SURVEY ON ENGRAVING CAPTIONS

Disha George
Assistant Professor
Dept. of Information Technology
Parul Institute of Engineering and Technology Vadodara, India.

Jignasha Parmar
Assistant Professor
Dept. of Information Technology
Parul Institute of Engineering and Technology Vadodara, India

Abstract*-– Engraving captions has been looking forward for the reliable manual work over the captions that are helpful in describing the picture with a derivative caption. The use of technology has depicted its role into which, it will be using the few concepts like Artificial intelligence which would manage the process of automation and the neural network which would be undergoing the recurrent neural network where it would mainly focus on the work using the highlights of both the supervised as well unsupervised method. This application will be helpful in maintaining the significance of the user over their picture with relevant captions. The emphasis of the given context will be totally defined by the process of describing the picture into its regulated and mandatory format. The image would be undertaken as the reference for the captions and would be engraved into the required format of the captions. The user would be availed by enhanced eye catching and focused captions that would be delivered by the main content provider. Requisites will be only on the requirement of user and the request that has to be responded by the content provider with the best and fascinating captions that would be derived from the event displayed on the image.

***Keywords*— Caption, image, RNN, CNN**

I. INTRODUCTION

The Social media is catching the rising trend of youth as well as other category of human being, this being the most effective nowadays requires an impact of caption that can describe the image/video in realistic and virtual manner. There were implementation on the field of describing pictures but it failed to incept the actual and trendy content.

We have read many research papers which shows different auto image captioning that are already constructed. They have some advantages but there are also disadvantages and limitations. Here we are mainly focusing on the attribute of describing the picture with a perfect caption. We would be undergoing through the manual as well as technical prospects for the betterment of the project. Description on something is as important as valuing the immature task over verb and noun clarification. Delivering the perfect caption for the photo has always been a very difficult task. Here, we categorize the method that has to be used in the caption generation. We would be describing the image with required terms and the mandatory things that it needs. Task will be focused on the feedback of the user which would be mainly noticing the social task and the pictures that are already in the scenario of the market.

Our main part of introduction would be drawing attention towards the pattern flow of the task that are going to be undergone by the project methods that are fetched with the picture and required accurate caption that describes the image. RNN is short memory and cannot predict for a long term as well as not helpful for long run. So, LSTM has been introduced to higher range of efficiency for longer sequences and prediction of appropriation.Gathering the current information and working on it would be magnifying the actual need that has to be struck and taken into consideration by the content deliverer and the application base as well.

II. THEORY OF ENGRAVING CAPTIONS

Caption is something that we usually need to think a lot and notice for a while to guess the catch of the moment in the picture. This app would be providing each and every type of captions related to the familiar emotion into it. The use of technical aspects that meet the need of the tentation of the user would be highlighted.

III. REVIEW OF LITERATURE

1.OBJECT HALLUCINATION IN IMAGE CAPTIONS

This generally focused on the illusions and dreams as well, where the generation of the caption would be according to their dream. In this the problem that was noticed was the actual captions cannot be generated where the process of illusion can create dillusions and other effects as well past decade. It is unclear to what extent captioning models actually rely on image content: as we show, existing metrics fall short awfully capturing the captions' relevance to the image .To measure object hallucination, we propose the CHAIR (Caption Hallucination Assessment with Image Relevance) metric. We have earlier described, how we deconstruct the Top-Down model to enable a controlled experimental setup. We rely on the deconstructed Top Down models to analyze the impact of model components on hallucination[1].Based on our results, we argue that the design and training of captioning models should be guided not only by cross-entropy loss or standard sentence metrics, but also by image relevance. Our CHAIR metric gives a way to evaluate the phenomenon of hallucination, but other image relevance metrics e.g. those that incorporate missed salient objects, should also be investigated. We believe that incorporating visual information in the form of ground truth objects in a scene (as opposed to only reference captions) helps us better understand the performance of captioning models.

