Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Image Captionbot for Assistive Technology

Safiya K. M.
Research Scholar, Department of Computer Science and Engineering
Sathyabama Institute of Science and Technology(Deemed to be University)
Chennai,India
Dr. R. Pandian
Associate Professor, Department of Electronics and Instrumentation
Sathyabama Institute of Science and Technology(Deemed to be University)
Chennai,India

Article Info
Page Number: 1629 – 1634
Publication Issue:
Vol. 71 No. 3s2 (2022)

Article History
Article Received: 22 April 2022
Revised: 10 May 2022
Accepted: 15 June 2022
Publication: 19 July 2022

Abstract— Generating small descriptions from the image is a very difficult task because of the complexity of image features and the vastness of the language contexts. An image may contain a wide variety of information and thus extracting the context of the information contained in the image and generation of the sentence using that context is a very complex task. However, the task can help blind people to understand the surrounding without others assistance. Deep learning techniques have emerged as a new trend in programming and can be utilized to develop this kind of system. In the project, we will be using VGG16, one of the best CNN architectures for image classification and for extracting features from images. An embedding layer and LSTM will be used for text description. And these two networks will be combined to form an image caption generation network. Then we will train the model using data prepared from the flickr8k dataset. The trained model will be used to generate caption from new images and the generated caption will be converted to audio for helping the blind.
Keywords— Deep learning, CNN,LSTM

I. INTRODUCTION

Even in this modern period, the differently abled people remains a vital element of our modern societies, where they struggle to integrate their daily lives with the greater society. As a result, their social advancement is hampered, and their contribution to commercial production is reduced to none. They were not included in crucial aspects of socioeconomic culture. Our aim is to decrease this constantly increasing gap and assist them to contribute more productively to their respective societies. This is achieved with the help of advancing modern technology.

For a normal person it is easy to determine the scene description and its contents when an image is given but the visually impaired or blind do not have this capacity. This system will describe the visual contents in natural language will help the impaired a lot. In this automate the way to take visual contents and will convert to natural language descriptions will empowered the impaired population that why the system will be a socially relevant one.

The capacity to produce natural language content descriptions of photographs is still one of the most difficult challenges for a machine to accomplish. It was made possible by recent breakthroughs in the field of computer vision. Image descriptions are more difficult to create than object identification and classification tasks because they require the descriptions of the

objects as well as the context in which they are located in the visual. The challenge of developing visual descriptions in natural language necessitates a combined knowledge of the language model and the visual model.

II. RELATED WORKS

Producing natural language descriptions of visual components is a difficult issue that has become a popular study topic in recent years. These are the systems in place.

Because of some limitations, they are easily broken and provide there are extremely few features.

Seung-Ho Han and Ho-Jin Choi proposed a novel in [1].

Domain-specific image captioning is a type of picture captioning model.A caption generator that creates a caption for a video,image created with visual and semantic attention (also known as a domain-specific caption)from the overall caption by substituting semantic ontology for the individual phrases in domain-specific terms in the generic caption.

Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston introduced a model in [2] that can understand image information while also providing feedback, captions that are interesting to read for humans In this case, in particular, achieve a on COCO, a new state-of-the-art in caption generating, and provide a new retrieval architecture, dubbed TransResNet.

N. Komal Kumar, D. Vigneswari, A. Mohan, and K. Mohan published [3], J. Yuvaraj and Laxman proposed a deep learning method for the use of neural networks to generate image captions is new.The proposed method was tested on a Flickr 8k dataset .The deep learning procedure that has been proposed created captions that were more descriptive than the original. Image caption generators that are already in use. It is possible to create a hybrid picture caption generator model for more accurate captions in the future

Feng Chen, Songxian Xie, Xinyi Li, Shasha Li, Jintao Tang, and Ting Wang suggested a method in [4] that brought attribute concept into the CNN-RNN architecture and improved performance while relying heavily on manually picked attributes. The CNN-RNN framework is used to propose a topic-guided neural image captioning model that incorporates a topic model.

Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele introduced an adversarial caption generator model that is deliberately trained to generate a variety of captions for photos in [5.] It's done by training the generator with an adversarial learning framework and a discriminator network designed to enhance diversity.

