2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

True-negative Label Selection for Large-scale Multi-label Learning

Atsushi Kanehira, Andrew Shin, Tatsuya Harada
The University of Tokyo
7-3-1 Hongo Bunkyo-ku, Tokyo Japan
Email: {kanehira, andrew, harada}@mi.t.u-tokyo.ac.jp

***Abstract*—In this paper, we focus on training a classifier from large-scale data with incompletely assigned labels. In other words, we treat samples with following properties: 1. assigned labels are definitely positive, 2. absent labels are not necessarily negative, and 3. samples are allowed to take more than one label. These properties are frequently found in various kinds of computer vision tasks, including image and video classification and retrieval.**

Many online algorithms for multi-label task employ label sampling, which selects a label pair that reduces the largest penalty to update the model, thereby avoiding waste of computation. In the setting above, however, there are "false-negative" labels, which are originally positive labels but regarded as negative. Since it is high likely for label sampling to select these labels as negative labels in the sampled pair, it may severely degrade classification performance.

In order to solve this problem while preserving convergence property of the online algorithms, we propose a novel label sampling approach, which aims to fetch "true-negative" labels via false-negativeness measure based on independently trained uni-class classifiers. Experimental results show the effectiveness of our approach.

I. Introduction

Training a classifier with high precision from large-scale data is crucial in computer vision. However, many kinds of data in real-world applications, especially in image or video recognition, frequently come with incompletely assigned labels, constituting a setting in which:

1) assigned labels are definitely positive,
2) absent labels are not necessarily negative, and
3) samples are allowed to take more than one label.

For example, we may have to build a classifier for images uploaded on SNS (Social Networking Site), using attached tags as labels, as in Fig. 1. Since tags are manually provided, as with "indoor" or "people" in the example, we can say that 1. assigned labels are positive. On the other hand, images also contain objects that are not provided as tags, such as "pet bottles", "smart-phones", or "chairs" in the example figure. As such, since it is rare to have images whose complete set of objects are provided as tags, data also have a property that 2. absent labels are not necessarily negative.

Recently, many online algorithms have been developed in order to handle large-scale data. Especially, in computer vision domain, Passive Aggressive [1] is often utilized in practice because of its performance and implementability. Naively applying it to multi-labeled data takes $O(|Y| \times (k - |Y|))$,

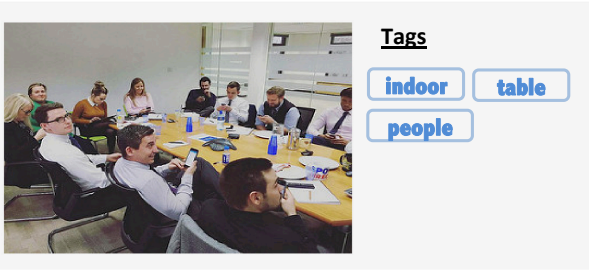

Fig. 1: Example of image and tags uploaded on SNS.

where $|Y|$ is the number of relevant labels for a sample and k is the total number of classes (We explain further in Sec. III), because the loss function which should be minimized includes all (positive label, negative label) pairs, and it is computationally expensive. To avoid waste of calculation, [1] relaxed loss function to have only one (positive label, negative label) pair which generates maximum penalty. In other words, the algorithm selects positive label having lowest score and negative label having highest score based on an old classifier before optimization for each step. Such label sampling approach based on maximum error reduction works effectively in practice and many methods follow it.

In our setting, however, positive labels may exist in absent labels (we call them "false-negative" labels throughout this paper), and such labels tend to be selected with high probability by maximum error reduction sampling strategy because they tend to have high score (intuitively, the images in which "dog" exists but without "dog" label tend to have high score for "dog" classifier.). Although one way to alleviate this problem may be sampling label pair randomly, the convergence of learning is slow.

