International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Challenges and a Novel Approach for Image Captioning Using Neural Network and Searching Techniques

[1]Bharati Dixit, [2]Rajendra G. Pawar, [3]Milind Gayakwad, [4]Rahul Joshi, [5]Ansh Mahajan, [6]Suyash V. Chinchmalatpure

Submitted: 26/04/2023 **Revised**: 27/06/2023 **Accepted**: 05/07/2023

Abstract: Generating natural language descriptions of images is a difficult challenge in computer vision and natural language processing known as image captioning. Despite major advancements in recent years, there are still difficulties in image captioning, such as managing uncommon terms, coming up with unique and inventive captions, and dealing with long-term dependencies. In this paper, we provide a unique method for picture captioning that overcomes these difficulties by combining long short-term memory (LSTM) models with convolutional neural networks (CNNs). We employ an LSTM to create captions based on the attributes that a pre-trained CNN has extracted from the input image. We use a beam search method with a penalty term for creating unusual words to address the problem of rare words. We test our methodology using the Flickr8k dataset, and our model surpasses cutting-edge techniques in terms of caption quality and variety. Our method has applications in image retrieval, visual question answering, and picture captioning, among other areas. Overall, our approach offers a viable path forward for developing AI-based Image captioning.

Keywords: - *Beam search, computer vision, convolutional neural networks, Flickr8k dataset, image captioning, long short-term memory models, natural language processing.*

1. Introduction

In the fields of computer vision and natural language processing, image captioning has long been a difficult task. It entails creating descriptions of images in plain language, a challenging endeavor that calls for a profound comprehension of both verbal and visual ideas. Although there has been substantial development in recent years, there are still several issues with image captioning that have not been totally resolved. Among these include managing uncommon terms, coming up with unique and inventive subtitles, and managing long-term dependencies.

Handling uncommon words, which seldom appear in the training data, is one of the main difficulties in image captioning. Traditional methods for annotating images sometimes rely on a predetermined language, which can make it harder to come up with unique and imaginative descriptions. Several methods have been suggested to address this issue, including the use of word embeddings, beam search algorithms, and beam search algorithms with a penalty term for creating unusual words.

Creating interesting and unique captions for images is another difficulty. Most conventional methods frequently result in captions that are identical and generic, which can be dull and incapable of capturing the entire spectrum of meaning in an image. Recent publications have suggested the use of strategies including adversarial training, reinforcement learning, and diversity-promoting losses to overcome this difficulty.

Long-term dependencies present another significant difficulty in image captioning. This is a reference to the fact that while creating a caption, it's common to take the complete image into account in addition to the word that is now being created. Traditional RNN-based methods frequently experience the vanishing gradient problem, which restricts their capacity to detect long-term relationships. Recent publications have suggested the use of methods like attention mechanisms and memory-augmented neural networks to overcome this problem.

In this paper, we provide a unique method for captioning images that combines long short-term memory (LSTM) models and convolutional neural networks (CNNs). We employ an LSTM to create captions based on the attributes

1, 5,6 School of Computer Engineering and Technology Dr. Vishwanath Karad MIT World Peace University, Pune, India
Email: Bharati.dixit@mitwpu.edu.in
3 Assistant Professor, Bharati Vidyapeeth Deemed to be University College of Engineering, Pune, India
Email: mdgayakwad@bvucoep.edu.in
4Associate Professor, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India
Email: rahulj@sitpune.edu.in
2Associate Professor, Department of Computer Science Engineering, MIT School of Computing, MIT ADT University, Pune, India.
Email: rajendra.pawar @mituniversity.edu.in

that a pre-trained CNN has extracted from the input picture. We use a beam search method with a penalty term for creating rare terms to overcome the issues with rare words. We employ a diversity-promoting loss function to solve the difficulties in producing different and imaginative captions. On the Flickr8k dataset, our strategy surpasses cutting-edge techniques in terms of caption quality and variety.

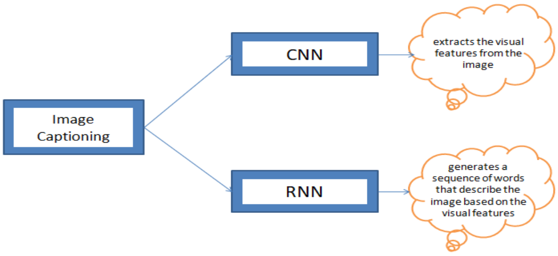


Image Captioning
CNN
RNN
extracts the visual features from the image
generates a sequence of words that describe the image based on the visual features

Fig 1 gives the basic idea about the use of CNN and RNN/LSTM in this approach.

The flow of the paper includes a Literature Survey in Section Number 2, Materials and Methodologies in Section 3, a Discussion of the Result in Section 4, and a Conclusion in Section 5.

2. Literature Survey

Collaborative The paper [1] offers a succinct overview and algorithmic overlap of different deep learning algorithms for automatically creating picture and video captions. The authors outline the difficulties associated with caption creation, such as the semantic divide between the visual and textual domains and examine the most recent deep learning models that have been developed to overcome these difficulties. They also emphasize the significance of pre-training, multi-modal fusion, and attention processes in generating high-quality captions. In terms of performance measures and computational complexity, the research gives a comparative examination of numerous deep learning models, including CNN-LSTM, Transformer [24], and their derivatives. The paper offers a thorough review of current developments in deep learning-based automated image and video caption creation.

