MKL based Local Label Diffusion for Automatic Image Annotation

Abhijeet Kumar, Anjali Anil Shenoy, and Avinash Sharma

CVIT, KCIS, IIIT Hyderabad
abhijeet.kumar@research.iiit.ac.in, anjali.shenoy@students.iiit.ac.in,
asharma@iiit.ac.in

Abstract. The task of automatic image annotation attempts to predict a set of semantic labels for an image. Majority of the existing methods discover a common latent space that combines content and semantic image similarity using the metric learning kind of global learning framework. This limits their applicability to large datasets. On the other hand, there are few methods which entirely focus on learning a local latent space for every test image. However, they completely ignore the global structure of the data. In this work, we propose a novel image annotation method which attempts to combine best of both local and global learning methods. We introduce the notion of neighborhood-types based on the hypothesis that similar images in content/feature space should also have overlapping neighborhoods. We also use graph diffusion as a mechanism for label transfer. Experiments on publicly available datasets show promising performance.

1 Introduction

Automatic image annotation is a multi-label prediction problem that attempts to predict a set of semantic labels based on visual content of a given image [39]. It has potential application in image retrieval [18], caption generation [10], image description and classification [29].

The basic assumption in image annotation is that the visual content of an image captures a wide variety of semantics at different levels of granularity. Additionally, the label co-occurrence patterns also model the semantic similarity between images. Therefore, existing methods have tried to model label-to-label [18], image-to-image [9] and image-to-label [4] similarities or a combination of them [17, 30].

In context of image-to-image and image-to-label similarities Nearest Neighbor (NN) based approaches have been largely successful and intuitive for image annotation. Recent methods either employ a global (metric) learning technique [9,28] or a local query specific model [14] for addressing the class-imbalance problem. However, while the former suffers from the problem of scalability due to global metric learning bottleneck, the latter fails to capture the global latent structure of the data as it is too focused on query specific neighborhood structure. An alternate approach in [22] addresses the class-imbalance by performing

scale-dependent label diffusion on global hypergraph in a transductive setup. However, their method also suffers from the scalability issue (due to SVD decomposition of large dense matrices). Many recent deep learning methods also propose to learn end-to-end network for solving image annotation task [21,32,37].

In this paper, we focus on bridging the gap between purely global and local modeling of the image annotation task. The key hypothesis is that similar images in feature space also have similar labels, hence two vicinal images in feature space should also have overlapping neighborhoods. Each of these neighborhoods (corresponding to an image) can be statistically characterized by constructing the label histogram of all their associated images. We refer to these label histogram features as Local Label Distribution (LLD) features. Thus, two similar images should have similar LLD feature representation which represent similarity in neighborhood. Hence we propose to learn a local label-transfer model for each such neighborhood-type (cluster) separately. This characterization of images by neighborhood-types also inherently captures the global latent structure of data.

Subsequently, each local model is formulated as Multiple Kernel Learning (MKL) task, using a family of multi-scale diffusion kernels. The MKL formulation minimizes the sum of squared error between the ground truth labels (known for each training image) and the labels predicted with multi-scale diffusion over the associated local graph. Such diffusion is performed by linearly combining a set of scale dependent diffusion kernels. A closed form solution exists for obtaining the optimal kernel combination coefficients (parameters of local model). Thus, MKL parameters per neighborhood-type are learnt over the training data. At test time, we construct and map the neighborhood structure of each query image to an existing neighborhood-type to retrieve the best parameter of local model. Finally, we construct the local graph for this query image and diffuse the label using these parameters for subsequent prediction.

1.1 Our Contributions

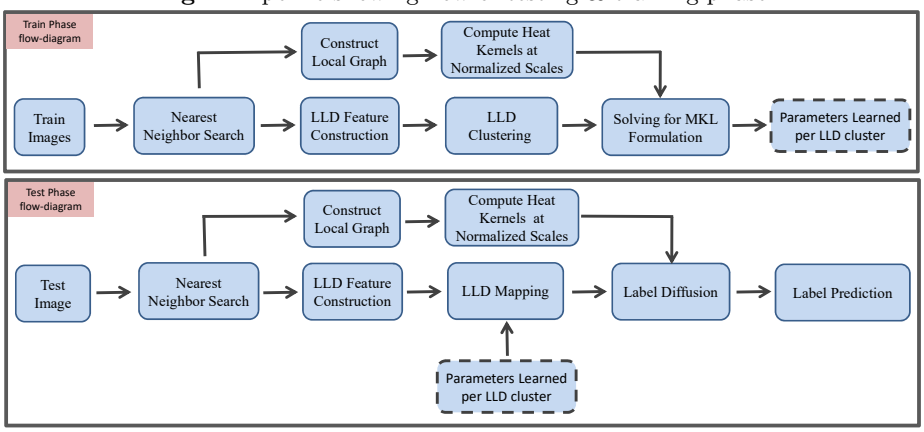
- We propose a new label histogram characterization (LLD features) of the image neighborhood enabling us to discover the neighborhood-types in the dataset.
- We propose a MKL formulation as local learning model and derived a closed form solution for obtaining the model parameters.
- We propose a diffusion scale normalization procedure for effectively combining diffusion over multiple graphs.

