International Journal of
AI Electronics and Nexus Energy
ISSN: 3070-0515 www.zesterapublications.com Original Research Paper

Image Caption Generation Using Deepseek API

[1]Margam Manasa,[2]Ms.G.Ramadevi
[1]MCA Student, Department Of Master of Computer Applications, Sri Chaitanya Institute Of Technology (Scit) Ponnekal ,Khammam
[1]Email : manasamargam10@gmail.com
[2]Assistant Professor, Department Of Master of Computer Applications, Sri Chaitanya Institute Of Technology (Scit) Ponnekal ,Khammam
[2]Email : www.rama1459@gmail.com

ABSTRACT

In the modern era of digital information, multimedia content such as images and videos dominates online communication, social media, e-commerce, and digital libraries. While humans can effortlessly perceive visual scenes and summarize them in natural language, automating this process has remained one of the most challenging tasks at the intersection of Computer Vision (CV) and Natural Language Processing (NLP). This project, titled 'Image Caption Generation Using DeepSeek API', addresses this challenge by designing, training, and implementing an intelligent web-based application capable of generating meaningful textual captions from uploaded images. The project adopts a dual-methodology framework to showcase both custom-trained models and state-of-the-art transfer learning. First, a custom Image Captioning model is developed and trained from scratch using an Encoder-Decoder architecture. A convolutional neural network (specifically, Keras InceptionV3 pre-trained on ImageNet) acts as the feature extractor, capturing high-dimensional spatial representations of input images. These visual features are then projected and combined with text token embeddings to feed a Recurrent Neural Network utilizing Long Short-Term Memory (LSTM) layers, which decodes the representation and generates captions sequentially using a custom tokenizer. Second, to achieve superior, production-ready accuracy, the web application integrates a pre-trained Salesforce BLIP (Bootstrapping Language-Image Pre-training) model. Implemented via Hugging Face Transformers in PyTorch, this model runs inference dynamically on a Flask web server, generating descriptive captions for images uploaded via a clean, user-friendly Bootstrap interface. In addition to raw caption generation, the architecture lays the foundation for a DeepSeek API post-processing layer.

Keywords: Image Caption Generation, Deep Learning, Computer Vision, Natural Language Processing (NLP), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Vision-Language Models, Salesforce BLIP, DeepSeek API, Transfer Learning, Transformer Architecture, Flask Web Application, Image Feature Extraction, Hugging Face Transformers, Artificial Intelligence (AI).

Received: 03-02-2026 | Accepted: 05-03-2026 | Published: 12-03-2026

International Journal of
AI Electronics and Nexus Energy
ISSN: 3070-0515 www.zesterapublications.com Original Research Paper

I. INTRODUCTION

Image captioning is a fundamental task in artificial intelligence that requires a deep understanding of both visual semantics and natural language structure. The goal is to design a software system that can automatically look at an input image and generate a grammatically correct, coherent, and contextually accurate textual sentence describing the objects, their attributes, and their relationships present in the scene. Historically, this task was tackled using heuristic rules and rigid templates, which frequently resulted in clunky, unnatural phrases. With the advent of modern deep learning and neural network architectures, researchers successfully applied encoder-decoder neural systems to bridge the gap between vision and text.

This project, titled 'Image Caption Generation Using DeepSeek API', implements a comprehensive, production-grade solution to this problem. It showcases a system designed with two complementary backends. The first is a custom-trained convolutional-recurrent model built on Keras and TensorFlow. This model uses the InceptionV3 convolutional network to extract spatial feature maps from images, maps them to a dense latent vector space, and uses a Long Short-Term Memory (LSTM) recurrent network to decode these features into sequential word tokens. The second backend is a modern vision-language transformer network, specifically Salesforce's BLIP (Bootstrapping Language-Image Pre-training) model, implemented in PyTorch. The application is integrated into a clean, modern web portal using the Flask framework, allowing users to upload images and instantly view accurate generated captions.

Furthermore, to push the boundaries of vision-language understanding, the project details the conceptual design for integrating the DeepSeek Large Language Model (LLM) API. Standard image captioning models generate concise, literal labels (e.g., 'a cat sitting on a table'). By feeding these base captions into the DeepSeek LLM, the system can post-process and expand them into rich, context-aware descriptions, compose creative stories, identify search engine keywords, and format the output into structured JSON representations. This demonstrates a complete, hybrid artificial intelligence pipeline that combines vision extraction, transfer learning, and large-scale language generation.

II. LITERATURE SURVEY

Evolution of Image Captioning Techniques

Automated image description has evolved through three major paradigms. Early attempts in the 2000s relied on template-based approaches. These algorithms first detected individual objects, actions, and locations using basic object detectors, and then mapped these elements into pre-defined grammatical templates (e.g., 'A [object1] is near a [object2] in the [location]'). While syntactically correct, these captions lacked linguistic variety and could not capture complex, dynamic relationships between objects.

