# A Hybrid Model for Automatic Image Annotation

Venkatesh N. Murthy
School of Computer Science
University of Massachusetts
Amherst, MA, USA
venk@cs.umass.edu

Ethem F. Can
School of Computer Science
University of Massachusetts
Amherst, MA, USA
efcan@cs.umass.edu

R Manmatha
School of Computer Science
University of Massachusetts
Amherst, MA, USA
manmatha@cs.umass.edu

## ABSTRACT

In this work, we present a hybrid model (SVM-DMBRM) combining a generative and a discriminative model for the image annotation task. A support vector machine (SVM) is used as the discriminative model and a Discrete Multiple Bernoulli Relevance Model (DMBRM) is used as the generative model. The idea of combining both the models is to take advantage of the distinct capabilities of each model. The SVM tries to address the problem of poor annotation (images are not annotated with all relevant keywords), while the DMBRM model tries to address the problem of data imbalance (large variations in number of positive samples). Since DMBRM does not work well with high-dimensional data, a Latent Dirichlet Allocation (LDA) model is used to reduce the dimensionality of vector quantized features before using it. The hybrid model's results are comparable to or better than the state-of-the-art results on three standard datasets: Corel-5k, ESP-Game and IAPRTC-12.

## Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models; I.2.10 [**Computing Methodologies**]: Artificial Intelligence-Vision and Scene Understanding

## General Terms

Algorithms, Experimentation

## Keywords

Image Annotation, SVM, Discrete MBRM, LDA

## 1. INTRODUCTION

There is a huge demand for automatic image/video annotation with increasing numbers of images and videos both in personal collections and on the Internet. For example, 100 hours of video are uploaded to YouTube everyday and on an average people upload 350 million photos to Facebook per day. One approach to retrieve or manage such large quantities of images/videos is to automatically annotate each test image with multiple keywords by training a statistical model on a labeled training set. Researchers have tried to address this problem by either using a discriminative model [14, 1] or a generative model [9, 5, 7]. Each of these techniques has its own advantages and disadvantages. In our model, we try to make use of both in order to gain maximum benefit.

In this work, we propose to automatically annotate the images using a hybrid discriminative/generative model. The proposed model has the advantage of learning both in a discriminative as well as a generative manner using a set of annotated training images. Different sets of global and local features are extracted for an image. We build models for each feature and later combine them appropriately. In the case of the Discrete MBRM model, we learn the joint probability of words and dimensionality reduced features similar to the relevance model. For a test image, this model can be used for assigning probability scores for words which best describes the image. Discretized features are modeled using the multinomial distribution and the words are modeled using a multiple Bernoulli distribution. In the case of SVM, a one-against-all model is built per word. Given a test image, we evaluate it against all the word models obtaining the corresponding probability scores. Finally, we fuse the probability scores appropriately from both models and assign the test image with the highest scoring probability words. We provide experimental results on three standard datasets, Corel-5k [3], IAPRTC-12 [11], ESP-Game [16] and show that we get better than state of the art results.

The rest of the paper is organized as follows, section 2 provides some related work , we present our proposed model in section 3 and discuss the experimental setup in section 4. In section 5, results and discussion are presented. Finally, with some conclusions in section 6.

## 2. RELATED WORK

Among all the proposed models, nearest neighbor and generative models are shown to be most successful. Early examples of these are relevance models - the Cross Media Relevance Model (CMRM) [7], the Continuous Relevance Model (CRM) [9] and the Multiple Bernoulli Model (MBRM) [5]. More recently, results have substantially improved by combining metric learning and nearest neighbors. Xiang et al. [18] focused on an approach based on Multiple Markov Random Fields (MRF) for semantic context modeling and learning in the context of automatic image annotation. Zhang et al. [20] proposed a regularization based


*ICMR* '14, Apr 01-04 2014, Glasgow, United Kingdom

feature selection algorithm to leverage both sparsity and clustering properties of features and incorporate it into the image annotation task. Nakayama et al. [12] focused on a distance called canonical contextual distance (CCD) and applied it to image annotation task. Feng et al. [4] and Llorente et al. [10] formulated the image annotation problem as a multi label ranking problem. Wu et al. [17] addressed the problem of class imbalance and weak labeling problem by a tag-completion technique for the training dataset based on some optimization criteria.

