Image Captioning using Artificial Intelligence

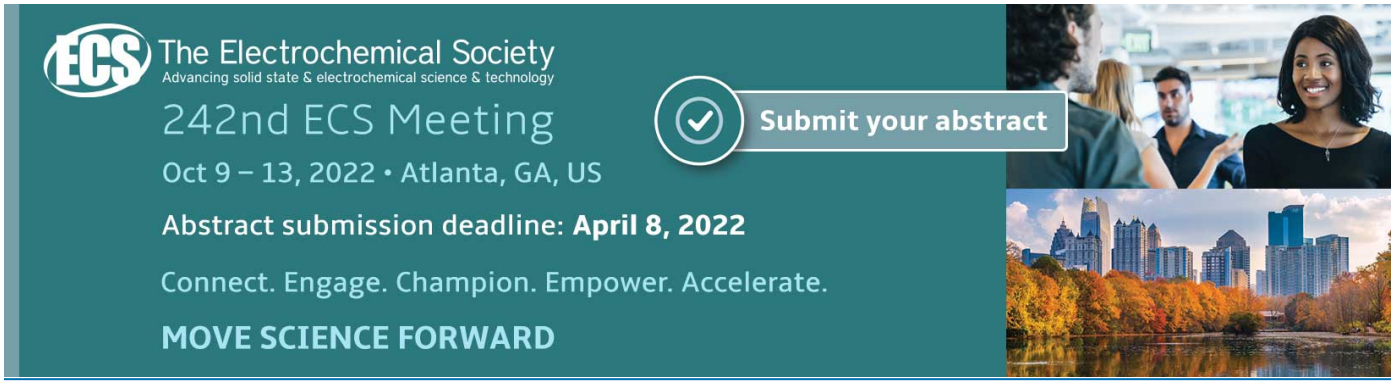


Image Captioning using Artificial Intelligence

[1]Yajush Pratap Singh, [2]Sayed Abu Lais Ezaz Ahmed, [3,*]Prabhishek Singh,
[4]Neeraj Kumar, [5]Manoj Diwakar

[1, 2, 3]Department of CSE, Amity School of Engineering and Technology,
Amity University Uttar Pradesh, Noida, India
[4]Sr. Editor, UoE Publication, India
[5]Department of CSE, Graphic Era deemed to be University, Dehradun, India

*Corresponding author email: prabhisheksingh88@gmail.com

***Abstract*—** In modern science there is a rapid development of artificial intelligence, image processing has gradually fascinated and inspired the attention of many researchers in the field of artificial intelligence and has become an interesting and demanding task. The main idea of Image caption is to automatically generate natural language descriptions according to the information observed in an image, this is an important portion of scene understanding, which combines all the knowledge and information available of computer vision and natural language processing. The use of image caption is broad and noteworthy, for example, the understanding of human-computer collaboration. This paper reviews the related methods and focuses on the attention mechanism, which plays a vital role in computer vision and is broadly used in image caption generation tasks. Furthermore, the advantages and the shortcomings of these methods are discussed, providing the commonly used datasets and evaluation criteria in this field. Finally, this paper proposes some open challenges in the image caption task.

***Keywords*—**image captioning, artificial intelligence, encoder, decoder, CNN

1. Introduction

Consistently, we experience an enormous number of pictures from different sources, for example, the web, news stories, archive graphs and promotions. These sources contain pictures that watchers would need to decipher themselves. Most pictures don't have a portrayal; however the human can generally comprehend them without their definite inscriptions. Be that as it may, machine needs to decipher some type of picture inscriptions if people need programmed picture subtitles from it.

1.1 Image Captioning

Since the time analysts began taking a shot at object acknowledgment in pictures, it turned out to be evident that just giving the names of the items perceived doesn't establish such a decent connection as a full human-like depiction. However long machines don't think, talk, and carry on like people, regular language depictions will stay a test to be illuminated. There have been numerous varieties and blends of various methods since 2014.

The main methodologies can be arranged into two streams. One stream takes a start to finish, encoder-decoder structure received from machine interpretation. For example, utilized a CNN to extricate significant level picture highlights and afterward took care of them into a LSTM to create subtitle went above and beyond by presenting the consideration component. The other stream applies a compositional structure. For instance, separated the subtitle age into a few sections: word identifier by a CNN, inscription competitors' age by a most extreme entropy model, and sentence re-positioning by a profound multimodal semantic model.

