Automated Image Captioning System

Sri Sai Jishnu Edara(21BCE6201)
Macha Naga Sai Vignesh(21BCE1834)
Viswateja Reddy B(21BCE1851)
Gajjarapu Satya Surya Vital Chowdary(21BCE6153)

PROBLEM STATEMENT

- The problem is to build an image captioning system that can generate descriptive captions for images automatically. The system should be able to understand the visual content of an image and generate human-like captions that accurately describe the scene or objects in the image.

OBJECTIVE

The objective of the project includes:

1. Developing a robust image captioning system that accurately describes the content of images.

2. Exploring effective methods for feature extraction from images to capture both visual and semantic information.

3. Designing and implementing a deep learning model architecture that can generate human-like captions based on the extracted features.

4. Training and optimizing the model to ensure it produces contextually relevant and accurate captions for a variety of images.

DATASET AND DESCRIPTION

Flickr 8k Dataset

8,000 photos with five distinct captions that clearly describe the important objects and events are included in this new benchmark collection for sentence-based image description and search. .. The photographs were picked from six distinct Flickr groups, and they were carefully chosen to show a range of settings and scenarios rather than any famous persons or places.

https://www.kaggle.com/datasets/adityajn105/flickr8k?select=Images

LITERATURE REVIEW

S.NO	TITLE	AUTHORS	FINDINGS/ FUTURE WORK
1	Boosting Image Captioning with Attributes https://openaccess.thecvf.com/content_ICCV_2017/papers/Yao_Boosting_Image_Captioning_ICCV_2017_paper.pdf	Ting Yao , Yingwei Pan , Yehao Li , Zhaofan Qiu and Tao Mei	LSTM-A, a novel image captioning architecture, integrates attributes with image representations via RNNs, achieving state-of-the-art performance on COCO (METEOR/CIDEr-D: 25.5%/100.2%). Effective attribute learning with Multiple Instance Learning further boosts accuracy. Different architecture variants explore the image-attribute relationship, showing promising results. Future work focuses on expanding attribute vocabulary and generating free-form sentences.
2	A New CNN-RNN Framework For Remote Sensing Image Captioning	Genc Hoxha, Farid Melgani, Jacopo Slaghenauffi	The study introduces a novel framework for remote sensing image captioning that combines both generation and retrieval-based approaches. The framework uses a CNN-RNN architecture with beam search to generate multiple captions for an image. The best caption is then selected based on its similarity to the reference captions of similar images. The results show that the proposed framework is promising and improves the quality of image descriptions. Future work includes developing more advanced scoring and retrieval mechanisms.
3	Image Caption Generation for News Articles https://aclanthology.org/2020.coling-main.176/	Zhishen Yang, Naoaki Okazaki	The paper proposes a Transformer model for news-image captioning, which generates descriptions of images based on both textual and visual features. The model outperforms existing methods and demonstrates that incorporating textual information from news articles improves the quality of the generated captions. The study suggests further exploration of visual content recognition and transferability of the news-image captioning model to other multimodal tasks.
4	Attribute Conditioned Fashion Image Captioning https://ieeexplore-ieee-org.egateway.chennai.vit.ac.in/document/9897417	Chen Cai, Kim-Hui Yap, Suchen Wang	The study introduces a new way to create personalized fashion image captions by using semantic attributes as a guide. Unlike traditional methods, this approach allows users to influence caption generation based on their preferences. The researchers curated a dataset, FACAD170K, for training and evaluating their method.

S.NO	TITLE	AUTHORS	FINDINGS/ FUTURE WORK
5	DEEP LEARNING APPROACHES ON IMAGE CAPTIONING: A REVIEW https://arxiv.org/pdf/2201.12944.pdf	Taraneh Ghandi, Hamidreza Pourreza, Hamidreza Mahyar	The review explores improved image captioning using deep learning and vision-language pre-training, offering a comprehensive review, addressing challenges, ranking methods, and suggesting future directions, with a particular focus on creating specialized visual assistants for visually impaired individuals.
6	An Image Captioning Model Based on SE-ResNest and EMSA https://ieeexplore-ieee-org.egateway.chennai.vit.ac.in/document/10332008	Rongrong Yuan, Haisheng Li	The paper introduces an image caption generation model that combines a ResNest-based encoder with a two-layer LSTM decoder enhanced with a Squeeze-and-Excitation module and multi-head attention, demonstrating improved accuracy and efficiency on Flickr8k and Flickr30k datasets, while acknowledging the need for further optimization in recognizing image details in future work.
7	Hybrid Feature and Sequence Extractor based Deep Learning Model for Image Caption Generation https://ieeexplore-ieee-org.egateway.chennai.vit.ac.in/document/9579897	Rohit Kushwaha, Anupam Biswas	The paper presents a hybrid deep learning model combining a VGG-19-based image feature extractor and an LSTM-based sequence processor for generating informative image captions, achieving robustness as evidenced by BLEU score comparisons with other models. The proposed model exhibits potential for generating scene descriptions from real images captured by vision-enabled eyewear for visually impaired individuals, suggesting future applications in assisting this demographic.
8	A Deep Neural Framework for Image Caption Generation Using GRU-Based Attention Mechanism https://arxiv.org/pdf/2203.01594.pdf	Rashid khana , M Shujah Islama , Khadija Kanwala , Mansoor Iqbal, Md. Imran Hossaina & Zhongfu Ye	The study introduces a joint model for image captioning that combines pre-trained convolutional neural networks (CNN) for feature extraction, a Gated Recurrent Unit (GRU) as the decoder, and an integrated Bahdanau attention mechanism. The model, evaluated on the MSCOCO dataset, demonstrates competitive performance against state-of-the-art approaches, successfully generating appropriate captions for images by focusing on specific portions through attention.