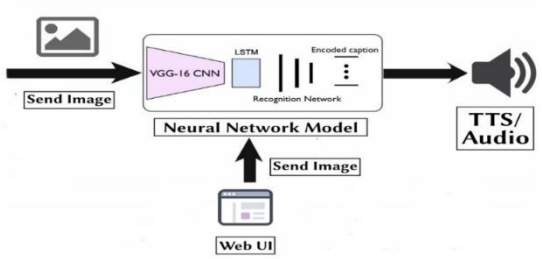
Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin introduced a new method for generating image descriptions in [6]. In contrast to current methods, which are primarily concerned with matching detailed phrasing, our approach focuses on improving overall quality, which includes semantic relevance, naturalness, and diversity. In comparison to the progressive MLE-based approach, the suggested methodology produced descriptions that were more natural, diversified, and semantically meaningful on MSCOCO and Flickr30k.

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng offer StyleNet as a new framework for creating attractive captions for photos and videos of various styles in [7]. On the recently gathered FlickrStyle10K image caption dataset, it reveals that StyleNet

outperforms existing algorithms for creating visual captions with radically diverse styles, as judged by both automatic and human review criteria.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher introduce a new adaptive attention encoderdecoder system in [8], which includes a decoder fallback option. Introduce a new LSTM extension that creates an additional'visual sentinel.' In this case, the model achieves state-of-the-art performance on typical picture captioning benchmarks. Many reviews are presented in literature by many researchers with respect to ecommerce applications in different domain [13][14][15]. This analysis will surely enable the researchers with the idea of deep learning technique in different applications [16][17][18][19]. Different issues also discussed in machine learning applications [20][21][22].

III. PROPOSED SYSTEM



A. OVERVIEW

Proposed method will be using VGG16, one of the best CNN architectures for image classification and for extracting features from images. An embedding layer and LSTM will be used for text description. And these two networks will be combined to form an image caption generation network. Then we will train the model using data prepared from the flickr8k dataset. The trained model will be used to generate caption from new images and the generated caption will be converted to audio for helping the blind**.**

3.2 MODULES

The proposed system consists of following main modules:

3.2.1 IMAGE FEATURE EXTRACTION

VGG16 model, one of the best CNN architectures for image classification is used for extracting features of an image. Using this pre-trained model VGG16, we first extract all the features of a particular image. VGG16 returns a vector of features and it is then saved as a file. Then we create a mapping of image id and image features.

3.2.2 TEXT PROCESSING

Initially convert the text into lower case and remove the punctuations, remove the words with numbers. After processing, create a vocabulary of text and save the vocabulary. May one image has more than one description, so a mapping is created between images and descriptions.

3.2.3 TOKENIZATION

In the text, add starting and ending tokens. The next step after adding tokens is to prepare a tokenizer for the text and save the tokenizer. Tokenize the description of the image that is with numbers. Then create image sequence, input sequence, output words for the corresponding image.

3.2.4 ARCHITECTURE CREATION

Two dense layers are used for image features that are for text descriptions. The layers are the embedding layer and the LSTM layer. These two networks will be combined to form an image caption generation network that is by concatenation.

3.2.5 TRAIN THE MODEL

We train the model in google colab and save the trained model.

3.2.6 IMAGE CAPTION GENERATION

The challenge of creating a human-readable written description for a given image is a difficult artificial intelligence problem, and caption creation is no exception. It necessitates expertise in both computer vision and natural language processing. Computer vision knowledge is required for picture recognition, while the natural language processing area is required for language modelling. To do so, load the image and use the pre-trained model VGG16 to extract the feature from it.

After loading the image, load the tokenizer and creae a caption generation function using the saved model and the tokenizer. At last, convert the generated caption to audio.

IV. CONCLUSION

As a result, we offer a system that delivers natural descriptions of provided images to the visually impaired population, allowing them to contribute more constructively to society. We create an encoder-decoder system in which a VGG-16 network is trained first, followed by an LSTM network, to produce a mapping from images to phrases.. The system performance is evaluated and shows a good results. In future we will try to improve accuracy and performance.

REFERENCES

1. Seung-Ho Han and Ho-Jin Choi ," Domain-Specific Image Caption Generator with Semantic Ontology" IEEE 2020
2. Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, Jason Weston, "Engaging Image Captioning via Personality" IEEE 2019
3. Soheyla Amirian, Khaled Rasheed , Thiab R. Taha , Hamid R. Arabnia, "Image Captioning with Generative Adversarial Network" IEEE 2019
4.]N. Komal Kumar , D. Vigneswari , A. Mohan , K. Laxman , J. Yuvaraj, "Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach" IEEE 2019
5. Yimin Zhou, Yiwei Sun, Vasant Honavar, " Improving Image Captioning by Leveraging Knowledge Graphs"IEEE 2019
6. Hoda, S. A. ., and D. A. C. . Mondal. "A Study of Data Security on E-Governance Using Steganographic Optimization Algorithms". International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 5, May 2022, pp. 13-21, doi:10.17762/ijritcc.v10i5.5548.
7. Feng Chen, Songxian Xie, Xinyi Li, Shasha Li, Jintao Tang, Ting Wang, "What topics do Images say: A Neural Image captioning model with Topic Representation" IEEE 2019.
8. Seung-Ho Han and Ho-Jin Choi, "Explainable Image Caption Generator Using Attention and Bayesian Inference" IEEE 2018

