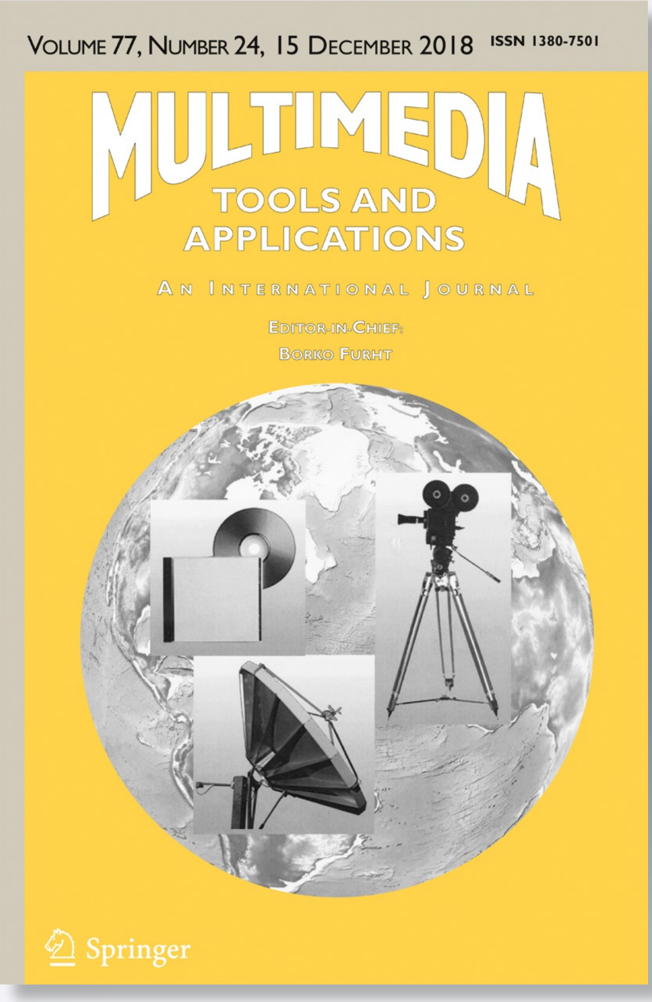
Automatic image annotation: the quirks and what works

Ayushi Dutta, Yashaswi Verma & C. V. Jawahar

Multimedia Tools and Applications
An International Journal

ISSN 1380-7501
Volume 77
Number 24

Multimed Tools Appl (2018)
77:31991-32011
DOI 10.1007/s11042-018-6247-3



Multimed Tools Appl (2018) 77:31991–32011
https://doi.org/10.1007/s11042-018-6247-3

Automatic image annotation: the quirks and what works

Ayushi Dutta[1] · Yashaswi Verma[2] · C. V. Jawahar[1]

Received: 24 August 2017 / Revised: 17 April 2018 / Accepted: 6 June 2018 /
Published online: 14 June 2018

Abstract Automatic image annotation is one of the fundamental problems in computer vision and machine learning. Given an image, here the goal is to predict a set of textual labels that describe the semantics of that image. During the last decade, a large number of image annotation techniques have been proposed that have been shown to achieve encouraging results on various annotation datasets. However, their scope has mostly remained restricted to quantitative results on the test data, thus ignoring various key aspects related to dataset properties and evaluation metrics that inherently affect the performance to a considerable extent. In this paper, first we evaluate ten state-of-the-art (both deep-learning based as well as non-deep-learning based) approaches for image annotation using the same baseline CNN features. Then we propose new quantitative measures to examine various issues/aspects in the image annotation domain, such as dataset specific biases, per-label versus per-image evaluation criteria, and the impact of changing the number and type of predicted labels. We believe the conclusions derived in this paper through thorough empirical analyzes would be helpful in making systematic advancements in this domain.

Keywords Image tagging · Empirical study · Evaluation metrics · Dataset analysis

✉ Ayushi Dutta
ayushi.dutta@research.iiit.ac.in

Yashaswi Verma
yashaswiv@iisc.ac.in

C. V. Jawahar
jawahar@iiit.ac.in

[1] CVIT, IIIT, Hyderabad, India

[2] CDS, IISc, Bangalore, India

1 Introduction

The last decade has witnessed an explosion of digital images on the Internet. As a result, it has become necessary to develop new technologies that can help in efficiently archiving and accessing large image collections. Since existing search engines are quite efficient in doing text-based search, it comes as a natural choice to address the above problem by associating images with text describing their semantics, such as discrete labels, short captions, or even full descriptions. Automatic image annotation is one of the fundamental problems that symbolizes the inter-play between visual and textual data. It aims at associating a set of discrete labels with a given image that describe its semantics, and has found applications in other visual understanding tasks such as image caption generation [13, 19, 39] and image retrieval [12, 28]. During the last decade, a large number of both non-deep-learning based [2, 3, 8, 9, 12, 18, 22, 23, 27–29, 36–38, 41, 42] as well as deep-learning based [10, 16, 17, 20, 24, 32, 40] image annotation techniques have been developed. However, almost all of these have primarily focussed on advancing quantitative results on the test sets of various image annotation datasets, without giving much attention to some of the core aspects related to dataset properties and evaluation metrics that internally affect the results to a significant extent.

In this paper, we attempt to investigate this situation. To summarize, our key contributions are:

1. First we evaluate ten benchmark image annotation techniques (five deep-learning based and five non-deep learning based) using the same baseline CNN features. These results can be useful in comparing future techniques addressing this task.
2. We thoroughly explore several aspects related to the performance evaluation metrics and five popular image annotation datasets (Corel-5K [7], ESP Game [1], IAPR TC-12 [11], NUS-WIDE [4] and MS-COCO [25]), and discuss what impacts these can have on quantitative performance.
3. We propose novel measures to quantify the degree of diversity (both image diversity as well as label diversity) in image annotation datasets. These are shown to relate with the performance of annotation methods, and can also be useful in developing new image annotation datasets.

As per our knowledge, this is the first study that investigates the above aspects related to the image annotation task. In the next section, we briefly describe the annotation techniques that we have considered in our analyzes. In Section 3, we detail the experimental settings and compare the results of various annotation techniques on popular annotation datasets. In Section 4, we extensively analyze various aspects related to datasets and performance evaluation metrics that can have significant impact while working in the real-world settings.

2 Label prediction models

We consider the following benchmark image annotation techniques in our analysis. As per our understanding, these techniques have made notable contributions in advancing this challenging area, and have also reported some of the best results on standard datasets.

1. **Joint Equal Contribution (JEC):** Based on the idea that similar images should share similar labels, Makadia et al. [27, 28] proposed a greedy nearest-neighbor based approach for image annotation. Given a test image, initially it transfers the labels

occurring in its nearest (or visually the most similar) training image in the order of their frequencies in the training set. Then it picks labels from additional neighbors, and considers their frequencies and co-occurrence with the initially assigned labels for further assignment.

2. **Tag Relevance (TagRel):** Li et al. [21] proposed an annotation approach based on the idea that for a given test image, the degree of relevance of a label is proportional to its frequency in the neighbor set of that image. However, in order to penalize very high frequency of a given label occurring in the neighborhood, it also takes into account the overall frequency of that label in the complete training set.
3. **TagProp:** Guillaumin et al. [12] proposed a weighted nearest neighbor model for image annotation. Given an image, the model takes weighted average of labels occurring in the neighbor set of that image and tries to maximize the likelihood of true labels during training. Additionally, to boost the performance of rare labels, per-label sigmoid functions are learned that act as wrapper functions over the baseline model.
4. **2PKNN:** Inspired by the success of JEC and TagProp, Verma and Jawahar [36, 38] proposed a two-step approach for image annotation. Given a test image, the first step constructs a balanced neighborhood that ensures a certain minimum number of occurrences of each label. Then in the second step, labels are propagated from the initially picked neighbors by computing a weighted average of their relevance based on visual similarity with the corresponding neighboring images. Analogous to TagProp, the motivation behind the first step is to improve the predictability of rare labels.
5. **SVM:** In [37], it was shown that simple binary (one-versus-rest) SVM classifiers [5] trained for each label could achieve superior performance than several state-of-the-art image annotation techniques. Precisely, to learn a classifier for a given label, all the (training) images annotated with that label are considered as positive samples and the rest as negative samples. Given a new image, all the classifiers are evaluated on it and their scores are calibrated using the Platt's normalization technique [31]. As discussed in [43], since this approach addresses the problem of multi-label image annotation on an individual label basis, it ignores the co-existence of other labels. Due to this, while it is conceptually simple and highly efficient, it fails to utilize label correlations.
6. **Deep Learning based techniques:** Recently, several deep neural network based techniques have been proposed for the multi-label image annotation task. Here, we focus on techniques that use different loss layers while training. Each loss function specifies a particular way of training the network, and how the network penalizes the differences in ground-truth and predicted labels. Precisely, we consider five loss functions: SoftMax [10], Sigmoid, Pairwise Ranking (or simply Ranking) [10], Weighted Approximate Ranking (or WARP) [10, 41], and the recently proposed Log-Sum-Exp Pairwise Ranking (or LSEP) [24].

