Prediction of Caption and Emoji of an Image using Deep Learning

Bindu R, Anuradha T

Abstract: ***With the invention of deep learning, there is a good progress in image classification. But automatic generation of captions for images is still a challenging problem and is in the initial stages of artificial intelligence research. Automatic description of images has applications in social networking and will be useful to visually impaired persons. This paper concentrates on designing a user-friendly web application framework which can predict the caption of an image using deep learning techniques. The verbs and objects present in the caption are used for forming the emoji and for predicting the major color of the image.***

Index Terms: caption generation, convolutional neural networks, deep learning, emoji prediction, progressive loading

I. INTRODUCTION

People are trying to generate captions for the images automatically for a long time using artificial intelligence. When an image is shown to different people, each can describe it differently but each description will have an appropriate meaning. Image captioning is the task of generating such a meaningful textual description when given an image[1]. The appropriate caption of the image can describe the objects present in the image, their attributes and actions. The two major tasks in image captioning [2] are to identify the correct verbs based on objects, their attributes and actions for which a deep learning model is trained. The second part is generating the syntactically correct statement which connects all the identified objects along with their attributes and actions. Once the model training was done, the system will predict a caption for the new image. This caption is evaluated using BLEU score [3]. With the help of Emoji2vec [4], using the verbs and objects present in the caption, to emoji is generated. This feature helps in getting additional attention from the users who used this system to predict their images. This paper concentrates on designing a user friendly web application where a user can select an image through a browser and get the details like color, caption and emoji of that image. This is designed using python Flask framework and is implemented at the backend by python and tensor flow.

II. LITERATURE REVIEW

The main work in image captioning lies in connecting both the vision and language model which helps in generating a caption by connecting all the objects in the image and framing a caption correctly without grammatical mistakes.

Different caption generating models contain only one step encoder and decoder where as Zhihao Zhu[5] implemented a model consisting of two encoders and decoders. In this preview and tell model, a sample caption is generated with the major identified objects at the end of first encoder stage. In the later stage, finalized caption is generated with the suitable sentence framing. Even though the model for captioning is very effective, it is difficult to describe entire image content in one single sentence. We can only mention the highlighted objects and their attributes but not in fine deeper way. Jonathan Krause's [6] model solves this problem. The hierarchical recurrent networks help in identifying the image keenly. This generates multiple captions which cover almost all the pixels in an image. Then all those captions are joined as a meaningful paragraph in the dense language model. But the only disadvantage in this model is lengthiness. Zhongliang Yang [7] uses the language convertor as a base concept. Encoder and decoder are needed to convert a sentence from one language to another. Here, the source is either an image or sentence , it need to be encoded and caption will be generated using a language decoder. Convolution a neural network is used for encoding and recurrent neural networks is used for decoding. Jiuxiang Gu [8] developed a new diverse model called stack captioning based on reinforced learning for decoder. This decodes a caption from dense to crisp. In reinforced learning the outcome from one step decoder will be sent to the upper level decoders and will be processed further. Vikram Mullachery [9] had made use of checkpoints in the image captioning model for developing captions for videos also. Zhihao Zhu [10] developed a model which first decides the topic to which this image belongs to. Later caption is generated based on that identified topic. This feedback type model is popularly known as topic guided captioning. Ankit Gupta [11] shows the difference in caption generated when RNN is used along with LSTN. Here RNN is used to decode the caption from the list of objects identified, and LSTM stores content for longer period. The well-known work in the field of image captioning is done by Google [12] and published through a paper work called "show and tell". MD Zahir Hossain [13] has done a comprehensive survey of deep learning for image captioning. The survey describes all the existing methods and classifications in models.

III. THEORITICAL BACKGROUND

A. Flickr8k Data Set

As image captioning in deep learning research so far is done through supervised learning[14], there are some bench mark datasets available like COCO ,

Revised Version Manuscript Received on 16 September, 2019.

Bindu R, Department of Information Technology, Veleagapudi Ramakrishna Siddhartha Engineering College, Vijayawada, AP, India. atadiparty@gmail.com

Anuradha T, Department of Information Technology, Veleagapudi Ramakrishna Siddhartha Engineering College, Vijayawada, AP, India.

Flickr8k, Flickr30k, visual genome, Instagram datasets [15] . Flickr8k dataset which can identify 1000+ objects along with their actions, is used to train the cnn model in the current research. This dataset consists of 8000 images in which 6000 images are used for training, 1000 for validation and 1000 for testing. It mainly contains animals and human images with each image having an identification number and described by 5 captions. Fig. 1 shows a sample image and the captions from Flickr dataset.

B. Convolution Neural Networks

There are different convolution nets available like VGG 16, VGG19, inception V3, Alexnet [16], [17]. For some of the nets code is already written in keras and for some, code need to be written in python. VGG 16 model is used in this paper. It is already a built in model in keras with 16 layers.