We propose a novel label sampling strategy combined with widely used Passive Aggressive algorithm, in order to avoid sampling false-negative labels, while preserving fast convergence. This enables the model to be updated with more accurate "true-negative" labels. Our approach trains an additional uni-class classifier to estimate false-negativeness of samples to explicitly avoid sampling negative label which is originally positive. Since a uni-class classifier is constructed from labeled samples only, we can use it as reliable side-information. In spite of simplicity of our proposed method,

978-1-5090-4846-5/16/$31.00 ©2016 IEEE

experimental results show that it adds robustness to trained model against label incompleteness while preserving fast convergence property.

Our main contributions are as following:

1) propose a novel label sampling approach for multi-label online learning from incompletely labeled data, and
2) demonstrate the effectiveness of our approach from experiments on several datasets.

In Sec. I, we describe the goals and contributions of this work. We then discuss related works in Sec. II. In Sec. III, we explain our label sampling approach based on false-negativeness measurement. In Sec. IV, experiments conducted to investigate the efficacy of our proposed method on several datasets are described and discussed. Then we conclude our work in Sec. V.

II. Related work

Many online learning algorithms have recently been proposed, and have also been widely utilized in computer vision field. [1] proposed online Passive Aggressive (PA) algorithm, and further proposed a soft version of PA, in which constraints are further relaxed and loss function is added, in order to avoid updating the model by a large margin when incorrect labels are sampled. [2] proposed Confidence Weighted (CW), in which they took the confidence of training weights for each dimension into consideration. [3] introduced Adaptive Regularization of Weight (AROW), which modified CW so that it is applicable to noise during training. [4] proposed Gaussian Herding (NHERD) by extending PA to second-order algorithm. [5] proposed Soft Confidence Weighted (SCW), which relaxed the conditions of constraints with the same motivation as PA. In order to apply online PA to multi-class multi-label problems, [1] selected one (positive, negative) pair of sample to update the model, instead of updating with all (positive, negative) pairs to train PA, which enabled efficient learning. Other algorithms in similar manner have been proposed to deal with multi-class multi-label domain [6]. As we have seen, various online learning algorithms have been proposed, but PA is known to be comparable to state-of-the-art online learning algorithm, such as SCW, in image recognition task [6], and has been widely used due to its simple implementability. In many large-scale datasets, however, labels attached are often incomplete, and contain false-negative labels. As discussed in Section 1, current label sampling approaches are prone to sample those labels, which degrades the accuracy of the model.

Some works have attempted to address label incompleteness in multi-label learning as label deficits [7]. [8] tried to eliminate the influence of label deficits in the optimization process by adding a regularization term to rank loss, which forces the difference between scores for positive and negative labels to be group-sparse. Subsequently, [9] used conditional Restricted Boltzmann Machine to denoise the label deficit. [10] simultaneously computes the classifier and reconstructs lacking labels, taking label's sparsity and correlation into consideration. Since [8], [9], [10] are batch algorithms that necessitate repetitive computations with all samples for updating the model, it is difficult to apply them to online setting in which each sample is processed one at a time and cannot be re-used for update. It is an advantage of our proposed model over the methods described in [8], [9], [10].

III. Proposed approach

In this section, we describe our proposed label sampling approach. Let $\mathcal{X} \subset \mathbb{R}^d$ be sample space and $\mathcal{Y} = \{0,1\}^k$ be the possible set of labels where d and k denote dimension of samples and the number of classes respectively. A dataset $S = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_1, \mathbf{y}_1), ..., (\mathbf{x}_N, \mathbf{y}_N)\}$ having N samples is generated from an unknown distribution on $\mathcal{X} \times \mathcal{Y}$. A set of relevant labels for sample $\mathbf{x}_t$ is denoted as Y_t. The weight of model is written as $\mathbf{W} \in \mathbb{R}^{d \times k}$, and its i-th column corresponding to classifier for i-th class is $\mathbf{w}_i$.