Makadia et al. [11] provided the baseline for image annotation based on the nearest neighbor. Later, results were improved by TagProp - a weighted nearest neighbor model [6]. Most recently, the 2PKNN model exemplified the state of the art results of TagProp by combining metric learning and nearest neighbor [15]. They show the best results specifically on three standard datasets: Corel-5k, ESP-GAME and IAPRTC-12. Discriminative models such as SML treated multi-labeling as a multi-class problem [1], but this suffers from class imbalance (insufficient training samples per label) and lots of overlap among class specific distributions. Recently, an SVM based model [14] proposed by Verma and Jawahar modified the SVM hinge loss function in order to handle confusing labels. But in our approach, we show that we are able to get better results without any modifications to the SVM model.

# 3. PROPOSED MODEL

Our method is based on discriminative and generative models. Here we provide the details of the models used in this study and further explain how we fuse these models.

## 3.1 Discriminative Model

Image annotation may be viewed as a variation of a multi-class problem in which a number of words are employed to annotate a test image. However, in the case of images sharing the same annotations, the creation of multi-class models is very difficult because different classes share the same descriptors yielding noisy discriminative hyper-planes. In this work, we focus on binary models rather than a multi-class model. In the case of binary models the intra-class dependencies are ignored unlike the multi-class models. Here we create a binary classification model per word in the vocabulary and then make use of its responses for annotation. While creating a model $M_{w_i}$ for word $w_i$ we assume that the images (in the training set) annotated with $w_i$ are positive examples (i.e. $y_i = +1$) and similarly the images that are not annotated with $w_i$ are assumed to be negative examples (i.e. $y_i = -1$). Employing binary classification models for words enables us to deal with the issue of images sharing the same word annotations.

If our vocabulary consists of a number of words $W = \{w_1, w_2, ..., w_n\}$ then we create $n$ binary models each of which provides a discriminative model for its corresponding word. For a test image we get $n$ responses representing the probability of having an annotation of a word. The standard evaluations [9, 7, 5, 6, 15] require five word annotations per image; therefore, we annotate a test image with the five words having the highest responses. Imbalanced positive examples might be a problem for the image annotation task, since every word might have a different number of annotated images. We normalize the responses of each binary models to deal with this imbalance problem. We first take the normal inverse cumulative distribution of the responses (i.e. the probabilities of having a word as an annotation) and then we map them back to [0,1].

## 3.2 Generative Model

We use a discrete MBRM model as opposed to the continuous model proposed in [5]. The reason for the discrete version is due to the fact that it helps in reducing the computational complexity. Let $V$ represent the annotation vocabulary and $W$ be any arbitrary set of words. Also, Let $J$ be an image in the training dataset $\mathcal{T}$. Each image is associated with different sets of dimensionality reduced feature vector and annotation words. where, each feature vector $f$ has a dimension $m$ and annotation words have dimension n $W = w_1, w_2 \cdots w_n$. For a test image, we extract its features and its distribution is known but we need to predict the words associated with it, formally given by $P(w|f)$. From Bayesian theory,

$$P(w|f) = \arg\max_{w} \frac{P(w,f)}{P(f)} \tag{1}$$

One possible solution to computing the joint distribution $P(w, f)$ is by taking an expectation over the entire training set of images see [7]. Mathematically, the joint probability is given by:

$$P(w,f) = \sum_{J \in \mathcal{T}} \{P_{\mathcal{T}}(J) \prod_{i=1}^{m} P(f_i|J) \prod_{w_i \in w} P(w_i|J) \times \prod_{w_i \notin w} 1 - P(w_i|J)\} \tag{2}$$

$P_{\mathcal{T}}(J)$ is kept uniform for all images in the training dataset. $P(f_i|J)$ are estimated using smoothed maximum likelihood estimates as follows:

$$P(f_i|J) = (1 - \alpha_J)\frac{n(f_i, J)}{n(f, J)} + (\alpha_J)\frac{n(f_i, \mathcal{T})}{n(f, \mathcal{T})} \tag{3}$$

Here $n(f_i, J)$ represents the number of times visterm (quantized feature value) occurrs in the training image $J$, $n(f, J)$ denotes total number of visterms in image $J$, $n(f_i, \mathcal{T})$ denotes number of times visterm occurrs in the entire training dataset $\mathcal{T}$ and $n(f, \mathcal{T})$ indicate total number of visterms in the entire training dataset $\mathcal{T}$. The smoothing parameter $\alpha$ is estimated using a validation dataset.

