International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-10 Issue-3, September 2021

Analysis of the Fuzziness of Image Caption Generation Models due to Data Augmentation Techniques

Kota Akshith Reddy, Satish C. J, Polsani Jahnavi, Chintapalli Teja Naveen, Gangapatnam Sai Ananya

Abstract: ***Automatic Image Caption Generation is one of the core problems in the field of Deep Learning. Data Augmentation is a technique which helps in increasing the amount of data at hand and this is done by augmenting the training data using various techniques like flipping, rotating, Zooming, Brightening, etc. In this work, we create an Image Captioning model and check its robustness on all the major types of Image Augmentation techniques. The results show the fuzziness of the model while working with the same image but a different augmentation technique and because of this, a different caption is produced every time a different data augmentation technique is employed. We also show the change in the performance of the model after applying these augmentation techniques. Flickr8k dataset is used for this study along with BLEU score as the evaluation metric for the image captioning model.***

Keywords: Automatic Image, Data Augmentation, Flickr8k dataset, BLEU score.

I.INTRODUCTION

Image Captioning is a complex but a very important task because it not only involves object understanding and also the relation of these objects with their environments. This scene understanding is one of the primary aims of Computer Vision and many big names like Microsoft, Apple, Google, etc are working towards improving these image captioning techniques. On top of understanding the relationships between different objects and their environments, these relations must be expressed in natural language. The fact that this problem requires knowledge from both the fields of CV and NLP makes it one of the more complex problems in the field of Deep Learning. In recent years, with the proliferation of the image data due to increased internet bandwidth and the advancement in the field of sensor technologies, there has been some increased encouragement and enthusiasm in development of some advanced techniques for a wide range of image understanding applications and image captioning is one of them.

Manuscript received on September 03, 2021.
Revised Manuscript received on September 13, 2021.
Manuscript published on September 30, 2021.
* Correspondence Author
Kota Akshith Reddy*, Department of Computer Science, Vellore Institute of Technology, Vellore, Tamil Nadu, India.
Satish C J, Department of Computer Science, Anna University, Tamil Nadu, India.
Jahnavi Polsani, Department of Computer Science, Vellore Institute of Technology, Vellore, Tamil Nadu, India.
Teja Naveen Chintapalli, Department of Computer Science, Vellore Institute of Technology, Vellore, Tamil Nadu, India.
Gangapatnam Sai Ananya, Department of Computer Science, Narayana Engineering College, Tamil Nadu, India.

There has been considerable research done in this particular field [1,2,3,4]. If implemented properly, this type of image captioning system can become a part of and improve the lives of visually impaired people. These systems can also aid in Robotic Vision. Many image captioning models [5,6] tend to have a two stage approach at solving this problem. In the first step, a deep CNN is trained on a large dataset of images to learn features for salient regions in the image. Then in the second step, neural network based captioning models are trained to use these features from the CNN to generate novel captions for the image.

Images taken in the real world are often noisy and may not be perfect for the image captioning model to work on. Some of the flaws in the image can be in the form of a blurry image, rotated image, images with no proper contrast between the object and the surroundings, occlusion, etc. As the first step involves extracting features from the image which are further sent to another neural network for generating the captions, any change in the quality of the image can hamper the ability of the image captioning model to generate relevant captions for the image. There has been a lot of research done on methods to improve the performance of the deep CNN models. One such method is data augmentation [7,8]. Deep CNNs require a lot of data to work with in order to eliminate overfitting [9]. Overfitting occurs when a neural network learns a function which perfectly models the training data, thereby becoming less generalized leading to poor performance on the training data. This can be avoided by giving a lot of data for the model to train upon but that is not always the case with most of the real world problems as data collection can be time consuming and expensive. This problem of limited data can be overcome using data augmentation. It encompasses a range of techniques that enhance the size of the training data and sometimes quality too. Even though data augmentation increases the amount of the training data and helps in increasing the performance of the model, these techniques are not usually used in this particular problem of image captioning mainly because of the two stage process. However some of the works [10,11] show some promising results in enhancing the performance of the captioning models by using data augmentation techniques. In this paper, our main aim is to study the effects of some of the major data augmentation techniques like rotating, flipping, blurring, etc on the performance of the image captioning models and see whether these techniques really help the model in extracting useful features and create relevant captions.

