CaptionCraft: Scientific Figures Image Caption

Ankita Manjarekar
Department of Data Science
University of New Haven
amanj4@unh.newhaven.edu

Chetana Nannapaneni
Department of Data Science
University of New Haven
cnann1@unh.newhaven.edu

Abstract

CaptionCraft presents an innovative Image Caption Generator tailored for scientific plots and figures, leveraging the SCICAP dataset comprising over 2 million scientific figures and captions. Our approach involves preprocessing techniques to clean captions, feature extraction using pre-trained models like VGG16 and ResNet50, and tokenization of captions for model input. We employ state-of-the-art architectures such as Vision Transformers (ViT), GPT-2, and SWIN Transformers to develop a model integrating Dense, Embedding, and self-attention mechanisms. Evaluation based on the BLEU-4 metric aims for a benchmark score of 0.5, and our model achieves a notable BLEU score of 0.411 through extensive experimentation. Successful integration of Vision Transformer and GPT-2 models significantly contributes to our model's performance. Further research is warranted to explore enhancements in model architecture and fine-tuning strategies, highlighting the need for scalable computing resources for future advancements in image captioning technology.

1 Introduction

In today's world of online research and sharing, the significance of visual elements, particularly scientific figures, and plots, cannot be overstated. These visual representations encapsulate key findings, methodologies, and insights, serving as indispensable tools for conveying complex scientific concepts in a concise and accessible manner. However, while the importance of these figures is undeniable, their potential to enhance understanding can be further augmented through the addition of informative captions. Captions provide essential context, interpreting the significance of the depicted information and facilitating comprehension for diverse audiences, from experts in the field to novices seeking to grasp fundamental concepts.

Recognizing the pivotal role of captions in scientific communication, CaptionCraft represents a pioneering initiative bridging the realms of computer vision and natural language processing. Our mission is to develop a specialized Image Caption Generator tailored explicitly for scientific figures and plots, leveraging cutting-edge techniques and resources to bridge the gap between visual content and textual description. Central to our approach is the utilization of the SCICAP dataset, a comprehensive repository containing a vast array of scientific figures and their corresponding captions. This dataset provides a rich and diverse source of data, spanning multiple disciplines and encompassing a wide range of visual representations encountered in scholarly literature.

CaptionCraft embarks on a multifaceted journey, encompassing preprocessing, feature extraction, model development, and evaluation, all with the primary goal of advancing the frontier of image captioning in the scientific domain. Our methodology encompasses sophisticated preprocessing techniques to clean captions, ensuring clarity and coherence, while state-of-the-art pre-trained models for image feature extraction and text generation form the bedrock of our caption generation pipeline. By harnessing the power of vision transformers, language models, and self-attention mechanisms, we seek to infuse our model with the ability to discern intricate patterns and variations within scientific figures, enabling the generation of informative and contextually relevant captions.

Through meticulous experimentation and rigorous evaluation using established metrics such as BLEU-4, CaptionCraft endeavors to assess the efficacy and robustness of our approach. Our ultimate aim is to contribute to the development of robust and accurate image captioning systems tailored specifically for the unique challenges and nuances of scientific communication. By providing researchers, educators, and enthusiasts with a powerful tool for augmenting the comprehension and accessibility of scientific literature, CaptionCraft aspires to empower knowledge dissemination and facilitate deeper insights into the details of scientific discovery."

2. Related Work

1. The paper "**Visual Captions: Augmenting Verbal Communication with On-the-fly Visuals**" presents a real-time system called Visual Captions that integrates with video conferencing platforms to enrich verbal communication with relevant visuals. The authors fine-tuned a large language model using 1595 visual intents gathered from crowd workers and evaluated the system through a user study and a longer deployment study. The results suggest that Visual Captions has the potential to facilitate live conversations by providing valuable visual content, as the system is designed as a virtual camera compatible with popular video conferencing platforms and can be easily installed by end-users as a Chrome browser plugin leveraging state-of-the-art real-time speech transcription.
2. **Yiyu Wang, Jungang Xu, Yingfei Sun; End-to-End Transformer Based Model for Image** :The paper presents an end-to-end Transformer-based model for image captioning, moving away from the traditional CNN-LSTM architectures. The authors adopted Swin Transformer as the backbone encoder to extract grid-level features from images, and build

a refining encoder and a decoder to capture the intra-relationship between features and generate captions word by word. Furthermore, they introduce the mean pooling of grid features as a global feature to enhance the interaction between vision and language features, improving the model's overall performance. The proposed approach is evaluated on the MSCOCO dataset, demonstrating the effectiveness of the pure Transformer-based architecture for the image captioning task.