$P(w_i|J)$ for each word is estimated using a Bayes estimate given by:

$$P(w_i|J) = \frac{\beta * 1_{w_i,J} + N_{w_i}}{\beta + N} \tag{4}$$

Here, $1_{w_i,J}$ is a indicator function for word $w_i$ occurring in image $J$. The smoothing parameter $\beta$ is estimated using a validation dataset resulting in a large number. $N_w$ is the number of training images containing $w_i$ and $N$ is the total number of training images.

### 3.2.1 LDA for Dimensionality Reduction

In our experiment, the feature sets are vector quantized and generally are large dimensional vectors. One of the main limitation of generative models such as CMRM, CRM or MBRM model is that their performance is limited by the dimensionality of the feature vector. Consider equation (2), in order to compute $P(f_i|J)$ we take a product over all the feature values because of the independence assumption.

Even though we use the log-sum-exp trick, its performance gets degraded. In order to overcome this we used a Latent Dirichlet Allocation model [8] to reduce the dimensionality. We treat each feature value in an image as a word and tried to summarize the words in the document by fewer topics. In other words, the LDA model gives us a compact representation of feature vectors. Experimentally we fixed the dimensionality of the feature vectors to be around 100 for all 14 features. These dimensionality reduced features were used only in the case of generative model whereas the feature dimensionality remained unchanged for the SVM model.

### 3.3 Fusion of Models

We described the discriminative and generative models in our method. Given that we make use of different descriptors we create separate models for each descriptor. Let $F = \{f_1, f_2, ..., f_m\}$ be a set of descriptors that we use in this work. Let $P_d(f_i)$ be the response of a discriminative model in terms of probabilities created for the descriptor $f_i$. Similarly let $P_g(f_i)$ be the response of a generative model in terms of probabilities created over the descriptor $f_i$. Then the final response $P_D$ for discriminative models is provided below;

$$P_D = \frac{1}{m}\sum_{i}^{m} P_d(f_i) \tag{5}$$

Similarly, the final response $G$ for generative models is as follows:

$$P_G = \frac{1}{m}\sum_{i}^{m} P_g(f_i) \tag{6}$$

The final response $P_R$ is based on the linear combination of discriminative and generative scores as follows:

$$P_R = (1-\lambda)P_D + \lambda P_G \tag{7}$$

## 4. EXPERIMENTAL SETUP

In this section, we provide information about datasets as well as the environmental settings used in our experiments.

As mentioned earlier, the hybrid model (SVM-DMBRM) is a fusion of a discriminative and a generative model. Each model is trained separately. For a test image $I$, we compute the probabilities for words based on its ability to best characterize the image using both models individually. Later, we fuse the normalized scores of the SVM and DMBRM model as given in Equation (7). Finally, the image is annotated with the top five (fixed annotations) words having high scores.

Experimental results are reported on the Corel-5K, ESP Game, and IAPRTC-12 datasets . These datasets have been widely used for reporting image annotation results. For better comparison of results, we follow the same train and test splits as TagProp. In Table 1, we provide information about the datasets used in this study in detail. While the Corel 5K dataset consists of about five thousand images, other datasets have more images and a larger vocabulary.

In order to create our discriminative and generative models we use the TagProp[6] features: histograms in RGB, HSV and LAB color space, SIFT descriptors extracted densely on a multi-scale grid and from Harris-Laplacain interest points along with four different features such as HOG2x2, LBP, Textons and Geotextons extracted using [19].

We make use of an SVM package, LIBSVM [2], to create our discriminative models. We set the regularization parameter -C- to a default parameter 1. Further, we employ the histogram intersection kernel which is shown to be successful in computer vision. Gensim[13], a python based library was used for implementing LDA. Empirically we fixed the reduced feature dimensionality to be a constant value of 100 across all the features.

In the literature, the standard evaluation technique used is based on a per word approach [9, 7, 5, 6, 15] . More recently, a per image evaluation was used by [17]. We follow the per word approach used by the majority of papers since it is a better evaluation technique. The per image evaluation used in [17] favors the most frequent words.[1] For instance, If we annotate the test images with the most frequent words in the training set, then we obtain a precision of 18 and a recall of 26 on the Corel-5k set which seems to be very high for an annotation using only the most frequent words of the training set.

