iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 9 **Issue:** VII **Month of publication:** July 2021

**DOI:** https://doi.org/10.22214/ijraset.2021.36267

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429*
*Volume 9 Issue VII July 2021- Available at www.ijraset.com*

# Image Captioning using Deep Learning for the Visually Impaired

Dr. A. M. Chandrashekhar[1], Akash Raj K R[2], Preetham Jain[3], Vinayaka Bhat[4], Nagarjun P R[5]
[1]*Assistant Professor, Department of Computer Science and Engineering JSS Science and Technology University, Karnataka, India*
[2, 3, 4, 5]*Undergraduate Student, Department of Computer Science and Engineering, JSS Science and Technology University, Karnataka, India*

***Abstract:*** ***Describing the content of an image has been a fundamental problem of Machine learning that connects computer vision and natural language processing. In recent years, the task of object recognition has advanced at an exceptional rate which in turn has made image captioning that much better and easier. In this paper, we have discussed the usage of image captioning using deep learning for the visually impaired. We have used Convolutional Neural Networks along with Long Short-Term Memory to train and generate captions for images along with a text-to-speech engine which makes the experience of visually impaired users who are browsing the internet much smoother. We discuss how the model was implemented, its different components and modules along with a result analysis conducted on a set of outputs peer reviewed by our colleagues, friends and professors.***

***Keywords: Recurrent Neural Network, Convolution Neural Network, Long short-term memory, Visually Impaired***

## I. INTRODUCTION

Recent progress in deep learning has enabled significant advancements in understanding the relationships between visual and language entities. Automatically generating captions to an image shows the understanding of the image by computers, which is a fundamental task of intelligence. Lately, the internet is playing a huge role in everyone's lives. The ease of access to the internet lets the majority of the population be able to use the internet. As a result, it also becomes important for visually impaired people to be able to access and get involved with everything happening on the internet. A lot of research has been put into text-to-speech models which have yielded promising results helping them to be able to read on the internet.[2] However, interpreting images is still a daunting task for them. Images play a vital role in conveying content and information which improves the overall experience of the user. Hence, there is a need to generate captions for images to help visually impaired people to be able to interpret images and use the internet with more freedom and improve their overall experience on the internet. So, we are developing an application that uses deep learning frameworks to generate captions for an image in a webpage upon user command and speak the caption out loud through text-to-speech from third party APIs. In a systematic literature review[4] done by Murk Chohan and others, they have found that using CNN along with LSTM would produce greater results of image caption relevance as opposed to ResNET and RNN. Thus, we opted to use a similar model with tuning of several hyper-parameters. Another review done by Parth Shah[5] and others, they have used inception v3 and LSTM in their show and tell model. Inception v3 was chosen as a CNN feature extraction model for the project.

## II. DATASET

Taking into account the time and resources available to us in terms of processing power and memory, we decided to use Flickr30k as our dataset for training and testing our model. The Flickr30k dataset has become a standard benchmark for sentence-based image description. It contains 31,000 images collected from Flickr, together with five reference sentences provided by human annotators[3]. The five sentences provide variety and help in making the model robust.

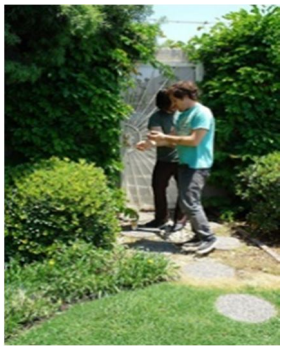


Figure 1. Example dataset image (1000092795.jpg)

Table 1. Caption data for the image 1000092795.jpg

| Image_name | Comment_number | Comment |
|---|---|---|
| 1000092795.jpg | 0 | Two young guys with shaggy hair look at their hands while hanging out in the yard . |
| 1000092795.jpg | 1 | Two young , White males are outside near many bushes . |
| 1000092795.jpg | 2 | Two men in green shirts are standing in a yard . |
| 1000092795.jpg | 3 | A man in a blue shirt standing in a garden . |
| 1000092795.jpg | 4 | Two friends enjoy time spent together . |