II. RELATED WORK

In this section, we see the overview of the research done on the development and improvement of the image captioning models and also look at the data augmentation approach employed to increase the size of the training data and its effect on image captioning models. If we look from a higher level, the works on image captioning follow one of the following three versions: template based [12,13,14], search based [15,16] and language based [17,18] . The first approach generates a caption using a template after detecting the objects and their relationship with the environment in the image. This approach detects words from image attributes and then classifies them into different parts of a sentence like subject, object, verb, etc and then by using a template, a caption will be generated. In [19], we can see this template-based approach , prepositions which help in determining the kind of relationship between the objects and also the attributes. Finally, sentences are created using this labeling. Similarly in [13], we see a template based approach of sentence generation and in this work a HMM is used to select a quadruplet containing the preposition, scene, verb and noun which describe the image. However, the main drawback of this type of method is that the human-made templates do not work for all types of images.

Search based approaches employ a technique wherein sentences are generated for an image by copying sentences from a pool of human generated sentences for similar images. This approach has an advantage over other methods in achieving human level descriptions because it is essentially copying from human generated sentences. However, it is not feasible to manually write sentences for all images when we decide to scale up. One such work which employs this type of search based image captioning is [15]. Similarity between a sentence and an image is calculated by mapping each to the meaning space and comparing the outcomes. This similarity score is then used to caption the image with a sentence or to find the images that have high similarity with a given sentence.

The third and final approach in image captioning is by using language based models. This can be done in one of the following two ways. The first way is to use a CNN to predict words that are most likely to be associated with or present in the caption [20]. The second way is to make use of the visual representation and translate using a language counterpart. Generally, a CNN is used to obtain the visual representation and a RNN is used to achieve the translation. One of the major advantages of such types of approaches is that they can be trained end to end [21].

Some of the works involving data augmentation in image captioning [10,11] showed promising results and increased the robustness of the model using data augmentation. In [10], data augmentation was used to negate the effects of motion blur and increase the robustness of the image captioning model to the blurring effect. [11] proposes data augmentation to increase the stability of the generated captions. In this work, we are going to find out the effects of data augmentation on image captioning models, thereby making it possible to propose whether these techniques are really helpful or not.

III.METHODS AND METHODOLOGY

A. Data:

The Flickr 8k dataset has been used to train the model. It has a total of 8000 images and each image has 5 captions attached with it. These captions give a clear textual description of the events and entities in the image. The dataset was available at Kaggle and the images were selected by hand to depict the variety of situations and scenes.

B. Model:

We have created a simple baseline model for image captioning using the encoder and decoder architecture without any attention mechanism. This was done for the sole purpose to check the ability of data augmentation to enhance the performance of such simple and baseline models. For feature extraction, we have used the ResNet50 [22] model. This model acts as an encoder and encodes any given image into a 2048 dimensional vector. And finally for the decoding part, LSTMs are used and the vector representations of the images are converted to sentences or textual descriptions. The architecture of the model can be seen in “Fig. 1”.

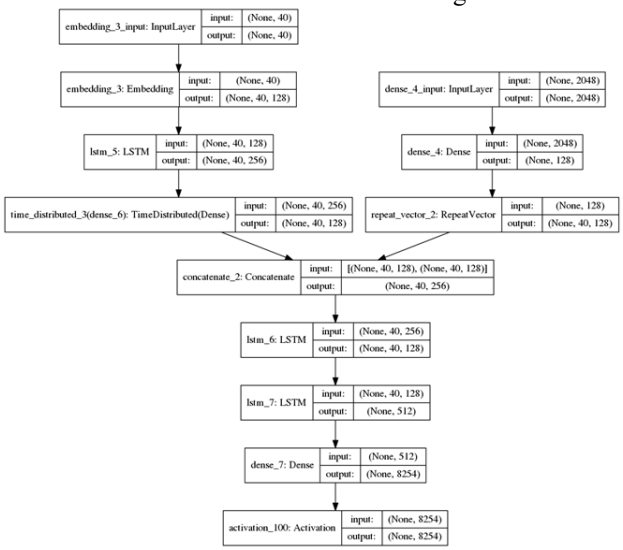

Fig. 1. Architecture of the image captioning model.