2 Literature Survey

2.1 Generative, Discriminative and Hybrid Models

Xinag et al. [34] proposed a Markov Random Field model, which captured many previously proposed generative models, but had an expensive training step as it learnt an MRF per label.

Fig. 1. Pipeline showing flow of testing & training phase



Discriminative models were proposed in [7, 8, 27, 35]. These methods learn label-specific models to classify an image belonging to the particular label. However they fail to capture label-to-label correlations. A hybrid model was presented in [20] combining a generative [6] and discriminative (SVM) model aimed at improving the number of labels recalled.

2.2 Nearest Neighbor Approaches

Though simple and highly intuitive, NN methods are among the best performing ones. [9] introduced metric learning to fuse an array of low-level features. They used cross entropy loss in addition to weighted (based on distance or rank) label propagation. Recently, [28] defined a Bayesian approach with two pass kNN for addressing the class-imbalance challenge and subsequently used an extension of existing LMNN approach [33] as metric learning to fuse different feature sets. One major limitation of these approaches is that they are global methods and heavily rely on metric learning construct, which makes it difficult to scale them to large datasets.

Alternatively, a local variant of NN method proposed in [14] performs the Non-negative Matrix Factorization (NMF [15]) of features from images in a smaller neighborhood, which are made to follow a consensus constraint. Here, the class-imbalance is dealt by means of weighting different feature matrices. However, this is purely a local approach and hence fails to capture the global structure of the latent space.

Recently, [25] and [28] report performance improvement over the NN methods by using cross modal embedding such as Canonical Correlation Analysis (CCA) or Kernel Canonical Correlation Analysis (KCCA). This is again an attempt to learn global common latent space. However selection of an appropriate kernel

function and scalability of KCCA poses a major challenge in these methods.

2.3 Graph Based Models

A graph based transductive method for explicitly capturing label-to-label and image-to-image similarities was proposed in [30]. Recently, [22] proposed a hypergraph diffusion based transductive method and exploited multi-scale diffusion to address the class-imbalance problem. However, both these methods are semi-supervised in the sense that they need access to all test data for prediction. Additionally the hypergraph diffusion method is not scalable due to requirement of storage and computation of eigen-decomposition of very large matrices.

2.4 Deep Learning Based Methods

Inspired by the recent success of deep neural net architectures in image classification [23,24], different approaches involving deep nets have been tried for multi-label classification [11,13,19,32]. [11] modeled these relationship on a hierarchical model exploiting Long Short Term Memory (LSTM) by incorporating inter-level and intra-level label correlations which were parsed using WordNet. CNN-RNN [32] learns a joint image-to-label embedding and label co-occurrence model in an end-to-end way. Semantically Regularised CNN-RNN (S-CNN-RNN) [37] improves on the CNN-RNN model by using a semantically regularized embedding layer as an interface between the CNN and RNN which enables RNN to capture the relational model alone. [13] proposed exploiting image metadata to generate neighbors of an image and blend visual information using neural-nets. Recent works in Deep nets capture label-to-label relationships more explicitly than before. Another recent work in [21] converted labels to a word2vec vector [19] and performed label transfer using nearest neighbor methods in embedding space computed using CCA or KCCA.

Recently deep-learning methods have been introduced in the context of graphs [31, 38, 40, 42]. Gated Graph Neural Network (GGNN) [40] is a LSTM variant for graphs which learns a propagation model that transfers information between nodes depending on the edge types. [41] introduced Graph Search Neural Network (GSNN) which improves GGNN [40] by diminishing the computational issues. GSNN is able to reason about the concepts by capturing the information flow between nodes in the noisy knowledge-graphs. GSNN is different from our model in the propagation modeling in graphs. We explicit model the label propagation with diffusion framework on graphs constructed from neighboring images while GSNN learns the network propagation parameters in the knowledge-graph.

3 Proposed Approach

This section will provide a detailed description of each individual module in the proposed train and test phase flow pipeline depicted in Figure 1. Let $\mathbf{X}^{tr} =$

$\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ be the feature-vector representation of training set of images with corresponding known ground truth labels $\mathbf{Z}^{tr} = \{\mathbf{z}_1, \cdots, \mathbf{z}_n\}$, where each $\mathbf{z}_i$ is a binary vector of size l denoting presence/absence of labels in the image $\mathbf{x}_i \in \mathbf{X}^{tr}$.

