International Journal on Applications in Information and Communication Engineering
Volume 1: Issue 8: August 2015, pp 41-45. www.aetsjournal.com ISSN (Online) : 2394-6237

Generating Text From Images Using Phrase-Based Model

R.Rajaguru

Abstract**—** Annotation based image retrieval is more feasible to fulfill user requirements in a straight forward manner. Modern multi-media documents are not merely collections of words but can be a vast collection of related text, images and audio references. Images that do not coincide with textual data cannot be retrieved. Analyzing the pictures in large collections is a crucial problem. Search engines on the web retrieve images without viewing their content, simply by matching user queries against thematically collocated textual information which in turn limits the applicability. Methods are proposed to automatically generate captions for a picture from a weakly labeled data. So as to annotate images and generate captions a probabilistic suggestion with abstractive and extractive caption generation model prevails. Indeed, the abstractive model compares favorably to handwritten captions and is often superior to extractive methods. However the system is designed which is been used to realize the features of the images locally and is less grammatical. A phrase-based probabilistic model is framed to generate captions for images. To resolve such criteria images are experimented with global features in thematically co-located documents related to document structure such as titles, section of articles and also to exploit syntactic information more directly.

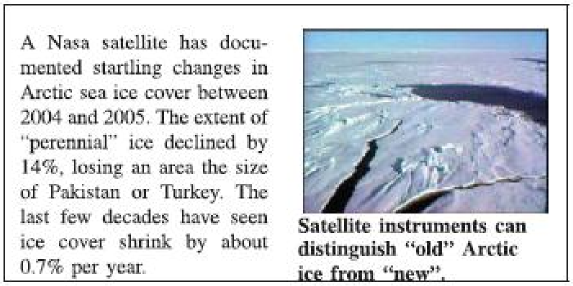
Keywords **—** image annotation, text summarization, abstractive and extractive topic models.

I. INTRODUCTION

Browsing pictures in large-scale is an important problem that has attracted much interest within information retrieval. Many of the search engines deployed on the web retrieve images without analyzing their content, simply by matching user queries against collocated textual information. Examples include metadata (e.g., the image's file name and format), user-annotated tags, captions, and, generally, text surrounding the image.

This limits the applicability of search engines (images that do not coincide with textual data cannot be retrieved), a great deal of work has focused on the development of methods that generate description words for a picture automatically. The literature littered with various attempts to learn the associations between image features and words using supervised classification, and models inspired by information retrieval.

A method that generates such descriptions automatically could therefore improve image retrieval by supporting longer and more targeted queries, by functioning as a short summary of the image's content. It could also assist journalists in creating descriptions for the images associated with their articles or in finding images that appropriately illustrate their text.



A Nasa satellite has documented startling changes in Arctic sea ice cover between 2004 and 2005. The extent of "perennial" ice declined by 14%, losing an area the size of Pakistan or Turkey. The last few decades have seen ice cover shrink by about 0.7% per year.

Satellite instruments can distinguish "old" Arctic ice from "new".

Fig. 1: Entry in BBC News database contains a document, an image and it's caption (in boldface)

Traditional approaches to object detection only look at local pieces of the image[13], as within regions around an interest point detector. However, such local pieces can be ambiguous, especially when the object of interest is small, or imaging conditions are otherwise unfavorable. This ambiguity can be reduced by using global features of the image which is called the "gist" of the scene as an additional source of evidence. By combining local and global features, significantly improved detection rates are obtained.

An approach for learning the semantics of images was devised by Jeon [5] which allows to automatically annotate an image with keywords and to retrieve images based on text queries. A training set of images with annotations is used to compute a joint probabilistic model for image features and also to predict the probability of generating a word given with image regions[2].Feng et al. Proposed relevance model to retrieve image in response to textual queries with some knowledge of semantics of pictures. It is a joint probability distribution of the word annotations and the image feature vectors which been computed using training set.

II. PROPOSED SYSTEM

The caption generation model adopts a two-stage approach where the image processing and surface realization are carried out sequentially. Image processing undergoes various steps like content selection which identifies what the image and accompanying article are about, whereas surface realization determines how to verbalize the chosen content.