This vanilla image captioning model consists mainly of 3 parts: the decoder, the photo feature extractor and the sequence processor. The Photo feature extractor helps in feature extraction. In this work, we have used the ResNet50 model for feature extraction. The sequence processor is nothing but a word embedding layer used to handle text data. The embedding layer is followed by a LSTM layer. Finally, the output of both these components is concatenated and sent to the decoder which again consists of 2 LSTM layers along with a dense and activation layer. The final layer is the vector of softmax probabilities over the entire vocabulary.

IV.RESULTS AND DISCUSSIONS:

We can see the performance of the image captioning model on the Flickr8k dataset before applying any data augmentation technique in Table I.

Table- I: Performance metrics of the image captioning model before data augmentation

Bleu score	Individual 1 gram score	Individual 2 gram score	Individual 3 gram score	Individual 4 gram score	Cumulative 1 gram score	Cumulative 2 gram score	Cumulative 3 gram score	Cumulative 4 gram score
0.55	0.43	0.1	0.03	0.01	0.43	0.18	0.09	0.05

A. Effect of change in brightness on image captioning:

To check whether there is any correlation between the change in the brightness and the ability of the captioning model in creating a caption through the varying brightness levels, we have conducted two experiments: the first one is to check the performance of the model on images whose brightness factor is increased by 2 and the second experiment involved images whose brightness has been reduced by a factor of 0.5. In the first case where the brightness has been increased by a factor of 2, the overall Bleu score has been increased for a total of 502 out of 1001 images- indicating the fact that increasing the brightness is indeed helping the model. However, there were several cases where the overall Bleu score decreased with the increase in the brightness level.

Table- II: Performance metrics of the image captioning model after increasing the brightness of the images by a factor of 2.

Bleu score	Individual 1 gram score	Individual 2 gram score	Individual 3 gram score	Individual 4 gram score	Cumulative 1 gram score	Cumulative 2 gram score	Cumulative 3 gram score	Cumulative 4 gram score
0.56	0.4	0.08	0.02	0.01	0.4	0.15	0.07	0.04

"Fig. 2" and "Fig. 3" show us an image before increasing and after increasing the brightness respectively.

Fig. 2. Before increasing the brightness

Fig. 3. After increasing the brightness

Table- III: Captions generated by the image captioning model for "Fig. 2" and "Fig. 3"

Generated caption for "Fig. 2"	Generated caption for "Fig. 3"
Black dog playing into the beach	Black dog playing in a water

Table- IV: Performance metrics of the image captioning model for "Fig. 2" and "Fig. 3"

Before/After	Bleu score	Individual 1 gram score	Individual 2 gram score	Individual 3 gram score	Individual 4 gram score	Cumulative 1 gram score	Cumulative 2 gram score	Cumulative 3 gram score	Cumulative 4 gram score
Before	0.56	0.5	0.2	0.02	0.03	0.5	0.31	0.13	0.09
After	0.9	0.66	0.02	0.02	0.03	0.66	0.11	0.07	0.05

Table III and Table IV certainly tell us that by increasing the brightness of the image, it is possible to enhance the performance of the model. However, this is not the case for each and every image as there were only 502 images out of 1001 with increased Bleu scores on increasing the brightness. The above experiment illustrates the fact that increasing the brightness of the image can actually help in boosting the performance of the model. We have also documented the performance of the model on images whose brightness has been reduced by a factor of 0.5. Table V shows the performance of the model on the images with reduced brightness levels. Even though decreasing the brightness did not have the same positive effect as increasing the brightness, an increase in Bleu scores was found in 466 out of 1001 images. "Fig. 4" and "Fig. 5" illustrate an example in which we can see this observation.

