Diverse Image Annotation

Baoyuan Wu[†,‡] Fan Jia[†] Wei Liu[‡] Bernard Ghanem[†]
[†]King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
[‡]Tencent AI Lab, Shenzhen, China
wubaoyuan1987@gmail.com fan.jia@kaust.edu.sa wliu@ee.columbia.edu bernard.ghanem@kaust.edu.sa

Abstract

In this work, we study a new image annotation task called diverse image annotation (DIA). Its goal is to describe an image using a limited number of tags, whereby the retrieved tags need to cover as much useful information about the image as possible. As compared to the conventional image annotation task, DIA requires the tags to be not only representative of the image but also diverse from each other, so as to reduce redundancy. To this end, we treat DIA as a subset selection problem, based on the conditional determinantal point process (DPP) model, which encodes representation and diversity jointly. We further explore semantic hierarchy and synonyms among candidate tags to define weighted semantic paths*. It is encouraged that two tags with the same semantic path are not retrieved simultaneously for the same image. This restriction is embedded into the algorithm used to sample from the learned conditional DPP model. Interestingly, we find that conventional metrics for image annotation (*e.g.*, precision, recall, and F_1 score) only consider an overall representative capacity of all the retrieved tags, while ignoring their diversity. Thus, we propose new semantic metrics based on our proposed weighted semantic paths. An extensive subject study verifies that the proposed metrics are much more consistent with human evaluation than conventional annotation metrics. Experiments on two benchmark datasets show that the proposed method produces more representative and diverse tags, compared with existing methods.*

1. Introduction

Image annotation aims to provide keyword tags to describe an image. It not only presents a simple way to understand the image's content, but also provides useful information for other tasks, such as object detection [12] or caption generation [6][33][14]. Many existing methods for image annotation are designed to produce a complete list of tags that cover all contents of an image, such as ML-MG [30] and FastTag [5]. We argue that such a complete list is unnecessary in many cases, especially when redundancy exists among the retrieved tags. We also argue that conventional metrics used to evaluate annotation methods (*e.g.*, precision, recall or F_1 score) should be modified to discourage such redundancy. We will motivate these two points with an example. In Figure 1, the drummer image has the following complete (ground truth) tag list: {"band", "music", "light", "man", "people", "person", "colors", "red", "wheel"}. Obviously, there are several redundancies in this list, including "people" and "person" or "colors" and "red". Clearly, this image can be described using a more compact list (*e.g.*, {"band", "light", "man", "red", "wheel"}), which describes the same content of the image as the complete list, but it is more diverse as it avoids redundancy. Moreover, there is usually an upper limit to the number of tags in the retrieved list for real-world applications. For example, in a crowd-sourced image annotation task, we may ask the human annotator to give at most k (*e.g.*, 3 or 5) tags for each image. Also, this upper limit naturally arises even when annotators are not confined to a specific number of tags, since they choose not to generate a longer list than necessary. As we will show in our experiments, the average size of the tag subsets to describe every image in ESP Game [26] and IAPRTC-12 [11] is around 5 (see Table 1). Based on our subject studies (see Section 5.4), annotators do tend to choose more diverse tags, which hints to the fact

Figure 1. This image with its complete tag list is extracted from the ESP Game [26] dataset. We also show the annotated tags by ML-MG [30] and our method using 3 and 5 tags, respectively. Obviously our tags are more representative and diverse than the tags of ML-MG. Note that the tag list of our method is obtained from sampling, so the tag orders in 3 and 5 tags could be different.

that they choose a compact tag list that covers as much of the image's content as they see necessary. However, when comparing this strategy to the top-k (in terms of individual tag prediction score) retrieved tags of automated annotation methods, we observe a serious discrepancy. For example, as shown in Figure 1, the top-3 tags of the recent annotation method ML-MG [30] are quite redundant. Similar to many other methods, ML-MG focuses on predicting highly representative *individual* tags, while ignoring diversity within the retrieved tag list.

Due to the discrepancy between human image annotations and those of existing methods, we propose a new task, called *diverse image annotation* (DIA), whose goal is to automatically generate a list of k (*e.g*. 3 or 5) tags that *jointly* cover as much useful information as possible in the image. We also propose a method that tackles DIA, by producing individually informative tags that are also *diverse*. As shown in Figure 1, the predicted tags of ML-MG and our method are quite different, primarily due to their diversities. To quantitatively evaluate methods for DIA, we have to propose a new measure that, on one hand, discriminates based on the aggregate semantic value of the retrieved tags and, on the other hand, correlates well with human judgment.

We treat the DIA task as a subset selection problem, where a k sized subset of tags should be retrieved from all possible tags. The conditional determinantal point process (DPP) [17] suitably models such a selection problem. DPP is a probabilistic distribution over subsets of a fixed ground set, and it enforces diversity among elements within a subset, by utilizing global negative correlations among them. The parameters of this DPP model are learned from training samples of images and tags. Once the DPP distribution is learned, the most probable (*i.e*., the most jointly representative and diverse) k tags can be sampled from it for each testing image. However, for meaningful sampling, we exploit semantic relationships between candidate tags, namely their semantic hierarchies and whether or not they are synonyms. These relationships are used to define *weighted semantic paths* for different tags. Two tags are discouraged to be sampled together, if they belong to the same path, thus reducing redundancy in annotation results. These semantic paths are also used to define a similarity measure between a retrieved tag list and the ground truth, which is shown to be more consistent with human annotation than conventional measures (*e.g*., precision, recall and F_1).

Contributions. Our main contributions are three-fold. **(i)** We propose a new task called *diverse image annotation*. **(ii)** We propose a new annotation method that treats DIA as a subset selection problem and uses the conditional DPP model, as well as, tag-specific semantic paths to address it. **(iii)** We define and validate (through subject studies) new semantic metrics to evaluate annotations based on the quality of representation and diversity.

2. Related Work

In this section, we review the main directions in image annotation, including: learning features, exploring tag correlations, designing loss functions, and handling incomplete tags in training data. Then, we show the connections and differences between DIA and these directions.

The **first** direction focuses on learning better image features for annotations, especially based on convolutional neural networks (CNNs) [18]. Such networks learn very promising features for many tasks, such as image classification [15] and object detection [23]. Global CNN-based image features have been used for image annotation too [13]; however, some recent work [10] [27] learns local features for detected bounding boxes, so as to extract more discriminative object-centric features rather than from background. The **second** direction focuses on exploring and exploiting tag correlations. As such, image annotation is treated as a multi-label learning problem, where tag correlations play a key role. Most common tag correlations involve tag-level smoothness [30, 32] (*i.e*., the prediction scores of two semantically similar tags should be similar in the same image), image-level smoothness [13, 30, 32, 20] (*i.e*., visually similar images have similar tags), low rank assumption [2] (*i.e*., the whole tag space is spanned by a lower-dimensional space), and semantic hierarchy [30, 25] (*i.e*. parent tags in a hierarchy are as probable as their children). Note that most of these methods only focus on positive tag correlations, while negative correlations have rarely been explored, such as mutual exclusion [7, 3] and diversity. The **third** direction focuses on designing loss functions that encourage certain types of annotation solutions, such as the (weighted) hamming loss [30, 34] or the pairwise ranking loss [1]. The **fourth** direction handles incomplete tags in training, which has been studied in many recent works [30, 34, 5, 29, 31, 19]. The basic idea is to utilize correlations between provided tags and missing ones to propagate information.