A. Multi-label Passive Aggressive

Firstly, we explain application of Passive Aggressive to multi-label task. Multi-label PA is a natural extension of multi-class PA proposed in [1]. It aims to minimize rank loss ℓ_{rank}, which imposes the penalty on scoring higher value for an irrelevant class than relevant one through minimization of surrogate hinge loss $\hat{\ell}_{\mathrm{rank}}$, with minimum weight's update. Thus, for sample $\mathbf{x}$, the model is updated as

$$\mathbf{W}^{(t+1)} = \arg\min_{\mathbf{W}} \ ||\mathbf{W} - \mathbf{W}^{(t)}||^2 + C\hat{\ell}_{\mathrm{rank}}(\mathbf{W}; (\mathbf{x}_t, Y_t))^2,$$

where

$$\hat{\ell}_{\mathrm{rank}}(\mathbf{W}; (\mathbf{x}_t, Y_t)) = \sum_{r \in Y_t} \sum_{s \notin Y_t} \left\{1 - (\mathbf{w}_r^T \mathbf{x}_t - \mathbf{w}_s^T \mathbf{x}_t)\right\}_+ \quad (1)$$

C is a hyper-parameter that controls trade-off between two functions. $(\cdot)_+$ is the hinge function that returns $\max(0, \cdot)$. Since naive optimization for this loss includes $|Y| \times (k - |Y|)$ updates, it is computationally expensive. In order to solve this problem, [1] proposed to reduce it to include only one (positive, negative) pair which has the highest penalty of (1). The loss function for optimization can be re-written as

$$\mathbf{W}^{(t+1)} = \arg\min_{\mathbf{W}} \ ||\mathbf{W} - \mathbf{W}^{(t)}||^2 + C\hat{\ell}_{\mathrm{mc}}(\mathbf{W}; (\mathbf{x}_t, Y_t))^2,$$

where

$$\hat{\ell}_{\mathrm{mc}}(\mathbf{w}; (\mathbf{x}_t, Y_t)) = \left\{1 - (\mathbf{w}_{r_t}^T \mathbf{x}_t - \mathbf{w}_{s_t}^T \mathbf{x}_t)\right\}_+ \quad (2)$$

and

$$r_t = \arg\min_{r \in Y_t} \ \mathbf{w}_r^T \mathbf{x}_t,$$
$$s_t = \arg\max_{s \notin Y_t} \ \mathbf{w}_s^T \mathbf{x}_t. \quad (3)$$

This relaxation makes it possible to compute a closed-form solution as following:

$$\mathbf{w}_{r_t}^{(t+1)} = \mathbf{w}_{r_t}^{(t)} + \tau_t \mathbf{x}_t$$
$$\mathbf{w}_{s_t}^{(t+1)} = \mathbf{w}_{s_t}^{(t)} - \tau_t \mathbf{x}_t \quad (4)$$

Algorithm 1 PA with proposed label sampling

INPUT: parameters C_1, C_2, λ. number of iteration T.
INITIALIZATION: $\mathbf{W}_{\mathrm{mc}}, \mathbf{W}_{\mathrm{uni}} \in \mathbb{R}^{d\times k}, \epsilon$
for $t = 1, 2, ..., N \cdot T$ **do**
  fetch (x_t, Y_t) from dataset.
  choose $r_t \in Y_t$ and $s_t \notin Y_t$ based on (15).
  compute τ_t as (17).
  update $\mathbf{W}_{\mathrm{mc}}$ as (4).
  update $\mathbf{W}_{\mathrm{uni}}$ and ϵ as (10) with respect to r_t.

where

$$\tau_t = \frac{\ell_{\mathrm{mc}}}{||\mathbf{x}_t||^2 + 1/2C} \tag{5}$$

By such relaxation, the convergence becomes significantly faster in practice. As mentioned in Sec. I, in our setting, positive labels may exist in absent labels, and such labels tend to be selected with high probability by maximum error reduction sampling strategy. Updates on such pairs will lead learning to wrong direction, and the classification performance will decrease. Such incorrect classification results from minimizing incorrect loss function. In other words, loss function ℓ_{rank} estimated from the sum of all (positive, negative) pairs can be decomposed into loss functions for (positive, true negative) and (positive, false negative) pairs. While the former is supposed to be treated as loss, we also end up incorrectly minimizing the latter that is not supposed to be treated as loss. By setting the former as $\ell_{\text{rank-true}}$ and the latter as $\ell_{\text{rank-false}}$, correct loss function becomes

$$\ell_{\text{rank-true}} = \ell_{\mathrm{rank}} - \ell_{\text{rank-false}}, \tag{6}$$

which should be minimized. However, since we cannot know false negative labels in advance, we approximate the false-negativeness $\mathcal{F}$ for each negative class, as explained below. Setting false-negativeness measure for class i as $\mathcal{F}_i$, we utilize

$$\ell_{\text{rank-false}} \approx \sum_{r\in Y}\sum_{s\notin Y} \mathcal{F}_s. \tag{7}$$