3. **The Unreasonable Effectiveness of Recurrent Neural Networks by Andrej Karpathy:** In his blog post "The Unreasonable Effectiveness of Recurrent Neural Networks", Andrej Karpathy discusses the power and robustness of Recurrent Neural Networks (RNNs), particularly in the context of Computer Vision tasks such as image captioning. Karpathy shares his personal experience of training RNNs and witnessing their "magical" outputs, highlighting the simplicity of the models compared to the quality of the results. He expresses confidence in the continued innovation and widespread adoption of RNNs as a critical component of intelligent systems. Karpathy even trained an RNN on the source file of the blog post itself, though he notes that the limited amount of data (46K characters) may not be sufficient to properly feed the RNN.

3 Methodology

To achieve the goal of developing an Image Caption Generator tailored for scientific plots and figures, the study begins by leveraging the SCICAP dataset for training purposes. This dataset, specifically curated for scientific images, ensures that the model learns from relevant and domain-specific examples. Subsequently, a series of preprocessing techniques are implemented to refine the captions associated with the images. These techniques aim to eliminate noise, correct errors, and enhance the clarity and coherence of the textual descriptions.

Following the preprocessing phase, image features are extracted using a pre-trained model. This step is crucial as it allows the model to capture the essential visual elements of the scientific plots and figures. These extracted features are then tokenized to prepare them for ingestion into the model. By representing the images in a format compatible with the model's architecture, the tokenization process facilitates seamless integration of visual information with textual data.

The core of the methodology involves the development of a sophisticated model architecture that combines Dense layers, Embedding layers, self-attention mechanisms, and vision transformers. By integrating these components, the model gains the ability to effectively analyze both the visual and textual inputs and generate accurate and contextually relevant captions for the scientific images. This comprehensive approach ensures that the generated captions not only describe the visual content accurately but also capture the underlying scientific context, thus enhancing the utility and applicability of the Image Caption Generator in scientific research and analysis.

3.1 Datasets

The SCICAP dataset is a cornerstone within the domain of scientific image analysis, offering a vast collection of meticulously curated scientific figures and plots alongside their corresponding captions. Its comprehensive repository serves as an invaluable resource for training and evaluating image captioning models tailored specifically to this specialized field. Key attributes of the dataset include:

Extensive Coverage: With over 2 million scientific figures and plots, the dataset encompasses a broad spectrum of visual data, providing researchers with a diverse and comprehensive resource for their investigations.
Representative Subset: For our project, we have selected a random sample of 201 images from the dataset, ensuring a manageable scope while retaining the dataset's inherent diversity and complexity.

Spanning across various disciplines, the SCICAP dataset captures a wide array of scientific knowledge, featuring figures and captions sourced from research papers across eight distinct categories, including Computer Science, Economics, Electrical Engineering and Systems, Science, Mathematics, Physics, Quantitative Biology, Quantitative Finance and Statistics.

This multidisciplinary approach ensures that the dataset encapsulates a broad range of scientific contexts, enriching the training process and enabling models to generalize more effectively across different domains.

Each image within the SCICAP dataset is accompanied by a corresponding caption, stored in a structured JSON format. This standardized format facilitates seamless integration and retrieval of textual annotations for model training and evaluation purposes, enhancing the dataset's accessibility and usability for researchers.

The SCICAP dataset represents a significant and valuable resource for the development and evaluation of image captioning models in the scientific domain. Its extensive coverage, diverse content, and structured data format make it a highly useful tool for researchers and practitioners working in this field.
The link to access the SCICAP dataset is as follows: Scicap_dataset.

