets: Corel5k, NUS-WIDE-LITE, and ESPGame. To generate the training dataset with missing labels, we randomly remove some labels in each training sample such that the number of labels in each sample is more than two. Table 1 shows the datasets used in our experiments. The columns in Table 1 named average and missing average represent the average number of assigned labels in each training sample and the average number of processed labels respectively.

5.2 Parameter Settings

Before feeding the images into the CNN, we reshape all the original images to a size of 127×127 pixels. The parameter α in Eq. (12) is set to 0.5 and β in Eq. (14) is set to 0.1. We obtained these values by performing a validated parameter sweep in the ranges [0, 1] and [0.001, 10] for α and β respectively.

Table 1 Datasets and their corresponding labels.

Dataset	Labels	#Training	#Test	Average	Missing Average
Corel5k	260	4500	499	3.4	2.5
NUS-WIDE-LITE	81	27807	27808	1.6	1.4
EspGame	268	18689	2081	3.7	3.0

Table 2 Comparison of the micro-F1 score on the Corel5k dataset.

Deep CNN methods	Top-3(%)	Top-5(%)	Probability ≥ 0.1(%)
Without regularization	17.25	17.75	17.12
Regularized with internal similarity alone	18.14	**18.23**	17.30
Regularized with external similarity alone	17.88	18.13	17.16
Proposed unified regularization	**18.52**	17.99	**17.76**

Table 3 Comparison of the micro-F1 score on the NUS-WIDE-LITE dataset.

Deep CNN methods	Top-3(%)	Top-5(%)	Probability ≥ 0.1(%)
Without regularization	5.95	6.31	4.22
Regularized with internal similarity alone	7.35	7.65	5.17
Regularized with external similarity alone	7.26	7.58	5.09
Proposed unified regularization	**7.43**	**7.73**	**5.37**

Table 4 Comparison of the micro-F1 score on the EspGame dataset.

Deep CNN methods	Top-3(%)	Top-5(%)	Probability ≥ 0.1(%)
Without regularization	9.42	10.13	8.49
Regularized with internal similarity alone	9.46	10.40	8.54
Regularized with external similarity alone	9.44	10.39	8.52
Proposed unified regularization	**9.68**	**10.74**	**8.87**

5.3 Performance Measures

The micro-F1 score is used to measure the performance of the proposed model for estimating multiple labels. This score measures the accuracy on the test datasets, considering both precision p and recall r, and is defined as

$$Micro - F_1 = 2 \times \frac{precision \times recall}{precision + recall}. \tag{15}$$

The performance on test datasets is calculated by

$$Micro - F_1 = \frac{1}{M} \sum_{i=1}^{M} \frac{2 \sum_{j=1}^{N} t_j^{(i)} y_j^{(i)}}{\sum_{j=1}^{N} t_j^{(i)} + \sum_{j=1}^{N} y_j^{(i)}}. \tag{16}$$

Here M and N are the number of test dataset samples and labels, respectively. The prediction is considered perfect when the score is 1 (perfect precision and recall) and worst when its value is 0. To select the most suitable labels, we use three kinds of label selection thresholds: labels whose predicted probability is above 0.1, the top-three ranked labels, and the top-five ranked labels. We compare the performance of the deep CNN model with unified regularization to those of a regularized deep CNN with internal label dependency alone, external label dependency alone, and a baseline CNN without any regularization term. Furthermore, to confirm the learning ability with incomplete labels, we remove some labels and observe the performance at different label removal rates (10%, 30%, and 50%) and compare the results in terms of the micro-F1-score against the baseline CNN.

