International Journal on Engineering Technology (InJET)

Volume 3, issue no. 1, Dec 2025, Page 271 - 278
Received Date: Aug 9, 2025
Accepted Date: Oct 26, 2025
ISSN: 3021-940X (Print)
DOI: https://doi.org/10.3126/injet.v3i1.87026

Automatic Image Captioning for Nepalese Socio-Cultural and Traditional Images Using CNN and RNN

Shayak Raj Giri[1], Sharad Kumar Ghimire[2*]

[1] *Department of Electronics and Computer Engineering, Pulchowk Campus, Tribhuvan University, Nepal*
email: shayakraj@ioe.edu.np
[2] *Department of Electronics and Computer Engineering, Pulchowk Campus, Tribhuvan University, Nepal*
email: skghimire@ioe.edu.np

Abstract

Image captioning is the process of generating textual description of an image. Automatically describing the content of images using natural language is a challenging task. Despite significant progress using deep learning architectures, most existing datasets and models are biased toward western cultural contexts, limiting their general applicability. This paper presents an approach for automatic image captioning focused on Nepalese socio-cultural and traditional contexts using Convolutional Neural Network (CNN) as an encoder and Long Short-Term Memory (LSTM) network as a decoder. A custom dataset consisting of 412 images with 1236 corresponding captions was developed, capturing local customs, festivals, and daily life. The model was evaluated on both the custom dataset and the standard Flickr8k dataset using Bilingual Evaluation Understudy BLEU-1 to BLEU-4 scores. The obtained accuracy was 90.421%, loss 3.4614%, BLEU-1 score 0.580268 and BLEU-4 score 0.300523. Same model was fitted with own dataset and achieved accuracy was 90.3519% with loss of 3.5082%, BLEU-1 score 0.569302 and BLEU-4 score 0.300328, showing competitive results. This highlights the value of culturally specific datasets.

Keywords: Image Captioning, Nepalese Culture, CNN, LSTM, BLEU Score

1. Introduction

Image captioning refers to the automated process of generating a natural language description that accurately reflects the content of an image. It is a challenging (Y. Wang 2017) and interdisciplinary task at the intersection of computer vision, natural language processing (X. He 2017), and deep learning. The ability to describe visual scenes in textual form has wide-ranging applications such as aiding visually impaired individuals, enhancing content-based image retrieval, improving surveillance systems, and enabling better human-computer interaction (R. Krishnan 2014). Human beings can effortlessly interpret the semantics of a scene with a single glance, drawing upon context, culture, and prior knowledge (A. Karpathy 2017). In contrast, machines must rely on computational models to parse and translate the raw pixel data into meaningful and coherent sentences. Recent advances in deep learning, particularly Convolutional Neural Networks (CNNs) for image understanding and Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) units, for sequential text generation have significantly improved the capabilities of automated image captioning systems. However, most existing image captioning models are trained on widely used datasets like Flickr8k, Flickr30k, and MSCOCO, which predominantly represent Western cultural contexts. As a result, these models fail to generalize well when applied to images with culturally specific features from underrepresented regions such as Nepal. This cultural gap limits the applicability of state-of-the-art captioning models for generating relevant and context-aware captions for Nepalese socio-cultural and traditional images. To address this issue, this study presents a novel image captioning system specifically designed for Nepalese socio-cultural and traditional contexts. The goal of this work is not only to build a functional image captioning system for a specific cultural domain but also to highlight the necessity of culturally inclusive datasets in training robust and context aware AI models. By promoting the use of

* *Corresponding author*

indigenous content in AI systems, this study contributes to a broader effort toward digital inclusion and the preservation of local cultural heritage

