IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 13, No. 3, September 2024, pp. 3111~3118
ISSN: 2252-8938, DOI: 10.11591/ijai.v13.i3.pp3111-3118

Contextual embedding generation of underwater images using deep learning techniques

Shivani Kerai, Ganesh Khekare
School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India

Article Info

Article history:

Received Jan 27, 2024
Revised Mar 1, 2024
Accepted Mar 16, 2024

Keywords:

Attention mechanism
Contextual embeddings
Convolution neural network
Image feature extraction
Underwater object detection

ABSTRACT

This article delves into the cutting-edge realm of artificial intelligence, specifically focusing on its application in marine research via underwater image analysis. It introduces an innovative, integrated approach that combines object detection with image captioning tailored for the aquatic domain. Central to this approach is the advanced technique of image feature extraction, complemented by the strategic implementation of attention mechanisms within neural networks. These mechanisms are key in enhancing the precision and contextual understanding of underwater imagery. The efficacy of this method is underscored by extensive experiments on diverse underwater datasets. Results show notable improvements in detecting and describing complex underwater scenes, thereby providing invaluable insights for marine biologists, environmentalists, and the broader scientific community. This exploration marks a significant advancement in marine research, offering a new lens through which the underwater world can be understood and preserved.

Corresponding Author:

Ganesh Khekare
School of Computer Science and Engineering, Vellore Institute of Technology
Vellore 632014, India
Email: khekare.123@gmail.com

1. INTRODUCTION

The enigmatic underwater realm, with its inherent complexities and vastness, poses significant challenges for exploration and analysis. Traditional methodologies often fall short of effectively capturing and interpreting the intricate details of marine environments. Addressing these challenges, this chapter introduces a novel approach that marries advanced machine-learning algorithms for object detection with sophisticated image captioning techniques. At the heart of this methodology lies the innovative use of feature extraction and attention mechanisms in neural networks. Feature extraction in underwater imagery is pivotal, as it allows for the precise identification of various elements within a complex aquatic environment. By harnessing the power of deep learning, the proposed approach effectively extracts salient features from images, enabling more accurate and detailed object detection. This step is crucial in understanding the diverse and often hidden aspects of marine life and topography. Moreover, the incorporation of attention mechanisms in neural networks significantly enhances the analysis. These mechanisms allow the model to focus on specific areas of an image, akin to the way human attention works. By doing so, it can prioritize relevant features and suppress less important information, leading to more accurate and contextually rich interpretations. This is especially important in underwater settings where varying light conditions, turbidity, and other factors can obscure important details.

The integration of these advanced techniques not only improves the accuracy of object detection but also enriches the contextual understanding of underwater scenes. This breakthrough in marine research technology offers a comprehensive and nuanced perspective of the underwater world, facilitating groundbreaking insights and discoveries. Through this approach, marine researchers and environmentalists can gain a deeper, more informed understanding of aquatic ecosystems, aiding in conservation efforts and enhancing our knowledge of the largely unexplored oceanic frontiers. Ueda *et al.* [1] presenting a model to improve multimodal representations by integrating different modalities, such as recognized objects, sections of optical character recognition (OCR), tokens from OCR, and full images. A further encoding module for linguistic information in OCR tokens, the use of pre-trained contrastive language-image pre-training (CLIP) models, and the addition of attention blocks to enhance TextCaps performance are some of the study's major achievements. The proposed method outperforms baseline approaches on the TextCaps dataset, with a notable increase in the consensus-based image description evaluation (CIDEr) score. The implications for underwater object detection and image captioning include the potential benefits of multimodal integration, enhanced OCR techniques, and attention mechanisms for more accurate analysis in underwater scenarios. Jing *et al.* [2] focusing on context-driven captioning for news images, the development of a sentence correlation analysis algorithm to extract relevant named entities, and an emphasis on understanding global semantic relations through a semantic knowledge graph. The model generates template captions with placeholders for entities, filled in using an entity-linking algorithm, resulting in richer information, and outperforming state-of-the-art models. Kandala *et al.* [3] introduces an innovative approach to image captioning using transformer models and multilabel classification, with potential contributions. The study explores a transmission-based system for isolated sensing photo tagging, leveraging its sequential data handling capabilities.