Table- V: Performance metrics of the image captioning model after decreasing the brightness of the images by a factor of 0.5.

Bleu score	Individual 1 gram score	Individual 2 gram score	Individual 3 gram score	Individual 4 gram score	Cumulative 1 gram score	Cumulative 2 gram score	Cumulative 3 gram score	Cumulative 4 gram score
0.55	0.4	0.09	0.02	0.01	0.4	0.17	0.08	0.05

"Fig. 4" and "Fig. 5" show us an image before decreasing and after decreasing the brightness respectively.

Fig. 4. Before decreasing the brightness

Fig. 5.After decreasing the brightness

Table- VI: Captions generated by the image captioning model for "Fig. 4" and "Fig. 5"

Generated caption for "Fig. 4"	Generated caption for "Fig. 5"
A person on a rock mountain	a person climbing a rock of rocks .

Table- VII: Performance metrics of the image captioning model for "Fig. 4" and "Fig. 5"

Before/After	Bleu score	Individual 1 gram score	Individual 2 gram score	Individual 3 gram score	Individual 4 gram score	Cumulative 1 gram score	Cumulative 2 gram score	Cumulative 3 gram score	Cumulative 4 gram score
Before	0.95	0.83	0.02	0.02	0.03	0.83	0.12	0.07	0.06
After	0.89	0.62	0.01	0.01	0.02	0.62	0.09	0.05	0.04

Table VI and Table VII shows the detrimental effect on the performance of the model when the brightness of the image is decreased. One obvious reason is that an image loses a lot of detail when the brightness is decreased. And yet, the model was able to boost the Bleu scores of 466 images. Therefore, it is sometimes possible to increase the performance of the model by decreasing the brightness of the image.

B. Effect of flipping the images on image captioning:

To check if the flipping of an image has any effect in the performance of the captioning model, the following two experiments were conducted: In the first experiment, the images were flipped horizontally (180 degrees along vertical axis) and in the second experiment, the images were flipped vertically (180 degrees along the horizontal axis). In the first experiment, Bleu score was increased for 461 out of 1001 images- again indicating the fact that flipping an image horizontally, sometimes, helps enhance the performance of the model. Table VIII shows the performance of the model on the horizontally flipped images.

Table- VIII: Performance metrics of the image captioning model after flipping the images horizontally .

Bleu score	Individual 1 gram score	Individual 2 gram score	Individual 3 gram score	Individual 4 gram score	Cumulative 1 gram score	Cumulative 2 gram score	Cumulative 3 gram score	Cumulative 4 gram score
0.54	0.42	0.1	0.03	0.01	0.42	0.18	0.09	0.05

"Fig. 6" and "Fig. 7" show us an image before horizontal flipping and after horizontal flipping respectively.

Fig. 6.Before horizontal flip

Fig. 7.After horizontal flip

Table- IX: Captions generated by the image captioning model for "Fig. 6" and "Fig. 7"

Generated caption for "Fig. 6"	Generated caption for "Fig. 7"
A dog is rushing in the water.	A dog is rushing in the snow .

Table- X: Performance metrics of the image captioning model for "Fig. 6" and "Fig 7"

Before/After	Bleu score	Individual 1 gram score	Individual 2 gram score	Individual 3 gram score	Individual 4 gram score	Cumulative 1 gram score	Cumulative 2 gram score	Cumulative 3 gram score	Cumulative 4 gram score
Before	0.75	0.88	0.07	0.07	0.06	0.88	0.08	0.07	0.07
After	1	1	1	1	1	1	1	1	1

Table VI shows us the generated captions by the model on "Fig. 6" and "Fig. 7". From Table VII, we can see that the model has scored a perfect score of 1.00, whose initial score was a 0.75, after flipping the image horizontally. This illustrates that flipping the image horizontally, in some cases, certainly helps in increasing the performance of the model. We have also documented the results of the model when the images were flipped vertically. We have observed an increase in the Bleu score of 511 images out of 1001 images, which was more than the case of images flipped horizontally. The model was performing better when images were flipped vertically than horizontally. Table XI shows the performance of the model on the images flipped vertically.