2. Related Work

Image captioning has evolved significantly over the past decade, transitioning from rule-based systems to modern deep learning approaches. Early works, such as that by Pan et al. (J.-Y. Pan 2004), attempted to generate captions by mapping image features like color and texture into text tokens (blob-token mappings). However, such techniques required extensive manual annotations and were not scalable for diverse datasets. Sivakrishna et al. (A. S. Reddy 2015) proposed a two-stage framework involving clustering for content selection and template-based sentence generation. While this method performed well for visually similar image clusters, it struggled with dissimilar or complex images. Harzig et al. (P. Harzig 2018) proposed a model to extend the image captioning by multimodal approach for market analysis. This model penalizes if brand name is not contained within the caption. Xu et al. (K. Xu 2016) further improved this architecture by introducing attention mechanisms that allowed the model to focus on specific image regions while generating each word. This significantly improved performance, especially on complex scenes. Farhadi et al. (A. Farhadi 2010) explored semantic alignment of images and captions by mapping both into a shared embedding space. Jacob et al. (J. Devlin 2015) introduced a nearest-neighbor method that retrieved captions from visually similar images, but lacked originality in caption generation. More advanced models began leveraging deep learning. Karpathy and Fei-Fei (A. Karpathy 2017) introduced a model that aligned visual image regions with segments of sentences using CNNs and Bidirectional Recurrent Neural Networks (BRNNs), enabling semantic understanding of image components. Vinyals et al. (O. Vinyals 2015) proposed a now-standard encoder-decoder framework using CNNs to extract image features and LSTMs to generate coherent natural language descriptions. In recent years, the field has shifted toward transformer-based and multimodal architectures that integrate vision and language understanding more holistically. BLIP (Bootstrapped Language–Image Pretraining) Li et al. (J. Li 2022) and OFA (One-For-All) Wang et al. (P. Wang 2022) unified vision–language tasks through large-scale pretraining, enabling improved generalization and zeroshot captioning capabilities. Similarly, Flamingo Alayrac et al. (J.-B. Alayrac 2023) introduced a vision–language model capable of few-shot learning across tasks using large multimodal datasets, while GPT-4V OpenAI, (OpenAI 2024) extended this approach to multimodal reasoning, combining visual and textual understanding in a single generative framework.

These advanced models achieve superior performance on large benchmark datasets such as MSCOCO and NoCaps, but they require massive computational resources and extensive data, making them unsuitable for low-resource environments. In contrast, this study adopts a lightweight CNN–RNN baseline that can be effectively trained and deployed in limited-data contexts. The proposed model focuses on culturally specific imagery, particularly Nepalese socio-cultural and traditional scenes addressing a gap in existing research where such localized datasets are rarely represented. By demonstrating the performance of a compact encoder–decoder model on a custom dataset, this work serves as a baseline for future transformer-based adaptation in low-resource, culturally contextual applications.

3. Methodology

The proposed system for automatic image captioning of Nepalese socio-cultural and traditional images follows deep learning-based encoder-decoder architecture shown in figure 1 and 2. The encoder extracts visual features using a Convolutional Neural Network (CNN), while the decoder generates captions using a Long Short-Term Memory (LSTM) (P. Shah 2015) network.

