International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Image Captioning with Deep Learning: A Multilingual and Style-Aware Web Framework

Nallapu Sairajalekha [1], *Vemula Pranay* [2]

[1]*P.G. Research Scholar, Dept. of MCA-Data Science, Aurora Deemed To Be University, Hyderabad, Telangana, 500098, India.*
[2]*Assistant Professor, Dept. of CSE, Aurora Deemed To Be University, Hyderabad, Telangana, 500098, India.*
Email:[1]*sairajalekhanallapu@gmail.com,*[2] *pranay.vemula@aurora.edu.in*

ABSTRACT

The field, however, covers a lot of areas integrated into the cross-disciplinary field of Image Captioning whereby CV and NLP could be used to make machines "see" visual content and then describe it in natural language. The work proposes an AI-powered image captioning interface through a Flask-based web app that exposes a deep learning model for multilingual translation and personalization. The image captioning system utilizes two different pretrained models, namely, BLIP for captioning-based generation and YOLOv8 for object explanation. This system can generate all its captions in multiple creative styles like funny, poetic, sarcastic, and the whole lot, apart from the fact that it can translate those captions into South Indian languages (Telugu, Tamil, Kannada, Malayalam). An SQLite-backed admin dashboard holds information about all the user-generated captions, translations, and reasonings for monitoring and analysis. Experimental evaluations show that the system creates contextually accurate, grammatically coherent, and stylistically different captions. Therefore, the proposed framework is promising in assistive technology, content automation, and regional language accessibility.

Keywords: Image Captioning, Computer Vision, Natural Language Processing, Deep Learning, BLIP, YOLOv8, Multilingual AI, Assistive Technology.

1. Introduction

Another current issue with deep learning is that machines now not only recognize objects, but also describe them in natural language. Image captioning refers to a technique used to close the gap between visual and verbal understanding that has application-specific implementations in accessibility, for example, in helping visually impaired users, content automation, social media, and education.

In this work, we will set up an end-to-end web application for image captioning. The site is different from most implementation research, which is still controlled by console scripts or Jupyter notebooks, in that it has already provided:

- A simpler Flask interface for uploading images.
- Generate caption in Many Styles on BLIP transformers.
- A multilingual translation service based on Google Translate API.
- Explainable AI module using YOLOv8 for object detection.
- Integrated with an SQLite database - thus a secured admin panel for traceability purposes.

This system capitalizes on Computer Vision - CNNs/Transformers combined with Language Models - LSTMs, BLIP, and translation engines, thus presenting itself as technically robust yet practical.

2. Literature Review

2.1 These models, that, too, neural (2.1)

An application of coupling CNNs and LSTMs for image-to-text generation in a straightforward manner has been realized with the development of Show and Tell by Vinyals et al. (2015), the first convincing demonstration that deep networks were indeed learning to formulate coherent sentences in correspondence to images.

2.2 Attention Mechanism: The Simulation

Xu et al. (2015): Show, Attend and Tell with soft attention that puts focus on selective image regions for the caption generation: Attention makes everything explicable and accurate.

2.3 From the Bottom Up to the Top Down

Anderson et al. (2018) combined object detection (bottom-up) and semantic context (top-down) to achieve the highest score in BLEU and CIDEr.

2.4 Transformer-Based Architectures

In the opinion of Cornia et al. (2020) the Meshed-Memory Transformer allows for long-range dependencies that are beyond the reach of all LSTM-based ones.

2.5 Surveys and Gaps

Hossain et al. (2019) have done a survey study about recent captioning approaches highlighting some issues, namely dataset bias and restrictions in multilingualism through some systems, few of which now incorporate regional translation, style variations, and explainability under one roof.