Our DIA task does not exactly belong to any of the above directions. However, there are connections and differences between DIA and these directions, which can help us understand DIA more clearly. The feature learning and loss function design directions can be seen to be independent to DIA. Any progress in these two directions can be seamlessly incorporated into our proposed annotation task. In this work, we adopt global CNN-based features to represent images and the softmax loss function. The second direction is the most related, as tag correlations also play a key role in DIA. However, the intrinsic difference is that existing work focuses on positive correlations, while DIA considers negative ones. Although mutual exclusion falls into this category too, it only involves a pair of tags. A related work presented in [22] utilizes the pairwise redundancy between tags in a sequential tag selection process, for the image retagging

task in the social media scenario. In contrast, DIA takes into account overall negative correlations across all tags. Interestingly, handling incomplete/missing tags seems to have an opposite goal as DIA, since the former seeks the complete tag list from a subset, while DIA targets for a subset from the complete list. However, they are not contradictory to each other because they target for different challenges. The motivation of handling incomplete tags is that the number of fully labeled images is insufficient, while most web images are partially labeled. Thus, learning from massive partially labeled images becomes valuable. But again, the tag diversity is not considered. In contrast, DIA provides a compact tag list that is not only individually representative but also diverse, thus, trying to bridge the gap between automatic image annotation and human annotation. Actually these two tasks can be combined together, where a complete tag list is firstly predicted, then DIA extracts a representative and diverse subset from it. Moreover, the DPP model has been applied to many computer vision tasks, where the diversity is required, such as image retrieval [16][17] and video summarization [9][35]. However, to the best of our knowledge, this work is the first attempt to applying DPP to image annotation.

3. Task and Model

3.1. Diverse Image Annotation (DIA)

The training image set is denoted as $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, where $\mathbf{x}_j \in \mathbb{R}^d$ is the d-dimensional feature representing the j^{th} image. For each image $\mathbf{x}_j$, a ground-truth tag subset $\mathcal{Y}_j \subset \mathcal{T} = \{1, 2, \ldots, m\}$ is also provided, with $\mathcal{T}$ being the whole tag set including m candidate tags. Our task is to learn a model based from all pairs $\{(\mathbf{x}_j, \mathcal{Y}_j)\}_{j=1}^n$ to predict a representative and diverse tag subset with at most k (a user defined integer) tags for each testing image.

3.2. Conditional DPP Model

The parametric conditional DPP with respect to an image and tag subset pair $(\mathbf{x}, \mathcal{Y})$ is formulated as follows [17]:

$$\mathcal{P}_{\mathbf{W}}(\mathcal{Y}|\mathbf{x}) = \frac{\det(\mathbf{L}_{\mathcal{Y}}(\mathbf{x}; \mathbf{W}))}{\det(\mathbf{L}(\mathbf{x}; \mathbf{W}) + \mathbf{I})}, \tag{1}$$

where the kernel matrix for all m tags $\mathbf{L}(\mathbf{x}; \mathbf{W}) \in \mathbb{R}^{m \times m}$ is positive semi-definite with $\mathbf{W}$ being its parameters. $\mathbf{L}(\mathbf{x}; \mathbf{W})$ can also be denoted as $\mathbf{L}_{\mathcal{T}}(\mathbf{x}; \mathbf{W})$. $\mathbf{I}$ is identity matrix. $\mathbf{L}_{\mathcal{Y}}(\mathbf{x}; \mathbf{W}) \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{Y}|}$ is a sub-matrix generated by extracting the rows and columns corresponding to the tags in $\mathcal{Y} \subset \mathcal{T}$ from $\mathbf{L}(\mathbf{x}; \mathbf{W})$. $\det(\mathbf{L}_{\mathcal{Y}})$ indicates the determinant of $\mathbf{L}_{\mathcal{Y}}$, and it is good for encoding negative correlations. Let us see an simple example that $\mathbf{L}_{\mathcal{Y}} = [a_{11}, a_{12}; a_{21}, a_{22}]$, and a_{11} and a_{22} indicate the individual scores of two tags respectively, while a_{12} and a_{21} denote tag correlations. Its determinant is $\det(\mathbf{L}_{\mathcal{Y}}) = a_{11}a_{22} - a_{12}a_{21}$. If $\det(\mathbf{L}_{\mathcal{Y}})$ is small, indicating this two tags are highly correlated, then the probability $\mathcal{P}_{\mathbf{W}}(\mathcal{Y}|\mathbf{x})$ is small; if $\det(\mathbf{L}_{\mathcal{Y}}) = 0$, indicating two tags are fully correlated, then $\mathcal{P}_{\mathbf{W}}(\mathcal{Y}|\mathbf{x})$ is 0. Regarding the general $\mathbf{L}_{\mathcal{Y}}$, if it is rank deficient because the included tags are highly correlated, then its probability is also 0. Obviously the model (1) discourages the tag subset with redundant tags.

Using the quality/diversity (here "quality" refers to "representation") decomposition [17], we have

$$\mathbf{L}_{ij}(\mathbf{x}; \mathbf{W})) = q_i(\mathbf{x}) \boldsymbol{\phi}_i(\mathbf{x})^\top \boldsymbol{\phi}_j(\mathbf{x}) q_j(\mathbf{x}), \tag{2}$$

where $\mathbf{W} = [\mathbf{w}_1\ \mathbf{w}_2\ \ldots \mathbf{w}_m]$ denotes the set of quality parameters, one for each tag. $q_i(\mathbf{x}; \mathbf{w}_i) = \exp(0.5 \mathbf{w}_i^\top \mathbf{x})$ is the quality term, indicating the individual score of $\mathbf{x}$ wrt tag i. $\boldsymbol{\phi}_i(\mathbf{x}) \in \mathbb{R}^m$ is a normalized diverse feature vector, with $\| \boldsymbol{\phi}_i(\mathbf{x}) \| = 1$. $\mathbf{S}(\mathbf{x}) = \boldsymbol{\phi}_i(\mathbf{x}) \boldsymbol{\phi}_i(\mathbf{x})^\top \in \mathbb{R}^{m \times m}$ is the similarity matrix among tags. In this work, we adopt a similarity matrix independent of $\mathbf{x}$, thus we denote it as $\mathbf{S}$ for clarity. Specifically, we adopt the cosine similarity,