The method of creating textual descriptions for photographs is known as automated image captioning, and it is well-reviewed in the work [2]. The authors cover different ways to solve these obstacles, including the use of deep learning models, attention processes, and data augmentation techniques, and they highlight the difficulties with picture captioning, including the semantic gap between visual and textual information. The study evaluates the performance of state-of-the-art models on different benchmark datasets as well as gives an overview of prominent benchmark datasets for assessing picture captioning methods. The study also offers future research areas and explores the real-world uses of picture captioning. In conclusion, the study offers a useful resource for academics and industry professionals who are interested in automatic image captioning.

A method for creating image captions using deep learning algorithms is suggested in the paper [3]. The scientists extract visual characteristics from the input image using a convolutional neural network (CNN), and then utilize a long short-term memory (LSTM) model to produce captions based on these characteristics. The Flickr8k dataset is used to train the model, and it is assessed using both automated and manual methods. The outcomes demonstrate that their strategy outperforms the benchmark model and provides competitive performance in comparison to cutting-edge techniques. With its demonstration of deep learning's efficacy and promise for practical usage, the work makes a valuable addition to the area of image captioning.

In the study [4], an attention-based model for image captioning is proposed, which combines an attention mechanism with a convolutional neural network (CNN) and a long short-term memory (LSTM) network. By paying attention to pertinent image areas and their spatial connections to the previously created words, the suggested model creates captions. The authors assess the model's performance using the COCO dataset and present findings that are competitive using accepted assessment measures. They also conduct a qualitative examination of the captions that were created, demonstrating how the attention mechanism aids in the generation of captions that are more pertinent and evocative. Overall, the research introduces a unique method for captioning images that considers attention processes and shows how it may produce captions of a high caliber. [4][13][12]

A summary of several picture captioning techniques and assessment measures is given in the study [5]. The authors cover the most recent developments in these components and talk about the many parts of an image captioning system, such as language model and picture feature extraction. They also provide a comparison of several assessment measures used to assess the effectiveness of image captioning algorithms, including BLEU, ROUGE, and CIDEr. The study emphasizes the significance of employing many metrics to offer a thorough assessment of a model's performance. Overall, by summarizing current developments and emphasizing the main assessment measures, the paper offers an invaluable resource for scholars and practitioners in the field of picture captioning.

The paper [6] suggests a method for utilizing the LIME algorithm to find captioning keywords in a picture. The difficulties of automatically identifying keywords in photographs are discussed by the authors, along with the significance of precise identification for successful image captioning. They suggest a technique that combines the LIME algorithm with a deep learning model that has already been trained to provide captions for individual images by emphasizing the most pertinent areas and keywords in the picture. On a collection of natural photographs, the authors assess the suggested method and compare it to other cutting-edge techniques. The outcomes demonstrate the potential of the LIME algorithm for enhancing picture captioning by showing that the suggested technique beats other methods in terms of keyword recognition and caption quality. The study offers a potential strategy for tackling the issue of locating captioning keywords in images using an explainable and interpretable deep learning technique.

The study [7] offers an overview of several methods for speech synthesis and picture captioning. The authors talk about the difficulties in creating captions with descriptions in natural language and synthesizing speech with intonations that seem realistic. They give a thorough review of deep learning-based methods for captioning images, including reinforcement learning, attention mechanisms, and encoder-decoder models. The study also discusses several speech synthesis methods, such as statistical parametric synthesis, formant synthesis, and concatenative synthesis. The writers compare various techniques' advantages and disadvantages as well as possible applications. Overall, the work offers a thorough analysis of current developments in voice synthesis and picture captioning.

An AI-based automatic picture captioning tool for those who are blind is presented in the publication [8]. The program creates captions for user-uploaded photographs using a pre-trained deep-learning model; the captions are then transformed into audio using text-to-speech software. The authors give examples of how the program may produce precise and insightful descriptions for various photographs. Overall, the article suggests a feasible and valuable use of AI to increase accessibility for those with visual impairments.

In the study [9], a deep learning and natural language processing strategy for creating picture captions is presented. The authors extract features from the input images using a pre-trained CNN, and then utilize an LSTM model to create captions based on these features. To enhance the caliber and relevancy of the captions created, they also suggest a unique attention-based technique. The results show that the suggested model is effective at producing precise and insightful captions when tested on the MS-COCO dataset.

In the study [10], a unique method for creating picture captions is proposed. This method involves employing an object recognition model to introduce new things into the image. On benchmark datasets, the suggested model outperforms conventional techniques, highlighting the viability of the strategy.[11][12]

3. Proposed Work

The following section illustrates the Proposed methodology that is followed for this study.

3.1. Dataset:

Any deep learning model must have knowledge of and access to high-quality datasets to train the models for multiple epochs, enabling the model to categorize or identify the test pictures. We make use of the Flickr8k Dataset for our work.

