

# Project Report

# Automated Image Annotation

---

Nishant Oli (MT2016096)

Suparna Ghanvatkar(MT2016138)

Swatantra Pradhan(MT2016140)

Vijay Aggarwal(MT2016152)

Vivek Mehta(MT2016155)

GROUP NO: 23



## Introduction

The problem we have tried to solve is - “Given an image, what are the words that can annotate the image?”. We are trying to find suitable annotations for the images automatically using Image features and Machine Learning models. The problem of automated image annotation finds its use in many applications. We have tried to find annotation for image based on the underlying assumption that the images are derived from some basic topic content. This possible distribution of topics which make up the image are the basis for our choice of approach towards the problem. As our basic assumption for the system is based on a generative approach, we have tried to find solutions which follow the above mentioned though process.



## Motivation

Exponentially growing photo collections motivate the needs for automatic image annotation for effective searching. The problem of automated image annotation finds its use in many fields. It can be used for keyword based image retrieval, or image description generation or image classification into appropriate categories, and many more applications. For all these applications, a central requirement is automated image annotation.

The internet is full of images consisting of strongly annotated(i.e. like available due to uploads by humans which have been manually tagged) and weakly annotated(i.e. like available on blogs, newspaper reports, inaccurate tags by humans, etc). Currently, the images resulting from a web search are due to these annotations present on the images which in many case do not describe the content of the image properly(weak annotation). Also, there are cases where the images do not have annotation. Such images need to be accessible by a simple search of appropriate keyword which can be achieved by good automated image annotations on the images available over the internet.

## Literature Review

### 1. An Overview of Automated Image Annotation Approaches

This paper basically performs the literature review for the techniques available for doing automated image annotation. There are many interesting approaches presented in the paper like annotations based on segmented data, variation based inference on region to annotation, continuous space relevance model. One of the last techniques mentioned was the technique which interested us the most – the one using latent dirichlet allocation for images.

### 2. Object recognition as machine translation : Learning a lexicon for a fixed image vocabulary


This paper considers the problem of image annotation as the problem of machine translation from visual words to the text based annotation words. A word to word mapping task as done for machine translation is undertaken and EM algorithm is used to re-estimate the probability of the annotation given the visual words present in the image iteratively.

### 3. Modeling annotated data

This paper proposes the algorithm of correspondence LDA. Firstly, an image is modeled using the bag of visual terms model, which is a simple model that represents a document as an order less set of terms. In the case of images, an image is therefore represented as an order less sequence of visual terms. Given a collection of images, the first task to perform is to identify a set of all visual terms used at least once in at least one image. This set will be called the Vocabulary. Once the Vocabulary is set now each image is represented as a vector with integer entries of length which is same as the size of the Vocabulary set. In CORR-LDA model, image features are firstly generated and subsequently words are generated. Indeed, N region features are generated. Then, for each of the M words, one of the regions is selected from the image and a corresponding word is drawn conditioned on the topic that generates the selected region.

### 4. Weakly supervised Image Annotation and Segmentation with objects and attributes

This paper considers dataset of images having segmentation labels. This approach uses non-parametric Bayesian model which learns from weakly annotated images as available widely over the internet. An important task undertaken in the paper is also to generate



suitable annotations from the weakly labeled image sets available. This is achieved by weakly supervised Markov Random Field Stacked Indian Buffet process (WS-MRF-SIBP) that models objects and attributes as latent factors and explicitly captures the correlation.

#### 5. Simultaneous Image Classification and Annotation

This paper proposes a modification to the supervised LDA algorithm (s-LDA). The s-LDA performs the task of classification given the class label. This approach combines the techniques of corr-LDA and s-LDA to generate the class as well as annotations for a given image. This paper uses the Label-Me dataset which is structured in the similar manner.

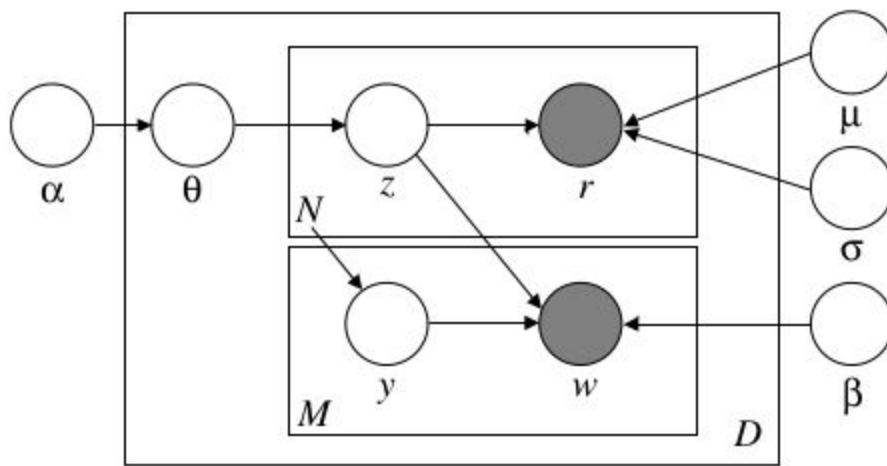
#### 6. Automatic Image Annotation and Retrieval Using the LDA Model