1.2 Text to Speech (TTS)

A book to-discourse (TTS) framework changes over typical language text into discourse. To begin with, it changes over crude content containing images like numbers and truncations into what might be compared to worked out words and partitions and denotes the content into prosodic units like

expressions, provisos, and sentences. At that point the synthesizer changes over the emblematic semantic portrayal into sound.

Text to Speech has for some time been a fundamental assistive innovation apparatus and its application here is critical and boundless. It permits ecological boundaries to be taken out for individuals with a wide scope of incapacities. The longest application has been in the utilization of screen perusers for individuals with visual impedance, however text-to-discourse frameworks are presently generally utilized by individuals with dyslexia and other perusing challenges just as by pre-proficient youngsters. They are likewise regularly utilized to help those with extreme discourse debilitation normally through a committed voice yield correspondence help.

1.3 Flutter

Flutter is an open-source UI programming advancement unit made by Google. It is utilized to create applications for Android, iOS, Windows, Mac, Linux, Google Fuchsia and the web. Flutter applications are written in the Dart language and utilize a considerable lot of the language's further developed highlights. On Windows, macOS and Linux through the semi-official Flutter Desktop Embedding venture, Flutter runs in the Dart virtual machine which includes an in the nick of time execution motor. While composing and investigating an application, Flutter utilizes Just in Time accumulation, considering "hot reload", with which changes to source records can be infused into a running application. Ripple expands this with help for stateful hot reload, where as a rule changes to source code can be reflected quickly in the running application without requiring a restart or any loss of state. This element as actualized in Flutter has gotten far reaching acclaim. UI plan in Flutter includes utilizing arrangement to amass/make "Gadgets" from different Widgets. The secret to understanding this is to understand that any tree of segments (Widgets) that is collected under a solitary form () technique is additionally alluded to as a solitary Widget. This is on the grounds that those more modest Widgets are likewise comprised of considerably more modest Widgets, and each has a form () technique for its own. This is the manner by which Flutter utilizes Composition.

Caption generation of an image is an interesting artificial intelligence process where we produce a descriptive sentence for a given image or sign. Image captioning has numerous applications such as in editing applications (encoding and decoding image fragments), usage in virtual assistants, for image indexing, for visually impaired persons, for social media, and several other natural language processing applications. Few of the applications are discussed here in our literature survey.

1.4 Uses

- There are many NLP applications right now, which extract insights/summary from a given text data or an essay etc. The same benefits can be obtained by people who would benefit from automated insights from images.
- A slightly (not-so) long term use case would definitely be, explaining what happens in a video, frame by frame.
- Would serve as a huge help for visually impaired people. Lots of applications can be developed in that space.
- Social Media. Platforms like facebook can infer directly from the image, where you are (beach, cafe etc), what you wear (color) and more importantly what you’re doing also (in a way). See an example to understand it better.
- Skin Vision: Lets you confirm whether a skin condition can be skin cancer or not.
- Google Photos: Classify your photo into Mountains, sea etc.
- Facebook: Using AI to classify, segmentate and finding patterns in pictures.
- A U.S. company is predicting crop yield using images from satellite.
- Picasa: Using facial Recognition to identify your friends and you in a group picture.
- Tesla/Google Self Drive Cars: All the self-drive cars are using image/video processing.
- Automatic Image captioning requires both Image analysis and neural network.

1.5 Challenges Faced In Image Captioning:

1.5.1 Problem: The problem of generating natural language descriptions of an image to describe the visual content has received much interest in the fields of computer vision and natural language

processing, driven by applications such as image indexing or retrieval, virtual assistants, image understanding and support of the visually impaired people.

The AI-based Image Captioning instrument creates comprehensible inscriptions or printed portrayals in the wake of understanding the pictures dependent on singular segments of the item and activities taken in them.

For any semblance of media and distributing organizations, creating n number of substance pieces every day, inscribing pictures has been a manual exertion up until this point. It requires critical exertion when there is a high volume of pictures that are distributed online in light of the fact that creating inscriptions that precisely depict the item and their relationship with their environmental factors isn't simple.

