An NLP Approach to Image And Image Captioning

Rohit Kumar
UMBC / Computer Science
rohitk1@umbc.edu

Abstract

Image indexing, Image captioning and Image retrieval is a hot topic of research in NLP and Computer Vision now-a-days, but the work on it have started decades ago. The context of the image captions matters a lot while talking of image processing. Presence of a name/text in a caption related to a given context provides useful information concerned with the image like who is associated with given image. Earlier work in the text and image processing have shown that both image and text are related to each other and thus are complementary. This paper discusses this image-text relationship along with the application of language model to image captioning and how it has helped to improve the accuracy in the image processing field. This paper also discusses automatic image captioning and retrieval of a picture using joint probability distribution for a word and feature function extracted for various regions of the image. Paper ends with discussing and comparing another approach for a captioned photographs using relational facts.

1 Introduction

Statistical Natural language processing has helped a lot in learning the semantics of image databases using image and text. It has been observed and a fact proved by K. Barnard and D.A. Forsyth (2001) that images and their linked annotated words are related to each other. We can see in K. Barnard and D.A. Forsyth (2001), how museum collections have been used to suggest/retrieve images for given captions and to generate words for any random images. This complimentary property of words and images have been utilized for the further research and work in the image captioning fields like we see in (Berg et al. (2004), (Berg et al. (2004), Lavrenko (2003)) etc. Our paper discusses the related works and findings in the field of text and image captioning and modeling using the results and findings from the first five paper mentioned in references section below. The paper Berg et al. (2004) talks about the approach to clean a noisy supervised data having inaccurately and ambiguously labelled face images. It makes use of clustering and merging procedure as discussed in the below section to obtain the accuracy of mapping an image to its correct label and thus removing noise. This accuracy is further increased with the addition of a language model for a given context as in Berg et al. (2004) on the clustering procedure. This paper further compares and evaluates both EM and maximum likelihood clustering and produces result as shown in Berg et al. (2004) highlighting the fact that application of language model with appearance model produces better results than using appearance model alone. Our paper further talks of assumption of every image as being divided into several regions, representation of each regions by continuous-valued feature vector and its mapping to a word that can be subsequently used in captioning and retrieval of an image as in Lavrenko (2003)). Finally Paper discusses a entirely different approach for indexing and retrieving a captioned photograph using relational facts represented in the form of triples. This uses a text-based approach for the automatic indexing and retrieval of digital photographs taken at crime scenes. This approach extracts relational facts from the captions represented in the form of triples of the form: ARG1-REL-ARG2 as in Pastra (2003)). This triples are used as image indexing terms as well as image retrieval mechanism. The retrieval process calculates the simi-

larity scores between query-triples and indexing triples.

2 Basic Approach and Methodology

2.1 Hierarchical Structure and Joint Probability

The museum collection used in K. Barnard and D.A. Forsyth (2001) consists of images like line drawings, paintings, and pictures of sculpture and ceramics. Lots of these images have associated text which varies from physical description to interpretation and mood to a great scale. These pictures are organized in a way to create a model of hierarchical structure so as to expose the maximum semantic structure to the user which encourages semantics through levels of generalization. This hierarchical model is a combination of the asymmetric clustering model (which maps documents into clusters), and the symmetric clustering model (which models the joint distribution of documents and features). The leaves of the hierarchy correspond to clusters. Node has special significance in the hierarchy. Each node has some probability of generating each word as well as probability of generating an image segment with given features. The leaf nodes represent a cluster. Documents belonging to a given cluster are generated by the nodes along the path from the leaf to the root node as shown in the Figure 1. Considering all the generated clusters, a document is mapped by summing over all the clusters and supported by the probability that the document is in the cluster. As explained in the paper K. Barnard and D.A. Forsyth (2001) the process for generating the set of observations D associated with a document d can be described by in mathematical terms as

$$P(D|d) = \sum_{c} P(c) \prod_{i \in D} \left(\sum_{l} P(i|l,c) P(l|c,d) \right) \quad (1)$$

where c is the indexes for clusters, i for items (words or image segments), and l for levels. D is a set of observations that includes both words and image segments. Expectation-Maximization algorithm by Dempster (1977) is used to train this model. This model produces a joint probability distribution for words and picture elements. This procedure aims to retrieve the images for given captions and also to generate words for images outside the training set.