Let $N$ be the number of images automatically annotated with a given word, $M$ be the number images correctly annotated with that word and let $K$ be the number of images having that word in ground truth annotation. Then in the case of per word evaluation, the recall and precision is calculated for every word in the test set, $recall = \frac{M}{K}$ and $precision = \frac{M}{N}$.

For each image, let $A$ be the number of words automatically annotated, $B$ be the number words correctly annotated and let $C$ be the number of words in ground truth annotation. In the case of a per image evaluation, recall and precision is calculated for every image in the test set, $recall = \frac{B}{C}$ and $precision = \frac{B}{A}$. Besides, the second type of evaluation always provide high numbers in both recall and precision, for an instance, even if we just happen to annotate the images with just five high frequency words then we still end up getting high recall and precision. Hence, we provide average precision and average recall scores that are directly comparable with most of the previous work involving per word evaluation.

For the linear combination of discriminative and generative models, we set the $\lambda$ parameter to a value of 0.5 (provides good balance for recall and precision). The parameter value $\lambda$ is a trade-off between high recall or precision, which can be chosen appropriately based on the requirements.

## 5. EXPERIMENTAL RESULTS AND DISCUSSION

Here we provide our experimental results and compare them with the previously reported numbers. Further we also discuss our results and findings.

### 5.1 Evaluation of Automatic Image Annotation

In Table 2, we provide our experimental results for three datasets; Corel-5K, ESP Game, and IAPRTC-12 in comparison with previously reported numbers. In the table, P represents the average precision, R represents the average recall, and N+ represents the non-zero recall (number of distinct words that are correctly assigned to the test image set). We provide three evaluations per datasets and our

[1] it also appears that [17] may have mixed up per word and per image evaluations.

Table 1: Details of the datasets used in this study.

| Dataset | Number of Images | Vocab. Size | Training Images | Test Images | Words Per Image | Images Per Word |
|---|---|---|---|---|---|---|
| Corel-5K | 5,000 | 260 | 4,500 | 500 | 3.4 | 58.6 |
| ESP Game | 20,770 | 268 | 18,689 | 2,081 | 4.7 | 362.7 |
| IAPRTC-12 | 19,627 | 291 | 17,665 | 1,962 | 5.7 | 347.7 |

Table 2: Experimental results of our methods as well as previously reported scores in three dataset; Corel-5K, ESP Game, and IAPRTC-12. P: Average Precision, R: Average Recall, N+: Number of distinct words that are correctly assigned to at least one test image. For all of the numbers the higher the better.

| | Corel-5K | | | ESP Game | | | IAPRTC-12 | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | P | R | N+ | P | R | N+ | P | R | N+ |
| CRM [9] | 16 | 19 | 107 | N/A | N/A | N/A | N/A | N/A | N/A |
| SML [1] | 23 | 29 | 137 | N/A | N/A | N/A | N/A | N/A | N/A |
| MRFA [18] | 31 | 36 | 172 | N/A | N/A | N/A | N/A | N/A | N/A |
| GS [20] | 30 | 33 | 146 | N/A | N/A | N/A | 32 | 29 | 252 |
| MBRM [5] | 24 | 25 | 122 | 18 | 19 | 209 | 24 | 23 | 223 |
| JEC [11] | 27 | 32 | 139 | 22 | 25 | 224 | 28 | 29 | 250 |
| CCD [12] | 36 | 41 | 159 | 36 | 24 | 232 | 44 | 29 | 251 |
| TagProp [6] | 33 | 42 | 160 | 39 | 27 | 239 | 46 | 35 | 266 |
| KSVM-VT [14] | 32 | 42 | 179 | 33 | **32** | 259 | 47 | 29 | 268 |
| 2PKNN [15] | **44** | 46 | 191 | 53 | 27 | 252 | 54 | **37** | 278 |
| SVM-DMBRM | 36 | **48** | **197** | **55** | 25 | **259** | **56** | 29 | **283** |

models outperform the previously reported scores on two of them. For Corel-5K dataset our model provides the highest recall and non-zero recall numbers(N+). In the case of ESP Game dataset, our models provide the highest scores for precision and N+. Similarly for the IAPRTC-12 dataset we again provide the highest precision and N+ scores.