1. The caption describes the content of the image directly or indirectly. Unlike traditional image annotation where keywords describe salient objects, captions supply more detailed information, not only about objects and their

R.Rajaguru, Department of Computer Science and Engineering, Fatima Michael College of Engineering and Technology, Anna University, Chennai,India. (Email: guruwizi@gmail.com)

attributes, but also events. For example, in Fig.1 the caption mentions the differentiation in ice covered area and the cause for its change too.

2. Since the images are implicitly rather than explicitly labeled, here do not assume that object can enumerate all present in the image nor that can create a detailed description of them. Instead, hope to model event-related information such as "what happened," "who did it," and "where" with the help of the news document.

Image Representation by Recognizing Global Features

Most object recognition systems have taken one of two approaches, using either global or local features exclusively. This may be in part due to the difficulty of combining a single global feature vector with a set of local features in a suitable manner [15]. In this paper, it's shown that combining local and global features is beneficial in an application where rough segmentations of objects are available.

Many object recognition systems use global features that describe an entire image. Most shape and texture descriptors fall into this category. Such features are attractive because they produce very compact representations of images, where each image corresponds to a point in a high dimensional feature space. As a result, any standard classifier can be used. On the other hand global features are sensitive to clutter and occlusion. As a result it is either assumed that an image only contains a single object, or that a good segmentation of the object from the background is available. Here in this paper an image often does contain a single object, but sometimes several organisms or particles are present. It is found that a simple global bimodal segmentation is usually effective for separating the plankton from the background, which tends to be significantly darker than the object. Here I use expectation maximization (EM) to fit a mixture of two Gaussians to the multimodal [6, 4] histogram of gray values for a given image.

Scale Invariant Feature T Algorithm to form Visual Vocabulary *(*V oc$_v$*)*

SIFT is one of the most effective interest point detectors, which uses scale-space extreme as interest points and provides a localized high-dimensional descriptor [14]. In general the existing object recognition algorithms can be classified into two categories: global and local features based algorithms. Global features based algorithms aim at recognizing an object as a whole. To achieve this, after the acquisition, the test object image is sequentially pre-processed and segmented. Calculation of SIFT image features is performed through the four consecutive steps which are briefly described in the following:

A. scale-space local extrema detection

The features locations are determined as the local extrema of Difference of Gaussians (DOG pyramid). To build the DOG pyramid the input image is convolved iteratively with a Gaussian kernel of σ = 1.6. The last convolved image is down-sampled and the convolving process is repeated. This procedure is repeated as long as the down sampling is possible. Each collection of images of the same size is called an octave. All octaves build together the so-called Gaussian pyramid, which is represented by a 3D function $L(x, y, \sigma)$.

B. Key point localization

The detected local extreme are good candidates for key points. Then, local extreme with low contrast and such that correspond to edges are discarded because they are sensitive to noise. Orientation assignment - once the SIFT-feature location is determined, a main orientation is assigned to each feature based on local image gradients. For each pixel of the region around the feature location the gradient magnitude and orientation are computed.

$$\mathrm{m(x,y)} = \sqrt{(L(x+1,y,\sigma)-L(x-1,y,\sigma))^2+(L(x,y+1,\sigma)-L(x,y-1,\sigma))^2}$$

$$\Theta(\mathrm{x,y}) = \arctan\left(\frac{(L(x,y+1,\sigma)-L(x,y-1,\sigma))}{(L(x+1,y,\sigma)-L(x-1,y,\sigma))}\right) \quad (1)$$

C. key point descriptor

The region around a key point is divided into 4X4 boxes. Then, for each box an 8 bins orientation histogram is established. From the 16 obtained orientation histograms, a 128 dimensional vector (SIFT-descriptor) is built.

