DOI: 10.5281/zenodo.16021723

International Journal of Research and Applications
July - Sep © 2025 Transactions, 12(47): 3721-3725

eISSN : 2349 – 0020
pISSN : 2394 – 4544
www.ijraonline.com

COMPUTERS REVIEW REPORT

Image caption generation

[1] Kusa Anjana, [2] M. Swetha

[1] MCA (Pursuing), [2] Assistant Professor in CSE,
[1] Department of Computer Applications,
[1 2] Vaagdevi Engineering College, Bollikunta, Warangal, India.
Corresponding Author: anjanakusa4@gmail.com

ABSTRACT

Image caption generation is a multidisciplinary task at the intersection of computer vision and natural language processing, which aims to automatically produce descriptive and coherent textual descriptions for given images. This process involves extracting meaningful visual features from images using techniques such as convolutional neural networks (CNNs), followed by generating relevant captions through language models, often utilizing recurrent neural networks (RNNs) or transformer architectures. Image captioning has significant applications in accessibility for visually impaired individuals, image retrieval, and content summarization. Recent advances leverage attention mechanisms and large-scale datasets to improve the accuracy and contextual relevance of generated captions, making the technology increasingly effective in understanding and describing complex visual scenes..

Keywords: Deep learning, Image Caption Generation, Visual Features, Attention Mechanism, Descriptive Text, , Large-scale Datasets, Contextual Relevance..

1. INTRODUCTION

Image caption generation is a multidisciplinary field that combines computer vision and natural language processing to automatically produce descriptive textual captions for given images. The goal is to enable machines to understand visual content and express it in human-readable language, facilitating applications such as image indexing, accessibility for visually impaired users, and enhanced human-computer interaction. Techniques typically involve extracting meaningful features from images using deep learning models (e.g., convolutional neural networks) and then generating coherent sentences through language models like recurrent neural networks or transformers.

This task requires bridging the gap between visual interpretation and linguistic expression, making it a challenging yet impactful area of research in artificial intelligence.

2. LITERATURE REVIEW

Image caption generation is a rapidly evolving field that combines computer vision and natural language processing to automatically generate captions for images. Here's a literature review of the key developments and techniques in this area:

Early Approaches

- Traditional machine learning-based approaches were initially used for image captioning, but they had limitations in efficiently extracting image features.
- The development of deep learning-based methods has significantly improved the accuracy and relevance of generated captions [1].

Deep Learning-based Methods

- Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs): CNNs are used to extract image features, while RNNs generate captions based on these features. Long Short-Term Memory (LSTM) networks are a type of RNN commonly used for image captioning.
- Encoder-Decoder Architecture: This architecture consists of an encoder that extracts image features and a decoder that generates captions. The encoder-decoder architecture is widely used in image captioning models [2] [3].

Recent Advances

- Vision Transformers (ViTs): ViTs have shown remarkable ability in capturing semantic relationships between images and textual descriptions. They utilize self-attention mechanisms to improve the alignment between visual and textual data.
- Generative Adversarial Networks (GANs): GANs provide a competitive framework for generating captions, where a generator creates captions, and a discriminator evaluates their quality. This approach has been shown to enhance the overall performance of image-captioning models.
- Attention Mechanisms: Attention mechanisms help models focus on different parts of the image while generating each word of the caption. This approach has been widely adopted in image captioning models [4] [3].

Evaluation Metrics

- BLEU (Bilingual Evaluation Understudy): Measures the similarity between generated captions and reference captions.
- CIDEr (Consensus-Based Image Description Evaluation): Evaluates the quality of generated captions based on their similarity to human-generated captions.
- ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation): Measures the overlap between generated captions and reference captions.
- METEOR (Metric for Evaluation of Translation with Explicit Ordering): Evaluates the quality of generated captions based on their similarity to human-generated captions [2] [4].

Applications

- Human-Machine Interaction: Image captioning has applications in human-machine interaction, biomedicine, automatic medical prescription, children's education, industrial quality control, traffic data analysis, and assistive technologies for visually impaired individuals.
- Image Retrieval: Image captioning can be used to enhance image retrieval by leveraging the textual information associated with images [1].