Table XI: Performance metrics of the image captioning model after flipping the images vertically.

Bleu score	Individual 1 gram score	Individual 2 gram score	Individual 3 gram score	Individual 4 gram score	Cumulative 1 gram score	Cumulative 2 gram score	Cumulative 3 gram score	Cumulative 4 gram score
0.57	0.38	0.07	0.02	0.01	0.38	0.14	0.07	0.04

"Fig. 8" and "Fig. 9" show us an image before flipping it vertically and after flipping it vertically.

Fig. 8. Before vertical flip

Fig. 9.After vertical flip

Table- XI: Captions generated by the image captioning model for "Fig. 8" and "Fig. 9"

Generated caption for "Fig. 8"	Generated caption for "Fig. 9"
A child in a shirt is carrying a grassy distance in a sand	A little boy on a rock.

Table- XII: Performance metrics of the image captioning model for "Fig. 8" and "Fig. 9"

Before/After	Bleu score	Individual 1 gram score	Individual 2 gram score	Individual 3 gram score	Individual 4 gram score	Cumulative 1 gram score	Cumulative 2 gram score	Cumulative 3 gram score	Cumulative 4 gram score
Before	0.74	0.31	0.006	0.007	0.007	0.31	0.04	0.02	0.01
After	0.91	0.71	0.01	0.02	0.02	0.71	0.1	0.06	0.04

Table XI shows the captions generated by the image captioning model on "Fig. 8" and "Fig. 9". From Table XII, we can clearly see that the model has performed well after the image has been flipped, attaining a Bleu score of 0.91 . Therefore, flipping the image vertically increases the performance of the model in some cases. One noteworthy observation after the above two experiments is that the model has actually performed well on the vertically flipped images when compared with the horizontally flipped images.

C. Effect of blurring on image captioning:

To check the effects of blurring on the performance of the image captioning models, we have blurred the images using a gaussian function with radius 2. It is observed that the Bleu scores of 443 images were increased after blurring and sending the images to the image captioning model. This indicates a positive effect of blurring, even though theoretically blurring should decrease the performance of the model because of an immense loss of detail. This interesting observation signifies the importance of not so used augmentation techniques like blurring in improving the performance of the model.

Table- XIII: Performance metrics of the image captioning model after applying a Guassian Blur function with a radius of 2

Bleu score	Individual 1 gram score	Individual 2 gram score	Individual 3 gram score	Individual 4 gram score	Cumulative 1 gram score	Cumulative 2 gram score	Cumulative 3 gram score	Cumulative 4 gram score
0.55	0.44	0.09	0.02	0.01	0.44	0.18	0.09	0.05

"Fig. 10" and "Fig. 11" show us an image before and after applying a gaussian blur filter with a radius of 2.

Fig. 10. Before blurring

Fig. 11.After blurring

Table- XIV: Captions generated by the image captioning model for "Fig. 10" and "Fig. 11"

Generated caption for "Fig. 10"	Generated caption for "Fig. 11"
A little girl in a flower playground with wooden backyard .	A little girl in a flower playground with wooden backyard .

Table- XV: Performance metrics of the image captioning model for "Fig. 10" and "Fig. 11"

Before/After	Bleu score	Individual 1 gram score	Individual 2 gram score	Individual 3 gram score	Individual 4 gram score	Cumulative 1 gram score	Cumulative 2 gram score	Cumulative 3 gram score	Cumulative 4 gram score
Before	0.47	0.72	0.06	0.04	0.02	0.72	0.06	0.05	0.04
After	0.36	0.54	0.04	0.03	0.02	0.54	0.04	0.04	0.03

Table XIV and Table XV clearly show us the detrimental effect in taking away the detail from the image by deliberately blurring it. But the interesting observation from this experiment is that the overall Bleu score did not change a lot before and after using blurring. However, there was an increase in Bleu scores of approximately half of the images in the dataset- signifying the positive effect blurring might have as a data augmentation technique in image captioning.