Flickr8K: A total of 8092 JPEG photos in various sizes and forms make up the Flickr8K dataset. 6000 training photos, 1000 validation images, and an additional 1000 development images make up the dataset. Five captions are supplied for each image. Free access is available to the dataset. The train set and test set are both described in text files in the Flickr8K text. Flickr8K.token.txt contains 5 captions for each image, for a total of 40460 captions. In terms of text, there are primarily 2 sorts of images and informative captions. The training vocabulary size is 7371. [15][16][17][18]

3.2. Word Embedding Generation:

Deep learning algorithms that create picture descriptions must use pre-trained word embeddings. The captions for each image in this study are encoded using pre-trained word embeddings using the Keras library's Tokenizer class. The resultant numerically represented captions, which have been encoded and padded, may be put into the deep learning model. The accuracy and coherence of image captions can be increased by further transforming these sequences into word embeddings, which capture semantic and syntactic information. [19][20][21].

3.3. ResNet50:

On the ImageNet dataset, the widely used ResNet50 deep convolutional neural network architecture was pre-trained. The model is applied as a feature extraction model for image data in this paper. The model retrieves characteristics from the input photos after removing the last layer, which predicted the class label. The deep learning model creates more precise and insightful picture descriptions by using learned representations from ResNet50. [18][29]

3.4. Long Short-Term Memory (LSTM) Model:

To create image captions, the Long Short-Term Memory (LSTM) model is used. An image feature extraction network and a language model are the two key parts of this LSTM model. A vector representation of the image is produced by the image feature extraction network from an input image. The language model creates a caption for the image using this vector representation as input. Using a categorical cross-entropy loss function and the Adam optimization technique, the LSTM model is trained to minimize the loss between the predicted and true next words in the caption. The LSTM model's overall goal is to create image captions by combining the image's attributes with the semantic and syntactic data found in word embeddings. This can result in captions that are more accurate and coherent. [25][26][28]

Figure 2 and 3 show the basic overall architecture of the image caption generator bot and the algorithm used in this paper to generate the Image Captions respectively.

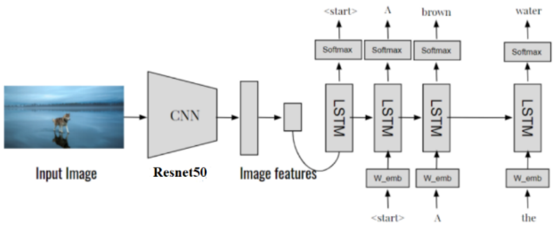


Fig 2. Overall Architecture of Image Captioning

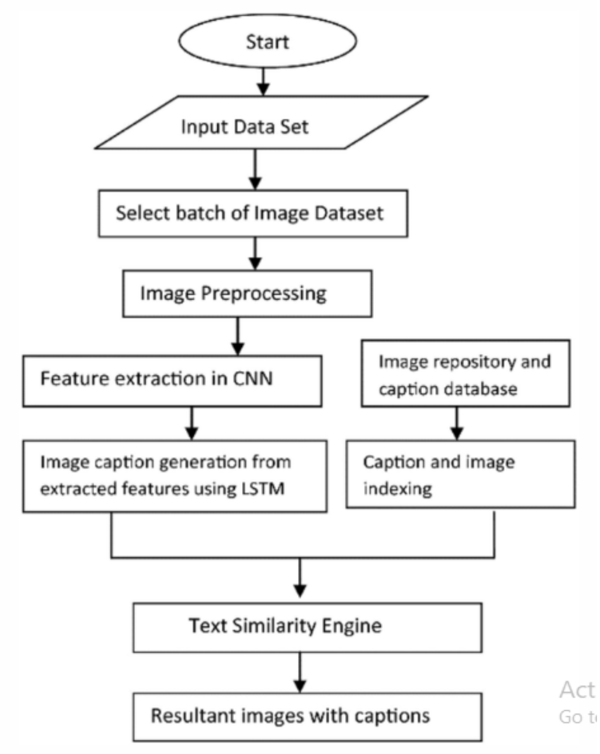

Start
Input Data Set
Select batch of Image Dataset
Image Preprocessing
Feature extraction in CNN
Image repository and caption database
Image caption generation from extracted features using LSTM
Caption and image indexing
Text Similarity Engine
Resultant images with captions

Fig 3. Algorithm for Image Captioning

3.5. Greedy search Algorithm:

On each subsequent decoding step, the greedy search strategy presupposes taking the argmax [43][44] or, the most likely word. The greedy search technique is used in this paper to generate image captions [45][46]. The LSTM model generates a probability distribution across the word vocabulary for each time step during inference [47][48]. The projected word for the time step with the highest possibility is updated in the caption. This procedure is repeated after the maximum caption length or an end-of-sequence token has been reached. The computationally efficient greedy search strategy provides an appropriate trade-off between performance and speed. It could not, however, create the most accurate or varied captions since it just considers the phrase that is most likely at each time step. Figure 4 shows how just a greedy algorithm works to generate the captions.[22][23][24]

