**Asian And Pacific Economic Review**
**ISSN: 1000-6052**

# IMAGE CAPTION GENERATION SYSTEM FOR LOW-RESOURCE TELUGU LANGUAGE

**Ravi Gedam**
Department of CSE, GH Raisoni University, Saikheda, India
**Pushpa Birha,**
Department of CSE
G.H.Raisoni College of Engineering and Management
Nagpur, India
**Pragati Agrawal**
Department of Data Science
G.H.Raisoni College of Engineering and Management
Nagpur, India
**Mr.Saurabh Vyas,**
Department of Computer Science and Engineering (Cyber Security)
G.H.Raisoni College of Engineering and Management
Nagpur, India
**Abhishek Kundu**
G H Raisoni University Saikheda, India

***Abstract*—This paper introduces an innovative approach to generating image captions for the low-resource Telugu language, an Indian language that is often overlooked. We address the scarcity of resources for the Telugu dataset by adapting the Flickr8K dataset through Google Translator, resulting in a custom Telugu dataset. Despite Telugu's cultural and regional significance, its limited resources pose challenges in advancing natural language processing, specifically in the domain of image caption generation. To accomplish the goal of caption generation, we employ the encoder-decoder framework by combining convolutional neural networks and recurrent neural networks. The experimentation involves testing on the designed TICD dataset and demonstrates the notable performance in terms of BLUE and ROUGE scores.**

***Index Terms*—Image Caption Generation, Low-resource Language, Convolutional Neural Network, Flickr8K, Attention.**

## I. Introduction

In the ever-evolving domain of computer vision [1], the generation of descriptive captions for images has emerged as a critical area of research, significantly enhancing the capacity of machines to comprehend and communicate visual content [2]. However, the majority of these advancements have predominantly focused on languages rich in linguistic resources [3] such as English, German, Hindi, etc, often neglecting the complexity associated with low-resource languages [3] such as Telugu, Tamil, Assamese, etc. High-resource languages, characterized by large linguistic datasets and comprehensive research attention, have been at the forefront of image caption generation tasks. In contrast, low-resource languages, with their scarcity of linguistic data and limited exploration, present unique challenges that demand specialized solutions. This paper bridges this gap by introducing a pioneering approach to image caption generation, specifically tailored for the Telugu language, a low-resource language of South India. The term high resource language denotes languages

well-endowed with linguistic resources, extensive datasets, and robust research infrastructure. These languages have served as the primary focus for the development of state- of-the-art image caption generation systems. On the other hand, low-resource languages encompass linguistic entities that lack comprehensive datasets and research attention. Telugu, the official language of the Indian states of Andhra Pradesh and Telangana, is spoken by around 75-80 million people globally. It stands as a quintessential example of a low-resource language, despite its immense cultural and regional significance in these states. Despite its preva- lence, the scarcity of comprehensive linguistic datasets has hindered progress in natural language processing tasks, particularly in the domain of image caption generation. This paper strives to address this void by presenting a meticulously crafted dataset specifically designed for Telugu, achieved through the translation of the widely uti- lized Flickr8K dataset utilizing the capabilities of Google Translator. The choice of Telugu as the focal point of this research is driven not only by its cultural signifi- cance but also by its underrepresentation in the existing body of research. Through this endeavor, we aim to shed light on the challenges and opportunities associated with low-resource languages, particularly in the South Indian linguistic landscape. In the subsequent sections, we discussed the intricacies of the dataset preparation pro- cess, illuminating the translation methodology employed to adapt the Flickr8K dataset for Telugu. Furthermore, we introduce a sophisticated prototype system leveraging the encoder-decoder framework, combining convolutional neural networks and recurrent neural network networks for image caption generation in the context of the Telugu language. II. Related Work