When we consider N+ scores, our model outperforms all of the previously reported techniques. N+ is a measure of how well does the system perform with the imbalanced positive example problem and also it indicates the number of distinct words used for annotation. Thus, our system with high N+ score is able to handle imbalanced data and the poor labeling problem more effectively.

Examples of our model's prediction matching with ground truth for all three datasets are provided in Figure 1. Note that since we are restricted to annotating with only five words, instances of IAPRTC-12 and ESP-game predictions are incomplete. From Figure 1, consider the second instance of ESP-game ($2^{nd}$ row, $2^{nd}$ column), it has human annotations: *light, planet, ship, space, star, sun* among which our model misses space since we are restricted to annotating with only five words.

Examples of our model failing to match the automatic annotations to ground truth for all three datasets are provided in Figure 2. From Figure 2 we can observe that in most of the instances, even though our algorithm produces relevant annotations they are sometimes found missing in the ground truth annotations. Consider the first example in the IAPRTC-12 dataset($3^{rd}$ row, $1^{st}$ column), we could see that there are some wrong/misleading human annotations such as tree, table and one which are noisy labels. Consider another example in the Corel-5k dataset ($1^{st}$ row, $1^{st}$ column), we observe that our model's annotation appears to be visually more accurate than the ground truth.

Table 3: Experimental results on different number of annotations (# ) of our method on three datasets; Corel-5K, ESP Game, and IAPRTC-12. P: Average Precision, R: Average Recall, N+: Number of distinct words that are correctly assigned to at least one test image. For all of the numbers the higher the better.

| | Corel-5K | | | ESP Game | | | IAPRTC-12 | | |
|---|---|---|---|---|---|---|---|---|---|
| # | P | R | N+ | P | R | N+ | P | R | N+ |
| 1 | 37 | 21 | 119 | 54 | 11 | 206 | 51 | 12 | 215 |
| 2 | 41 | 31 | 156 | 59 | 16 | 240 | 57 | 18 | 259 |
| 3 | 40 | 39 | 177 | 58 | 20 | 252 | 58 | 23 | 273 |
| 4 | 38 | 44 | 189 | 56 | 22 | 254 | 57 | 26 | 278 |
| 5 | 36 | 48 | 197 | 55 | 25 | 259 | 56 | 29 | 283 |
| 10 | 31 | 56 | 207 | 49 | 31 | 261 | 50 | 37 | 285 |
| 20 | 26 | 64 | 221 | 41 | 40 | 264 | 42 | 47 | 285 |

In Table 3 we provide the experimental results of our methods for different number of annotations. Even though the evaluation is performed with a fixed number of annotations (five) per image, here we provide the experimental results using different number of annotations i.e. 1, 2, 3, 4, 5, 10, and 20 per image to better understand the system performance. Table 3 if we consider five words annotations and ten word annotations then the precision scores decrease and recall scores increase as expected. Interestingly, the small change in N+ scores (259 to 261 for ESP Game and 283 to 285 for IAPRTC-12) indicates that our model is able to make use of almost all distinct words from the vocabulary for annotating test images even in the case of predicting five

**Automatic annotation:** *f-16, jet, plane, sky, smoke*
**True annotation:** *sky, jet, plane, smoke*

**Automatic annotation:** *needles, petals, cactus, blooms, flowers*
**True-annotation:** *flowers, needles, blooms, cactus*


**Automatic annotation:** *formula, wall, cars, tracks, crafts*
**True annotation:** *wall, cars, tracks, formula*

**Automatic annotation:** *athlete, water, swimmers, pool, people*
**True annotation:** *water, people, pool, swimmers*

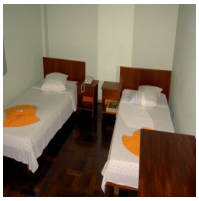
**Automatic annotation:** *paper, old, wood, brown, small*
**True annotation:** *brown, old, paper, wood*