$$\mathbf{S}(i, j) = \frac{1}{2} + \frac{\langle \mathbf{t}_i, \mathbf{t}_j \rangle}{2 \|\mathbf{t}_i\|_2 \|\mathbf{t}_j\|_2} \in [0, 1] \ \ \forall\, i, j \in \mathcal{T}, \tag{3}$$

where the tag representation $\mathbf{t}_i \in \mathbb{R}^{50}$ is derived from the GloVe algorithm [21]. Then, Eq (1) can be reformulated as

$$\mathcal{P}_{\mathbf{W}}(\mathcal{Y}|\mathbf{x}) = \frac{\prod_{i \in \mathcal{Y}} [\exp(\mathbf{w}_i^\top \mathbf{x})] \det(\mathbf{S}_{\mathcal{Y}})}{\sum_{\mathcal{Y}' \subset \mathcal{T}} \prod_{i \in \mathcal{Y}'} [\exp(\mathbf{w}_i^\top \mathbf{x})] \det(\mathbf{S}_{\mathcal{Y}'})}, \tag{4}$$

where $\mathbf{S}_{\mathcal{Y}} \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{Y}|}$ and $\mathbf{S}_{\mathcal{Y}'} \in \mathbb{R}^{|\mathcal{Y}'| \times |\mathcal{Y}'|}$ are sub-matrices of $\mathbf{S}$ corresponding to tag subsets $\mathcal{Y}, \mathcal{Y}' \subset \mathcal{T}$.

3.3. Learning

Assuming that the diversity kernel $\mathbf{S}$ is given and following [17], we learn the parameter $\mathbf{W}$ by minimizing the negative log likelihood with an ℓ_2 regularization,

$$\begin{aligned} \mathcal{L}(\mathbf{W}) =& -\frac{1}{n} \sum_j^n \log \mathcal{P}_{\mathbf{W}}(\mathcal{Y}_j|\mathbf{x}_j) + \frac{\eta}{2} \sum_i^m \| \mathbf{w}_i \|_2^2 \\ =& \frac{\eta}{2} \sum_i^m \| \mathbf{w}_i \|_2^2 - \frac{1}{n} \sum_j^n \Big[\sum_{i \in \mathcal{Y}_j} \mathbf{w}_i^\top \mathbf{x}_j - \log \det(\mathbf{S}_{\mathcal{Y}_j}) \Big] \\ &+ \frac{1}{n} \sum_j^n \log \Big[\sum_{\mathcal{Y}'_j \in \mathcal{T}} \prod_{i \in \mathcal{Y}'_j} [\exp(\mathbf{w}_i^\top \mathbf{x}_j)] \det(\mathbf{S}_{\mathcal{Y}'_j}) \Big]. \end{aligned} \tag{5}$$

It is easy to prove that $\mathcal{L}(\mathbf{W})$ is a convex function wrt each $\mathbf{w}_i$. Thus, we adopt a simple gradient-based method to min-

imize $\mathcal{L}(\mathbf{W})$. The gradient wrt $\mathbf{w}_i$ is computed as follows:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\mathbf{w}_i} =& \eta \mathbf{w}_i - \frac{1}{n}\sum_j^n \mathbf{x}_j \mathbb{I}_{i\in\mathcal{Y}_j} + \\ & \frac{1}{n}\sum_j^n \sum_{\mathcal{Y}'_j\in\mathcal{T}} \frac{\exp(\mathbf{w}_i^\top \mathbf{x}_j)\det(\mathbf{S}_{\mathcal{Y}'_j})\mathbf{x}_j \mathbb{I}_{i\in\mathcal{Y}'_j}}{\sum_{\mathcal{Y}'_j\in\mathcal{T}}\prod_{i'\in\mathcal{Y}'_j}[\exp(\mathbf{w}_{i'}^\top \mathbf{x}_j)]\det(\mathbf{S}_{\mathcal{Y}'_j})} \\ =& \eta \mathbf{w}_i + \frac{1}{n}\sum_j^n \mathbf{x}_j \big[\sum_{\mathcal{Y}'_j\in\mathcal{T}} \mathcal{P}_{\mathbf{W}}(\mathcal{Y}'_j|\mathbf{x}_j)\mathbb{I}_{i\in\mathcal{Y}'_j} - \mathbb{I}_{i\in\mathcal{Y}_j}\big],\end{aligned} \tag{6}$$

where indicator function $\mathbb{I}_{i\in\mathcal{Y}_j}$ is 1 if $i \in \mathcal{Y}_j$, otherwise 0. $\sum_{\mathcal{Y}'_j\in\mathcal{T}} \mathcal{P}_{\mathbf{W}}(\mathcal{Y}'_j|\mathbf{x}_j)\mathbb{I}_{i\in\mathcal{Y}'_j}$ can be seen as the marginal probability of $\mathbf{x}_j$ wrt tag i. It is equivalent to the diagonal entry $\mathbf{K}_{ii}$ of the marginal kernel $\mathbf{K}(\mathbf{x}_j;\mathbf{W}) = \mathbf{L}(\mathbf{x}_j;\mathbf{W})/(\mathbf{L}(\mathbf{x}_j;\mathbf{W})+\mathbf{I})$, with $\mathbf{K}_{ii}(\mathbf{x}_j) = \sum_{i'=1}^m \frac{\lambda_{i'}}{\lambda_{i'}+1}\boldsymbol{\upsilon}_{i'}(i)^2$, where $\lambda_{i'}$ and $\boldsymbol{\upsilon}_{i'}$ are the i'-th eigenvalue and eigenvector of the kernel $\mathbf{L}(\mathbf{x}_j;\mathbf{W})$ respectively, derived from the SVD decomposition: $\mathbf{L}(\mathbf{x}_j;\mathbf{W}) = \sum_{i'=1}^m \lambda_{i'}\boldsymbol{\upsilon}_{i'}\boldsymbol{\upsilon}_{i'}^\top$. Then we obtain the gradient as follows:

$$\frac{\partial \mathcal{L}}{\mathbf{w}_i} = \eta \cdot \mathbf{w}_i + \frac{1}{n}\sum_j^n \mathbf{x}_j \big[\mathbf{K}_{ii}(\mathbf{x}_j) - \mathbb{I}_{i\in\mathcal{Y}_j}\big] \tag{7}$$

Given $\frac{\partial \mathcal{L}}{\mathbf{w}_i}$, the back-propagation and stochastic gradient descent algorithm [24] are adopted to optimize $\mathbf{W}$.