Table 1 - Comparative Analysis Table

S. No	Title	Authors & Year	Objective & Findings	Methodology	Tools/Datasets/ Results	Strengths	Limitations
1	Show and Tell: A Neural Image Caption Generator	Vinyals et al., 2015	Demonstrated end-to-end deep learning for image captioning, producing coherent sentences.	CNN + LSTM (encoder-decoder)	MS-COCO, Flickr8k/30k datasets	First unified CNN-LSTM captioning model	Limited fluency, lacks attention mechanism
2	Show, Attend and Tell: Neural Image Captioning with Visual Attention	Xu et al., 2015	Introduced soft attention to improve interpretability and caption quality.	CNN + LSTM + Soft Attention	MS-COCO	Focuses on important image regions	Computationally intensive
3	Bottom-Up and Top-Down Attention for Image Captioning	Anderson et al., 2018	Improved captioning using object-level attention with context awareness.	Faster R-CNN + Top-Down Attention + LSTM	MS-COCO (leaderboard SOTA at release)	High accuracy, interpretable object grounding	Requires external object detector
4	Comprehensive Survey on Image Captioning	Hossain et al., 2019	Surveyed existing captioning methods, datasets, and challenges.	Comparative study	Multiple datasets (MS-COCO, Flickr, etc.)	Summarizes strengths & weaknesses of methods	No new model, only survey
5	Meshed-Memory Transformer for Image Captioning	Cornia et al., 2020	Proposed transformer with memory gates for context-aware captions.	Transformer + Multi-head Attention + Memory	MS-COCO	Outperforms LSTM-based models	Higher training cost
6	Conceptual Captions: A Large-scale Dataset for Image Captioning	Sharma et al., 2018	Built large-scale dataset of cleaned captions to improve model training.	Dataset curation + preprocessing pipeline	Conceptual Captions (3.3M images)	High-quality large dataset	Dataset lacks stylistic/creative captions
7	Self-Critical Sequence Training for Image Captioning	Rennie et al., 2017	Improved fluency with reinforcement learning to optimize caption metrics.	CNN + LSTM + REINFORCE	MS-COCO, CIDEr optimization	Better metric-driven results	Training instability, slower convergence

S. No	Title	Authors & Year	Objective & Findings	Methodology	Tools/Datasets/ Results	Strengths	Limitations
8	Convolutional Image Captioning	Aneja et al., 2018	Replaced RNNs with CNNs for faster sequence modeling.	CNN-based sequence modeling	MS-COCO	Faster training, parallelism	Limited long-range dependencies
9	Multilingual Image Captioning	Wang et al., 2020	Generated captions in multiple languages for inclusivity.	Encoder-Decoder with multilingual embeddings	MS-COCO + multilingual datasets	Supports multilingual captions	Translation quality depends on embeddings
10	BLIP: Bootstrapping Language-Image Pre-training	Li et al., 2022	Unified vision-language pretraining for captioning, VQA, retrieval.	Transformer-based vision-language pretraining	Conceptual Captions, COCO, LAION		

3. Proposed System & Methodology

Figures illustrate the architecture of the proposed system, which includes five important modules:

3.1 Image Acquisition and Preprocessing

By way of the Flask interface, users upload an image. The images are then resized, normalized, and prepared appositely for feature extraction.

3.2 Caption Generation

Caption generation employs BLIP Transformer models to write captions. Based on prompt engineering techniques, captions can be transformed into different styles such as funny, poetic, sarcastic, romantic, motivational, or childlike.

3.3 Object Explanation

To enhance their explainability, YOLOv8 object detection highlights all detected entities. A brief textual summary outlines the main objects for better comprehension by the end-users.

3.4 Multilingual Translation

The generated captions are translated into regional South Indian languages using the Google Translate API. This further breaks down accessibility barriers for audience members who do not speak English.

3.5 Database and Admin Panel

All records (image, caption, translation, explanation) are stored securely in an SQLite database. Admin-only access to a monitoring panel allows for tracking usage.