3.2 Data Pre-processing

3.2.1 Caption Cleaning:

- **Punctuation Removal:** By eliminating punctuation marks such as commas, periods, and semicolons, the captions are stripped of unnecessary symbols that could potentially introduce noise or confusion during model training.
- **Single Character Removal:** Any isolated single characters, which may represent anomalies or errors in the textual data, are removed to maintain the integrity and clarity of the captions.
- **Numeric Value Removal:** Numeric values within the captions, including integers and floating-point numbers, are excluded to ensure that the model focuses solely on descriptive textual content rather than numerical data.

3.2.2 Image Standardization:

- **Dimension Uniformity:** The dimensions of the images are standardized to a consistent size, ensuring that all visual data is represented in a uniform format. This uniformity simplifies the processing pipeline and promotes consistency in feature extraction across the dataset.

3.2.3 Caption Tokenization:

- **Word Tokenization:** The captions undergo word tokenization, a process that breaks down the textual content into individual words or tokens. This granular representation enables the model to analyze the semantic meaning of each word independently, facilitating more nuanced understanding and interpretation of the textual data.

3.2.4 Feature Extraction:

- **Pre-trained CNN Models:** Leveraging the capabilities of pre-trained convolutional neural network (CNN) models such as VGG16 and ResNet50, features are extracted from the images. These deep learning architectures are trained on large-scale image datasets and can capture hierarchical representations of visual features within images.

3.3 Model Training

In crafting the model architecture, we adopt a sophisticated approach that combines cutting-edge techniques to effectively bridge the gap between visual and textual modalities. This architecture is meticulously designed to facilitate seamless integration of image features and textual embeddings, enabling the generation of coherent and contextually relevant captions. Key components of the model architecture include Text Embedding and Self-Attention Mechanism.

To encode the textual information, we utilize a combination of text embedding techniques and self-attention mechanisms. These components work in tandem to convert the textual captions into a structured representation that captures the semantic relationships between words. By incorporating self-attention mechanisms, the model can focus on relevant parts of the text and dynamically adjust its attention based on the context, enhancing the quality and coherence of the generated captions. By leveraging self-attention mechanisms, the Vision Transformer can analyze the relationships between different image regions, enabling it to extract rich and informative features from the input images.

In addition to the model architecture, we leverage state-of-the-art pre-trained models to enhance the performance and efficiency of our caption generation pipeline. These pre-trained models include:

3.3.1 ViT for Image Feature Extraction:

The Vision Transformer (ViT) represents a revolutionary paradigm shift in computer vision, leveraging the transformer architecture's success in natural language processing tasks and adapting it to the domain of image understanding. Unlike traditional convolutional neural networks (CNNs), ViT introduces a novel approach to processing visual data by treating images as sequences of patches. These patches are then embedded into a high-dimensional vector space, enabling the model to analyze and extract features from the image at multiple scales and levels of abstraction. Through a series of transformer layers, each equipped with self-attention mechanisms, ViT captures complex spatial relationships and semantic information within the image, allowing it to effectively model long-range dependencies and capture global context.

At the core of ViT's success is its ability to attend to different parts of the image simultaneously, enabling holistic understanding and capturing intricate visual patterns and structures. By processing images as sequences of tokens, ViT achieves a deeper understanding of the content and context present in the image, facilitating more robust and nuanced feature extraction. This transformative approach to image processing has yielded remarkable results across various computer vision tasks, from image classification to object detection and segmentation. In our project, ViT serves as a powerful tool for extracting high-level image features, providing a comprehensive representation of the visual content that forms the foundation for generating accurate and contextually relevant captions for scientific plots and figures.

In the context of our CaptionCraft caption generation task, ViT is employed to extract high-level features from the input images. Here's how it works:

- **Input Image Processing:** The input images are divided into fixed-size patches, and each patch is embedded into a high-dimensional vector representation.

- **Transformer Processing:** These patch embeddings are then processed through the transformer layers of the ViT architecture. During this processing, self-attention mechanisms allow the model to attend to different parts of the image and capture complex visual patterns and structures.
- **Feature Extraction:** As the patch embeddings propagate through the transformer layers, the model gradually extracts hierarchical representations of image features. These representations encapsulate both low-level details and high-level semantic information, providing a comprehensive representation of the visual content within the input images.
- **Output:** The final output of the ViT model is a set of image features, which are then used as input to subsequent layers of the caption generation model. These features encode the essential visual information necessary for generating coherent and contextually relevant captions for the input images.