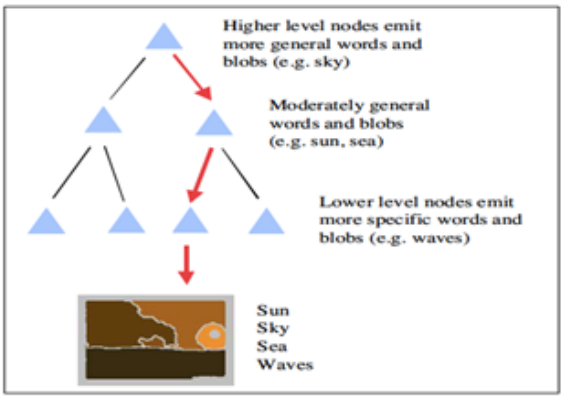

Figure 1: Taken from K. Barnard and D.A. Forsyth (2001): Illustration of the generative process implicit in the statistical model. Each document has some probability of being in each cluster. To the extent that it is in a given cluster, its words and segments are modeled as being generated from a distribution over the nodes on the path to the root corresponding to that cluster.

2.2 Clustering and Merging Approach

This approach is discussed in Berg et al.!(2004) for word and image mapping. It makes use of a data set consisting of 44,773 face images extracted out of half a million captioned news images, with a set of names associated with it (automatically extracted from the associated caption). A lot of images are there in the data set which does not contain correct name. Paper uses a clustering procedure (Modified K-Means Clustering) to identify labelling ambiguities and faces with incorrect labels. Paper discusses how the discussed approach is different from face recognition. It aims to identify discriminant coordinates which can be used to distinguish between faces, even if it is not present in data set rather than to classify the faces. Therefore it uses kPCA/LDA methodology instead of building a multi-class classifier. Kernel principal components analysis (kPCA) is used to reduce the dimensionality of data and linear discriminant analysis (LDA) is used to project data into a space suitable for discrimination task. As the data set is very large to do kPCA directly on the kernel matrix, we therefore use an approximation to calculate the eigenvectors of K. The clustering process randomly map every image to one of its extracted names. The mean of image vectors is calculated to each assigned name (which represent a cluster). Each image is reassigned to the closest mean of its

extracted names. And the process is repeated 2-3 times till convergence. The clusters with fewer than three images are removed. A merging procedure is used thereafter, that identifies all different names referring to the same individual based on images and links name. The table below shows this merging procedure proposed merges. Finally extracted data set is noise free to the extent possible, with minimum error rates and correctly labelled faces.

Proposed Merges	
President Bush	President George
Donald Rumsfeld	Defense Secretary Donald Rumsfeld
State Colin Powell	Colin Powell
President Bush	Richard Myers
President Bush	United Nations
Defense Donald Rumsfeld	Donald Rumsfeld
Venezuelan President Hugo Chavez	Hugo Chavez

Figure 2: Taken from Berg et al.(2004): Table shows multiple names often refer to the same person. Merging Procedure links names to people based on images. If two names have the same associated face, then they must refer to the same person. The above pairs are the top name merges proposed by the system. Merges are proposed between two names if the clusters referring to each name contain similar looking faces

2.3 Applying Language Model to Clustering

This section discussed about the accuracy results when a language model is applied to the clustering procedure discussed above. The aim of this experiment is to take very large collection of news images and captions as a semi-supervised input and and produce a data set of faces labeled with names. When we apply a language model as in Berg et al. (2004) based on the given context to the clustering process, we find that the result improves further. Same data set is used for this experiment as above. A face detector and named entity recognizer are used to identify potential faces and names respectively. An appearance model is created/learnt for each name using kPCA and LDA, and chances of a name mapping to pictures depends on the given context. Named entity recognizer sometimes generates words or strings like 'Winter Olympics'. These are not naming because their appearance is in quite different contexts from actual people. However, assignment of these ambiguous names to a face is unlikely because the model assigns such names a low probability. Also, in case of multiple names being assigned to a given face, probability of each name depend on their contexts. EM and maximum likelihood clustering, both are evaluated here and shows that using language model with appearance model produces better result than using appearance model alone. For a face clustering and language model, EM computes an expected set of face-name relations and connections keeps on iterating and updating the clusters and language model until there is relation between the two. The parameters of this model are P (face | name) and P (pictured | context). The probability of being pictured in case of multiple context cues can be extracted using Bayes rule as shown below in the mathematical form.