To repeat this conduct day by day and putting significant chance to it, at that point, turns into another test itself. Something that can be mechanized with the assistance of AI and neural organizations.

1.6 Solution

To defeat this issue, we have concocted a picture subtitling instrument dependent on Artificial Intelligence and Convolutional Neural Networks to disentangle and computerize the inscription age.

The device has been prepared adequately with in excess of 10,000 pictures. With expanding mindfulness on openness and giving better client experience, the device can be utilized for both picture subtitling and creating alt text. Intended to convey precise and solid picture subtitling, the outcomes can be utilized to give.

Undertakings can utilize it, either by Transferring a picture from their information base or give a URL to the apparatus, or Incorporating their preferred device with the CMS programming. The instrument, distinguishes the picture, first based on labels. It pre-measures the information and tokenizes the inscriptions. Part of the arrangement that we have constructed depends on calculations, AWS containerized arrangement, flexible compartments, an article distinguishing proof module, and API-based functionalities. This picture subtitling device will mechanize the errand of interpreting the picture to depict them in regular sentences, improving work process and effectiveness. Man-made reasoning can gain constantly from past encounters and adjust to changes-production it the most appropriate for creating important inscriptions in the long run.

Fig 1- Shows a patient detecting skin infection using image captioning app

Fig 2- It is a social networking app which uses image captioning method

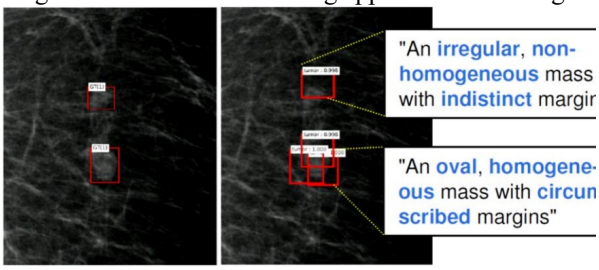


Fig 3- We can use image captioning to check any internal complication in our body.

2. Methodology

Before a picture is worth a thousand words. A human can explain a complex picture without much thought. But for a machine it is much more complex as machines cannot just explain a complex picture. But in recent time, we've gotten a bit nearer – various AI frameworks that can consequently create subtitles to precisely depict pictures the first occasion when it sees them were built. This sort of framework could inevitably enable specially challenged individuals to get pictures, give captions content to pictures in parts of the reality while making it simpler for everybody to look on Google for pictures. Ongoing exploration has enormously improved item identification, characterization, and naming. However, precisely portraying a complex picture requires a more profound portrayal of what's happening in the scene, catching how the different how the different elements in the pictures are in context to each other and making an interpretation of everything into common sounding language.

At first the objectives and aim of the Research paper is understood. After understanding the main objective of the Research paper, different past works were researched in search of finding the best model for preparing the image captioning model. Different model architectures were studied to find the best architecture which gives the maximum accuracy for the captions generated by the machines. As already discussed, it is really difficult for the machines to generate captions for very complex pictures so it is really important to find a model which can generate captions for complex pictures as accurate as possible. Different machine learning and AI algorithms associated with image processing and image captioning were studied. Different research papers and articles about those algorithms were studied to find the best possible way of finding the captions for complex pictures as much as possible [15-18].

From earlier time there was much efforts to construct computer-generated natural captions of pictures combining current progressive techniques in each computer vision and linguistic communication process to create an entire image description approach. However, in recent times the combination of recent computer vision and language models into one trained model to taking the picture and directly manufacturing an individual's decipherable sequence of words to explain it. This thought originates from late advances in machine interpretation between dialects, where a Recurrent Neural Network (RNN) changes, state, a French sentence into a vector portrayal, and a second RNN utilizes that vector portrayal to create an objective sentence in German [19-22].

Presently, imagine a scenario where we supplanted that first RNN and its information words with a profound Convolutional Neural Network (CNN) prepared to order protests in pictures. Regularly, the CNN's last layer is utilized in a last Softmax among known classes of articles, allotting a likelihood that each item may be in the picture. In any case, in the event that we eliminate that last layer, we can rather take care of the CNN's rich encoding of the picture into a RNN intended to deliver phrases. We would then be able to prepare the entire framework straightforwardly on pictures and their subtitles, so it expands the probability that portrayals it delivers best match the preparation depictions for each picture [23-24].