S.NO	TITLE	AUTHORS	FINDINGS/ FUTURE WORK
9	Research on Image Tibetan Caption Generation Method Fusion Attention Mechanism https://ieeexplore-ieee-org.egateway.chennai.vit.ac.in/document/10348351	Jianjun Xia, Xin Yang, Qiong Ni, Dingguo Gao	The study introduces an image Tibetan caption generation method that utilizes VGG19, InceptionV3, and ResNet101 as backbone networks, incorporating an improved attention mechanism to assign different weights to key objects in the image, leading to superior performance on the Flickr8K and Flickr30K-tic Tibetan caption datasets. The future work includes incorporating traditional Tibetan cultural thangkas and mural images for algorithm research, aiming to enhance understanding, cultural exchange, and intelligent information processing in Tibetan areas.
10	LSTM-VGG-16: A Novel and Modular Model for Image Captioning Using Deep Learning Approaches https://jespublication.com/upload/2021-V12I1116.pdf	Tirumanisetty Venkata Sneha, Dr.S.Jhansi Rani	The paper explores three image captioning methods using deep neural networks, showcasing the effectiveness of models such as VGG16 for feature extraction and LSTM for caption generation. The future work aims to enhance accuracy by leveraging advanced algorithms like Faster RCNN and SSD, particularly focusing on forest land cover detection using high GPU and TensorFlow Object Detection API.
11	A Linear Sub-Structure with Co-Variance Shift for Image Captioning https://ieeexplore-ieee-org.egateway.chennai.vit.ac.in/document/9654828	Shaik Rafi, Ranjita Das	The paper introduces a novel image captioning model that effectively addresses the Gradient Diminishing problem by employing LSTM to fuse local and global characteristics, achieving state-of-the-art accuracy on the Flickr 8k dataset. The proposed model, incorporating GLoVe embedding, Inception V3, and a Linear Sub-Structure, lays the groundwork for future enhancements, suggesting the extension of the approach to handle graphical moving images for dynamic text generation.
12	Medical image captioning via generative pretrained transformers https://www.nature.com/articles/s41598-023-31223-5	Alexander Selivanov, Oleg Y. Rogov, Daniil Chesakov, Artem Shelmanov, Irina Fedulova & Dmitry V. Dylov	The paper introduces a novel model combining Show-Attend-Tell and GPT-3 language models for medical image captioning, providing comprehensive radiology reports with pathologies, their locations, and 2D heatmaps. The results demonstrate efficient applicability to chest X-ray image captioning, outperforming Transformer-based decoders. Future work could explore advancements in interactive training, like active learning and dialog-based models, to further enhance the performance of medical image captioning and support ongoing research in this domain.