We evaluate the above label prediction models using two state-of-the-art convolutional neural network architectures (GoogLeNet [34] and ResNet [15]), which were initially trained on the ImageNet 1000-class classification dataset [33]. For each deep-learning based method, we fine-tune the weights between the penultimate and soft-max layers using the corresponding loss function. For non-deep-learning based methods, we use the features from the penultimate layers of these networks. Following [35, 38], these features are transformed into an embedding space learned using Kernel Canonical Correlation Analysis (KCCA) [14]. This embedding is expected to provide a semantically richer representation than using raw features. Precisely, first we compute a similarity (or kernel) matrix between visual features using an exponential kernel and a kernel matrix between textual (binary labels) features using a linear kernel, and then adopt the implementation of [14] to learn the embedding space using the two kernel matrices. We will publish our code and data to facilitate reproducibility and future comparisons.

3 Experiments

3.1 Datasets

We use the following benchmark datasets in our analyzes, a subset of which has been used by almost all the existing techniques in their evaluations.

1. **Corel-5K** [7]: It contains 4,500 training and 499 testing images. Each image is annotated with up to 5 labels, with 3.4 labels per image on an average. This is one of the oldest image annotation datasets, and was considered as the de facto benchmark for evaluation until recently. Since most of the recent image annotation techniques are based on deep neural networks and require large amount training data, there has been a decline in the usage of this dataset.
2. **ESP Game** [1]: This dataset contains 18,689 training and 2,081 testing images, with each image being annotated with up to 15 labels and 4.7 labels on an average. It was formed using an on-line game where two mutually unknown players are required to assign labels to a given image, and score points for every common label. This way, several participants perform the manual annotation task, thus making this dataset quite challenging.
3. **IAPR TC-12** [11]: It contains 17,665 training and 1,962 testing images. Each image is annotated with up to 23 labels, with 5.7 labels per image on an average. In this dataset, each image is associated with a long description in multiple languages. Makadia et al. [27, 28] extracted nouns from the descriptions in the English language and treated them as annotations. Since then, it has been used extensively for evaluating image annotation methods.
4. **NUS-WIDE** [4]: This is the largest publicly available image annotation dataset, containing 269,648 images downloaded from Flickr. The vocabulary contains 81 labels, with each image being annotated with up to 3 labels. On an average, there are 2.40 labels per image. Following the earlier papers [10, 35], we discard the images without any label. This leaves us with 209,347 images, that we split into $\sim$ 125K images for training and $\sim$ 80K for testing by adopting the split originally provided by the authors of this dataset.
5. **MS-COCO** [25]: This is the second largest popular image annotation dataset, and is primarily used for object recognition in the context of scene understanding. It contains 82,783 training images and 80 labels, with each image being annotated with 2.9 labels on an average. For this dataset, the ground-truth of the test set is not publicly available. Hence, we consider the validation set containing 40,504 images as the test set in our experiments.

3.2 Evaluation metrics

To analyze annotation performance, we consider both per-label as well as per-image evaluation metrics. While per-label evaluation metrics have been popularly used to evaluate image annotation models for over a decade [7, 8, 12, 28, 36–38], some recent papers also use per-image metrics [10, 17, 26, 35]. Below we describe the metrics in these two categories:

3.2.1 Per-label evaluation metrics

Here, we consider per-label precision, recall and mean average precision (mAP). Given a label, let it be present in the ground-truth of m_1 images, and during testing let it be predicted for m_2 images out of which m_3 predictions are correct ($m_3 \leq m_1$ and $m_3 \leq m_2$). Then the precision for this label will be m_3/m_2, and recall will be m_3/m_1. These values are averaged over all the labels in the vocabulary to get average (percentage) per-label precision ($\mathrm{P_L}$) and average per-label recall ($\mathrm{R_L}$) respectively. From these, we compute average per-label F1 score ($\mathrm{F1_L}$), which is the harmonic mean of $\mathrm{P_L}$ and $\mathrm{R_L}$; i.e., $\mathrm{F1_L} = 2 \times \mathrm{P_L} \times \mathrm{R_L}/(\mathrm{P_L} + \mathrm{R_L})$. We also consider the N+ metric, that counts the number of labels with positive recall (or, how many labels in the vocabulary are correctly predicted for at least one test image). Additionally, we compute label-centric mAP ($\mathrm{mAP_L}$) that measures the quality of image-ranking corresponding to each label.

3.2.2 Per-image evaluation metrics

Here, we consider per-image precision, recall and mAP. Given an image, let there be n_1 labels present in its ground-truth, and during testing a model predicts n_2 labels out of which n_3 predictions are correct ($n_3 \leq n_1$ and $n_3 \leq n_2$). Then the precision for this image will be n_3/n_2, and recall will be n_3/n_1. These values are averaged over all the images in the test set to get average (percentage) per-image precision $\mathrm{P_I}$ and average per-image recall $\mathrm{R_I}$. From these two scores, we compute per-image F1 score ($\mathrm{F1_I}$), which is the harmonic mean of $\mathrm{P_I}$ and $\mathrm{R_I}$. We also compute image-centric mAP ($\mathrm{mAP_I}$) that measures the quality of label-ranking corresponding to each image.

3.2.3 Label assignment

Unless stated otherwise, we follow the earlier papers [8, 10, 12, 35, 36] and assign the top 5 labels to each test image in the Corel-5K, ESP Game and IAPR TC-12 datasets, and the top 3 labels in the NUS-WIDE and MS-COCO datasets for evaluating all the metrics, except for $\mathrm{mAP_L}$ and $\mathrm{mAP_I}$ that are evaluated on the complete ranked list of all the (test) images and labels respectively.

3.3 Model comparison

In Table 1 (for NUS-WIDE), Table 2 (for MS-COCO) and Tables 3 and 4 (for Corel-5K, ESP Game and IAPR TC-12), we compare the annotation performance of different label prediction models. As discussed earlier, since per-image evaluation metrics have become popular only recently, and most of the recent papers have reported results only on the NUS-WIDE dataset, we show results for per-image metrics only for the methods considered in this paper in Tables 2, 3 and 4.

In general, we can observe significant variations in the results on different datasets. This is becuase of the differences in how these datasets were created and their vocabularies. For datasets with large vocabularies (Corel-5K, ESP Game and IAPR TC-12), the results are usually lower than those with small vocabularies (NUS-WIDE and MS-COCO). Another

Table 1 Performance comparison of various annotation models (deep-learning based models are marked by '*') on the NUS-WIDE dataset