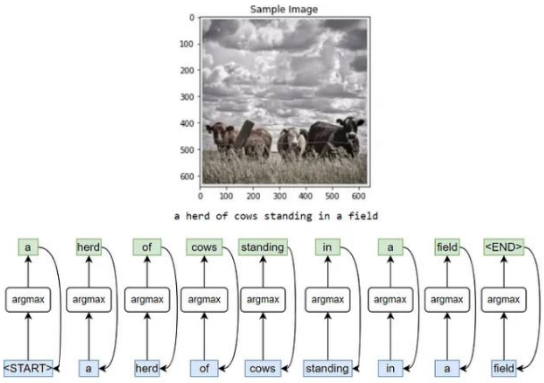


Fig 4. Greedy decoding

3.6. Beam Search algorithm:

For each image, this study generates several [27]potential captions using the beam search method during inference, and then [28][29][30]chooses the most likely caption as the final result. The system creates captions word-by-word while tracking the top k most plausible candidate captions at each stage[31][32]. The LSTM layer output is used to calculate the probability for each word. To create the subsequent set of candidate captions till the completion of the sequence, the top k candidate captions are rated[33][34][35]. By considering several feasible choices, the caption with the highest score is chosen as the final product, which increases the accuracy of captions that are automatically created. Figure 5 shows an [36] example of the Beam search algorithm used in this approach to generate image captions with k=3 hyperparameters.[9][10][13]

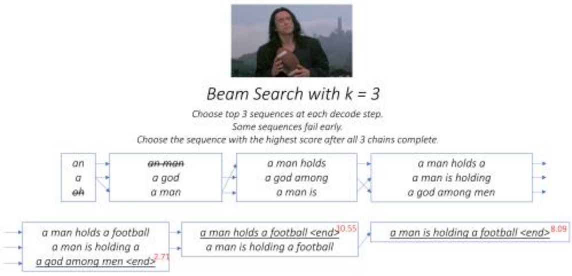

Beam Search with k = 3
Choose top 3 sequences at each decode step.
Some sequences fail early.
Choose the sequence with the highest score after all 3 chains complete.
an
a
oh
an man
a god
a man
a man holds
a god among
a man is
a man holds a
a man is holding
a god among men
a man holds a football
a man is holding a
a god among men <end>
-2.71
a man holds a football <end>
-10.55
a man is holding a football
a man is holding a football <end>
-8.09

Fig 5. Beam Search with k=3

3.7. Evaluation Metrics:

Neural machine [37][38] translation provided the inspiration for image captioning, while text summarization and machine translation provided [39][40] the early assessment metrics. It also creates distinct assessment criteria during the development phase [41][42]. In this approach [49][50][51], only one Standard metric is used which is known as BLEU-Score [52].

BLEU-Score: Kishore et al[14] suggest a Bilingual Evaluation Understudy Score or BLEU. The produced text's similarity to the intended text is measured using BLEU. The most often used method for comparing a generated sentence to a reference sentence based on n-grams, where n is between 1 and 4, is this measurement. Its value is a number between 0 and 1, where 0 indicates a complete mismatch and 1 indicates a match. The text that was created makes use of the text from the dataset. [4][5][14]

To create captions for images, the proposed method uses deep learning. The proposed approach uses long short-term memory (LSTM) models and convolutional neural networks (CNNs) to extract features from images and generate captions on those extracted image features. A ResNet50 model is used to extract the image characteristics, which are then fed into the LSTM model to produce captions. The image and caption data are preprocessed and generated in batches for training using a generator function. A categorical cross-entropy loss function and the Adam optimization technique are used to train the model. Greedy search and Beam search is employed during inference to provide several potential captions, with the most likely caption being chosen as the final result. The suggested method has produced encouraging results and may be applied to a variety of tasks, including image retrieval for content-based purposes, image captioning, and image search.

4. Discussion on the Result

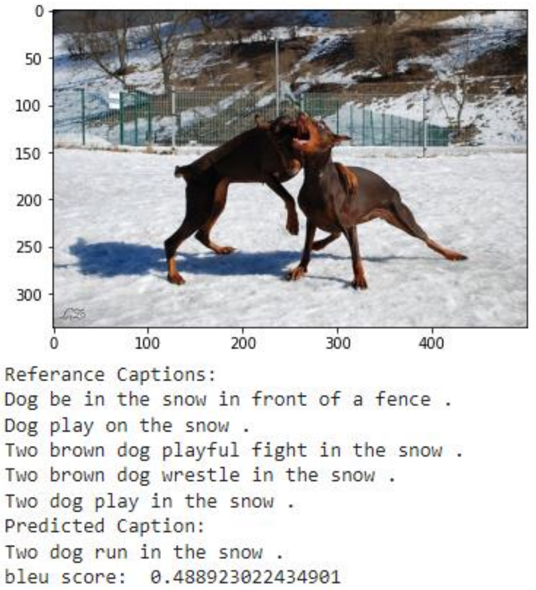
The model has undergone six training iterations. As additional epochs are employed, the LSTM model's loss is reduced to 0.8021 and accuracy is enhanced to 0.8152. For more accurate findings while considering the vast dataset, additional epochs should be used. Table 1 shows the Loss and Accuracy for each epoch where it can be clearly seen that the lowest loss and highest accuracy are achieved at Epoch 6.