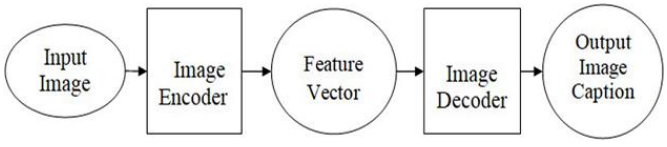

Figure 1. System Block Diagram

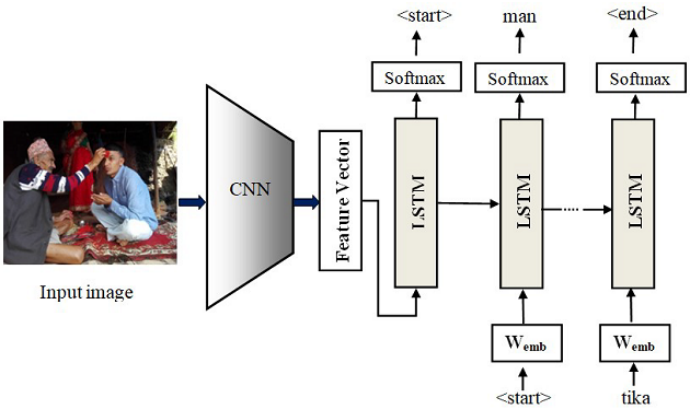

Figure 2. Elaborated System Block Diagram

3.1 Dataset Preparation

A custom dataset was developed consisting of 412 culturally rich images representing Nepalese festivals, attire, rituals, and rural life. Each image was annotated with three descriptive captions, resulting in a total of 1,236 captions. Images were sourced from public domains and personal photography. The dataset was divided into training (312 images) and testing (100 images) sets. Additionally, the Flickr8k dataset was used to compare model performance.

3.2 Image Preprocessing and Feature Extraction

All images were resized to 224×224×3 to match the input dimensions of the CNN. VGG19, pre-trained on ImageNet, was used as the feature extractor. The output from the second last fully connected layer, a 4096-dimensional feature vector, was used to represent each image. These vectors were stored and reused for efficient training.

3.3 Caption Preprocessing and Tokenization

Captions were cleaned by removing punctuation, converting to lowercase, removing tokens with numbers in them and removing hanging 's' and 'a'. Special tokens such as start and end were added to denote sentence boundaries. Tokenization was applied to convert words into integer sequences, and padding was used to ensure uniform input length.

3.4 Model Architecture

The model consists of three primary components:

- Feature Extraction Layer: The 4096-dimensional image feature vector (provided by fc7) is passed through a dense layer with ReLU activation to reduce dimensionality to 512.
- Sequence Processing Layer: Captions are processed through an embedding layer and passed into a unidirectional LSTM with 512 units. Padding masks are applied to ignore pad tokens during training.
- Decoder and Output Layer: The image features and embedded caption vectors are combined and passed through a dense layer with softmax activation to predict the next word in the sequence. Cross-entropy loss is used as the objective function.

3.5 Long Short-Term Memory *(LSTM)*

Long Short-Term Memory (LSTM) acts as the decoder model. When extracted features are transferred to it, LSTM cell provides ability to store, process and generate sequential information, which helps to generate the sentences with keeping previous words in context. LSTM is special type of RNN, which is introduced to overcome the vanishing gradient problem of normal RNN network. A common LSTM network consists of three logical units called gates such as input gate, forget gate and output gate. The core part of an LSTM model is its memory unit c, which keeps the whole information about the previously generated texts, image features and track functionality of all three gates. It utilizes gating mechanisms to control the flow of information into and out of this cell, enabling both short and long-term context learning. The LSTM is governed by:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad \text{(Equation 1)}$$

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \quad \text{(Equation 2)}$$

$$o_t = \sigma(W_0 x_t + U_0 h_{t-1} + b_0) \quad \text{(Equation 3)}$$

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad \text{(Equation 4)}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \quad \text{(Equation 5)}$$

$$h_t = o_t \circ \tanh(c_t) \quad \text{(Equation 6)}$$

where $\sigma(\cdot)$ is the sigmoid activation, $\tanh(\cdot)$ the hyperbolic tangent, and $\circ$ denotes element-wise multiplication.

Each LSTM unit contains four sets of weight matrices W_f, W_i, W_c, W_o corresponding to the *forget*, *input*, *cell*, and *output gates*, respectively, along with their associated bias vectors b_f, b_i, b_c, b_o. Each weight matrix has dimensions of (512 × (512 + E)), where E is the embedding size.

