International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 02 | Feb 2021 www.irjet.net p-ISSN: 2395-0072

Image Captioning and Visual Question Answering for the Visually Impaired

Yash Jeswani [1], Siddhesh Sawant[2], Hrushikesh Makode[3], Ankush Rathi[4] ,Prof.Sumitra Jakhete[5]
sajakhete@pict.edu, yashjeswani2420@gmail.com, makodehrushikesh@gmail.com, rathi.ankush2438@gmail.com, sid123.sawant@gmail.com
[1,2,3,4,] *Student,,Dept Information Technology,Pune Institute of Computer Technology (PICT), Pune, Maharashtra.*
[5]*Professor,Dept Information Technology,Pune Institute of Computer Technology (PICT), Pune,Maharashtra, India*

Abstract: *For people, it's straightforward for us to take a look at an image and give the response to the answer for any questions utilizing our insight. In any case, there additionally are situations, for example, a visually impaired user or an intelligence, any place they need to effectively evoke visual information given a picture. We might want to help blind people to beat their day by day visual difficulties and separate social availability obstructions. The main purpose of our project is Image captioning and VQA, Image captioning is to get a caption for an Image.Image Captioning is to get an inscription or caption for a picture. Picture inscribing needs to decide objects in the picture, activities, their relationship and a couple of quiet highlights that might be absent inside the picture. While distinguishing, the accompanying advance is to get a most relevant and transient description for the picture that must be grammatically and semantically right. It utilizes each CNN for identification of objects and language process ways for description and on its description users (visually impaired) will raise any questions, we tend to propose the task of free-form and open-ended VQA. Given an image and a characteristic language question concerning the picture, the task is to deliver a right regular language answer.*

Keywords - Object Detection,Fully connected neural networks(FCCNs),Longs hort-Term Memory, Convolutional Neural Network,Image captioning,VQA

I. INTRODUCTION

One of the complex important tasks for the computer vision community is to combine various tools for high-level scene interpretation, such as image captioning and visual question answering. Such technologies have the potential to assist people who are blind or visually impaired. With regard to an image, image captioning is the process of generating a textual description and visual question answering is aimed at answering questions about it.

We propose a Machine Learning application for this to deal with the same. For image captioning, an LSTM network is fed with vectorized representation of image by a pre-trained CNN to generate captions. For question answering, the vectorized representations of image and textual question are combined to generate the answer.

We hope this work will help visually impaired people overcome their daily visual challenges.

II. LITERATURE SURVEY

Literature Survey was conducted in order to study and obtain knowledge from previous researches and surveys. Some papers were classified based on tools/software's used, algorithms used and their corresponding data sets (if any) used along with the corresponding platform on which they were deployed. Papers are also mentioned describing the comparison between various existing Image Captioning and Visual Question Answering methodologies along with various Datasets as follows.

1)Image Captioning

Retrieval based and template based image captioning methods are adopted mainly in early work. Due to great progress made in the field of deep learning **[1]**, recent work begins to rely on deep neural networks for automatic image captioning. In this section, we will review such methods. Even though deep neural networks are now widely adopted for tackling the image captioning task, different methods may be based on different frameworks. Therefore, we classify deep neural network based methods into subcategories on the basis of the main framework they use and discuss each subcategory respectively.

A. Retrieval and template based methods augmented by neural networks:

To retrieve description sentences for a query picture, Socheret al. propose to utilize dependency-tree recursive neural networks to address phrases and sentences as compositional vectors. They utilize another deep neural network **[2]** as a visual model to extract features from images.Obtained multimodal features are mapped into a common space by using a max-margin objective function.After training, correct image and sentence pairs in the common space will have larger inner products and vice versa. Finally, sentence recovery is performed dependent on similarities between representations of images and sentences in the common space.

Karpathy et al. propose to embed sentence fragments and image fragments into a common space for ranking sentences for a query image **[3]**. Addressing both picture sections and sentence parts as feature vectors, the creators plan an organized max-edge objective, which incorporates a global ranking term and a fragment alignment term, to map visual and textual data into a common space.

In **[4]** Ma's framework incorporates three sorts of parts, for example picture CNNs to encode visual information coordinating CNNs to together address visual and textual data and multilayer insights to score similarity of visual and textual information.