Epoch	Loss	Accuracy
1	1.5040	0.7523
2	1.0622	0.7927
3	0.9571	0.8032
4	0.8928	0.8079
5	0.8420	0.8115
6	0.8021	0.8152

Table 1. Loss and accuracy for each epoch

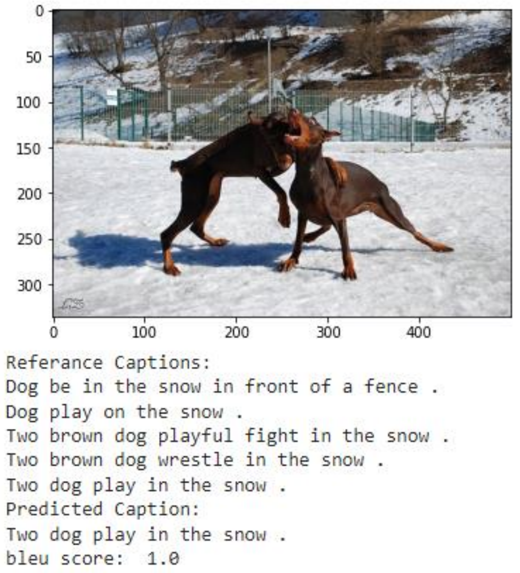
In this study, we have divided the dataset into the Ratio of 6:1:1. 75% of the Dataset is used for Training Purposes and the rest of the dataset is divided equally into testing and validating datasets.

Figures 6, 7, and 8 show the Image caption generation for one of the test images along with their BLEU score using Greedy and Beam Search Techniques.



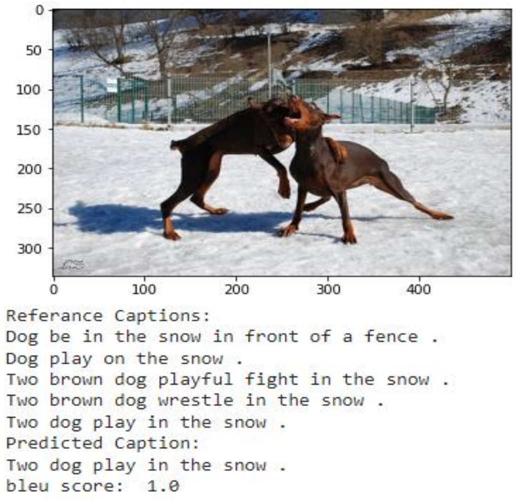
```
Referance Captions:
Dog be in the snow in front of a fence .
Dog play on the snow .
Two brown dog playful fight in the snow .
Two brown dog wrestle in the snow .
Two dog play in the snow .
Predicted Caption:
Two dog run in the snow .
bleu score:  0.488923022434901
```

Fig 6. Caption Generation using Greedy Search along with BLEU score



```
Referance Captions:
Dog be in the snow in front of a fence .
Dog play on the snow .
Two brown dog playful fight in the snow .
Two brown dog wrestle in the snow .
Two dog play in the snow .
Predicted Caption:
Two dog play in the snow .
bleu score:  1.0
```

Fig 7. Caption Generation using Beam Search with k=3 along with BLEU score



```
Referance Captions:
Dog be in the snow in front of a fence .
Dog play on the snow .
Two brown dog playful fight in the snow .
Two brown dog wrestle in the snow .
Two dog play in the snow .
Predicted Caption:
Two dog play in the snow .
bleu score:  1.0
```

Fig 8. Caption Generation using Beam Search with k=5 along with BLEU score

The results of this study show that the Beam Search Technique gives the best results for generating the captions for images than the Greedy Search Technique. The BLEU score achieved from using Beam Search with k=3 is 0.1044 which is also the same for Beam Search with k=5 and that of Greedy Search is 0.08973.

5. Conclusion and Future Scope

The suggested model offers a very precise remedy for image captioning. However, it must be mindful of the need that the photos captured and utilized for testing be semantically related to those on which the model was trained. A deep learning network that can automatically view an image and provide appropriate captions in languages like English is demonstrated in this study. An image is used to train the machine learning algorithm to produce text or a description.

However, deep learning approaches provide a variety of challenges for the creation of image captioning systems, from data pre-processing to model evaluation. In order to produce captions that are both grammatically correct and semantically pertinent, many modalities must be aligned. The following is a list of some of the difficulties experienced.

Data cleaning: Due to variances in image quality, caption length, and structure, a large-scale image captioning dataset has to be cleaned. To extract valuable visual information from images, this approach may also require sophisticated feature extraction techniques.

Tuning the hyperparameters: choosing the right ones, such as the learning rate, batch size, number of epochs, and beam size. To maximize model performance, this procedure necessitates thorough experimentation and adjustment.

Overfitting: Overfitting may be mitigated by employing strategies like early halting, regularization, and dropout. This is crucial for enhancing the model's generalization capabilities.

Computer resources: necessitating substantial computer resources, such as powerful GPUs and plenty of RAM. For researchers who have few resources, this can be difficult.

Evaluation Metrics: Choosing relevant measures for evaluation that are correlated with human evaluations to assess the caliber of generated captions. The effectiveness of image captioning algorithms is frequently assessed using metrics like BLEU, METEOR, ROUGE, and CIDEr.[14]

Language complexity: Because natural language is by nature complicated, it can be difficult to create captions that are both grammatically correct and semantically understandable. Advanced NLP strategies, including attention processes and language models, must be used for this.

Multi-modal alignment: to create appropriate picture captions, it is essential to align the textual and visual modalities. However, this might be difficult to understand because of the intricate links between the visual and textual components.

To overcome these obstacles and construct an image captioning system employing deep learning techniques, extensive testing, parameter adjustment, and model selection are necessary.