S.NO	TITLE	AUTHORS	FINDINGS/ FUTURE WORK
13	Visual attention based on long-short term memory model for image caption generation https://ieeexplore-ieee-org.egateway.chennai.vit.ac.in/document/7979342	Shiru Qu, Yuling Xi, Songtao Ding	The paper introduces an attention-based L-STM model for image caption generation, achieving state-of-the-art performance on benchmark datasets. The attention mechanism enhances interpretability, and future work aims to achieve a comprehensive understanding of images through unsupervised data, indicating a direction toward more advanced and holistic image captioning systems.
14	Comparison of VGG and ResNet used as Encoders for Image Captioning https://ieeexplore-ieee-org.egateway.chennai.vit.ac.in/document/9108880	Viktar Atliha, Dmitrij Šešok	The paper compares VGG and ResNet as encoders for image captioning, revealing that ResNet outperforms VGG in achieving higher BLEU-4 scores with fewer training epochs. This emphasizes the significant impact of the encoder on model performance. Future work could explore further improvements in image captioning models by experimenting with different encoder architectures and evaluating their effectiveness in diverse tasks.
15	Generating Caption for Image using Beam Search and Analyzation with Unsupervised Image Captioning Algorithm https://ieeexplore-ieee-org.egateway.chennai.vit.ac.in/document/9432245	Prashant Giridhar Shambharkar, Priyanka Kumari, Pratik Yadav, Rajat Kumar	The paper presents an image caption generator using CNN-RNN with encoder-decoder architecture, incorporating pre-trained imagenet models for feature extraction, and comparing prediction methods such as beam search and argmax. Additionally, an unsupervised image captioning model is introduced, leveraging over two million sentences from Shutterstock, demonstrating promising results even without image sentence labels. Future work could focus on refining the unsupervised model, exploring diverse datasets, and potentially integrating the developed caption generator into mobile applications to assist differently-abled individuals relying on text-to-speech features.
16	Triple Sequence Generative Adversarial Nets for Unsupervised Image Captioning https://ieeexplore-ieee-org.egateway.chennai.vit.ac.in/document/9414335#full-text-header	Yucheng Zhou, Wei Tao, Wenqiang Zhang	The paper introduces a novel unsupervised image captioning method, Triple Sequence Generative Adversarial Nets, focusing on the correspondence between images and sentences through an image generator, discriminator, and sentence generator. The experimental results show significant improvements over baselines, indicating the effectiveness of the proposed model. Future work could explore further enhancements to the model and evaluate its performance on diverse datasets to establish its generalizability across various image-captioning scenarios.

S.NO	TITLE	AUTHORS	FINDINGS/ FUTURE WORK
17	Stack-VS: Stacked Visual-Semantic Attention for Image Caption Generation https://ieeexplore-ieee-org.egateway.chennai.vit.ac.in/document/9174742	Ling Cheng, Wei Wei, Xianling Mao, Yong Liu, Chunyan Miao	The paper introduces Stack-VS, a novel multi-stage framework for image caption generation, combining top-down and bottom-up attention models with a stack decoder. Extensive experiments on MSCOCO show significant improvements over state-of-the-art methods. Future work could explore incorporating natural language inference for more reasonable captions and experimenting with different architectures, such as graph convolution networks, to further enhance captioning performance.
18	Adaptive Hard Example Mining for Image Captioning https://ieeexplore-ieee-org.egateway.chennai.vit.ac.in/document/8803418	Yongzhuang Wang, Yangmei Shen, Hongkai Xiong, Weiyao Lin	The paper introduces an adaptive hard example mining method for image captioning, utilizing reinforcement learning with a beam search algorithm. The approach automatically selects and focuses on hard examples, leading to improved model performance without hyper-parameter tuning. Future work could explore extending this approach to other natural language processing tasks and evaluating its performance on diverse datasets.
19	Image Caption Method Based on Graph Attention Network with Global Context https://ieeexplore-ieee-org.egateway.chennai.vit.ac.in/document/9886239	Jiahong Sui, Huimin Yu, Xinyue Liang Ping Ping	The paper introduces an image caption method based on a graph attention network with global context, improving image caption quality by capturing both global and local features. The approach successfully utilizes a graph attention network to optimize image representation. Future work could involve extending the model to incorporate textual context information, enhancing semantic expression, and further improving image caption performance.
20	Image Caption Enhancement with GRIT, Portable ResNet and BART Context-Tuning https://ieeexplore-ieee-org.egateway.chennai.vit.ac.in/document/10185494	Wuyang Zhang, Jianming Ma	The paper introduces an image captioning architecture combining GRIT, Portable ResNet, and BART, providing detailed captions with contextual information. The approach enhances accuracy and speed compared to conventional models, offering potential applications in areas like parking management. Future work involves improving object detection models, exploring grammar tuning on larger corpora, and considering automated corpus generation for BART context-tuning to further enhance the system's capabilities.

SUMMARY OF THE FINDINGS

- Diverse models, such as LSTM-A, Stack-VS, and Transformer-based approaches, showcase advancements in image captioning, achieving state-of-the-art performance on various datasets.
- Attention mechanisms, GPT-3 integration, and innovative architectures contribute to enhanced image caption quality, demonstrating the versatility of these techniques.
- The combination of CNN-RNN and transformer architectures, along with unsupervised methods, introduces flexibility and adaptability to different image captioning scenarios.
- Several models address challenges like gradient diminishing, feature extraction, and the semantic gap, providing comprehensive solutions for accurate and detailed image descriptions.
- The research explores image captioning beyond conventional domains, such as medical images, news articles, and personalized fashion, indicating the applicability of these models in diverse contexts.
- Ethical considerations, including bias in image captioning, need more attention and careful exploration.