3.1 Nearest Neighbors Search

This module performs feature space NN search for a given image in order to discover a group of similar images. Instead of performing a global search, we opt for quantizing the feature space image representation by employing the paralleizable k-means clustering algorithm [1] on training data and finding a fixed g number of clusters in an offline manner. From all these clusters, we subsequently find the top η closest cluster centers in feature space, then we perform the local NN search in those clusters by retrieving a fixed δ number of similar images from each of them. All such retrieved images form the neighbourhood of a given image. We denote this set of nearest neighbor images obtained with this method for a given image $\mathbf{x}$ as $\mathcal{N}(\mathbf{x})$. Note that $|\mathcal{N}(\mathbf{x})| \leq \eta\delta$, as a cluster can have less than δ images. This naturally provides diversity and scalability over the exhaustive NN search.

3.2 Local Graph Construction

This module constructs an undirected weighted graph $\mathcal{G}(V, E, \mathcal{W})$ for an image $\mathbf{x}$ using the neighborhood $\mathcal{N}(\mathbf{x})$, where the input image and each of the selected images in $\mathcal{N}(\mathbf{x})$ images become nodes of the graph.

We construct a local graph by connecting each node to its k most similar nodes using an inverse euclidean distance similarity function over the corresponding image features. We use the standard Gaussian kernel over the feature space as the similarity function, i.e., $Sim(x_1, x_2) = \exp(-||x_1, x_2||_{l2}/\sigma^2)$ to assign weights to these edges.

Note that here $|V| = \zeta + 1 \leq \eta\delta + 1$ where $|\mathcal{N}(\mathbf{x})| = \zeta$

3.3 LLD Feature Construction, Clustering and Mapping

This module first constructs an l-dimensional histogram feature $\mathcal{F}(\mathbf{x})$ for each image representing the LLD of a given image $\mathbf{x}$, by taking the sum of ground truth labels of all the samples in $\mathcal{N}(\mathbf{x})$:

$$\mathcal{F}(\mathbf{x}) = \sum_{\forall \mathbf{x_i} \in \mathcal{N}(\mathbf{x})} \mathbf{z}_i. \tag{1}$$

We employ parallelizable k-means clustering algorithm over the LLD features corresponding to all training images to get a fixed number of cluster-centers (c) which act as representatives of the neighborhood-types. To map an image to a neighborhood-type we just need to compute its closest neighborhood-type (cluster-centers) in this l-dimensional space. As discussed in Section 1, we discover the clusters in LLD space and learn a local model per cluster.

[1] https://github.com/serban/kmeans

3.4 Diffusion Kernel

In this module we construct a family of diffusion kernels at different diffusion scales for a given local graph computed in the previous module. In each weighted graph $\mathcal{G}(V, E, \mathcal{W})$, training images with labels acts as heat sources that are diffused/propagated to all the other nodes in the graph where we aggregate the information and subsequently use for prediction.

Graph Laplacian. For (dyadic) undirected weighted graphs, diffusion kernels are derived from the spectra (constituted by both eigenvalues & eigenvectors) of the graph Laplacian matrix [1]. The unnormalized Laplacian L of a weighted undirected graph $\mathcal{G}$ with adjacency matrix A is defined as:

$$L = D - A = U \Lambda U^T \tag{2}$$

where D is the diagonal matrix with each diagonal entry $d_{ii} = \sum_{i=1}^{n} A_{ij}$, U is the matrix of eigenvectors and Λ is the diagonal matrix of corresponding real positive eigenvalues of the Laplacian matrix, i.e. $\Lambda = Diag(\lambda_1, \cdots, \lambda_{\zeta+1})$ and $0 = \lambda_1 \leq \lambda_2 \cdots \leq \lambda_{\zeta+1}$ for a connected graph.

Diffusion kernel is a positive semi-definite, non-linear family of kernels and can be used for defining distances over non-euclidean spaces and for multiscale-multilabel-diffusion in graphs. It has been used for multi-scale label diffusion over graphs [36], and a variety of other applications 3D Shape Matching [3] and Robotics graphSLAM [2]

Subsequently, the scale dependent diffusion kernel matrix is defined as:

$$H(t) = U e^{-\Lambda t} U^T \tag{3}$$

where $t > 0$ is the parameter of the diffusion. Every entry $H(i, j, t)$ of the diffusion kernel can be interpreted as the amount of heat diffused from node v_j to node v_i at scale t while considering v_j as a point heat source of unit magnitude. We use the diffusion kernel matrix at m different scales to diffuse each label.