Note that if we replace $\mathbf{S}$ by the the identity matrix $\mathbf{I}$, then the above learning can be seen as a standard multi-label learning, where the tag subset is transformed to a label powerset. We refer the reader to [36] for details about label powerset based multi-label learning. In this case, the tag correlations are not utilized at all. In fact, $\mathbf{S}$ in the DPP model serves to encourage negative correlations between tags, since it penalizes the probability of the subset including highly-correlated tags. Thus, a subset with representative (from $\mathbf{q}$) and diverse (from $\mathbf{S}$) tags is encouraged.

4. Sampling

Given the learned conditional DPP model, we can produce a representative and diverse tag subset for each testing image by sampling from the learned distribution. Before detailing the sampling algorithm in Section 4.3, we first introduce a few pertinent concepts, namely semantic hierarchy, synonyms (Section 4.1) and weighted semantic path (4.2) because they play important roles in sampling.

4.1. Semantic Hierarchy and Synonyms

Semantic hierarchy (SH) was explored in ML-MG [30] for image annotation. It describes the semantic dependencies among tags. For example, "woman" is a "person". Figure 2-left shows a part of the semantic hierarchies of ESP Game [26]. Please refer to [30] for the detailed definition of semantic hierarchy. In ML-MG, it is assumed that if the descendant tag (*e.g.*, "woman") exists, then all its ancestor tags (*e.g.*, "person" and "people") must exist too. In contrast, the usage of SH in our sampling algorithm is different. We assume that two tags with semantic dependency cannot be selected together, thus, reducing redundancy. Also, we will use SH to define semantic metrics for DIA evaluation.

Synonyms indicates the state when two tags have the same or similar meaning, such as "people" and "person", or "rock" and "stone". We find that in many benchmark image datasets, such as ESP Game [26] and IAPRTC-12 [11], there are many pairs of synonymous tags, according to Wordnet [8]. In [28], synonyms are utilized to modify the evaluation metric. In this work, synonyms are not only used to define semantic metrics, but also utilized to discourage synonymous tags from being selected simultaneously.

4.2. Weighted Semantic Path

Here, we propose a new concept, called *weighted semantic path* (SP), to merge the idea of SH and synonyms together. We present a simple example in Figure 2 to illustrate some definitions of SP. Firstly, as shown in Figure 2-left, We can find some directed paths among the 5 candidate tags, such as ["person"→"woman"→"lady"]. However, some paths may represent the same or similar meaning, as their constituent tags are synonyms, such as ["person"→"woman"→"lady"] and ["people"→"woman"→"lady"]. We propose that if two directed paths are only different at synonymous tags, then they should be merged into one path, such as ["person", "people")→"woman"→"lady"], as shown in Figure 2-middle. All semantic paths corresponding to the whole tag set $\mathcal{T}$ is denoted as $SP_{\mathcal{T}} = \{sp_1, \ldots, sp_r\}$.

For each semantic path, we focus on two of its importance properties, namely the hierarchy layers and tag weights. See the first path shown in Figure 2-middle, the tag layers are $[(2,2),1,0]$ respectively. As for tag weights, we seek a model that scores a tag based on the content it represents for an image. To motivate such a model, we make two observations. **(i)** A descendant tag can tell more specific information than its ancestor tags (*e.g.* "women" is more informative than "person"). Therefore, the weight of the descendant tag should be higher than the weights of its ancestor tags. **(ii)** The number of descendants of each tag is also considered. For example, in the SH of IAPRTC-12, "person" has 15 descendants, while "sport" has 3. As such, one can assume that "person" is less discriminative than "sport". Thus, we model the tag weight to be inversely proportional to its number of descendants. Combining both observations, we define the weight of tag y_i in path sp_j as $\omega_{ij} = \tau^{l_{ij}}/|d_i|$, where $|d_i|$ indicates the number of descendants of y_i, l_{ij} represents the layer of y_i in sp_j, and $\tau \in (0,1)$ denotes the decay factor between

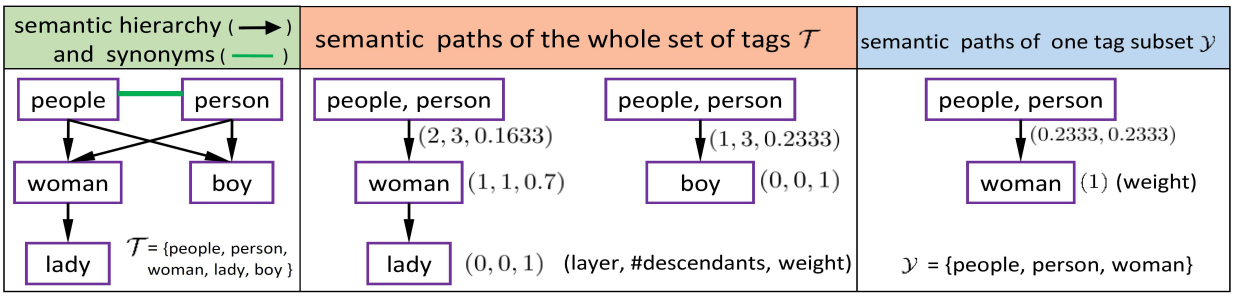

Figure 2. An example of constructing the semantic paths from the semantic hierarchy and synonyms. **Left**: The semantic hierarchy and synonyms of the whole tag set $\mathcal{T}$. **Middle**: The semantic paths of $\mathcal{T}$, and tag weights in each path. **Right**: The semantic paths of one tag subset $\mathcal{Y}$, and tag weights in each path.

layers. In this work, we set $\tau = 0.7$. Consequently, the weight of tag y_i in the whole set of semantic paths is defined as the sum of its weights in all semantic paths, *i.e.*, $\omega_i = \sum_j^{|SP|} \omega_{ij}$. As shown in Figure 2-middle, the weight of "people" is $0.3966 = 0.1633 + 0.2333$, as it exists in two paths. The weights of all tags can be concatenated into one vector: $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_m)$.

In the above paragraph we have introduced the construction of the semantic paths $SP_{\mathcal{T}}$ of the whole of tags $\mathcal{T}$. In the following, we also define the semantic paths $SP_{\mathcal{Y}}$ of the tag subset $\mathcal{Y}$ of one image, as shown in Figure 2-right, where we set $\mathcal{Y} = \{$"people", "person", "woman"$\}$. Firstly, from $SP_{\mathcal{T}}$, we crop the partial paths where tags in $\mathcal{Y}$ occur, *i.e.*, [("person", "people")→"woman"]. Then, to ensure the leaf tag weight in each path of $SP_{\mathcal{Y}}$ to be 1 (such that each independent path tells the same amount of content), we should adjust the tag weight. Thus, the weight of "woman" is changed from 0.7 to 1, and the change factor is $1/0.7$. Using the same factor, we adjust the weights of "people" and "person" from 0.1633 to $0.2333 = 0.1633 * (1/0.7)$.

