the same time, we modify our surface realization component so that it takes note of the image annotation probabilities. Intuitively, we hope the new language model will prefer words that have high image annotation probabilities while are likely to appear in a sentence according to the background language model.

$$P(w1, w2, \dots, wn) = \prod_{i=1}^{n} P(wi \in H | wi \in D) \cdot P(len(H) = n) \cdot \prod_{i=2}^{n} P(wi \mid wi-1) \quad \dots\dots\dots\dots(2)$$

where wi is a word that may appear in headline H, D the document being summarized and P(len(H)=n) a headline length distribution model. The above model can be easily adapted to our image caption generation task. Content selection is now the probability of a word appearing in the caption given the image and its associated document which we obtain from the output of our image annotation model.

$$P(w1, w2, \dots, wn) = \prod_{i=1}^{n} P(wi \in C | I, D) \cdot P(len(C) = n) \cdot \prod_{i=3}^{n} P(wi \mid wi-1, wi-2) \quad \dots\dots\dots(3)$$

Where C is the caption, I the image, D the accompanying document.

6. Implementation Details

1) Input image and article: The image and the article are the inputs
2) Content Extraction: Only relevant content is extracted.
3) Summarization : Abstractive summarization is benefitial to create words or phrases for the entire article with its respective image.
4) Text annotation model: Finds out important keywords and phrases using parts of speech.
5) Image annotation model: learns to automatically label images under assumption that the images and the surrounded text are shared set of latent variables.Our annotation model takes these topic distributions into account by searching frequrnt keywords for that image and its relevant document.I have compared LDA and its alternative perceptron.And as per the theory, it is found that LDA is best suitable.Also I have compared 3 different types of LDA namely, word overlap, standard vector space model, and Txt LDA and finally found that mix LDA is better.Also I can use here SIFT algorithm, which is a part of LDA.
6) Caption Generation: After comparing extractive and abstractive caption generation processes I conclude that abstractive is better one since it provides a word based or phrase based caption whereas, extractive results in a single sentence and there is a probability of generating a single sentence that may not clearify the whole article.Also they are not concise and they are not catchy.
   a) Word based : Content selection is used as the probability of a word appearing in the headline given that the similar word occurs in the corresponding document.Surface realizations also take the length of a caption into account to generate output of reasonable length.
   b) Phrase based : In word based generation process there is no guarantee that the caption will be compatible.to overcome this I need phrases to capture long-range dependencies.

Figure.1 Proportion of caption words given a rating by human judges.

7. Evaluation

As this paper explores the feasibility of automatic caption generation for images in news domain it is having some limitations like images that don't co-inside with textual data can't be retrieved. We need to follow some evaluation techniques to get oue work done without any manual involvement and human authored grammar. At testing time we will input 1 news articles and image, then system will generate caption. 1) Image Database: Database created with huge amount of images & articles. The dataset also covers wide range of topics including national and international topics related to science , sports, politics, technology, etc.

1) Content selection : Identifies important keywords from text and images. Text frequency calculator is used to extract these keywords which will also find out the importance of that keyword in that article. One word is read at a time and accordingly hash table is built.
2) Surface Realization: Verbalize the keywords extracted from text & image.
3) Stemming process is used to reduce the variant forms of a word into a common single word.
4) Stop word removal is used to remove most frequently occurring meaningless words.
5) SIFT: Describe local features in an image.

8. Result

In this paper I have explained the possibility of generating captions automatically for particularly news domain because the dataset for news domain is easily available.Image annotation model converts image features into self understandable keywords which are frequently used to guide the generation process.Finally sentence is generated by combining phrases generated by abstractive method. Following image and respective article is taken from the standard dataset:

Image:

Article

They denounced the government violation of a May peace agreement and expressed concern for the safety of civilians. This is the first time since the peace deal that the Sudanese army has been confirmed as fighting in Darfur. The attack was against the rebel Justice and Equality Movement, (JEM) which did not sign the deal. Some 2m people have fled their homes in Darfur since conflict began in 2003, and tens of thousands of people are reported to have been killed in ongoing violence. Both government and militia troops had been observed massing near the western town of Geneina before the attack on Friday. An assortment of armed groups that remained outside of the peace agreement, including Chadian elements, are known to be taking shelter in the Jebel Moon mountains. The attack is confirmation that Darfur's conflict has changed in nature, the BBC's Jonah Fisher reports from Sudan. The two signatories to the peace agreement - the government and the Minni Minnawi faction of the Sudan Liberation Movement - are using the agreement as a springboard to attack those outside the deal. The SLM Minnawi has launched a wide-ranging offensive against their former rebel allies and supporters, leaving at least 80 people killed and thousands displaced. JEM spokesman Ahmed Hussein Adam said the Sudanese government was systematically attacking groups who had refused to sign for peace. SLM Minnawi was the main rebel group in the Darfur conflict, and the only one that signed the deal on 5 May with the Sudanese government.

9. Conclusion

I have introduced the task of automatic caption generation for news domain. This concept virtualizes the inheritance of computer vision and NLP. Already it is available on internet resource to retrieve images with respect to the user queries. Here the results shows caption generated by weakly labeled data without any costly manual involvement. Captions are treated as labels for images. These captions can be used to learn the correspondances between textual and visual modalities and act as a gold standard. We built our dataset from resources that are publicly available on the internet without manual post processing.

10. Future Scope

A news in the form of video involves key frames from streaming video data. So instead I can use this strategy for video purpose. Currently I treat the image regions or ROI as bags of words, which could be extended to bigrams according to their spatial relations. Also I could however, improve grammaticality more globally by generating a dependency graph. In this work, images were preprocessed by extracting primarily local feature representations (e.g., color, texture, corners, SIFT features, etc.), without considering more global representations, such as spatial relationship among different regions. An obvious extension would be taking spatial information into account when dealing with image representations. Currently, we treat the image regions or detected regions of interest as bags-of-words, which could be extended to bigrams according to their spatial relations.

The image caption generation task has been formulated as a two-step approach, where the image content extraction and caption generation are carried out sequentially. A more general model should integrate the two steps in a unified framework. Indeed, an avenue for future work would be to define a phrase-based model for both image annotation and caption generation.

References

[1] Yansong Feng, Member, IEEE, and Mirella Lapata, Member, IEEE "*Automatic Caption Generation for News Images*,"IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL.35, NO. 4, APRIL 2013
[2] D. Lowe, "*Object Recognition from Local Scale-Invariant Features*,"Proc. IEEE Int'l Conf. Computer Vision, pp. 1150-1157,1999.
[3] P. He´de, P.A. Moe¨llic, J. Bourgeoys, M. Joint, and C. Thomas, "*Automatic Generation of Natural Language Descriptions for Images*," Proc. Recherche d'Information Assiste´e par Ordinateur,2004.
[4] A. Farhadi, M. Hejrati, A. Sadeghi, P. Yong, C. Rashtchian, J.Hockenmaier, and D. Forsyth, "*Every Picture Tells a Story: Generating Sentences from Images*," Proc. 11th European Conf. Computer Vision, pp. 15-29, 2010.

Author Profile

Kalyani Zinjurde received the B.E degree in Computer Engineering from PESCOE in 2010 and persuing M.E degree in computer Engineering ,Working as a lecturer in DIETMS, Aurangabad (Mah.), India.