Method	Per-label metrics					Per-image metrics			
	$\mathrm{P_L}$	$\mathrm{R_L}$	$\mathrm{F1_L}$	$\mathrm{mAP_L}$	N+	$\mathrm{P_I}$	$\mathrm{R_I}$	$\mathrm{F1_I}$	$\mathrm{mAP_I}$
Johnson [17]	54.74	57.30	55.99	61.88	–	53.46	75.10	62.46	80.27
Hu [16] (1)	57.02	59.82	58.39	67.20	–	56.84	78.78	66.04	89.99
Hu [16] (2)	58.30	60.63	59.44	69.24	–	57.05	79.12	66.30	82.53
Liu [26]	71.73	61.73	66.36	–	–	77.41	76.88	77.15	–
Gong [10]	31.65	35.60	33.51	–	80	48.59	60.49	53.89	–
Ren [32]	37.74	40.15	38.91	–	81	52.23	65.03	57.93	–
Wang [40]	40.50	30.40	34.73	–	–	49.90	61.70	55.18	–
Liu [26]	55.65	50.17	52.77	–	–	70.57	71.35	70.96	–
Using *GoogLeNet*									
SoftMax*	45.16	51.72	48.22	46.45	81	52.98	74.92	62.07	79.95
Sigmoid*	45.91	52.18	48.85	53.97	81	53.84	75.69	62.92	81.19
Ranking*	44.49	51.70	47.82	45.41	81	52.84	74.27	61.75	79.34
WARP*	43.91	53.17	48.09	46.04	81	53.03	74.56	61.98	79.54
LSEP*	44.29	53.46	48.45	49.32	81	53.64	75.54	62.73	80.91
JEC	37.15	40.91	38.94	21.63	80	29.36	69.32	41.25	61.68
TagRel	39.75	59.27	47.58	49.15	81	49.87	70.71	58.49	73.55
TagProp	48.84	58.10	53.07	53.81	80	51.52	73.16	60.46	76.90
2PKNN	52.49	52.28	52.38	51.96	81	45.30	64.77	53.31	67.82
SVM	46.56	52.38	49.30	51.84	81	53.31	74.96	62.31	79.87
Using *ResNet*									
SoftMax*	44.36	51.88	47.82	47.94	80	53.61	75.68	62.76	80.73
Sigmoid*	46.97	53.11	49.85	55.72	81	54.64	76.74	63.83	82.20
Ranking*	45.73	52.82	49.03	49.30	81	54.09	75.96	63.19	81.28
WARP*	44.74	52.44	48.28	48.96	81	53.81	75.48	62.83	80.65
LSEP*	43.30	53.98	48.05	51.09	81	54.34	76.41	63.52	81.86
JEC	39.71	40.60	40.15	22.05	80	49.27	69.84	57.78	62.83
TagRel	40.27	60.26	48.28	50.87	81	50.86	71.95	59.60	74.85
TagProp	49.45	59.13	53.86	54.55	81	52.37	74.21	61.41	78.03
2PKNN	47.94	55.76	51.55	52.97	81	51.90	73.00	60.67	77.46
SVM	46.35	54.02	49.89	53.55	81	54.23	76.00	63.30	80.97

reason is the diversity in these datasets (in terms of both images as well as labels), that we will analyze and discuss in Section 4.2.

In Table 1, the first two blocks show the performances reported by some recent techniques, all of which are based on some end-to-end trainable deep neural network. Note that while the techniques in the second block make use of only the available training data (images and their labels), those in the first block also make use of additional meta-data such as either social tags [17, 26] or the WordNet hierarchy [16]. Due to this, while these are able to achieve significantly higher results than others, these can not be compared directly. From

Table 2 Performance comparison of various annotation models (deep-learning based models are marked by '*') on the MS-COCO dataset

Method	Per-label metrics					Per-image metrics			
	P_L	R_L	$F1_L$	mAP_L	N+	P_I	R_I	$F1_I$	mAP_I
Using *GoogLeNet*									
SoftMax*	58.87	57.41	58.13	59.47	80	58.32	71.76	64.35	80.34
Sigmoid*	61.14	58.87	59.98	67.74	80	60.16	73.22	66.05	82.37
Ranking*	58.84	57.18	57.99	61.48	80	58.65	71.48	64.43	80.22
WARP*	59.21	56.44	57.79	61.07	80	58.27	71.00	64.01	79.80
LSEP*	60.16	59.24	59.70	63.92	80	59.80	73.05	65.77	82.04
JEC	53.03	42.68	47.30	29.63	80	49.52	61.52	54.87	55.82
TagRel	52.36	57.61	54.86	62.87	80	54.20	67.07	59.95	73.35
TagProp	60.35	56.82	58.53	63.13	80	57.23	70.13	63.02	77.97
2PKNN	71.00	49.25	58.16	61.03	80	50.20	61.50	55.28	69.74
SVM	61.71	59.07	60.36	68.33	80	60.11	73.03	65.94	81.75
Using *ResNet*									
SoftMax*	56.77	56.20	56.49	57.24	80	57.88	71.05	63.79	79.69
Sigmoid*	58.82	57.92	58.37	66.74	80	59.78	72.62	65.58	81.72
Ranking*	56.75	55.75	56.25	58.75	80	58.00	70.54	63.66	79.42
WARP*	57.09	55.31	56.19	58.11	80	57.54	70.03	63.18	78.93
LSEP*	57.35	58.66	57.99	61.41	80	59.52	72.61	65.41	81.51
JEC	56.06	45.53	50.25	32.53	80	51.64	64.29	57.27	59.77
TagRel	55.67	59.67	57.60	63.30	80	55.66	68.67	61.49	74.40
TagProp	63.11	58.29	60.61	63.46	80	58.17	71.07	63.98	78.52
2PKNN	63.77	55.70	59.46	62.72	80	54.13	66.95	59.86	75.48
SVM	60.27	57.67	58.94	65.52	80	59.33	72.08	65.09	80.38

Table 1, we observe that both non-deep-learning methods as well as deep-learning based methods (that do not use additional meta-data) give comparable results.

In Table 2, we compare the performance of various methods on the MS-COCO dataset. Here, we observe that generally TagProp and SVM achieve the best results among the non-deep-learning based methods, and Sigmoid and LSEP achieve the best results among the deep-learning based methods. Similar to the NUS-WIDE dataset, the best results using both deep as well as non-deep methods are comparable.

In Tables 3 and 4, we compare the performance of the five non-deep-learning based methods on small-scale datasets using GoogLeNet and ResNet features respectively. Here, we do not consider deep-learning based methods since they require large amount of data for proper training. From the results, we observe that 2PKNN generally achieves the best performance.

In general, for all the datasets and methods, we can observe that the scores corresponding to per-image metrics are higher than per-label metrics. We will study this trend in Section 4.1, where we will empirically show that per-image metrics show some bias toward good performance on frequent labels.

Table 3 Performance comparison of various annotation methods on Corel-5K, ESP Game and IAPR TC-12 datasets using *GoogLeNet* features

Method	Per-label metrics					Per-image metrics			
	P_L	R_L	$F1_L$	mAP_L	N+	P_I	R_I	$F1_I$	mAP_I
Corel-5K									
JEC	41.70	44.95	43.27	37.29	161	45.97	64.92	53.76	50.87
TagRel	40.64	46.43	43.34	41.81	167	45.29	63.76	52.96	58.56
TagProp	37.88	42.79	40.19	43.15	155	46.05	65.27	54.00	60.13
2PKNN	46.10	52.85	49.25	53.18	197	44.48	62.60	52.01	57.89
SVM	36.64	46.29	40.90	53.15	158	48.42	68.77	56.83	67.30
ESP Game									
JEC	45.15	31.39	37.03	21.66	239	41.85	47.06	44.31	35.83
TagRel	38.89	42.05	40.41	38.81	252	43.57	48.52	45.92	49.91
TagProp	44.48	41.23	42.79	38.95	250	44.17	49.77	46.80	51.44
2PKNN	45.48	42.20	43.78	41.40	260	43.89	49.43	46.50	51.57
SVM	44.21	36.07	39.73	41.33	245	47.13	52.97	49.88	55.70
IAPR TC-12									
JEC	44.52	27.77	34.20	20.08	226	49.62	47.92	48.76	38.08
TagRel	45.07	42.21	43.59	39.71	267	50.66	49.01	49.82	52.40
TagProp	49.13	41.73	45.13	44.18	270	50.39	49.18	49.78	55.40
2PKNN	50.77	41.64	45.75	46.39	275	50.41	48.72	49.55	56.39
SVM	51.13	30.81	38.45	46.67	235	54.41	52.63	53.50	60.60

4 Analysis

Now we analyze various aspects of image annotation datasets and performance evaluation metrics by considering the ten annotation methods discussed in Section 2 as the working examples wherever required.

4.1 Per-label versus Per-image evaluation

In per-image metrics each test image contributes equally, and thus they tend to get biased toward performance on frequent labels. In contrary, each label contributes equally in per-label metrics, due to which they tend to get affected by performance on rare labels. It is important to note that the issue of imbalance in label frequencies (also called class-imbalance) in image annotation datasets has attained some attention in the past [12, 36, 38].