4.3. DPP Sampling with Weighted Semantic Paths

Here, we present the sampling algorithm based on the learned conditional DPP model (see Section 3.3), and the weighted semantic paths SP. The pseudo-code is shown in Algorithm 1. The inputs are the testing image feature $\mathbf{x}$, the learned parameters $\mathbf{W}$, the similarity matrix $\mathbf{S}$, two integers k_1, k_2 with $m > k_1 > k_2 > 0$, semantic paths $SP_{\mathcal{T}}$ with tag weights $\boldsymbol{\omega}$. The output is the tag subset $\mathcal{Y}_{k_2}$ with at most k_2 tags for this testing image. Although the number k_2 should be provided as a priori, it is not a strict requirement. As k_2 only serves as a upper limit of the sampled tags, rather than requiring exactly k_2 tags. In practice, k_2 can be determined according to user's requirement or experience.

Algorithm 1 is a modified version of the standard k-DPP sampling algorithm [17], by embedding the weighted semantic paths in Line 7-9. It consists of two stages. The **first** stage ranges from Line 1 to Line 3, to compute the eigenvalues $\{\lambda_j\}$ and the elementary symmetric polynomials $\{e_k^N\}$, and it is the normalization term of the k-DPP model, about

Algorithm 1: DPP Sampling with Weighted Semantic Paths

Input: $\mathbf{x}, \mathbf{W}, \mathbf{S}, k_1, k_2, SP_{\mathcal{T}}, \boldsymbol{\omega}$.
Output: A tag subset $\mathcal{Y}_{k_2}$.

1 compute the quality score $q_i(\mathbf{x}; \mathbf{w}_i) = \exp(\frac{1}{2}\mathbf{w}_i^\top \mathbf{x})$, and the kernel matrix $\mathbf{L} = \mathrm{diag}(\mathbf{q}) \cdot \mathbf{S} \cdot \mathrm{diag}(\mathbf{q})$ with $\mathbf{q} = (\ldots; q_i(\mathbf{x}, \mathbf{w}_i); \ldots) \in \mathbb{R}^m$;
2 determine the tag set $\mathcal{Y}_{k_1}$ corresponding to the largest k_1 entries in $\mathbf{q}$, and the sub-kernel $\mathbf{L}_{\mathcal{Y}_{k_1}} = \mathbf{L}(\mathcal{Y}_{k_1}, \mathcal{Y}_{k_1})$;
3 compute eigenvalues $\{\lambda_j\}$ of $\mathbf{L}_{\mathcal{Y}_{k_1}}$, and $e_k^N = \sum_{\mathcal{Y}_k \subset [N], |\mathcal{Y}_k| = k} \prod_{j \in \mathcal{Y}_k} \lambda_j$ for $N = 0, 1, \ldots, k_1$ and $k = 0, 1, \ldots, k_2$;
4 **for** $t = 1, \ldots, 10$ **do**
5 $\mathcal{Y}_t = \emptyset, l = k_2$
6 **for** $i = k_1, \ldots, 1$ **do**
7 **if** $\mathcal{Y}_{k_1}(i)$ *is in the same semantic path in* $SP_{\mathcal{T}}$ *with any tag in* $\mathcal{Y}_t$ **then**
8 **skip** to the next iteration
9 **end**
10 **if** $u \sim U[0, 1] < \lambda_i \frac{e_{l-1}^{i-1}}{e_l^i}$ **then**
11 $\mathcal{Y}_t \leftarrow \mathcal{Y}_t \cup \mathcal{Y}_{k_1}(i), \quad l \leftarrow l - 1$
12 **end**
13 **if** $l = 0$ **then** Break **end**
14 **end**
15 compute tag weights $\boldsymbol{\omega}_{\mathcal{Y}_t} = \boldsymbol{\omega}(\mathcal{Y}_t)$, and the weight summation $\overline{\omega}_{\mathcal{Y}_t} = \sum_j^{|\mathcal{Y}_t|} \boldsymbol{\omega}_{\mathcal{Y}_t}(j)$
16 **end**
17 **return** $\mathcal{Y}_{k_2} = \arg\max_{\mathcal{Y}_{t'}} \overline{\omega}_{\mathcal{Y}_{t'}}$.

which we refer the readers to [17] for more details. $\{\lambda_j\}$ and $\{e_k^N\}$ will be used to compute the sampling probability of each tag. Note that we pick k_1 candidate tags with the k_1 largest entries in $\mathbf{q}$, and the output k_2 tags are sampled from these k_1 tags, rather than from $\mathcal{T}$. As a result, the sampling cost is significantly reduced, and most negative labels will not be sampled, with the cost that some positive tags may also be removed. The **second** stage is sampling (Line 4-17). Since the sampling may produce different subsets, we run 10 samplings to produce 10 subsets. Line 7-9 ensures that two tags in the same semantic path will not be selected together. $\lambda_i e_{l-1}^{i-1} / e_l^i$ in Line 10 indicates the probability of adding tag $\mathcal{Y}_{k_1}(i)$ into the subset, given the current subset $\mathcal{Y}_t$. At the end of each sampling process, the tag weight

Data	C1	C2	C3	C4	C5	C6	C7	C8
ESP Game [26]	18689	2081	268	597	129	9	106	4.56
IAPRTC-12 [11]	17495	1957	291	536	178	11	139	5.85

Table 1. Details of the semantic hierarchies, synonyms and semantic paths of two benchmark datasets. The notations C1 to C8 indicate the numbers of: training images, testing images, candidate tags, feature dimension, parent-child pairs in SH, synonyms pair, semantic paths corresponding to the whole set of tags, average semantic paths corresponding to the tag subset of each image.

summation in the sampled subset is computed (see Line 15). Finally, we pick the subset with the largest weight summation among 10 sampled subsets (see Line 17), as we believe that the larger weight summation indicates more contents.

5. Experiments

5.1. Experimental Settings

Datasets. We run experiments on two benchmark datasets in image annotation, namely ESP Game (20770 images, 268 tags) [26] and IAPRTC-12 (19452 images, 291 tags) [11]. Regarding features, we extract a 4096-dimensional feature vector for each image, using the pre-trained VGG-F model[1] [4]. Then, we perform dimensionality reduction using PCA to maintain 80% of the feature variance. As described in Section 4, we construct the semantic hierarchies[2], synonyms and the weighted semantic paths. The basic statistics of these terms in two datasets are shown in Table 1. The complete set of semantic hierarchies, synonyms, weighted semantic paths of both datasets are provided in **supplementary material**. Note that we find many repeating images in IAPRTC-12, so we remove these redundant images (170 training and 5 testing images) in experiments.