This paper proposes the technique of using LDA and vector quantization (VQ). The LDA is performed on segmented images obtained by using the watershed algorithm. These segmented regions are quantized using k-means and the appropriate annotations are generated. This paper gave an inspiration for using the unsupervised LDA algorithm for our project.

## Our Model Explanation

We had initially planned to use the corr-LDA model, but due to a lack of API for the same, we went ahead with using LDA for our project. A simple LDA cannot be used for achieving our task and so we tried to achieve what corr-LDA tried to do using the simple LDA.

The model for corr-LDA is:



On observing the model and reading the paper, we observed that at the base level, we are considering a similar generative model for the  $N$  visual words and  $M$  textual annotations available as the simple LDA.

The annotated word  $w$  is estimated using the  $y$  which is derived using uniform sample of each image region. We can consider the annotations themselves to exhibit the lda type of model and have underlying topics determine the word chosen for annotation rather than a uniform distribution assumed.

The topic space of both these image regions and the annotated words is different but has to again inherently be similar as they represent similar region. So they must be derived from similar topics.

Once this generative model is in place, we need some kind of relation or estimation for the translation between the models. The basic missing piece is that the visual word to annotation

type of relation must be established. We can use principles of machine translation for it, but we experimented with a 2PKNN approach which estimates probabilities for each annotation being present for the image.

For each annotation, K similar images are determined, given a test image. We can consider a simple posterior for estimating the probability where J is test image, I are the similar images determined.

$$P(J|w_k) = \sum_I \theta_{J,I} P(w_k, I)$$

The  $\theta$  gives the similarity between J and I and  $P(w_k, I)$  is 0 or 1 depending on whether image I has the label  $w_k$ .

For every image, we try to build a probability distribution of each word in textual vocabulary by considering the above probability and the probability obtained by LDA.

### Advantage of our model

The procedure followed in all the literature is to use highly specific images for training i.e. images of fish, water, etc and have a single annotation for it as available in corel-5k and then on a new test image, essentially the task of object detection and identification is done. But, we can build a very rich annotation dataset if we utilize the commonly available images and their annotations, like available on flickr or pininterest etc. Our model essentially builds on this philosophy.

It can be also noted that, the probability estimation done on the annotations also ensures that the disparity in the word annotation and image distribution is normalized while inferencing.

Thus, we have tried to combine the benefits of corr-LDA along with the normalizing for annotation words and their relation to visual words.

## Current Work

### Datasets explored

We have experimented with various datasets as each dataset is tailored to a specific task. A report of the datasets explored:

1. Corel-5k: The images in this dataset are classified into one of the 10 classes - hence is suitable for multi-class-classification problem and not suitable for our task as we want multiple annotations for each image.
2. Label-Me: This dataset is having user-specified labels and also is basically a categorizing problem. The interface allows us to define region wise annotation. This would require manual effort and hence is not used for the project.
3. IAPR-T12 dataset: This dataset has text-based comments for each image. This dataset has been used for our project task.
4. NUS-WIDE dataset: This dataset has annotations meant as tags for the image and thus are consisting of not just the content but descriptive colors, etc.

### Algorithms

As the NUS-WIDE dataset is too large, we have performed reservoir sampling to randomly obtain the images so that images of all various categories are obtained uniformly. This is used for the dataset creation.

### Feature Extraction

1. SIFT descriptors for the image are found
2. To form a visual vocabulary, we perform k-means on all the SIFT descriptors stacked together. The k is set to the size of the visual vocabulary.
3. Each image's SIFT descriptor is quantized and identified as one of the word of the visual vocabulary determined in the previous step.
4. A histogram of the visual words for each image is computed and used as feature.

### Annotation extraction (required for IAPR-T12 dataset)

1. Identify the description in the annotation file.



2. POS tagging using NLTK on the words obtained in description. Only the words tagged as NN and NNP are forwarded.
3. Stopword removal.
4. Lemmatization and stemming using WordNetLemmatizer
5. The stemmed words obtained are used as annotations and the vocabulary generation and document-word matrix is generated accordingly.