This is different from [22] as we use multiple scales for all labels rather than using specific scales for different frequency type of labels.

Diffusion Scale Normalization. Instead of manually choosing diffusion scales for all local graphs, we propose to find a set of normalized scales per graph using the structure of the local graph. This structure is captured by the spectrum (eigenvalues) of the graph. Such a normalization is very important as the diffusion scale value is relative to the graph structure/topology. Interestingly, if we see the plot of exponential function

$$f(\lambda, t) = e^{-\lambda t} = \theta \tag{4}$$

in Figure 2, we see that one can find the normalized values of the diffusion scale parameter t (varying from smaller to larger values) by fixing the values of θ and index of λ.

$$t = -log(\theta)/\lambda \tag{5}$$

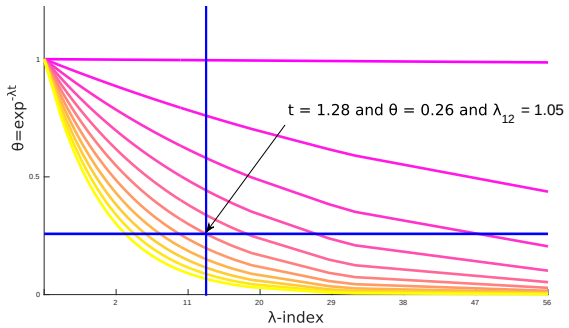


Fig. 2. Diffusion scale normalization on a sample graph of 56 nodes. In this case we use the 12th ($= 0.2 * |56|$) eigenvalue and $\theta = 0.26$ for computing the scale normalized t value. In general a set of θ values will be chosen for defining bank of diffusion kernels

3.5 MKL Formulation for Label Diffusion

In this section, we outline our MKL formulation for learning the label diffusion parameters locally for each LLD cluster/neighborhood type.

Let $\mathbf{X}_c^{tr} \subset \mathbf{X}^{tr}$ be the subset of n_c training images and $\mathbf{Z}_c^{tr} \subset \mathbf{Z}^{tr}$ be the ground truth labels in c-th LLD cluster. We can write $\mathbf{X}_c^{tr} = [\mathbf{x}_1, \cdots, \mathbf{x}_{n_c}]$ and $\mathbf{Z}_c^{tr} = [\mathbf{z}_1, \cdots, \mathbf{z}_{n_c}]$.

For each image $\mathbf{x}_k \in X_{tr}^c$ there is a local graph $\mathcal{G}_k$ (Section 3.2) with $\zeta_k + 1$ nodes where the node with the last index is $\mathbf{x}_k$ itself. Let $\mathbf{z}_k$ be the ground truth label of $\mathbf{x}_k$ and $\mathbf{Y}_k = [\mathbf{y}_1 \cdots \mathbf{y}_l]$ be the transpose matrix of ground truth labels (i.e., $[\mathbf{z}_1 \cdots \mathbf{z}_k]^T$) for all the other ζ_k images (nodes) in the local graph appended with a $\mathbf{0}$ row vector representing labels for $\mathbf{x_k}$ itself. Note that $\mathbf{y}_i$ is a column vector of dimension $(\zeta_k + 1) \times 1$. The last row is appended for compatibility in matrix multiplication.

Let $\tau = [t_1, \cdots, t_m]^T$ be the set of m normalized diffusion scale parameters used for defining the diffusion kernels: $H_k(t_1), \cdots, H_k(t_m)$ (Section 3.4). Here each $H_k(t_i)$ is a $(\zeta_k + 1) \times (\zeta_k + 1)$ dimensional matrix. Let $\mathbf{e} = [0, \cdots, 0, 1]$ be a $(\zeta_k + 1) \times 1$ dimensional vector, then $h_k^i = e^T H_k(t_i)$ represents the last row of the (symmetric) diffusion kernel matrix.

It is important to note that since the training image $\mathbf{x}_k$ is kept at the last position in the index order in graph $\mathcal{G}_k$, only the last row of $H_k(t_i)$ (i.e., h_k^i) is sufficient to perform label diffusion (at scale t_i) from all other images (nodes) in the graph.

