Bulletin of Electrical Engineering and Informatics
Vol. 14, No. 4, August 2025, pp. 2861∼2870
ISSN: 2302-9285, DOI: 10.11591/eei.v14i4.8885

Exploring deep learning approaches for image captioning to mimic human understanding

Maheen Islam[1], Mahedi Hassan Ratul[1], Rezaul Haque[1], Sazzad Hossain Rony[1], Azharul Huq Asif[1], Tanni Mittra[1], Md Miskat Hossain[1], Mahamudul Hasan[2]
[1]Department of Computer Science and Engineering, Faculty of Science and Engineering, East West University, Dhaka, Bangladesh
[2]Department of Computer Science and Engineering, University of Minnesota Twin Cities, Minneapolis, United States

Article Info

Article history:

Received Jun 21, 2024
Revised Feb 9, 2025
Accepted Mar 9, 2025

Keywords:

Caption generation
Context dataset
Deep learning
Image captioning
Image encoding
Microsoft common objects in context

ABSTRACT

Image captioning has emerged as a vital research area in computer vision, aiming to enhance how humans interact with visual content. While progress has been made, challenges like improving caption diversity and accuracy remain. This study proposes transfer learning models and RNN algorithms trained on the microsoft common objects in context (MS COCO) dataset to improve image captioning quality. The models combine image and text features, utilizing ResNet50, VGG16, and InceptionV3 with LSTM, and BiLSTM. Performance is measured using metrics such as BLEU, ROUGE, and METEOR for greedy and beam search. The InceptionV3+BiLSTM model outperformed others, achieving a BLEU score of over 60%, a METEOR score of 28.6%, and a ROUGE score of 57.2%. This research contributes to building a simple yet effective image captioning model, providing accurate descriptions with human-like understanding. The error was analyzed to improve results while discussing ongoing research aimed at enhancing the diversity, fluency, and accuracy of generated captions, with significant implications for improving the accessibility and searchability of visual media and informing future research in this area.

Corresponding Author:

Mahamudul Hasan
Department of Computer Science and Engineering, University of Minnesota Twin Cities
200 Union St SE, Minneapolis, 55455, Minnesota, United States
Email: munna09bd@gmail.com

1. INTRODUCTION

Deep learning has revolutionized object detection with methods like convolutional neural networks (CNNs) and region-based CNNs, including Fast R-CNN, Faster R-CNN, and YOLO, becoming primary tools [1]. Transfer learning has further reduced the training data requirement by leveraging pre-trained models on large datasets [2]. Image captioning generates textual descriptions of images and complements object detection, offering a comprehensive understanding of image content. Applications include making visual content accessible to visually impaired people, improving image searchability, enhancing content understanding, and enabling automated image tagging [3], [4].

Recent studies in image captioning have focused on improving caption accuracy, fluency, and diversity by incorporating attention mechanisms and additional information. Attention-based models, such as those in [5], focus on critical image regions, resulting in improved performance on datasets like COCO. For example, [5] achieved a BLEU-4 score of 50.4 using a Hard-Attention model. Similarly, [6] introduced SCA-CNN with

spatial and channel-wise attention, achieving a BLEU-1 score of 71.9. Other approaches, such as bottom-up and top-down attention [7] and semantic attention [8], further enhanced performance by capturing object relationships and fine-grained attributes.

Modern image captioning models combine CNNs to extract image features with RNNs or Transformer models to generate captions, evaluated using metrics like BLEU, ROUGE, METEOR, and CIDEr [9]. Challenges remain in improving diversity, fluency, and efficiency while handling rare objects and complex relationships [10], [11]. Real-time performance and resource efficiency also need attention [12].

This research aims to overcome all the problems mentioned in literatures by improving the diversity, fluency, accuracy, and efficiency of image captioning models. Using transfer learning and RNN algorithms, we trained models on the microsoft common objects in context (MS COCO) dataset [13]. Our study introduces models that effectively describe image content, including objects, scenes, and relationships, evaluated with traditional metrics. The contributions of our research are as follows:

- Build a simple image captioning architecture based on lightweight transformer learning (ResNet50, VGG16, and InceptionV3) and RNN (LSTM and BiLSTM) models that can achieve accurately describe the content of an image with human-level understanding.
- Achieve higher performance gain on the basis of BLEU, recall-oriented understudy for gisting evaluation (ROUGE), and metric for evaluation of translation with explicit ordering (METEOR) score for greedy search and beam search.
- Provide a comparative analysis of the models' performance and adequate error analysis to improve the results.