In this paper, we propose a novel method in which false-negativeness is estimated by one-class classifier. In the next subsection, we explain uni-class Passive Aggressive.

B. Uni-class Passive Aggressive

Here we explain Uni-class Passive Aggressive proposed in [1]. It calculates representative point from only positive samples. The update rule is written as

$$\mathbf{w}^{(t+1)} = \arg\min_{\mathbf{w}} \ ||\mathbf{w} - \mathbf{w}^{(t)}||^2 + C\ell_{\mathrm{uni}}(\mathbf{w}, \epsilon; \mathbf{x}_t)^2 \tag{8}$$

$$\ell_{\mathrm{uni}}(\mathbf{w}, \epsilon; \mathbf{x}) = \begin{cases} 0 & (\text{if } ||\mathbf{w} - \mathbf{x}|| - \epsilon < 0) \\ ||\mathbf{w} - \mathbf{x}|| - \epsilon & (\text{otherwise}) \end{cases} \tag{9}$$

This formalization also has a closed-form solution as

$$\mathbf{w}^{(t+1)} = \left(1 - \frac{\tau_t}{||\mathbf{w}_t - \mathbf{x}_t||}\right)\mathbf{w}^{(t)} + \left(\frac{\tau_t}{||\mathbf{w}_t - \mathbf{x}_t||}\right)\mathbf{x}_t \tag{10}$$

where

$$\tau_t = \frac{\ell_{\mathrm{uni}}}{1 + 1/2C} \tag{11}$$

ϵ can be learned together by appending it to the last element of weight vector $\mathbf{w}$. Since it does not leverage negative class information, it has poor discriminative power. Even so, it has desired property for the data in our setting: we can apply it without influence of label incompleteness because it only uses "reliable" positive samples. Therefore, it will be helpful to utilize obtained classifiers as side information. Our approach trains uni-class classifier simultaneously with main classifier and uses obtained weight to measure false-negativeness of samples.

C. Proposed label sampling

Our proposed algorithm is summarized in Algorithm 1. It aims to fetch a pair of labels, whose update reduces much loss, while avoiding false-negative labels and thus sampling true negative labels. In order to achieve that, we propose to measure false-negativeness of sample $\mathbf{x}_t$ for class s as

$$\mathcal{F}_s = \epsilon_s - ||\mathbf{h}_s - \mathbf{x}_t|| \tag{12}$$

where $\mathbf{h}_s$ and ϵ_s are representative point of class s and its radius trained by uni-class PA respectively. Intuitively, the closer to the representative point the sample is, the more likely it is false-negative. Using (6), (7), and (12), and applying hinge loss as surrogate function, we obtain

$$\hat{\ell}_{\text{rank-true}} = \sum_{r\in Y_t}\sum_{s\notin Y_t} \left\{1 - (\mathbf{w}_r^T\mathbf{x}_t - \mathbf{w}_s^T\mathbf{x}_t) - \lambda\mathcal{F}_s\right\}_+ \tag{13}$$

Our proposed label sampling is naturally derived by relaxing (13) to include one label pair by maximum error reduction strategy as

$$\hat{\ell}_{\mathrm{mc}} = \left\{1 - (\mathbf{w}_{r_t}^T\mathbf{x}_t - \mathbf{w}_{s_t}^T\mathbf{x}_t) - \lambda\mathcal{F}_{s_t}\right\}_+ \tag{14}$$

where

$$r_t = \arg\min_{r\in Y_t} \ \mathbf{w}_r^T\mathbf{x}_t, \tag{15}$$

$$s_t = \arg\max_{s\notin Y_t} \ \mathbf{w}_s^T\mathbf{x}_t - \lambda\mathcal{F}_s. \tag{16}$$