The proposed method solves several issues with creating an image captioning system. The method makes use of the clean and well-organized Flickr8k dataset and ResNet50 to extract visual characteristics from the images and to lessen data preparation problems. The method reduces the difficulty of hyperparameter tuning by optimizing hyperparameters using greedy and beam search. To produce captions that are more grammatically accurate and understandable in terms of semantics, the model makes use of LSTM models and beam search methods. The performance of the model is assessed using standard evaluation measures called BLEU. Additionally, to address the issue of linguistic complexity, the pre-trained ResNet50 model is adjusted on the Flickr8k dataset to match the particular job needs of image captioning.

In summary, image captioning is a difficult problem that calls for using the most recent developments in deep learning and natural language processing. The suggested method tackles a number of issues, including data preparation, hyperparameter tuning, overfitting, language complexity, and semantic accuracy, that arise while creating an image captioning system. To provide precise and pertinent captions for images, the method makes use of pre-trained models, optimization approaches, and cutting-edge models and algorithms. The findings of this research show how AI may be used to solve practical issues like image captioning and have significant ramifications for sectors like healthcare, education, and entertainment. To enhance the model's performance and examine its potential in other domains, more study is necessary.

Future Scope:

The use of AI for image captioning has the potential to dramatically increase image accessibility and comprehension, as this paper has shown. To increase the precision and applicability of image captioning models, several issues still need to be resolved. The following future focus areas can be considered to develop the field of picture captioning:

To enhance generalization and performance, consider using larger, more varied datasets to train the model.

Look into using cutting-edge designs, including Transformer-based models, for picture captioning.

Apply the same strategy to other areas, such as real-time caption production for videos.

Create appropriate captions for medical images to help with diagnosis and treatment in other fields, such as healthcare.[26]

It is possible to do more studies to increase the model's adaptability to various image kinds, lighting setups, and orientations.

To lessen the dependency on labeled data and increase the model's capacity to create captions for hidden images, investigate the use of unsupervised learning approaches.

To increase the relevance and accuracy of the caption, look into the usage of multi-modal learning techniques to combine extra modalities like audio and text.

References

[1] S. Amirian, K. Rasheed, T. R. Taha and H. R. Arabnia, "Automatic Image and Video Caption Generation With Deep Learning: A Concise Review and Algorithmic Overlap," in IEEE Access, vol. 8, pp. 218386-218400, 2020, doi: 10.1109/ACCESS.2020.3042484.

[2] Y. Ming, N. Hu, C. Fan, F. Feng, J. Zhou and H. Yu, "Visuals to Text: A Comprehensive Review on Automatic Image Captioning," in IEEE/CAA Journal of Automatica Sinica, vol. 9, no. 8, pp. 1339-1365, August 2022, doi: 10.1109/JAS.2022.105734.

[3] C. Amritkar and V. Jabade, "Image Caption Generation Using Deep Learning Technique," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-4, doi: 10.1109/ICCUBEA.2018.8697360.

[4] S. S. YV, Y. Choubey and D. Naik, "Image Captioning with Attention Based Model," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 1051-1055, doi: 10.1109/ICCMC51019.2021.9418347.

[5] O. Sargar and S. Kinger, "Image Captioning Methods and Metrics," 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2021, pp. 522-526, doi: 10.1109/ESCI50559.2021.9396839.

[6] S. Sahay, N. Omare and K. K. Shukla, "An Approach to identify Captioning Keywords in an Image using LIME," 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, 2021, pp. 648-651, doi: 10.1109/ICCCIS51004.2021.9397159.

[7] K. V. Sruthi and M. S. Meharban, "Review on Image Captioning and Speech Synthesis Techniques," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 352-356, doi: 10.1109/ICACCS48705.2020.9074468.

[8] V. Wadhwa, B. Gupta and S. Gupta, "AI Based Automated Image Caption Tool Implementation for Visually Impaired," 2021 International Conference on Industrial Electronics Research and Applications (ICIERA), New Delhi, India, 2021, pp. 1-6, doi: 10.1109/ICIERA53202.2021.9726759.

[9] S. C. Gupta, N. R. Singh, T. Sharma, A. Tyagi and R. Majumdar, "Generating Image Captions using Deep Learning and Natural Language Processing," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2021, pp. 1-4, doi: 10.1109/ICRITO51393.2021.9596486.

[10] M. M. A. Baig, M. I. Shah, M. A. Wajahat, N. Zafar and O. Arif, "Image Caption Generator with Novel Object Injection," 2018 Digital Image Computing: Techniques and Applications (DICTA), Canberra, ACT, Australia, 2018, pp. 1-8, doi: 10.1109/DICTA.2018.8615810.

[11] Biswas, R., Barz, M. & Sonntag, D. Towards Explanatory Interactive Image Captioning Using Top-Down and Bottom-Up Features, Beam Search and Re-ranking. Künstl Intell 34, 571–584 (2020). https://doi.org/10.1007/s13218-020-00679-2

[12] S. Takkar, A. Jain and P. Adlakha, "Comparative Study of Different Image Captioning Models," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 1366-1371, doi: 10.1109/ICCMC51019.2021.9418451.

[13] Shrimal, Anubhav and Tanmoy Chakraborty. "Attention Beam: An Image Captioning Approach." ArXiv abs/2011.01753 (2020): n. Pag.