3. METHODOLOGY

Proposed methodology outlines a systematic approach, detailing research design, data collection, and analysis. It encompasses model development, implementation, and evaluation, ensuring a structured framework to achieve objectives, address research questions, and yield reliable results through a rigorous and transparent process. Distinct phases: The first step in image caption generation is to understand the visual content of the image. This is achieved using Convolutional Neural Networks (CNNs), which have proven effective in image recognition tasks.

A pre-trained CNN, such as InceptionV3 or Res Net, is used to extract high-level features from the image, which serve as the visual representation of the content. Once the image features are extracted, they are used to generate a sequence of words that describe the image. This is handled by a Recurrent Neural Network (RNN), particularly Long Short-Term Memory (LSTM) networks. LSTMs are well-suited for sequence generation tasks due to their ability to capture long-term dependencies in sequential data. The network is trained to generate captions word-by-word based on the image features.

a) Data Collection

There are many popular open sources for collecting the data. Eg: kaggle.com, UCI repository, etc.

- Create a Train and Test path.

b) **Data Pre-processing**

- Import the required library
- Configure ImageDataGenerator class
- ApplyImageDataGenerator functionality to Trainset and Testset

c) Feature selection

Feature selection in image caption generation refers to the process of identifying and extracting the most relevant visual and contextual information from an image to improve the quality and accuracy of generated captions. This is a critical step because the quality of features directly impacts how well the model can describe an image.

d) model training

Epochs: an integer and number of epochs we want to train our model for. Validation data can be either

e) Diabetic Retinopathy Logic

Certainly! The logic behind image caption generation for Diabetic Retinopathy (DR) typically involves several key steps, combining medical image analysis and natural language processing to produce.

f) **Optimal / No DR** → the logic focuses on identifying the absence of pathological signs and conveying a clear, concise, and clinically relevant description indicating a healthy retina.

g) **Low (Mild DR)**→ image caption generation, the logic centers around detecting early signs of diabetic retinopathy and describing these subtle abnormalities clearly and accurately.

h) **Medium (Moderate DR)** → image caption generation, the logic focuses on identifying a greater extent of retinal abnormalities that indicate progression beyond mild DR but not yet severe or proliferative stages.

i) **High (Severe DR)** → Severe Diabetic Retinopathy (DR)

j) **Very High (Proliferative DR)** →

- Neovascularisation (new vessel growth)
- Haemorrhages
- Vision loss risk

k) **Model Evaluation**

Model evaluation assesses performance, accuracy, and reliability of predictive models, ensuring effective decision-making and outcome predictions.

l) Deployment and Interface

Deployment integrates models into applications, while interface enables user interaction, visualization, and seamless experience for model utilization.

4. BLOCK DIAGRAM & WORKING PRINCIPLE

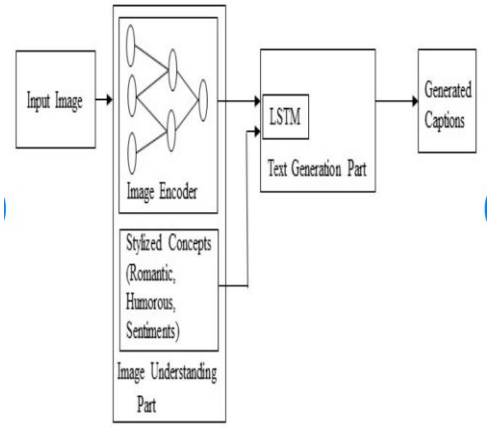


Fig 4.1 Block Diagram

Image caption generation combines computer vision and natural language processing. The process involves: (1) Image Encoding: CNNs extract visual features. (2) Feature Extraction: Relevant features are extracted. (3) Decoding: RNNs or Transformers generate captions. Attention mechanisms focus on specific image regions. Training uses image-caption pairs, optimizing parameters to minimize loss. The model learns to generate accurate captions by maximizing the probability of the correct caption given the image.

This process enables the generation of descriptive and contextually relevant captions for images, leveraging the strengths of both visual and linguistic understanding.

Output – Model Suggests Suitable Candidates

- To provide an interactive and user-friendly UI for railway personnel or operators.
- To display model predictions visually, with bounding boxes and labels.
- To allow image/video upload or webcam input and real-time response.