B. Image captioning based on multimodal learning

Kiros et al. propose to utilize a neural language model which is conditioned on image inputs to generate captions for images. In their method, the log-bilinear language model **[5]** is adjusted to multimodal cases. In a natural language processing problem, a language model is utilized to predicate the probability of generating a word conditioned on recently produced words.

Karpathy and Fei-Fei present a way to deal with align image areas represent by a Convolutional Neural Network and sentence sections represent by a Bidirectional Recurrent Neural Network **[6]** to become familiar with a multimodal Recurrent Neural Network model to generate descriptions for image regions.After representing image regions and sentence fragments by utilizing corresponding neural networks, an structured goal is utilized to map visual and textual information into a common space.

Chen and Zitnick propose to dynamically construct a visual representation of a picture as an caption is being produced for it, so that drawn out visual ideas can be remembered during this process **[7].**

C. Image captioning based on the encoder-decoder framework

Inspired by recent advances in neural machine translation **[8]** the encoder-decoder framework is adopted to generate captions for images.In image captioning methods under this framework, an encoder neural network first encodes an image into an intermediate representation,then a decoder recurrent neural network takes the intermediate representation as input and generates a sentence word by word.

Kiros et al. introduce the encoder-decoder framework into image captioning research to unify joint image-text embedding models and multimodal neural language models, so that given an image input, a sentence output can be generated word by word **[9]** like language translation.

With the same inspiration from neural machine translation, Vinyals et al. use a deep Convolutional Neural Network as an encoder to encode images and use Long Short-Term Memory(LSTM) Recurrent Neural Networks to decode obtained image features into sentences **[10].** With the above framework,the authors formulate image captioning as predicting the probability of a sentence conditioned on an input image.

Similar to Vinyals's work , Donahue et al. also adopt a deep Convolutional Neural Network for encoding and Long Short-Term Memory Recurrent Networks for decoding to generate a sentence description for an input image**[11]**. The difference is that instead of inputting image features to the system only at the initial stage, Donahue et al. provide both image feature and context word feature to the sequential model at each time step.

2)Visual Question Answering

A VQA system is studied under the domain of computer vision. In the last few years the popularity of the VQA system has increased many folds.

In **[12]**, They have introduced a dataset called Figure-QA dataset along-with a baseline model.

So in **[13]** a new dataset was introduced VQA v2 which contains twice the number of images and every question having at-least two different answers.

In **[14],** bottom up top-down attention model which later won the 2017 VQA challenge. After image classification and object detection, image captioning became the main focus of experts.

Later other models like **[15]** were also introduced which support passing relational facts along-with answers to the model while training. This helped make the system more semantically capable and answer questions containing why other than what, how, which etc.

In **[16]** proposed a multimodal pooling method to concatenate text and image models. Deep Convolutional Neural Networks are preferred over other techniques for image captioning and VQA systems, as they contain various layers which can represent various details of an image.

We have come a long way from digit recognition to the state of art Pythia**[17]** VQA system by facebook.A lot of work has been done in the field. Various approaches are proposed like Multimodal fusion,Compositional approaches,Question-Aware models, etc.

In **[18]** they proposed how these networks can be used for a huge number of images.After image caption generation, interacting with those images was the next popular task which resulted into VQA systems.Using questions and answers along-with images while training gives the model capability for answering when asked questions.

III.PROPOSED MODEL AND DATASET

A) Image Captioning (CNN+LSTM)

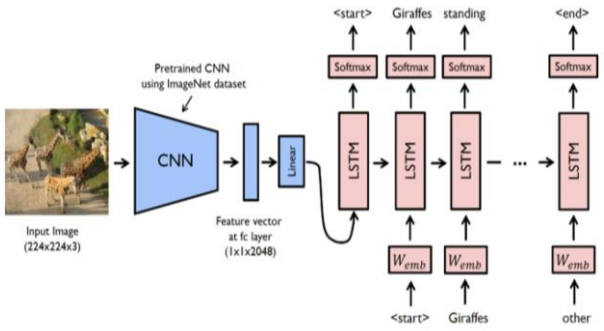

Fig.1: Common Architecture of Image captioning

- **Extract the features from an image:**
  - For feature Extraction,CNN pre-trained on ImageNet is used.
  - The Extract feature vector of CNN is linearly transformed and this feature vector is used as an initial input to the RNN.
- **Language Based Model(RNN):**
  - This model translates the features and objects given by our image based model to a Caption.