Parameters. The parameters of stochastic gradient descent for learning $\mathbf{W}$ are set as follows: the initial learning rate is 20, and the decay is 0.02. The learning rate is updated every 50 iterations with momentum 0.9 and batch size 1024. The maximum number of epochs is 5 and the parameter of the ℓ_2 regularization is $\eta = 0.0001$ (see Eq (5)).

Compared Methods. We first compare with existing image annotation methods, namely ML-MG [30] and LEML[34]. Also, we compare three variants of the proposed method, including DPP-I-topk, DPP-S-topk, and DPP-S-sampling. DPP-I-topk denotes the case when the $\mathbf{S}$ matrix is replaced by identity during the learning phase, and then the tags with the top-k largest quality scores are retrieved. DPP-S-topk denotes the case where we learn the conditional DPP model with $\mathbf{S}$, then retrieve tags with the top-k largest quality scores. DPP-S-sampling means that we learn the conditional DPP model with $\mathbf{S}$, and then use Algorithm 1 to retrieve at most k tags for the testing image.

[1]It is downloaded from http://www.vlfeat.org/matconvnet/ pretrained/

[2]The semantic hierarchies of ESP Game and IAPRTC-12 are provided by the author of ML-MG [30].

Algorithm 2: Semantic Metrics

Input: The ground-truth tag subset $\mathcal{Y}$, the predicted tag subset $\mathcal{Y}'$.
Output: P_{sp}, R_{sp} and F_{1-sp}.

construct the semantic paths $SP_{\mathcal{Y}}$ and $SP_{\mathcal{Y}'}$;
for $sp_j \in SP_{\mathcal{Y}}$ **do**
  for $y_i \in sp_j$ **do**
    if $y_i \in \mathcal{Y}'$ **then**
      $s_{y_i,j} = \omega_{y_i,j}$
    else
      $s_{y_i,j} = 0$
    end
  end
  $s_j = \max_{y_i \in sp_j} s_{y_i,j}$
end
$\mathrm{P}_{sp} = \sum_j^{|SP_{\mathcal{Y}}|} s_j / |SP_{\mathcal{Y}'}|$;
$\mathrm{R}_{sp} = \sum_j^{|SP_{\mathcal{Y}}|} s_j / |SP_{\mathcal{Y}}|$;
$\mathrm{F}_{1-sp} = 2(\mathrm{P}_{sp} \cdot \mathrm{R}_{sp})/(\mathrm{P}_{sp} + \mathrm{R}_{sp})$;

5.2. Semantic Metrics

Many evaluation metrics have been used in image annotation and multi-label learning, such as the example-based precision, recall and F_1 score [36]. However, these metrics are not very suitable for the DIA task, as they treat every tag equally and independently. In other words, they focus on evaluating representation, but ignoring diversity. Thus, we propose semantic metrics to evaluate representation and diversity jointly. Semantic metrics are defined based on the semantic paths (see Section 4.2), rather than individual tag metrics. Algorithm 2 shows how to compute the scores of semantic metrics for one testing image. In experiments we will report the average score over all testing images.

5.3. Results

The results evaluated by semantic metrics on ESP Game and IAPRTC-12 are shown in Table 2. Since the compared methods belong to different categories, in the following we present the comparisons with different groups. The **first** category is ML-MG, which utilizes the linear inequality constraint to encourage the tag order satisfying the semantic hierarchy. Thus, the ancestor tags are always ranked before their descendant tags. Besides, ML-MG also utilizes the tag co-occurrence to encourage similar tags to have similar scores. Then the tags in one semantic path will be ranked close to each other. As a result, if we pick the top-3 or top-5 tags from the tag ranking list of ML-MG, it is expected that more ancestor tags (corresponding to lower weights in the semantic path), and tags from fewer paths will be obtained. Such tags will cover less-representative and less information. This is why ML-MG shows the worst performance evaluated by semantic metrics. In the **second** category, LEML utilizes the empirical risk minimization (ERM) framework with a decomposed loss over each tag; DPP-I-topk can be seen as a label-powerset-based multi-label method. They don't consider the ranking relationships (as

did in ML-MG), neither the tag diversity (as did in DPP-S-sampling). Thus their performances range between ML-MG and DPP-S-topk, DPP-S-sampling. The **last** category includes DPP-S-topk and DPP-S-sampling, which takes the diversity into account. The difference is: given the learned DPP model with $\mathbf{S}$, DPP-S-sampling utilizes Algorithm 1 to obtain the tag subset, while DPP-S-topk chooses the top-k tags according to the quality scores. They shows the best performance in most cases. In details, in the case of 3 tags, DPP-S-sampling gives significant improvements of F_{1-sp} scores over DPP-S-topk: 13.94% on ESP Game, 19% on IAPRTC-12. This verifies the efficacy of the proposed sampling algorithm. It is notable that DPP-S-sampling always shows the best recall R_{sp} (see Line 13 in Algorithm 2) than other methods. The reason is DPP-S-sampling encourages to cover more diverse tags from different semantic paths, thus its nominator value s_j of R_{sp} is very likely to be higher than the values of other methods. Besides, the denominator value $|SP_{\mathcal{Y}}|$ of R_{sp}, i.e., the number of ground-truth semantic paths, is same for all methods. Hence, DPP-S-sampling gives higher recall than others. However, we also observe there is a significant decreasing on P_{sp} of DPP-S-sampling, from 3 tags to 5 tags. The first reason is that Algorithm 1 ensures the number of semantic paths of the sampled subset (*i.e.*, $|SP_{\mathcal{Y}'}|$) to equal to the number of included tags, as two tags in the same path cannot be selected simultaneously. In contrast, the number of paths of the subset produced by DPP-S-topk and other compared methods is likely to be smaller than the number of included tags, as tags in the same path could be selected together. Thus, when computing P_{sp} (see Line 12 in Algorithm 2), the denominator value $|SP_{\mathcal{Y}'}|$ of DPP-S-sampling will not be smaller than the values of other methods (actually it is larger at most times). Meanwhile, since the 3 tags and 5 tags are sampled from the top-6 and top-8 candidate tags respectively (see Algorithm 1) for DPP-S-sampling, if the additional 2 candidate tags don't include positive tags in different semantic paths, or just one, then the nominator value of P_{sp} will not increase much. Hence, P_{sp} of DPP-S-sampling could be lower than the one of DPP-S-topk, in the case of 5 tags.

Moreover, the comparison between DPP-S-topk and DPP-I-topk could highlight the influence of $\mathbf{S}$. $\mathbf{S}$ will influence the tag ranking, i.e., the quality scores of two similar (or highly related) tags should be not to close. As shown in Table 2, DPP-S-topk shows improvements over DPP-I-topk in most cases. It tells that $\mathbf{S}$ indeed contributes to produce more representative and diverse tags. But meanwhile, the limited improvements reminds us that this $\mathbf{S}$ derived from the cosine similarity between GloVe vectors are not perfect enough. Exploring a better $\mathbf{S}$ will be a future direction of our research. Due to the space limit, in the **supplementary material** we provide some additional results, including: a) the evaluation results by conventional metrics, b) the results of combining our sampling algorithm with ML-MG and LEML, to verify the diversity contribution of the sampling algorithm, and c) quality results of some images with predicted tag subsets, as well as the evaluation scores.