3. Model Architecture

In this image captioning model we will be using the combination of two separate architecture that is Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). In this model we will be using a special kind of RNN which is LSTM (Long Short Term Memory). We will be using LSTM because it is a special kind of RNN which uses memory cell as result we will be able to store the information for a longer time. Fundamentally, CNN is utilized to create include vectors from the spatial information in the pictures and the vectors are taken care of through the completely associated straight layer into the RNN engineering so as to produce the consecutive information or grouping of words that in the end create depiction of a picture by applying different picture preparing strategies to discover the pattern in a picture.

3.1 Encoder-CNN

Presently, we're utilizing the CNN as an element extractor that packs the enormous measure of extraction contained in the first picture into a more modest portrayal. This CNN is regularly called the encoder on the grounds that it encodes the substance of the picture into a more modest feature vector. At that point we can handle this feature vector and use it as an underlying contribution to the following RNN.

3.2 Decoder-RNN

The work of the RNN is to decode the process vector and transform it into a succession of words. Accordingly, this segment of the organization is regularly called a decoder.

3.3 Tokenizing Captions

The RNN part of the captioning network is prepared on the subtitles in the Flikcr8 k dataset. We're expecting to prepare the RNN to foresee the following expression of a sentence dependent on past words. Yet, how precisely would it be able to prepare on string information? Neural nets don't do well with strings. They need an all-around characterized mathematical alpha to successfully perform back-propagation and figure out how to create comparable yield. In this way, we need to change the captions related with the picture into a rundown of tokenized words. This tokenization transforms any string into a rundown of words.

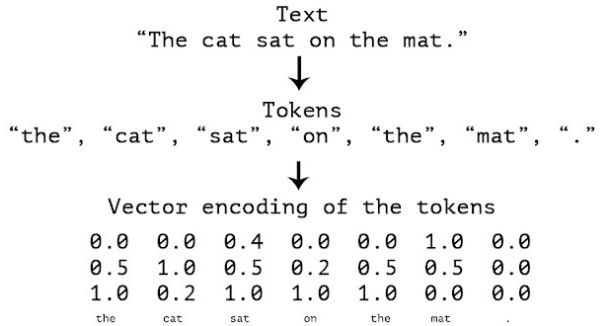


Fig 4. Tokenizing Captions

3.4 Working of Tokenization

To start with, we iterate through the entirety of the preparation subtitles and make a word reference that maps all novel words to a mathematical file. Thus, every word we go over will have a relating whole number worth that can discover in this word reference. The words in this word reference are alluded to as our vocabulary. The vocabulary ordinarily additionally incorporates a couple of uncommon tokens.

3.5 Embedding Layer

There's one more advance before these words get sent as contribution to a RNN and that is the embedding layer, which changes each word in a subtitle into a vector of an ideal steady shape.

3.6 Words to Vectors

Now, we realize that we can't feed words into a LSTM now and anticipate that it should have the option to produce the correct option. These words initially should be transformed into a mathematical portrayal so an organization can utilize typical misfortune capacities and enhancers to compute how "close" an anticipated word and ground truth word are? In this way, we normally transform a grouping of words into a succession of mathematical qualities; a vector of numbers where each number guides to a particular word in our vocabulary.

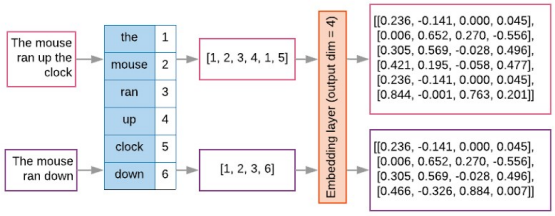


Fig 5. Encoding layer

3.7 Training the RNN Decoder Model with suitable parameters

The Decoder will be made of LSTM cells which is useful for recalling the extensive successions of words. Each LSTM cell is hoping to see a similar state of the information vector at each time-step. The absolute first cell is associated with the yield include vector of the CNN encoder. The contribution to the RNN for all future time steps will be the individual expressions of the preparation inscription. Along these lines, toward the beginning of preparing, we have some contribution from our CNN, and LSTM cell with starting state. Presently the RNN has two duties:

- To Remember spatial data from the input feature vector.
- To Predict the following word.