9. Mirza Muhammad Ali Baig, Mian Ihtisham Shah, Muhammad Abdullah Wajahat, “Image Caption Generator with Novel Object Injection” IEEE 2018.
10. Agarwal, D. A. . (2022). Advancing Privacy and Security of Internet of Things to Find Integrated Solutions. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 8(2), 05–08. https://doi.org/10.17762/ijfrcsce.v8i2.2067
11. Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, Bernt Schiele, “Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training” IEEE 2017
12. Bo Dai,Sanja Fidler, Raquel Urtasun, Dahua Lin,“Towards Diverse and Natural Image descriptions via a Conditional GAN” IEEE 2017
13. Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, Li Deng, “StyleNet: Generating attractive Visual Captions with Styles” IEEE 2017.
14. J. . Hermina, N. S. . Karpagam, P. . Deepika, D. S. . Jeslet, and D. Komarasamy, “A Novel Approach to Detect Social Distancing Among People in College Campus”, Int J Intell Syst Appl Eng, vol. 10, no. 2, pp. 153–158, May 2022.
15. Jiasen Lu , Caiming Xiong , Devi Parikh, Richard Socher, “Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning” IEEE 2017.
16. Kanyadara Saakshara, Kandula Pranathi, R.M. Gomathi, A. Sivasangari, P. Ajitha, T. Anandhi, "Speaker Recognition System using Gaussian Mixture Model", 2020 International Conference on Communication and Signal Processing (ICCSP), pp.1041-1044, July 28 - 30, 2020.
17. R. M. Gomathi, P. Ajitha, G. H. S. Krishna and I. H. Pranay, "Restaurant Recommendation System for User Preference and Services Based on Rating and Amenities," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), 2019, pp. 1-6, doi: 10.1109/ICCIDS.2019.8862048.
18. Subhashini R , Milani V, "IMPLEMENTING GEOGRAPHICAL INFORMATION SYSTEM TO PROVIDE EVIDENT SUPPORRT FOR CRIME ANALYSIS", Procedia Computer Science, 2015, 48(C), pp. 537–540
19. Harish P, Subhashini R, Priya K, "Intruder detection by extracting semantic content from surveillance videos", Proceeding of the IEEE International Conference on Green Computing, Communication and Electrical Engineering, ICGCCEE 2014, 2014, 6922469
20. P. M. Paithane and D. Kakarwal, “Automatic Pancreas Segmentation using A Novel Modified Semantic Deep Learning Bottom-Up Approach”, Int J Intell Syst Appl Eng, vol. 10, no. 1, pp. 98–104, Mar. 2022.
21. Sivasangari, A., Krishna Reddy, B.J., Kiran, A., Ajitha, P.(2020), “ Diagnosis of liver disease using machine learning models”, ISMAC 2020, 2020, pp. 627–630, 9243375
22. Sivasangari, A., Nivetha, S., Pavithra,, Ajitha, P., Gomathi, R.M. (2020),” Indian Traffic Sign Board Recognition and Driver Alert System Using CNN”, 4th International Conference on Computer, Communication and Signal Processing, ICCCSP 2020, 2020, 9315260
23. Ajitha, P., Lavanya Chowdary, J., Joshika, K., Sivasangari, A., Gomathi, R.M., "Third Vision for Women Using Deep Learning Techniques", 4th International Conference on Computer, Communication and Signal Processing, ICCCSP 2020, 2020, 9315196

24. Ajitha, P.Sivasangari, A.Gomathi, R.M.Indira, K."Prediction of customer plan using churn analysis for telecom industry",Recent Advances in Computer Science and Communications,Volume 13, Issue 5, 2020, Pages 926-929.
25. Gowri, S. and Divya, G., 2015, February. Automation of garden tools monitored using mobile application. In International Confernce on Innovation Information in Computing Technologies (pp. 1-6). IEEE.
26. Gowri, S., and J. Jabez. "Novel Methodology of Data Management in Ad Hoc Network Formulated Using Nanosensors for Detection of Industrial Pollutants." In International Conference on Computational Intelligence, Communications, and Business Analytics, pp. 206-216. Springer, Singapore, 2017.