2. METHOD

This study aims to improve the quality and efficiency of image captioning models by introducing several new models that incorporate transfer learning and RNN algorithms trained on the MS COCO dataset. These models integrate image and text features and utilize lightweight transfer learning methods with ResNet50, VGG16, and InceptionV3 architectures alongside RNN approaches including LSTM and BiLSTM. Performance assessment of these models is conducted using standard image captioning evaluation metrics such as BLEU, ROUGE, and METEOR, employing greedy search and beam search strategies. Figure 1 shows us the flowchart of the proposed model.

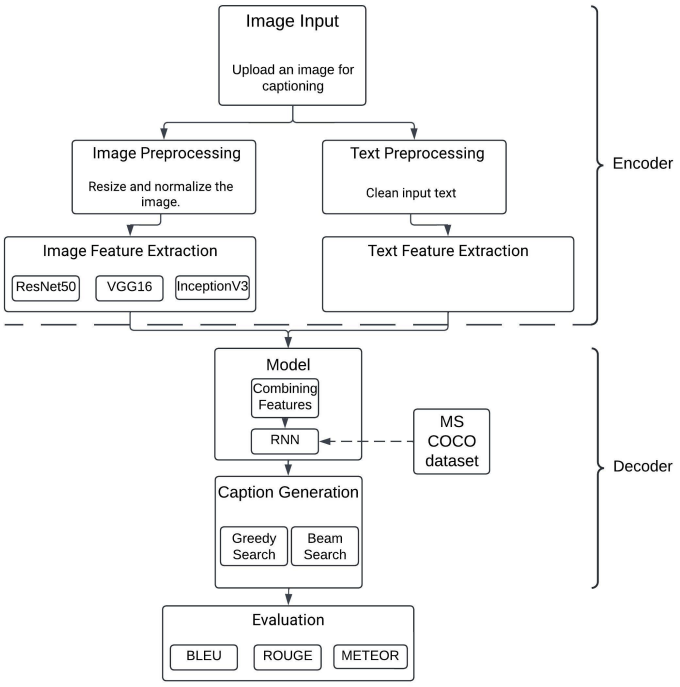

Figure 1. A flowchart of the proposed model

2.1. Data description and analysis

Image captioning datasets are collections of images paired with corresponding textual descriptions that accurately reflect the image content. Image captioning datasets contain images from various sources, such as photographs, illustrations, or animations. Captions in the dataset can vary in length and complexity, ranging from simple one-word labels to detailed, multi-sentence descriptions of the scene. In this study, we used the Microsoft COCO dataset, which contains 330,000 images paired with 5 human-generated captions each, to train and test our image captioning models. The dataset include annotations such as object and attribute labels, scene graphs, and semantic segmentations, which can improve model performance and interpretability. However, the dataset poses several challenges for image captioning models, such as handling the diversity of objects and scenes, generating accurate and semantically meaningful captions, incorporating additional information, such as scene graphs or semantic segmentation, and accurately describing multiple objects and their relationships in a single caption. Figure 2 shows an example of a COCO dataset image paired with its caption.