Recall that as discussed in Section 3.2.3, we assign a fixed number of labels to each test image during evaluation. However, since several test images may have either more or less labels in the ground-truth, no method can achieve perfect performance. In order to study the relative trade-off between per-label and per-image metrics, we try to evaluate the best performance achievable by each method. For this, we assume to know what labels are incorrect predictions for each test image using a given annotation method, and then replace these labels by either the most frequently occurring, or the least frequently occurring, or

Table 4 Performance comparison of various annotation methods on Corel-5K, ESP Game and IAPR TC-12 datasets using *ResNet* features

Method	Per-label metrics					Per-image metrics			
	P_L	R_L	$F1_L$	mAP_L	N+	P_I	R_I	$F1_I$	mAP_I
Corel-5K									
JEC	37.87	43.04	40.29	34.94	151	46.41	65.80	54.43	51.64
TagRel	37.01	45.23	40.71	40.13	157	46.25	65.43	54.20	59.16
TagProp	36.54	42.62	39.34	49.23	145	47.81	67.92	56.12	62.10
2PKNN	47.21	54.95	50.78	54.6	201	45.41	64.30	53.23	59.08
SVM	36.79	43.84	40.01	54.32	155	48.94	69.64	57.48	68.13
ESP Game									
JEC	47.14	30.43	36.99	21.92	235	42.69	48.22	45.29	37.20
TagRel	43.12	41.84	42.47	39.97	255	44.55	49.87	47.06	51.37
TagProp	48.48	41.78	44.88	41.04	253	45.70	51.60	48.47	53.48
2PKNN	46.17	44.22	45.17	42.89	262	46.04	52.25	48.95	54.22
SVM	47.11	35.02	40.17	43.46	236	48.99	55.55	52.06	58.61
IAPR TC-12									
JEC	48.69	27.16	34.87	20.52	226	51.34	49.49	50.40	39.86
TagRel	48.29	41.35	44.55	41.70	265	53.03	51.31	52.15	55.50
TagProp	52.27	40.97	45.93	45.82	266	52.88	51.28	52.07	57.72
2PKNN	53.54	42.58	47.43	48.88	281	53.21	51.43	52.30	59.32
SVM	51.97	28.16	36.53	48.71	224	54.41	52.37	53.37	61.05

randomly chosen incorrect labels to satisfy the requirement of 3/5 label assignment. Note that in this analysis, we can not evaluate mAP_L and mAP_I.

In Tables 5 and 6, we show the performance of various methods in terms of F1 scores using GoogLeNet and ResNet respectively. Here, “True” denotes the actual performance obtained by each method, and “Rare”, “Freq.”, and “Rand” denote the scores obtained by filling the empty slots with rare, frequent and random incorrect labels respectively. In case of “Ground” (ground-truth), the “True” performance is 100% since here we relax the constraint of assigning exactly 3/5 labels and evaluate over the ground-truth labels themselves. In one sense, these results can be thought of as upper-bounds achievable using various methods. In case of $F1_L$, we can observe that the performance of each method improves significantly when we replace the incorrect predictions by either the most rare or the most frequent (incorrect) labels. This is expected because very few labels tend to get highly penalized. Due to this, while their performances drop, that of each of the remaining labels improves, thus significantly improving the overall performance. However, when we randomly assign incorrect labels, the penalty spreads across all the labels and this leads to significant drop in the performance. In contrast, in case of $F1_I$, we can observe that the performance of each method improves significantly when we replace the incorrect predictions by the most frequent labels. However, the performance of each method generally remains close to its actual (“True”) performance when we replace incorrect predictions by either the most rare or randomly chosen (incorrect) labels.

These results show that per-image metrics are biased toward rewarding correct predictions of frequently occurring labels. Moreover, as long as the *number* of incorrect

Table 5 Comparing the actual per-label and per-image performance (using *GoogLeNet*) of various label prediction models (deep-learning based models are marked by '*') with those obtained by replacing incorrect predictions with rare/frequent/random incorrect labels

	Method	Per-label F1 score ($F1_L$)				Per-image F1 score ($F1_I$)			
		True	Rare	Freq.	Rand	True	Rare	Freq.	Rand
Corel-5K	Ground	100.00	99.25	99.55	71.59	100.00	82.61	82.61	82.61
	JEC	43.27	52.07	51.74	33.87	53.76	53.80	56.42	54.33
	TagRel	43.34	54.53	53.61	34.53	52.96	53.09	57.01	53.42
	TagProp	40.19	50.59	49.47	32.46	54.00	54.14	56.52	54.14
	2PKNN	49.25	62.27	61.98	38.15	52.01	52.11	55.61	52.51
	SVM	40.90	52.90	52.15	34.72	56.83	56.92	59.08	57.11
ESP Game	Ground	100.00	99.26	99.56	81.50	100.00	88.60	88.60	88.60
	JEC	37.03	47.38	46.81	29.02	44.31	44.36	49.37	44.92
	TagRel	40.41	58.92	58.44	36.10	45.92	45.98	52.40	46.62
	TagProp	42.79	57.99	57.43	35.82	46.80	46.87	52.22	47.25
	2PKNN	43.78	59.22	59.14	36.10	46.50	46.54	52.15	47.02
	SVM	39.73	52.76	51.92	32.99	49.88	49.95	53.69	50.47
IAPR TC-12	Ground	100.00	99.26	99.71	87.58	100.00	92.83	92.83	92.83
	JEC	34.20	41.30	41.11	28.21	48.76	48.88	52.72	49.28
	TagRel	43.59	57.83	58.06	38.33	49.82	49.89	55.97	50.27
	TagProp	45.13	57.70	57.81	38.37	49.78	49.87	55.65	50.32
	2PKNN	45.75	57.90	57.95	38.07	49.55	49.63	54.10	50.04
	SVM	38.45	44.95	44.64	31.10	53.50	53.62	55.91	54.05
NUS-WIDE	Ground	100.00	98.56	99.11	62.33	100.00	79.88	79.90	79.88
	SoftMax*	48.22	68.33	67.97	36.41	62.07	62.08	65.12	62.48
	Sigmoid*	48.85	68.57	68.30	36.49	62.92	62.94	65.36	63.33
	Ranking*	47.82	68.34	67.94	36.25	61.75	61.76	64.82	62.20
	WARP*	48.09	69.44	69.19	36.99	61.98	61.99	65.12	62.43
	LSEP*	48.45	69.58	69.40	37.17	62.73	62.75	65.51	63.16
	JEC	38.94	57.17	56.62	28.81	41.25	56.74	61.01	57.37
	TagRel	47.58	74.31	74.42	39.42	58.49	58.50	65.80	59.07
	TagProp	53.07	73.61	72.96	39.51	60.46	60.48	65.13	60.95
	2PKNN	52.38	68.03	68.64	33.88	53.31	53.32	58.65	54.06
	SVM	49.30	68.95	68.48	36.60	62.31	62.32	64.89	62.72
MS-COCO	Ground	100.00	98.36	99.19	82.9	100.00	85.46	85.46	85.46
	SoftMax*	58.13	72.90	72.73	52.06	64.35	64.41	66.75	64.84
	Sigmoid*	59.98	73.98	73.82	53.41	66.05	66.12	67.92	66.47
	Ranking*	57.99	72.74	72.52	51.96	64.43	64.51	66.49	64.94
	WARP*	57.79	72.09	71.96	51.28	64.01	64.08	66.26	64.52
	LSEP*	59.70	74.23	74.13	53.61	65.77	65.83	67.88	66.22
	JEC	51.05	63.34	62.93	42.21	57.22	57.30	59.80	57.99
	TagRel	58.20	74.77	74.94	52.85	61.55	61.61	67.70	62.09
	TagProp	61.44	73.71	73.68	52.90	64.21	64.27	66.54	64.74
	2PKNN	62.05	72.33	72.22	50.77	61.47	61.54	66.85	62.07
	SVM	60.37	73.11	73.02	52.59	65.28	65.34	66.94	65.72

(Refer Section 4.1 for details)

Table 6 Comparing the actual per-label and per-image performance (using *ResNet*) of various label prediction models (deep-learning based models are marked by ‘*’) with those obtained by replacing incorrect predictions with rare/frequent/random incorrect labels