ISSUES AND GAPS IDENTIFIED

- Scalability Challenges: Efficient scaling of models for real-world applications remains a challenge, impacting broader adoption.
- Interpretability Concerns: Ensuring captions align with human intuition and are interpretable poses a persistent challenge.
- User-Centric Evaluation: Limited emphasis on user-centric evaluation hinders understanding of practical applicability and user satisfaction.
- Efficiency Optimization: Challenges related to computational efficiency may hinder wider adoption in resource-constrained environments.
- Transferability Issues: Evaluating how well models transfer across diverse domains and datasets needs further exploration.
- External Context Integration: Limited exploration of incorporating external context in image captioning poses a potential gap in understanding contextual relevance.
- Real-Time Constraints: Investigating strategies to meet real-time requirements for generating captions in dynamic environments remains an unexplored aspect in some models.

TEAM ROLE'S

- Viswateja Reddy B
  - Research and implement the image captioning model architecture.
  - Experiment with different neural network architectures suitable for image captioning.
- Macha Naga Sai Vignesh
  - Gather and preprocess relevant datasets for training and testing.
  - Collaborate with MLE to define specifications for the dataset.
- Sri Sai Jishnu Edara
  - Implement the chosen model using a deep learning framework.
  - Optimize the model's architecture for better performance.
- Vital Chowdary
  - Implement data augmentation techniques to enhance the diversity of the dataset.
  - Perform exploratory data analysis to identify patterns and insights in the dataset.

REFERENCES

1. Ting Yao , Yingwei Pan , Yehao Li , Zhaofan Qiu and Tao Mei, “Boosting Image Captioning with Attributes”
2. Genc Hoxha, Farid Melgani, Jacopo Slaghenauffi, “A New CNN-RNN Framework For Remote Sensing Image Captioning”
3. Zhishen Yang, Naoaki Okazaki, “Image Caption Generation for News Articles”
4. Chen Cai, Kim-Hui Yap, Suchen Wang, “Attribute Conditioned Fashion Image Captioning”
5. Taraneh Ghandi, Hamidreza Pourreza, Hamidreza Mahyar, “DEEP LEARNING APPROACHES ON IMAGE CAPTIONING: A REVIEW”
6. Rongrong Yuan, Haisheng Li, “An Image Captioning Model Based on SE-ResNest and EMSA”
7. Rohit Kushwaha, Anupam Biswas, “Hybrid Feature and Sequence Extractor based Deep Learning Model for Image Caption Generation”
8. Rashid khana , M Shujah Islama , Khadija Kanwala , Mansoor Iqbal, Md. Imran Hossaina & Zhongfu Ye, “A Deep Neural Framework for Image Caption Generation Using GRU-Based Attention Mechanism”
9. Jianjun Xia, Xin Yang, Qiong Ni, Dingguo Gao, “Research on Image Tibetan Caption Generation Method Fusion Attention Mechanism”
10. Tirumanisetty Venkata Sneha, Dr.S.Jhansi Rani, “LSTM-VGG-16: A Novel and Modular Model for Image Captioning Using Deep Learning Approaches”

REFERENCES

10. Shaik Rafi, Ranjita Das, “A Linear Sub-Structure with Co-Variance Shift for Image Captioning”

11. Alexander Selivanov, Oleg Y. Rogov, Daniil Chesakov, Artem Shelmanov, Irina Fedulova & Dmitry V. Dylov, “Medical image captioning via generative pretrained transformers”

12. Shiru Qu, Yuling Xi, Songtao Ding, “Visual attention based on long-short term memory model for image caption generation”

13. Viktar Atliha, Dmitrij Šešok, “Comparison of VGG and ResNet used as Encoders for Image Captioning”

14. Prashant Giridhar Shambharkar, Priyanka Kumari, Pratik Yadav, Rajat Kumar, “Generating Caption for Image using Beam Search and Analyzation with Unsupervised Image Captioning Algorithm”

15. Yucheng Zhou, Wei Tao, Wenqiang Zhang, “Triple Sequence Generative Adversarial Nets for Unsupervised Image Captioning”

16. Ling Cheng, Wei Wei, Xianling Mao, Yong Liu, Chunyan Miao, “Stack-VS: Stacked Visual-Semantic Attention for Image Caption Generation”

17. Yongzhuang Wang, Yangmei Shen, Hongkai Xiong, Weiyao Lin, “Adaptive Hard Example Mining for Image Captioning”

18. Jiahong Sui, Huimin Yu, Xinyue Liang Ping Ping, “Image Caption Method Based on Graph Attention Network with Global Context”

19. Wuyang Zhang, Jianming Ma, “Image Caption Enhancement with GRIT, Portable ResNet and BART Context-Tuning”