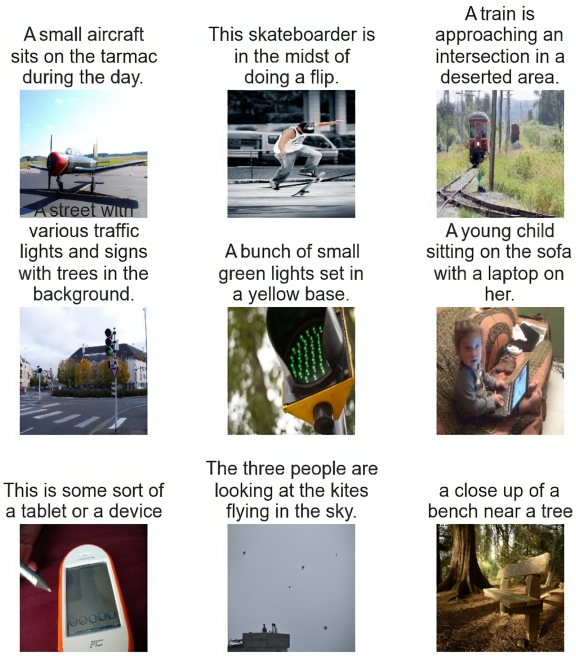

Figure 2. Number of reviews for each class

2.2. Image and caption pre-processing

In image captioning, image and text preprocessing are essential for effective processing by machine learning models. Proper preprocessing can have a significant impact on model performance [14]. The study uses OpenCV for image preprocessing and NLTK for text preprocessing. Common steps include resizing images, normalizing pixel values, tokenizing captions, converting words to numerical representations, and padding or truncating captions to a fixed length. Furthermore, we created a word cloud representation of the captions to identify which words have the most frequent occurrence, as shown in Figure 3. From the figure it can be seen mostly frequent words are stop words such as “in”, “on”, and “with”. In addition, there can be seen a frequent use of sports keywords such as “tennis”, “baseball”, and “skateboard”. To build an adequate image captioning model we need to reduce the caption dimensionality with the use of proper text preprocessing techniques. Removing punctuation and special characters can improve model efficiency [15]. Additionally, proper preprocessing helps to improve model performance and convergence, reduce overfitting, and increase the model’s ability to generalize.

2.3. Image encoding

In this research, the ResNet-50 [16], [17], VGG-16 [18], [19], and InceptionV3 models were used for image captioning by utilizing their pre-trained CNN layers to extract features from images. InceptionV3

outperforms all other models. The image encoding model acts as a feature extractor, learning to recognize and extract important features from the image, such as objects, scenes, and attributes.

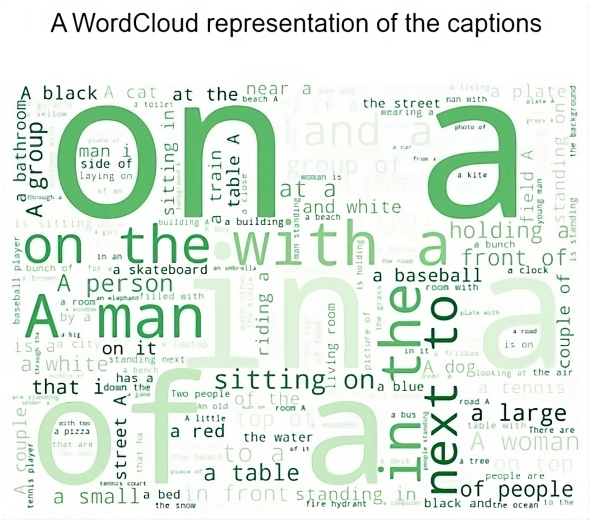

Figure 3. Frequency of words

2.3.1. InceptionV3

The model has a unique modular architecture composed of multiple Inception modules that can learn both local and global features from an image. These modules are built to learn different levels of abstraction, from low-level features such as edges and curves to high-level features that capture more abstract information about the image. In the image encoding process, InceptionV3 uses these modules to extract features from the image, which are then converted into a feature vector that can be used as input to the caption generation process [20]. Compared to ResNet and VGG16, InceptionV3 has a more flexible architecture [21], enabling it to learn a wider range of features from the image, resulting in better image representations and improved performance on image captioning tasks. Additionally, InceptionV3 is computationally efficient and has relatively few parameters, making it a suitable choice for real-time image captioning applications.

2.4. Language decoder

Image captioning involves extracting image features using methods like ResNet50, VGG16, or InceptionV3, followed by generating textual descriptions with recurrent neural networks (RNNs) [22]. These networks use image features as context to produce captions. Captions are preprocessed into numerical representations by assigning unique indices to words and padding sequences to a uniform length of 40. With a vocabulary size of 11,632 words, two RNN variants, LSTM and BiLSTM, are employed as language decoders.