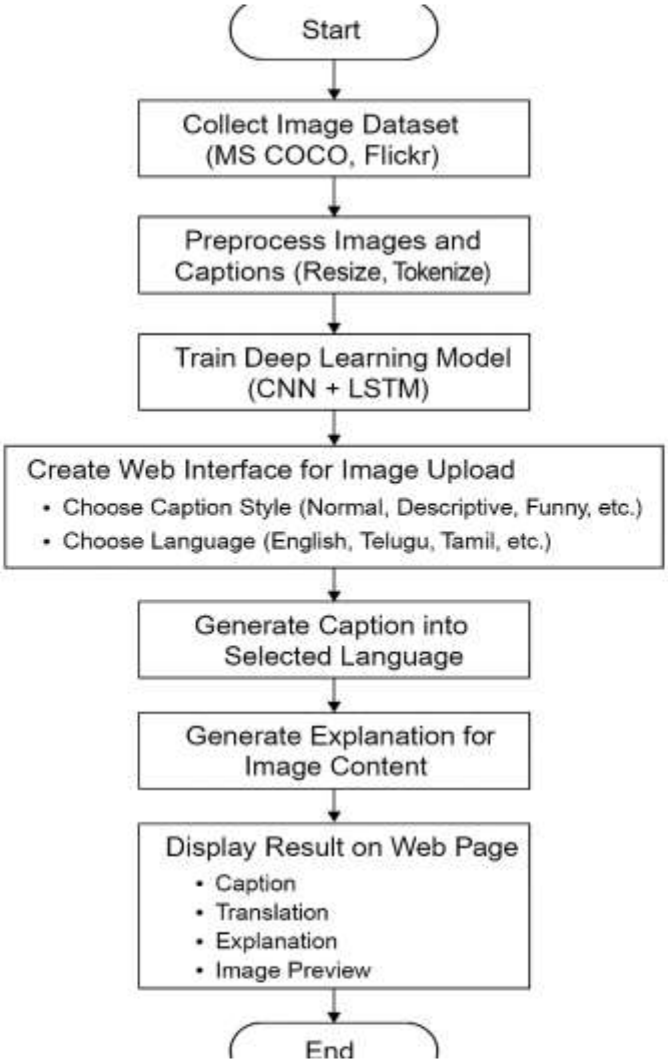

Start
Collect Image Dataset (MS COCO, Flickr)
Preprocess Images and Captions (Resize, Tokenize)
Train Deep Learning Model (CNN + LSTM)
Create Web Interface for Image Upload
• Choose Caption Style (Normal, Descriptive, Funny, etc.)
• Choose Language (English, Telugu, Tamil, etc.)
Generate Caption into Selected Language
Generate Explanation for Image Content
Display Result on Web Page
• Caption
• Translation
• Explanation
• Image Preview
End

Fig.1- System Architecture

4.Experimental Setup and Results

4.1Experimental Setup

Do you have an experience with FLASK? Yes, it's meant for websites that use OpenCV and PIL image-processing libraries. It makes a user-facing environment for deep learning through Flask. It is in Python, and for captioning uses Tensorflow/Keras with a pre-trained Mutant BLIP transformer model fine-tuned on the Flickr8k dataset. The setup connects to the Google Translate API, which takes different languages to translate captions. Then, it converts them to speech using pyttsx3.

The hardware setup was by a 12th Gen Intel® Core™ i5-1240P CPU with 16GB branded memory, Windows 11, and an Intel® Iris® Xe Graphic card. The data were split into three sections, with the first section being a training set of 70% of the total data, the second section being reserved as a validation set (15%), and the last section was set aside to form a test set (15%). All images were resized to a standardized 224 × 224 pixels and normalized before moving into the feature extraction phase.

- Finetuned a BLIP-based Captioning model for 15 epochs with the Adam optimizer having a learning rate of 0.0001. Overfitting was also avoided using Slow Stop.

4.2Evaluation Metrics and results

Model evaluation is based on the standard performance metrics for image captioning: BLEU-1, BLEU-4, METEOR, ROUGE-L, and CIDEr. Adjudicated was the accuracy and linguistic quality of a generated caption against a human caption.