Since we know the ground truth labels for image $\mathbf{x}_k$, and if we take β_j^c to represent the diffusion contributions of the diffusion kernels $h_k^i \forall i \in \{1, \cdots, m\}$ for j^{th} label in the c^{th} cluster, we can obtain an MKL formulation as a minimization criterion:

$$\min_{\beta_c} \sum_{k=1}^{n_c} ||\mathbf{\Gamma}_k \beta_c - \mathbf{z}_k||^2, \tag{6}$$

where,

$$\beta_c^j = [\beta_c^{j1}, \beta_c^{j2}, ..\beta_c^{jm}]_{(1\times m)} \tag{7}$$

$$\beta_c = [\beta_c^1, \beta_c^2, ..\beta_c^l]_{(lm\times 1)}^T \tag{8}$$

$$\mathit{\Gamma}_k = \begin{bmatrix} (\mathbf{M}_k \mathbf{y}_1)^T & & & \\ & \ddots & & \\ & & (\mathbf{M}_k \mathbf{y}_r)^T & \\ & & & \ddots \\ & & & & (\mathbf{M}_k \mathbf{y}_l)^T \end{bmatrix}_{(l\times lm)} \tag{9}$$

and

$$\mathbf{M}_k = [{h_k^1}^T, {h_k^2}^T, \ldots, {h_k^m}^T]_{(m\times(\zeta_k+1))}^T. \tag{10}$$

By combining Eq. 6, 8, 9 with simple algebraic manipulations, we can write the simplified MKL formulation as:

$$\min_{\beta_c} ||\hat{\mathbf{\Gamma}}_c \beta_c - \mathcal{Z}_c||^2, \tag{11}$$

where,

$$\hat{\mathbf{\Gamma}}_c = [\mathbf{\Gamma}_1^T, \cdots, \mathbf{\Gamma}_{n_c}^T]^T \tag{12}$$

is a $(n_c l \times lm)$ matrix and

$$\mathcal{Z}_c = [z_1^T, \cdots, z_{n_c}^T]^T \tag{13}$$

is a $(n_c l \times 1)$ vector.

The minimization of proposed MKL formulation Eq. 11 can be achieved by finding the optimal value of parameter β_c in a closed form manner as:

$$\beta_c = (\hat{\mathbf{\Gamma}}_c^T \hat{\mathbf{\Gamma}}_c + \epsilon I)^{-1} \hat{\mathbf{\Gamma}}_c^T \mathcal{Z}_c, \tag{14}$$

It is important to note that the $\hat{\mathbf{\Gamma}}_c$ is a very sparse and low rank matrix. This sparse structure can be useful in efficiently computing its singular value decomposition and hence the pseudoinverse of the $\hat{\mathbf{\Gamma}}_\mathbf{c}$ which is relatively inexpensive as KCCA or hypergraph Laplacian eigenvector-decomposition.

Since our method solves MKL at cluster level in a closed form manner, it is scalable to large data.

3.6 Label Diffusion & Prediction

For a test image $\mathbf{x_q}$, $\mathcal{F}(x_q)$ is used to find s nearest LLD neighborhood-types represented as $\mathbf{P} = [\mathbf{p}_1, \cdots, \mathbf{p}_s]$. The associated pre-learned MKL parameters $[\beta_1, \cdots, \beta_s]$ for the selected clusters are used for independent diffusion over the local graph $\mathcal{G}_q$ on $\mathbf{x}_q$. This is achieved by first computing the respective $\mathbf{\Gamma}_q$ matrix sized $(l \times lm)$ and then multiplying it with the pre-learned parameter vector β_c sized $(lm \times 1)$. Finally, we take an average of these diffused values. Label diffusion is performed as:

$$\mathbf{z}_q^{\text{diffused}} = \frac{1}{s} \sum_{c=1}^{s} \mathbf{\Gamma}_q \beta_c \tag{15}$$

where $\mathbf{z}_q^{\text{diffused}}$ is the $(l \times 1)$ vector of the diffused labels. Finally, a fixed set of r labels is predicted for $\mathbf{x}_q$ by choosing labels corresponding to top r values in $\mathbf{z}_q^{\text{diffused}}$.

4 Datasets

Table 1 provides details of datasets used in our experiments.

Table 1. Dataset details: Row 2-4 contain basic dataset information, Row 5-6 denote the statics in the order- median, mean and max.

Dataset Information	PascalVOC-2007 [5]	MIRFlickr-25k [12]
Total Number of Images	9963	25000
Vocabulary Size	20	38
Train/Test Split	5011/4092	12500/12500 [25]
Labels/Images	1, 1.51, 6	5, 4.7, 17
Images/Labels	5, 4.7, 17	995.5, 1560, 5216

4.1 PascalVOC-2007

PASCAL Visual Object Classes (VOC) challenge [5] datasets are widely used as the benchmark for multi-label classification. The VOC 2007 dataset contains 9963 images. We follow a train/test split of 5011/4952 as in [32].

4.2 MIRFlickr-25k

This dataset contains images downloaded from Flickr and was introduced in [12] for evaluating keyword-based image retrieval. It consists of 25000 images and we follow an equal split of train and test (12500 images each) as used in [25]. 419 images in the dataset do not have any of the 38 semantic label annotations. Metadata, GPS and EXIF information are also provided in the dataset but we do not use any of these in our method.

