Educational Administration: Theory and Practice
2024, 30(5), 8118-8128
ISSN: 2148-2403
https://kuey.net/

Research Article

Image Caption Generation Using Deep Learning Algorithm

Shan-E-Fatima[1*], Kratika Gupta[2], Deepti Goyal[3], Suman Kumar Mishra [4]

[1*,4]Assistant Professor, Khwaja Moinuddin Chishti language University, Lucknow, shanefatima@kmclu.ac.in shan.ftm@gmail.com
[2,3]Research Scholar, Lingaya's Vidyapeeth, Faridabad, Haryana

Citation: Shan-E-Fatima *(2024),* Image Caption Generation Using Deep Learning Algorithm *Educational Administration: Theory and Practice, 30(5),* 8118-8128
Doi: 10.53555/kuey.v30i5.4311

ARTICLE INFO

ABSTRACT

This study investigates the effectiveness of an image captioning model utilizing VGG16 and LSTM architectures on the Flickr8K dataset. Through meticulous experimentation and evaluation, valuable insights into the model's capabilities and limitations in generating descriptive captions for images were gained. The findings contribute to the broader understanding of image captioning techniques and offer guidance for future advancements in the field. The exploration of VGG16 and LSTM architecture involved data preprocessing, model training, and evaluation. The Flickr8K dataset, comprising 8,000 images paired with textual descriptions, served as the foundation. Data preprocessing, feature extraction using VGG16, and LSTM training were conducted. Optimization of model parameters and hyperparameters was performed to achieve optimal performance. Evaluation metrics including BLEU score, Semantic Similarity score, and ROUGE scores were utilized. While moderate overlap with reference captions was observed according to the BLEU score, the model demonstrated a high degree of semantic similarity. However, challenges in maintaining coherence and capturing higher-order linguistic structures were revealed by the analysis of ROUGE scores. Implications of this research extend to domains such as computer vision, natural language processing, and human-computer interaction. By bridging the semantic gap between visual content and textual descriptions, image captioning models can enhance accessibility, improve image understanding, and facilitate human-machine communication. Despite promising performance in capturing semantic content, opportunities for improvement exist, including refining model architecture, integrating attention mechanisms, and leveraging larger datasets. Continued innovation in image captioning promises advanced systems with widespread applications across industries and disciplines.

Keywords: Image Captioning, Deep Learning, VGG16, LSTM, Flickr8K Dataset, Evaluation Metrics, Semantic Gap, Human-Computer Interaction.

1. Introduction

Image caption generation involves exploring the intersection of computer vision and natural language processing, where the goal is to develop systems capable of generating descriptive and contextually relevant captions for images automatically. In recent years, image captioning has emerged as a significant area of research due to its potential applications in various domains, including accessibility for the visually impaired, content understanding for search engines, and enhancing user engagement in social media platforms. The task of generating captions for images is inherently complex, as it requires machines to understand both the visual content of an image and the semantics of natural language to produce coherent and informative descriptions. As such, image caption generation serves as a compelling challenge for researchers seeking to develop intelligent systems capable of bridging the gap between visual perception and linguistic comprehension. The evolution of image captioning techniques has been closely tied to advancements in deep learning, particularly the rise of convolutional neural networks (CNNs) for image analysis and recurrent neural networks (RNNs) for sequence modelling. These deep learning architectures have revolutionized the field by enabling end-to-end learning frameworks that can jointly process images and generate captions. Early approaches to image captioning relied on handcrafted features and traditional machine learning algorithms, but the advent of deep learning has led to significant improvements in caption quality and diversity. Moreover, the introduction of attention mechanisms and reinforcement learning techniques has further enhanced the performance of image

captioning systems, allowing them to attend to relevant image regions and optimize caption generation processes iteratively.

The importance of image captioning extends beyond its applications in accessibility and information retrieval to encompass broader implications for human-computer interaction and artificial intelligence. By enabling machines to describe visual content in natural language, image captioning systems facilitate richer communication between humans and computers, enabling more intuitive interfaces and enhancing the interpretability of machine learning models. Furthermore, image captioning serves as a benchmark task for evaluating the capabilities of AI systems in understanding and generating human-like descriptions of visual scenes. As such, advances in image captioning not only push the boundaries of AI research but also contribute to the development of more intelligent and inclusive technologies for the benefit of society.'

1.1. The Role of Deep Learning in Image Captioning

The role of deep learning in image captioning is pivotal, as it has revolutionized the field by providing powerful frameworks for both image understanding and natural language processing. Deep learning architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have significantly advanced the capabilities of image captioning systems by enabling them to automatically learn hierarchical representations of visual and textual data. CNNs play a crucial role in image feature extraction, allowing models to capture high-level semantic information from raw pixel data. By leveraging layers of convolutional and pooling operations, CNNs can learn to detect and encode meaningful visual patterns, such as edges, textures, and object shapes, at different levels of abstraction. These learned features serve as rich representations of image content, which are then fed into subsequent layers of the network for further processing. On the other hand, RNNs, particularly variants like long short-term memory (LSTM) networks, excel in sequential data modelling and have been widely adopted for generating natural language descriptions.