$$P(pictured|C_1, C_2, ..., C_n) = \frac{P(picturedC_1)...P(pictured|C_n)}{P(pictured)^{n-1}} \quad (2)$$

2.4 Joint Probability including Image Regions

This approach discusses the way to learn the semantics of images, and idea of automatically annotating and retrieving the image based on text queries as discussed in paper (2003). In this experiment, it is assumed that each image is segmented into several regions and each region can be represented by a feature vector. Single pixel or region is hard to interpret as it may be more biased and not the generalized one, so we take image regions. A joint probabilistic model is created for words and image features which can be used to predict the word given the image region and this can further used to achieve the objective of automatic annotation of image and subsequently its retrieval. Other way, we can say that the assumption and consideration of different regions in an image provides context while the linking of words with different image regions through probability provides meaning. This can be explained and understood in the way explained in paper (2003) as, when we tell a word tiger then the association of a region having grass or water increased and the association with the region showing the interior of an aircraft is decreased. This model does not require any clustering procedure and directly links continuous image features with words and hence does not suffer

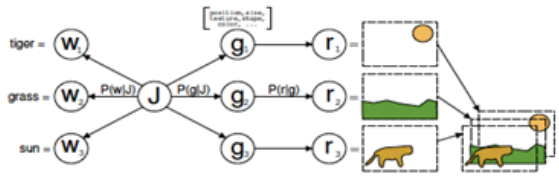

Figure 3: Taken from Lavrenko (2003): A generative model of annotated images. Word wj in the annotation are i.i.d. sampled from the underling multinomial. Image pixels are produced by first picking a set of i.i.d. feature vectors (g1,...gn), then generating image regions (r1,...rn) from the feature vectors, and finally stacking the regions on top of each other.

from the granularity issues. The model is called Continuous-space Relevance Model (CRM). The paper (2003) describes the approach to learn joint probability distribution over the regions of some image and the words in its annotation required for Image Annotation(calculating word if we know regions) and Image Retrieval (retrieving required imaged based on ranking of images in the collection according to their conditional likelihood of having the query as annotation).

2.5 Extracting Pragmatic Relation in Tuples Form

This approach discusses ways of indexing and retrieving a captioned photograph by using advanced NLP approach to extract relational facts instead of syntactical relations from captions as in Katerina (2003). Experiment includes British Police Information Technology Organisation Common Data Model and a collection of formal reports produced by scene of crime officers (SOCO) to develop OntoCrime (a structure relevant to SOC investigation like physical evidence, trace evidence, weapon). This prototype called SOCIS in the original paper, extends itself to the existing available approaches like key-word approach and creates a relational fact having 'pragmatic relation' between objects depicted in the photographs, thus eliminating the complexity of written text. This relation connects entities in the form of triples: (Argument1, Relation, Argument2). Arguments have the form Class : String. Class can be called as the immediate supercalss for the entity belonging to OntoCrime. For the triples to match, they should have identical relation slot. Finally, a score is computed. These triples are used as image indexing terms as well as image retrieval using free text queries. Similarity scores are computed between query triples and indexing-triples for retrieval mechanism. SOCIS performs stream of tasks in the following order starting with pre-processing (e.g., tokenization, POS tagging, named entity recognition and classification, etc.), parsing and naive semantic interpretation, inference and then triple extraction, extracting pragmatic relations in Tuples Form. The formula used for computing the similarity between query term $T1 = (Class_1 : Arg_1, Rel, Class_2 : Arg_2)$ and indexing term $T2 = (Class_3 : Arg_3, Rel, Class_4 : Arg_4)$ is as follows:

$$\begin{aligned} &Sim(T_1, T_2) = \\ &\alpha_1 * OntoSim(Class_1, Class_3) + \\ &\alpha_2 * OntoSim(Class_2, Class_4) + \\ &\alpha_3 * ArgSim(Arg_1, Arg_3) + \\ &\alpha_4 * ArgSim(Arg_2, Arg_4) \end{aligned} \tag{3}$$