These descriptors in turn is further restricted as groups expressed in a bag-of words format. At present both visual images and textual factors are transformed into a same status thus forming a mixed document called d_{Mix}

D. Document Representation

Image annotation model makes use of probabilities estimated by Latent Dirichlet's Allocation(LDA) which uses a three level hierarchical Bayesian model

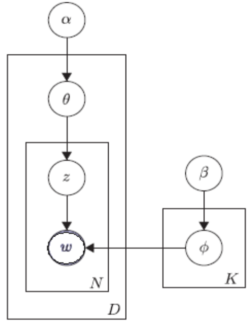


Fig. 2: LDA topic model:The variables in lower corner refer to the number of samples.

This corpus IN Fig.2.,consists of D documents which is been modeled using mixture of K topic models The generative process for a document d is given as,

Choose $\theta \mid \alpha \sim Dir(\alpha)$
For $n \in 1,2,\ldots\ldots N;$
Choose topic $z_n \mid \theta \sim Mult(\theta),$
Choose a word $\omega_n \mid z_n, \beta_{1:K} \sim Mult(\beta\, z_n)$

Each entry in $\beta_{1:K}$ is a distribution over words indicating a topic definition. The mixing proportion over topics θ is drawn from a Dirichlet's prior with a parameter α role is to create a smoothed topic distribution.

$$P(d|\alpha,\beta)=\int_{\theta}P(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_k}P(z_k|\theta)P(w_n|z_k,\beta)\right)\mathrm{d}\theta \quad (2)$$

An LDA model trained on a document yields two sets of parameters, $P(\omega|Z_{1:k})$ the word probabilities for given topics and $P(Z_{1:k}|d)$,the topic proportions for each document and its completely dependent on that document too. For a longer way the corpus determines the predictive probabilities for a unseen document as,

$$\mathrm{p}(\square|\mathrm{d}_{\mathrm{new}}) \approx \sum_{k=1}^{K} P(\omega \mid z_k)\frac{\gamma_k}{\sum_{j=1}^{K}\gamma_j} \qquad (3)$$

Where $P(\omega|Z_{1:k})$ are word probabilities over topics $Z_{1:k}$ learned during model training

E. Image Annotation Model

Along with former image annotations a conditional model is given with an image I and a set of keywords *W*, and must find the subset $W_I (W_I \subseteq W)$ which actually describes image I as follows:

$$W_I^{*}=arg\ \max_{W} P(W \mid I) \qquad (4)$$

Since image I and document D are given jointly as a set of visual and textual terms , it can be obviously result can be simplified and given as a mixed document d_{Mix} directly in estimating the conditional probabilities $\mathrm{P}(\omega_{\mathrm{t}}|\mathrm{I,D})$:

$$\mathrm{P}(\omega_{\mathrm{t}}|\mathrm{I,D}) \approx P(\omega_{\mathrm{t}} \mid d_{Mix}) \qquad (5)$$

For an unseen image-document pair , it is also possible to approximately conclude the topic proportions as $P(Z_{1:k} \mid dMix_{new})$ with (3).Fairly for an unseen image d_I and its follower document d_D , the estimated topic proportions are completely based on variational inference with an approximate algorithm. The topic proportions are further smoothed with probabilities based on each modalities.

$$\mathrm{P}^{*}(\mathrm{z}_{1:k}|\mathrm{d}_{\mathrm{Mix}}) \approx \mathrm{q}_1\mathrm{P}(\mathrm{z}_{1:k}|\mathrm{d}_{\mathrm{Mix}})+\mathrm{q}_2\mathrm{P}(\mathrm{z}_{1:k}|\mathrm{d}_{D})+\mathrm{q}_3\mathrm{P}(\mathrm{z}_{1:k}|\mathrm{d}_{I}) \ (6)$$

Where $\mathrm{P}(\mathrm{z}_{1:k}|\mathrm{d}_{D})$ and $\mathrm{P}(\mathrm{z}_{1:k}|\mathrm{d}_{I})$ are to be inferred on d_D and d_I respectively and q_1, q_2, q_3 are smoothing parameters. q_3 is a shorthand for $(1\text{-}\mathrm{q}_1\text{-}\mathrm{q}_2)$.This probabilistic topic model formulation is highly advantageous when generating captions for images in which both content selection and surface realization can be easily integrated.