## III. CAPTIONING MODEL

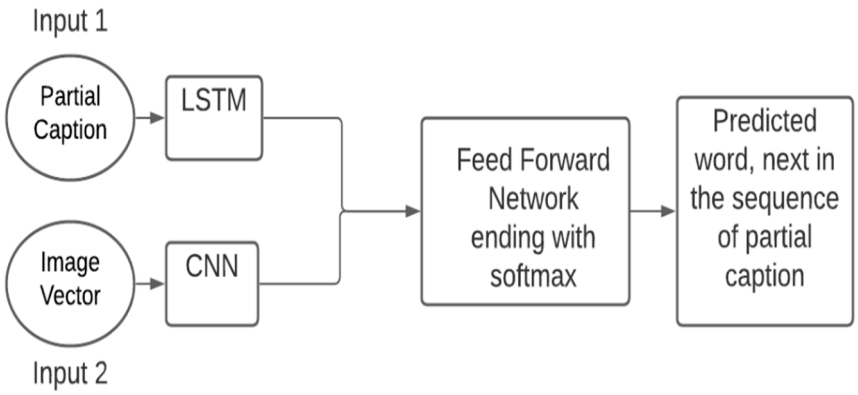



Figure 2. Captioning model overview

The task of image captioning can be divided into two modules logically – one is an image-based model – which extracts the features and nuances out of our image, and the other is a language-based model – which translates the features and objects given by our image-based model to a natural sentence.

For our image-based model (viz encoder) – we usually rely on a Convolutional Neural Network model. And for our language-based model (viz decoder) – we rely on a Recurrent Neural Network.

We have opted for transfer learning by using a pretrained CNN namely InceptionV3 to extract the features from our input image. The image is first reshaped to the proper size(299x299) as required and is fed into the InceptionV3 model which gives a fixed-length feature vector.

In the Flickr30k dataset, each image has five captions describing them. Since we are using supervised learning, these captions are used as the target results during our training phase. The captions are preprocessed by removing punctuations, special characters, indexing etc. The feature vector along with its already predicted words is then fed into the RNN/LSTM to get the next word. Corresponding changes are made to the weights based on the result predicted and the actual result. This process is repeated until the model reaches the end of the sentence.

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429*
*Volume 9 Issue VII July 2021- Available at www.ijraset.com*

## IV. CAPTION GENERATOR SYSTEM ARCHITECTURE

### *A. Image Preprocessing*

Given images can be of any size but the image will be converted into 299x299 shape to feed the InceptionV3 since the InceptionV3 model requires the mentioned size.

### *B. Text Preprocessing*

Each image in the dataset has 5 corresponding captions associated with it. Basic preprocessing like removing special tokens, eliminating words which contain numbers and lowercasing all the alphabets was done. All the unique words were taken from the captions to generate the vocabulary which was further reduced based on a threshold frequency of the number of occurrences of the words in all the captions.

### *C. Feature Detection Module*

The image vector given as input to the InceptionV3 model will generate a feature vector of shape (1, 2048) is converted to (2048, ). We have opted for transfer learning by using a pretrained CNN namely InceptionV3 to extract the features from our input image. InceptionV3 is pretrained with ImageNet dataset to classify images.

| Layer (type) | Output Shape | Param # |
|---|---|---|
| input_4 (InputLayer) | (None, 34) | 0 |
| input_3 (InputLayer) | (None, 2048) | 0 |
| embedding_2 (Embedding) | (None, 34, 200) | 330400 |
| dropout_3 (Dropout) | (None, 2048) | 0 |
| dropout_4 (Dropout) | (None, 34, 200) | 0 |
| dense_2 (Dense) | (None, 256) | 524544 |
| lstm_2 (LSTM) | (None, 256) | 467968 |
| add_2 (Add) | (None, 256) | 0 |
| dense_3 (Dense) | (None, 256) | 65792 |
| dense_4 (Dense) | (None, 1652) | 424564 |

Total params: 1,813,268
Trainable params: 1,813,268
Non-trainable params: 0

Figure 3. Model Summary

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429*
*Volume 9 Issue VII July 2021- Available at www.ijraset.com*