We realize that the absolute first word it produces ought to consistently be the <start> token and the following word ought to be those in the preparation subtitle. At each time step, we take a gander at the current subtitle word as info and consolidate it with the shrouded condition of the LSTM cell to create a yield. This yield is then passed to the completely associated layer that creates a dispersion that speaks to the most probable next word. We feed the following word in the inscription to the organization, etc until we come to the <end>token. The shrouded condition of a LSTM is a component of the information token to the LSTM and the past state likewise alluded to as the repeat work. The repeat work is characterized by loads and during the preparation cycle, this model uses back-spread to refresh these loads until the LSTM cells figure out how to create the right next word in the inscription given

the current information word. Similarly as with most models, you can likewise exploit clustering the preparation information. The model updates its loads after each preparation group with the bunch size is the quantity of picture subtitle sets sent through the organization during a solitary preparing step. When the model has prepared, it will have gained from many picture caption pairs combines and ought to have the option to create subtitles for new picture information.

3.8 Complete CNN-RNN Architecture

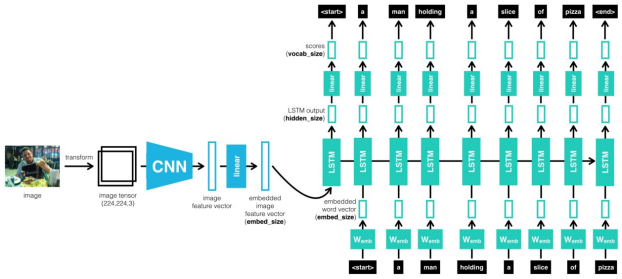
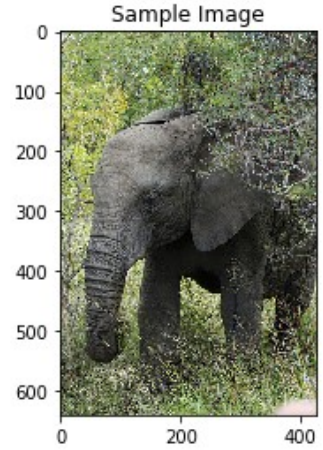


Fig 6. CNN Architecture

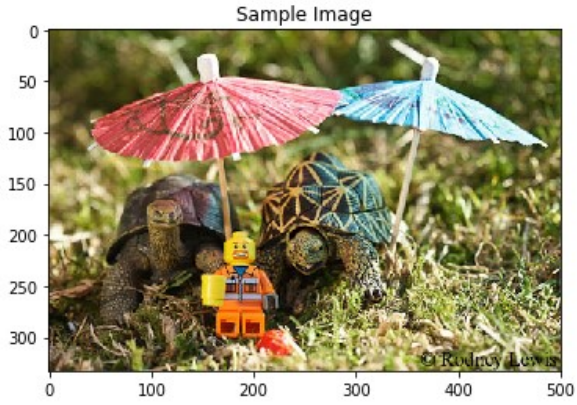


An elephant standing in a field of grass

Fig 7. Sample image (a)

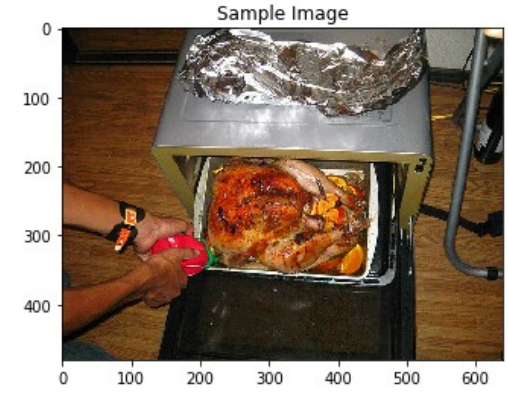
A train is coming down the tracks in a city .

Fig 8. Sample image (b)



A bunch of umbrellas that are sitting in the grass

Fig 9. Sample image (c)



A person is cutting into a pizza with a knife . .

Fig 10. Sample image (d)

4. Results and Discussion

Image captioning model is a model in which we give an image to the system and in return it gives us a sentence in return that is 80 to 90 percent accurate. We achieved this with the help of CNN and RNN deep learning models. We took this Research paper because It has various real world applications associated with it. One of them is that we can search for similar types of images by using this Research paper. One of the other is it can help blind people .All these applications I will be discussing in future scope.