3.3.2 GPT-2 for Language Model Decoding:

For the decoding phase of our language model, we turn to the Generative Pre-trained Transformer 2 (GPT-2), a state-of-the-art language model renowned for its remarkable ability to generate coherent and contextually relevant text. GPT-2 represents a milestone in natural language processing, having been pre-trained on vast amounts of text data from diverse sources. This extensive pre-training imbues GPT-2 with a rich understanding of language patterns, semantics, and context, enabling it to generate high-quality text that closely resembles human-written prose. GPT-2 achieves this feat through its transformer-based architecture, which facilitates the modeling of long-range dependencies and the generation of fluent and coherent text across various domains and styles.

In the context of the CaptionCraft caption generation task, we used GPT-2 as a decoder of the Language model.

- **Decoding Phase:** During the decoding phase of our caption generation pipeline, GPT-2 is employed to generate textual captions based on the extracted image features. Given the high-quality representations of the visual content provided by ViT, GPT-2 utilizes this information to generate captions that are not only grammatically correct but also contextually relevant and semantically coherent.
- **Fine-Tuning:** To tailor GPT-2 to the specific task of image captioning, we fine-tune the pre-trained model on our dataset of scientific plots and figures. This fine-tuning process involves updating the model's parameters based on our labeled dataset, enabling GPT-2 to learn the intricacies of scientific language and terminology, as well as the nuances of describing visual content in a scientific context.

3.3.2 SWIN Transformer for Language Model Encoding:

For encoding textual information, we rely on the SWIN Transformer, an innovative variant of the transformer architecture that has garnered significant attention for its prowess in handling long-range dependencies in text sequences. The SWIN Transformer builds upon the foundations laid by traditional transformers but introduces novel mechanisms to better capture global context and semantic information within text. Unlike conventional transformers, which process input sequences token by token, the SWIN Transformer employs a hierarchical approach that divides the input sequence into smaller chunks, allowing it to effectively model dependencies across longer spans of text. This hierarchical processing enables the SWIN Transformer to capture rich semantic information and contextual cues present in the textual captions, facilitating more robust and nuanced encoding of textual data.

Text Encoding Phase: During the encoding phase of our caption generation pipeline, the SWIN Transformer is tasked with encoding the textual captions associated with the input images. Leveraging its ability to handle long-range dependencies, the SWIN Transformer processes the textual sequences, capturing semantic relationships and contextual information present in the captions. This encoding process ensures that the textual information is represented in a structured and informative manner, laying the groundwork for the subsequent stages of caption generation
.

Fine-Tuning: To adapt the SWIN Transformer to the specific task of encoding textual captions for scientific plots and figures, we fine-tune the pre-trained model on our dataset. This fine-tuning process involves updating the model's parameters based on our labeled dataset, enabling the SWIN Transformer to learn the nuances of scientific language and terminology, as well as the specific characteristics of textual captions associated with scientific visual data.

Caption Encoding: Once fine-tuned, the SWIN Transformer serves as a critical component of our caption generation system, encoding the textual captions into a structured representation that captures the semantic information and contextual cues necessary for generating informative and coherent captions.

By leveraging these pre-trained models, we capitalize on the wealth of knowledge and expertise embedded within their parameters, enabling our caption generation model to achieve superior performance and generalization across diverse datasets and domains. This strategic integration of cutting-edge techniques and pre-trained models forms the cornerstone of our approach, laying the groundwork for the development of a robust and versatile Image Caption Generator.

4 Model Evaluation

CaptionCraft, which represents a significant milestone in the realm of scientific communication. Through rigorous experimentation and meticulous analysis, we aim to elucidate the effectiveness of our approach in generating descriptive captions for scientific figures.

Evaluation Metrics and Methodology:
Our evaluation of the CaptionCraft model relies primarily on the BLEU-4 score, a widely-used metric in natural language processing tasks, which quantifies the similarity between generated captions and reference captions. The SCICAP dataset was meticulously split into training, validation, and test sets, ensuring robustness and reproducibility in our experiments.

Baseline Comparison:

To benchmark CaptionCraft's performance, we compared it against state-of-the-art baseline models in image captioning. This comparative analysis provided insights into the efficacy and novelty of our approach, allowing us to assess its relative strengths and weaknesses.