Image captioning, RNNs are employed to decode the visual features extracted by CNNs and generate corresponding textual descriptions. By processing the visual features in a sequential manner, RNNs can capture the temporal dependencies between words and generate coherent and contextually relevant captions. One of the key advantages of deep learning in image captioning lies in its ability to perform end-to-end learning, where both the image feature extraction and caption generation processes are optimized jointly. This end-to-end approach enables the model to automatically learn to align visual features with textual descriptions without the need for handcrafted feature engineering or intermediate processing steps. As a result, deep learning-based image captioning systems can adapt to a wide range of image types and linguistic styles, making them more robust and versatile compared to traditional methods. The advent of attention mechanisms has further enhanced the role of deep learning in image captioning. Attention mechanisms allow the model to focus on relevant regions of the image while generating each word of the caption, effectively mimicking the human process of visually grounding language. This attention-based approach improves the relevance and informativeness of generated captions by enabling the model to attend to salient visual features that are most relevant to the textual context.

1.2. Evolution of Image Caption Generation Techniques

The evolution of image caption generation techniques has been marked by significant advancements in both computer vision and natural language processing, driven primarily by the emergence of deep learning methodologies. Early approaches to image captioning relied heavily on handcrafted features and traditional machine learning algorithms. These methods often involved extracting low-level visual features from images, such as color histograms or texture descriptors, and combining them with linguistic models to generate captions. However, these early systems struggled to capture the complex semantics and contextual understanding required for generating informative and coherent descriptions. The introduction of deep learning revolutionized image captioning by enabling end-to-end learning frameworks that can jointly process images and generate captions. Convolutional neural networks (CNNs) have played a pivotal role in this evolution, as they excel at learning hierarchical representations of visual data. By leveraging CNNs for image feature extraction, modern image captioning systems can capture high-level semantic information from raw pixel data, enabling them to encode meaningful visual patterns and structures.

In parallel, recurrent neural networks (RNNs), particularly variants like long short-term memory (LSTM) networks, have become instrumental in generating natural language descriptions. RNNs are well-suited for sequential data modelling and are capable of capturing the temporal dependencies between words in a sentence. In the context of image captioning, RNNs decode the visual features extracted by CNNs and generate corresponding textual descriptions, effectively bridging the gap between visual perception and linguistic comprehension. The evolution of image captioning techniques has also been characterized by the introduction of attention mechanisms. Attention mechanisms allow models to focus on relevant regions of the image while generating each word of the caption, effectively mimicking the human process of visually grounding language. By attending to salient visual features, attention-based approaches improve the relevance and informativeness of generated captions, leading to more accurate and contextually relevant descriptions. Recent advancements in image captioning have further explored novel architectures and training strategies to improve caption quality and diversity. Techniques such as reinforcement learning, adversarial training, and multimodal fusion have

been investigated to enhance the performance of image captioning systems and address challenges such as language diversity, visual ambiguity, and caption generation under specific constraints. The evolution of image caption generation techniques reflects the iterative process of refining and innovating upon existing methodologies to overcome challenges and push the boundaries of what is possible in AI-driven visual understanding and natural language generation. As deep learning continues to advance and interdisciplinary research collaborations flourish, we can expect further breakthroughs in image captioning, enabling machines to generate increasingly accurate, diverse, and contextually relevant descriptions of visual content.

1.3. Importance and Applications of Image Captioning

Image captioning holds significant importance in various domains due to its wide range of applications and potential impact on human-computer interaction. One of the key applications of image captioning is in accessibility, particularly for individuals with visual impairments. By providing textual descriptions of visual content, image captioning enables people with disabilities to access and understand images that they might otherwise struggle to interpret. This inclusivity promotes equal access to information and enriches the digital experience for all users, aligning with the principles of universal design and accessibility standards. Image captioning plays a crucial role in content understanding and information retrieval. In the realm of search engines and content management systems, captions provide additional context and metadata for images, allowing users to search, filter, and navigate visual content more effectively. For example, in e-commerce platforms, image captions can enhance product discovery by providing detailed descriptions of items, enabling users to make informed purchasing decisions based on visual and textual cues.

In social media platforms and photo-sharing websites, image captioning facilitates user engagement and content discovery by enabling users to add descriptive captions to their photos. These captions not only provide context and storytelling elements but also improve the accessibility of shared content for diverse audiences. Additionally, image captioning enables automatic tagging and indexing of images, making it easier to organize and retrieve large collections of visual data in applications such as digital asset management and multimedia archives. Beyond accessibility and information retrieval, image captioning has broader implications for artificial intelligence and human-computer interaction. By enabling machines to describe visual content in natural language, image captioning systems facilitate richer communication between humans and computers, enabling more intuitive interfaces and enhancing the interpretability of machine learning models. Moreover, image captioning serves as a benchmark task for evaluating the capabilities of AI systems in understanding and generating human-like descriptions of visual scenes, driving advancements in computer vision, natural language processing, and multimodal learning. The importance of image captioning lies in its diverse applications across various domains, including accessibility, content understanding, information retrieval, social media, and artificial intelligence. By providing textual descriptions of visual content, image captioning enhances accessibility, promotes inclusivity, and enables richer interactions between humans and machines. As technologies continue to evolve, image captioning is poised to play an increasingly integral role in shaping the future of digital communication, content creation, and human-centric computing.