**Automatic annotation:** *planet, ship, light, sun, star*
**True annotation:** *light, planet, ship, space, star, sun*

**Automatic annotation:** *boat, ocean, sea, sky, water*
**True annotation:** *boat, ocean, sea, ship, sky, water*

**Automatic annotation:** *ocean, cloud, sky, water, sea*
**True annotation:** *blue, cloud, ocean, sea, sky, water*

**Automatic annotation:** *stripe, fog, train, roof, mountain*
**True annotation:** *fog, mountain, roof, stripe, train*



**Automatic annotation:** *room, bedcover, bed, towel, wood*
**True annotation:** *bed, bedcover, room, towel, wall, wood*

**Automatic annotation:** *bay, beach, cloud, building, street*
**True annotation:** *bay, beach, building, cloud, street, tree*

**Automatic annotation:** *flag, side, woman, man, sky*
**True annotation:** *flag, man, side, sky, wall, woman*

Figure 1: Examples of our model's automatic annotation matching with ground truth for all three datasets. Each row corresponds to a different dataset, First row: Corel-5k, Second row: ESP-Game, Third row: IAPRTC-12.

**Automatic annotation:** *water, city, town, peaks, tower*
**True annotation:** *sky, water, reflection, castle*

**Automatic annotation:** *pillar, stone, road, shadows, sculpture*
**True annotation:** *buildings, shadows, stone, pillar*

**Automatic annotation:** *crystals, lion, ice, fruit, town*
**True annotation:** *ice, frost, frozen*

**Automatic annotation:** *beach, water, lake, island, ships*
**True annotation:** *sky, water, beach, sand*

**Automatic annotation:** *guy, rock, mountain, people, man*
**True annotation:** *green, man, people, tree*

**Automatic annotation:** *painting, anime, eat, smoke, art*
**True annotation:** *art, blue, colors, painting, picture, red*

**Automatic annotation:** *game, dark, small, ice, album*
**True annotation:** *computer, dark, game, picture, purple*


**Automatic annotation:** *tree, sketch, internet, shop, icon*
**True annotation:** *tree*

**Automatic annotation:** *shrub, couch, bush, path, ridge*
**True annotation:** *grass, hill, landscape, mountain, one, table, tree*

**Automatic annotation:** *girl, child, couple, backpack, cloth*
**True annotation:** *child, girl, hair, head*

**Automatic annotation:** *fjord, landscape, ridge, mountain, village*
**True annotation:** *lake, landscape, meadow, ridge, sky*

**Automatic annotation:** *tourist, group, couple, sea, lot*
**True annotation:** *coast, lookout, sea, team, tourist*

Figure 2: Examples of our model failing to match the automatic annotations to ground truth for all three datasets. Each row corresponds to a different dataset, First row: Corel-5k, Second row: ESP-Game, Third row: IAPRTC-12.

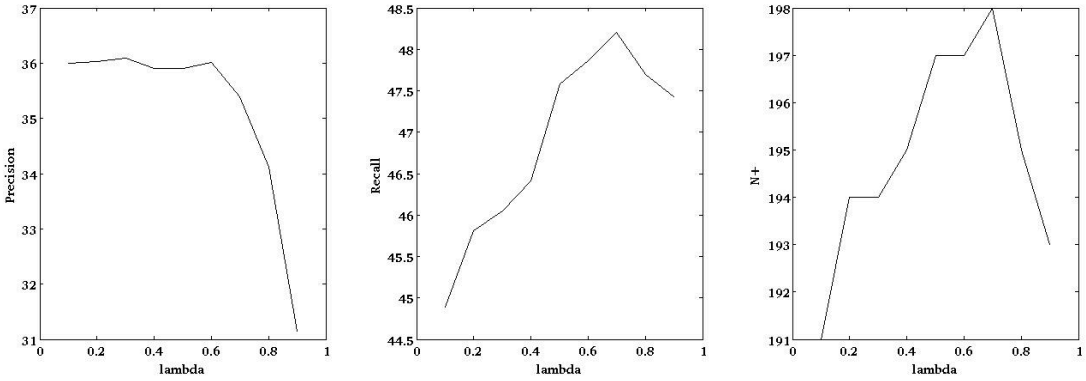



Figure 3: Precision, Recall, and N+ scores with different $\lambda$ (lambda) values.

Table 4: Single word retrieval results for three datasets; Corel-5k, ESP Game, and IAPRTC-12

| Corel-5K | | ESP Game | IAPRTC-12 |
|---|---|---|---|
| Method | **MAP** | **MAP** | **MAP** |
| CRM [9] | 26 | N/A | N/A |
| MBRM [5] | 30 | N/A | N/A |
| TagProp [6] | 42 | 40 | 28 |
| JEC [11] | 35 | 21 | 27 |
| SVM-DMBRM | **57** | **71** | **73** |

words.