Figure 3. Sample Images from the Prepared Dataset

3.6 Training and Optimization

The model was trained using the Adam optimizer with a learning rate tuned for convergence. Training was performed on both local machines and Google Colab (with GPU acceleration) to manage computational requirements. Dropout regularization was used to mitigate overfitting, and early stopping was applied based on validation performance

3.7 Evaluation *Metrics*

Model performance was assessed using the BLEU (Bilingual Evaluation Understudy) (M. Kilickaya 2016), (K. Papineni 2002) score, which compares generated captions with reference captions using n-gram precision. Both BLEU-1 and BLEU-4 scores were computed to capture unigram and higher-order n-gram accuracy. Additional insights were obtained through qualitative analysis of generated captions.

4. Results and Evaluation

The proposed image captioning model was evaluated on both the custom dataset of Nepalese socio-cultural and traditional images and the benchmark Flickr8k dataset. The evaluation was conducted using both quantitative metrics (BLEU scores, accuracy, loss) and qualitative inspection of generated captions.

Table 1. BLEU Scores on Flicker8K Dataset

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Mao et al. (J. Mao 2014)	0.5778	0.2751	0.2307	-
Google NIC (A. Karpathy 2017)	0.63	0.41	0.27	-
Karpathy (A. Karpathy 2017)	0.579	0.383	0.245	0.16
Our Model	0.580268	0.409	0.376	0.301

4.1 Experimental Setup

The model was trained using a custom dataset comprising 412 images and 1236 captions. 312 images were used for training, and 100 for testing. For validation, subsets of the Flickr8k dataset were also used. Training was conducted using the Adam optimizer with categorical cross-entropy loss, a learning rate of 0.001, and a batch size of 32. Training was done for 100 epochs, with dropout set to 0.1 to reduce overfitting.

4.2 Quantitative Evaluation

Performance was primarily assessed using BLEU (Bilingual Evaluation Understudy) scores. BLEU-1 and BLEU-4 were used to evaluate unigram and 4-gram matching accuracy between predicted and reference captions.

- Custom Dataset Results: BLEU-1 Score: 0.5693 BLEU-4 Score: 0.3003 Final Test Accuracy: 90.35%.
- Flickr8k Dataset Results: BLEU-1 Score: 0.580268 BLEU- 4 Score: 0.3005.

These scores are competitive with state-of-the-art models such as NIC (Google), Karpathy et al., and Mao et al., particularly when evaluated on a culturally specific and relatively small dataset.

4.3 Comparison with Existing Models

The model outperformed Karpathy et al. and Mao et al. in BLEU-4 score and was close to Google NIC in BLEU-1, demonstrating strong performance despite using a smaller and culturally specific dataset.

4.4 Loss and Accuracy Analysis

For the different size of dataset (number of images 100, 200, 300, 400 and 2000), loss and accuracy during each epoch was recorded in the log directory and records were plotted. The resultant graphs were obtained as follows:

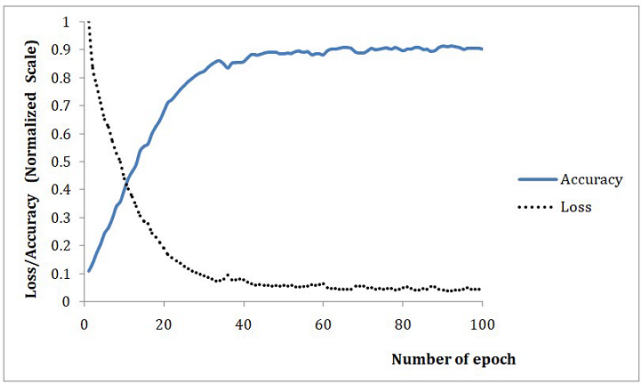

Figure 4. Loss/Accuracy vs. Epoch (Own Dataset)