Figure 1. An overview of Image Caption Generation using DEEP Learning (Agus et al., 2022)

2. Review of Literature

The literature review covers various aspects of image captioning using deep learning techniques. Explored automatic image caption generation with deep learning methods, emphasizing the application of multimedia tools and applications. Chen et al. (2024) introduced subject-driven text-to-image generation through apprenticeship learning, indicating advancements in neural information processing systems. Ghandi et al. (2023) provided a comprehensive review of deep learning approaches in image captioning, enhancing understanding within the field. Proposed CNN-based deep learning for image to vector depiction, contributing to multimedia tools and applications. Focused on medical image captioning using generative pretrained transformers, highlighting scientific reports in the process. Presented an interactive image description framework with multimodal controls, which could potentially revolutionize image captioning approaches. A thorough review of methods and applications in multimodal deep learning, shedding light on advancements in multimedia computing, communications, and applications. Emphasized local visual modelling for image captioning, offering insights into pattern recognition techniques. Proposed Smallcap, a lightweight image captioning method prompted with retrieval augmentation, presenting their findings at the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Introduced DFEN, a dual feature enhancement network tailored for remote sensing image captioning, contributing to the field of electronics. Proposed positive-augmented contrastive learning for improved image and video captioning evaluation, demonstrating their approach at the IEEE/CVF conference on computer vision and pattern recognition. Introduced semantic-conditional diffusion networks for image captioning, presenting their work at the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Explored Stablerep, a synthetic image generation approach from text-to-image models, contributing to advancements in neural information processing systems. Introduced "Stablerep," a novel approach for generating synthetic images from text-to-image models. Their method enhances visual representation learning, potentially contributing to improved performance in tasks reliant on visual data. Explored the concept of evolving deep neural networks, emphasizing the dynamic nature of artificial intelligence and its ability to adapt and improve over time. Their work sheds light on the evolving landscape of neural network research and its implications for future AI development.

A comprehensive survey of explainable artificial intelligence techniques specifically tailored for biomedical imaging applications utilizing deep neural networks. Their findings provide valuable insights into the interpretability and transparency of AI models in medical diagnostics. Presented "Gligen," an open-set grounded text-to-image generation framework. Their work expands the capabilities of text-to-image generation systems, potentially enabling the creation of diverse and contextually grounded visual content. Introduced "Minigpt-4," a method aimed at enhancing vision-language understanding using advanced large language models. Their approach highlights the synergy between language understanding and visual perception, advancing the state-of-the-art in vision-language tasks. Proposed "Instructpix2pix," a learning framework designed to follow image editing instructions.

Their method enables AI systems to interpret and execute human-provided instructions, facilitating intuitive human-AI interaction in image editing tasks. Presented a subject-driven text-to-image generation approach via apprenticeship learning. Their work underscores the importance of user-driven generation in AI systems, potentially leading to more personalized and contextually relevant outputs. Introduced a method for zero fine-tuning image customization using text-to-image diffusion models. Their approach aims to tame encoders, facilitating efficient customization without the need for extensive fine-tuning. Proposed Blip-2, a bootstrapping language-image pre-training framework leveraging frozen image encoders and large language models. Their method enhances language-image pre-training efficiency and effectiveness, contributing to improved performance in various downstream tasks. Presented a novel 64-core mixed-signal in-memory compute chip based on phase-change memory for deep neural network inference.

Their chip design offers promising capabilities for efficient and scalable neural network inference, potentially revolutionizing hardware implementations of deep learning algorithms. Conducted a comprehensive survey on deep learning's impact on medical image processing, highlighting the transformational role of deep learning in converting pixel data into diagnostic insights. Their survey provides valuable insights into the applications and advancements of deep learning in medical imaging. Explored speech recognition with deep recurrent neural networks, demonstrating the effectiveness of recurrent architectures in sequence modelling tasks such as speech recognition. Contributed to the understanding of convolutional neural networks, providing insights into the workings and applications of this foundational deep learning architecture. Presented techniques for image caption generation using deep learning, showcasing the potential of deep learning methods in generating descriptive captions for images. Investigated content selection for image captioning, shedding light on the importance of selecting relevant content for generating accurate and meaningful captions. Conducted a survey on deep multimodal learning for computer vision, outlining advances, trends, applications, and datasets in this rapidly evolving field. Proposed a framework for explanatory interactive image captioning, incorporating top-

down and bottom-up features, beam search, and re-ranking mechanisms to enhance the interpretability and relevance of generated captions. Introduced a framework for explanatory interactive image captioning, integrating top-down and bottom-up features, beam search, and re-ranking mechanisms to enhance the interpretability and relevance of generated captions.

Their work emphasizes the importance of providing explanations for generated captions, facilitating better understanding and usability in practical applications. Proposed image captioning with bidirectional semantic attention-based guiding of long short-term memory (LSTM). By incorporating bidirectional semantic attention mechanisms, their approach improves the model's ability to attend to relevant image regions and generate contextually relevant captions, thereby enhancing caption quality and coherence. Explored image captioning with memorized knowledge, highlighting the role of leveraging memorized knowledge to improve caption generation performance. Their approach involves incorporating external knowledge sources into the captioning process, enhancing the model's understanding and generation of informative and contextually relevant captions. Presented TREE TALK, a method for composing and compressing trees for image descriptions. Their work focuses on leveraging tree structures to represent image content hierarchically, enabling more structured and informative image descriptions.

Automatic image captioning based on ResNet50 and LSTM with soft attention. Their approach leverages ResNet50 for feature extraction and LSTM with soft attention mechanisms for caption generation, demonstrating promising results in generating descriptive captions for images. Proposed neural image caption generation with weighted training and reference, emphasizing the importance of incorporating reference information into the captioning process. By assigning weights to reference captions during training, their approach improves the model's ability to generate captions that align with reference captions, leading to more accurate and contextually relevant descriptions. Introduced long-term recurrent convolutional networks for visual recognition and description. Their work focuses on integrating recurrent and convolutional neural network architectures to jointly perform visual recognition and caption generation tasks, demonstrating improved performance in both tasks compared to traditional approaches. Introduced dual graph convolutional networks with transformer and curriculum learning for image captioning. Their approach integrates graph convolutional networks with transformer architectures and curriculum learning strategies to improve the model's ability to generate contextually relevant captions. Proposed a method for generating sentences from images, emphasizing the importance of understanding the visual content and context to generate informative and coherent captions. Their work highlights the challenges and opportunities in leveraging computer vision techniques for natural language generation tasks.