In Figure 3 we provide the precision, recall, and N+ scores for different values of $\lambda$. The main purpose of this study was to show the model's unique capability to get the desired performance for fixed number of annotation words, by just varying the $\lambda$ parameter. When the $\lambda$ value is close to 0, SVM models dominate the final scores, on the other hand if the parameter is close to 1, then the DMBRM models dominate the final scores. The precision score decrease when $\lambda$ gets larger. Recall scores increase when $\lambda$ gets larger.

## 5.2 Evaluation of Ranked Retrieval for Single Word Queries

In this section, we provide the retrieval results per word query for all three datasets using our proposed model. For a given query, our system return the images automatically annotated with that word. Further, these images are ranked according to their annotation scores. Finally, we compute the mean average precision (**MAP**) for this ranked list. **MAP** results for all three datasets are reported in Table 4 . Since, the state of the art results paper [15] does not report retrieval in terms of **MAP**, we cannot directly compare to them. Hence, we compare our results with the next best results provided by Tagprop [6]. Our **MAP** scores on all the three datasets are significantly better than their best reported scores by a factor of 1.35 for Corel-5k, 1.77 for ESP-game and 2.7 for IAPR-12 dataset.

# 6. CONCLUSION

In terms of image annotation evaluation, we showed that a hybrid model combining both discriminative and generative model capabilities gives results comparable to the state of the art on three challenging datasets and always perform better in annotating with the number of distinct words (N+ measure). In addition, our proposed model significantly outperforms state of the art results in terms of ranked retrieval results evaluation. Limitations of the MBRM model of using high dimensional features were overcome by using LDA for dimensionality reduction. We showed that our proposed model is able to address the problem of data imbalance and poor annotation which are prevalent in the real world. Our future work will investigate unsupervised feature learning to replace handcrafted features and also, we will focus on more efficient way of combining these models.

# 7. ACKNOWLEDGMENT

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #IIS-0910884. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

# 8. REFERENCES

[1] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(3):394–410, Mar. 2007.

[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[3] P. Duygulu, K. Barnard, J. F. G. d. Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European Conference on Computer Vision-Part IV*, ECCV '02, pages 97–112, London, UK, UK, 2002. Springer-Verlag.

[4] S. Feng and R. Manmatha. A discrete direct retrieval model for image and video retrieval. In *Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval*, CIVR '08, pages 427–436, New York, NY, USA, 2008. ACM.

[5] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR'04, pages 1002–1009, 2004.

[6] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *In ICCV*, 2009.

[7] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 119–126, 2003.

[8] S. Lacoste-Julien, F. Sha, and M. I. Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. In *Advances in neural information processing systems*, pages 897–904, 2008.

[9] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *in NIPS*. MIT Press, 2003.

[10] A. Llorente, R. Manmatha, and S. Rüger. Image retrieval using markov random fields and global image features. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '10, pages 243–250, New York, NY, USA, 2010. ACM.

[11] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *Proceedings of the 10th European Conference on Computer Vision: Part III*, ECCV '08, pages 316–329, Berlin, Heidelberg, 2008. Springer-Verlag.

[12] H. Nakayama. *Linear distance metric Learning for large-scale generic image recognition*. PhD thesis, The University of Tokyo, Japan, 2011.

[13] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, 2010.

[14] Y. Verma and C. Jawahar. Exploring svm for image annotation in presence of confusing labels. In *Proceedings of the 24th British Machine Vision Conference*, 2013.

[15] Y. Verma and C. V. Jawahar. Image annotation using metric learning in semantic neighborhoods. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part III*, ECCV'12, pages 836–849, Berlin, Heidelberg, 2012. Springer-Verlag.

[16] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 319–326, New York, NY, USA, 2004. ACM.

[17] L. Wu, R. Jin, and A. K. Jain. Tag completion for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):716–727, 2013.

[18] Y. Xiang, X. Zhou, T. Chua, and C. Ngo. A revisit of generative models for automatic image annotation using Markov random fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[19] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3485 –3492, june 2010.

[20] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and D. Metaxas. Automatic image annotation using group sparsity. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.