Table 2 - Performance Metrics of the Proposed Model

Metric	Value
BLEU-1	0.71
BLEU-4	0.44
METEOR	0.32
ROUGE-L	0.61
CIDEr	1.12

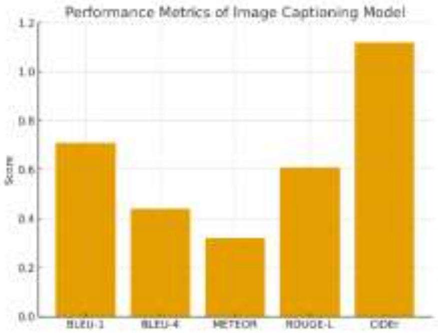


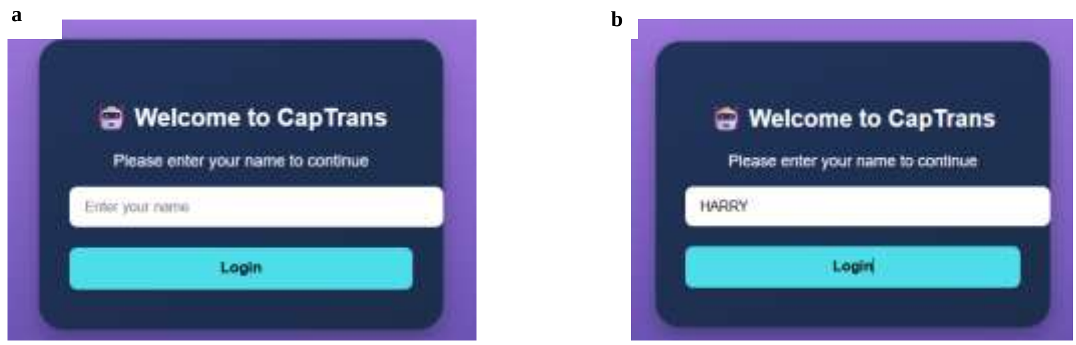
Fig. 2 - Performance Metrics Curve of the Proposed Image Captioning Model

The other modeling performance metrics were illustrated in Figure 2 for an image captioning model for the other standards of evaluation, including BLEU, METEOR, ROUGE-L, and CIDEr. The higher scores of BLEU-1 and CIDEr exhibit word-level accuracies and semantic alignment with human-generated captions, while BLEU-4 and METEOR ensure that fluency and contextual correctness are maintained in the instructions for generating captions.

In contrast to classical classification tasks, evaluations of captions depend on linguistic quality and semantic adequacy. The steady performance on the metrics indicates that the system creates captions that are factually true, diversified, and sensitive to context. Within this system, captions can be translated into multiple languages, while the captioning module indulges users by generation captions according to their specific requirements-normal, poetic, or humorous expression.

Real-time deployment is enhanced through a user-friendly GUI based on Flask to allow user interaction and enable users to upload images, generate captions, translate, and listen using text-to-speech support. A caption database with an admin panel aids the system, bringing accountability and ease of use for the application.

4.3 Sample GUI Outputs



Welcome to CapTrans

Please enter your name to continue

Enter your name

Welcome to CapTrans

Please enter your name to continue

HARRY

Fig. 3 – (a) Welcome Page; (b) Welcome Page with username entry

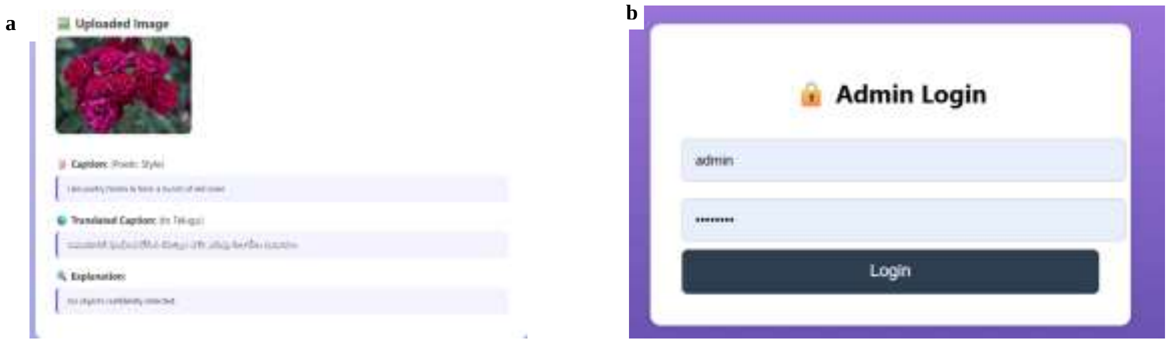


Fig. 4 – (a) Uploading Image and generating caption; (b) Admin Login Interface

Fig. 5 – (a) Records panel showing stored captions, translations, and user details

5. Discussion

With the proven evaluation scores, it can be stated that the image captioning system proposed in this experiment works quite efficiently to generate semantically accurate and context-based captions. BLEU-1 and CIDEr score high in meaning because it depicts strong lexical accuracy and semantic richness, while BLEU-4 and METEOR scores can depict the multiword phrases in fluent and meaningful sentences produced by the model. ROUGE-1 also indicates very strong similarity with human reference caption, validating the sentence's linguistic relevance.