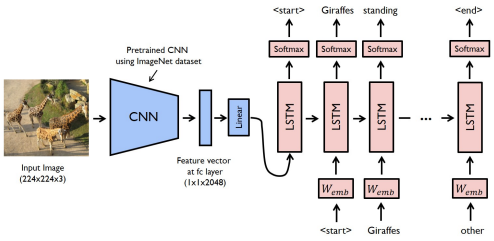


Fig 11. Working

4.1 Training phase

For the first part, CNN extracts the objects and patterns from the images or vector. Now it gives an array of nos. as output that can be used to give as input to RNN. For the second part, we have our source and target. For example, let say we have a sentence as “A dog in garden” and the source comes

out to be['A', 'dog', 'in', 'garden'] and target will be [" 'A', 'dog', 'in', 'the','garden'"]. Using source sequence and the target one we can train our CNN model.

4.2 Test phase

Now comes the testing phase so we ran 8000 images with captions on our model to train our model so that it gives correct output to us. It is somewhat similar to the training phase. One more thing we do is take the BLEU score of the captions that the model generated and that score used to predict how accurately our model is predicting the images .

This as you can see is the architecture of our Image captioning model . What CNN does is it has 3 layers input layer , hidden layer and output layer . So it helps in recognising patterns and gives us features of the image in return . So CNN gives us a feature vector of the images that we are providing in our model and RNN on the other hand takes a sequence of data that CNN gave and in return gives the correct caption.

RNN has a current caption positioner that allows RNN to give a complete sentence.

4.3 Encoder

So talking about our encoder that is our CNN (Convolutional Neural Network) that takes images as input and gives us features of that by using its hidden layers it tries to find a pattern and hence extracting all the objects present in the image . So this step is done again and again like 3 to 4 times and to improve accuracy we usually do this.

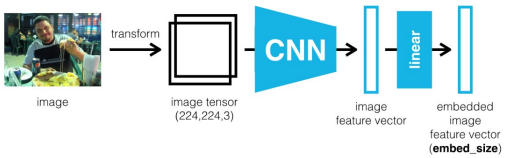


Fig 12. Encoder and decoder

4.4 Decoder

Talking about RNN the array of nos. the generated CNN model is given to RNN for further processing of images and it has multiple layers which can be used to get the caption by using those as input. After passing array of nos. as input to the first hidden layer so it performs some mean operation on it and gives us some output and that output is then taken as input to the next input hidden layer in RNN and at the end finally with the help of current caption positioner we get the output.

5. Conclusion

Image captioning has made huge advances as of late. Late work dependent on profound learning procedures has brought about an advancement in the image captioning. The content depiction of the picture can improve the proficiency of recovery of complex picture, the extending application extent of visual comprehension in the fields of medication, security, military and different fields, which has an expansive application prospect. Simultaneously, the hypothetical structure and examination techniques for picture inscribing can advance the improvement of the hypothesis and utilization of picture comment and visual inquiry replying cross media recovery, video subtitling and video discourse, which has significant scholastic and useful application esteem.

In this paper, we have looked into the various image captioning techniques. We have discussed the various algorithms like CNN, RNN, LSTM. We examined the various earlier methods used and what is their shortcomings and this shortcoming can be overcome.

We discussed the model architecture and the methodology and the how different algorithms can be incorporated to prepare a model with output of maximum accuracy.

Albeit profound learning-based image captioning techniques have accomplished a striking advancement as of late, a powerful picture subtitling strategy that can produce excellent captions for virtually all pictures is yet to be accomplished. With the approach of novel profound learning network models, programmed image captioning will stay a functioning examination region for quite a while.