Data	metric→ / method↓	3 tags			5 tags		
		P_{sp}	R_{sp}	F_{1-sp}	P_{sp}	R_{sp}	F_{1-sp}
ESP Game	ML-MG [30]	30.51	16.55	19.73	36.61	29.63	30.59
	LEML [34]	45.16	23.61	28.31	41.82	33.87	34.58
	DPP-I-topk	47.39	23.77	29.02	44.79	35.37	36.77
	DPP-S-topk	**48.07**	23.93	29.34	**45.33**	35.6	**37.04**
	DPP-S-sampling	42.37	**30.48**	**33.43**	36.15	**40.1**	35.96
IAPRTC-12	ML-MG [30]	35.74	17.99	21.89	41.95	29.56	31.98
	LEML [34]	43.03	19.54	24.86	**47.27**	29.76	33.67
	DPP-I-topk	42.88	20.24	25.3	46.64	31.06	34.35
	DPP-S-topk	42.95	20.2	25.32	47.14	31.13	**34.56**
	DPP-S-sampling	**44.01**	**25.16**	**30.13**	38.91	**34.21**	34.23

Table 2. Results ($\%$) evaluated by semantic metrics on ESP Game and IAPRTC-12. The higher value indicates the better performance, and the best result in each column is highlighted in bold.

5.4. Subject Study

To evaluate the efficacy of the proposed semantic metrics, a subject study via Amazon Mechanical Turk is conducted for two algorithmic comparisons: DPP-S-sampling vs ML-MG and DPP-S-sampling vs DPP-S-topk. For each image, we present two tag subsets produced by two methods, and ask the human to judge "which tag subset can tell more useful contents about the image". To avoid the random choice by the annotator, we pick a subset of testing images for the study as follows. According to the computed F_{1-sp} values of the two tag subsets, if both values are larger than 0.2 (*i.e.*, they both are representative of the testing image content), and the absolute difference between two values is larger than 0.15 (*i.e.*, there is enough difference between two results such that the annotator does not need to choose randomly), then this image is picked. We collect 7 judgements from 7 different persons, for each testing image. Then we determine the better subset through majority vote, and set the better one as 1, while the other as 0. Consequently, we obtain a binary vector for the first tag subset produced by method-1 over all testing images. Meanwhile, we compute the evaluation scores of this two subsets, using the semantic metric F_{1-sp} and the conventional metric F_1. Using this two scores, we also obtain two binary vectors of method-1 respectively. Then we compute the consistencies (i.e., $1-$hamming loss) between the binary vector from subject study and the two binary vectors from F_{1-sp} and F_1. The higher consistency (from 0 to 1)indicates the metric is more close to human evaluation.

The subject study results of DPP-S-sampling vs ML-MG on ESP Game are shown in the top sub-table of Table 3. In the case of 3 tags, 437 images are studied. DPP-S-sampling wins at 250 images, while ML-MG wins at 187 images, according to the subject study. We show present the judgment by the standard F_1 score and F_{1-sp}. F_1 is consistent

Data ↓	# tags →	3 tags				5 tags			
ESP Game	metric↓	DPP-S-sampling wins	ML-MG wins	equivalent	consistency	DPP-S-sampling wins	ML-MG wins	equivalent	consistency
	subject study	250	187	0	100%	494	53	0	100%
	conventional F_1	16 / 19	123 / 210	208	31.81%	46 / 49	47 / 394	104	17%
	F_{1-sp}	231 / 351	67 / 86	0	68.19%	341 / 357	37 / 190	0	69.1%
	metric↓	DPP-S-sampling wins	DPP-S-topk wins	equivalent	consistency	DPP-S-sampling wins	DPP-S-topk wins	equivalent	consistency
	subject study	445	47	0	100%	447	82	0	100%
	conventional F_1	40 / 41	37 / 239	212	15.65%	45 / 47	74 / 376	106	22.5%
	F_{1-sp}	324 / 341	30 / 151	0	71.95%	254 / 280	56 / 249	0	58.6%
IAPRTC-12	metric↓	DPP-S-sampling wins	ML-MG wins	equivalent	consistency	DPP-S-sampling wins	ML-MG wins	equivalent	consistency
	subject study	251	91	0	100%	388	116	0	100%
	conventional F_1	15 / 20	52 / 162	160	19.59%	19 / 28	83 / 374	102	20.24%
	F_{1-sp}	193 / 256	28 / 86	0	64.62%	237 / 291	62 / 213	0	59.33%
	metric↓	DPP-S-sampling wins	DPP-S-topk wins	equivalent	consistency	DPP-S-sampling wins	DPP-S-topk wins	equivalent	consistency
	subject study	269	108	0	100%	333	121	0	100%
	conventional F_1	19 / 21	66 / 171	185	22.55%	22 / 28	98 / 339	87	26.43%
	F_{1-sp}	213 / 270	51 / 107	0	70.03%	192 / 234	79 / 220	0	59.69%

Table 3. Subject study on ESP Game and IAPRTC-12, of DPP-S-sampling vs. ML-MG, and DPP-S-sampling vs. DPP-S-topk. The numbers "231 / 351" corresponding to the metric F_{1-sp} and "DPP-S-sampling wins" in the top sub-table mean: according to the score of F_{1-sp}, DPP-S-sampling wins at 351 images, among which DPP-S-sampling also wins at 231 images according to subject study, *i.e.*, the number of consistent judgments between F_{1-sp} and subject study. The consistency 68.19% is computed by $(231+67)/(351+86)$.

with subject study at 139 images, i.e., 31.81% consistency. F_{1-sp} is consistent with subject study at 298 images, i.e., 68.19% consistency. Note that the conventional F_1 judges the DPP-S-sampling tags and the ML-MG tags are equivalent at 208 images, since every tag is treated equally and independently. As long as the numbers of correct tags in two tag subsets are the same, then their F_1 scores will be same. In contrast, as each tag in each semantic path is of different weight when calculating F_{1-sp}, it is less likely to give the same score to two different tag subsets. This subject study tells the semantic metric F_{1-sp} is much closer to human annotation than the standard F_1 score. The results of DPP-S-sampling vs DPP-S-topk on ESP Game are shown in the second sub-table of Table 3. This comparison is more challenging, since the tags between DPP-S-sampling and ML-MG are quite different, while the tags between DPP-S-sampling and DPP-S-topk are more similar, and at many images they are different at the tags within the same semantic path. Even in this case, F_{1-sp} gives much higher consistencies than F_1 in subject study, *i.e.*, 71.95% vs 15.65% at 3 tags and 58.6% vs 22.5% at 5 tags. The subject study results on IAPRTC-12 are shown in Table 3. F_{1-sp} always gives much higher consistency with human performance than F_1 in all cases. Above comparisons tell that the proposed semantic metrics are much more consistent with human annotation than the standard metrics, and that they are suitable for quantitative DIA evaluation. Moreover, in all cases the human annotator judges that DPP-S-sampling wins at many more images than the conventional methods. This validates the good performance of DPP-S-sampling for DIA.