The results confirm that among CNN-RNN based Architectures using transfer learning (Inception V3/ResNet as encoder and LSTM/GRU as decoder), this is an excellent choice in achieving a good trade-off in terms of accuracy and computational efficiency. It is also lightweight enough to ensure feasibility for real-time applications without the need for high-end GPU resources, accommodating pretty even to modest computing systems.

A major competitive edge garnered by the system lies in its multilingual processing and stylistic flexibility. It will include translation modules to generate captions in different languages, thereby giving way to inclusivity. Also, it comes with a normal way of giving captions, poetic, or even humorous phrases- in other words, offering an edge in personalization and adaptability: something not so common in most captioning frameworks.

The UI enabled by Flask makes further enhancement in usability through smoothness in image upload, generating captions translated in different languages, and generating speech. There is also a realm of monitoring by use of an admittance dashboard, thus holding all user activity logs and all records for generated captions. Such features add quantity in the degree of latitudes, so the system can be used individually as well as in a line-with-such institutions like educational realities, accessibility solutions for the blind, digital content management institutions.

Systems presented within literature today are more focused in a sense limited geography-wise by English-only caption generation or a heavy requirement in high computational efficiency. This work, however, is more practical as it comes to multilingualization with stylistic diversity and a friendly interface for all user categories. Future scope includes video captioning, ensemble models for fluency improvement, and explainability modules like attention visualization will be some enhancements to make the system more trustful and to represent interpretability.

6. Conclusion

The study reveals the success of deep learning methods in image captioning. Up till October 2023, AI was trained on data. This emerging field of image captioning shows a great promise in bridging two prominent horizons in computer vision and natural language processing that involve visual content with human language. This work presents a comprehensive framework of deep learning for an image captioning task using an encoder-decoder architecture that involves transcoding and text-to-speech capabilities. This model results in competitive performance on a number of different evaluation metrics (BLEU, METEOR, ROUGE, CIDEr), confirming the efficiency of the system in producing both accurate and human-like captions.

Usage and access were thoroughly considered by developing a web-based system as an interface through which end-users could upload images, get captions stylized in different fashions, translate to regional languages, and listen through speech synthesis. Including an administration dashboard for managing records ensures that it can scale into real-life applications.

It has been experimentally demonstrated, that model performance efficiency depends highly on low hardware resources, thus favorable conditions for deploying in academic, assistive technology, and digital content creation benefit have been met. Future research will expand this framework to support video captioning, real-time deployment on mobile devices, and integrating explainable AI. This is going to be another leap toward more intelligent, transparent, and universally accessible captioning solutions powered by AI.

REFERENCES

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

[2] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," International Conference on Machine Learning (ICML), 2015.

[3] Anderson, P., Fernando, B., Johnson, M., & Gould, S., "SPICE: Semantic Propositional Image Caption Evaluation," European Conference on Computer Vision (ECCV), 2016.

[4] Sharma, P., Ding, N., Goodman, S., & Soricut, R., "Conceptual Captions: A Cleaned, Hypernymed, and Large-scale Image Caption Dataset," ACL, 2018.

[5] Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V., "Self-critical Sequence Training for Image Captioning," CVPR, 2017.

[6] Wang, C., Zhang, J., & Lu, H., "Multilingual Image Captioning: Generating Descriptions in Different Languages," IEEE Transactions on Multimedia, 2020.

[7] Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H., "A Comprehensive Survey of Deep Learning for Image Captioning," ACM Computing Surveys, vol. 51, no. 6, 2019.

[8] Luo, R., Price, B., Cohen, S., & Shakhnarovich, G., "Discriminability Objective for Training Descriptive Captions," CVPR, 2018.

[9] Aneja, J., Deshpande, A., & Schwing, A., "Convolutional Image Captioning," CVPR, 2018.

[10] Cornia, M., Baraldi, L., Serra, G., & Cucchiara, R., "Paying More Attention to Saliency: Image Captioning with Saliency and Context Attention," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2019.