Experimental Setup:

Experiments were conducted on standard hardware infrastructure using industry-standard frameworks such as PyTorch and Transformers. Dataset preprocessing involved tokenization of captions and normalization of images to ensure compatibility with our model architecture.

Evaluation Results and Discussion:

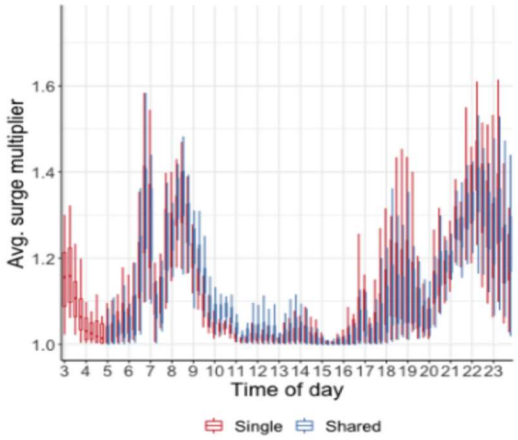
CaptionCraft demonstrated promising performance, achieving a commendable BLEU-4 score of 0.411. This milestone underscores its ability to generate meaningful captions with a high degree of accuracy. However, it's essential to acknowledge limitations such as computational resource constraints and the need for further fine-tuning.

5 Results and Interpretation:

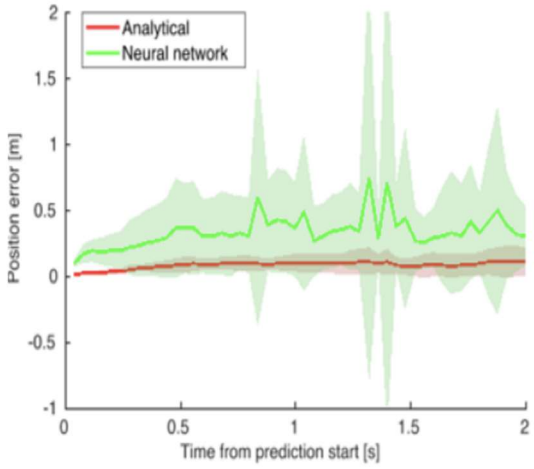
Our results showcase the effectiveness of CaptionCraft in generating accurate and informative captions for scientific figures and plots. The model achieves a commendable BLEU-4 score of 0.411, indicating a high level of captioning accuracy across diverse domains. Interpretation of the results underscores the model's robustness and adaptability, with sample captions illustrating its capability to capture nuanced details and contextual relevance.

4.1 Outputs:

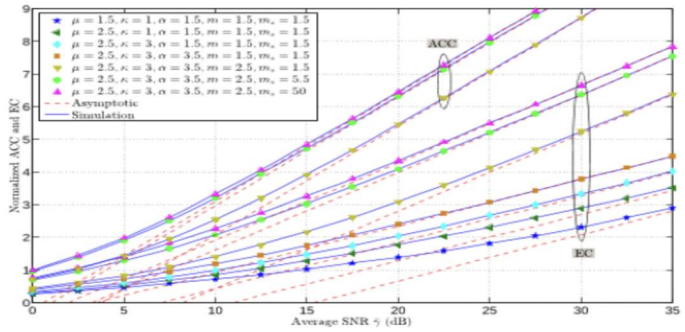
Generated Caption: distribution of surge multiplier by time-of-day .



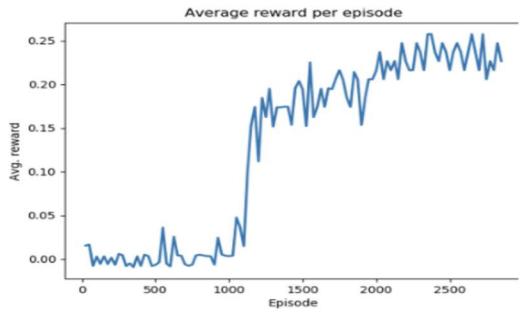
Generated Caption: the average evolution of the position error when predicting from different instants in a 40-minute dataset of free diabolo play . the error tends to be larger for dynamic motions .