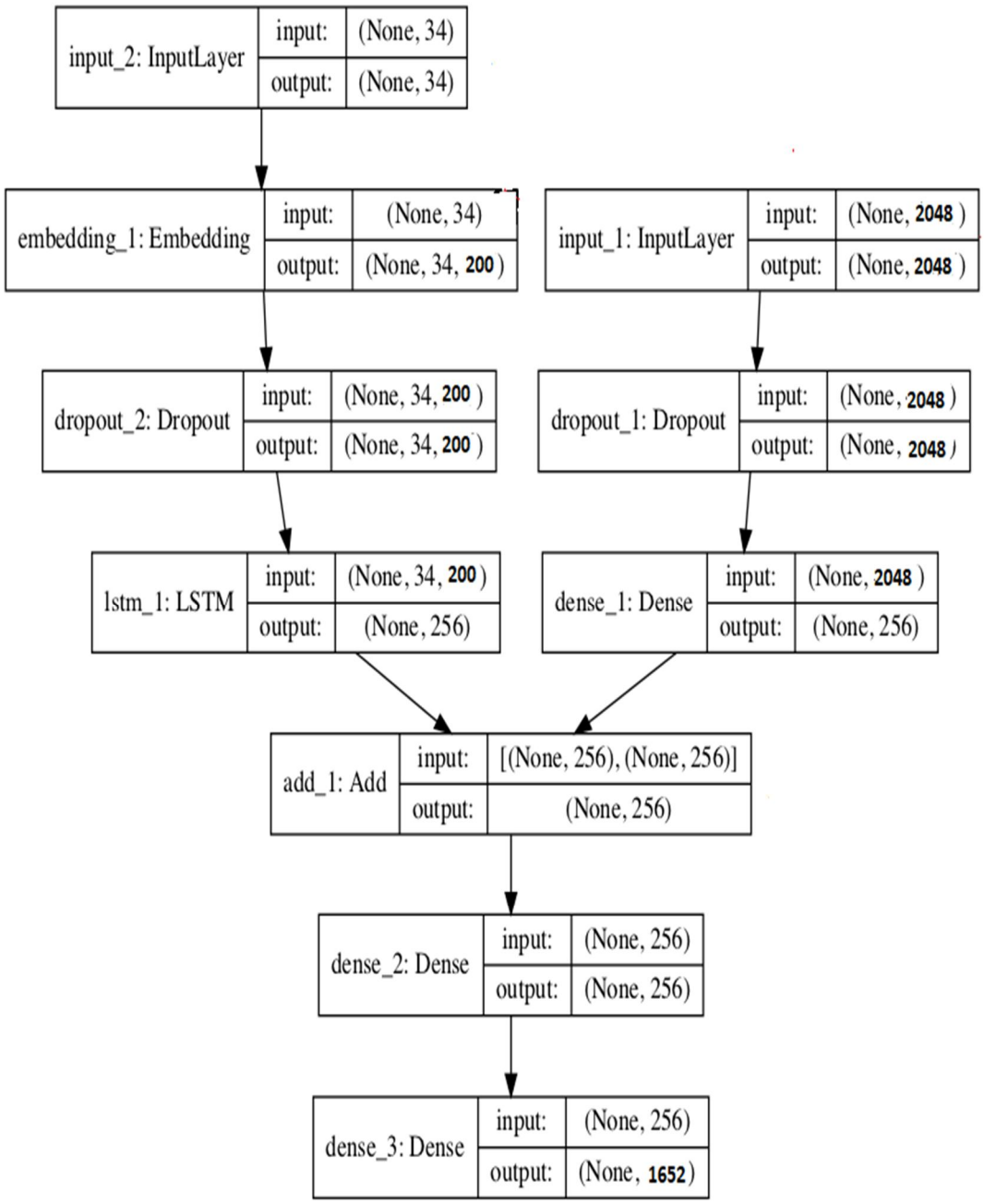



Figure 4. Model architecture

*D. Caption Generation Module*

Lazy loading using a generator function was implemented to run the model efficiently with a huge dataset of 30000 images. An embedding matrix will be used to fit all the words in the current caption corpus and is fed into the network along with the image feature vector. In order to form a sentence for the image, we need to know the features of the image and feed them into our gated RNN (LSTM). The CNN (InceptionV3) model is used to extract the features of the image which produces a fixed-size vector. The LSTM uses the fixed-size vector and the captions given for the image and produces the next word in the sentence as the result. The idea of taking a word from the given caption also as an input along with the image vector ensures that the result obtained is much similar to the caption in terms of its grammar and sentence structure. It helps in making the sentence more meaningful and legible compared to forming a sentence with just classification of objects in the image.