The novel use of multilevel categorization as a supplementary thing proves effective in handling limited training data and improving overall model performance. The multitask framework, featuring a common transformer encoder and separate decoders for caption generation and multilabel classification, allows effective training with limited data. Experimental validation on the University of California-merced caption dataset demonstrates the method's effectiveness over traditional captioning approaches. Im *et al.* [4] presents an innovative method for captioning images by integrating bidirectional analysis for contextual characteristics into a bidirectional content-adaptive recurrent unit (Bi-CARU) model, which captures important image features. To improve relationship extraction and determination, an attention layer is added, which focuses on the features generated by content-adaptive recurrent units (CARU's) context-adaptive gate. The suggested convolutional neural network (CNN) to Bi-CARU system works superior to existing systems, according to iterative data, highlighting its competence in extracting contextual information and offering comprehensive image descriptions. Yeh *et al.* [5] introduces a unique approach to underwater object detection, emphasizing the joint learning of object detection models by color conversion and object detection through a lightweight deep neural network. contributes by introducing a joint learning approach, addressing color distortion in underwater images with a lightweight model suitable for battery-powered autonomous underwater vehicles (AUVs). The proposed model generates underwater images for training, overcoming data scarcity challenges, and includes an effective color conversion module to enhance object detection while maintaining computational efficiency. Experimental validation demonstrates the model's promising results. Galassi *et al.* [6] offers a comprehensive overview of attention mechanisms in neural architectures, specifically focusing on their applications in natural language processing (NLP). As it explores advanced attention models that can play a pivotal role in image captioning and object detection. The key contributions include a systematic overview and taxonomy of attention models, a unified model for attention architectures, and discussions on empirical applications across various NLP tasks, along with insights into future directions and challenges in the field. Liu *et al.* [7] covers all the important topics related to underwater object detection in detail. The study discusses difficulties found in underwater settings, like target blur, low contrast, and color offset.

Using the swin transmission as the solid backup, integrating the path aggregation network for improvised different level attribute mixture, and enhancing region of interests (ROI). pooling to align ROIs are some of the major achievements, implementation of online hard example mining (OHEM) for efficient training, and comprehensive experimental validation and comparisons showcasing the superior performance of the proposed algorithm. The literature review spans critical topics, including underwater object detection challenges, the swin transformer and deep learning applications, feature fusion techniques, ROI pooling variants, hard example mining, comparative studies of object detection models, and challenges in underwater image processing. This comprehensive exploration contextualizes the paper's contributions within the broader landscape, emphasizing its significance in addressing the specific challenges of underwater object detection through advanced deep-learning approaches. Wang *et al.* [8] makes a substantial contribution to the detection of underwater targets by combining deep learning methods with polarization imaging. The technique reduces the effect of backscattered light. To address issues including complicated clustering, lower side identification, and ultimate cluster. The suggested approach maximizes detection and feature extraction by combining polarization gradient and edge detection algorithms, demonstrating improved performance in clustering settings. It underscores the complexity of underwater target detection and the limitations of traditional imaging methods.

The methodology employs deep learning, constructing a neural network with input from four polarization component images to map target characteristics. Experimental results, conducted in simulated turbid underwater environments, demonstrate the method's superiority in detecting multiple material targets, particularly in scenarios with overlapping targets. It emphasizes the novelty of the approach, attributing its effectiveness to the innovative use of polarization parameters and the integration of deep learning techniques, positioning it as a promising method for polarized target detection in challenging underwater environments. Li *et al.* [9] provides a comprehensive solution to enhance underwater target detection and positioning for remotely operated vehicles (ROVs) [10]. Addressing challenges posed by poor visibility and dynamic underwater environments, the authors propose a vision-based approach combining the you only look once transformer (YOLO-T) algorithm, an enhanced version of YOLOv5, with a target positioning algorithm. YOLO-T [11] incorporates the ghost module and squeeze-and-excitation (SE) attention module for improved speed and accuracy, and image processing techniques further enhance detection results. The target positioning algorithm [12] estimates both position and attitude using a monocular camera, involving artificial underwater target design and feature point detection [13]. Extensive experiments in various settings validate the accuracy and efficiency of the proposed approach, emphasizing YOLO-T's effectiveness in target detection and the positioning algorithm's robustness in different underwater conditions. Stability tests confirm continuous stability and robustness, which are crucial for real-world applications [14]. The paper concludes that the approach significantly enhances underwater target detection and positioning, suggesting future work to refine the YOLO-T algorithm for underwater targets and extend the method to real-time tracking in diverse underwater missions [15]. Research contributes significantly to underwater robotics by integrating advanced deep learning algorithms with innovative positioning techniques, promising advancements in underwater robotic applications [16].