B)VQA:

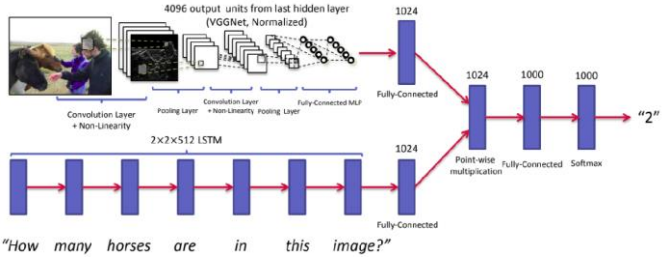

Fig.2: Common Architecture of VQA

The task can be divided into three parts:

- ❖ Extract features from the image
- ❖ Extract features from the question
- ❖ Combine the features to generate an answer

In the case of image features, CNN pre-trained on ImageNet is used.For question features, LSTM encoders are used.Both these features are then combined by element wise product(pointwise Multiplication).Then it is passed through a fully connected layer to generate an answer.

C) Dataset:

Within the last 10 or a ton of years, almost twenty in public available captioning datasets have been produced to help the occasion of programmed image captioning models. The general dependable guideline has been to join a singular amount scope of models, relying on scratching pictures from the net to a great extent from Flickr to help the extension from numerous thousand to a large number of subtitled pictures in such datasets.To help adjust the vision local area to focus on tending to the genuine interests of people who need picture inscriptions, we tend to rather focus on presenting an inscribing dataset that rises out of a characteristic use case.This remembers subtitled pictures for paper articles and given by local escorts concerning pictures of traveler areas . As opposed to these past works, we center around a clear use case (i.e., inscribing blind picture takers' pictures) and our new dataset is fundamentally bigger (i.e., contains almost 40,000 pictures versus 3,361 and 20,000).For model, this is frequently anyway well known visual discernment datasets , scene acknowledgment datasets and characteristic acknowledgment datasets were made. discerning that machine-driven techniques trust such enormous scope datasets to direct what thoughts they realize, an issue arises of anyway well the substance in such invented datasets reflect the interests four D. Gurari et al. of genuine clients of picture portrayals administrations. We tend to direct examinations between normal vision datasets and our new dataset to supply such knowledge. This examination is successful each for featuring the value of existing datasets to help a genuine use case and uncovering how vision datasets might be improved.

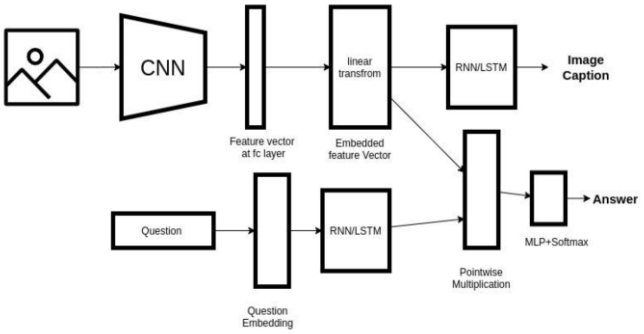

Fig.3: Overview of proposed model

IV. SYSTEM DESIGN

The user should be able to click an image with a button, and user will ask for the description and the description for the particular image will be generate and this description text converted text to speech, same image is given to VQA model and then user can ask the spoken question, once it fed in to the model,it would return an answer to the user. The answer would be supported by a speech assistant.The system has been designed to serve as an assisting technology for the blind and visually impaired.

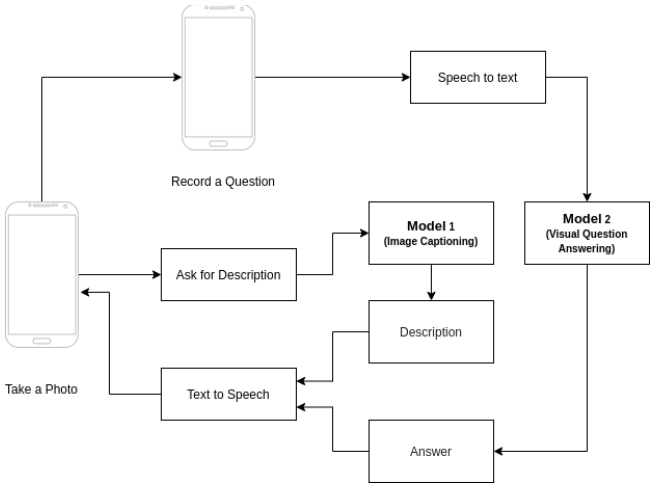

Fig.4 - System Design

V. FUTURE SCOPE

It can be used to describe video in real time. Generating Passage from real time video and then User again can ask Questions Based on the passage generated and then our product will answer(VQA).