	Method	Per-label F1 score ($\mathrm{F1_L}$)				Per-image F1 score ($\mathrm{F1_I}$)			
		True	Rare	Freq.	Rand	True	Rare	Freq.	Rand
Corel-5K	Ground	100.00	99.26	99.55	70.65	100.00	82.62	82.62	82.62
	JEC	40.29	49.92	49.06	33.19	54.43	54.57	56.64	54.99
	TagRel	40.71	52.19	51.37	34.95	54.20	54.33	57.38	54.98
	TagProp	39.34	49.16	47.87	32.53	56.12	56.31	57.73	56.56
	2PKNN	50.78	64.36	63.96	39.39	53.23	53.32	57.01	53.84
	SVM	40.01	50.89	50.13	34.06	57.48	57.58	59.21	57.72
ESP Game	Ground	100.00	99.26	99.56	81.74	100.00	88.60	88.60	88.60
	JEC	36.99	46.26	45.50	28.44	45.29	45.35	49.80	45.85
	TagRel	42.47	58.71	58.42	35.92	47.06	47.11	53.09	47.56
	TagProp	44.88	58.83	58.08	36.61	48.47	48.53	52.99	49.14
	2PKNN	45.17	61.38	61.06	37.83	48.95	49.00	53.43	49.34
	SVM	40.17	50.90	50.20	32.18	52.06	52.12	55.00	52.62
IAPR TC-12	Ground	100.00	99.26	99.71	87.75	100.00	92.83	92.83	92.83
	JEC	34.87	40.77	40.41	28.21	50.40	50.53	53.67	51.03
	TagRel	44.55	56.95	57.01	38.45	52.15	52.24	56.89	52.63
	TagProp	45.93	56.71	56.73	38.56	52.07	52.16	56.85	52.52
	2PKNN	47.43	59.19	59.12	39.43	52.30	52.39	55.39	52.80
	SVM	36.53	41.87	41.23	29.04	53.37	53.52	55.14	53.86
NUS-WIDE	Ground	100.00	98.57	99.12	62.36	100.00	79.88	79.88	79.88
	SoftMax*	47.82	68.43	67.75	36.74	62.76	62.77	65.33	63.17
	Sigmoid*	49.85	69.44	69.08	37.26	63.83	63.84	66.18	64.20
	Ranking*	49.03	69.25	68.84	37.07	63.19	63.20	65.74	63.60
	WARP*	48.28	68.86	68.51	36.72	62.83	62.84	65.33	63.21
	LSEP*	48.05	70.14	69.82	37.70	63.52	63.53	66.10	63.90
	JEC	40.15	58.14	57.57	29.43	57.78	57.80	61.60	58.37
	TagRel	48.28	74.85	75.12	40.19	59.60	59.61	66.25	60.13
	TagProp	53.86	74.14	74.10	40.19	61.41	61.42	65.93	61.84
	2PKNN	51.55	71.15	71.31	37.94	60.67	60.68	63.71	61.12
	SVM	49.89	70.19	69.82	37.64	63.30	63.31	65.60	63.68
MS-COCO	Ground	100.00	98.36	99.20	82.94	100.00	85.46	85.46	85.46
	SoftMax*	56.49	72.08	71.76	51.26	63.79	63.88	66.26	64.33
	Sigmoid*	58.37	73.35	73.06	52.81	65.58	65.65	67.35	66.02
	Ranking*	56.25	71.65	71.37	50.87	63.66	63.74	65.76	64.20
	WARP*	56.19	71.43	71.04	50.34	63.18	63.26	65.68	63.70
	LSEP*	57.99	73.84	73.68	53.14	65.41	65.48	67.47	65.90
	JEC	50.25	63.27	62.66	42.01	57.27	57.37	59.77	58.04
	TagRel	57.60	74.70	74.68	52.78	61.49	61.56	67.47	62.12
	TagProp	60.61	73.66	73.42	52.54	63.98	64.05	66.36	64.45
	2PKNN	59.46	71.70	71.44	49.77	59.86	59.94	66.06	60.49
	SVM	58.94	73.10	72.87	52.44	65.09	65.16	66.86	65.53

(Refer Section 4.1 for details)

predictions remains the same, the performance will not get seriously affected *irrespective* of what labels are incorrectly assigned. In contrast, per-label metrics are expected to provide comparatively better insights about annotation performance for both rare as well as frequently occurring labels, even though the performance scores corresponding to rare labels might be somewhat noisy in the sense that they are based on only a handful of test images.

To further study the bias of per-image metrics toward frequent labels, we assign to all the test images in a dataset (i) the same 3/5 most rare labels, (ii) the same 3/5 most frequent labels, and (iii) 3/5 randomly chosen labels. The fixed set of frequent/rare labels is chosen based on the frequencies of labels in the training subset of a given dataset. While assigning random labels, we pick different randomly chosen labels for each image. Table 7 shows the performance obtained using these label assignment techniques. We can observe that the performance is negligibly low in all the cases, except in the case of per-image metrics with frequent label assignment. Precisely, simply by assigning the same three/five most frequent labels to all the test images, we can achieve quite significant $F1_I$ scores. This is because as we can see in Fig. 1, in real-world datasets the label distributions follow the Zipf's law, due to which there are a small number of frequently occurring labels and a large number of rare labels.

From the above results, we arrive at the conclusion that for the image annotation problem, per-label metrics should be preferred over per-image metrics in general. Additionally, this analysis also suggests another issue in the evaluation schemes that have been followed for over a decade in the image annotation domain, that require each test image to be annotated with a pre-defined fixed number of labels rather than doing variable number of label assignments, and thus warrants more discussion in future work.

Table 7 Performance by assigning the three most rare, the three most frequent, and three randomly chosen labels to each test image

Dataset	Method	Per-label metrics				Per-image metrics		
		P_L	R_L	$F1_L$	N+	P_I	R_I	$F1_I$
Corel-5K	Rare	0.00	1.92	0.00	5	0.24	0.33	0.28
	Frequent	0.34	1.92	0.57	5	17.59	25.50	20.82
	Random	1.16	0.99	1.07	23	1.28	1.68	1.45
ESP Game	Rare	0.00	1.86	0.00	5	0.15	0.13	0.14
	Frequent	0.31	1.86	0.53	5	16.51	18.72	17.55
	Random	1.90	1.80	1.84	96	1.83	1.94	1.88
IAPR TC-12	Rare	0.00	1.72	0.01	5	0.32	0.46	0.37
	Frequent	0.33	1.72	0.55	5	19.10	17.04	18.01
	Random	2.04	1.75	1.88	105	2.05	1.84	1.94
NUS-WIDE	Rare	0.00	3.70	0.00	3	0.04	0.06	0.05
	Frequent	1.06	3.70	1.65	3	28.71	39.37	33.21
	Random	2.95	3.68	3.28	80	2.96	3.71	3.30
MS-COCO	Rare	0.01	3.75	0.02	3	0.33	0.37	0.35
	Frequent	0.93	3.75	1.49	3	24.87	24.29	24.57
	Random	3.48	3.61	3.54	80	3.46	3.59	3.52

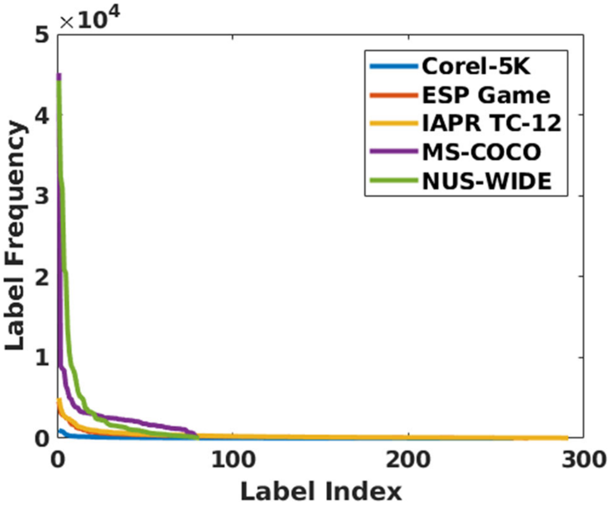


Fig. 1 Frequency of labels in the training sets of each of the five datasets sorted in decreasing order (best viewed in color)

4.2 Dataset diversity

Here we analyze various aspects in the context of diversity in image annotation datasets.

4.2.1 Label diversity

To study this, we compute two measures: percentage unique label-sets and novel label-sets. The former computes what percentage of labels-sets are unique in the ground-truth of the test data, and the latter computes what percentage of label-sets are novel (i.e., not seen in the training data). Note that while computing the percentage of novel label-sets, we omit the uniqueness criterion; i.e., multiple test images can have the same novel label-set.

Table 8 shows the values of these two measures for various datasets. We observe that ESP Game and IAPR TC-12 datasets offer maximum diversity in terms of both unique as well as novel label-sets. However, for the NUS-WIDE dataset, we observe that there are only 12.3% of unique label-sets in the test set, and only 6.7% of test images that have novel label-sets in their ground-truth that are not seen in the training set. This indicates that there is a lack of label diversity in the NUS-WIDE dataset. This is because though it has a large number of images, its vocabulary contains only 81 labels with just around 2.40 labels per image, indicating a low degree of *multi-labelness* in this dataset.