Overall, this paper outlines a pioneering approach to exploring the complexities of underwater environments through the integration of advanced machine learning and image processing techniques [17]. Traditional methods often struggle to capture and interpret the intricate details of marine settings effectively [18]. Each contribution, from improving multimodal integration [19] and enhancing OCR [20] techniques to innovative uses of deep learning for color distortion correction in underwater imagery [21], underscores the potential of this approach to address the specific challenges faced in underwater exploration [22]. Notably, the application of deep learning in conjunction with polarization imaging and advanced detection algorithms [23] shows promise in overcoming visibility [24] and scattering issues common in marine environments. Through such technological advancements, the methodology not only promises to enhance our understanding of aquatic ecosystems but also contributes to the development of more effective conservation strategies and underwater robotic applications, marking a significant leap forward in the field of marine exploration and research.

2. METHOD

2.1. Image preprocessing

The initial stage in our underwater image analysis involves a systematic image preprocessing protocol designed to standardize the input data for machine learning analysis. This step is critical for ensuring consistency across the dataset, a fundamental requirement for the neural network's effective performance. Each image is resized to a uniform dimension of 299×299 pixels. This uniformity is crucial for input consistency across the information. Resolution is generalized to a standard scale, increasing the throughput of system testing by ensuring that input data operates on a common scale. To augment the diversity of our dataset and thereby enhance the model's generalization capability, we apply several image augmentation techniques. These include rotation (at angles of 0°, 90°, 180°, and 270°), horizontal and vertical flipping, and random cropping. Each technique is applied to simulate varied environmental conditions and perspectives encountered in underwater imagery.

2.2. Feature extraction with pretrained convolution neural network

Following preprocessing, we extract relevant features from the images using the Inception ResNetV2 model, pretrained on the ImageNet dataset. This stage is vital for identifying high-level features that are crucial for the accurate analysis and understanding of underwater scenes. The Inception ResNetV2 model is chosen for its depth and complex architecture, which is highly effective in capturing detailed textures and patterns in the images. Utilizing TensorFlow and Keras libraries, we leverage Inception ResNetV2's pretrained capabilities on the ImageNet dataset to enhance feature detection and extraction. This approach allows the model to effectively identify and highlight key aspects of the underwater imagery, which are crucial for subsequent analysis phases. In the context of marine research, these extracted features play a vital role in identifying species, assessing habitat conditions, and monitoring changes in the underwater environment. The use of a pretrained CNN like Inception ResNetV2 not only streamlines the feature extraction process but also ensures a high level of accuracy and reliability in the analysis of complex underwater scenes.

2.2.1. Inception ResNetV2

The utilization of TensorFlow and Keras libraries provides a solid foundation for model development, capitalizing on their capabilities for efficient computation and neural network construction. The deep CNN component, specifically the Inception ResNetV2 model, serves as a powerful tool for extracting intricate features from input images. A CNN design called Inception ResNetV2 integrates the ideas of residual networks and the Inception architecture. To better utilize the benefits of residual connections for training convergence and model performance, Google unveiled it as an addition to the Inception family of models. Because the model has been pre-trained on the ImageNet dataset, it has access to a plethora of knowledge that improves its capacity to identify and record significant patterns in underwater imagery. Inception ResNetV2 architecture is shown in Figure 1.