2. IMAGE CAPTIONING

Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. Research in this area spans and connects numerous domains, such as computer vision, natural language processing, and machine learning. A description must capture not only the objects contained in an image, but it also must express how these objects relate to each other as well as their attributes and the activities they are involved in[4]. Recent advances in machine translation have shown an approach that can be used to circumvent these problems and solve the main problem directly. The task in machine translation is to transform a sentence written in a source language, into its translation T in the target language, by maximizing p(T|S). LSTM(Long Short Term Model) is a type of RNN's used for remembering information over long period of time[4]. c GRU's are a recent development which are growing in popularity as they train faster and give better performance on less-data.

3. DEFOILING FOILED IMAGE CAPTIONING

We address the task of detecting foiled image captions, i.e. identifying whether a caption contains a word that has been deliberately replaced by a semantically similar word, thus rendering it inaccurate with respect to the image being described. Solving this problem should in principle require a fine-grained understanding of images to detect linguistically valid perturbations in captions. In such contexts, encoding sufficiently descriptive image information becomes a key challenge. In this paper, we demonstrate that it is possible to solve this task using simple, interpretable yet powerful based on representation. We presented an object-based image representation derived from explicit object detectors/gold annotations to tackle the task of classifying foiled captions. The hypothesis was that such models provide the necessary semantic information for the task, while this information is not explicitly present in CNN image embeddings commonly used in V2L tasks. We achieved state of- the-art performance on the task, and also provided a strong upper-bound using gold annotations.

4. PROPOSED METHODOLOGY

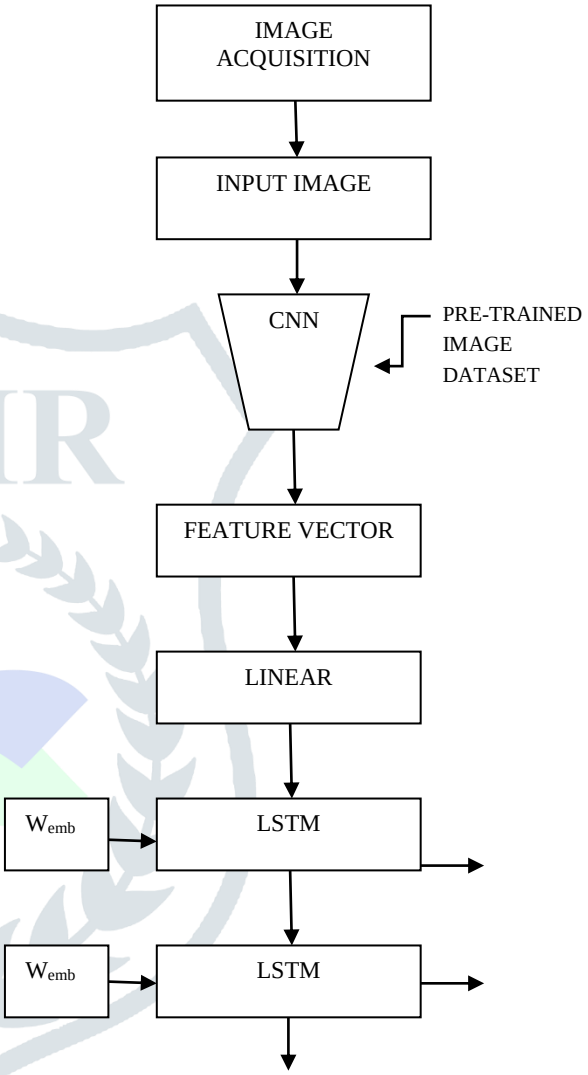

Fig: Flowchart of Proposed Methodology

5. A MULTIMODEL CAPTION GENERATOR

In this work, we showcase the Image2Text system, which is a real-time captioning system that can generate human-level natural language description for any input image. We formu- late the problem of image captioning as a multimodal trans- lation task. Analogous to machine translation, we present a sequence-to-sequence recurrent neural networks (RNN) model for image caption generation[9]. Different from most existing work where the whole image is represented by a convolutional neural networks (CNN) feature, we propose to represent the input image as a sequence of detected objects to serve as the source sequence of the RNN model. Based on the captioning framework, we develop a user-friendly system to automatically generated human-level captions for users. The system also enables users to detect salient objects in an