C. Evaluation Metrics

After the model is trained, whenever an image is passed to the model, it generates the captions. The accuracy of the caption predicted is measured using different techniques like BLUE [3], SPICE, ROUGE [18] etc. In this paper BLUE score is used. The score ranges from 0 to 1 with score 0 representing invalid or less similar cation and 1 represents an accurate caption.

"A man is doing tricks on a bicycle on ramps in front of a crowd"
"A man on a bike executes a jump as part of a competition while the crowd watches"
"A man rides a yellow bike over a ramp while others watch"
"Bike rider jumping obstacles"
"Bike rider jumps off a ramp"

Fig. 1. sample image along with five captions from flickr8k dataset.

IV. EXPERIMENTAL WORK

As the main task of the proposed work is to design a user friendly web application framework to caption the required images, the first task is to design the web page to facilitate the user to select the images for captioning. This was done using Flask [19], a framework available in python with tensor flow as the backend. Fig. 2 shows the architectural diagram of the proposed work.

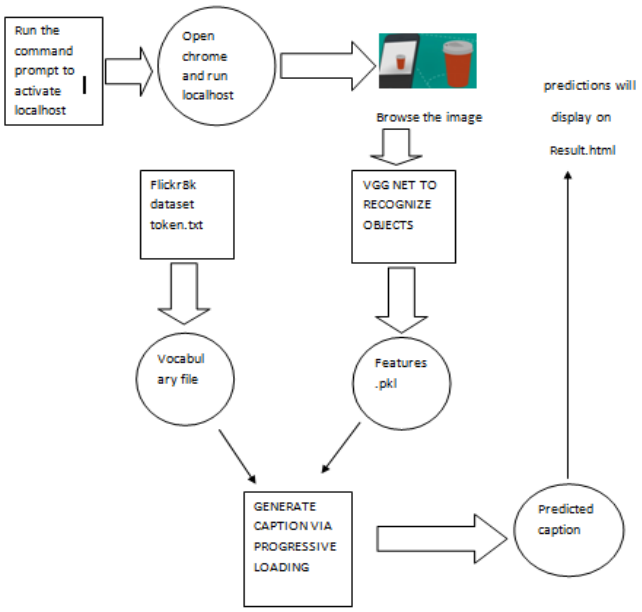

Fig. 2. Architectural Diagram of the proposed work

The background work majorly contains two modules. 1. Vision module and 2. NLP module. In the vision module image is scanned pixel by pixel and then with the help of pre trained VGG network and training dataset, objects present in the image will be predicted. The object with maximum area occupied will be sent to the color prediction. The model is trained with the help of color recognition utility present in python to identify seven colors namely black, white, grey, blue, green, red and yellow. The predicted color output will be one among this. For caption generation, the system should also detect the actions present in the image. VGG 16 model will encode the extracted features and they will be stored in pickle file . It has different objects and descriptions of those objects present in the image. Then this pickle file will be given to second module called natural language processing module which consists of LSTM [20] . It decodes the data and appropriate caption will be generated. The next part of the work is to generate the emoji. for this, the predicted caption is split into words and words are listed down in vector format with python inbuilt word2vec and this vector is sent to identify the related emoji in emoji2vec. Finally all these predictions will be sent to result.html and displayed to the user.

V. EXPERIMENTAL RESULTS

The web application framework designed by Flask has the option to choose an image file for which the user need to get the caption. The initial page displayed to the user is as shown in Fig. 3. Once the user clicks on 'choose file' button, he will be displayed with different images as shown in Fig. 4 from which he can select an image file. After selecting the image, user need to click the 'caption me' button. Then the web application shows the location from where the app has got the image, the predicted color of the image, predicted caption and emoji of the image as shown in Fig. 5.

As the system is trained for black, white, yellow, red, blue and green colors and as in the image selected for experimentation, the major part of the image has green color , the color predicted was green. And the major object in the image was dog, the system shown the name of the dog breed type and dog emoji.

VI. CONCLUSIONS

Image Captioning and emoji creation was successfully implemented using Deep Learning models. But this work can be further extended to recognize more colors. Emoji prediction is done by splitting words in predicted caption. But emoji can be predicted directly from image without using language model.