Numerous articles discuss creating captions for images, but there's limited research on this topic for the Tel- ugu language. For instance, Rohan and colleagues [4] developed an advanced method using a combination of InceptionV3, VGG16, and ResNet50 convolutional neural networks along with a transformer architecture that uses multihead attention. This helps the model understand the context of images, handle language complexities, and establish important connections between visual and text elements. They tested the model on translated ver- sions of well-known datasets like Flickr8k, Flickr30k, and MSCOCO, showing impressive results, especially in terms of BLEU metrics. In Tamil and Telugu caption generation tasks, the model achieved a high BLEU-1 score of 65.16 and 66.79, respectively. Addressing the lack of a system for generating image captions in Assamese, Nath and team [5] present a significant challenge for AI-NLP re- searchers. They used an encoder-decoder framework with Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Their evaluation utilized the English datasets Flickr30k and Coco Captions, marking a crucial early step in this research area. Since there was no existing collection of descriptions for pictures in the Myanmar language, Aung and colleagues [6] took the initiative to create one manually. They expanded the Flickr8k dataset to form a Myanmar image caption dataset. Their method involved using a generative merge model that combined convolutional neural networks and Long-Short Term Memory (LSTM) specifically designed for creating captions in Myanmar. To evaluate how well their approach worked, they compared two traditional models for extracting features: VGG OxfordNet 16-layer and 19-layer. They used BLEU scores and 10-fold cross- validation on the Myanmar image caption dataset for per- formance assessment. Kinghorn and team [7] introduced an inventive deep learning framework that focuses on regions to generate descriptions for images. The structure integrates a detector for regional objects, prediction of attributes through a recurrent neural network, and a language generator with dual RNNs for encoding and de- coding. Their method concentrates on a localized strategy, aiming to improve existing methods by focusing on specific regions of images that contain people and objects. Rig- orous evaluation using the IAPR TC-12 dataset showed remarkable performance, surpassing state-of-the-art tech- niques across various evaluation metrics. Fei and col- laborators [8] proposed an improved framework for Non- autoregressive (NA) prediction to speed up the process of generating captions for images. Their decoding module

includes a positional alignment mechanism for arranging words that describe the identified content in the given image. Experimental results on publicly available datasets demonstrated that their model outperforms conventional NA captioning models. It showed superior performance while achieving comparable results to autoregressive image captioning models but with a notable increase in speed. Parikh and colleagues [9] suggest two different ways to figure out what's in a picture using computer vision and turn that understanding into sentences that make sense. They introduce a model that combines a convolutional neural network and a Gated Recurrent Unit (GRU) to create accurate image captions. The results of experi- ments on different sets of data, including using the BLEU evaluation method on the MS-COCO 2017 dataset, show that the suggested model works well, with a BLEU-4 score of 53.5. Dash and team [10] stress the importance of recognizing the main subject of an image. They propose a new way of creating captions using a special kind of computer model (encoder-decoder framework) based on deep learning. Their method is compared to other top models using the MSCOCO 2017 training dataset, and it performs better in generating captions, as seen through BLEU, CIDEr, ROGUE-L, and METEOR scores. Zhang and collaborators [11] build upon the same idea but make improvements by using a more powerful decoder and smoothly combining spatial and adaptive attention into the Transformer architecture. They replace the usual LSTM decoder with a Transformer to speed up the train- ing process, and this leads to significant improvements in results, as shown in experiments using the Flickr30k dataset.

## III. Dataset

The Telugu Image Captioning Dataset (TICD) is in- troduced as a specialized resource for advancing image caption generation in the Telugu language, an under- represented South Indian language. Comprising 8,000 images sourced from Flickr8K, each accompanied by mul- tiple Telugu captions translated via Google Translator, TICD captures the linguistic nuances and cultural context specific to Telugu speakers. The dataset, encompassing variations in grammar and syntax, serves as a valuable tool for researchers to train and evaluate models, fostering the development of culturally sensitive natural language processing technologies for Telugu. Freely available for academic use, TICD contributes to the inclusivity of artificial intelligence in linguistically diverse environments, emphasizing the significance of linguistic diversity in ad- vancing technology.

## IV. System Architecture

The design of the suggested system architecture is il- lustrated in Figure 1, mainly depending on the encoder- decoder framework. In this proposed model, convolutional neural networks encode image features, while recurrent neural networks encode the image captions represented as word sequences. Subsequently, the encoded image un- dergoes processing in a text feature decoder, sequentially forecasting the caption. During each word generation in the caption, the model employs attention to emphasize the utmost significant aspects of the image.

### *A. Image Features Encoder*

We employ transfer learning to preprocess raw files, utilizing a pre-trained CNN-based system. This system, initially trained, processes images, generating encoded im- age vectors that encapsulate the crucial features. Herein, for extracting image features, we opted for pre-trained models, specifically VGG16 [12] and EfficientNetB3 [13], both trained on the ImageNet dataset. VGG16, with 16 layers, is simpler compared to EfficientNetB3, which is not only easier to optimize but also exhibits enhanced per- formance with increasing network depth. We exclusively utilized the encoded image features from

hidden layers, discarding the final output layer of the pre-trained models, as it contains classification results.