5.4 Results and Discussion

We found that the model trained by the new combined objective function has better prediction ability than the one using the original objective function. To evaluate our results comparatively, we set up the experimental analyses according to previous works utilizing external [12] and internal [13] label similarity. Tables 2, 3 and 4 show the results evaluated by measuring the micro-F1 score at 30% label removal rates on different datasets based on the three kinds of label selection criteria. The proposed deep CNN model with unified regularization achieves higher F1-scores than the other methods, as highlighted in the tables. Furthermore, it can be observed that the deep CNN model with either internal similarity regularization or external similarity regularization produces better scores than the baseline CNN without a regularization term. On the Corel5k dataset, the micro-F1 score for labels with probability over 0.1% for the proposed method is higher than that of the baseline CNN, by more than 0.64%. Similarly, the score of the proposed method is higher than the deep CNN model with a regularization term of internal [13] and external [12] similarity by more than 0.30% and 0.6%, respectively. The approach presented in [31], using pairwise label correlations for multi-label classification achieved lower accuracy (12.1%) on the Corel5k dataset. Our unified approach modelling label relations on the same dataset achieves an acceptable improvement on label prediction. On the NUS-LITE, and EspGame datasets, the micro-F1 scores based on the labels with prob-

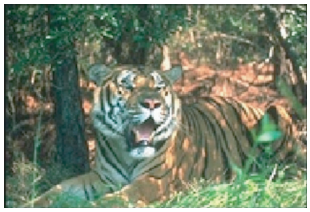

Fig. 4 Example image of Corel5k dataset and predicted labels with probability ≥ 0.1%. Given labels: cat, tiger, tree. Baseline CNN: cat, tiger, flowers, grass. Proposed: cat, **tree**, grass, tiger, **forest**

Fig. 5 Example image from NUS-WIDE-LITE dataset and predicted labels with probability ≥ 0.1%. Given labels: lake. Baseline CNN: clouds, mountain. Proposed: clouds, mountain, **valley**, **sky**

Fig. 6 Example image from EspGame dataset and predicted labels with probability ≥ 0.1%. Given labels: guitar, woman, music, hair, sing. Baseline CNN: guitar, man, music, light, hair, sing, singer. Proposed: guitar, music, hair, sing, **singer, man, light, band**

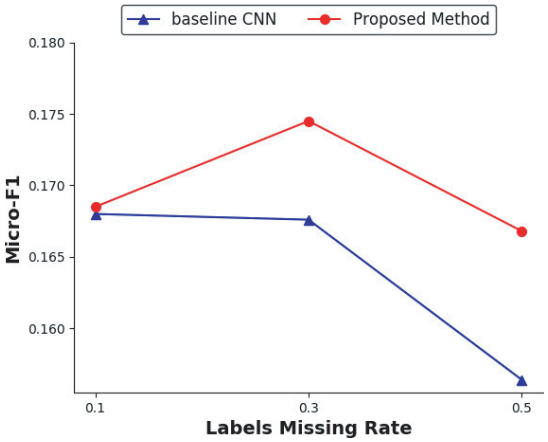

Fig. 7 Comparison of the micro-F1 score of the proposed method with that of the baseline CNN using different missing-label rates on the Corel5K dataset (probability ≥ 0.1%)

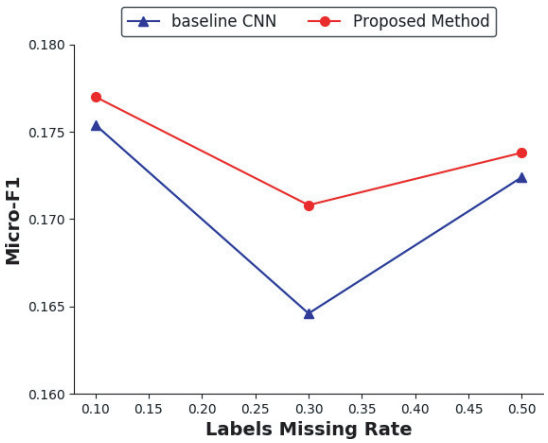

Fig. 8 Comparison of the micro-F1 score of the proposed method with that of the baseline CNN using different missing-label rates on the Corel5K dataset (Top-3)