LSTMs process sequences while maintaining past information, whereas BiLSTMs consider both past and future contexts, enabling more coherent captions. The decoder predicts each word by utilizing the previous hidden state and predicted word, outputting a probability distribution over the vocabulary. The captioning process halts upon predicting an end-of-sequence token or reaching the maximum length. The network is trained end-to-end by minimizing the difference between predicted and ground-truth captions, with post-processing applied to refine fluency and correct grammatical errors.

2.5. Evaluation

In this study, BLEU, METEOR, and ROUGE metrics was utilized to evaluate the performance of the models [23]. These metrics are widely used evaluation metric for image captioning models that measures the similarity between the generated captions and the ground-truth captions. However, BLEU score has some limitations, such as sensitivity to the length of the captions and potential for missing important information in the captions [24]. Therefore, it is often used in combination with METEOR and ROUGE, to provide a more comprehensive evaluation of image captioning models [25].

3. RESULT AND DISCUSSION

3.1. Performance comparison

The Table 1 lists the performance of image captioning models based on the BLEU, ROUGE, and METEOR evaluation metrics for greedy and beam search technique. Looking at the results, we can see that the Beam search algorithm performs better than the greedy search algorithm in terms of all evaluation metrics. it seems that the InceptionV3+BiLSTM model using beam search is the most effective model among those tested, achieving the highest scores in most metrics.

Table 1. Performance of image captioning models on greedy and beam search

Search	Algorithm	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE
Beam	InceptionV3+BiLSTM	0.524	0.584	0.602	0.571	0.286	0.572
	InceptionV3+LSTM	0.519	0.561	0.589	0.552	0.254	0.525
	VGG16+LSTM	0.517	0.545	0.556	0.521	0.221	0.495
	ResNet+LSTM	0.498	0.501	0.521	0.482	0.21	0.478
Greedy	InceptionV3+BiLSTM	0.514	0.541	0.551	0.525	0.276	0.551
	InceptionV3+LSTM	0.501	0.521	0.531	0.432	0.248	0.526
	VGG16+LSTM	0.498	0.531	0.512	0.481	0.2331	0.484
	ResNet+LSTM	0.477	0.494	0.532	0.461	0.221	0.475

Figure 4 shows the performance of models on all the opted evaluation metrics. In terms of BLEU scores, the InceptionV3+BiLSTM model using beam search achieves the highest scores, with BLEU-4 score of 0.571 and BLEU-3 score of 0.602. The worst BLEU score is achieved by ResNet+LSTM model using greedy search, with BLEU-4 score of 0.461. However, the BLEU-4 score is the lowest among all metrics, indicating that the models are struggling to generate longer and more complex sentences. For METEOR score, the InceptionV3+BiLSTM model using Beam search obtains the highest score of 0.286, while ResNet+LSTM model using beam search obtains the lowest score of 0.21.In terms of ROUGE score, the InceptionV3+BiLSTM model using beam search obtains the highest score of 0.572, while the ResNet+LSTM model using greedy search obtains the lowest score of 0.475.