*B. Word Sequences Encoder*

We utilize Tensorflow to tokenize our sentences, ex- tracting tokens from the top 25000 words. These tokens undergo processing through an Embedding layer with an embedding size of 256, followed by an RNN utilizing Gated Recurrent Units. The GRU, introduced by Kyunghyun Cho et al. [14], has proven successful in applications such as machine translation and sequence generation. The GRU represents an advancement over the traditional Recurrent Neural Network. To address the vanishing gradient problem, standard RNNs employ update and reset gates, dictating the cell's behavior. In the GRU model, the memory cell retains information from each time step, reflecting the observed input. The update and reset gates act as vectors directing specific information to the output, designed to preserve input from earlier time steps and discard irrelevant data for forecasting. The GRU differs from Long Short-Term Memory [15] primarily in the number of gates. GRU features two gates, reset and update, while LSTM incorporates three gates: input, output, and forget. Due to its fewer gates, GRU is less complex than LSTM, exhibiting a 29.29% faster model training speed on the same dataset. In terms of results, GRU surpasses LSTM in scenarios involving long text and relatively small data sources but falls short in other cases.

*C. Attention Mechanism*

In our experiments, we implemented Bahdanau Atten- tion, as detailed in the cited research articles [16]. For VGG16, the configuration of attention features amounts to 49, while for EfficientNetB3, it reaches 100. The attention layer receives the context vector produced by the last hidden layer of the pre-trained model. Following this, the GRU employs the context vector as input to generate an image description, demonstrating superior performance to traditional CNN and RNN architectures using Long Short-Term Memory as a decoder. During each keyframe, the Attention module is provided with the encoded image and the hidden state from the Decoder in the preceding timesteps. It calculates an attention score, assigning weights to each pixel in the encoded image. Pixels with higher weights indicate a greater likelihood of being part of the output word at the subsequent timestep. For instance, when forming the phrase "ఒక బాలుడు బంతిని తనుస్తునాన్ు (A boy is kicking the ball)", the pixels corresponding to the boy are accentuated during the generation of the word ``అబాబ్ (boy)", while those related to the ball are highlighted for the word "బంతి (ball)". Attention is a mechanism focused on a specific aspect of information while disregarding other perceptible details. It serves as a guide for the model, indicating where to concentrate to generate the appropriate word rather than considering the entire image. The decoder, relying on the hidden state, directs attention to particular regions of the image at time t, employing spatial image features to compute the context vector.

*D. Caption Generation*

The caption generation system represents a sophisti- cated architecture designed to generate descriptive cap- tions for images. At its core, the system utilizes a decoder, a critical component equipped with a Dense layer, and Rectified Linear Unit (ReLU) activation. This design aims to enhance the system's ability to decipher and interpret visual features embedded in images. The inclusion of a Dense layer is instrumental, providing a crucial mechanism for processing and transforming the encoded visual information. Elaborately interlinked into the system is the incorporation of attention weights within the Dense layer. These attention weights play a pivotal role in determining the significance of different parts of the input, allowing the model to focus on

specific regions of the image. This attention mechanism ensures that the system can effectively capture and utilize relevant visual prompts when generating captions.

The intricate data begins with the picture feature en- coder, a component responsible for extracting essential features from the input images. The output from this encoder serves as the foundation for the subsequent stages of caption generation. The Dense layer, enriched with at- tention weights, takes this output and generates a softmax prediction for the next word in the sequence. The predic- tion process is meticulous, involving consideration of each word in the vocabulary. The system selects the word with the highest probability, shaping the unfolding narrative in the caption. This mechanism is crucial for generating coherent and contextually relevant descriptions, ensuring that the generated captions align with the content of the images. A noteworthy departure from conventional image processing approaches lies in the utilization of encoded attributes instead of raw images in the image caption algo- rithm. This strategic shift enhances efficiency and enables the system to encapsulate essential visual information in a condensed form.