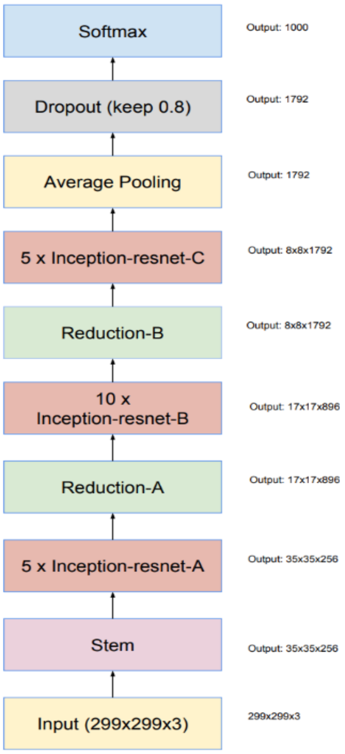


Figure 1. Inception ResNetV2 architecture

2.2.2. Encoder

The input image is processed by the encoder component, which then converts it into a fixed-dimensional representation. Because convolutional neural networks are good at extracting hierarchical features from visual data, they are frequently utilized as image encoders. The encoder in this design is a pre-trained CNN, like Inception ResNetV2. The last layer of the encoder summarizes the important details from the input image by capturing high-level characteristics and creating a context vector. The invisible level [25] h_t is determined as shown in (1). The encoded model system of the proposed system is depicted in Figure 2.

$$h_t = f(W^{(hh)}h_{t-1} + W^{(hx)}x_t) \tag{1}$$

2.2.3. Attention mechanism

The attention mechanism integrated into the model plays a pivotal role in refining the caption generation process by directing the model's focus to pertinent regions within the input image. Two specific types of attention mechanisms, namely additive and standard, are strategically employed to enhance the model's performance. The additive attention mechanism is instrumental in dynamically allocating weights to different parts of the image, allowing the model to prioritize specific regions based on their relevance to the context. This adaptability ensures that the model concentrates more on salient features or objects within the image, contributing to the level of contextually accurate tags. Simultaneously, the standard authentication system is employed to refine the model's attention allocation process further. This mechanism aids in the systematic examination of different regions of the image, enabling the model to capture nuanced details and relationships that might be crucial for generating precise and meaningful captions. By incorporating both additive and standard attention mechanisms, the model benefits from a comprehensive approach to image analysis during the captioning process. This dual attention strategy enhances the model's capability to synthesize captions that not only accurately describe the content of the image but also reflect a nuanced understanding of the contextual relationships among various elements within the visual data. Ultimately, the attention mechanisms contribute significantly to the overall contextual accuracy and richness of the generated image captions. By combining advanced machine learning algorithms for object detection with sophisticated image captioning techniques, this new methodology aims to overcome these challenges. Central to this approach is the use of feature extraction and attention mechanisms in neural networks, which are crucial for identifying and analyzing various elements within the aquatic environment accurately. An attention mechanism that concentrates on various areas of the input image is integrated into the caption decoder. The decoder selectively focuses on specific segments of the input sequence by using attention. For every example, the attention receives a series of data as a source and outputs an "attention" vector.

$$\alpha_{ts} = \frac{\exp(score(h_t, \bar{h}_s))}{\sum_{s'=1}^{S} \exp(score(h_t, \bar{h}_{s'}))} \text{ [Attention weights]} \tag{2}$$

In (2) uses a softmax efficiency of the encoder's final series to determine the attention weights or α_{ts}.

$$c_t = \sum_s \alpha_{ts} \bar{h}_s \text{ [Context vector]} \tag{3}$$

The context vector is determined by (3) using the weighted addition of the encoded finals. Where s is encoder index, t is decoder index, α_{ts} is weights of attention, $\bar{h}_s$ is series of encoded finals which are going to come, h_t is decoded state attending to the seq. c_t is final content terminology, and α_t is resultant output by combining the "context" and "query".

$$score(h_t, \bar{h}_s) = \begin{cases} h_t{}^T W \bar{h}_s & \text{[Luong's multiplicative style]} \\ v_a{}^T \tanh(W_1 h_t + W_2 \bar{h}_s) & \text{[Bahdanu's additive style]} \end{cases} \tag{4}$$

The score function (4), which calculates the scalar logit-score for a key-query pair, is the last stage. For this reason, the multiplicative style of Luong and the additive style of Bahdanu are frequently employed.