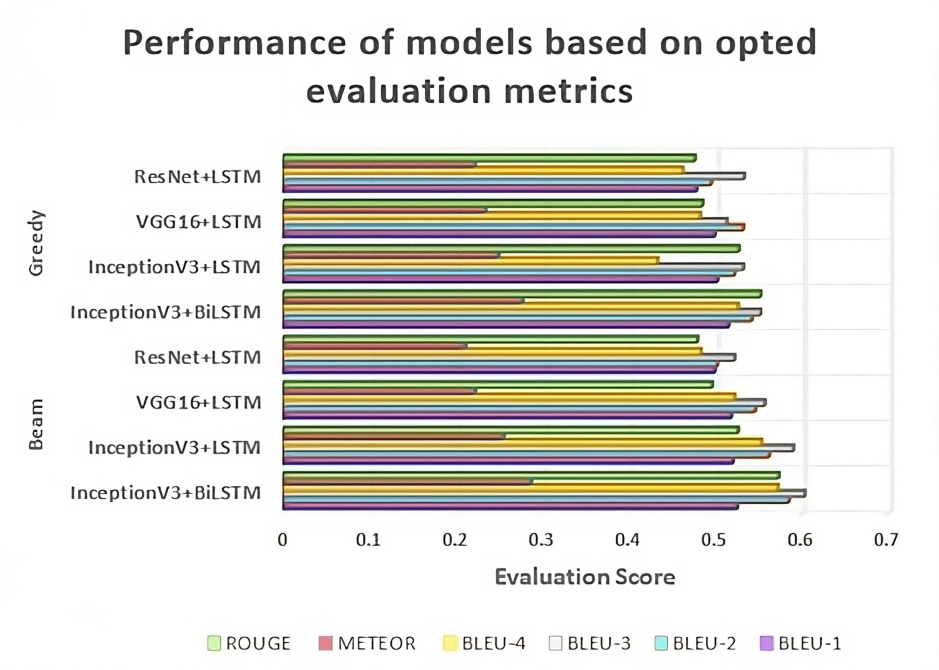

Figure 4. Evaluation score for each model's based on greedy and beam search

Figure 5 shows an example of predicted captions using InceptionV3+BiLSTM model on the testing set. The model achieves the highest BLEU score of 0.602 which is a quality of translation which is often better than human.

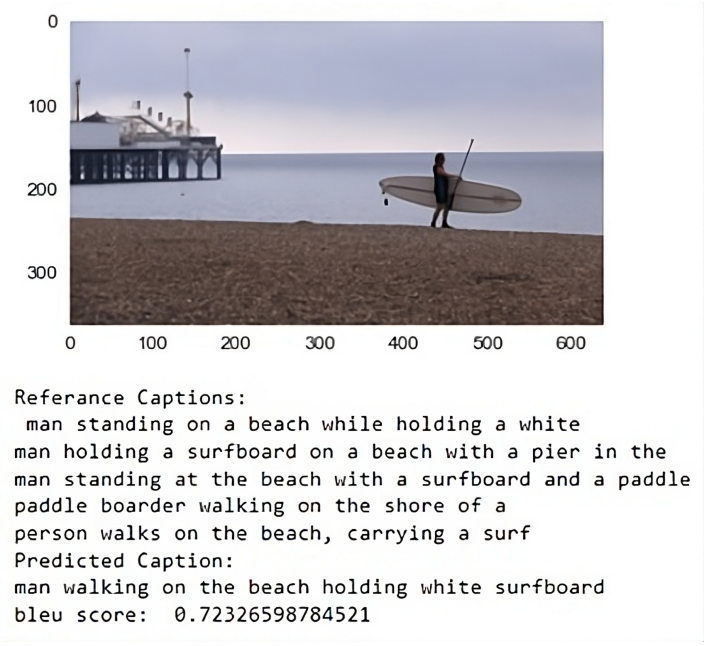

Figure 5. Example caption predictions using the highest performing image captioning model

3.2. Performance validation

The study evaluated the performance of the models by assessing their training and validation accuracy and loss. Figure 6 illustrate the gap between the training and validation loss, for every epoch of the InceptionV3+BiLSTM model, which was the best-performing one. The learning curves for each model indicate a significant improvement in performance with each epoch step, as both training and validation loss decrease over time.

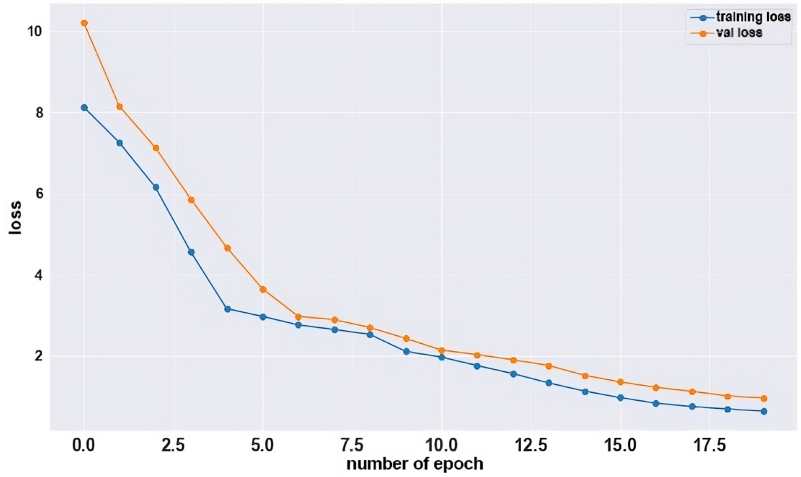

Figure 6. Difference of training and validation loss for each epoch of InceptionV3+BiLSTM model