where, $OntoSim(X, Y)$ is the inverse of the length between X and Y in OntoCrime, and $ArgSim(X, Y)$ is computed using the formula:

$$\begin{aligned} &ArgSim(X, Y) = \\ &\beta_1 * Match(X_Head, Y_Head) + \\ &\beta_2 * Match(X_Qual, Y_Qual) + \\ &\beta_3 * Match(X_Adj, Y_Adj) \end{aligned} \tag{4}$$

where $Match(X, Y)$ is 1 when $X = Y$ and 0 when $XisnotequaltoY$. The weighs $alpha_i$ and $beta_i$ have to be experimentally identified.

3 Results

Clustering is very useful and successful for a very large data set of diverse images and image annotations. We see above in 2.1 that text and image are complementary to each other and can be linked using Joint Probability distribution. While using clustering and merging procedures, we observe error rate is minimized in case of mid-level pruning (i.e, removing cluster with lesser than three images or cluster whose likelihood is very small). Error rate is probability of face being incorrect given an individual and a face.

We see following accuracy percentage while labelling of pictures like using only language model gives an accuracy of 85%, using EM to learn a language model gives 76% accuracy while using a

maximum likelihood clustering gives 84%. Again, we see the maximum likelihood clustering outperforms EM. Going further, we see that introducing a natural language model into face clustering produces much better results than clustering on appearance alone. We get an accuracy of 67% using only appearance model but when we add language model to appearance its accuracy is reached to 77%. On the other hand, the maximum likelihood assignment used for labeling instead of EM, improves the results from 72% to 77%.

For checking the performance of model discussed in 2.4, original paper talks of comparison between the annotation performance of the four models i.e., the Co-occurrence Model, the Translation Model, CMRM, and the model proposed in this paper (CRM). The result clearly shows that the CRM outperforms the other models substantially and is the only model among four which on using every word in the test set produces reasonable mean recall and mean precision numbers. Model discussed in 3.5 is tested on a corpus of 500 captions by evaluating the triple extraction and inference mechanism, and an accuracy of 80% is obtained.

4 Conclusion

Text and image features are both equally important in the clustering process. Using both text and image features, we derive a joint probabilistic model, which can be used both to retrieve images for the given text, and to annotate images automatically. We see above that a large noisy data set containing semi-supervised set of faces can be turned into a well supervised set and thus produces a clean clusters of many name and people. Application of language model to the appearance produces a model with better accuracy than using appearance model alone for labelling of pictures. Concept of various image region extracted out of an image is talked about and it is shown that the continuous feature vector extracted out of these image regions can be linked with the words using Joint probability model which produces one of the best model and also avoid usage of clustering process. Finally, we discuss about another approach to indexing and captioning of a photograph which uses pragmatic relations from a caption for creation of triple and this triple is further used for image indexing and retrieval. This approach is not explored much and much reseacrh needs to be done in this field.

5 Future Works

Future work will include the extension of the works above to larger data sets for both training and test data. This is required for both the better coverage and an evaluation of how such algorithms extend to large data sets. Improved feature sets may also lead to substantial improvements in performance. The retrieval mechanism for SOCIS is currently being implemented. More research needs to be done in this field to integrate all the methods discussed above in one and derive more accurate image-text processing results.

References

K. Barnard, P. Duygulu, and D. Forsyth. 2001. Clustering art. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pages II–II.

T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Yee-Whye Teh, E. Learned-Miller, and D. A. Forsyth. 2004. Names and faces in the news. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–II.

Tamara L. Berg, Alexander C. Berg, Jaety Edwards, and D. A. Forsyth. 2004. Who's in the picture? In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'04, pages 137–144, Cambridge, MA, USA. MIT Press.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

V. Lavrenko, R. Manmatha, and J. Jeon. 2003. A model for learning the semantics of pictures. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, pages 553–560, Cambridge, MA, USA. MIT Press.

Katerina Pastra, Horacio Saggion, and Yorick Wilks. 2003. Nlp for indexing and retrieval of captioned photographs. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 2*, EACL '03, pages 143–146, Stroudsburg, PA, USA. Association for Computational Linguistics.