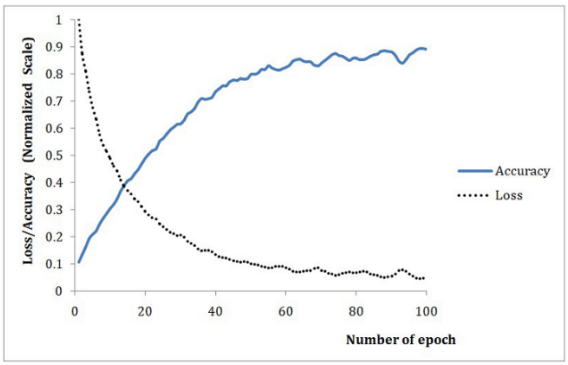

Figure 5. Loss/Accuracy vs. Epoch (Flickr8k Dataset)

4.5 Qualitative Results

The model successfully generated semantically accurate and culturally appropriate captions for test images.

Some failure cases included captions that mismatched gender or object due to feature similarity among culturally similar images (e.g., butter churning vs. pottery). These errors decreased when training data was augmented with similar examples.

Figure 6. Generated Caption: "Mother is carrying her baby on back."

Figure 7. Generated Caption: "Boys are playing with swing during Dashain."

5. Discussion

The results demonstrate that the proposed image captioning model performs effectively on both the custom Nepalese sociocultural dataset and the standard Flickr8k dataset. Despite the relatively small size of the custom dataset, the model achieved competitive BLEU scores and high test accuracy, validating the use of a CNN-LSTM framework for culturally specific image captioning tasks. One of the major contributions of this work is the development of a localized image-caption dataset representing Nepalese cultural practices, attire, festivals, and daily life. Existing datasets like MSCOCO and Flickr8k lack such diversity, which often leads to semantically incorrect captions when used on non-Western imagery. The improvement in performance on the custom dataset illustrates the importance of context-aware training data in machine-generated language. While the model produced many accurate and meaningful captions, some incorrect or ambiguous results were observed. These errors were primarily due to:

- Limited training samples for certain cultural events or objects,
- High visual similarity between different scenes (e.g., pottery vs. butter churning),
- And the lack of linguistic diversity in the reference captions.

Increasing the dataset size and including more varied annotations could significantly reduce these limitations. Additionally, applying attention mechanisms or transformer-based architectures (e.g., BLIP, ViT-GPT) could further enhance the model's ability to distinguish fine-grained details in complex scenes. The comparison with state-of-the-art models confirms that domain adaptation using even a small but relevant dataset can yield comparable or superior results. This highlights the broader significance of building culturally inclusive datasets for AI models in underrepresented regions. Overall, this study contributes not only a working image captioning system for Nepalese images but also a framework for how similar efforts can be replicated in other culturally rich, low-resource settings.

6. Conclusion and Future Work

This paper presented an image captioning system specifically designed for Nepalese socio-cultural and traditional images. By constructing a custom dataset and employing a CNN-LSTM architecture, the system was able to generate meaningful captions that reflected local cultural contexts, something that general models trained on standard datasets often fail to achieve. Quantitative evaluation using BLEU scores and qualitative analysis of generated captions demonstrated the model's effectiveness. The results showed that even with a relatively small but culturally relevant dataset, it is possible to achieve performance comparable to or better than existing state-of-the-art models on similar-sized benchmark datasets. This work highlights the importance of building localized datasets for AI applications in linguistically and culturally diverse regions. It contributes not only a functional captioning model but also a methodology for developing AI systems that are inclusive and context-aware. To improve and extend this research, the following directions are proposed for future works:

- Dataset Expansion: Increase the size and diversity of the dataset by incorporating more images across different ethnic groups, regions, festivals, and traditional practices. Annotating each image with multiple captions from various annotators can also enrich linguistic variability.
- Use of Transformer-based Models: Implement and compare advanced architectures such as BLIP, ViT-GPT, or OFA that have recently shown superior performance in vision-language tasks.
- Multilingual Captioning: Extend the model to support Nepali and other local languages for broader accessibility and relevance.
- Attention Mechanisms: Incorporate spatial or semantic attention to enable the model to focus on key objects or regions in the image during caption generation.
- Human Evaluation: Involve human raters from cultural and linguistic backgrounds to assess the semantic quality, grammaticality, and cultural relevance of generated captions.