D. Effect of rotating an image on image captioning:

To check whether the correlation between rotating an image and the performance of image captioning model on the rotated images, we have conducted an experiment wherein images were rotated by an angle of forty five degrees in the counterclockwise direction. Rotating an image had the least positive effect when compared with the above augmentation techniques in enhancing the performance of the model. Bleu score was increased for 408 out of 1001 images after applying this augmentation technique. Table XVI shows the overall performance of the model on the rotated images.

Table- XVI: Performance metrics of the image captioning model after rotating the images by forty five degrees in the counterclockwise direction

Bleu score	Individual 1 gram score	Individual 2 gram score	Individual 3 gram score	Individual 4 gram score	Cumulative 1 gram score	Cumulative 2 gram score	Cumulative 3 gram score	Cumulative 4 gram score
0.48	0.42	0.07	0.02	0.01	0.42	0.16	0.07	0.04

"Fig. 12" and "Fig. 13" show us an image before and after rotating it by an angle of 45 degrees in the counterclockwise direction.

Fig. 12. Before rotation

Fig. 13.After rotation

Table- XVII: Captions generated by the image captioning model for "Fig. 12" and "Fig. 13"

Generated caption for" Fig. 12"	Generated caption for "Fig. 13"
Two dogs fighting of in the sand .	A group of of mountain climbers climbers across a mountain .

Table- XVIII: Performance metrics of the image captioning model for "Fig. 12" and "Fig. 13"

Before/After	Bleu score	Individual 1 gram score	Individual 2 gram score	Individual 3 gram score	Individual 4 gram score	Cumulative 1 gram score	Cumulative 2 gram score	Cumulative 3 gram score	Cumulative 4 gram score
Before	0.72	0.87	0.05	0.03	0.02	0.87	0.07	0.05	0.04
After	0.42	0.27	0.01	0.01	0.01	0.27	0.05	0.03	0.02

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Table XVII and Table XVIII show us a decline in the performance of the model. One reason might be that there is a lot of blank space in the images after rotation and this might be having a negative effect on the performance of the image captioning model. If we could extrapolate the image to the entire frame, then it might be possible to improve the performance to some extent. And yet, even with the black blobs in the image, the model was still able to improve the Bleu scores of 408 out of 1001 images- showing the importance of rotation as the data augmentation model.

V.CONCLUSION

This work signifies the importance of data augmentation techniques and the extent to which they help increase the performance of the image captioning models. We have analysed the effect of the following data augmentation techniques: Changing the brightness, Flipping- both horizontal and vertical, Blurring and Rotation. The performance metric used to check the performance of the model was the Bleu score and the changes in these scores were observed while employing each and every data augmentation technique. Though there were negative effects on the performance of the model with some techniques like blurring and rotation, they helped in improving the Bleu scores for a significant amount of images- concluding the fact that one has to employ each and every data augmentation technique that is there in the toolshed and see which one works best on the problem at hand.

DECLARATION OF CONFLICTING INTERESTS

The authors declare no conflict of interest in doing this research work.