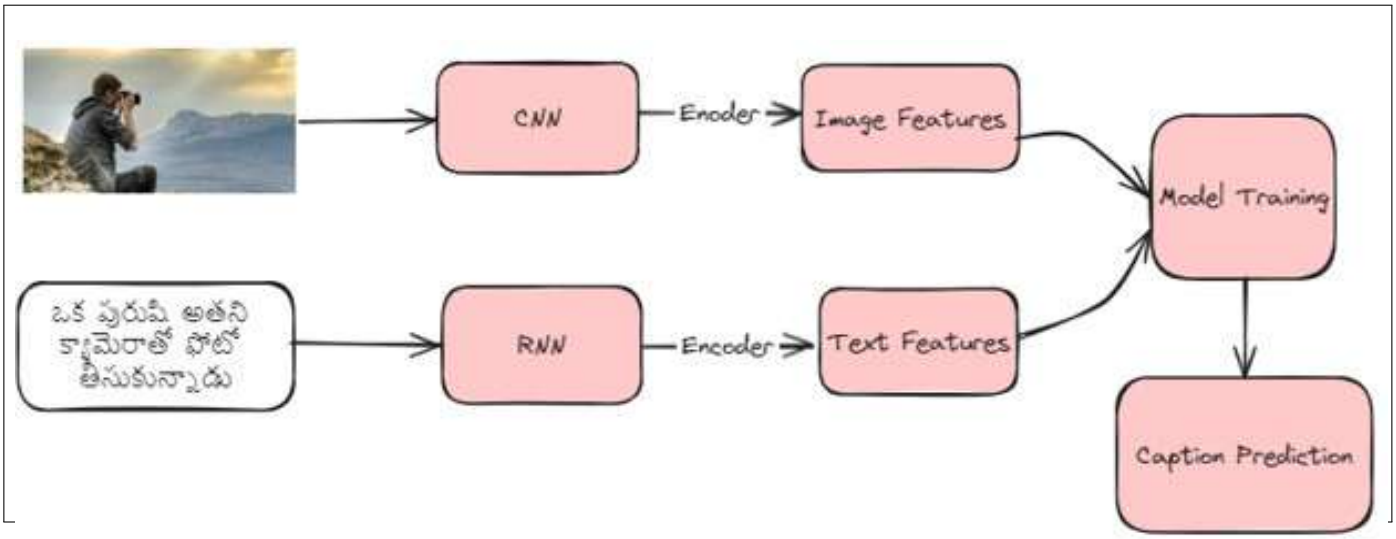


Figure 1. Workflow of the proposed system architecture

Simultaneously, the system is fed with corresponding target captions for each encoded image during the training phase. This process employs the Teacher Forcing method, a technique where the target word becomes the subse- quent input to the decoder for predicting the next word. Through iterative cycles, the system refines its predictive capabilities, learning to associate visual features with linguistic elements. This comprehensive approach results in the generation of captions that not only complement the intended description but also exhibit a meticulous understanding of the visual content. The system, by decoding visual information and leveraging the interplay between the encoder and decoder, achieves a synthesis of visual and linguistic elements. The final token or caption is a testament to the system's prowess in seamlessly merging these components, offering a glimpse into the evolving area of advanced image captioning systems.

## V. Experimental Results

Assessing the outcomes of the proposed methodology for the Telugu language proves to be a fine task, demanding scrutiny from both quantitative and qualitative perspec- tives. The systems yielded

results showcase noteworthy metrics, setting a significant precedent for other exist- ing frameworks in the field of Telugu language caption generation. Various evaluative criteria are applied in the context of image captioning tasks, as expounded in the literature. The widely acknowledged BLEU score [17] assumes eminence among these metrics, alongside the commonly employed Rouge score [14], utilized to appraise the image caption generation systems efficacy by juxtaposing its captions with a set of reference summaries. The evaluative procedure entails subjecting the model to testing on the Flickr30k and COCO 2017 datasets, encompassing 2000 and 5000 test images, respectively, to comprehensively assess the proposed model's effectiveness.

Rigorous recording of BLEU scores for the test dataset is conducted. Subsequent experiments unfold on our amal- gamated dataset, comprising 150k images in the training set and 7000 test images in the test set. The scrupulous monitoring and presentation of both BLEU and ROUGE scores are methodically documented in the accompanying Table I.

Table I
Experimental Results of the Proposed Approach

| Model | BLEU | ROUGE |
|---|---|---|
| VGG16 | 0.3012 | 0.19 |
| Efficient | 0.3219 | 0.20 |

The dataset primarily consists of images featuring hu- man subjects, and their captions are remarkably similar. However, the model, having been trained extensively on these similar human subjects, struggles to accurately de- scribe and distinguish non-human subjects during testing. Additionally, translating compound English sentences to Telugu faces limitations. The anticipated improvement in accuracy through a combined system, measured by BLEU and rouge scores, yielded average results.. Nevertheless, this outcome provides an opportunity for future research to enhance results after merging datasets. Some challenges may be addressed by employing multi-head attention, allowing the model to focus on more than one region. In the field of image captioning, transformers face hurdles due to variations in data processing order and the attention mechanism's ability to provide context for any position in the input sequence.