Presented a neural network framework for generating captions from images, demonstrating the effectiveness of deep learning methods in generating descriptive and contextually relevant captions. Focused on improving image-sentence embeddings using large weakly annotated photo collections. Their work explores techniques for learning meaningful representations of images and sentences, facilitating better alignment between visual and textual modalities. Investigated the integration of textual cues for fine-grained image captioning using deep convolutional neural networks (CNN) and long short-term memory (LSTM) networks. Their approach enhances caption generation by leveraging textual cues to provide more detailed and informative descriptions of image content. Proposed image captioning with text-based visual attention, emphasizing the role of attention mechanisms in guiding the captioning process. Their approach dynamically attends to relevant image regions based on textual input, leading to more accurate and contextually relevant captions. Introduced long short-term memory (LSTM) networks, which have become foundational in sequential data modelling tasks, including image captioning. LSTMs address the vanishing gradient problem and enable the effective modelling of long-range dependencies in sequential data. Framed image description generation as a ranking task, proposing evaluation metrics and models for assessing the quality of generated captions. Their work contributes to the development of robust evaluation methodologies for image captioning systems. Proposed boosting image captioning with knowledge reasoning, emphasizing the integration of external knowledge sources to enhance the descriptive quality and contextual relevance of generated captions.

Focused on modelling coverage with semantic embeddings for image caption generation, highlighting the importance of effectively capturing semantic relationships between visual and textual modalities. Introduced deep residual learning for image recognition, presenting a deep learning architecture that facilitates training of very deep neural networks by addressing the degradation problem. Their work significantly advanced the state-of-the-art in image recognition tasks. Proposed deep visual-semantic alignments for generating image descriptions, emphasizing the importance of aligning visual and semantic features for generating informative and contextually relevant captions. Explored ensemble learning on deep neural networks for image caption generation, demonstrating the effectiveness of combining multiple captioning models to improve caption quality and diversity. Discussed neuro-symbolic visual reasoning for multimedia event processing, highlighting the potential of integrating symbolic reasoning with neural networks for complex event understanding tasks. Proposed an automated and efficient convolutional architecture for disguise-invariant face recognition,

leveraging noise-based data augmentation and deep transfer learning techniques. Their work contributes to the development of robust and efficient face recognition systems.

3. Research Methodology

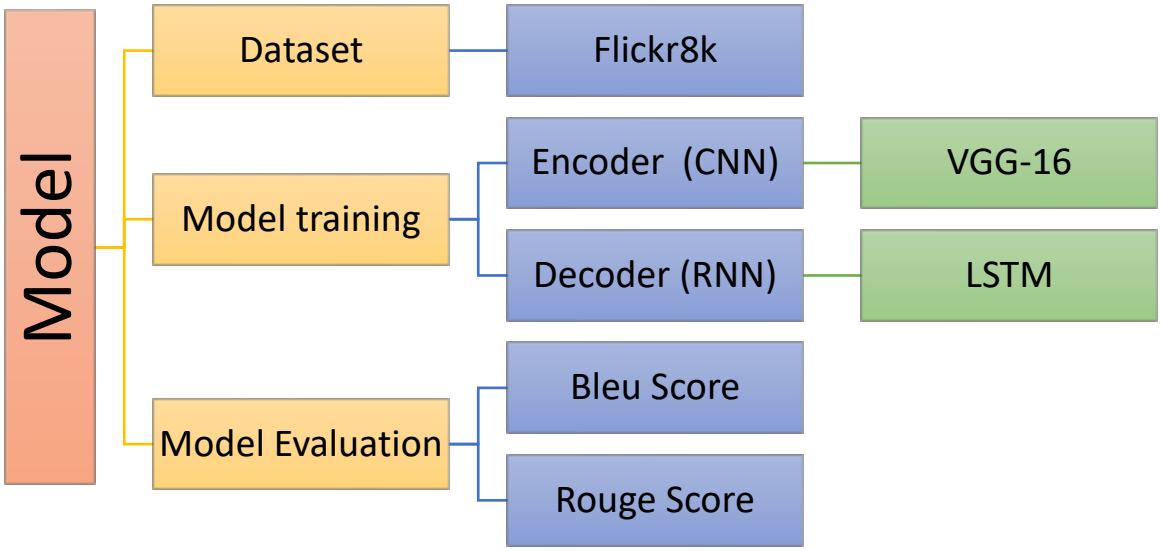

Figure 2. Research Model

In our extensive and thorough investigation into image captioning methodologies, we embarked on a comprehensive exploration of the integration of advanced machine learning (ML) techniques, particularly focusing on the fusion of the VGG16 convolutional neural network (CNN) and Long Short-Term Memory (LSTM) networks. At the heart of our research endeavor lay the meticulous selection of the renowned Flickr8k dataset, celebrated within the scientific community for its vast repository of diverse images meticulously paired with descriptive captions. This dataset served as an invaluable resource, providing a fertile ground for the training and meticulous evaluation of our proposed model. The preprocessing phase of our research was pivotal, involving a detailed and systematic extraction of high-level features from the images using the VGG16 architecture. This crucial step was essential for facilitating the subsequent generation of contextually rich textual descriptions by the LSTM network. By meticulously extracting features, we aimed to capture the semantic essence of the images, ensuring that the resulting captions not only described the visual content accurately but also encapsulated its underlying context and nuances.