References

1. J. Aneja, A. Deshpande, and S. Alexander, “Convolutional image captioning,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 2018.
2. T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, “Boosting image captioning with attributes,” in Proceedings of the IEEE Conference on International Conferenceon Computer Vision, pp. 4904–4912, Las Vegas, NV, USA, June 2016.
3. M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, “Areas of attention for image captioning,” in Proceedings of the IEEE Conference on International Conference on Computer Vision, pp. 1251–1259, Venice, Italy, October 2017.
4. H. R. Tavakoli, R. Shetty, B. Ali, and J. Laaksonen, “Paying attention to descriptions generated by image captioning models,” in Proceedings of the IEEE Conference on International Conference on Computer Vision, pp. 2506–2515, Venice, Italy, October 2017.
5. A. Mathews, L. Xie, and X. He, “SemStyle: learning to generate stylised image captions using unaligned text,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 2018.
6. T.-H. Chen, Y.-H. Liao, C.-Y. Chuang, W.-T. Hsu, J. Fu, and M. Sun, “Show, adapt and tell: adversarial training of cross-domain image captioner,” in Proceedings of the IEEE Conference on International Conference on Computer Vision and Pattern Recognition, pp. 521–530, Honolulu, HI, USA, July 2017.
7. C. C. Park, B. Kim, and G. Kim, “Towards personalized image captioning via multimodal memory networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 99, p. 1, 2018.
8. X. Chen, Ma Lin, W. Jiang, J. Yao, and W. Liu, “Regularizing RNNs for caption generation by reconstructing the past with the present,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 2018.
9. R. Zhou, X. Wang, N. Zhang, X. Lv, and L.-J. Li, “Deep reinforcement learning-based image captioning with embedding reward,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1151–1159, Honolulu, HI, USA, July 2017.

10. Q. You, Z. Zhang, and J. Luo, “End-to-end convolutional semantic embeddings,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5735–5744, Salt Lake City, UT, USA, June 2018.
11. A. Aker and R. Gaizauskas, “Generating image descriptions using dependency relational patterns,” in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, vol. 49, no. 9, pp. 1250–1258, Uppsala, Sweden, July 2010.
12. S. Li, G. Kulkarni, T. L. Berg, and Y. Choi, “Composing simple image descriptions using web-scale N-grams,” in Proceeding of Fifteenth Conference on Computational Natural Language Learning, pp. 220–228, Association for Computational Linguistics, Portland, OR, USA, June 2011.
13. Yang, L., & Hu, H. (2019). Adaptive syncretic attention for constrained image captioning. Neural Processing Letters, 50(1), 549-564.
14. Oluwasanmi, A., Aftab, M. U., Alabdulkreem, E., Kumeda, B., Baagyere, E. Y., & Qin, Z. (2019). CaptionNet: Automatic end-to-end siamese difference captioning model with attention, 106773-106783.
15. Singh, Prabhishek, and Raj Shree. "Statistical modelling of log transformed speckled image." International Journal of Computer Science and Information Security 14.8 (2016): 426.
16. Singh, Prabhishek, and Raj Shree. "Quantitative Dual Nature Analysis of Mean Square Error in SAR Image Despeckling." International Journal on Computer Science and Engineering (IJCSE) 9.11 (2017): 619-622.
17. Diwakar, Manoj, and Prabhishek Singh. "CT image denoising using multivariate model and its method noise thresholding in non-subsampled shearlet domain." Biomedical Signal Processing and Control 57 (2020): 101754.
18. Singh, Prabhishek, and Raj Shree. "A New Computationally Improved Homomorphic Despeckling Technique of SAR Images." International Journal of Advanced Research in Computer Science 8.3 (2017).

19. Diwakar, Manoj, et al. "Latest trends on heart disease prediction using machine learning and image fusion." Materials Today: Proceedings (2020).
20. Dhaundiyal, Rashmi, et al. "Clustering based Multi-modality Medical Image Fusion." Journal of Physics: Conference Series. Vol. 1478. No. 1. IOP Publishing, 2020.
21. Kumar, Neeraj, et al. "Flood risk finder for IoT based mechanism using fuzzy logic." Materials Today: Proceedings (2020).
22. Jindal, Muskan, et al. "A novel multi-focus image fusion paradigm: A hybrid approach." Materials Today: Proceedings (2020).
23. Diwakar, Manoj, et al. "A comparative review: Medical image fusion using SWT and DWT." Materials Today: Proceedings (2020).
24. Maurya, Awadhesh Kumar, Ajay Kumar, and Neeraj Kumar. "Improved chain based cooperative routing protocol in wsn." Journal of Physics Conference Series. Vol. 1478. No. 1. 2020.