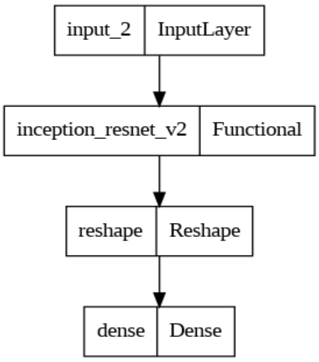
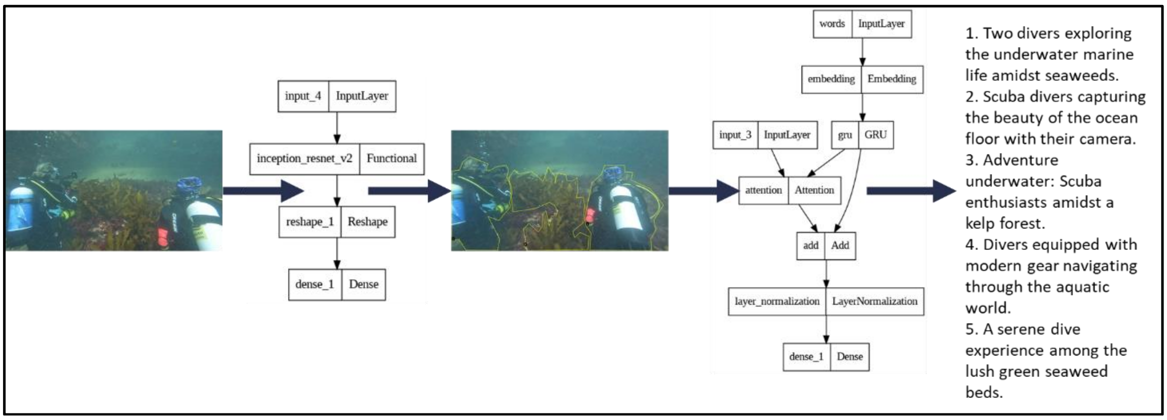


Figure 2. Encoder model architecture of proposed model

3. RESULTS AND DISCUSSION

In this system, the encoder's role is to discern and extract significant features from an image. This task is accomplished using a deep CNN based on Inception ResNetV2, renowned for its efficiency in processing visual information. Following this, the decoder comes into play, where it is enhanced by an attention system. This setup enables the system to concentrate on perticular things of the image while generating a sequence of words, resulting in a coherent and relevant caption. The decoder employs a RNN with GRU, a sophisticated framework known for its proficiency in handling sequential data and producing precise descriptive captions. This architecture is a strategic blend of advanced neural network techniques, ensuring the generation of accurate and context-aware captions for a wide range of underwater photos. The pipeline block diagram of tag producing system is as shown in Figure 3.

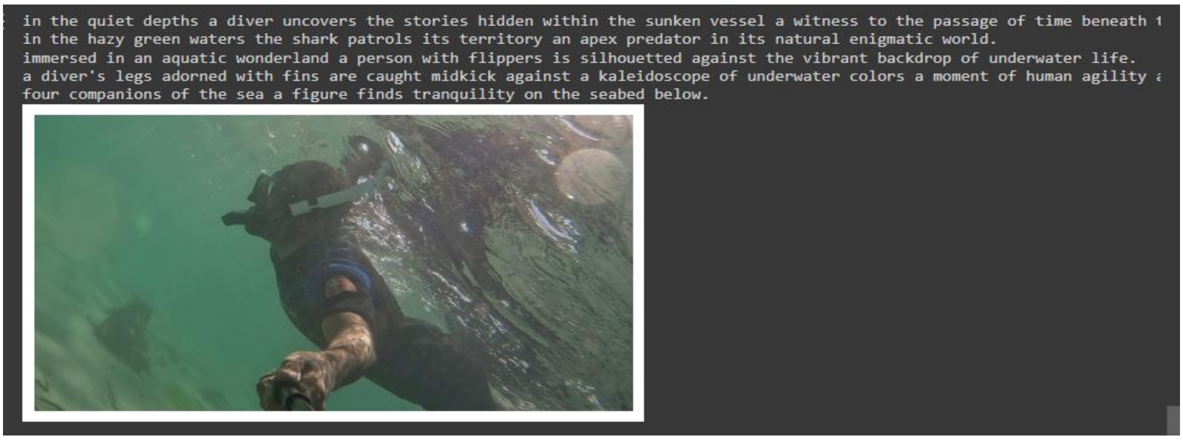


1. Two divers exploring the underwater marine life amidst seaweeds.
2. Scuba divers capturing the beauty of the ocean floor with their camera.
3. Adventure underwater: Scuba enthusiasts amidst a kelp forest.
4. Divers equipped with modern gear navigating through the aquatic world.
5. A serene dive experience among the lush green seaweed beds.