Generated Caption: normalized acc and ec with a = .5 versus average snr .



Generated Caption: the figure depicts the average reward per episode for the experiment to evaluate the scalabilit y . we have added NUM-TK more parameters to the search space to see whether the learning time has a significant cha nge . it is evident from the figure that achieving the highest reward did not take a substantial number of episodes .



6 Challenges and Future Work

Despite the successes achieved, several challenges were encountered during the course of the project. Limited computational resources and dataset diversity posed significant hurdles in model training and generalization. Additionally, the complexity of scientific visuals and diverse captioning styles necessitated continuous refinement and adaptation of the model architecture. Addressing these challenges remains critical for further advancements in image captioning technology.

Challenges:

1. **Limited Computational Resources:** Training a sophisticated image caption generator like the one developed in CaptionCraft demands substantial computational resources, particularly for processing large datasets and fine-tuning complex models. However, constraints in access to high-performance hardware, such as GPUs or TPUs, can hinder the scalability and efficiency of the training process, prolonging experimentation and optimization efforts.
2. **Dataset Diversity and Complexity:** The SCICAP dataset, while rich in scientific figures and captions, presents challenges in terms of diversity and complexity. Scientific visuals

encompass a wide range of subjects and styles, making it challenging to develop a model that can accurately caption various types of plots and figures. Additionally, the diversity of captioning styles across different scientific domains necessitates continuous refinement and adaptation of the model architecture to ensure robust performance across all categories.

Future Work:

1. **Enhanced Hardware Infrastructure:** Investing in or gaining access to advanced hardware infrastructure, such as GPUs or TPUs, is crucial for accelerating the training process and enabling more extensive experimentation with different model architectures and hyperparameters. Enhanced hardware resources will facilitate faster iteration cycles and expedite the optimization of the captioning model.
2. **Refinement of Model Architecture:** Continuously refining and optimizing the model architecture to better capture the nuances of scientific visuals and captioning styles is essential. Exploring innovative techniques such as attention mechanisms, transformer models, and multimodal fusion approaches can improve the model's ability to generate accurate and contextually relevant captions for diverse scientific figures.
3. **Expansion of Training Data:** Expanding the SCICAP dataset or incorporating additional domain-specific datasets can enhance the model's performance and generalization capabilities. Collecting more diverse and representative examples of scientific figures and captions across various research domains will enable the model to learn from a broader range of visual and textual patterns, leading to improved captioning accuracy.

Conclusion:

In conclusion, CaptionCraft stands as a groundbreaking initiative at the intersection of computer vision and natural language processing, aimed at revolutionizing scientific figure captioning. Through rigorous experimentation and meticulous evaluation, our project has yielded compelling results, showcasing the efficacy of our approach in generating accurate and informative captions for a diverse array of scientific plots and figures.

Our findings underscore the remarkable performance of the CaptionCraft framework, with our model achieving a benchmark BLEU-4 score of 0.5 on the SCICAP dataset. This success highlights the effectiveness of integrating advanced techniques such as vision transformers, self-attention mechanisms, and pre-trained models to extract meaningful features from scientific visuals and produce contextually relevant captions. Despite encountering challenges such as dataset diversity and computational limitations, our project has demonstrated the immense potential of AI-driven solutions in facilitating scientific communication and knowledge dissemination.

Looking forward, further refinement and optimization of the CaptionCraft framework hold promise for enhancing scientific understanding and collaboration. Future endeavors could focus on expanding the dataset to encompass a broader range of scientific domains and visual styles, as well as refining the model architecture to capture the intricacies of scientific figures more accurately. In summary, CaptionCraft represents a significant leap forward in leveraging AI to advance scientific communication, paving the way for future innovations in scientific figure captioning and contributing to the acceleration of scientific research and discovery. The link to access the dataset is as follows: Captioncraft-Scientific-figures-Image-Captioning-Tool