The second phase saw the emergence of retrieval-based methods. Given a query image, the algorithm computed global visual similarities (using hand-crafted features like SIFT or Gabor filters) against a database of pre-captioned images. The caption of the most similar image was then retrieved and assigned to the new image. Although this produced natural-sounding sentences, the

International Journal of
AI Electronics and Nexus Energy
ISSN: 3070-0515 www.zesterapublications.com Original Research Paper

model could not generate new combinations of descriptions for novel objects.

The current paradigm is neural network-based image captioning, inspired by neural machine translation. Instead of translating from one language to another, these models translate an image into its corresponding description. Using an Encoder-Decoder architecture, a Convolutional Neural Network (CNN) acts as the vision encoder, converting an image into a fixed-size feature vector, and a Recurrent Neural Network (RNN) or Long Short-Term Memory (LSTM) network acts as the decoder, outputting words sequentially. This approach allows the system to generate novel captions for unseen combinations of objects.

Deep Learning in Computer Vision & Feature Extraction

Modern computer vision relies heavily on Convolutional Neural Networks (CNNs) to extract features. A CNN consists of layers of convolutional filters that apply mathematical operations to local receptive fields of an image. Early layers capture low-level features such as edges, gradients, and textures, while deeper layers detect high-level semantic concepts like object parts and full categories. Pooling layers reduce spatial dimensions, improving translation invariance and computational efficiency.

In this project, Keras's InceptionV3 model is used as a feature extractor in `ModelTraining.py`. Pre-trained on the ImageNet database (containing over 1.2 million images across 1,000 categories), InceptionV3 utilizes 'Inception blocks' that run convolutions with different filter sizes (1x1, 3x3, 5x5) and max pooling operations in parallel. This design allows the model to capture features at multiple spatial scales. For feature extraction, the final classification layer is removed, and the output of the second-to-last layer (a 2048-dimensional dense vector representing the abstract visual content) is saved. This vector is then fed into the text decoder.

III. EXISTING SYSTEM

The existing image caption generation systems primarily rely on traditional deep learning models such as CNN-LSTM encoder-decoder architectures or template-based image description techniques. These systems use Convolutional Neural Networks (CNNs) to extract visual features from images and Long Short-Term Memory (LSTM) networks to generate captions sequentially. Although these approaches can produce basic image descriptions, they often generate short, generic, and literal captions that fail to capture detailed context, object relationships, or semantic richness. Earlier template-based and retrieval-based methods are even more limited, as they depend on predefined sentence structures or retrieve captions from visually similar images rather than generating new descriptions.

Furthermore, conventional image captioning systems generally lack integration with advanced vision-language transformer models and large language models. As a result, they struggle to produce human-like, context-aware, and descriptive captions suitable for applications such as digital accessibility, content indexing, and automated metadata generation. Many existing solutions also require extensive training on large datasets, have limited adaptability to diverse image domains, and do not provide enhanced outputs such as detailed descriptions, keywords, or structured metadata. These limitations reduce their effectiveness in real-world applications where richer and more informative image understanding is required.

IV. PROPOSED SYSTEM

The proposed system introduces an intelligent web-based image caption generation framework that combines deep

International Journal of
AI Electronics and Nexus Energy
ISSN: 3070-0515 www.zesterapublications.com Original Research Paper

learning, vision-language transformers, and large language model integration to produce accurate and context-aware image descriptions. The system incorporates two complementary approaches: a custom CNN-LSTM encoder-decoder model for learning image-caption relationships and a pre-trained Salesforce BLIP (Bootstrapping Language-Image Pre-training) model for high-quality caption generation. A Flask-based web application provides an interactive interface where users can securely upload images, process them through the captioning model, and instantly receive generated captions. This hybrid architecture enhances caption quality while ensuring scalability, efficiency, and ease of deployment.

To further improve the generated output, the proposed system conceptually integrates the DeepSeek API as a post-processing layer. After the BLIP model generates a base caption, the DeepSeek large language model expands it into a more descriptive and context-aware paragraph, extracts relevant keywords, and produces structured outputs that can be used for metadata generation, search engine optimization, and content indexing. The system supports standard image formats, performs automatic image preprocessing, and leverages modern frameworks such as TensorFlow, PyTorch, Hugging Face Transformers, and Flask to provide a reliable, production-ready image captioning solution suitable for applications including digital accessibility, multimedia management, and intelligent image retrieval.

