IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 2, Issue 2, Apr-May, 2014
ISSN: 2320 – 8791(Impact Factor: 1.479)
www.ijreat.org

Analyzing The Performance Of Automatically Captioned Images Using Probabilistic Approach

Pondsingh Jackie Shrine G[1], Sravan Yadav Eadala[2]

[1]PG Scholar, Department of Computer Science, SRM University, Chennai, India

[2]Assistant Professor, Department of Computer Science, SRM University, Chennai, India

Abstract

Images are automatically captioned in two stages, Content Selection and Surface Realization. Content Selection is based on probabilistic model that suggests the keywords for the image and description in the Article. Surface realization techniques which determines how to verbalize the chosen keywords. In Content Selection the keyword assumption is based on probabilistic image annotation model that the images and the surrounding text are generated by a latent variables or topics. The surface realization emphasizes the caption generation in extractive and abstractive methods. Whereas in extractive method a sentence is generated from the keywords obtained through the former method discussed above, the abstractive method creates the caption which may be from the word based or phrase based caption generation approach. So the idea here is to use phrases for surface realization. The approach is to analyze the performance of automatically captioned images using probabilistic approach. Indeed the output of the abstractive model is obtained using phrases rather than words.

Keywords: *caption, content selection, surface realization, annotation model, probabilistic approach, Latent Dirichlet Allocation.*

1. Introduction

The growth of Digital information is going as far from our imagination on the internet. Many websites publish images with their descriptions. Pictures in large scale and heterogeneous collections will be a overhead to the search engine for information retrieval. Search engines usually retrieve the images without analyzing the content of the image. Images become more informative when it's being annotated, captioned and the text surrounding the image. The literature shows various attempts to learn the relationship between the images and the words using supervised classification techniques [1],[2] and latent variable models [4], [5],[6] and models for information retrieval [7],[8]. Keyword based information retrieval are the popular one although the use of more linguistically meaningful descriptions have a better good reasons for image retrieval engines. Keywords can be ambiguous, an image which is annotated with words like Tiger, lake, deer could depict a tiger holding a deer near the lake of water which would the relation between the words explicit.

2. Image And Document Representation

2.1 System Design

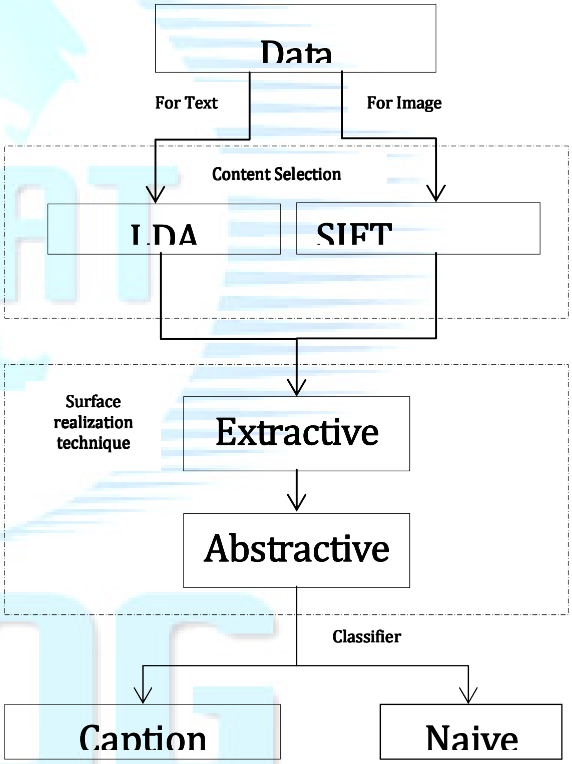

Fig: 1 System Design

2.2 Scale Invariant Feature Transform

As Fig:1depicts extracting the features from the image is one of the fundamental aspects of many problems in computer vision, the image features that may have many properties that make them suitable for matching images. The features are invariant to image scaling and rotation, the features are unique, which allows a single feature to be correctly matched with high probability against a large database of unique features, providing a basis for object and scene recognition [9]. To generate the set of image features several steps are there to be followed.

1) Scale-space extrema detection

Initially search is over all scales and image locations. It is implemented using a difference-of-Gaussian function to identify potential interest points that are invariant to scale and orientation.

2) Keypoint localization

At each object location, a detailed model is fit to determine location and scale. Then the keypoints are selected based on their measures of their stability.

3) Orientation assignment

One or more orientations are assigned to each keypoint location based on local image gradient directions. All operations for the future are performed on image data that has been transformed relative to the defined orientation, scale, and location for each feature, thereby providing invariance to all these transformations.

4) Keypoint descriptor

The local image gradients are measured at the selected scale in the region around each and every keypoint. These keypoints are then transformed into a representation that allows for significant levels of local shape distortion and change in illumination.