Table 8 Label diversity in test data in terms of percentage "unique" and "novel" label-sets

Dataset	Corel-5K	ESP Game	IAPR TC-12	NUS-WIDE	MS-COCO
Unique	86.8	95.2	95.3	12.3	27.2
Novel	48.9	82.5	73.2	6.7	17.3

4.2.2 Image diversity

One natural expectation from an image annotation dataset is that its test set should contain compositionally novel images that are not seen in entirety in the training set [6]. To study this

Table 9 Performance comparison (using *GoogLeNet*) of various label prediction models (deep-learning based models are marked by '*') over the 20% most overlapping test subsets of various datasets (refer Section 4.2.2 for details)

	Method	Per-label metrics				Per-image metrics			
		P_L	R_L	$F1_L$	mAP_L	P_I	R_I	$F1_I$	mAP_I
Corel-5K	JEC	67.42	74.61	70.83	68.47	62.20	84.42	71.63	71.53
	TagRel	68.02	78.20	72.76	75.58	63.20	85.50	72.68	83.58
	TagProp	70.23	78.25	74.02	76.27	64.60	87.50	74.33	85.42
	2PKNN	71.27	83.22	76.78	83.46	65.80	89.33	75.78	87.06
	SVM	66.32	81.22	73.02	83.81	65.00	88.42	74.92	86.30
ESP Game	JEC	59.77	51.88	55.55	45.97	55.92	64.79	60.03	51.78
	TagRel	55.46	63.50	59.21	64.05	53.91	60.96	57.22	64.49
	TagProp	61.04	64.74	62.84	65.84	56.83	64.96	60.63	69.18
	2PKNN	62.16	64.05	63.09	64.52	57.84	66.46	61.85	68.94
	SVM	56.90	61.44	59.08	64.93	59.47	68.02	63.46	71.12
IAPR TC-12	JEC	55.24	46.62	50.56	43.10	64.27	65.57	64.92	56.84
	TagRel	66.27	65.45	65.85	71.11	65.09	66.19	65.63	73.57
	TagProp	68.75	67.35	68.04	76.03	65.14	66.12	65.62	76.01
	2PKNN	68.22	66.27	67.23	74.18	66.26	67.02	66.63	77.09
	SVM	63.57	56.74	59.96	76.24	68.50	69.55	69.02	79.78
NUS-WIDE	SoftMax*	59.80	65.98	62.74	63.53	61.75	81.68	70.33	87.64
	Sigmoid*	60.18	65.76	62.84	69.02	62.48	82.33	71.05	88.93
	Ranking*	57.75	66.30	61.73	62.91	61.79	81.54	70.30	87.64
	WARP*	58.08	67.26	62.34	62.94	61.94	81.69	70.46	87.48
	LSEP*	59.23	66.77	62.77	65.60	62.51	82.38	71.08	88.86
	JEC	56.54	49.29	52.67	38.02	56.55	75.62	64.71	71.22
	TagRel	58.44	70.99	64.12	66.16	59.69	79.20	68.08	82.76
	TagProp	65.64	71.83	68.59	68.93	60.72	80.44	69.20	85.18
	2PKNN	65.26	69.97	67.53	66.94	58.74	77.97	67.00	81.63
	SVM	63.90	67.79	65.79	66.67	61.57	81.16	70.02	86.95
MS-COCO	SoftMax*	60.62	65.89	63.15	64.98	57.36	89.61	69.95	93.96
	Sigmoid*	59.34	66.92	62.90	69.64	57.70	89.87	70.28	94.47
	Ranking*	56.62	66.74	61.26	65.55	57.19	89.34	69.74	93.97
	WARP*	58.15	65.86	61.77	65.30	56.89	89.01	69.41	93.71
	LSEP*	60.40	67.08	63.57	66.50	57.71	89.93	70.30	94.51
	JEC	60.91	56.91	58.84	44.74	52.23	83.64	64.31	78.14
	TagRel	56.44	65.63	60.69	65.83	53.45	85.07	65.65	87.68
	TagProp	69.09	63.50	66.18	66.14	56.27	88.23	68.72	92.13
	2PKNN	66.74	65.39	66.06	66.48	54.40	86.24	66.71	91.41
	SVM	69.09	63.80	66.34	65.01	56.20	88.03	68.60	91.99

phenomenon, we identify compositionally novel as well as compositionally similar images in the test sets of various datasets and evaluate the performance of different methods on these subsets. To do so, we bin the images in the test set of a given dataset based on their

Table 10 Performance comparison (using *ResNet*) of various label prediction models (deep-learning based models are marked by '*') over the 20% most overlapping test subsets of various datasets (refer Section 4.2.2 for details)

	Method	Per-label metrics				Per-image metrics			
		P_L	R_L	$F1_L$	mAP_L	P_I	R_I	$F1_I$	mAP_I
Corel-5K	JEC	73.50	78.88	76.10	74.73	64.00	86.58	73.59	72.79
	TagRel	72.38	79.02	75.55	77.48	64.20	86.83	73.82	85.70
	TagProp	70.93	79.55	74.99	85.88	64.80	87.83	74.58	86.87
	2PKNN	77.51	87.53	82.22	88.44	68.40	92.83	78.76	90.02
	SVM	68.22	81.64	74.33	88.81	66.60	90.17	76.61	89.42
ESP Game	JEC	58.54	51.67	54.89	45.86	55.68	65.63	60.25	52.94
	TagRel	57.23	62.79	59.88	65.68	54.53	63.23	58.56	66.56
	TagProp	62.59	64.83	63.69	67.21	56.83	66.68	61.37	69.59
	2PKNN	62.03	65.87	63.89	65.90	58.18	68.39	62.87	70.44
	SVM	57.26	59.61	58.41	67.73	60.09	70.90	65.05	74.60
IAPR TC-12	JEC	57.97	46.19	51.42	42.49	66.21	67.04	66.62	58.61
	TagRel	65.99	62.63	64.27	71.58	67.53	67.74	67.63	75.93
	TagProp	68.15	64.74	66.40	76.52	66.61	67.24	66.92	77.98
	2PKNN	69.27	64.97	67.05	75.58	68.29	68.70	68.49	78.97
	SVM	61.25	50.91	55.60	77.15	68.85	68.95	68.90	80.76
NUS-WIDE	SoftMax*	59.96	67.45	63.49	66.84	62.87	81.68	71.05	88.51
	Sigmoid*	61.95	68.61	65.11	72.55	63.93	82.69	72.11	90.30
	Ranking*	60.27	68.24	64.01	67.37	63.49	82.26	71.66	89.64
	WARP*	59.83	67.86	63.59	67.14	63.30	82.01	71.45	89.06
	LSEP*	59.34	68.82	63.73	69.23	63.76	82.54	71.94	90.01
	JEC	59.38	51.27	55.02	40.04	58.37	76.62	66.26	72.47
	TagRel	61.54	73.08	66.82	70.13	61.37	79.83	69.39	84.72
	TagProp	67.10	74.23	70.49	72.15	62.05	80.78	70.19	86.53
	2PKNN	67.33	70.76	69.00	70.41	61.50	79.91	69.51	85.55
	SVM	64.72	69.88	67.20	69.72	63.32	81.88	71.42	88.69
MS-COCO	SoftMax*	55.55	66.04	60.34	66.06	57.61	89.90	70.22	94.23
	Sigmoid*	55.85	66.54	60.73	71.42	57.92	90.16	70.53	94.69
	Ranking*	53.86	66.07	59.34	65.44	57.55	89.77	70.14	94.26
	WARP*	53.98	66.00	59.39	65.17	57.24	89.45	69.81	93.95
	LSEP*	56.08	66.99	61.06	67.82	58.02	90.30	70.65	94.79
	JEC	61.84	55.72	58.62	45.09	52.80	84.71	65.05	79.22
	TagRel	59.11	65.26	62.03	67.08	54.25	85.96	66.52	88.57
	TagProp	68.28	63.73	65.93	67.31	56.69	88.73	69.18	92.40
	2PKNN	63.95	64.70	64.32	66.29	53.93	85.81	66.23	90.44
	SVM	63.85	63.40	63.63	66.84	56.96	88.99	69.46	92.92

feature similarity with the training images. For each test image, we compute its Euclidean distance with every training image, and then take the mean of the distance with the 50 closest images. This mean distance measures the degree of visual overlap of each test image with

Table 11 Performance comparison (using *GoogLeNet*) of various label prediction models (deep-learning based models are marked by '*') over the 20% least overlapping test subsets of various datasets (refer Section 4.2.2 for details)