6. Conclusions

This work studied a new task called diverse image annotation (DIA), where an image is annotated using a limited number of tags that attempt to cover as much semantic image information as possible. This task inherently requires that the few retrieved tags are not only representative of the image but also diverse. To this end, we treated the new task as a subset selection problem and model it using a conditional DPP model, which naturally incorporates the representation and diversity jointly. Further, we proposed a modified DPP sampling algorithm, which incorporates semantic paths. We also proposed new metrics based on these semantic paths to evaluate the quality of the diverse tag list. The experiments on two benchmarks demonstrate that our proposed method is superior to those state-of-the-art image annotation approaches. An extensive subject study validates the claim that our proposed semantic metrics are much more consistent with human annotation than traditional metrics.

However, many interesting issues about the new diverse image annotation (DIA) task deserve to be studied in the future. Firstly, the similarity matrix $\mathbf{S}$ in the DPP model is assumed to be pre-computed in this work. That is why the contribution of $\mathbf{S}$ is not very significant, compared with the contribution of semantic paths in sampling. In future work, we plan to learn $\mathbf{S}$ and $\mathbf{W}$ jointly. Secondly, there is till a sizeable gap between the semantic metrics and human evaluation. To bridge this gap, we will focus on updating the way that the semantic paths are constructed and weighted, based on more detailed analysis of the path structure and tag weights. We will make the new semantic metrics available to the community as an online toolkit[3]. Consequently, the evaluation of DIA can be standardized for fair comparison amongst future annotation methods.

Acknowledgements. This work is supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research. Baoyuan Wu is partially supported by Tencent AI Lab. We thank Fabian Caba for his help in conducting the online subject studies.

[3] https://sites.google.com/site/baoyuanwu2015/

References

[1] S. S. Bucak, R. Jin, and A. K. Jain. Multi-label learning with incomplete class assignments. In *CVPR*, pages 2801–2808. IEEE, 2011. 2

[2] R. S. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino. Matrix completion for multi-label image classification. In *NIPS*, pages 190–198, 2011. 2

[3] X. Cao, H. Zhang, X. Guo, S. Liu, and D. Meng. Sled: Semantic label embedding dictionary representation for multi-label image annotation. *IEEE Transactions on Image Processing*, 24(9):2746–2759, 2015. 2

[4] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *BMVC*, 2014. 6

[5] M. Chen, A. Zheng, and K. Weinberger. Fast image tagging. In *ICML*, pages 1274–1282, 2013. 1, 2

[6] X. Chen and C. Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *CVPR*, pages 2422–2431, 2015. 1

[7] X. Chen, X.-T. Yuan, Q. Chen, S. Yan, and T.-S. Chua. Multi-label visual classification with label exclusive context. In *ICCV*, pages 834–841, 2011. 2

[8] C. Fellbaum. *WordNet*. Wiley Online Library, 1998. 4

[9] B. Gong, W.-L. Chao, K. Grauman, and F. Sha. Diverse sequential subset selection for supervised video summarization. In *NIPS*, pages 2069–2077, 2014. 3

[10] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894*, 2013. 2

[11] M. Grubinger, P. Clough, H. Müller, and T. Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International Workshop OntoImage*, pages 13–23, 2006. 1, 4, 6

[12] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*, pages 297–312. Springer, 2014. 1

[13] J. Johnson, L. Ballan, and L. Fei-Fei. Love thy neighbors: Image annotation by exploiting image metadata. In *ICCV*, pages 4624–4632, 2015. 2

[14] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, 2016. 1

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 2

[16] A. Kulesza and B. Taskar. k-dpps: Fixed-size determinantal point processes. In *ICML*, pages 1193–1200, 2011. 3

[17] A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*, 2012. 2, 3, 5

[18] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989. 2

[19] Y. Li, B. Wu, B. Ghanem, Y. Zhao, H. Yao, and Q. Ji. Facial action unit recognition under incomplete data based on multi-label learning with missing labels. *Pattern Recognition*, 60:890–900, 2016. 2

[20] S. Liu, S. Yan, T. Zhang, C. Xu, J. Liu, and H. Lu. Weakly supervised graph propagation towards collective image parsing. *IEEE Transactions on Multimedia*, 14(2):361–373, 2012. 2

[21] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–43, 2014. 3

[22] X. Qian, X.-S. Hua, Y. Y. Tang, and T. Mei. Social image tagging with diverse semantics. *IEEE transactions on cybernetics*, 44(12):2493–2508, 2014. 2

[23] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 2

[24] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1, 1988. 4

[25] A.-M. Tousch, S. Herbin, and J.-Y. Audibert. Semantic hierarchies for image annotation: A survey. *Pattern Recognition*, 45(1):333–345, 2012. 2

[26] L. Von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326. ACM, 2004. 1, 4, 6

[27] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan. Cnn: Single-label to multi-label. *arXiv preprint arXiv:1406.5726*, 2014. 2

[28] J. Weston, S. Bengio, and N. Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning*, 81(1):21–35, 2010. 4

[29] B. Wu, Z. Liu, S. Wang, B.-G. Hu, and Q. Ji. Multi-label learning with missing labels. In *ICPR*, 2014. 2

[30] B. Wu, S. Lyu, and B. Ghanem. Ml-mg: Multi-label learning with missing labels using a mixed graph. In *ICCV*, pages 4157–4165, 2015. 1, 2, 4, 6, 7

[31] B. Wu, S. Lyu, and B. Ghanem. Constrained submodular minimization for missing labels and class imbalance in multi-label learning. In *AAAI*, pages 2229–2236, 2016. 2

[32] B. Wu, S. Lyu, B.-G. Hu, and Q. Ji. Multi-label learning with missing labels for image annotation and facial action unit recognition. *Pattern Recognition*, 48(7):2279–2289, 2015. 2

[33] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *CVPR*, 2016. 1

[34] H.-F. Yu, P. Jain, P. Kar, and I. Dhillon. Large-scale multi-label learning with missing labels. In *ICML*, pages 593–601, 2014. 2, 6, 7

[35] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Summary transfer: Exemplar-based subset selection for video summarizatio. *CVPR*, 2016. 3

[36] M.-L. Zhang and Z.-H. Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2014. 4, 6