VI. CONCLUSION

The objective of our project was to discuss the system design and methodology to be adopted to design an assistive technology for the blind and visually impaired. We compared different Image Captioning and VQA architectures and proposed a free real-time application which is user friendly.

VII.REFERENCES

[1] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553)(2015) 436–444.

[2] Q. V. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. Corrado,J. Dean, A. Y. Ng, Building high-level features using large scale unsupervised learning, in: International Conference on Machine Learning,2012.

[3] A. Karpathy, A. Joulin, F. Li, Deep fragment embeddings for bidirectional image sentence mapping, in: Advances in Neural Information Processing Systems 27 (NIPS), Vol. 3, 2014, pp. 1889–1897.

[4] L. Ma, Z. Lu, Lifeng, S. H. Li, Multimodal convolutional neural networks for matching image and sentences, in: IEEE International Conference on Computer Vision, 2015, pp. 2623–2631

[5] A. Mnih, G. Hinton, Three new graphical models for statistical language modelling, in: Proceedings of the 24th international conference on Machine learning, 2007, pp. 641–648

[6] M. Schuster, K. Paliwal, Bidirectional recurrent neural networks, IEEE Transactions on Signal Processing 45 (11) (1997) 2673–2681.

[7] X. Chen, C. Zitnick, Mind's eye: A recurrent visual representation for image caption generation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2422–2431

[8] N. Kalchbrenner, P. Blunsom, Recurrent continuous translation models,in: Conference on Empirical Methods in Natural Language Processing,2013.

[9] R. Kiros, R. Salakhutdinov, R. Zemel, Unifying visual-semantic embeddings with multimodal neural language models, arXiv preprint arXiv:1411.2539.

[10] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: Lessons learned from the 2015 mscoco image captioning challenge, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (4).

[11] J. Donahue, L. Hendricks, S. Guadarrama, M. Rohrbach, S. Venu-gopalan, Long-term recurrent convolutional networks for visual recognition and description, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625–263

[12] Kahou, Samira Ebrahimi and Michalski, Vincent and Atkinson, Adam and Kadar, Akos and Trischler, Adam and Bengio, Yoshua. "Fig- ureqa: An annotated figure dataset for visual reasoning". arXiv preprint arXiv:1710.07300,2017.

[13] Goyal, Yash and Khot, Tejas and Summers-Stay, Douglas and Batra, Dhruv and Parikh, Devi. "Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering, 2017 IEEE Conference on Computer [14] Das, Abhishek and Agrawal, Harsh and Zitnick, Larry and Parikh, Devi and Batra, Dhruv. "Human attention in visual question answering: Do humans and deep networks look at the same regions". Computer Vision and Image Understanding, vol. 163, pages 90–100, 2017, Elsevier.

[15] Lu, Pan and Ji, Lei and Zhang, Wei and Duan, Nan and Zhou, Ming and Wang, Jianyong. "R-VQA: Learning Visual Relation Facts with Semantic Attention for Visual Question Answering". arXiv preprint arXiv:1805.09701, 2018.

[16] Fukui, Akira and Park, Dong Huk and Yang, Daylen and Rohrbach, Anna and Darrell, Trevor and Rohrbach, Marcus. "Multimodal compact bilinear pooling for visual question answering and visual grounding". arXiv preprint arXiv:1606.01847. 2016.

[17] Jiang, Yu and Natarajan, Vivek and Chen, Xinlei and Rohrbach, Marcus and Batra, Dhruv and Parikh, Devi."Pythia v0. 1: the winning entry to the vqa challenge 2018, arXiv preprint arXiv:1807.09956, 2018.

[18] Stanislaw Antol, Aishwarya Agrawal,Jiasen Lu, Margaret Mitchell, Dhruv Batra,C. Lawrence Zitnick,Devi Parikh "VQA: Visual Question Answering"2016 arXiv preprint arXiv:1505.00468v7,2016