2.3 Latent Dirichlet Allocation

The representation of LDA can be as a three-level hierarchical Bayesian model. As shown in the Fig:1 Given a corpus consisting of D documents, each document is modeled using a mixture over K topics (which is assumed to follow a multinomial distribution with a Dirichlet prior), which are characterized as distributions over words. Probably the words in the document are generated by sampling repeatedly a topic according to the topic distribution, and selecting a word given with the chosen topic.

Blei et al. [10] describe the generative process for a document d as follows:

1. Choose $\theta|\alpha$: $dir(\alpha)$,
2. For $n\in 1,2....,N$:
   a. Choose topic $z_1|\theta$: $Mult(\theta)$,
   b. Choose a word $w_n|z_n,\beta_{1:k}$: $Mult(\beta_{z_n})$,

Where each entry of $\beta_{1:k}$ is a distribution over words which may indicate the topic definitions. The mixing proportion over topics θ is taken from the prior distribution with holds the parameters α whose role is to create a smoothed topic distribution. After α and β are sampled then each document is generated according to the topic proportions $z_{1:k}$ and word probabilities over topics β. The probability of a document *d* in a corpus can be obtained as follows

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

2.4 Extractive Caption Generation

Extractive caption generation method draws inspiration from previous work on automatic summarization, most of which focuses on extracting a sentence from a paragraph (see [11] and [12] for comprehensive overviews). The aim is to create a summary simply by identifying and subsequently concatenating the most important sentences in a document. From the Fig:1 without a great deal of linguistic analysis, there is a possibility of creating summaries for a wide range of documents and texts, which is independent of style, text type, and subject matter. In our model we need to extract only a single sentence. And important that this sentence must be maximally similar to the description keywords generated by the SIFT model. We are able to represent the content of an image in two ways, such as a ranked list of keywords and as a distribution of topics. There are different ways of the operational similarity between a sentence and each of these representations. Hence the word Overlap-Based Sentence Selection is the most intuitive way of measuring the similarity between image keywords and document sentences is word overlapping,

$$Overlap(W_I \cap S_d)=\frac{|W_I \cap S_d|}{|W_I \cup S_d|} \qquad (2)$$

Where w_I the set of keywords is suggested by our image annotation model and S_d a sentence in the document. The selected caption is then the sentence that has the highest overlap with the image keywords.

3. Sentence Selection Based On Topic

The probabilistic topic model with images and documents rendered into a bag of visual and textual words and represented as a probability distribution over a set of latent topics. Here, the similarity between word and a sentence can be broadly measured by the extent to which they share the same topic distributions [13]. For example, we may use the Kull back-Leibler divergence to Measure the difference between two distributions

$$KL(p,q)=\sum_{j=1}^{k} p_j \log_2 \frac{p_j}{q_j} \qquad (3)$$

Where p and q are the parameter for the image topic distribution P_d Mix and sentence topic distribution S_d, respectively. We obtain the image topic distribution according to the mixed document (using both the image and the document). When doing inference on the document sentence, we also take its neighboring sentences into account to avoid estimating the topic proportions on short sentences inaccurately. The KL divergence is asymmetric and, in many applications, it is preferable to apply a symmetric measure such as the Jensen Shannon (JS) divergence. The latter measures the "distance" between p and q through the average of p and q are as follows:

$$JS(p,q)=\frac{1}{2}\left[KL\left(\frac{(p+q)}{2}\right)+KL\left(q,\frac{(p+q)}{2}\right)\right] \qquad (4)$$

3.1 Abstractive Caption Generation

Although extractive methods yield naturally grammatical captions and require relatively little linguistic analysis, there is often no single sentence in the document that uniquely describes the image's content. In most of the cases the keywords are found in the document but interspersed across multiple sentences. Here the selected sentences make for long captions which are not concise and overall not as catchy as human-written captions. Considering these reasons, we are using the abstractive caption generation and present models based on Phrase-Based Caption Generation. The model specified in [14] will generate captions with those function words. Anyway there is no guarantee that these will be compatible with their surrounding context . To avoid these problems, we turn our attention to phrases which are naturally associated with function words capture long-range dependencies. In which phrases have been previously used in abstractive summarization. For example, Zhou and Hovy [15] first identify a list of keywords which are then used to extract phrases from the document and the phrases are linked together to create headlines using a set of handwritten rules. Based on this approach, Soricut and Marco [16] identify a list of keywords but also use syntactic information (extracted from parse trees) to build syntactically driven phrases around the extracted keywords. Finally, Wan et al. [17] extract dependencies from the input document and join them together using n-grams. The selection of content from individual words to phrases poses additional difficulties for surface realization which is based on language models are typically built from individual words rather than phrases and as a result they do not take phrase adjacency constraints into account. This model relies on phrases which we obtain from the output of a dependency parser whereas a phrase is simply a head and its dependents except verbs, where we record only the head (otherwise, an entire sentence could be a phrase).The dependencies we consider the heads are nouns, verbs, and prepositions, as these constitute 80 percent of all dependencies contained in our caption data. We define a bag-of-phrases model for caption generation by modifying the content selection and caption length components as follows:

$$P(\rho_1,\rho_2,\ldots,\rho_m)\approx\prod_{j=1}^{m}P(\rho_j\in C|\rho_j\in D) \qquad (5)$$

$$\cdot P\left(len(c)=\sum_{j=1}^{m}len(\rho_j)\right)$$

$$\cdot\prod_{i=3}^{L}P_{adap}(w_i|w_{i-1},w_{i-2})$$

Where $L=\sum_{j=1}^{m}len(\rho_j)$. The term P(ρ$_j$ЄC|ρ$_j$ЄD) models the probability of phrase ρ$_j$ appearing in the caption given that it also appears in the document and is estimated as

$$P(\rho_j\in C|\rho_j\in D)=\prod_{w_j\in\rho_j}\rho(w_j\in C|w_j\in D) \qquad (6)$$

Where w$_j$ is a word in the phrase ρ$_{j.}$ One problem with the models discussed thus far is that words or phrases are independent of each other. It is up to the trigram model to enforce coarse ordering constraints. We therefore attempt to take phrase adjacency constraints into account by estimating the probability of phrase j attaching to the right of phrase i as

$$P(\rho_j|\rho_i)=\sum_{w_i\in\rho_i}\sum_{w_j\in\rho_j}p(w_j|w_i)$$

$$=\frac{1}{2}\sum_{w_i\in\rho_i}\sum_{w_j\in\rho_j}\left\{\frac{f(w_i,w_j)}{f(w_i,-)}+\frac{f(w_i,w_j)}{f(-,w_j)}\right\} \qquad (7)$$

Where *f* (w$_{i,}$ w$_j$) is the probability of a phrase containing word w$_j$ appearing to the right of a phrase containing word w$_i$, f(w$_i$, w$_j$) indicates the number of times two phrases containing w$_i$ and w$_j$ are adjacent, f(w$_i$) Þ is the number of times w$_i$ appears on the left of any phrase, and f(-,w$_i$) the number of times it appears on the right. After integrating the adjacency probabilities into (22), the caption generation model becomes

$$P(\rho_1,\rho_2,\ldots..,\rho_m)\approx\prod_{j=1}^{m}P(\rho_j\in C|\rho_j\in D) \qquad (8)$$

$$\cdot\prod_{j=2}^{m}P(\rho_j|\rho_{j-1})$$

$$\cdot P\left(len(C)=\sum_{j=1}^{m}len(\rho_j)\right)$$

$$\cdot\prod_{i=3}^{\sum_{j=2}^{m}len(\rho_j)}P_{adap}(w_i|w_{i-1},w_{i-2})$$

The model in (25) takes long distance dependency constraints into account and has some notion of syntactic structure through the use of attachment probabilities. As it has a primitive notion of caption length estimated by

$$P\left(len(c)=\sum_{j=1}^{m}len(\rho_j)\right) \qquad (9)$$

it will invariably generate captions of similar (phrase) length.

4. Evaluation Method

Our evaluation followed the experimental methodology proposed by toon calders [19]. We are given an image I with its associated document and asked to automatically produce suitable keywords for it. We consider the keywords are assigned to the features of the image as the annotations for image I and compare them against the standard captions. The bag of words obtained from the above model is treated with the Natural Language Processing technique and caption is generated. Model performance is evaluated using a Naïve Bayes classifier where precision, recall, and F1. In the image annotation task, precision 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. These measures are averaged over all items in the test set. In addition to F1, we also report Mean Average Precision (mAP), an evaluation measure commonly used in information retrieval. Mean average precision is the mean of the Average Precision (AP) of a set of queries. The AP of a query is the average of the precision scores at the rank locations of each relevant document (or image in our case).

5. Conclusions

In this paper, we introduced the method for automatic caption generation for images with its descriptions. The task fuses insights from computer vision and natural language processing and it may promise for various multimedia applications, such as image and video retrieval and for individuals with visual impairment. We have approached this task by leveraging the vast resource of images available on the Internet and exploiting the fact that many of these co-occur with textual information (i.e., captions and text documents). So it is possible to learn a caption generation model from weakly labeled data without costly manual involvement. The dataset that contains real-world images and exhibits a large vocabulary including both object names and abstract keywords; instead of manually creating annotations, image captions are treated as labels for the image. A key aspect of our approach is to allow both the visual and textual modalities to influence the generation task. This is achieved through an image annotation model that characterizes pictures in terms of description keywords that are subsequently used to guide the caption generation process. The abstractive model defined over phrases yields promising results. It generates captions that are more grammatical than a closely related word-based system and manages to capture the list of the image (and document) as well as the captions written by humans.