F. Caption Generation Model

Extractive caption generation-the idea is to create a summary simply by identifying and subsequently concatenating the most important sentences in a document. For the task of caption generation only extraction of a single sentence is needed and the sentence must be maximally similar to the description keywords generated by the annotation model. Given the probabilistic nature of image annotation model, the content of an image is represented in two ways, i.e., as a ranked list of keywords and as a distribution of topics.

G. Word Overlap-Based Sentence Selection

The best way of measuring the similarity between image keywords and document sentence is word-overlap

$$Overlap(W_I,S_d)=\frac{|W_I \cap S_d|}{|W_I \cup S_d|} \qquad (7)$$

Here W_I is the set of keywords suggested by image annotation model and $\mathrm{S_d}$ is a sentence in the document. The sentence with the highest overlap with the image keywords is been selected as caption.

H. Vector Space-Based Sentence selection

Keywords of the image and sentences in the document is represented in vector-space and similarity between the two is been computed. The word-sentence co-occurrence matrix is been created in which each row represents a word and each column represents a sentence.The matrix cells are weighted by *tf**idf*.

$$\mathrm{sim}(\left(\overrightarrow{W_I},\ \overrightarrow{S_d}\right)=\frac{\overrightarrow{W_I}\,\overrightarrow{S_d}}{\overline{|W_I||S_d|}} \qquad (8)$$

I. Topic-Based Sentence Selection

The probabilistic topic model generated is made up of images and documents given as a bag of visual words and textual words and is represented as a distribution over a set of latent topics. The difference between two distributions *p* and *q* for sharing same topic distribution is measured by some mathematical divergence. The Kullback-Leibler (KL) divergence may be asymmetric at times. It is better to use Jensen-Shannon (JS) divergence to calculate average distance between two distributions as,

$$\mathrm{JS}(p,q)=\frac{1}{2}\left[KL\left(p,\frac{(p+q)}{2}\right)+KL\left(q,\frac{(p+q)}{2}\right)\right] \qquad (9)$$

J. Abstractive caption generation

Although extractive method yield grammatical captions and require relatively little linguistic analysis. There is often no single sentence in the document that uniquely describes the image's content also the keywords are found in the document but placed across multiple sentences. Those selected sentences make for long captions, which are not concise and overall not as catchy as human-written captions. For these reasons, abstractive caption generation is done and present models based on single words but also phrases.

K. Word-Based Caption Generation

Content Selection is modeled as the probability of a word appearing in the headline given that the same word appears in the corresponding document and is independent of other words in headline. An adaptive language model is been used that modifies an *n-gram* model with local unigram probabilities Word-Based Caption Generation.

Content Selection is modeled as the probability of a word appearing in the headline given that the same word appears in the corresponding document and is independent of other words in headline.

An adaptive language model is been used that modifies an *n-gram* model with local unigram probabilities

$$P(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} P(w_i \in C | I, \mathcal{D}) \cdot P(len(C) = n) \cdot \prod_{i=3}^{n} P(w_i | w_{i-1}, w_{i-2}), \qquad (10)$$

The length of the component is modeled as a normal distribution, which in turn modulates the caption length

L. Phrase-Based Caption Generation

A Bag-of-Phrase model is defined by modifying the content selection and caption length component. Since phrases are longer and their combinations are subjected to the topic model being formed the *adjacency* constraints are considered for estimating the probability of the phrases P_j which is been attached to the right of the phrase P_i. After integrating the adjacency probabilities caption generation model becomes as follows,

$$P(\rho_1, \rho_2, \ldots, \rho_m) \approx \prod_{j=1}^{m} P(\rho_j \in C | \rho_j \in \mathcal{D}) \qquad (11)$$
$$\cdot \prod_{j=2}^{m} P(\rho_j | \rho_{j-1})$$
$$\cdot P(len(C) = \sum_{j=1}^{m} len(\rho_j))$$
$$\cdot \prod_{i=3}^{\sum_{j=1}^{m} len(\rho_j)} P_{adap}(w_i | w_{i-1}, w_{i-2}).$$