λ is the hyper-parameter, which controls the influence of false-negativeness measure on the label sampling. When $\lambda = 0$, the proposed algorithm is completely equal to the normal Passive Aggressive algorithm. Weight's update is the same as (4) except

$$\tau_t = \frac{\hat{\ell}_{\mathrm{mc}}}{||\mathbf{x}_t||^2 + 1/2C} \tag{17}$$

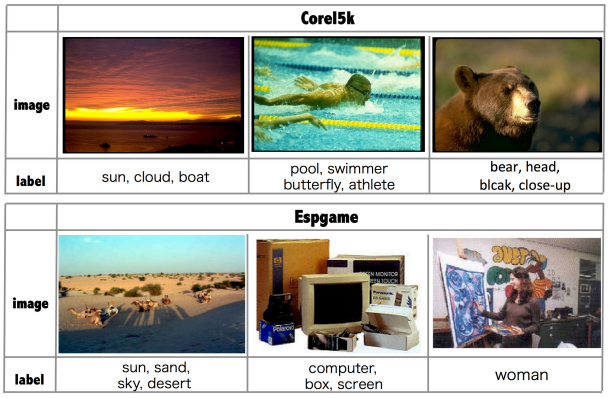

Fig. 2: Example images of Corel5k and Espgame.

TABLE I: Datasets used in ExperimentA and ExperimentB

Dataset	Train	Test	Class	Labels per sample
Corel5k	4500	499	260	3.40
Espgame	18689	2081	268	4.68

IV. Experiment

A. ExperimentA: Influence of hyper-parameter

In order to examine the influence of the hyper-parameters on the performance of our proposed method, we performed experiments on image annotation datasets, namely Corel5k [11] and Espgame [12]. We used 7-th layer's activation of CNN trained on ILSVRC2012 [13] dataset, and used AlexNet [14] as our network architecture. We deliberately set some of the labels in training data as unknown, and performed the experiment on fully-labeled test data. We varied the percentage of unknown labels in training data by every 10% within the range $0-80\%$, where each percentage was applied to the entire training data randomly. As for the evaluation metric, average precision over the samples was employed. Our proposed method involves three hyper-parameters, namely trade-off parameter of PA (denoted by C_1), uni-class classifier's parameter (denoted by C_2), and λ to indicate how much weight should be assigned to the metrics of false-negativeness. In order to examine the influence of newly added hyper-parameters C_2 and λ, we performed the following experiments:

(i) Fix λ at $\lambda = 1$, and vary C_2 from $C_2 = [0.001, 0.01, 0.1, 1, 10]$.
(ii) Fix C_2 at $C_2 = 0.01$, and vary λ from $\lambda = [0, 0.3, 0.5, 0.7, 1.0]$.

In both experiments, different values of C_1 were applied from $C_1 = [0.001, 0.01, 0.1, 1, 10]$, and the value yielding the best result was chosen.

Results from each experiment are shown in Fig. 3 and Fig. 4 respectively. In the experiment where C_2 was adjusted, it was found that C_2 = 0.1 or 0.01 yields the best result for Corel 5k, and C_2 = 0.01 or 0.001 for Espgame. We conjecture that, if C_2 is too large, the uni-class classifier updates the model by an excessively large margin when confronted with samples far apart, whereas, if C_2 is too small, update of the model becomes slow. In both cases, the model cannot find a good representation point of the class.

Results from the experiments where λ was adjusted show that our method outperforms the original PA regardless of the value of λ. In both datasets, small λ tends to increase the accuracy when the number of unknown labels is small, and large λ does a better job when the number of unknown labels is large. It can be interpreted that, as a larger number of unknown labels increases the number of false-negative labels, the necessity to exclude them from the sampling also increases, thus requiring a deficit-invariant training. The reason that our proposed method outperforms the original PA even when the percentage of unknown labels is 0% may be that there exist some false-negative labels in the original dataset. Such false-negative labels are frequently found in Espgame in particular, which may be attributed to the way it was constructed. Specifically, since two players provide the labels from the screen, from which only overlapping labels are accepted, it is highly likely that some labels that are originally positive may have been classified as negative, enabling our proposed method to work even without any deliberately generated noise.