V. SYSTEM ARCHITECTURE

The proposed Image Caption Generation Using DeepSeek API system follows a client-server architecture that integrates deep learning, computer vision, natural language processing, and web technologies to generate meaningful captions for uploaded images. The workflow begins when the user accesses the web application through the frontend interface and uploads an image. The web interface, developed using HTML, CSS, Bootstrap, and Flask templates, provides a simple and user-friendly environment for image uploading, viewing generated captions, and managing application features. After receiving the image, the request is forwarded to the Flask backend server, which acts as the central controller of the entire system.

The Flask backend is responsible for managing all application logic. It receives the uploaded image, validates the file type, stores it temporarily, and forwards it to the preprocessing module. During preprocessing, the image is resized, normalized, converted into tensor format, and transformed into a representation suitable for deep learning models. This preprocessing stage ensures that the input image satisfies the requirements of the caption generation models while maintaining consistency and improving inference accuracy.

Once preprocessing is completed, the processed image is sent to the model inference engine. The architecture supports two caption generation approaches. The first approach utilizes a custom CNN-LSTM model, where a Convolutional Neural Network (CNN) extracts meaningful visual features from the image, and a Long Short-Term Memory (LSTM) network sequentially generates descriptive text based on those extracted features. The second approach employs the pre-trained Salesforce BLIP (Bootstrapping Language-Image Pre-training) model, which combines vision and language understanding using transformer architecture to generate more accurate and context-aware image captions. Either of these models produces a base caption describing the visual content of the uploaded image.

International Journal of
AI Electronics and Nexus Energy
ISSN: 3070-0515 www.zesterapublications.com Original Research Paper

After the initial caption is generated, the caption is passed to the DeepSeek API for post-processing and enhancement. Instead of displaying only a short and literal description, the DeepSeek large language model expands the caption into a richer and more informative description. It improves contextual understanding, generates detailed explanations, extracts important keywords, and produces structured outputs that can be utilized for metadata creation, content indexing, search engine optimization, and accessibility applications. This enhancement significantly improves the overall quality and usefulness of the generated captions.

The system also incorporates a dedicated data storage module for maintaining uploaded images, trained models, tokenizers, and caption history. Uploaded images are stored in the uploads directory, while trained models and tokenizers are saved separately for future inference without requiring repeated training. Additionally, generated captions and related metadata are maintained in a caption history database, enabling efficient retrieval, analysis, and future reference. This modular storage mechanism improves scalability, maintainability, and system performance.

Finally, the enhanced caption is returned to the Flask server and displayed to the user through the web interface. The user can immediately view the generated description along with the enriched contextual information. The entire workflow—from image upload to caption generation and enhancement—occurs seamlessly within a single integrated framework. The architecture combines Python, Flask, TensorFlow, PyTorch, Hugging Face Transformers, Bootstrap, and the DeepSeek API to deliver a scalable, efficient, and intelligent image caption generation system capable of producing high-quality captions for various real-world applications, including digital accessibility, multimedia management, content recommendation, and intelligent image retrieval.