	Method	Per-label metrics				Per-image metrics			
		P_L	R_L	$F1_L$	mAP_L	P_I	R_I	$F1_I$	mAP_I
Corel-5K	JEC	27.32	30.82	28.96	27.84	32.40	48.92	38.98	37.74
	TagRel	25.74	32.03	28.54	31.34	31.60	47.67	38.01	45.44
	TagProp	25.02	31.41	27.85	34.74	33.60	51.92	40.80	44.67
	2PKNN	26.51	34.97	30.16	51.54	30.40	47.25	37.00	38.99
	SVM	24.88	37.27	29.84	49.67	36.80	56.17	44.47	53.21
ESP Game	JEC	17.97	14.61	16.12	12.15	27.34	31.36	29.21	23.36
	TagRel	21.14	22.18	21.65	22.18	31.22	35.77	33.34	34.51
	TagProp	20.57	19.87	20.22	22.09	31.51	35.99	33.60	34.97
	2PKNN	21.25	20.60	20.92	24.14	29.93	33.64	31.68	34.12
	SVM	18.60	15.69	17.02	23.39	33.67	38.13	35.76	39.18
IAPR TC-12	JEC	16.21	11.65	13.56	10.70	33.28	31.46	32.34	22.77
	TagRel	24.24	22.16	23.16	22.80	36.03	34.48	35.24	34.67
	TagProp	23.45	19.88	21.52	27.43	36.95	36.40	36.67	38.79
	2PKNN	22.06	16.52	18.89	29.33	35.06	33.60	34.32	38.50
	SVM	15.99	11.83	13.60	28.93	39.64	38.26	38.94	43.46
NUS-WIDE	SoftMax*	23.32	25.14	24.19	18.86	42.30	67.14	51.90	67.45
	Sigmoid*	23.02	26.17	24.49	22.49	42.86	67.72	52.49	68.26
	Ranking*	20.40	24.06	22.08	16.86	41.31	64.89	50.48	65.15
	WARP*	20.94	25.80	23.11	17.74	41.59	65.53	50.89	65.76
	LSEP*	21.95	27.31	24.34	19.79	42.68	67.53	52.30	67.88
	JEC	16.91	24.06	19.86	10.10	37.29	58.77	45.63	48.18
	TagRel	17.53	35.18	23.40	20.07	38.12	60.38	46.74	58.92
	TagProp	25.17	27.97	26.50	22.78	41.72	66.01	51.12	65.02
	2PKNN	29.08	19.87	23.61	21.02	34.02	53.66	41.64	50.03
	SVM	21.41	24.16	22.70	20.90	42.53	66.96	52.02	67.17
MS-COCO	SoftMax*	39.52	35.72	37.52	32.87	48.57	54.17	51.21	62.78
	Sigmoid*	42.50	38.53	40.42	39.44	51.17	56.38	53.65	65.58
	Ranking*	39.35	33.93	36.44	32.67	48.22	52.49	50.26	61.33
	WARP*	38.69	33.19	35.73	32.70	47.75	51.70	49.65	60.79
	LSEP*	42.17	39.32	40.69	37.04	50.61	56.03	53.18	64.90
	JEC	34.46	27.82	30.79	17.23	43.14	47.25	45.13	45.01
	TagRel	36.95	41.32	39.01	37.27	46.90	52.40	49.49	56.81
	TagProp	42.77	38.42	40.48	37.18	49.69	54.76	52.10	61.56
	2PKNN	49.74	34.05	40.43	39.15	46.03	50.42	48.13	58.22
	SVM	42.37	37.32	39.69	39.13	51.00	55.83	53.31	63.93

the training images, with larger mean distance denoting less overlap and vice-versa. Using this, we pick the 20% most overlapping and 20% least overlapping images from the test set of each dataset.

Table 12 Performance comparison (using *ResNet*) of various label prediction models (deep-learning based models are marked by '*') over the 20% least overlapping test subsets of various datasets (refer Section 4.2.2 for details)

	Method	Per-label metrics				Per-image metrics			
		P_L	R_L	$F1_L$	mAP_L	P_I	R_I	$F1_I$	mAP_I
Corel-5K	JEC	19.68	25.69	22.29	23.31	27.80	45.75	34.58	35.15
	TagRel	19.25	28.24	22.89	25.59	27.80	45.42	34.49	37.62
	TagProp	20.32	29.35	24.01	38.17	32.20	52.58	39.94	44.31
	2PKNN	21.03	28.24	24.11	47.12	25.80	43.08	32.27	38.44
	SVM	18.86	29.40	22.98	43.29	32.60	53.25	40.44	49.64
ESP Game	JEC	20.91	16.11	18.20	12.92	29.21	33.06	31.02	24.32
	TagRel	21.26	23.29	22.22	22.90	31.99	36.29	34.01	35.52
	TagProp	22.31	22.36	22.34	25.18	33.91	38.43	36.03	38.19
	2PKNN	24.00	25.22	24.60	27.36	32.76	37.27	34.86	38.23
	SVM	18.39	16.86	17.59	26.54	37.27	42.65	39.78	43.39
IAPR TC-12	JEC	20.12	14.44	16.81	11.89	36.39	35.86	36.12	26.43
	TagRel	25.81	24.19	24.97	25.35	38.78	38.66	38.72	38.40
	TagProp	24.59	20.78	22.52	29.85	39.39	39.55	39.47	41.42
	2PKNN	28.58	22.01	24.87	31.80	38.68	38.06	38.36	42.39
	SVM	16.72	12.12	14.05	31.01	39.95	39.49	39.72	44.75
NUS-WIDE	SoftMax*	20.33	25.84	22.76	18.37	41.82	67.57	51.66	67.38
	Sigmoid*	23.46	27.01	25.11	22.96	42.71	68.74	52.69	68.73
	Ranking*	21.84	25.99	23.73	18.90	41.94	67.32	51.69	67.12
	WARP*	21.78	27.21	24.20	19.18	41.45	66.26	51.00	66.03
	LSEP*	20.79	28.89	24.18	20.17	42.22	68.07	52.12	68.02
	JEC	16.96	24.82	20.15	11.24	37.25	59.87	45.93	48.28
	TagRel	17.62	37.71	24.02	20.05	38.07	61.19	46.94	59.21
	TagProp	24.16	30.37	26.91	22.11	41.25	66.42	50.90	65.01
	2PKNN	22.63	28.05	25.05	20.10	40.14	63.74	49.26	63.75
	SVM	19.31	24.40	21.56	20.70	42.17	67.39	51.88	67.41
MS-COCO	SoftMax*	38.04	34.46	36.16	30.70	47.43	52.92	50.02	61.57
	Sigmoid*	41.22	37.91	39.50	38.28	50.13	55.20	52.54	64.32
	Ranking*	38.82	33.03	35.69	31.06	46.81	50.93	48.78	60.05
	WARP*	39.82	31.65	35.27	30.53	46.13	49.85	47.92	59.09
	LSEP*	40.18	39.50	39.83	34.87	49.66	55.27	52.32	63.92
	JEC	42.37	37.32	39.69	39.13	51.00	55.83	53.31	63.93
	TagRel	36.46	42.08	39.07	37.19	45.97	51.47	48.56	56.02
	TagProp	42.18	39.15	40.61	37.06	48.59	53.68	51.01	60.65
	2PKNN	45.89	33.71	38.87	36.59	43.76	48.35	45.94	56.09
	SVM	41.01	37.98	39.44	39.19	49.82	54.83	52.20	63.32

Performances of various methods on these two sets are shown in Tables 9, 10, 11 and 12. From these results, we can make following observations: (i) The performances of all the methods significantly improve on the "20% most" set, and significantly reduce on the "20% least" set compared to that on the full test set. While this is the case in all the datasets, the degree of relative variations in performance is minimum in the ESP Game dataset. This indicates that though all the datasets lack compositional diversity in their images, the ESP Game dataset seems to suffer the least from this. However, since it is an order of magnitude smaller than the NUS-WIDE dataset, this also motivates to create new large-scale datasets that would contain compositionally novel images in their test sets. (ii) The reduction in performance corresponding to per-image metrics on the "20% least" set is much less compared