Our caption generation model follows a two-stage approach where the image is processed and then text are carried out sequentially. Indeed, an avenue for future work would be to define a phrase-based model for both image annotation and caption generation, e.g., by exploiting recent work in detecting visual phrases (e.g., [18]). We also believe that our approach would benefit from more detailed linguistic and nonlinguistic information. However, our future work could improve grammaticality more globally by generating a well-formed tree (or dependency graph).

Acknowledgments

The authors are grateful to express sincere thanks and gratitude to our Professor and HOD, Dr.E.Poovammal, Department of Computer Science and Engineering, SRM University for her encouragement and the facilities that were offered to me for carrying out this project. And we take this opportunity to thank our Director, Dr.C.Muthamizchelvan, Ph.D., Faculty of Engineering and Technology, SRM University for providing us with excellent infrastructure that is required for the development of our project.

References

[1] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang, “Image Classification for Content-Based Indexing,”IEEE Trans. ImageProcessing,vol. 10, no. 1, pp. 117-130, 2001.

[2] A.W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain,“Content-Based Image Retrieval at the End of the Early Years,”IEEE Trans. Pattern Analysis and Machine Intelligence,vol. 22, no. 12,pp. 1349-1380, Dec. 2000.

[3] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth, “ObjectRecognition as Machine Translation: Learning a Lexicon for aFixed Image Vocabulary,”Proc. Seventh European Conf. ComputerVision,pp. 97-112, 2002.

[4] D. Blei, “Probabilistic Models of Text and Images,” PhDdissertation, Univ. of Massachusetts, Amherst, Sept. 2004.

[5] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M.Jordan, “Matching Words and Pictures,” J. Machine LearningResearch,vol. 3, pp. 1107-1135, 2002.

[6] C. Wang, D. Blei, and L. Fei-Fei, “Simultaneous Image Classification and Annotation,”Proc. IEEE Conf. Computer Vision and PatternRecognition,pp. 1903-1910, 2009.

[7] V. Lavrenko, R. Manmatha, and J. Jeon, “A Model for Learning theSemantics of Pictures,” Proc. 16th Conf. Advances in NeuralInformation Processing Systems, 2003.

[8] S. Feng, V. Lavrenko, and R. Manmatha, “Multiple BernoulliRelevance Models for Image and Video Annotation,”Proc. IEEEConf. Computer Vision and Pattern Recognition,pp. 1002-1009, 2004.

[9] D. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,”Int'l J. Computer Vision,vol. 60, no. 2, pp. 91-110, 2004.

[10] D. Blei, A. Ng, and M. Jordan, “Latent Dirichlet Allocation,” J. Machine Learning Research,vol. 3, pp. 993-1022, 2003

[11] K. Sparck Jones, “Automatic Summarizing: Factors and Directions,” Advances in Automatic Text Summarization,I. Mani andM.T. Maybury, eds., pp. 1-33, MIT Press, 1999.

[12] I. Mani,Automatic Summarization.John Benjamins Publishing Co., 2001.

[13] M. Steyvers and T. Griffiths, “Probabilistic Topic Models,”A Handbook of Latent Semantic Analysis,T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, eds. Psychology Press, 2007.

[14] A. Makadia, V. Pavlovic, and S. Kumar, “Baselines for Image Annotation,”Int'l J. Computer Vision, vol. 90, no. 1, pp. 88-105, 2010

[15] L. Zhou and E. Hovy, “Headline Summarization at ISI,”Proc. HLT-NAACL Text Summarization Workshop and Document Understanding Conf.,pp. 174-178, 2003.

[16] R. Soricut and D. Marcu, “Stochastic Language Generation Using WIDL-Expressions and Its Application in Machine Translation and Summarization,”Proc. 21st Int'l Conf. Computational Linguistics and the 44th Ann. Meeting Assoc. for Computational Linguistics, pp. 1105-1112, 2006

[17] S. Wan, R. Dale, M. Dras, and C. Paris, “Statistically Generated Summary Sentences: A Preliminary Evaluation of Verisimilitude Using Precision of Dependency Relations,”Proc. Workshop Using Corpora for Natural Language Generation,2005

[18] A. Sadeghi and A. Farhadi, “Recognition Using Visual Phrases,” Proc. IEEE Conf. Computer Vision and Pattern Recognition,pp. 1745-1752, 2011.

[19] Naïve Bayes classifier Evaluation , Toon Calders (t.) http://wwwis.win.tue.nl/~tcalders/teaching/datamining09/slides/DM09-02-Classification.pdf