## Training

1. The histogram of words obtained for each image is used to train the Latent Dirichlet Allocation model (package lda in python which uses collapsed Gibbs sampling). The output of LDA algorithm is the per document-topic proportions, etc.
2. The per document-topic proportions for each image are considered and for each image, the important topics are detected by finding a threshold for each image based on the highest topic estimate and lowest topic estimate for each document. According to this threshold which is different for each document, a 1-0 based vector is generated where 1 represents that the topic at the index is important for the given document 4 e.g. [0.2, 0.01, 0.02, 0.77] gives 0.01 as minimum and 0.77 as maximum and the threshold is determined to be 0.39. So the vector becomes: [0,0,0,1] i.e. only topic 4 is important for the document.
3. For these vectors of important topics detected, a clustering using k-means is performed to cluster together the documents having similar important topics.
4. LDA on the textual annotations
5. Important topic determination for each image based on the annotation LDA.
6. Clustering using k-means on the important topics in annotation space determined.

The LDA parameters in the visual as well as textual words are determined using the training phase.

## Inferencing

1. Find the K most similar images for each word in textual vocabulary and store along with their cosine similarity with the test image.
2. Fit the test image into the LDA model for visual words and obtain point estimates of the topic proportions for the new image.
3. Find important topics for the test image and fit it into one of the clusters.
4. For all images in the cluster determined in the visual space, the similarity of the trained images and the new image is detected using hamming distance of the 0-1 vector and appropriately weighted belongingness to the annotation cluster space is determined. i.e. the annotation cluster to which the most similar trained image belongs is given the maximum weight and the probability of annotation space clusters is obtained.

- 
5. Find the probability estimate for each word in visual vocabulary to be annotated to given test image using the result of step 1.
  6. Multiply the probability estimate obtained in step 5 and the topic-word probability estimate for the important topics determined in step 4.
  7. Annotate the image with the top words of this probability estimate obtained in step 6.

## Results

We have run our model on two datasets- IAPR-T12 and NUS-WIDE dataset. Due to system and memory constraints, we have currently trained only on very small number of images. So our results can be considered representative of the functionality of our model.

\*To access the results, datasets, kindly use the link <https://drive.google.com/open?id=0B15hiCPPynGsRXpLVGpGczU4bIU>.

NUS-WIDE dataset is not included due to size constraints.

### Initial Small Dataset

Initially, our model only tried to approximate using LDA on both annotation and visual space. This was initially trained on a meagre 50 images and tested on 4 images. The results gave a confidence in the approach to further refine the model.



annotation: school, camera, blue, bit



annotation: foreground, left, building, top, background, right, house, gate



annotations: building, column, flag, lawn



annotation: school, camera, bit, tourist

In this the basic problem was the problem nearly turned into topic classification problem and images classified into similar topic got same label. This was due to not considering each label individually for its contribution.

### Results on IAPR-T12 dataset - Result Set 1

We enhanced our model to be the current model and tried to obtain results on a larger dataset of 355 training images and 34 test images. This corrected the problem of the previous model by estimating probability for each annotation along with consideration for topic allocation.

Parameter values:

- ☐  $K(\text{no. of similar images to determine}) = 5$
- ☐  $n_{\text{topics for visual words}} = 8$
- ☐  $n_{\text{topics for annotations}} = 10$
- ☐  $k \text{ no. of clusters based on important topics in visual space} = 10$
- ☐  $k \text{ no. of clusters based on important topics in annotation space} = 17$



annotations: road gravel roof jungle



annotations: road gravel roof jungle



annotations: hill sea view beach



annotations: mountain landscape view sky

Though the new model was designed to overcome the issue of nearly classifying into topic and allotting same annotation to all images in the topic, we observe that this has not happened. On studying the probability determined by K similar images for each annotation, we observe that it is nearly similar as the K is very huge compared to the dataset size used. For every annotation around maximum of 5-6 images are only being allotted in the dataset and we have effectively considered all the images and reduced it into the previous model. The new probability values effectively get marginalized out.

### Results on IAPR-T12 dataset - Result Set 2

To overcome the problem faced in the previous iteration, we changed the value of K.

Parameter values:

- ☐ K(no. of similar images to determine) = 2
- ☐ n\_topics for visual words = 8
- ☐ n\_topics for annotations = 10
- ☐ k no. of clusters based on important topics in visual space = 10
- ☐ k no. of clusters based on important topics in annotation space = 17



annotations: mountain landscape view lake



annotations: view sky city jungle



annotations: road gravel picture group





annotations: table wooden round restaurant

We observe that the classification type of pattern is broken. But still there is a strong relation in the co occurrence of the words. To handle this, reducing the weightage of the topic model derived probabilities would prevent it from become a classification type of problem.

We also observe that the model is identifying and annotating scenes or images it has been trained on properly. For topics or categories with only few images in training set, the model is finding it difficult to annotate.