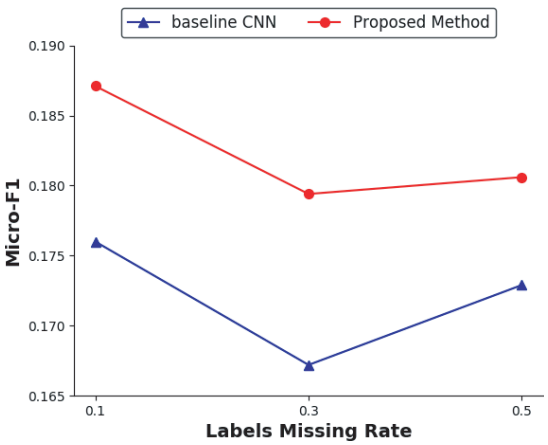

Fig. 9 Comparison of the micro-F1 score of the proposed method with that of the baseline CNN using different missing-label rates on the Corel5K dataset (Top-5)

ability over 0.1% are higher (in the range of 5.37%-8.87%) for the unified regularization and lower (in the range of 4.22%-8.49%) for the baseline CNN, demonstrating the efficiency of the proposed unified regularization technique in deep CNN for predicting missing labels.

Furthermore, to evaluate the performance of the unified regularization model against the baseline CNN without regularization, we visualize and compare the predicted labels across three different dataset images used in this study, as shown in Figs. 4, 5 and 6. The labels with predicted probability over 0.1 are considered for the comparison of the methods. The additional labels predicted by the proposed method are shown in bold. The results show that the proposed model can accurately capture labels that are related to the original given label set as well as to the visual content of the objects appearing in Fig. 4. The proposed model notably predicts "tree" and "forest", which are not predicted by the baseline CNN. It clearly demonstrates exemplars on which our proposed method improves the baseline predictions. Similarly, the proposed method detects the missing labels "valley" and "sky" for the given label "lake" in Fig. 5. However, the baseline CNN without regularization predicted a limited number of missing labels. The proposed unified co-occurrence approach can effectively infer the labels "sky" and "valley" from "clouds" and "mountain", re-

Table 5 Model sensitivity to parameters α and β on the Corel5k dataset (micro-F1 score for top-3 labels %).

	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
$\beta = 0.01$	17.48	**17.74**	17.59	17.39	17.12
$\beta = 0.1$	17.08	16.71	**18.52**	16.63	16.82
$\beta = 1$	17.64	17.74	17.09	18.02	**18.14**

spectively, because it can infer labels based on similarities in the internal space as well as from the external world of common sense. Though the micro-F1 score of the proposed approach is lower than that of the baseline CNN, the probability of predicting missing labels is more accurate, precise and relevant to the visual content of the objects appearing in Fig. 5.

Finally, we explain the results for Fig. 6, which includes a large number of object classes. In terms of the micro-F1 score, the proposed approach outperforms the baseline CNN in terms of accurate and very relevant missing labels related to the visual content of the objects. The experimental results indicate that the internal and external distributions of label similarity are more appropriate for detecting missing labels when the number of training labels is small and the number of object classes is large. In addition, the learning ability of the proposed model with different missing-label rates, measured in terms of the micro-F1 score, compared to that of the baseline CNN on the Corel5K dataset is shown in Fig. 7, Fig. 8 and Fig. 9. We observe that the proposed model outperforms the baseline CNN across different percentages of missing labels.

An analysis of the effect of parameters α and β on the performance of the model is shown in Table 5. The results show that as the value of β increases from 0.01 to 1 exponentially, the optimal value of alpha changes from 0.25 to 1.0. This means that the two parameters interact with each other, with bigger values of β requiring bigger values of α. We reason that at larger values of β, increased contribution of the external label similarity term leads to too much contradiction with the ground truth.

6. Conclusions

This study proposed a novel approach to address the multi-label image annotation problem with missing labels. A unified approach based on both internal and external label dependencies was presented, utilized in a graph Laplacian regularization term in order to reliably reconstruct missing label information from available labels and visual input. Due to the combined approach of label dependency representation, our proposed approach obtained distinctively more accurate results than other competing methods described in this study. Experimental analyses on three popular datasets revealed that our approach performs better than the baseline CNN without regularization. For future work, we plan to extend our unified framework approach with other neural network structures. Moreover, we will test our proposed approach on other benchmark data sets to further evaluate its effectiveness for the recovery of missing labels.