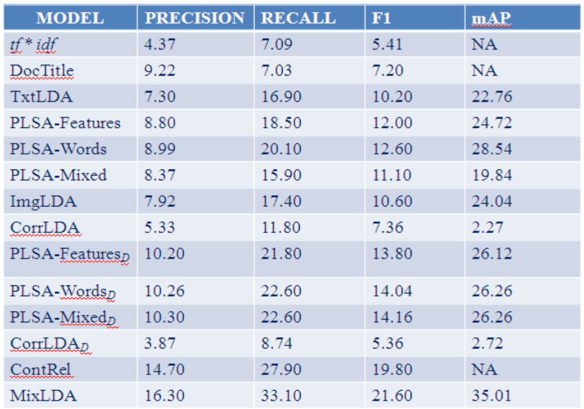
This model which is been formed takes much more long distance dependency into consideration and will generate captions invariable of length.

III. EVALUATION METHOD

An unannotated image *I* with its associated document is considered and keywords are automatically produced for those images by the model created. Model performance is evaluated by measures like *Precision* which is the percentage of correctly annotated words over all annotations that the system suggested. *Recall* is the percentage of correctly annotated words over the number of genuine annotations in the test data.*F1* is the harmonic mean of precision and recall. Mean Average Precision (mAP) is the mean of average precision of set of queries. These measures are compared with the created comparative models in Table1. over all annotations that the system suggested. *Recall* is the percentage of correctly annotated words over the number of genuine annotations in the test data.*F1* is the harmonic mean of precision and recall.Among the trained models the formed MixLDA outperforms other than all models. With these as an initiative the probability, P<*0.01* is further trained to compete with human created gold standard captions and the keywords are extracted and displayed in Table2 and captions generated are shown in Table 3.

Table I

SCORES REPORTED AS PERCENTAGE AS PLSA AND CORRLDA MODELS TRAINED ON IMAGE – CAPTION – DOCUMENT

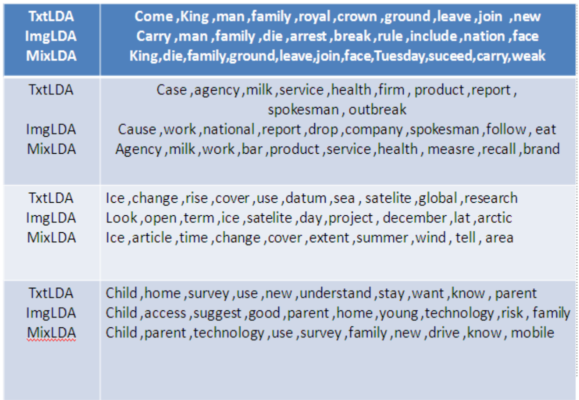


MODEL	PRECISION	RECALL	F1	mAP
tf idf*	4.37	7.09	5.41	NA
DocTitle	9.22	7.03	7.20	NA
TxtLDA	7.30	16.90	10.20	22.76
PLSA-Features	8.80	18.50	12.00	24.72
PLSA-Words	8.99	20.10	12.60	28.54
PLSA-Mixed	8.37	15.90	11.10	19.84
ImgLDA	7.92	17.40	10.60	24.04
CorrLDA	5.33	11.80	7.36	2.27
PLSA-Features$_D$	10.20	21.80	13.80	26.12
PLSA-Words$_D$	10.26	22.60	14.04	26.26
PLSA-Mixed$_D$	10.30	22.60	14.16	26.26
CorrLDA$_D$	3.87	8.74	5.36	2.72
ContRel	14.70	27.90	19.80	NA
MixLDA	16.30	33.10	21.60	35.01

PLSA-Words trained on image –caption tuples is far the best model in the comparison made with other models.