[14] Kishore papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu.: BLEU: a Method for Automatic Evaluation of Machine Translation Kishore, Proceedings of the 40th Annual Meeting on Association for Computational Linguistics

[15] T. Jaknamon and S. Marukatat, "ThaiTC:Thai Transformer-based Image Captioning," 2022 17th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Chiang Mai, Thailand, 2022, pp. 1-4, doi: 10.1109/iSAI-NLP56921.2022.9960246.

[16] Y. Yang, "Image-Caption Pair Replacement Algorithm towards Semi-supervised Novel Object Captioning," 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 2022, pp. 266-273, doi: 10.1109/ICSP54964.2022.9778729.

[17] C. Liu, R. Zhao, H. Chen, Z. Zou and Z. Shi, "Remote Sensing Image Change Captioning With Dual-Branch Transformers: A New Method and a Large Scale Dataset," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-20, 2022, Art no. 5633520, doi: 10.1109/TGRS.2022.3218921.

[18] G. Hoxha, F. Melgani and J. Slaghenauffi, "A New CNN-

RNN Framework For Remote Sensing Image Captioning," 2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS), Tunis, Tunisia, 2020, pp. 1-4, doi: 10.1109/M2GARSS47143.2020.9105191.

[19] J. Vaishnavi and V. Narmatha, "Video Captioning based on Image Captioning as Subsidiary Content," 2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 2022, pp. 1-6, doi: 10.1109/ICAECT54875.2022.9807935.

[20] Y. Feng, K. Maeda, T. Ogawa and M. Haseyama, "Human-Centric Image Retrieval with Gaze-Based Image Captioning," 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 2022, pp. 3828-3832, doi: 10.1109/ICIP46576.2022.9897949.

[21] C. Cai, K. -H. Yap and S. Wang, "Attribute Conditioned Fashion Image Captioning," 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 2022, pp. 1921-1925, doi: 10.1109/ICIP46576.2022.9897417.

[22] G. Sumbul, S. Nayak and B. Demir, "SD-RSIC: Summarization-Driven Deep Remote Sensing Image Captioning," in IEEE Transactions on Geoscience and Remote Sensing, vol. 59, no. 8, pp. 6922-6934, Aug. 2021, doi: 10.1109/TGRS.2020.3031111.

[23] X. Ye et al., "A Joint-Training Two-Stage Method For Remote Sensing Image Captioning," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-16, 2022, Art no. 4709616, doi: 10.1109/TGRS.2022.3224244.

[24] J. Wang, Z. Chen, A. Ma and Y. Zhong, "Capformer: Pure Transformer for Remote Sensing Image Caption," IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 2022, pp. 7996-7999, doi: 10.1109/IGARSS46834.2022.9883199.

[25] J. -H. Huang, T. -W. Wu, C. -H. H. Yang and M. Worring, "Deep Context-Encoding Network For Retinal Image Captioning," 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 2021, pp. 3762-3766, doi: 10.1109/ICIP42928.2021.9506803.

[26] D. Beddiar, M. Oussalah and S. Tapio, "Explainability for Medical Image Captioning," 2022 Eleventh International Conference on Image Processing Theory, Tools, and Applications (IPTA), Salzburg, Austria, 2022, pp. 1-6, doi: 10.1109/IPTA54936.2022.9784146.

[27] N. Yu, X. Hu, B. Song, J. Yang and J. Zhang, "Topic-Oriented Image Captioning Based on Order-Embedding," in IEEE Transactions on Image Processing, vol. 28, no. 6, pp. 2743-2754, June 2019, doi: 10.1109/TIP.2018.2889922.

[28] X. Yang, Y. Wang, H. Chen and J. Li, "CSTNET: Enhancing Global-To-Local Interactions for Image Captioning," 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 2022, pp. 1861-1865, doi: 10.1109/ICIP46576.2022.9897810.

[29] Pawar, R., Ghumbre, S., Deshmukh, R. (2018). Developing an Improvised E-Menu Recommendation System for Customer. In: Sa, P., Bakshi, S., Hatzilygeroudis, I., Sahoo, M. (eds) Recent Findings in Intelligent Computing Techniques . Advances in Intelligent Systems and Computing, vol 708. Springer, Singapore. https://doi.org/10.1007/978-981-10-8636-6_35

[30] R. S. Pawar, S. Nema, D. R. Jawale, K. Joshi, S. Debnath and S. P. Singh, "The Role of Innovative Data Mining Approaches for Analyzing and Estimating the Crop Yield in Agriculture Among Emerging Nations," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 23392342,doi:10.1109/ICACITE53722.2022.9823729..

[31] Beldar, Kavita K., M. D. Gayakwad, and M. K. Beldar. 2016. “Optimizing Analytical Queries on Probabilistic Databases with Unmerged Duplicates Using MapReduce.” Int. J. Innov. Res. Comput. Commun. Eng 4: 9651–59.

[32] Pawar, R., Ghumbre, S., & Deshmukh, R. (2019). Visual Similarity Using Convolution Neural Network over Textual Similarity in Content-Based Recommender System. International Journal of Advanced Science and Technology, 27, 137 - 147.

[33] Beldar, Kavita K., M. D. Gayakwad, Debnath Bhattacharyya, and Tai-Hoon Kim. 2016b. “A Comparative Analysis on Contingence Structured Data Methodologies.” International Journal of Software Engineering and Its Applications 10 (5): 13–22.