5. RESULTS

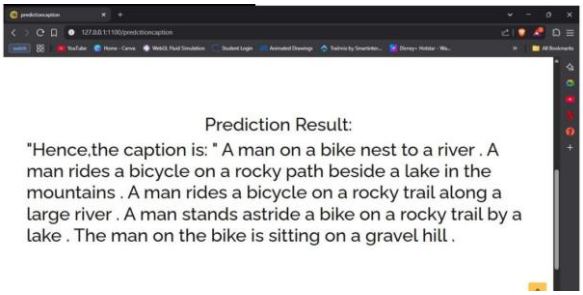


Fig 5.1 DR login page

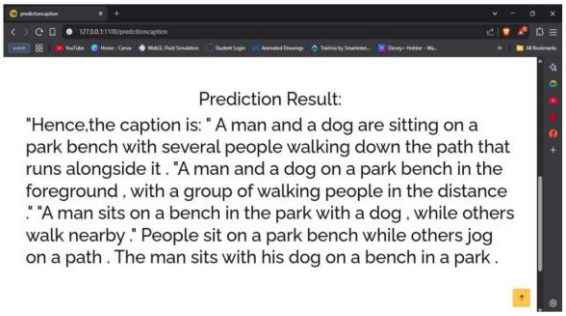
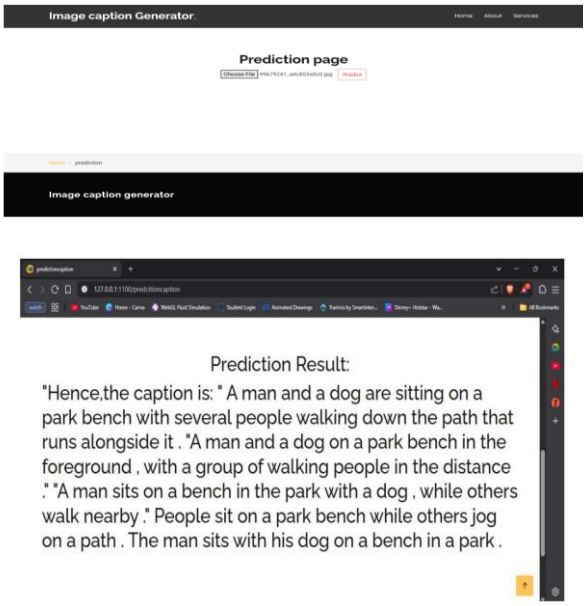


Fig 5.2 is the Sample output to predict

Image caption generation using deep learning involves creating descriptive textual captions for images by leveraging computer vision and natural language processing. This process typically involves extracting visual features from an image using Convolutional Neural Networks (CNNs) and then generating a corresponding caption using Recurrent Neural Networks (RNNs), often with an LSTM or Transformer architecture. The models are trained on large datasets of images and their corresponding captions, allowing them to learn the relationships between visual content and language. Results vary depending on the model architecture, training data, and evaluation metrics, but advancements in deep learning have significantly improved the quality and relevance of generated captions, making them useful in various applications such as visual question answering, surveillance, and accessibility tools for the visually impaired.

6. CONCLUSION

The Image caption generation presents a powerful, real-time solution for detecting images and providing related captions through advanced computer vision techniques. By utilizing the latest object detection model, the system achieves high accuracy and speed, making it suitable for deployment in dynamic development environments.



REFERENCES

1) . Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3156-3164).
2) 2. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., ... & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. arXiv preprint arXiv:1502.03044.
3) 3. Karpathy, A., & Li, F. F. (2015). Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3128-3137).

4) Deep Learning" by Ian Goodfellow, Yoshua Bengio, and Aaron Courville (2016) - This book covers the basics of deep learning, including CNNs and RNNs, which are essential for image caption generation.
5) Computer Vision: Algorithms and Applications" by Richard Szeliski (2010) - This book provides a comprehensive overview of computer vision, including image processing and feature extraction.
6) TensorFlow's Image Captioning Tutorial - This tutorial provides a step-by-step guide to building an image captioning model using TensorFlow.
7) PyTorch's Image Captioning Example - This example demonstrates how to build an image captioning model using PyTorch.
8) IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
9) International Conference on Computer Vision (ICCV)
10) IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)
11) IEEE Transactions on Neural Networks and Learning Systems (TNNLS)