Fig. 3. Initial Page of the web App

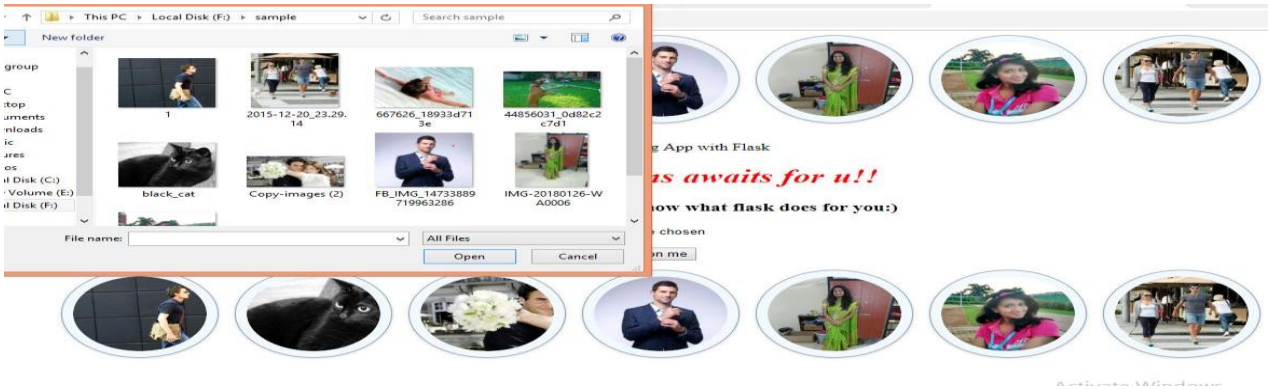

Fig. 4. Image Selection Page

Fig. 5. Final output- color, caption, emoji predictor

REFERENCES

1. Pan JY, Yang HJ, Duygulu P, Faloutsos C. "Automatic image captioning". In: IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No. 04TH8763) 2004 Jun 27 (Vol. 3, pp. 1987-1990). IEEE.
2. You Q, Jin H, Wang Z, Fang C, Luo J. "Image captioning with semantic attention". In: Proceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 4651-4659).
3. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "Bleu: a method for automatic evaluation of machine translation". In: Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL), pages 311–318, July 2002
4. Eisner B, Rocktäschel T, Augenstein I, Bošnjak M, Riedel S." emoji2vec: Learning emoji representations from their description". arXiv preprint arXiv:1609.08359. 2016 Sep 27.
5. Zhihao zhu, zhan xue, and Zejian Yuan. "Think and Tell: Preview network for Image captioning". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.5561–5570.
6. Shuang Bai and Shan An. 2018. "A Survey on Automatic Image Caption Generation". Neurocomputing.
7. Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, Barbara Plank, et al. 2016. "Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures". Journal of Artificial Intelligence Research (JAIR) 55,409–442.
8. XinleiChenandCLawrenceZitnick.2015.Mind'seye:"A recurrent visual representation for image caption generation". In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2422–2431.
9. Bo Dai, Dahua Lin, Raquel Urtasun, and Sanja Fidler.2017. "Towards Diverse and Natural Image Descriptions via a Conditional GAN". In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR).2989–2998.
10. Andrej Karpathy and Li Fei-Fei. 2015. "Deep visual-semantic alignments for generating image descriptions". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3128–3137.

11. Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim.2017. “Attend to You: Personalized Image Captioning with Context Sequence Memory Networks”. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR).6432–6440.
12. Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek.2017. “Areas of Attention for Image Captioning”. In: Proceedings of the IEEE international conference on computer vision.1251–1259.
13. Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-JiaLi.2017. “Deep Reinforcement Learning-based Image Captioning with Embedding Reward” .In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR).1151–1159.
14. Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and JieboLuo.2016. “Image captioning with semantic attention”. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.4651–4659.
15. Vinyals O, Toshev A, Bengio S, Erhan D. “Show and tell: Lessons learned from the 2015 mscoco image captioning challenge”. IEEE transactions on pattern analysis and machine intelligence. 2016 Jul 7;39(4):652-63.
16. Simon M, Rodner E, Denzler J. “Imagenet pre-trained models with batch normalization”. arXiv preprint arXiv:1612.01452. 2016 Dec 5.
17. Bhavya Sai V, Narasimha Rao G, Ramya M, Sujana Sree Y, Anuradha T. “Classification of skin cancer images using TensorFlow and inception v3”, International Journal of Engineering and Technology Vol 7, No 2.7 (2018)
18. Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, pages 74–81, July 2004.
19. Armin Ronacher Flask, Web Development, One Drop at a Time [online] Available: http://flask.pocoo.org/ © Copyright 2010 - 2019 Armin Ronacher
20. Pranjal Srivastava Essentials of Deep Learning : Introduction to Long Short Term Memory DECEMBER 10, 2017 [online] Available: https://www.analyticsvidhya.com/blog/2017/12/fundamentals-of-deep-learning-introduction-to-lstm/ [Accessed]: 10th Jan 2019

AUTHORS PROFILE

Bindu. R Final year B.Tech student in Department of Information Technology, VR Siddhartha Engineering College.

Anuradha. T Ph.D from Acharya Nagarjuna University, Working as professor in VR Siddhartha Engineering College, AP, India. Research Areas Data Mining and Machine Learning. Published more than 40 research papers in reputed journals and conferences.