GT: {animal, road, zebra}	**GT:** {computer}	**GT:** {map}	**GT:** {person, soccer}	**GT:** {railroad, sky}
TagProp: {animal, sky, zebra}	**TagProp:** {computer, sky, water}	**TagProp:** {map, sky, water}	**TagProp:** {person, soccer, sports}	**TagProp:** {clouds, railroad, sky}
Sigmoid: {animal, grass, zebra}	**Sigmoid:** {computer, person, window}	**Sigmoid:** {map, tree, window}	**Sigmoid:** {grass, person, soccer}	**Sigmoid:** {plants, railroad, sky}
GT: {nighttime, sky}	**GT:** {window}	**GT:** {toy}	**GT:** {road, sand}	**GT:** {leaf}
TagProp: {person, sky, wedding}	**TagProp:** {reflection, sky, water}	**TagProp:** {plants, road, sky}	**TagProp:** {flags, sand, sky}	**TagProp:** {sky, tree, water}
Sigmoid: {book, nighttime, sky}	**Sigmoid:** {reflection, water, window}	**Sigmoid:** {animal, clouds, military}	**Sigmoid:** {ocean, sand, snow}	**Sigmoid:** {leaf, sky, water}

Fig. 2 Examples from the "most" overlapping images (top) and the "least" overlapping images (bottom) from the NUS-WIDE dataset. For each image, its ground-truth labels (GT) and the labels predicted using TagProp and Sigmoid methods are shown. The labels in blue are the ones that match with the ground-truth labels (best viewed in color)

to per-label metrics. This again demonstrates that per-label metrics may be more informative than per-image metrics for evaluation in the image annotation task. Figure 2 some example images from the NUS-WIDE dataset along with their ground-truth and predicted labels. Here, we observe that for the images that are more similar to training images, the number of predicted labels that match with the ground-truth labels is higher than those for the images that are less similar.

5 Discussion and conclusions

While it is close to two decades since the problem of image annotation has been studied [30], improving upon the quantitative results has always remained the key focus. In this paper, through detailed experimental analyzes on five popular image annotation datasets, we have made an attempt to highlight some of the core yet mostly overlooked issues related to dataset construction and popularly used evaluation metrics in this domain. Our two key observations are: (i) among all the datasets, the ESP Game dataset offers the maximum label and image diversity, and and is least influenced by the impact of frequent labels on the performance, and (ii) per-label metrics should be preferred over per-image metrics for comparing image annotation techniques in general. Based on these observations, we would like to emphasize the importance of taking careful considerations with respect to these aspects when developing new datasets and techniques for the image annotation task in the future.

Acknowledgments Yashaswi Verma would like to thank the Department of Science and Technology (India) for the INSPIRE Faculty Award 2017.

References

1. Ahn LV, Dabbish L (2004) Labeling images with a computer game. In: ACM SIGCHI Conference on human factors in computing systems
2. Carneiro G, Chan AB, Moreno PJ, Vasconcelos N (2007) Supervised learning of semantic classes for image annotation and retrieval. IEEE Trans Pattern Anal Mach Intell 29(3):394–410
3. Chen M, Zheng A, Weinberger KQ (2013) Fast image tagging. In: ICML
4. Chua TS, Tang J, Hong R, Li H, Luo Z, Zheng Y (2009) NUS-WIDE: A real-world web image database from National University of Singapore. In: ACM CIVR
5. Cristianini N, Shawe-Taylor J (2000) An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods. Cambridge University Press, Cambridge
6. Devlin J, Cheng H, Fang H, Gupta S, Deng L, He X, Zweig G, Mitchell M (2015) Language models for image captioning: The quirks and what works. In: ACL
7. Duygulu P, Barnard K, de Freitas JFG, Forsyth DA (2002) Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: ECCV
8. Feng SL, Manmatha R, Lavrenko V (2004) Multiple Bernoulli relevance models for image and video annotation. In: CVPR
9. Fu H, Zhang Q, Qiu G (2012) Random forest for image annotation. In: ECCV, pp 86–99
10. Gong Y, Jia Y, Leung TK, Toshev A, Ioffe S (2014) Deep convolutional ranking for multilabel image annotation. In: ICLR
11. Grubinger M, Clough PD, Müller H, Deselaers T (2006) The IAPR benchmark: A new evaluation resource for visual information systems. In: International Conference on Language Resources and Evaluation. http://www-i6.informatik.rwth-aachen.de/imageclef/resources/iaprtc12.tgz
12. Guillaumin M, Mensink T, Verbeek J, Schmid C (2009) TagProp: Discriminative metric learning in nearest neighbour models for image auto-annotation. In: ICCV
13. Gupta A, Verma Y, Jawahar CV (2012) Choosing linguistics over vision to describe images. In: AAAI
14. Hardoon DR, Szedmak S, Shawe-Taylor J (2004) Canonical correlation analysis: An overview with application to learning methods. Neural Comput 16(12):2639–2664

15. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: CVPR
16. Hu H, Zhou GT, Deng Z, Liao Z, Mori G (2016) Learning structured inference neural networks with label relations. In: CVPR
17. Johnson J, Ballan L, Fei-Fei L (2015) Love thy neighbors: Image annotation by exploiting image metadata. In: ICCV
18. Kalayeh MM, Idrees H, Shah M (2014) NMF-KNN: Image annotation using weighted multi-view non-negative matrix factorization. In: CVPR
19. Kuznetsova P, Ordonez V, Berg AC, Berg TL, Choi Y (2012) Collective generation of natural image descriptions. In: ACL
20. Li Z, Tang J (2016) Weakly supervised deep matrix factorization for social image understanding. IEEE Trans Image Process 26(1):276–288
21. Li X, Snoek CGM, Worring M (2009) Learning social tag relevance by neighbor voting. Trans Multi 11(7):1310–1322
22. Li Z, Liu J, Xu C, Lu H (2013) Mlrank: Multi-correlation learning to rank for image annotation. Pattern Recogn 46(10):2700–2710
23. Li Z, Liu J, Tang J, Lu H (2015) Robust structured subspace learning for data representation. IEEE Trans Pattern Anal Mach Intell 37(10):2085–2098
24. Li Y, Song Y, Luo J (2017) Improving pairwise ranking for multi-label image classification. In: CVPR
25. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollar P, Zitnic CL (2014) Microsoft COCO: Common objects in contex. In: ECCV
26. Liu F, Xiang T, Hospedales TM, Yang W, Sun C (2017) Semantic regularisation for recurrent image annotation. In: CVPR
27. Makadia A, Pavlovic V, Kumar S (2008) A new baseline for image annotation. In: ECCV
28. Makadia A, Pavlovic V, Kumar S (2010) Baselines for image annotation. Int J Comput Vis 90(1):88–105
29. Moran S, Lavrenko V (2014) A sparse kernel relevance model for automatic image annotation. Int J Multimed Inf Retr 3(4):209–219
30. Mori Y, Takahashi H, Oka R (1999) Image-to-word transformation based on dividing and vector quantizing images with words. In: MISRM'99 First international workshop on multimedia intelligent storage and retrieval management
31. Platt JC (2000) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in large margin classifiers
32. Ren Z, Jin H, Lin ZL, Fang C, Yuille AL (2015) Multi-instance visual-semantic embedding. CoRR arXiv:1512.06963
33. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252
34. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: CVPR
35. Uricchio T, Ballan L, Seidenari L, Bimbo AD (2016) Automatic image annotation via label transfer in the semantic space. CoRR arXiv:1605.04770
36. Verma Y, Jawahar CV (2012) Image annotation using metric learning in semantic neighbourhoods. In: ECCV
37. Verma Y, Jawahar CV (2013) Exploring SVM for image annotation in presence of confusing labels. In: BMVC
38. Verma Y, Jawahar CV (2017) Image annotation by propagating labels from semantic neighbourhoods. Int J Comput Vis 121(1):126–148
39. Verma Y, Gupta A, Mannem P, Jawahar CV (2013) Generating image descriptions using semantic similarities in the output space. In: CVPR Workshop
40. Wang J, Yang Y, Mao J, Huang Z, Huang C, Xu W (2016) CNN-RNN: A unified framework for multi-label image classification. In: CVPR
41. Weston J, Bengio S, Usunier N (2011) WSABIE: Scaling up to large vocabulary image annotation. In: IJCAI
42. Zhang S, Huang J, Huang Y, Yu Y, Li H, Metaxas DN (2010) Automatic image annotation using group sparsity. In: CVPR, pp 3312–3319
43. Zhang M, Zhou Z (2014) A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng 26(99):1819–1837

Ayushi Dutta is currently pursuing M.S. by research in Computer Science and Engineering at IIIT Hyderabad. Her areas of interest are computer vision and machine learning.

Yashaswi Verma received his Bachelor's degree in 2011 and PhD degree in 2017, both in Computer Science and Engineering from IIIT Hyderabad. His broad areas of research include computer vision and applied machine learning.

C. V. Jawahar is a Professor at IIIT Hyderabad, India. His areas of research include robotic and computer vision, machine learning and document image analysis.