Figure 3. Pipeline architecture of caption generation model

3.1. Caption generation and model evaluation

The model's caption generation process incorporates a probabilistic approach, a strategy aimed at introducing variability and creativity in the language used to describe the visual content. This approach is particularly pertinent in the context of underwater scenes, which are inherently diverse, dynamic, and sometimes unpredictable. In a probabilistic caption generation approach, the model does not deterministically produce a single fixed caption for a given image. Instead, it introduces an element of randomness or uncertainty into the captioning process. The model assigns probabilities to different words or phrases, allowing for a range of potential captions to be generated for the same image. By adopting a probabilistic strategy, the model introduces linguistic variety in the generated captions. This means that for a single image, the model has the capacity to produce multiple, diverse captions during different runs or instances of caption generation. The resulatant probabilistic caption generation is as shown in Figure 4.



```
in the quiet depths a diver uncovers the stories hidden within the sunken vessel a witness to the passage of time beneath
in the hazy green waters the shark patrols its territory an apex predator in its natural enigmatic world.
immersed in an aquatic wonderland a person with flippers is silhouetted against the vibrant backdrop of underwater life.
a diver's legs adorned with fins are caught midkick against a kaleidoscope of underwater colors a moment of human agility
four companions of the sea a figure finds tranquility on the seabed below.
```

Figure 4. Resultant probabilistic caption generation

Our model exhibits a marked improvement in extracting salient features from underwater imagery and generating contextually relevant captions. The probabilistic approach to caption generation not only introduces linguistic diversity but also enhances the relevance and accuracy of the captions produced. As delineated in Table 1, the progressive optimization of our model versions-from image_captioning_model_v1_6_30_finetune to image_captioning_model_v4_40_200_finetune demonstrates substantial advancements in both training and validation metrics, with the latter achieving a training accuracy of 87.57% and a validation accuracy of 92.10%. Comparing our results with existing studies, it's evident that the integration of an attention system significantly augments the system's capability to concentrate on pertinent aspects of an image, thereby enriching the quality and relevance of generated captions. This approach mitigates the limitations observed in traditional CNN-based models, which may overlook dynamic and context-specific elements in underwater scenes.

Table 1. Model evaluation

Model	Training loss	Validation loss	Training accuracy	Validation accuracy
image_captioning_model_v4_40_200_finetune	0.1243	0.0790	0.8757	0.9210
image_captioning_model_v3_30_150_finetune	0.1788	0.1257	0.8212	0.8743
image_captioning_model_v2_10_50_finetune	0.2208	0.1976	0.7792	0.8021
image_captioning_model_v1_6_30_finetune	0.4391	0.2775	0.5609	0.7225

Despite these advancements, our approach is not without its limitations. The model's performance, while superior in controlled settings, may vary in real-world applications due to the unpredictable nature of underwater environments. Further, the complexity of the model architecture necessitates substantial computational resources, potentially limiting its accessibility for real-time applications. The promising results of this study pave the way for future research to explore the integration of more nuanced probabilistic models and advanced attention mechanisms. Future work could also investigate the application of this framework in other complex imaging contexts, potentially broadening the scope and utility of artificial intelligence (AI) driven caption generation. In summary, this study underscores the efficacy of combining deep learning architectures with attention mechanisms to enhance automated image captioning, particularly in the nuanced and variable domain of underwater imagery. Our findings not only give to the present state of wisdom but also highlight the potential for upcoming technologies in the field of AI-driven image analysis. This research exemplifies how technological advancements can be leveraged to improve our understanding and interaction with the marine ecosystem, ultimately contributing to its preservation and study.