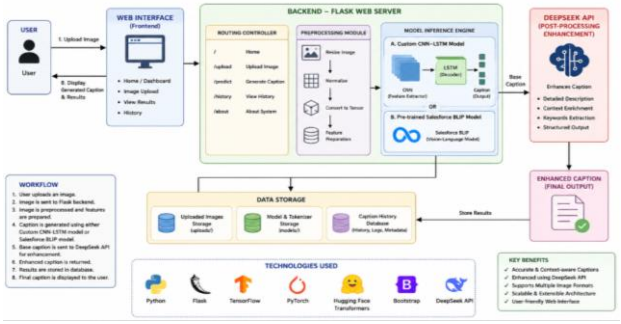

Fig 5.1: System Architecture

VI. IMPLEMENTATION

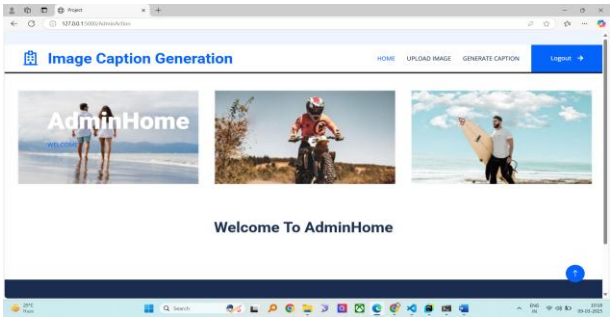

Fig 6.1: Admin Dashboard

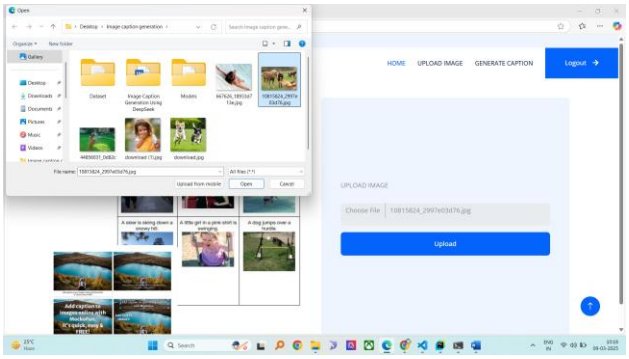

Fig 6.2: Upload Image

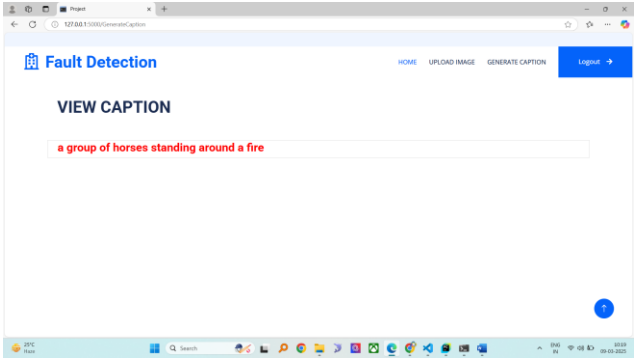

Fig 6.3: Generated Caption

VII. CONCLUSION

The Image Caption Generation Using DeepSeek API project successfully demonstrates an intelligent and efficient framework for automatically generating meaningful textual descriptions from images by integrating computer vision and natural language processing techniques. The system combines a custom CNN-LSTM model with the state-of-the-art Salesforce BLIP vision-language model to produce accurate image captions, while the DeepSeek API further enhances these captions by generating detailed, context-aware descriptions and relevant keywords. The Flask-based web application provides a simple and interactive interface for uploading images and viewing generated results, making the system practical for real-world applications such as digital accessibility, multimedia content management, image indexing, and automated metadata generation. Overall, the proposed architecture offers a scalable, reliable, and user-friendly solution that significantly improves the quality and usefulness of automated image captioning while providing a strong foundation for future advancements in multimodal artificial intelligence.

VIII. FUTURE SCOPE

The future scope of the Image Caption Generation Using DeepSeek API project lies in enhancing its intelligence, scalability, and real-time applicability. The system can be extended by integrating more advanced multimodal foundation models capable of understanding complex scenes, emotions, object interactions, and contextual relationships with greater accuracy. Support for multilingual caption generation and speech synthesis can improve accessibility for users across different languages and for visually impaired individuals. The application can also be expanded to generate captions for videos by incorporating temporal analysis and real-time object tracking. Future enhancements may include cloud-based deployment for large-scale processing, personalized caption generation, emotion and sentiment analysis, automatic hashtag and metadata creation, and integration with smart surveillance, social media platforms, digital libraries, and e-commerce systems. Additionally, optimizing the models for mobile and edge devices while improving inference speed and reducing computational requirements will make the system more efficient, scalable, and suitable for a wider range of real-world applications.

IX. REFERENCES

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
DOI: 10.1109/CVPR.2015.7298935

[2] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," Proceedings of ICML, 2015.
DOI: 10.48550/arXiv.1502.03044

[3] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 664–676, 2017.
DOI: 10.1109/TPAMI.2016.2598339

[4] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning," Proceedings of CVPR, 2017.
DOI: 10.1109/CVPR.2017.324

[5] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question

Answering," Proceedings of CVPR, 2018.
DOI: 10.1109/CVPR.2018.00633

[6] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of NAACL-HLT, 2019.
DOI: 10.48550/arXiv.1810.04805

[7] A. Vaswani et al., "Attention Is All You Need," Advances in Neural Information Processing Systems (NeurIPS), 2017.
DOI: 10.48550/arXiv.1706.03762

[8] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," Proceedings of ICML, 2022.
DOI: 10.48550/arXiv.2201.12086

[9] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," Proceedings of ICML, 2021.
DOI: 10.48550/arXiv.2103.00020

[10] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," Proceedings of ICLR, 2015.
DOI: 10.48550/arXiv.1412.6980

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," Proceedings of CVPR, 2016.
DOI: 10.1109/CVPR.2016.308

[12] O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, 2015.
DOI: 10.1007/s11263-015-0816-y

[13] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," International Conference on Learning Representations (ICLR), 2014.
DOI: 10.48550/arXiv.1312.6114

[14] M. Abadi et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," 2016.
DOI: 10.48550/arXiv.1603.04467

[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," Proceedings of ECCV, 2014.
DOI: 10.1007/978-3-319-10602-1_48