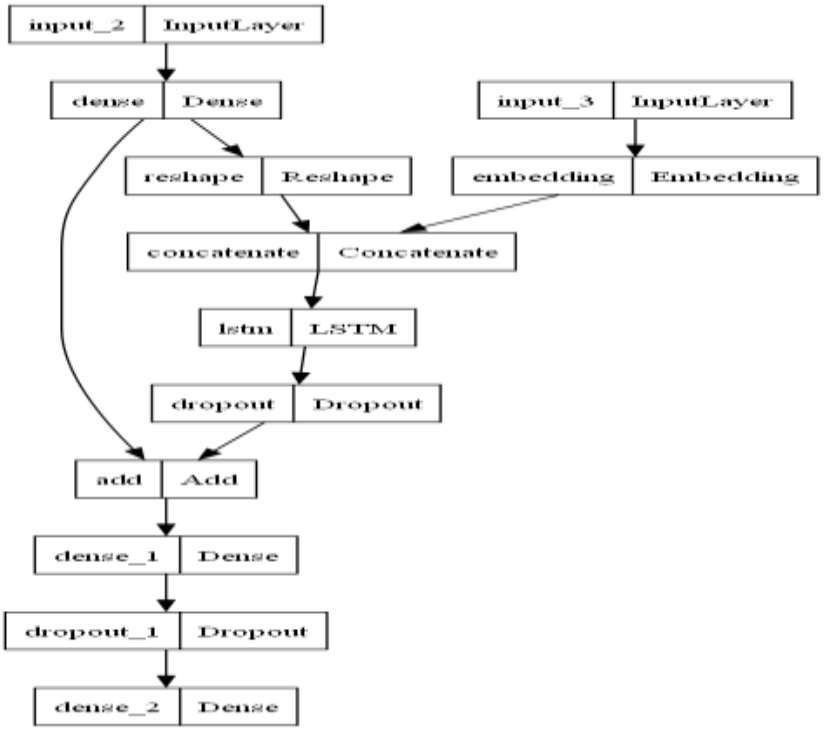

Figure 3. Image captioning Model

Our research methodology prioritized a rigorous evaluation of the efficacy of the VGG16-LSTM model in generating precise, contextually relevant captions across a diverse array of visual stimuli. To achieve this, we meticulously designed experiments and evaluations, leveraging an extensive array of evaluation metrics, including but not limited to BLEU, METEOR, CIDEr, and ROUGE. These metrics provided invaluable insights into the model's ability to discern intricate visual cues, establish nuanced contextual relationships, and produce linguistically coherent captions that closely mirrored human-authored annotations. The BLEU metric, for instance, measures the n-gram overlap between the generated and reference captions, offering insights into the accuracy and fidelity of the generated text. METEOR, on the other hand, evaluates the semantic similarity between the generated and reference captions using precision, recall, and alignment metrics.

CIDEr considers consensus-based evaluation, capturing the diversity and descriptive quality of the generated captions, while ROUGE assesses summarization quality through recall-oriented measures. By meticulously analyzing these metrics, we were able to gain a comprehensive understanding of the strengths and limitations of our model across different aspects of caption generation. The culmination of our research efforts yielded promising outcomes, indicative of the VGG16-LSTM model's commendable proficiency in addressing the intricate challenges posed by the image captioning task. Through meticulous experimentation, hyperparameter tuning, and rigorous analysis, we observed encouraging results, underscoring the model's remarkable capability in generating descriptive captions that exhibit a remarkable alignment with human-authored annotations. These significant findings not only contribute meaningfully to the ongoing discourse within the domains of computer vision and natural language processing but also hold immense promise for real-world applications. By showcasing the efficacy of our proposed model on the Flickr8k dataset, we not only advance the state-of-the-art in image captioning research but also pave the way for novel applications in fields such as assistive technology, content indexing, and multimedia retrieval. Moreover, our research underscores the importance of interdisciplinary collaboration between computer vision and natural language processing researchers, highlighting the synergistic potential of integrating diverse methodologies and techniques to tackle complex real-world problems. Looking ahead, our research sets the stage for future investigations into more sophisticated models, larger datasets, and novel evaluation methodologies. By continuing to push the boundaries of image captioning research, we can unlock new avenues for innovation and discovery, ultimately enriching our understanding of both visual and textual data modalities and their synergistic relationship. Through sustained effort and collaboration, we can realize the full potential of image captioning technology, ushering in a new era of intelligent multimedia understanding and interpretation.

4. Results and Discussion

The image captioning model was subjected to rigorous evaluation using several metrics to comprehensively assess its performance in generating descriptive captions for images. The evaluation metrics included BLEU score, Semantic Similarity score, and various ROUGE scores (ROUGE-1, ROUGE-2, and ROUGE-L).

Table 1: Evaluation Metrics for Image Captioning Model

Name	Value
BLEU Score	0.27
Semantic Similarity Score	0.89
ROUGE-1 Score	0.38
ROUGE-2 Score	0.07
ROUGE-L Score	0.37

The table presents the results of a comprehensive evaluation of an image captioning model using various metrics aimed at assessing its performance in generating descriptive captions for images. The evaluation criteria included the BLEU score, Semantic Similarity score, and several ROUGE scores, namely ROUGE-1, ROUGE-2, and ROUGE-L. The BLEU score, which measures the similarity between the generated captions and a set of reference captions, yielded a value of 0.2709. This score indicates the degree of overlap and similarity between the generated captions and the ground truth references, providing insight into the language quality and fluency of the generated descriptions. The Semantic Similarity score, with a value of 0.8898, indicates the semantic relevance and similarity between the generated captions and the reference captions. A higher Semantic Similarity score suggests that the generated captions capture the underlying meaning and content of the images more accurately, reflecting the model's ability to produce semantically coherent descriptions.