3.3. Model architectures and training strategies

In recent years, the image captioning field has experienced growth due to advancements in deep learning models and the availability of large-scale image and caption datasets. This study aimed to create models based on transfer learning algorithms capable of accurately describing image content and producing human-like captions, with applications in accessibility, multimedia retrieval, and robot navigation. Transfer learning algorithms for image encoding and RNN for language decoding have surpassed human-like machine translation. While attention-based models have good BLEU scores, they are computationally slow. Transfer learning

models, on the other hand, are data and computationally efficient, offering better performance with fewer training epochs. However, to increase the size and diversity of the image dataset, data augmentation techniques such as random cropping, flipping, and rotation can be utilized. In addition, the models were trained for only 20 epochs, but higher epoch training can result in improved performance, as seen from the increasing accuracy and decreasing loss in each epoch.

4. CONCLUSION

Image captioning generates a natural language description of an image by mapping visual information to a textual representation. Recent research focuses on improving models with multimodal representations, attention mechanisms, and pre-trained models.However, challenges remain in improving models' robustness to diverse visual content and incorporating light-weighted models for better performance. The goal of this study research is to develop image captioning models based on transfer learning algorithms that can generate accurate and human-like captions, which requires the integration of both computer vision and NLP techniques. For that purpose, we created four image captioning models where ResNet50, VGG16, InceptionV3 models were utilized to encode the image feature and LSTM, BiLSTM models to generate captions from the encoder image. Out of all the experimental models, the InceptionV3+BiLSTM model performed with the highest BLEU score of over 60%, which can be considered better text generation than human. In future works, we want to fine-tune pre-trained models on large datasets to build more effective image captioning models, as it allows the model to leverage pre-existing knowledge and improve performance on the task. Additionally, we want to explore the use of language models in image captioning, and have made significant progress in generating captions that are not only accurate, but also coherent and semantically consistent.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the valuable feedback and constructive suggestions provided by the anonymous reviewers.

FUNDING INFORMATION

The authors declare that no funding was received for this research.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Maheen Islam	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	✓
Mahedi Hassan Ratul	✓		✓	✓		✓			✓					
Rezaul Haque	✓								✓					
Sazzad Hossain Rony					✓					✓				
Azharul Huq Asif	✓									✓				
Tanni Mittra		✓	✓	✓	✓	✓			✓	✓	✓	✓	✓	
Md Miskat Hossain	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Mahamudul Hasan	✓	✓	✓	✓	✓	✓		✓	✓	✓		✓	✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal Analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject Administration
Fu : **Fu**nding Acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest regarding the publication of this paper. No financial, personal, or professional relationships influenced the work reported in this manuscript.

INFORMED CONSENT

Not applicable. This study did not involve human participants, and therefore informed consent was not required.

ETHICAL APPROVAL

Not applicable. This study did not involve human participants or animal subjects, and therefore, ethical approval was not required.

DATA AVAILABILITY

The dataset used in this study is publicly available:

– The Microsoft COCO Captions dataset, which supports the findings of this research, can be accessed at https://paperswithcode.com/dataset/coco-captions. For citation details, refer to reference [26].

REFERENCES

[1] M. Hasan, N. Vasker, Md M. Hossain, Md I. Bhuiyan, J. Biswas, and M. R. A. Rashid, "Framework for fish freshness detection and rotten fish removal in Bangladesh using mask R–CNN method with robotic arm and fisheye analysis," *Journal of Agriculture and Food Research*, vol. 16, 2024, doi: 10.1016/j.jafr.2024.101139.

[2] M. Hon and N. M. Khan, "Towards Alzheimer's disease classification through transfer learning," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Nov. 2017, pp. 1166–1169, doi: 10.1109/BIBM.2017.8217822.