This study serves as a foundational step toward developing AI systems that reflect and respect local cultures and can be adapted for other underrepresented regions around the world.

References

A. Farhadi, et al. 2010. "Every picture tells a story: Generating sentences from images, in Computer Vision - ECCV." *Lecture Notes in Computer Science, vol. 6314.* Springer. 15–29.

A. Karpathy, and L. Fei-Fei. 2017. "Deep visual-semantic alignments for generating image descriptions." *IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 4.* 664–676.

A. S. Reddy, N. Monolisa, M. Nathiya, and D. Anjugam. 2015. "Automatic caption generation for annotated images by using clustering algorithm." *Proc. Int. Conf. Innovations Inf., Embedded Commun. Syst. (ICIIECS).* 1–5.

J. Devlin, S. Gupta, R. Girshick, M. Mitchell, and C. L. Zitnick. 2015. "Exploring nearest neighbor approaches for image captioning." *arXiv preprint arXiv:1505.04467.*

J. Li, D. Li, C. Xiong, and S. C. Hoi. 2022. "BLIP: Bootstrapped Language–Image Pre-training for Unified Vision–Language Understanding and Generation." *Proceedings of the 39th International Conference on Machine Learning (ICML), Baltimore, USA* 12888–12900.

J. Mao, W. Xu, Y. Yang, J. Wang, and A. Yuille. 2014. "Explain images with multimodal recurrent neural networks." *arXiv preprint arXiv:1410.1090.*

J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, A. Barron, A. Hasson, J. Carreira, and A. Zisserman. 2023. "Flamingo: A Visual Language Model for Few-Shot Learning." *Advances in Neural Information Processing Systems (NeurIPS), vol. 36.*

J.-Y. Pan, H. Yang, C. Faloutsos, and P. Duygulu. 2004. "Automatic image captioning." *Proc. IEEE Int. Conf. Multimedia Expo (ICME), vol. 3.* 1987–1990.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. "BLEU: a method for automatic evaluation of machine translation." *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics.* 311–318.

K. Xu, et al. 2016. "Show, attend and tell: Neural image caption generation with visual attention." *arXiv preprint arXiv:1502.03044v3.*

M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem. 2016. "Reevaluating automatic metrics for image captioning." *Proc. 15th Eur. Conf. Comput. Vis. (ECCV).*

O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2015. "Show and tell: A neural image caption generator." *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR).* 3156–3164.

OpenAI. 2024. "GPT-4 Technical Report, arXiv preprint arXiv:2303.08774."

P. Harzig, S. Zwicklbauer, M. Redeker, and R. Stiefelhagen. 2018. "Multimodal image captioning for marketing analysis." *Proc. IEEE Conf. Multimedia Inf. Process. Retrieval (MIPR).* 319–324.

P. Shah, V. Bakrola, and S. Pati. 2015. "Image captioning using deep neural architectures." *Proc. Int. Conf. Innovations Inf., Embedded Commun. Syst.* 1–4.

P. Wang, A. Yang, R. Men, Q. Lin, L. Bai, M. Zhou, J. Lin, and H. Wang. 2022. "OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework." *Proceedings of the 39th International Conference on Machine Learning (ICML).*

R. Krishnan, R. Saini, V. Sivaraman, and V. Ganti. 2014. "AutoCaption: Automatic caption generation for personal photos." *Proc. IEEE Winter Conf. Appl. Comput. Vis.* 1050–1057.

X. He, and D. Li. 2017. "Deep learning for image-to-text generation." *IEEE Signal Process. Mag., vol. 34, no. 6.* 109–116.

Y. Wang, Z. Lin, C. Scott, G. Cottrell, and X. Song. 2017. "Skeleton key: Image captioning by skeleton-attribute decomposition." *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR).* 7378–7387.