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Number 16K00239.

References

[1] P. Liu, X. Qiu, J. Chen, and X. Huang, "Deep Fusion LSTMs for Text Semantic Matching," Proceedings of 54th Annual Meeting of the Association for Computational Linguistics, pp.1034–1043, 2016.
[2] L. Xu, Z. Wang, Z. Shen, Y. Wang, and E. Chen, "Learning Low-Rank Label Correlations for Multi-label Classification with Missing Labels," Proceedings of IEEE International Conference on Data Mining, pp.1067–1072, 2014.
[3] Y. Zhu, J.T. Kwok, and Z.-H. Zhou, "Multi-Label Learning with Global and Local Label Correlation," IEEE Trans. Knowl. Data Eng., vol.30, no.6, pp.1081–1094, 2018.
[4] V. Ordonez, J. Deng, Y. Choi, A.C. Berg, and T.L. Berg, "From Large Scale Image Categorization to Entry-Level Categories," Proceedings of IEEE International Conference on Computer Vision, pp.2768–2775, 2013.
[5] J. Johnson, L. Ballan, and L. Fei-Fei, "Love Thy Neighbors: Image Annotation by Exploiting Image Metadata," Proceedings of IEEE International Conference on Computer Vision, pp.4624–4632, 2015.
[6] N. Ueda and K. Saito, "Parametric Mixture Models for Multi-Labeled Text," Proceedings of Advances in Neural Information Processing Systems, pp.721–728, 2003.
[7] C.-W. Lee, W. Fang, C.-K. Yeh, and Y.-C.F. Wang, "Multi-Label Zero-Shot Learning With Structured Knowledge Graphs," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1576–1585, 2018.
[8] R. Rehurek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," Proceedings of LREC 2010 Workshop on New Challenges for NLP Frameworks, pp.45–50, 2010.
[9] Wikipedia contributors, Plagiarism - Wikipedia, The Free Encyclopedia, 2018. https://en.wikipedia.org/
[10] C. Hayashi, "Multidimensional quantification, I," Proceedings of Proceedings of the Japan Academy, vol.30, no.2, pp.61–65, 1954.
[11] M. Tomas, S. Ilya, C. Kai, C. Greg, and D. Jeffrey, "Distributed Representations of Words and Phrases and Their Compositionality," Proceedings of 26th International Conference on Neural Information Processing Systems, pp.3111–3119, 2013.
[12] Y. Zhao, J. Miyao, and T. Kurita, "Multi-label Image Annotation via CNN with Graph Laplacian Regularization based on Word2Vec," Proceedings of International Workshop on Frontiers of Computer Vision, 2018.
[13] J. Mojoo, K. Kurosawa, and T. Kurita, "Deep CNN with Graph Laplacian Regularization for Multi-label Image Annotation," Proceedings of International Conference Image Analysis and Recognition, pp.19–26, 2017.
[14] H. Yu, P. Jain, P. Kar, and I. Dhillon, "Large-scale Multi-label Learning with Missing Labels," Proceedings of 31st International Conference on Machine Learning, pp.593–601, 2014.
[15] B. Wu, S. Lyu, and B. Ghanem, "Constrained submodular minimization for missing labels and class imbalance in multi-label learning," Proceedings of 30th AAAI Conference on Artificial Intelligence, pp.2229–2236, 2016.
[16] B. Wu, S. Lyu, and B. Ghanem, "ML-MG: Multi-label Learning with Missing Labels Using a Mixed Graph," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.4157–4165, 2015.
[17] X. Zhang, R. Henao, Z. Gan, Y. Li, and L. Carin, "Multi-Label Learning from Medical Plain Text with Convolutional Residual