image, and retrieve similar images and corresponding description from database. We develop an image captioning system, which allows the users to 1) generate human- level natural language description of any input image, 2) detect objects in the given image alongside caption generation, and 3) retrieve similar images and descriptions from a database, which holds over 120,000 image- description pairs, all in real time. The core of the system is a combination of pre-trained deep convolutional neural networks for object detection, and recurrent neural networks for caption generation.

6. RECONSTRUCTION NETWORK FOR VIDEO CAPTION

In this paper, the problem of describing visual contents of a video sequence with natural language is addressed. Unlike previous video captioning work mainly exploiting the cues of video contents to make a language description, we propose a reconstruction network (RecNet) with a novel encoder-decoder- reconstructor architecture, which leverages both the forward (video to sentence) and backward (sentence to video) flows for video captioning. Specifically, the encoder-decoder makes use of the forward flow to produce the sentence description based on the encoded video semantic features. Two types of reconstructions are customized to employ the backward flow and reproduce the video features based on the hidden state sequence generated by the decoder. The generation loss yielded by the encoder-decoder and the reconstruction loss introduced by the reconstruction are jointly drawn into training the proposed RecNet[9] in an end-to-end fashion. Experimental results on benchmark datasets demonstrate that the proposed reconstruction can boost the encoder-decoder models and leads to significant gains in video caption accuracy. In this paper, we proposed a novel RecNet with the encoder-decoder-reconstruction architecture for video captioning.

IV. OBJECTIVE AND SCOPE

- Focus on the caption that has to be realistic as well as productive.
- To draw the attention of the user on its positive side.
- Mainly witness the progress over the aspects of caption generation.
- To reach the extent of the content where we could match the expectation of the user as well as the delivering of the caption.
- Categorize the general overview of the application that will divide the task into various different occasion in general.
- Go through the current social market and rule over the photos with the best and trending description in a new and elegant way.

V. ADVANTAGES

- Fast, reliable and reluctant from the unwanted tasks.
- The more better way of describing something would increase the popularity of the user.
- Comprehensive way of managing the photo would be highlighted.

VI. CONCLUSION

It would be easy for the user to go through the example and get a look how their actual caption should be and would not have to wait for their friend or colleague to get their caption done within a required time. Would be standardizing the actual use of technology as well influence the users with the fascinating captions that they would always like to deal with and keep it in their content expression of the picture. Here, the application would be playing the main role in content delivering and helping the users in letting them generate the basic and exact view for their picture. At the end of the day, we would simply say, it will be the application that will provide self written captions which would play the mesmerizing role on the part of users and will also dwell with technological terms for the betterment and management of the application

VII. REFERENCES

1 Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation, European Conference on Computer Vision, pages 382–398

2 Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi, et al. Video in sentences out. Uai, 2012.

3 D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate,ICLR, 2015.

4 P. Anderson, B. Fernando, M. Johnson, and S. Gould. Spice: Semantic propositional image caption evaluation. In ECCV, 2016.

5 C.-Y. Lin. Rouge: A package for automatic evaluation of summaries,ACL Workshop, 2004.2014

6 Y. Pan, Y. Li, T. Yao, T. Mei, H. Li, and Y. Rui. Learning deep intrinsic video representation by exploring temporal coherence and graph structure, IJCAI, 2016,Barry Evans, Clive E Sabel.

7 Karpathy and L. Fei-Fei. Deep visual semantic alignments for generating image descriptions, CVPR, 2015.

8 O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator, CVPR, 2015.

9 M. Lamb, A. G. A. P. Goyal, Y. Zhang, S. Zhang, A.C. Courville, and Y. Bengio. Professor forcing: A new algorithm for training recurrent networks, NIPS,2016.