Table 2. Training Process

| | | Xi | Yi |
|---|---|---|---|
| i | Image feature vector | Partial Caption | Target word |
| 1 | Image_1 | startseq | the |
| 2 | Image_1 | startseq the | black |
| 3 | Image_1 | startseq the black | cat |
| 4 | Image_1 | startseq the black cat | sat |
| 5 | Image_1 | startseq the black cat sat | on |
| 6 | Image_1 | startseq the black cat sat on | grass |
| 7 | Image_1 | startseq the black cat sat on grass | endseq |
| 8 | Image_2 | startseq | the |
| 9 | Image_2 | startseq the | white |
| 10 | Image_2 | startseq the white | cat |
| 11 | Image_2 | startseq the white cat | is |
| 12 | Image_2 | startseq the white cat is | walking |
| 13 | Image_2 | startseq the white cat is walking | on |
| 14 | Image_2 | startseq the white cat is walking on | road |
| 15 | Image_2 | startseq the white cat is walking on road | endseq |

*E. Deployment Model*

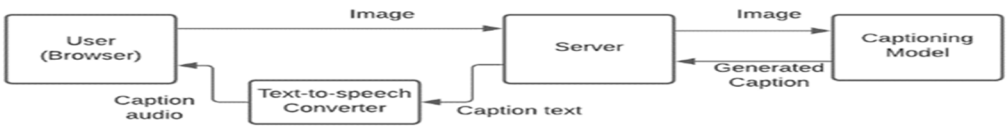



Figure 5. Deployment Model

The system consists of four components namely, user (Browser), server, captioning model and text-to-speech converter. The above figure shows the overall flow of execution.

The user first uploads the image for which they want the caption from their browser. It is then processed and sent to the Captioning Model for generating the description. The model processes the image using predefined weights and generates a caption for the given image which is sent back to the user. The user also has the option of converting the generated caption into audio form by using the Text-to-speech converter which reads the caption text. The goal of our design is to keep user interaction to a minimum and simple with our website.

*F. Text-to-Speech Generator*

Using an off-the-shelf TTS engine seemed a better and faster way to achieve TTS. We used a third party free TTS service provider responsivevoice.org API by responsivevoice.org to implement the TTS in the system. Users can choose to hear the caption out loud if they are unable to read the caption due to their visual impairment.

## V. RESULT ANALYSIS

Initially we had decided to test our model accuracy using BLEU scores but later as the project progressed and we started learning more about the BLEU scores core implementation and got to know that BLEU scores will be given based on the word presence and only shallow comparisons will be made as opposed to deep comparisons where we compare the meanings of sentences instead of just word occurrence count. This kind of implementation affects our sentence predictions negatively since even if the sentences are relevant to the images, using different words with similar meaning would drastically reduce the BLEU scores for our model. Hence, we decided not to include BLEU scores for the result analysis of our model. Instead of that, peer rating would appropriately fit the purpose of result analysis of our model. We prepared google forms with randomly picked caption image-caption pairs and let random people rate the relevance of the caption to the image. 20 images were chosen and 75 people rated the caption relevance to the image. The rating values that were considered to be whole numbers from 1 to 5 with 1 being the least relevant caption to the image and 5 being the most relevant caption to the image. Since 75 ratings were received for 20 images each, a total of 1500 ratings were gathered. These ratings were averaged to get the value of 3.223333333 out of 5 or 64.46% of relevance score. Below figures (figure 6,7,8) are some of the results obtained by our model.