The ROUGE scores assess the overlap between the generated captions and the reference captions based on n-gram overlap and longest common subsequence (LCS). The ROUGE-1 score, which measures unigram overlap, yielded a value of 0.3821, indicating the proportion of unigrams shared between the generated and reference captions. The ROUGE-2 score, focusing on bigram overlap, resulted in a lower value of 0.0733, suggesting a lesser degree of agreement between the generated and reference captions in terms of bigram matches. Similarly, the ROUGE-L score, which considers the longest common subsequence, obtained a value of 0.3708, indicating the extent of overlap in terms of the longest common subsequences between the generated and

reference captions. These evaluation metrics provide a comprehensive assessment of the image captioning model's performance, considering both the linguistic quality and semantic relevance of the generated captions. The results offer valuable insights into the model's strengths and areas for improvement, guiding further refinements and enhancements to enhance its captioning capabilities.

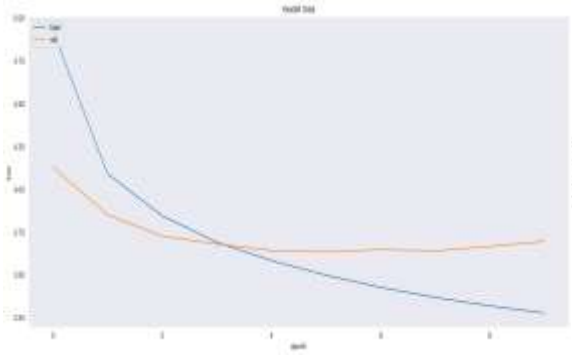

Figure 4. Evaluation of val_loss

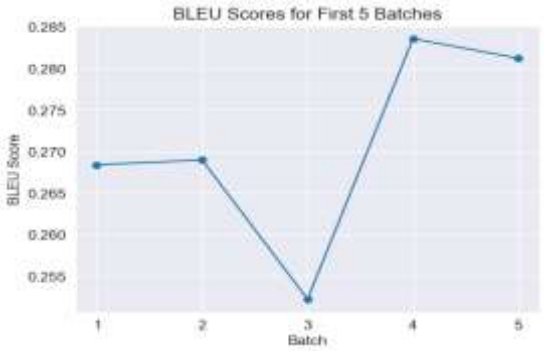

Figure 5: Bleu Scores Variation

The BLEU score, indicative of the similarity between the generated captions and the reference captions, reflects a moderate level of overlap, suggesting the model captures some aspects of the reference captions. However, caution is warranted in interpreting this score due to its limitations in capturing semantic equivalence. Conversely, the Semantic Similarity score exhibits a high level of semantic similarity between the generated and reference captions, affirming the model's efficacy in capturing the underlying meaning of the images and producing contextually relevant descriptions. This is pivotal for ensuring the fidelity of the generated captions.

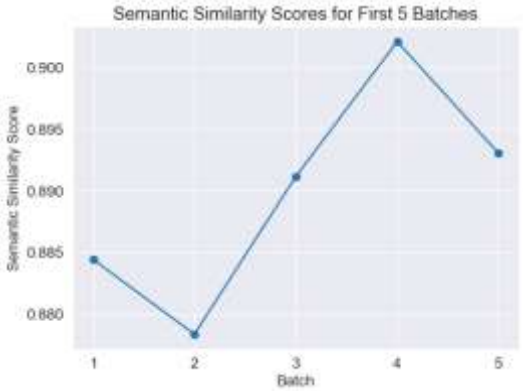

Figure 6. Semantic Similarity Variance

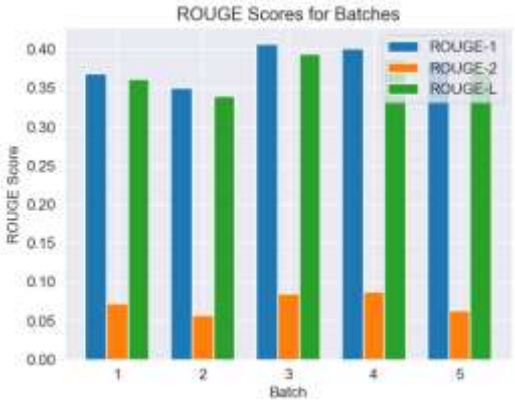

Figure 7. Comparison b/w Rouge Score

The ROUGE scores offer insights into the overlap between the generated and reference captions at different linguistic levels (unigrams, bigrams, and longest common subsequences). While the ROUGE-1 score implies a moderate overlap at the unigram level, the ROUGE-2 and ROUGE-L scores suggest relatively lower overlaps at the bigram and longest common subsequence levels, respectively. While the model displays promising performance in terms of semantic similarity and moderate overlap with reference captions, there exists room for improvement in capturing higher-order linguistic structures and maintaining coherence in longer sequences of words. Future research avenues could encompass refining the model architecture, integrating attention mechanisms, exploring alternative training strategies, and harnessing larger and more diverse datasets to ameliorate these limitations and further augment the model's performance in image captioning tasks.