B. ExperimentB: Convergence property

In order to test the convergence property of our proposed method, we compared the performances of the three following methods on the same datasets as in Experiment A:

- **Basic**: original Passive Aggressive. Sample class pairs according to (3) and update their weights according to (4). We fix $C_1 = 0.01$.
- **Random**: identical to **Basic**, except sampling class pair is performed randomly.
- **Proposed**: Passive Aggressive with proposed class sampling described in Sec. III-C. We fix $C_1 = 0.01$, $C_2 = 0.01$, and vary λ from $\lambda = [0.5, 1.0]$.

We attempted two percentages for setting unknown labels; 0% and 30% of the labels in the training data. Features and evaluation metrics were identical as in Experiment A.

Results are shown in Fig. 5. Throughout the graphs, it is shown that **Proposed** converges at roughly identical speed as the **Basic**, while being consistently faster than **Random** model. Also, when there are unknown labels, **Basic** method converges at a low accuracy, while the accuracy of the proposed method is comparable to that of **Random**. From these results, we can say that our proposed method succeeds in both preventing false-negative labels from being sampled, and training with high efficiency.

C. ExperimentC: Application to data in the wild

ImageCLEF2014: In ImageCLEF2014 [15] image annotation dataset, images come with information from the web page they were extracted from, instead of labels. Thus, corresponding mete-data such as filenames are provided so that appropriate labels can be presumed. As shown in Fig. 6, since meta-data are provided by the uploaders of the images, they tend

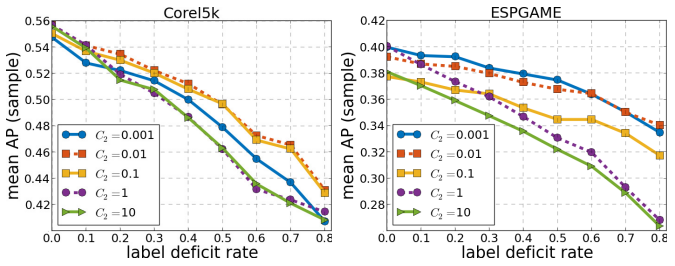

Fig. 3: Results of Experiment A (i) on Corel5k (left) and Espgame (right).

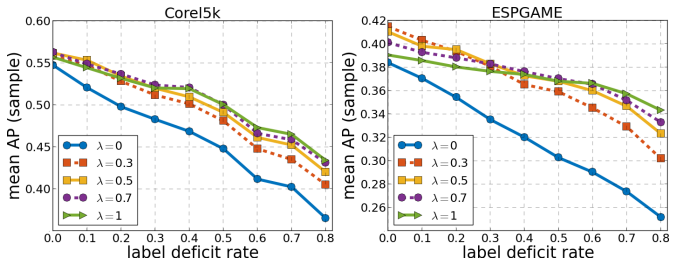

Fig. 4: Results of Experiment A (ii) on Corel5k (left) and Espgame (right).

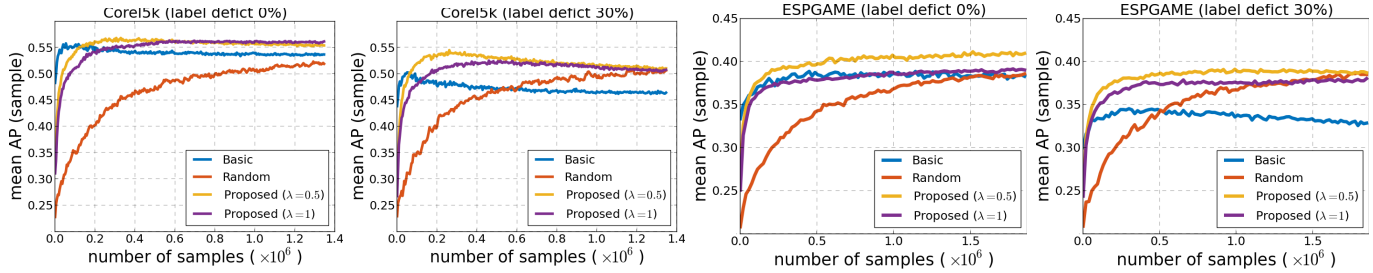

Fig. 5: Two leftmost figures show the results on Corel5k dataset with label deficit of 0% and 30% respectively. Two on the right-hand side show the results on Espgame with the same setting of label deficits as Corel5k.