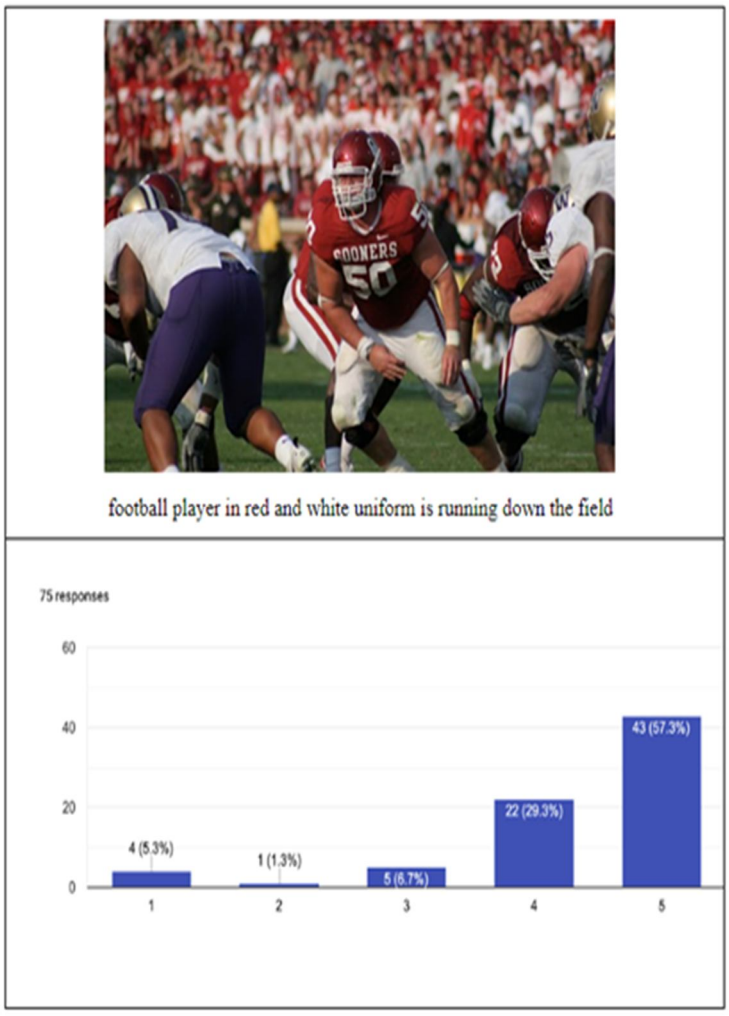



Figure 6. Sample input 1

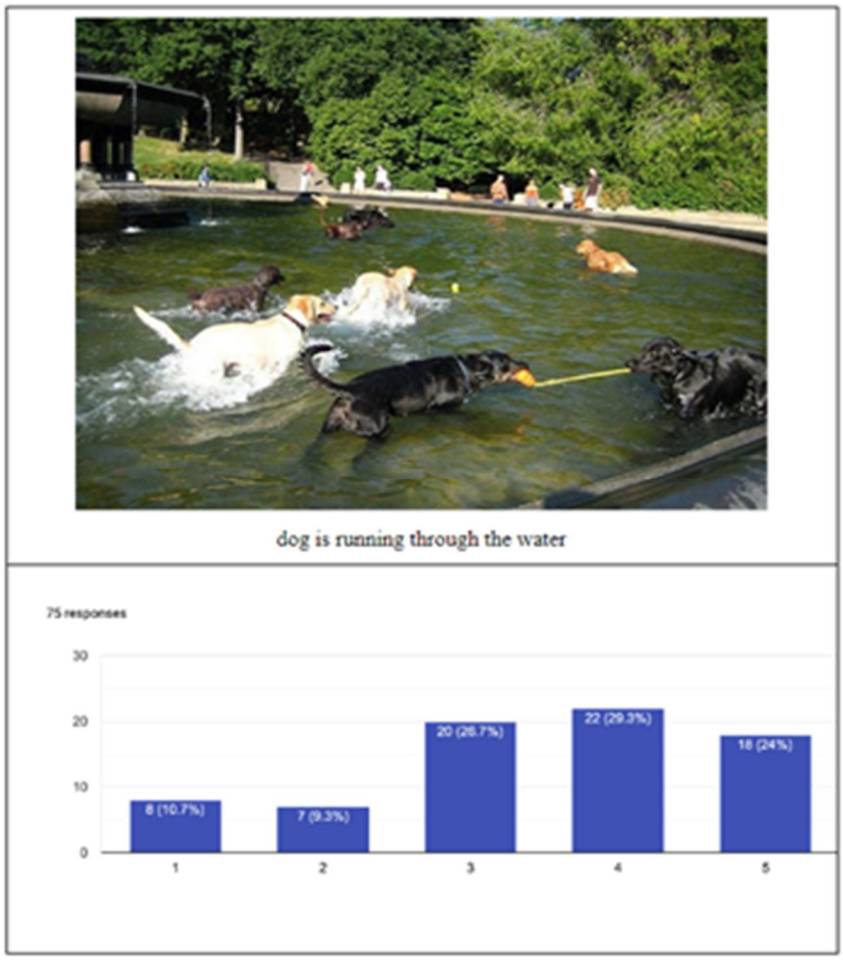



Figure 8. Sample input 2

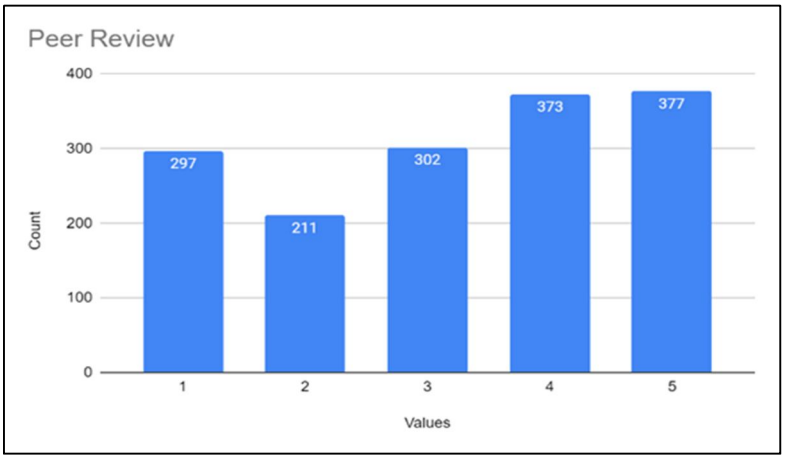


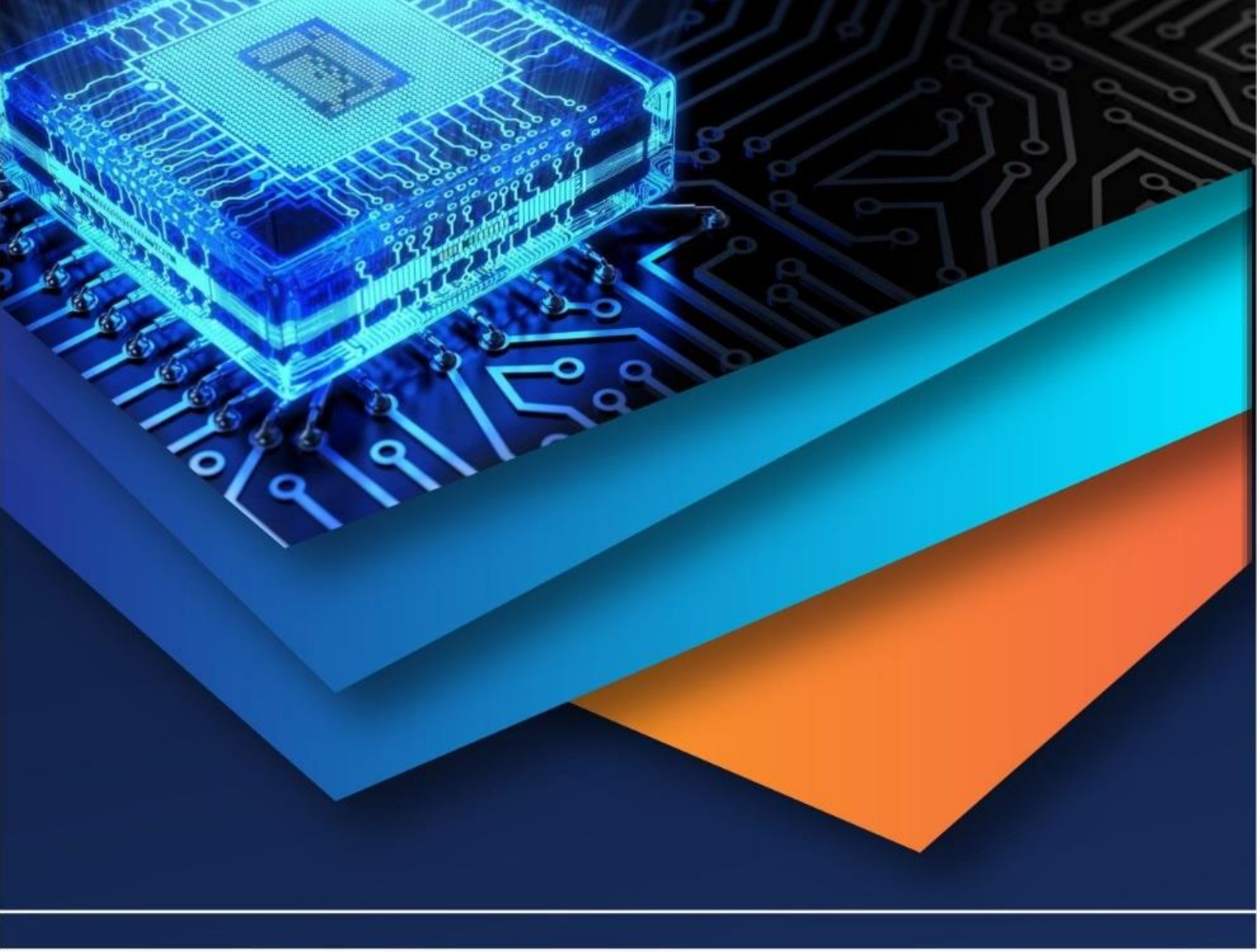

Figure 6. Peer review analysis summary

## VI. CONCLUSION

Image captioning has many advantages in almost every complex area of Artificial Intelligence. The main use case of our model is to help visually impaired people browse the internet and make it easy for them to comprehend images on the internet. We have used pre-trained models and powerful deep learning frameworks such as Convolutional Neural Network and Long Short-Term Memory to train and caption the images. We have mainly trained this dataset on Flickr30k and have received satisfying results for the most part.