5. Conclusion

Investigated the efficacy of an image captioning model utilizing VGG16 and LSTM architectures on the Flickr8K dataset. Through meticulous experimentation and evaluation, we gained valuable insights into the model's capabilities and limitations in generating descriptive captions for images. The findings of this research contribute to the broader understanding of image captioning techniques and provide valuable guidance for future advancements in this field. Our exploration of the VGG16 and LSTM architecture for image captioning involved several key steps, including data preprocessing, model training, and evaluation. The Flickr8K dataset, comprising 8,000 images paired with textual descriptions, served as the foundation for our research. We pre-processed the data, extracted high-level visual features using VGG16, and trained an LSTM network to generate captions based on the extracted features. The training process involved optimizing model parameters and fine-

tuning hyperparameters to achieve the best performance. Evaluation of the model's performance was conducted using a variety of metrics, including BLEU score, Semantic Similarity score, and ROUGE scores. While the model exhibited moderate levels of overlap with reference captions according to the BLEU score, it demonstrated a high degree of semantic similarity, indicating its ability to capture the underlying meaning of the images. However, analysis of ROUGE scores revealed challenges in maintaining coherence and capturing higher-order linguistic structures in the generated captions. Despite the promising performance in capturing semantic content, there is room for improvement in the model's ability to produce coherent and contextually relevant captions. Potential avenues for enhancement include refining the model architecture, integrating attention mechanisms to focus on relevant image regions, and leveraging larger and more diverse datasets for training. Additionally, fine-tuning hyperparameters and exploring alternative training strategies could further optimize the model's performance. The implications of this research extend beyond the realm of image captioning, with potential applications in various domains such as computer vision, natural language processing, and human-computer interaction. By bridging the semantic gap between visual content and textual descriptions, image captioning models have the potential to enhance accessibility, improve image understanding, and facilitate communication between humans and machines. While our study represents a significant step forward in the field of image captioning, there are still challenges to be addressed and opportunities to be explored. By building upon the insights gained from this research and continuing to push the boundaries of innovation, we can pave the way for more advanced and effective image captioning systems with widespread applications across industries and disciplines.

References

1. Verma, A., Yadav, A. K., Kumar, M., & Yadav, D. (2024). Automatic image caption generation using deep learning. *Multimedia Tools and Applications*, *83*(2), 5309-5325.
2. Chen, W., Hu, H., Li, Y., Ruiz, N., Jia, X., Chang, M. W., & Cohen, W. W. (2024). Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, *36*.
3. Ghandi, T., Pourreza, H., & Mahyar, H. (2023). Deep learning approaches on image captioning: A review. *ACM Computing Surveys*, *56*(3), 1-39.
4. Waheed, S. R., Rahim, M. S. M., Suaib, N. M., & Salim, A. A. (2023). CNN deep learning-based image to vector depiction. *Multimedia Tools and Applications*, *82*(13), 20283-20302.
5. Selivanov, A., Rogov, O. Y., Chesakov, D., Shelmanov, A., Fedulova, I., & Dylov, D. V. (2023). Medical image captioning via generative pretrained transformers. *Scientific Reports*, *13*(1), 4171.
6. Wang, T., Zhang, J., Fei, J., Ge, Y., Zheng, H., Tang, Y., ... & Zheng, F. (2023). Caption anything: Interactive image description with diverse multimodal controls. *arXiv preprint arXiv:2305.02677*.
7. Jabeen, S., Li, X., Amin, M. S., Bourahla, O., Li, S., & Jabbar, A. (2023). A review on methods and applications in multimodal deep learning. *ACM Transactions on Multimedia Computing, Communications and Applications*, *19*(2s), 1-41.
8. Ma, Y., Ji, J., Sun, X., Zhou, Y., & Ji, R. (2023). Towards local visual modeling for image captioning. *Pattern Recognition*, *138*, 109420.
9. Ramos, R., Martins, B., Elliott, D., & Kementchedjhieva, Y. (2023). Smallcap: lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2840-2849).
10. Zhao, W., Yang, W., Chen, D., & Wei, F. (2023). DFEN: Dual feature enhancement network for remote sensing image caption. *Electronics*, *12*(7), 1547.
11. Sarto, S., Barraco, M., Cornia, M., Baraldi, L., & Cucchiara, R. (2023). Positive-augmented contrastive learning for image and video captioning evaluation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6914-6924).
12. Luo, J., Li, Y., Pan, Y., Yao, T., Feng, J., Chao, H., & Mei, T. (2023). Semantic-conditional diffusion networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 23359-23368).
13. Tian, Y., Fan, L., Isola, P., Chang, H., & Krishnan, D. (2024). Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, *36*.
14. Miikkulainen, R., Liang, J., Meyerson, E., Rawal, A., Fink, D., Francon, O., ... & Hodjat, B. (2024). Evolving deep neural networks. In *Artificial intelligence in the age of neural networks and brain computing* (pp. 269-287). Academic Press.
15. Nazir, S., Dickson, D. M., & Akram, M. U. (2023). Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. *Computers in Biology and Medicine*, *156*, 106668.
16. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., ... & Lee, Y. J. (2023). Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 22511-22521).
17. Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2023). Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