REFERENCES

1. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., & Yuille, A. (2014). Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632.
2. Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3128-3137).
3. Karpathy, A., Joulin, A., & Fei-Fei, L. (2014). Deep fragment embeddings for bidirectional image sentence mapping. arXiv preprint arXiv:1406.5679.
4. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).
5. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6077-6086).
6. Huang, L., Wang, W., Chen, J., & Wei, X. Y. (2019). Attention on attention for image captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4634-4643).
7. Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621.
8. Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 1-48.
9. Inoue, H. (2018). Data augmentation by pairing samples for images classification. arXiv preprint arXiv:1801.02929.
10. Bujimalla, S., Subedar, M., & Tickoo, O. (2021). Data augmentation to improve robustness of image captioning solutions. arXiv preprint arXiv:2106.05437.
11. Aldabbas, H., Asad, M., Ryalat, M. H., Malik, K. R., & Qureshi, M. Z. A. (2019). Data Augmentation to Stabilize Image Caption Generation Models in Deep Learning.
12. Mitchell, M., Dodge, J., Goyal, A., Yamaguchi, K., Stratos, K., Han, X., … & Daumé III, H. (2012, April). Midge: Generating image descriptions from computer vision detections. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (pp. 747-756).
13. Yang, Y., Teo, C., Daumé III, H., & Aloimonos, Y. (2011, July). Corpus-guided sentence generation of natural images. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (pp. 444-454).
14. Kuznetsova, P., Ordonez, V., Berg, A., Berg, T., & Choi, Y. (2012, July). Collective generation of natural image descriptions. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 359-368).
15. Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010, September). Every picture tells a story: Generating sentences from images. In European conference on computer vision (pp. 15-29). Springer, Berlin, Heidelberg.
16. Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., … & Mitchell, M. (2015). Language models for image captioning: The quirks and what works. arXiv preprint arXiv:1505.01809.
17. You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image captioning with semantic attention. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4651-4659).
18. Yao, T., Pan, Y., Li, Y., & Mei, T. (2017). Incorporating copying mechanism in image captioning for learning novel objects. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6580-6588)
19. Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., … & Berg, T. L. (2013). Babytalk: Understanding and generating simple image descriptions. IEEE transactions on pattern analysis and machine intelligence, 35(12), 2891-2903.
20. Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollár, P., … & Zweig, G. (2015). From captions to visual concepts and back. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1473-1482).
21. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).
22. Chu, Y., Yue, X., Yu, L., Sergei, M., & Wang, Z. (2020). Automatic image captioning based on ResNet50 and LSTM with soft attention. Wireless Communications and Mobile Computing, 2020.
23. NOTE: All the pictures used in this study(Fig [1-13]) are taken from the Flickr8k dataset.

AUTHORS PROFILE

Kota Akshith Reddy, is a fourth year Computer Science major at Vellore Institute of Technology, Vellore, Tamil Nadu, India. His research interests lie primarily in the field of Computer Vision and Natural Language Processing. He is currently working on improving the performance of Image Captioning models through data augmentation. He has one publication in the field of Natural Language Processing.

Satish C J, received his Ph.D. degree from VIT. He received his Master of Engineering degree from Anna University and Bachelor of Engineering degree from Madras University Tamilnadu, India. He was with Tata Consultancy Services for five years developing and maintaining software systems. His research interests include Recommendation Systems, Software Engineering, Software Visualization and Software Documentation Management

Jahnavi Polsani, is adept at learning new things and enthusiastic to solve real-life problems using the latest technologies. She is currently pursuing B.Tech 4th year in CSE from Vellore Institute of Technology, Vellore. Her areas of interest are Problem Solving, DBMS, Data Analytics,

Machine Learning, Deep Learning and Cryptography. She has a significant ability to provide valuable insights with a data-driven approach.

Teja Naveen Chintapalli, is passionate about learning new things and keen to solve real-life challenges by adhering the present-day technologies. He is Currently Pursuing his 4th year B.Tech (CSE) in information security at Vellore institute of Technology, Vellore. His area of interests are Cyber security, Data Privacy, Network security, Image processing and Machine Learning. He has a special skillset As an Cyber security Analyst, while making a significant contribution to different projects related to security.

Gangapatnam Sai Ananya, is a curious Computer Science undergraduate pursuing her fourth-year BTech in Narayana Engineering College. She has been working on several projects based on the Internet of Things and has been a part of Paper presentations at various Tech symposiums. She is currently developing her skills in machine learning and solving few daily life issues through basic technologies.