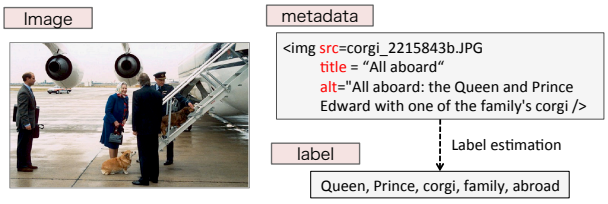

Fig. 6: Examples of estimated labels from metadata.

to contain information relevant to the images in most cases. However, they mostly contain only partial information of the image, and the obtained data frequently contain false-negative labels. In this experiment, labels are assigned by following steps:

1) extract words from image tag of the html file corresponding to the image.
2) let S be the set of these words and synonyms and hyponyms of them obtained from WordNet [16].
3) let T be the set of target classes.
4) assign labels $Y = T \cap S$ to the image.

Fig. 8 displays examples of the images and the labels obtained from web meta-data. From 500k samples of (image, meta-data), we extracted 121,131 samples for training that contain at least one label. For test, we used publicly available validation dataset consisting of 100k images of 108 classes.

Sentiment Dataset [17]: We examined whether our proposed method can be extended to a setting where the labels are of subjective nature, such as sentiments of the image. Labels of subjective nature inevitably form a setting with properties discussed in Sec 1., since obtained positive labels are reliable but the subjective nature permits many of the absent labels to be positive as well. For example, images with label "cute" may just as well be labeled as "funny" or "lovely," which may not have been provided as positive labels. Noting that viewers' comments toward images frequently correspond to the sentiment of the images, The authors [17] collected images from Flickr and DeviantArt, and collected associated comments, extracting adjectives describing the sentiments of the images. A series of Natural Language Processing techniques was employed, aided by SentiWordNet [18], to make sure that the adjectives correctly correspond to the sentiment of the image. The authors filtered out the adjectives whose positive and negative scores on SentiWordNet were both lower than the threshold, and also filtered out the adjectives that were negated (e.g., "not funny"), or were used to describe the speaker (e.g., "Im serious"), and manually removed the ones that are too general (e.g., "good", "great"). 20 most common adjectives were finally selected as the possible labels for the images. Table IV shows the obtained sentiment classes with their scores on SentiWordNet in three sentiment polarities (positive, negative, objective), along with the number of images in each class. Some examples are shown in Fig. 7. Images are treated in multi-label setting, containing up to 5 labels. In the experiment, we randomly sampled 100k images from our dataset. We refer to this dataset as Sentiment Dataset [1].

We compare our method to PA and SCW by repeating each experiment three times and comparing the average values. Results for each dataset are shown in Table II and Table III. In both datasets, our method outperforms others. Since these datasets contain a substantial amount of false-negative labels, our method proves to be effective.

[1] http://www.mi.t.u-tokyo.ac.jp/static/projects/sentidata/

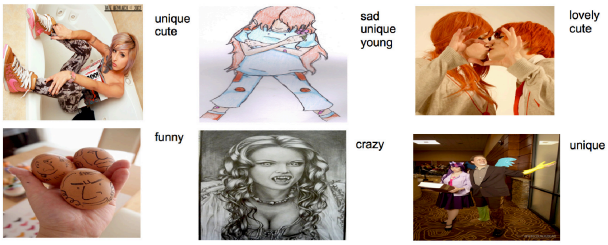

Fig. 7: Example images of Sentiment Dataset.

Fig. 8: Example images of ImageCLEF2014.