4. CONCLUSION

The findings presented in this paper underscore the transformative potential of machine learning technologies in the realm of marine research, setting a new benchmark for the field. By leveraging advanced pre-trained models like Inception ResNetV2 and incorporating sophisticated attention mechanisms, our approach has demonstrated significant improvements in the accuracy and contextual relevance of automated image captions over previous methodologies. This enhancement is not merely incremental; it represents a paradigm shift in the precision and applicability of AI tools for marine studies. Our model outperforms previous work by harnessing the latest advancements in deep learning to capture the nuanced intricacies of marine imagery. The integration of attention mechanisms allows for a more granular understanding of complex marine scenes, enabling the model to identify and describe specific elements with unprecedented accuracy. This leap in performance is pivotal for marine conservation efforts, as it provides researchers and practitioners with a powerful tool to monitor, analyze, and protect the marine environment with an efficiency and scale hitherto unachievable. The future scope of this research is vast and promising. The next steps consist of increasing data to include a larger number of marine ecosystems and species, further refining the model's ability to discern and articulate the rich biodiversity of our oceans. Additionally, integrating real-time data analysis capabilities could revolutionize how we respond to environmental challenges, allowing for proactive rather than reactive conservation strategies. The potential for cross-disciplinary applications is also immense, ranging from enhancing climate change models with more accurate oceanic data to improving sustainable fishing practices through a better understanding of marine habitats. Moreover, the advancement of explainable artificial intelligence (XAI) techniques presents an exciting frontier for making these complex models more transparent and accessible to a wider audience, including policymakers, conservationists, and the public. By demystifying the workings of AI in marine research, we can foster a more informed and collaborative approach to tackling the pressing environmental difficulties in front of our earth.

REFERENCES

[1] A. Ueda, W. Yang, and K. Sugiura, "Switching text-based image encoders for captioning images with text," *IEEE Access*, vol. 11, pp. 55706–55715, 2023, doi: 10.1109/ACCESS.2023.3282444.

[2] Y. Jing, X. Zhiwei, and G. Guanglai, "Context-driven image caption with global semantic relations of the named entities," *IEEE Access*, vol. 8, pp. 143584–143594, 2020, doi: 10.1109/ACCESS.2020.3013321.

[3] H. Kandala, S. Saha, B. Banerjee, and X. X. Zhu, "Exploring transformer and multilabel classification for remote sensing image captioning," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2022.3198234.

[4] S. K. Im and K. H. Chan, "Context-adaptive-based image captioning by Bi-CARU," *IEEE Access*, vol. 11, pp. 84934–84943, 2023, doi: 10.1109/ACCESS.2023.3302512.

[5] C. H. Yeh *et al.*, "Lightweight deep neural network for joint learning of underwater object detection and color conversion," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 11, pp. 6129–6143, 2022, doi: 10.1109/TNNLS.2021.3072414.

[6] A. Galassi, M. Lippi, and P. Torroni, "Attention in natural language processing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 10, pp. 4291–4308, 2021, doi: 10.1109/TNNLS.2020.3019893.

[7] J. Liu, S. Liu, S. Xu, and C. Zhou, "Two-stage underwater object detection network using swin transformer," *IEEE Access*, vol. 10, pp. 117235–117247, 2022, doi: 10.1109/ACCESS.2022.3219592.

[8] G. Wang *et al.*, "Polarization-enhanced underwater detection method for multiple material targets based on deep-learning," *IEEE Photonics Journal*, vol. 15, no. 6, pp. 1–6, 2023, doi: 10.1109/JPHOT.2023.3326158.

[9] Y. Li, W. Liu, L. Li, W. Zhang, J. Xu, and H. Jiao, "Vision-based target detection and positioning approach for underwater robots," *IEEE Photonics Journal*, vol. 15, no. 1, pp. 1–12, 2023, doi: 10.1109/JPHOT.2022.3228013.

[10] L. Bai, W. Zhang, X. Pan, and C. Zhao, "Underwater image enhancement based on global and local equalization of histogram and dual-image multi-scale fusion," *IEEE Access*, vol. 8, pp. 128973–128990, 2020, doi: 10.1109/ACCESS.2020.3009161.

[11] N. Deluxni, P. Sudhakaran, Kitmo, and M. F. Ndiaye, "A review on image enhancement and restoration techniques for underwater optical imaging applications," *IEEE Access*, vol. 11, pp. 111715–111737, 2023, doi: 10.1109/ACCESS.2023.3322153.

[12] G. Khekare and Midhunchakkravarthy, "Smart image recognition system for the visually impaired people," in *2023 International Conference on Energy, Materials and Communication Engineering (ICEMCE)*, 2023, pp. 1–6. doi: 10.1109/ICEMCE57940.2023.10434130.

[13] G. Khekare *et al.*, "Optimizing network security and performance through the integration of hybrid GAN-RNN models in SDN-based access control and traffic engineering," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 12, pp. 596–606, 2023, doi: 10.14569/IJACSA.2023.0141262.