Models," Proceedings of 3rd Machine Learning for Healthcare Conference, pp.280–294, 2018.
[18] Z. Xue, S. Antani, L.R. Long, and G.R. Thoma, "Automatic multi-label annotation of abdominal CT images using CBIR," Proceedings of SPIE 10138, Medical Imaging, 2017.
[19] W.-S. Chu, F.D. Torre, and J.F. Cohn, "Learning Spatial and Temporal Cues for Multi-Label Facial Action Unit Detection," Proceedings of 12th IEEE International Conference on Automatic Face and Gesture Recognition, pp.25–32, 2017.
[20] J. Shao, K. Kang, C.C. Loy, and X. Wang, "Deeply learned attributes for crowded scene understanding," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.4657–4666, 2015.
[21] K. Kang, W. Ouyang, H. Li, and X. Wang, "Object Detection From Video Tubelets With Convolutional Neural Networks," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.817–825, 2016.
[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick, "Microsoft COCO: Common Objects in Context," Proceedings of European Conference on Computer Vision, vol.8693, pp.740–755, 2014.
[23] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Proceedings of Advances in Neural Information Processing Systems 25, pp.1097–1105, 2012.
[24] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang, "Correlative multi-label video annotation," Proceedings of 15th international conference on Multimedia, pp.17–26, 2007.
[25] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative Adversarial Text to Image Synthesis," Proceedings of 33rd International Conference on Machine Learning, pp.1060–1069, 2016.
[26] Z.-J. Zha, T. Mei, J. Wang, Z. Wang, and X.-S. Hua, "Graph-based semi-supervised learning with multiple labels," J. Vis. Commun. Image. Represent., vol.20, no.2, pp.97–103, 2009.
[27] D. Kingma and M. Welling, "Auto-encoding variational bayes," Proceedings of International Conference on Learning Representations, 2014.
[28] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2Image: Conditional Image Generation from Visual Attributes," Proceedings of European Conference on Computer Vision (ECCV), vol.9908, pp.776–791, 2016.
[29] H. Hu, G.-T. Zhou, Z. Deng, Z. Liao, and G. Mori, "Learning Structured Inference Neural Networks with Label Relations," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2960–2968, 2016.
[30] J. Deng, S. Satheesh, A. Berg, and F. Li, "Fast and Balanced: Efficient Label Tree Learning for Large Scale Object Recognition," Proceedings of Advances in Neural Information Processing Systems 24, pp.567–575, 2011.
[31] J. Huang, G. Li, S. Wang, Z. Xue, and Q. Huang, "Multi-Label Classification by Exploiting Local Positive and Negative Pairwise Label Correlation," Neurocomputing, vol.257, pp.164–174, 2017.

Jonathan Mojoo he received a master's degree in Information Engineering from Hiroshima University, Japan, in 2018. His research interests include deep learning, object detection, image understanding, and image retrieval.

Yu Zhao he received a master's degree in Information engineering from Hiroshima University, Japan, in 2019. His research interests include image processing, machine learning and anomaly detection.

Muthu Subash Kavitha she received a Ph.D. in Information engineering from Hiroshima University, Japan, in 2012. Currently she is a JSPS research fellow with the Pattern recognition and machine learning lab, Department of Information Engineering, Hiroshima University, Japan. Her research interests include Image processing algorithms for Image pattern analysis, machine learningand deep learning.

Junichi Miyao he received a Ph.D. degree in Engineering from Hiroshima University in 1987, and is presently an associate professor at Hiroshima University. He has worked on image compression in multimedia processings, and pattern recognition. IEICE, IPA, IEEE member.

Takio Kurita he received the B.Eng. degree from Nagoya Institute of Technology and the Dr.Eng. degree from the University of Tsukuba, in 1981 and in 1993, respectively. From 1981 to 2000, he was a research scientist at the Electrotechnical Laboratory. From 2001 to 2009, he was a Deputy Director of Neuroscience Research Institute, National Institute of Advanced Industrial Science and Technology. He is currently a Professor at Hiroshima University. His current research interests are statistical pattern recognition and its application to image recognition. He is a member of the IEEE, IEICE, JNNS, JSAI, and ITE.