TABLE II: Experimental results on ImageCLEF2014

Method	mean F (sample)	mean F (class)
PA	0.344	0.306
SCW	0.324	0.274
proposed	**0.362**	**0.325**

TABLE III: Experimental results on Sentiment Dataset

Method	mean AP (sample)	meanAP (class)
PA	0.316	0.184
SCW	0.339	0.197
proposed	**0.388**	**0.234**

Class	POS	NEG	OBJ	Images
angry	0	.875	.125	25,824
beautiful	.750	0	.250	254,905
crazy	.625	(.500)	-	37,810
creepy	0	.875	.125	28,830
cute	.625	0	.375	325,606
dirty	0	.750	.250	16,417
funny	.500	(.500)	-	85,590
gorgeous	.750	0	.250	71,712
handsome	.625	0	.375	28,404
hot	.625	0	.375	48,486
lovely	.625	0	.375	123,004
sad	.125	.750	.125	75,263
scary	0	.750	.250	30,773
sexy	.625	0	.375	72,186
simple	.875	(.500)	-	46,874
stunning	.750	(.625)	-	24,049
ugly	0	.750	.250	21,840
unique	.500	0	.500	24,981
weird	0	.250	.750	51,072
young	.625	.250	.125	39,612

TABLE IV: Sentiment score and number of images for each class. Numbers in parentheses indicate that additional sense of word was retrieved to obtain its sentiment score.

V. Conclusion

In this paper, our goal is to train a classifier from large-scale samples with incompletely assigned labels, which are frequently found in various kinds of multimedia data, including images and videos.

In order to make learning robust and efficient on data under such setting, we proposed a novel label sampling approach for large-scale multi-label learning, which performs true-negative label mining via utilizing independently trained uni-class classifiers as false-negativeness measure. We also conducted experiments on several datasets and demonstrated the effectiveness of our approach.

Acknowledgment

This work was supported by CREST, JST.

References

[1] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *JMLR*, vol. 7, pp. 551–585, 2006.
[2] K. Crammer, M. Dredze, and F. Pereira, "Confidence-weighted linear classification for text categorization," *JMLR*, vol. 13, no. 1, pp. 1891–1926, 2012.
[3] K. Crammer, A. Kulesza, and M. Dredze, "Adaptive regularization of weight vectors," in *NIPS*, 2009.
[4] K. Crammer and D. D. Lee, "Learning via gaussian herding," in *NIPS*, 2010.
[5] S. C. Hoi, J. Wang, and P. Zhao, "Exact soft confidence-weighted learning." in *ICML*, 2012.
[6] Y. Ushiku, M. Hidaka, and T. Harada, "Three guidelines of online learning for large-scale visual recognition," in *CVPR, 2014*.
[7] A. Kanehira and T. Harada, "Multi-label ranking from positive and unlabeled data," in *CVPR*, 2016.
[8] S. S. Bucak, R. Jin, and A. K. Jain, "Multi-label learning with incomplete class assignments," in *CVPR, 2011*. IEEE, 2011.
[9] X. Li, F. Zhao, and Y. Guo, "Conditional restricted boltzmann machines for multi-label learning with incomplete labels," in *AISTATS, 2015*, 2015, pp. 635–643.
[10] F. Zhao and Y. Guo, "Semi-supervised multi-label learning with incomplete labels," in *IJCAI*, 2015.
[11] P. Duygulu, K. Barnard, J. F. de Freitas, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *ECCV*, 2002.
[12] L. Von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2004.
[13] J. Deng, A. Berg, S. Satheesh, H. Su, A. Khosla, and L. Fei-Fei, "Imagenet large scale visual recognition competition (ilsvrc2012)," 2012.
[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS, 2012*, 2012.
[15] M. Villegas and R. Paredes, "Overview of the imageclef 2014 scalable concept image annotation task," in *CLEF 2014 Evaluation Labs and Workshop, Online Working Notes*, 2014.
[16] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
[17] A. Shin, Y. Ushiku, and T. Harada, "Image captioning with sentiment terms via weakly-supervised sentiment dataset," in *BMVC*, 2016.
[18] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining." in *LREC*, vol. 10, 2010, pp. 2200–2204.