The scenery based scenes are getting annotated properly whereas the human based images are getting mixed up. From the train dataset we can see that as it had a lot of scenic images so the annotation was carried out better than the others. So we can infer that a larger training set will facilitate the annotation.

### Results on IAPR-T12 dataset - Result Set 3

To overcome the problem faced in the previous iteration, we changed the value of k.

Parameter values:

- ☐ K(no. of similar images to determine) = 2
- ☐ n\_topics for visual words = 8
- ☐ n\_topics for annotations = 10
- ☐ k no. of clusters based on important topics in visual space = 10
- ☐ k no. of clusters based on important topics in annotation space = 10



We observed more or less same results as in Result Set2.



Annotations: view group photo picture



Annotations: front house view entrance



Annotations: view group photo picture

### Results on IAPR-T12 dataset - Result Set 4

To overcome the problem faced in the previous iteration, we changed the value of k.

Parameter values:

- ❑ K(no. of similar images to determine) = 2
- ❑ n\_topics for visual words = 8
- ❑ n\_topics for annotations = 10
- ❑ k no. of clusters based on important topics in visual space = 10
- ❑ k no. of clusters based on important topics in annotation space = 25

We observed that increasing the value of k does not improve the results, instead we got same annotations for all the image. So we can conclude that for certain range of values of k, this gives appropriate results.

### Results on NUS-WIDE dataset

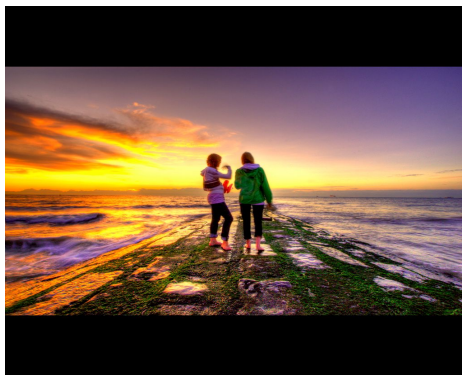
The NUS-WIDE dataset has very high quality images and also the annotations are not very appropriate. Resizing of the images is performed as even 177 images were giving memory error otherwise. The resizing also led to a decrease in quality of annotations, coupled with the fact that the annotations are not very proper as they are very abstract and relevant on the dataset itself, our model performs a little poorly in this scenario. Further it can be said that training over the complete dataset will give good results as for a fraction of dataset taken randomly the annotations are very abstract for training.



Annotations: explore interestingness orange



Annotations: nature flower animal nikon



Annotations: car train politics railroad

Thus, we can conclude that our model has performed fair enough even in the scenario of very low training images, and the model is expected to give better results with bigger training sets.

## Future Scope and Analysis of Shortcomings

### Major challenges and shortcomings in the project

1. Semantic Gap between low-level image features and the context of the image used for annotation. Though, over the course of the project, we have come to realize that we can also try to get cues from features like colors, textures for the similarity measures.
2. Lack of correspondence of keyword and region is a major problem. The annotation is for the whole image but is basically derived from a certain region. The correspondence between this creates an inherent assumption fault in our model as we assume the image gets annotated from any region and we do not estimate the region. To overcome this challenge a few teams have used manually region-wise annotated data.
3. We cannot identify or annotate using words the model has not been trained on. This requires huge dataset and subsequently, a very high computation cost, which was not in the scope of mini-project.

### Future scope

An implementation of the actual correspondence-LDA along with the probability estimate using 2PKNN approach mentioned in the model might give more interesting results as that would still follow our paradigm. A deep learning based system also can be explored, but the current methodologies also learn using single object type images like in corel-5k which is a limiting factor for real-world usage. So a different approach which better suits the real world need can be explored.



## References

1. Sumathi, C.Lakshmi Devasena, and Hemalatha, An Overview of Automated Image Annotation Approaches, International Journal of Research and Reviews in Information Sciences, 2011
2. P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary, In Seventh European Conference on Computer Vision
3. D. Blei, Michael, and M. I. Jordan. Modeling annotated data, Proceedings of the 26th annual international ACM SIGIR conference
4. J. Jeon, V. Lavrenko and R. Manmatha, Automatic Image Annotation and Retrieval using Cross-Media Relevance Models, Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval
5. Dr. Simon J. Prince, Computer Vision: Models, Learning and Inference
6. Zhang, Lu, Chan, Li, Automatic Image Annotation and Retrieval Using the LDA Model, IJCES 2011
7. Shi, Yang, Hospedales, Xiang, Weakly supervised Image Annotation and Segmentation with objects and attributes, IEEE Transactions on Pattern Analysis and Machine Intelligence
8. Wang, Blei, Fei-Fei, Simultaneous Image Classification and Annotation, IEEE Conference on Computer Vision and Pattern Recognition, 2009