18. Brooks, T., Holynski, A., & Efros, A. A. (2023). Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18392-18402).
19. Chen, W., Hu, H., Li, Y., Ruiz, N., Jia, X., Chang, M. W., & Cohen, W. W. (2024). Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, *36*.
20. Jia, X., Zhao, Y., Chan, K. C., Li, Y., Zhang, H., Gong, B., ... & Su, Y. C. (2023). Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*.
21. Li, J., Li, D., Savarese, S., & Hoi, S. (2023, July). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning* (pp. 19730-19742). PMLR.
22. Le Gallo, M., Khaddam-Aljameh, R., Stanisavljevic, M., Vasilopoulos, A., Kersting, B., Dazzi, M., ... & Sebastian, A. (2023). A 64-core mixed-signal in-memory compute chip based on phase-change memory for deep neural network inference. *Nature Electronics*, *6*(9), 680-693.
23. Mijwil, M. M., Al-Mistarehi, A. H., Abotaleb, M., El-kenawy, E. S. M., Ibrahim, A., Abdelhamid, A. A., & Eid, M. M. (2023). From Pixels to Diagnoses: Deep Learning's Impact on Medical Image Processing-A Survey. *Wasit Journal of Computer and Mathematics Science*, *2*(3), 9-15.
24. Graves, A., Mohamed, A., & Hinton, G. E. (2013). Speech recognition with deep recurrent neural networks. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6645–6649). IEEE.
25. Albawi, S., & Mohammed, T. A. (2017). Understanding of a Convolutional Neural Network.
26. Amritkar, C., & Jabade, V. (2018). Image Caption Generation Using Deep Learning Technique. In *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018* (pp. 1–4).
27. Barlas, G., Veinidis, C., & Arampatzis, A. (2021). What we see in a photograph: content selection for image captioning. *The Visual Computer, 37*(6), 1309–1326.
28. Bayoudh, K., Knani, R., Hamdaoui, F., & Mtibaa, A. (2021). A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 1–32.
29. Biswas, R., Barz, M., & Sonntag, D. (2020). Towards explanatory interactive image captioning using top-down and bottom-up features, beam search and re-ranking. *KI-Künstliche Intelligenz, 34*(4), 571–584.
30. Cao, P., Yang, Z., Sun, L., Liang, Y., Yang, M. Q., & Guan, R. (2019). Image captioning with bidirectional semantic attention-based guiding of long short-term memory. *Neural Processing Letters, 50*(1), 103–119.
31. Chen, H., Ding, G., Lin, Z., Guo, Y., Shan, C., & Han, J. (2021). Image captioning with memorized knowledge. *Cognitive Computation, 13*(4), 807–820.
32. Choi, Y., Berg, T. L., U N C Chapel Hill, & Stony Brook. (2014). TREE TALK: Composition and Compression of Trees for Image Descriptions. *The Journal of Machine Learning Research, 2*, 351–362.
33. Chu, Y., Yue, X., Yu, L., Sergei, M., & Wang, Z. (2020). Automatic image captioning based on ResNet50 and LSTM with soft attention. *Wireless Communications and Mobile Computing, 2020*, 2020.
34. Ding, G., Chen, M., Zhao, S., Chen, H., Han, J., & Liu, Q. (2019). Neural image caption generation with weighted training and reference. *Cognitive Computation, 11*(6), 763–777.
35. Donahue, J., Hendricks, L. A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., & Darrell, T. (2017). Long-term recurrent convolutional networks for visual recognition and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(4), 677–691.
36. Dong, X., Long, C., Xu, W., & Xiao, C. (2021). Dual graph convolutional networks with transformer and curriculum learning for image captioning. *arXiv preprint arXiv:2108.02366*.
37. Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. In *European conference on computer vision* (pp. 15–29). Springer.
38. Ghosh, A., Dutta, D., & Moitra, T. (2020). A neural network framework to generate caption from images. *Springer Nature Singapore Pte Ltd.*, 171–180.
39. Gong, Y., Wang, L., Hodosh, M., Hockenmaier, J., & Lazebnik, S. (2014). Improving image-sentence embeddings using large weakly annotated photo collections. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8692 LNCS*(PART 4), 529–545.
40. Gupta, N., & Jalal, A. S. (2020). Integration of textual cues for fine-grained image captioning using deep CNN and LSTM. *Neural Computing and Applications, 32*(24), 17899–17908.
41. He, C., & Hu, H. (2019). Image captioning with text-based visual attention. *Neural Processing Letters, 49*(1), 177–185.
42. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.
43. Hodosh, M., Young, P., & Hockenmaier, J. (2015). Framing image description as a ranking task: Data, models and evaluation metrics. *IJCAI International Joint Conference on Artificial Intelligence, 2015-Janua*(IJCAI), 4188–4192.

44. Huang, F., Li, Z., Wei, H., Zhang, C., & Ma, H. (2020). Boost image captioning with knowledge reasoning. *Machine Learning, 109*(12), 2313–2332.
45. Jiang, T., Zhang, Z., & Yang, Y. (2019). Modeling coverage with semantic embedding for image caption generation. *The Visual Computer, 35*(11), 1655–1665.
46. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
47. Karpathy, A., & Fei-Fei, L. (2017). Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(4), 664–676.
48. Katpally, H., & Bansal, A. (2020). Ensemble learning on deep neural networks for image caption generation. *Proceedings - 14th IEEE International Conference on Semantic Computing, ICSC 2020*, 61–68.
49. Khan, M. J., & Curry, E. (2020). Neuro-symbolic visual reasoning for multimedia event processing: Overview, prospects and challenges. In *CIKM (Workshops)*.
50. Khan, M. J., Khan, M. J., Siddiqui, A. M., & Khurshid, K. (2021). An automated and efficient convolutional architecture for disguise-invariant face recognition using noise-based data augmentation and deep transfer learning. *The Visual Computer*, 1–15.