## VI. Conclusion and Future Scope

Conclusion and Future Scope In this manuscript, we introduced an encoder-decoder framework designed for the generation of image captions in the Telugu language. Our approach involved a thorough examination of the Telugu alphabet, considering its representation in the Devanagari script and drawing connections to the other low-resource Indian languages. The current model is constructed upon distinct language and visual understand- ing components. Although our primary emphasis was on translating English sentences, our overarching goal is to develop a comprehensive Gold Dataset specifically for Telugu image captions. Given the growing popularity of Transformer-based frameworks in various NLP tasks, we foresee that exploring image captioning with Attention-based Transformers could be a valuable avenue for future experimentation.

References

[1] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopa- padakis, "Deep learning for computer vision: A brief review," *Computational intel ligence and neuroscience*, vol. 2018, no. 1, p. 7068349, 2018.

[2] H. Wang, Y. Zhang, and X. Yu, "An overview of image caption generation methods," *Computational intel ligence and neuro- science*, vol. 2020, no. 1, p. 3062706, 2020.

[3] S. Ranathunga, E.-S. A. Lee, M. Prifti Skenduli, R. Shekhar, M. Alam, and R. Kaur, "Neural machine translation for low- resource languages: A survey," *ACM Computing Surveys*, vol. 55, no. 11, pp. 1–37, 2023.

[4] R. Reddy, S. Gunti, S. Sridevi *et al.*, "Multilingual image cap- tioning: multimodal framework for bridging visual and linguistic realms in tamil and telugu through transformers," 2023.

[5] P. Nath, P. K. Adhikary, P. Dadure, P. Pakray, R. Manna, and S. Bandyopadhyay, "Image caption generation for low-resource assamese language," in *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)*, 2022, pp. 263–272.

[6] W. P. Pa, T. L. Nwe *et al.*, "Automatic myanmar image caption- ing using cnn and lstm-based language model," in *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Col laboration and Computing for Under-Resourced Languages (CCURL)*, 2020, pp. 139–143.

[7] P. Kinghorn, L. Zhang, and L. Shao, "A region-based image caption generator with refined descriptions," *Neurocomputing*, vol. 272, pp. 416–424, 2018.

[8] Z.-c. Fei, "Fast image caption generation with position align- ment," *arXiv preprint arXiv:1912.06365*, 2019.

[9] H. Parikh, H. Sawant, B. Parmar, R. Shah, S. Chapaneri, and D. Jayaswal, "Encoder-decoder architecture for image caption generation," in *2020 3rd International Conference on Commu- nication System, Computing and IT Applications (CSCITA)*. IEEE, 2020, pp. 174–179.

[10] S. K. Dash, S. Acharya, P. Pakray, R. Das, and A. Gelbukh, "Topic-based image caption generation," *Arabian Journal for Science and Engineering*, vol. 45, no. 4, pp. 3025–3034, 2020.

[11] W. Zhang, W. Nie, X. Li, and Y. Yu, "Image caption generation with adaptive transformer," in *2019 34rd youth academic annual conference of Chinese association of automation (YAC)*. IEEE, 2019, pp. 521–526.

[12] R. Mulyawan, A. Sunyoto, and A. H. M. Muhammad, "Pre- trained cnn architecture analysis for transformer-based indone- sian image caption generation model," *JOIV: International Journal on Informatics Visualization*, vol. 7, no. 2, pp. 487–493, 2023.

[13] M. H. Rodoshi, M. U. Ahmed, M. S. Ashraf, M. G. H. Mim, and A. Khanam, "Automated image caption generator in bangla using multimodal learning," Ph.D. dissertation, Brac University,

2023.

[14] T. H. Nguyen, K. Cho, and R. Grishman, "Joint event extrac- tion via recurrent neural networks," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016, pp. 300–309.

[15] H. Al Fatta, U. Fajar *et al.*, "Captioning image using con- volutional neural network (cnn) and long-short term memory (lstm)," in *2019 International Seminar on Research of Infor- mation Technology and Intel ligent Systems (ISRITI)*. IEEE, 2019, pp. 263–268.

[16] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine trans- lation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[17] C. Callison-Burch, M. Osborne, and P. Koehn, "Re-evaluating the role of bleu in machine translation research," in *11th con- ference of the european chapter of the association for computa- tional linguistics*, 2006, pp. 249–256.