[3] S. Bai and S. An, "A survey on automatic image caption generation," *Neurocomputing*, vol. 311, pp. 291–304, Oct. 2018, doi: 10.1016/J.NEUCOM.2018.05.080.

[4] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018, doi: 10.1109/TGRS.2017.2776321.

[5] K. Xu *et al.*, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in *32nd International Conference on Machine Learning (ICML 2015)*, Feb. 2015 vol. 3, pp. 2048–2057.

[6] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Nov. 2017, pp. 6298–6306, doi: 10.1109/CVPR.2017.667.

[7] P. Anderson *et al.*, "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 6077-6086, doi: 10.1109/CVPR.2018.00636.

[8] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image Captioning with Semantic Attention," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 4651-4659, doi: 10.1109/CVPR.2016.503.

[9] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, vol. 9909, pp. 382–398, doi: 10.1007/978-3-319-46454-1_24.

[10] Q. Wang, J. Wan, and A. B. Chan, "On Diversity in Image Captioning: Metrics and Methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 2, pp. 1035–1049, Feb. 2022, doi: 10.1109/TPAMI.2020.3013834.

[11] V. Milewski, M.-F. Moens, and I. Calixto, "Are scene graphs good enough to improve Image Captioning?," *arXiv*, Sep. 2020, doi: 10.48550/arXiv.2009.12313.

[12] M. Zhang, Y. Yang, H. Zhang, Y. Ji, H. T. Shen, and T.-S. Chua, "More is better: Precise and detailed image captioning using online positive recall and missing concepts mining," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 32–44, Jan. 2019, doi: 10.1109/TIP.2018.2855415.

[13] T. Y. Lin *et al.*, "Microsoft COCO: Common objects in context," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8693, part 5, pp. 740–755, 2014, doi: 10.1007/978-3-319-10602-1_48.

[14] B. Bajracharya and D. Hua, "A preprocessing method for improved compression of digital images," *Journal of Computer Sciences and Applications*, vol. 6, no. 1, pp. 32-37, 2018, doi: 10.12691/jcsa-6-1-4.

[15] C. P. Chai, "Comparison of text preprocessing methods," *Natural Language Engineering*, vol. 29, no. 3, pp. 509-553, 2023, doi: 10.1017/S1351324922000213.

[16] Y. Chu, X. Yue, L. Yu, Mi. Sergei, and Z. Wang, "Automatic image captioning based on ResNet50 and LSTM with soft attention," *Wireless Communications and Mobile Computing*, 2020, pp. 1-7, doi: 10.1155/2020/8909458.

[17] E. Cetinic, "Towards generating and evaluating iconographic image captions of artworks," *Journal of Imaging*, vol. 7, no. 8, p. 123, 2021, doi: 10.3390/jimaging7080123.

[18] S. Mascarenhas and M. Agarwal, "A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Clas-

sification," *2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON)* Bengaluru, India, 2021, pp. 96-99, doi: 10.1109/CENTCON52345.2021.968794.
[19] H. Maru, T. Chandana, and D. Naik, "Comparison of Image Encoder Architectures for Image Captioning," in *5th International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2021, pp. 740-744, doi: 10.1109/ICCMC51019.2021.9418234.
[20] S. Degadwala, D. Vyas, H. Biswas, U. Chakraborty, and S. Saha, "Image Captioning Using Inception V3 Transfer Learning Model, in *2021 6th International Conference on Communication and Electronics Systems (ICCES)*, Coimbatre, India, 2021, pp. 1103-1108, doi: 10.1109/ICCES51350.2021.94891.
[21] S. R. Shah, S. Qadri, H. Bibi, S. M. W. Shah, M. I. Sharif, and F. Marinello, "Comparing inception V3, VGG 16, VGG 19, CNN, and ResNet 50: a case study on early detection of a rice disease," *Agronomy*, vol. 13, no. 6, p. 1633, 2023, doi: 10.3390/agronomy13061633.
[22] L. Yang, H. Wang, P. Tang, and Q. Li, "CaptionNet: A Tailor-made Recurrent Neural Network for Generating Image Descriptions," *IEEE Transactions on Multimedia*, vol. 23, pp. 835-845, 2021, doi: 10.1109/TMM.2020.2990074.
[23] G. Luo, L. Cheng, C. Jing, C. Zhao, and G. Song, "A thorough review of models, evaluation metrics, and datasets on image captioning," *IET Image Processing*, vol. 16, no. 2, pp. 311-332, 2022, doi: 10.1049/ipr2.12367.
[24] H. R. Tavakoliy, R. Shetty, A. Borji, and J. Laaksonen, "Paying Attention to Descriptions Generated by Image Captioning Models," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 2506-2515, doi: 10.1109/ICCV.2017.272.
[25] Z. K. Tawfeeq, M. S. Thesis, A hybrid deep learning model for image captioning, Department of Computer Engineering, Karabuk University, Karabuk, Turkey, 2024.
[26] X. Chen *et al.*, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint*, 2015, doi: 10.48550/arXiv.1504.00325.