5 Experiments & Results

5.1 Features and Evaluation Method

Deep-learning based features have proven to be effective in image representation [22, 28] and hence we use outputs from *fc7-layer* of VGG16 network (pre-trained on ImageNet) [23] to represent an image. To analyze the annotation performance, we consider precision, recall, F1-score, average precision (AP) and mean average precision (mAP). We predict a fixed number of r labels per image which is set to be the mean number of labels per image in the dataset. Let a label w_i be present in $m1$ images as ground-truth and is predicted for $m2$ images where $m3$ of them are correct. The precision for w_i is $m3/m2$ and recall is $m3/m1$. Mean precision (P) and recall (R) is the precision and recall values averaged over all the labels. $F1$ measure is the harmonic mean of P and R. We also report AP and mAP by evaluating ranking of all the images.

5.2 Experiments

We set $s = 3$ for PascalVOC-2007 and $s = 5$ for MIRFlickr-25K. Additionally we set $k = 6$ in kNN graph construction in section 3.2 and $m = 100$. θ is chosen as m equally spaced values between 0.001 to 1.0 (corresponding t will vary from large to small scales of diffusion) and the index of the eigenvalue is chosen as the closest integer value greater than $0.2 \times |V|$. Performance variation observed for varying m from 32 to 100 was less than 1%.

We find the values of the hyper-parameters via cross-validation by dividing the train dataset into two parts (5 : 1 ratio) while maximizing $F1$ and the best performing parameters on validation set were used to evaluate performance on the test data. For η and δ we explore from the following set $\{5, 12, 20, 28\}$ to find the best performing values. We also vary the number of cluster centers in LLD (c) and the number of clusters in image-feature space clustering (g). The best performing values for MIRFlickr-25k were found to be $\eta = 5$, $\delta = 12$, $c = 100$ and $g = 30$ and for PascalVOC-2007 were found to be $\eta = 5$, $\delta = 20$, $c = 45$ and $g = 20$.

5.3 Results

Table 2 shows the performance comparison of the proposed method with existing methods that uses VGG16 features. The obvious understanding one can make here is that there is non-agreement between F1 and mAP measures. The mAP considers the global ranking of all images corresponding to each label instead of just considering top r labels for computation of average precision.

We can see that our method performs very close to the state of the art 2PKNN method and also has similar mAP. This small disparity in performance can be attributed to the fact that our method does not consider KCCA and metric learning type of fully global operations.

Table 2. Comparison of popular methods on different evaluation metrics for MIRFlickr-25k Dataset for $\mathbf{r = 5}$

Method	$P@r$	$R@r$	$F1@r$	mAP
TagRel [16]	41.5	72.1	52.7	68.9
TagProp [18]	45.5	70.1	55.2	70.8
2PKNN [28]	46.4	70.9	56.1	66.5
SVM [26]	38.8	72.4	50.5	72.7
HHD [22]	-	-	-	75.0
Our Method	51.0	59.9	55.1	66.3

Table 3. Label specific (Average Precision in %) for all labels, mAP, P@r, R@r and F1@r with $\mathbf{r = 2}$ on PascalVOC-2007 dataset.

	CNN-RNN [32]	Our Method
plane	96.7	92.8
bike	83.1	84.7
bird	94.2	91.3
boat	92.8	81.7
bottle	61.2	41.3
bus	82.1	83.9
car	89.1	89.0
cat	94.2	86.3
chair	64.2	55.7
table	70.0	68.4

	CNN-RNN [32]	Our Method
cow	83.6	71.9
dog	92.4	86.7
horse	91.7	89.4
motor	84.2	82.7
person	93.7	91.8
plant	59.8	54.0
sheep	93.2	75.1
sofa	75.3	57.1
train	99.7	92.6
tv	78.6	66.9

	CNN-RNN [32]	Our Method
P@r	-	53.8
R@r	-	77.7
F1@r	-	63.6
mAP	84.0	77.2

Fig. 3. Distribution of label frequency in MIRFlickr-25k and PascalVOC-2007 test images.