[14] J. Zhang *et al.*, "Marine organism detection based on double domains augmentation and an improved YOLOv7," *IEEE Access*, vol. 11, pp. 68836–68852, 2023, doi: 10.1109/ACCESS.2023.3287932.

[15] G. Khekare, C. Masudi, Y. K. Chukka and D. P. Koyyada, "Text Normalization and Summarization Using Advanced Natural Language Processing," *2024 International Conference on Integrated Circuits and Communication Systems (ICICACS)*, Raichur, India, 2024, pp. 1-6, doi: 10.1109/ICICACS60521.2024.10498983.

[16] X. Ye *et al.*, "A joint-training two-stage method for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022, doi: 10.1109/TGRS.2022.3224244.

[17] D. Xiang, H. Wang, D. He, and C. Zhai, "Research on histogram equalization algorithm based on optimized adaptive quadruple segmentation and cropping of underwater image (AQSCHE)," *IEEE Access*, vol. 11, pp. 69356–69365, 2023, doi: 10.1109/ACCESS.2023.3290201.

[18] F. Wu, Z. Cai, S. Fan, R. Song, L. Wang, and W. Cai, "Fish target detection in underwater blurred scenes based on improved YOLOv5," *IEEE Access*, vol. 11, pp. 122911–122925, 2023, doi: 10.1109/ACCESS.2023.3328940.

[19] W. Zhao, X. Wu, and J. Luo, "Cross-domain image captioning via cross-modal retrieval and model adaptation," *IEEE Transactions on Image Processing*, vol. 30, pp. 1180–1192, 2021, doi: 10.1109/TIP.2020.3042086.

[20] G. Khekare, P. Verma, and S. Raut, "The smart accident predictor system using internet of things," in *Cloud IoT*, Boca Raton: Chapman and Hall/CRC, 2022, pp. 163–175. doi: 10.1201/9781003155577-14.

[21] T. Wang *et al.*, "Underwater image enhancement based on optimal contrast and attenuation difference," *IEEE Access*, vol. 11, pp. 68538–68549, 2023, doi: 10.1109/ACCESS.2023.3292275.

[22] M. A. Kastner *et al.*, "Imageability- and length-controllable image captioning," *IEEE Access*, vol. 9, pp. 162951–162961, 2021, doi: 10.1109/ACCESS.2021.3131393.

[23] C. Y. Li, J. C. Guo, R. M. Cong, Y. W. Pang, and B. Wang, "Underwater image enhancement by Dehazing with minimum information loss and histogram distribution prior," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5664–5677, 2016, doi: 10.1109/TIP.2016.2612882.

[24] Y. Li, J. Li, Y. Li, H. Kim, and S. Serikawa, "Low-light underwater image enhancement for deep-sea tripod," *IEEE Access*, vol. 7, pp. 44080–44086, 2019, doi: 10.1109/ACCESS.2019.2897691.

[25] C. Liu, Z. Mao, T. Zhang, A. A. Liu, B. Wang, and Y. Zhang, "Focus your attention: a focal attention for multimodal learning," *IEEE Transactions on Multimedia*, vol. 24, pp. 103–115, 2022, doi: 10.1109/TMM.2020.3046855.

BIOGRAPHIES OF AUTHORS

Shivani Kerai currently pursuing a Master of Technology in Computer Science Engineering with specialization in artificial intelligence and machine learning from Vellore Institute of Technology, Vellore, Tamil Nadu, India. She also received a B.E. (Computer) degree from Gujarat Technological University, India in 2021. Her research interests are machine learning, deep learning, computer vision, natural language processing, image processing, and neural networks. She can be contacted at email: shivani.kerai155@gmail.com.

Ganesh Khekare holds a Doctor of Computer Science and Engineering from Bhagwant University, India in 2021. He is a postdoc fellow from Lincoln University, Malaysia. He also received his B.E. and M.E. (CSE) from Nagpur University, India in 2010 and 2013, respectively. He is currently an associate professor at the Department of Computer Science & Engineering at Vellore Institute of Technology, Vellore, India. His research includes artificial intelligence, machine learning, data science, networks, and internet of things. He has published over 70 papers in international journals and conferences. He has done 5 patents and 10 copyrights. He is an active member of various professional societies like ACM, ISTE, IEEE senior member, and IEI. He can be contacted at email: khekare.123@gmail.com.