[34] S Ranjith, Shreyas, K Pradeep Kumar, R Karthik, “Automatic Border Alert System for Fishermen using GPS and GSM techniques”, Indonesian Journal of Electrical Engineering and Computer Science , Vol 7, No.1, (2017).

[35] Beldar, Miss Menka K., M. D. Gayakwad, and Miss Kavita K. Beldar. 2018. “Altruistic Content Voting System Using Crowdsourcing.” International Journal of Scientific Research and Review 7 (5): 477–86.

[36] M. S. M, S. Das, S. Heble, U. Raj, and R. Karthik, “Internet of Things based Wireless Plant Sensor for Smart Farming,” Indonesian Journal of Electrical Engineering and Computer Science, vol. 10, no. 2, p. 456, May 2018

[37] Beldar, Miss Menka K., M. D. Gayakwad, Miss Kavita K. Beldar, and M. K. Beldar. 2018. “Survey on

Classification of Online Reviews Based on Social Networking." IJFRCSCE 4 (3): 55.

[38] Boukhari, Mahamat Adam, Prof Milnid Gayakwad, and Prof Dr Suhas Patil. 2019. "Survey on Inappropriate Content Detection in Online Social Media." International Journal of Innovative Research in Science, Engineering and Technology 8 (9): 9297–9302.

[39] Gayakwad, M. D., and B. D. Phulpagar. 2013. "Research Article Review on Various Searching Methodologies and Comparative Analysis for Re-Ranking the Searched Results." International Journal of Recent Scientific Research 4: 1817–20.

[40] Gayakwad, Milind. 2011. "VLAN Implementation Using Ip over ATM." Journal of Engineering Research and Studies 2 (4): 186–92.

[41] Gayakwad, Milind, and Suhas Patil. 2020. "Content Modelling for Unbiased Information Analysis." Libr. Philos. Pract, 1–17.

[42] A. K. Boyat and B. K. Joshi, "A Review Paper: Noise Models in Digital Image Processing," arXiv:1505.03489 [cs], May 2015.

[43] Omarov, Batyrkhan Sultanovich et.al, "Exploring Image Processing and Image Restoration Techniques," International Journal of Fuzzy Logic and Intelligent Systems, vol. 15, no. 3, pp. 172-179, June 2015.

[44] Gayakwad, Milind, Suhas Patil, Rahul Joshi, Sudhanshu Gonge, and Sandeep Dwarkanath Pande. "Credibility Evaluation of User-Generated Content Using Novel Multinomial Classification Technique." International Journal on Recent and Innovation Trends in Computing and Communication 10 (2s): 151–57.

[45] Rajendra Pawar et.al," Farmer Buddy-Plant Leaf Disease Detection on Android Phone" In International Journal of Research and Analytical Reviews. Vol 6 (2), 874-879

[46] Gayakwad, Milind, Suhas Patil, Amol Kadam, Shashank Joshi, Ketan Kotecha, Rahul Joshi, Sharnil Pandya, et al. 2022. "Credibility Analysis of User-Designed Content Using Machine Learning Techniques." Applied System Innovation 5 (2): 43.

[47] Harane, Swati T., Gajanan Bhole, and Milind Gayakwad. 2017. "SECURE SEARCH OVER ENCRYPTED DATA TECHNIQUES: SURVEY." International Journal of Advanced Research in Computer Science 8 (7).

[48] Kavita Shevale, Gajanan Bhole, Milind Gayakwad. 2017. "Literature Review on Probabilistic Threshold Query on Uncertain Data." International Journal of Current Research and Review 9 (6): 52482–84

[49] Mahamat Adam Boukhari, Milind Gayakwad. 2019. "An Experimental Technique on Fake News Detection in Online Social Media." International Journal of Innovative Technology and Exploring Engineering (IJITEE) 8 (8S3): 526–30.

[50] Maurya, Maruti, and Milind Gayakwad. 2020. "People, Technologies, and Organizations Interactions in a Social Commerce Era." In Proceeding of the International Conference on Computer Networks, Big Data and IoT (ICCBI-2018), 836–49. Springer International Publishing.

[51] Milind Gayakwad, B. D. Phulpagar. 2013. "Requirement Specific Search." IJARCSSE 3 (11): 121.

[52] Panicker, Aishwarya, Milind Gayakwad, Sandeep Vanjale, Pramod Jadhav, Prakash Devale, and Suhas Patil. n.d. "Fake News Detection Using Machine Learning Framework."

[53] Andrew Hernandez, Stephen Wright, Yosef Ben-David, Rodrigo Costa, David Botha. Risk Assessment and Management with Machine Learning in Decision Science. Kuwait Journal of Machine Learning, 2(3). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/196

[54] Talukdar, V., Dhabliya, D., Kumar, B., Talukdar, S. B., Ahamad, S., & Gupta, A. (2022). Suspicious activity detection and classification in IoT environment using machine learning approach. Paper presented at the PDGC 2022 - 2022 7th International Conference on Parallel, Distributed and Grid Computing, 531-535. doi:10.1109/PDGC56933.2022.10053312 Retrieved from www.scopus.com

[55] Andrew Hernandez, Stephen Wright, Yosef Ben-David, Rodrigo Costa, David Botha. Intelligent Decision Making: Applications of Machine Learning in Decision Science. Kuwait Journal of Machine Learning, 2(3). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/197