Table 2

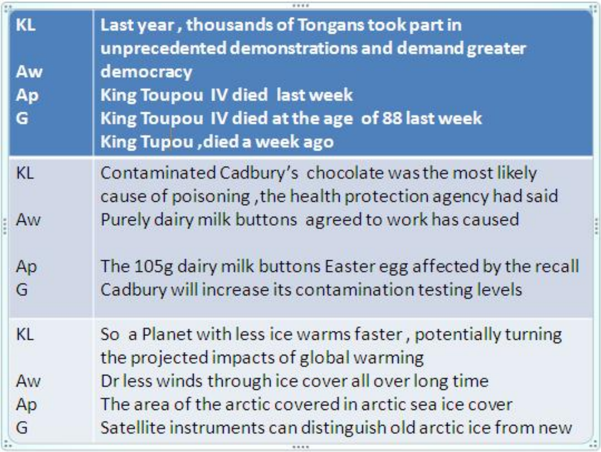
IMAGE ANNOTATIONS CREATED WITH EXACT MATCH IN THE DOCUMENT IN GOLD STANDARDS



TxtLDA ImgLDA MixLDA	Come ,King ,man ,family ,royal ,crown ,ground ,leave ,join ,new Carry ,man ,family ,die ,arrest ,break ,rule ,include ,nation ,face King,die,family,ground,leave,join,face,Tuesday,suceed,carry,weak
TxtLDA	Case ,agency ,milk ,service ,health ,firm , product ,report , spokesman , outbreak
ImgLDA	Cause ,work ,national ,report ,drop ,company ,spokesman ,follow , eat
MixLDA	Agency ,milk ,work ,bar ,product ,service ,health , measre ,recall ,brand
TxtLDA	Ice ,change ,rise ,cover ,use ,datum ,sea , satelite ,global ,research
ImgLDA	Look ,open ,term ,ice ,satelite ,day ,project , december ,lat ,arctic
MixLDA	Ice ,article ,time ,change ,cover ,extent ,summer ,wind , tell , area
TxtLDA	Child ,home ,survey ,use ,new ,understand ,stay ,want ,know , parent
ImgLDA	Child ,access ,suggest ,good ,parent ,home ,young ,technology ,risk , family
MixLDA	Child ,parent ,technology ,use ,survey ,family ,new ,drive ,know , mobile

Table 3

CAPTIONS GENERATED BY MODEL CREATED



KL Aw Ap G	Last year , thousands of Tongans took part in unprecedented demonstrations and demand greater democracy King Toupou IV died last week King Toupou IV died at the age of 88 last week King Tupou ,died a week ago
KL Aw Ap G	Contaminated Cadbury's chocolate was the most likely cause of poisoning ,the health protection agency had said Purely dairy milk buttons agreed to work has caused The 105g dairy milk buttons Easter egg affected by the recall Cadbury will increase its contamination testing levels
KL Aw Ap G	So a Planet with less ice warms faster , potentially turning the projected impacts of global warming Dr less winds through ice cover all over long time The area of the arctic covered in arctic sea ice cover Satellite instruments can distinguish old arctic ice from new

IV. CONCLUSION

The novel task of automatic caption generation for news images is introduced. The task fuses insights from computer vision and natural language processing such as image and video retrieval, development of tools supporting news media management. The results show that it is possible to learn a caption generation model from weakly labeled data without costly manual intervention.

REFERENCES

[1] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth, "*Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary,*" Proceedings Seventh European Conf. Computer Vision, pp. 97-112, 2002

[2] S. Feng, V. Lavrenko, and R. Manmatha, "*Multiple Bernoulli Relevance Models for Image and Video Annotation*," Proceedings IEEE Conf. Computer Vision and Pattern Recognition, pp. 1002-1009, 2004.

[3] Y.Feng,M.Lapata, "*Automatic caption generation for news images*" IEEE Trans. on Pattern analysis and machine Intelligence, vol. 35, no. 4, pp. 797-812, 2013.

[4] L. Ferres, A. Parush, S. Roberts, and G. Lindgaard, "*Helping People with Visual Impairments Gain Access to Graphical Information through Natural Language: The igraph System*," Proc.,11th Intl' Conf. Computers Helping People with Special Needs,Springer pp. 1122-1130, 2006.