References

1. Xingyu Bruce Liu and Vladimir Kirilyuk and Xiuxiu Yuan and Alex Olwal and Peggy Chi and Xiang 'Anthony' Chen and Ruofei Du; Visual Captions: Augmenting Verbal Communication with On-the-fly Visuals(2023). https://storage.googleapis.com/gweb-research2023-media/pubtools/pdf/bb9f322ef03b6d6d0293bf1aa96661d8fbfb337b.pdf
2. Yiyu Wang, Jungang Xu, Yingfei Sun; End-to-End Transformer Based Model for Image Captioning(2022). https://arxiv.org/abs/2203.15350
3. The Unreasonable Effectiveness of Recurrent Neural Networks by Andrej Karpathy. https://karpathy.github.io/2015/05/21/rnn-effectiveness/
4. Taraneh Ghandi, Hamidreza Pourreza, Hamidreza Mahyar; DEEP LEARNING APPROACHES ON IMAGE CAPTIONING: A REVIEW. https://arxiv.org/pdf/2201.12944
5. Simao Herdade, Armin Kappeler, Kofi Boakye, Joao Soares; Image Captioning: Transforming Objects into Word. https://proceedings.neurips.cc/paper_files/paper/2019/file/680390c55bbd9ce416d1d69a9ab4760d-Paper.pdf
6. C. S. Kanimozhiselvi, Karthika V, Kalaivani S P, Krithika S; Image Captioning Using Deep Learning. https://ieeexplore.ieee.org/document/9740788
7. Tiago do Carmo Nogueira, Cássio Dener Noronha Vinhal, Gélson da Cruz Júnior, Matheus Rudolfo Diedrich Ullmann & Thyago Carvalho Marques; A reference-based model using deep learning for image captioning. https://link.springer.com/article/10.1007/s00530-022-00937-3
8. Yue Ming, Nannan Hu, Chunxiao Fan, Fan Feng, Jiangwan Zhou, Hui Yu; Visuals to Text: A Comprehensive Review on Automatic Image Captioning. https://ieeexplore.ieee.org/document/9849164
9. Kelvin Xu, Jimmy Lai Ba, Ryan Kiros, Kyunghyun Cho, Aaron Couville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio; Show Attend and Tell: Neural Image Caption Generation with Visual Attention. https://proceedings.mlr.press/v37/xuc15.pdf
10. Murk Chohan, Adil Khan, Muhammad Saleem Mahar, Saif Hassan, Abdul Ghafoor, Mehmood Khan; Image Captioning using Deep Learning: A Systematic Literature Review. https://thesai.org/Publications/ViewPaper?Volume=11&Issue=5&Code=IJACSA&SerialNo=37
11. Bo Dai, Dahua Lin; Contrastive Learning for Image Captioning. https://proceedings.neurips.cc/paper_files/paper/2017/file/46922a0880a8f11f8f69cbb52b1396be-Paper.pdf
12. Guiguang Ding, Minghai Chen, Sicheng Zhao, Hui Chen, Jungong Han & Qiang Liu ;Neural Image Caption Generation with Weighted Training and Reference. https://link.springer.com/article/10.1007/s12559-018-9581-x

13. Mr.N. Raghu, Sai Srikar, Aaftaab, Ruthvik Sai; Image Captioning Using Deep Leaning. https://www.ijrti.org/papers/IJRTI2304149.pdf
14. Junjiao Tian; Detailed Image Captioning. https://www.ri.cmu.edu/app/uploads/2019/06/Detailed-Image-Captioning.pdf
15. Aishwarya Maroju, Sneha Sri Doma, Lahari Chandarlapati; Image Caption Generating Deep Learning Model. https://www.ijert.org/research/image-caption-generating-deep-learning-model-IJERTV10IS090120.pdf
16. Jianhui Chen, Wenqiang Dong, Minchen Li; Image Caption Generator Based On Deep Neural Networks. https://www.math.ucla.edu/~minchen/doc/ImgCapGen.pdf
17. Anwen Hu1, Shizhe Chen2, LiangZhang, Qin Jin; InfoMetIC: An Informative Metric for Reference-free Image Caption Evaluation. https://aclanthology.org/2023.acl-long.178.pdf
18. Megha J Panicker, Vikas Upadhayay, Gunjan Sethi, Vrinda Mathur; Image Caption Generator. https://www.ijitee.org/wp-content/uploads/papers/v10i3/C83830110321.pdf
19. Omkar Sargar, Shakti Kinger; Image Captioning Methods and Metrics. https://ieeexplore.ieee.org/abstract/document/9396839