BIOGRAPHIES OF AUTHORS

Maheen Islam is currently serving as an Associate Professor in the Department of Computer Science and Engineering at East West University, Bangladesh. She received the B.Sc. and M.Sc. degrees from the University of Dhaka, Bangladesh, in 1998 and 1999, respectively, and the Ph.D. degree in Wireless Mesh Networking from the Department of Computer Science and Engineering, University of Dhaka, in 2017. Her research interests include wireless mesh networks, wireless sensor networks, cognitive radio networks, and software-defined networks. She is a member of the Green Networking Research Group (GNR), IEEE, and BWIT. She can be contacted at email: maheen@ewubd.edu.

Mahedi Hassan Ratul has completed his Bachelor's degree from the Department of Computer Science and Engineering at East West University, Bangladesh. His research interests include machine learning, natural language processing, artificial intelligence, and data analysis. He can be contacted at email: mahammad.ratul1004@gmail.com.

Rezaul Haque has completed his Bachelor's degree from the Department of Computer Science and Engineering at East West University, Bangladesh. His research interests include deep learning, computer vision, image processing, and pattern recognition. He can be contacted at email: rezaulh603@gmail.com.

Sazzad Hossain Rony has completed his Bachelor's degree from the Department of Computer Science and Engineering at East West University, Bangladesh. His research interests include machine learning, natural language processing, artificial intelligence, and data analysis. He can be contacted at email: sazzad.cse050@gmail.com.

Azharul Huq Asif has completed his Bachelor's degree from the Department of Computer Science and Engineering at East West University, Bangladesh. His research interests include machine learning, internet of things (IoT), artificial intelligence, and data analysis. He can be contacted at email: asif.huq1998@gmail.com.

Tanni Mittra is a Senior Lecturer at the Department of Computer Science and Engineering, East West University, Dhaka, Bangladesh, where she has been a faculty member since 2016. She graduated with a first-class honours B.Sc. degree in Computer Science and Engineering from Khulna University, Bangladesh, in 2011, and an M.Sc. in Computer Science and Engineering from BUET, Bangladesh in 2015. Her research interests are primarily in the area of machine learning, deep learning, NLP, and knowledge graph, where she is the author/co-author of over 20 research publications. She can be contacted at email: tanni@ewubd.edu.

Md Miskat Hossain has completed his Bachelor's degree from the Department of Computer Science and Engineering at East West University, Bangladesh. His research interests include deep learning, computer vision, image processing, the internet of things (IoT), natural language processing, artificial intelligence, sensor networks, and real-time data analytics. Alongside his academic pursuits, he actively contributes to research and has published work in various journals. He can be contacted at email: miskatbd42@gmail.com.

Mahamudul Hasan is currently a Research Assistant at the University of Minnesota Twin Cities, USA. Previously, he served as a Senior Lecturer at East West University in Dhaka, Bangladesh, from 2017 to 2024. He earned both his M.S. and B.Sc. degrees in Computer Science and Engineering from the University of Dhaka, Bangladesh. His research interests encompass recommender systems, machine learning, deep learning, blockchain, large language models, and natural language processing. He can be contacted at email: munna09bd@gmail.com.