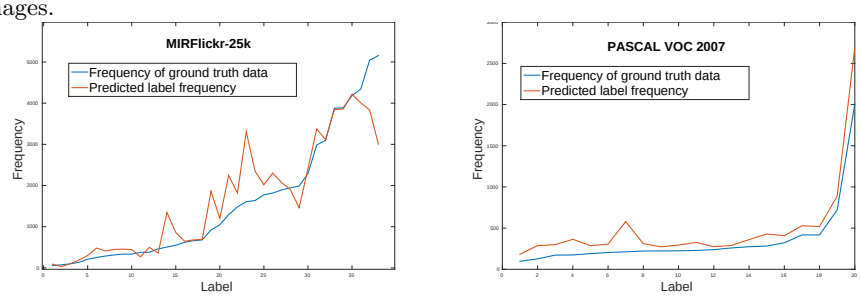


Figure 3 shows the distribution of both ground-truth and predicted label frequency in test images on two datasets. For MIRFlickr-25k, a large section of predicted label frequency curve (including low and high frequency labels) closely overlaps with that of ground truth. However, the medium frequency labels (in the middle) are over-predicted at the cost of suppression of few frequent occurring labels (right tail). This is prevalent due to the nature of our approach, but we accept this as a trade-off in order to address the issue of class imbalance by accurately predicting lower frequency labels in images. In case of PascalVOC-2007, we observe our curve running parallel to the one of ground truth but with a shift of a few units. This shift is due to the difference that on an average only 1.5 labels are associated with image in the ground truth annotation but we predict 2 labels per image.

Figure 4 shows the qualitative results on a few examples from the MIRFlickr-25k dataset. Row1 depicts images which consists of frequent labels in the ground-truth(*male*, *people*) while Row2 consists of images with no ground-truth and Row3 contains images with rare labels (*baby*, *portrait*) in the ground-truth. From Row1 and Row3 we observe that our method performs well on the frequent and rare labels. We also observe that for images with no ground-truth (Row2) the predictions are semantically relevant to the image-content. Labels in red-color denote tags which are not present in the ground-truth of the image but are semantically meaningful. This indicates towards the incomplete-labeling in the dataset.

6 Conclusion

We have proposed a novel solution for automatic multi-label image annotation. Our method exploits the empirical observation that similar images have similar neighborhood-types in terms of their label distribution. We have introduced the notion of neighborhood-type and proposed to learn local MKL models per cluster/neighborhood-type with closed form learning solution. The MKL formulation exploits the multi-scale diffusion where we also proposed a novel diffusion scale normalization to be able to combine diffusion at different local graphs. The overall formulation is scalable as we have mainly proposed local models while clustering is employed twice (in original VGG16 feature space and as well as in histogram/LLD space). Finally, we have shown promising results on publicly available dataset.

As part of future work it will be interesting to explore the hypergraphs in the local space to model the higher order correlations between labels explicitly in conjunction with image correlations. Local metric learning construct in forming and/or manipulating LLD clusters can be used which may provide insights into the inherent nature of the problem.

7 Acknowledgments

We thank Yashashwi Verma for his helpful feedback.

Fig. 4. Qualitative results for label prediction on MIRFlickr-25k dataset. Row1: images with frequent labels(*male*, *people*); Row2: images with no ground truth label given; Row3: images with rare labels(*baby*, *potrait*). Green: Labels present in ground-truth and our model's predictions (true positives) , Blue: Incorrect predictions by the model and Red: Labels are not present in ground truth but are semantically meaningful. Red and blue labels combined form False Positives

References

1. Mikhail Belkin and Partha Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. Neural Computing 15, 6 (2003), 1373 – 1396.
2. Sayantan Datta, Siddharth Tourani, Avinash Sharma, K. Madhav Krishna. SLAM Pose-graph Robustification via Multi-scale Heat-Kernel Analysis. In CDC (2016)
3. Avinash Sharma, Radu Horaud, Jan Cech, Edmond Boyer. Topologically-robust 3d shape matching based on diffusion geometry and seed growing. In CVPR (2011)
4. Gustavo Carneiro, Antoni B Chan, Pedro J Moreno, and Nuno Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. IEEE TPAMI 29, 3 (2007), 394 – 410
5. Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. IJCV 88 (2010), 303 – 338
6. Shaolei Feng, R. Manmatha, and Victor Lavrenko. Multiple Bernoulli Relevance Models for Image and Video Annotation. In CVPR (2004).
7. Hao Fu, Qian Zhang, and Guoping Qiu. Random Forest for Image Annotation. In ECCV (2012)
8. David Grangier and Samy Bengio. A Discriminative Kernel-Based Approach to Rank Images from Text Queries. IEEE TPAMI 30 (2008), 1371 – 1384.
9. Matthieu Guillaumin, Thomas Mensink, Jakob J. Verbeek, and Cordelia Schmid. TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation. In ICCV (2009).
10. Ankush Gupta, Yashaswi Verma, and C. V. Jawahar. Choosing linguistics over vision to describe images. In AAAI (2012)
11. Hexiang Hu, Guang-Tong Zhou, Zhiwei Deng, Zicheng Liao, and Greg Mori. Learning Structured Inference Neural Networks with Label Relations. CoRR abs/1511.05616 (2015).
12. Mark J. Huiskes and Michael S. Lew. The MIR flickr retrieval evaluation. In Multimedia Information Retrieval (2008)
13. Justin Johnson, Lamberto Ballan, and Li Fei-Fei. Love Thy Neighbors:Image Annotation by Exploiting Image Metadata. CoRR abs/1508.07647 (2015).
14. Mahdi Kalayeh, Haroon Idrees, and Mubarak Shah. NMF-KNN: Image Annotation using Weighted Multi-view Non-negative Matrix Factorization. In CVPR (2014).
15. Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In NIPS (2000).
16. Xirong Li, Cees G. M. Snoek, and Marcel Worring. Learning Social Tag Relevance by Neighbor Voting. IEEE Transactions on Multimedia 11 (2009), 1310 – 1322.
17. Jing Liu, Mingjing Li, Qingshan Liu, Hanqing Lu, and Songde Ma. Image annotation via graph learning. Pattern Recognition 42 (2009), 218 – 228.
18. Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. Baselines for image annotation. In IJCV (2010)
19. Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. CoRR abs/1301.3781 (2013)
20. Venkatesh N. Murthy, Ethem F. Can, and R. Manmatha. A Hybrid Model for Automatic Image Annotation. In ICMR. (2014)
21. Venkatesh N. Murthy, Subhransu Maji, and R. Manmatha. Automatic Image Annotation using Deep Learning Representations. In ICMR (2015)
22. Venkatesh N. Murthy, Avinash Sharma, Visesh Chari, and R. Manmatha. Image Annotation using Multi-scale Hypergraph Heat Diffusion Framework. In ICMR (2016).

23. Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR abs/1409.1556 (2014).
24. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. CoRR abs/1409.4842 (2015).
25. Tiberio Uricchio, Lamberto Ballan, Lorenzo Seidenari, and Alberto Del Bimbo. Automatic Image Annotation via Label Transfer in the Semantic Space. CoRR abs/1605.04770 (2016)
26. Jakob Verbeek, Matthieu Guillaumin, Thomas Mensink, and Cordelia Schmid. Image Annotation with TagProp on the MIRFLICKR Set. In ACM MIR (2010)
27. Yashaswi Verma and C. V. Jawahar. Exploring SVM for Image Annotation in Presence of Confusing Labels. In BMVC (2013).
28. Yashaswi Verma and C. V. Jawahar. Image Annotation by Propagating Labels from Semantic Neighbourhoods. IJCV (2016), 1 – 23
29. Hua Wang, Heng Huang, and Chris H. Q. Ding. Image annotation using multi-label correlated Greens function. In ICCV (2009)
30. Hua Wang, Heng Huang, and Chris H. Q. Ding. Image annotation using bi-relational graph of images and semantic labels. In CVPR (2011).
31. D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams. Convolutional networks on graphs for learning molecular fingerprints. NIPS(2015).
32. Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu 2016. CNN-RNN: A Unified Framework for Multi-label Image Classification. CoRR abs/1604.04573 (2016)
33. Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. JMLR 10, Feb (2009), 207 – 244
34. Yu Xiang, Xiangdong Zhou, Tat-Seng Chua, and Chong-Wah Ngo. A revisit of Generative Model for Automatic Image Annotation using Markov Random Fields. In CVPR (2009)
35. Hao Zhang, Alexander C. Berg, Michael Maire, and Jitendra Malik. SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition. In CVPR (2006).
36. Arthur D. Szlam and Mauro Maggioni and Ronald R. Coifman. Regularization on Graphs with Function-adapted Diffusion Processes. JMLR 9 (2008) 1711 – 1739
37. Feng Liu, Tao Xiang, Timothy M Hospedales, Wankou Yang and Changyin Sun. Semantic Regularisation for Recurrent Image Annotation. In CVPR (2017).
38. F. Scarselli, M. Gori, A. C. Tsoi, and G. Monfardini. The graph neural network model. IEEE Transactions on Neural Networks (2009)
39. Xirong Li, Tiberio Uricchio, Lamberto Ballan, Marco Bertini, Cees G. M. Snoek, Alberto Del Bimbo. Socializing the semantic gap: A comparative survey on image tag assignment, refinement and retrieval. CSUR 49, 1 (2016)
40. Y. Li and R. Zemel. Gated graph sequence neural networks. ICLR (2016).
41. Marino, Kenneth and Salakhutdinov, Ruslan and Gupta, Abhinav. The More You Know: Using Knowledge Graphs for Image Classification. CVPR (2017).
42. M. Henaff, J. Bruna, and Y. LeCun. Deep convolutional networks on graph-structured data. arXiv preprint arXiv:1506.05163 (2015).