## REFERENCES

[1] Ahmed, Faruk & Mahmud, Md Sultan & Al-Fahad, Rakib & Alam, Shahinur & Yeasin, M.. (2018). Image Captioning for Ambient Awareness on a Sidewalk. 85-91. 10.1109/ICDIS.2018.00020.
[2] Mullachery, V., & Motwani, V. (2018). Image Captioning. ArXiv, abs/1805.09137.
[3] Xu, K. et al. Show, attend and tell: Neural image caption generation with visual attention. In Proc. International Conference on Learning Representations. ArXiv, abs/1502.03044 (2015).
[4] Chohan, Murk & Khan, Adil & Mahar, Muhammad & Hassan, Saif & Ghafoor, Abdul & Khan, Mehmood. (2020). Image Captioning using Deep Learning: A Systematic Literature Review. International Journal of Advanced Computer Science and Applications. 11. 10.14569/IJACSA.2020.0110537.
[5] Shah, Parth & Bakrola, Vishvajit & Pati, Supriya. (2017). Image Captioning using Deep Neural Architectures. 10.1109/ICIIECS.2017.8276124.
[6] Yang, L., & Hu, H. (2019). Adaptive syncretic attention for constrained image captioning. Neural Processing Letters, 50(1), 549-564.
[7] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research 47 (2013), 853–899.
[8] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. 2015. Guiding the long-short term memory model for image caption generation. In Proceedings of the IEEE International Conference on Computer Vision. 2407–2415.
[9] L. Anne Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell, “Deep compositional captioning: Describing novel object categories without paired training data,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1–10.
[10] K. Shinohara, C. L. Bennett, and J. O. Wobbrock, “How designing for people with and without disabilities shapes student design thinking,” in Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility. ACM, 2016, pp. 229–237.
[11] A. M Chandrashekhar , K. Raghuveer, “Diverse and Conglomerate Modi-operandi for Anomaly Intrusion Detection Systems”, International Journal of Computer Application (IJCA**)** Special Issue on “Network Security and Cryptography (NSC)”, 2011.
[12] Puneeth L Sankadal , A. M. Chandrashekhar , Prashanth Chillabatte, “Network Security situation awareness system” International Journal of Advanced Research in Information and Communication Engineering(IJARICE), Volume 3, Issue 5, May 2015.
[13] Sanjana K G , A. M. Chandrashekhar, " comparison of techniques used to provide data security in cloud" international journal for research and development in technology, volume 7, issue 6,June 2017
[14] A. M. Chandrashekhar, Arpitha M.G."Big data challenges and it's tools" international Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), Vol 6, issue 6,June 2017.
[15] Huda Mirza Saifuddin, A. M. Chandrashekhar, Spoorthi B.S, “Exploration of the ingredients of original security” International Journal of Advanced Research in Computer Science and Applications(IJARCSA), Volume 3, Issue 5, May 2015.
[16] A.M.Chandrashekhar, Thejaswini.S, "Comparative analysis of Indoor Positioning System Using Bluetooth and Wi-Fi",International Journal for Innovative Research in Science & Technology (IJIRST),Volume 4,Issue 1,June 2017.
[17] Koushik P, A. M. Chandrashekhar, Jagadeesh Takkalakaki, “Information security threats, awareness and coginizance” International Journal for Technicle research in Engineering(IJTRE), Volume 2, Issue 9, May 2015.
[18] A. M. Chandrashekhar, Ngaveni Bhavi, Pushpanjali M K, “Hierarchical Group Communication Security”, International journal of Advanced research in Computer science and Applications (IJARCSA), Volume 4,Issue 1,Feb-2016
[19] A. M. Chandrashekhar, Jagadish Revapgol, Vinayaka Pattanashetti, “Security Issues of Big Data in Networking”, International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Volume 2, Issue 1, JAN-2016.
[20] A. M. Chandrashekhar, Chitra K V, Sandhya Koti, “Security Fundamentals of Internet of Things”, International Journal of Research (IJR), Volume 3, Issue no1, JAN-2016
[21] Sowmyashree K.K, A. M. Chandrashekhar, Sheethal R.S, “Pyramidal aggregation on Communication security” International Journal of Advanced Research in Computer Science and Applications (IJARCSA), Volume 3, Issue 5, May 2015.
[22] Yadunandan Huded, A. M. Chandrashekhar, Sachin Kumar H S, “Advances in Information security risk practices” International Journal of Advanced Research in data mining and Cloud computing (IJARDC), Volume 3, Issue 5 May 2015.
[23] Prashanth G M, A. M. Chandrashekhar, Anjaneya Bulla, “Secured infrastructure for multiple group communication” International Journal of Advanced Research in Information and Communication Engineering (IJARICE), Volume 3, Issue 5, May 2015.
[24] A. M. Chandrashekhar, Jagadish Revapgol, Vinayaka Pattanashetti, “Security Issues of Big Data in Networking”, International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Volume 2, Issue 1, JAN-2016.
[25] A. M Chandrashekhar , Naveen J., Chethana S., Charith S. “Tackling Counterfeit Medicine Through an Automated System Integrated with QR Code”. In: Hemanth J., Bestak R., Chen J.IZ. (eds) Intelligent Data Communication Technologies and Internet of Things. Lecture Notes on Data Engineering and Communications Technologies, vol 57. Springer, Singapore.
[26] A M. Chandrashekhar, Muktha G,”On Demand Feedback Analysis for Certification Process”, International Journal on Recent and Innovation Trends in Computing and Communication,ISSN: 2321 ,Volume: 5.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY