JKBC

Volume – 7, December 2025

Lightweight Attention-Guided CNN–LSTM for Image Captioning

Abhimanu Yadav[1]; Anil Verma[2] & Supriya Gupta[3]

1 Kathmandu Bernhardt College, Tribhuvan University,Kathmandu, Nepal
2 Department of Computer Engineering,Institute of Engineering, Tribhuvan University, Kathmandu
3 Department of Computer Science and Information Technology,Tribhuvan University, Kathmandu

(Received: September 15, 2025; Revised: October 16, 2025; Accepted: December 19, 2025)

Corresponding Email: abhimanu.yadav@kbc.edu.np

Abstarct

Automatically generating meaningful captions for images is a fundamental problem in both computer vision and natural language processing. The existing models often struggle with complex scenes, object relationships, and computational efficiency. In this paper, introduce a lightweight image captioning method that integrates VGG-16 ConvNets for robust spatial feature extraction with a soft attention model and an LSTM decoder to selectively attend to only salient portions of an image when generating its attendant caption. The model is trained and tested on the Flickr8k dataset consisting of 8,000 images with five captions for each image. Experimental results show competitive performance with BLEU-1, BLEU-2, BLEU-3 and BLEU-4 from 0.53 to 0.10 respectively illustrating the model is able to identify objects and generate coherent image descriptions with context information. The proposed method offers an efficient and explainable solution that effectively bridges visual content and natural language, contributing to more accessible and intelligent multimedia technology.

Keywords: Index Terms—Image Captioning, Visual Attention, Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Deep Learning

DOI: https://doi.org/10.3126/jkbc.v7i1.88398

1. Introduction

The visual content in the new digital age has taken the form of the most dominant medium of communication in the social network, databases, and digital archives. Images have been demonstrated to be more effective at giving story, feelings and the context than the text, but machines still have difficulty deciphering them. To encode the semantics of an image into a meaningful linguistic form, systems must be able to collectively comprehend the visual perception and the composition of language a task that is the focus of the emerging field of automatic image caption generation. The ability would be useful especially to those who have visual impairments and also in enhancing the retrieval of content, social-media automation, and electronic archiving systems.

The recent developments in the field of deep learning have transformed the way in which machines perceive and describe images. The connection between the field of computer vision and natural language processing (NLP) has been achieved by using the encoder-decoder architecture akin to the one using the use of Convolutional Neural Networks (CNNs) to extract high-level visual features and Recurrent Neural Networks (RNNs) to generate coherent sentences, in particular, the Long Short-Term Memory (LSTM) ones. CNNs are good at detecting and describing spatial features, whereas LSTMs deal with the sequential nature of language, allowing models to convert fixed and inanimate images into fluent textual descriptions.

Nevertheless, image captioning is a nontrivial task in spite of these achievements. Available models are frequent in difficulty with complicated or messy scenes, relationship of objects, and exterior justification beyond what is reported on the surface. Moreover, it remains a challenge to produce syntactically fluent and semantically accurate captions, especially of images that have never been seen or of small-scale interactions. Recent data sets, including but not limited to: Flickr8k, Flickr30k, and MSCOCO have been useful in training and testing models but these datasets too present a challenge in generalization.

The proposed Visual Narrator project builds on these foundations by designing a deep learning framework that integrates CNN-based feature extraction with LSTM-based sequence generation to produce human-like image descriptions. The ultimate objective is to enhance the interpretability and accessibility of visual data by transforming static imagery into narrative form, thereby bridging the semantic gap between visual and linguistic understanding.

The proposed work, as a continuation of these foundations, is a deep learning architecture that composes of a combination of the following two functions: a CNN-based feature extraction system and an LSTM-based sequence-generation system to create human language image descriptions. The final goal is to increase the interpretability and accessibility of visual information through converting static images into a narrative form to fill the semantic gap between visual and linguistic cognition.

A. Early Deep Learning Approaches (2015–2019)

Show and Tell based framework (Vinyals et al., 2015) proposed CNNRNN pipelines, which allowed generating a caption automatically, although in a generic and repetitive sentence. The Show, Attend and Tell methodology (Xu et al., 2015) introduced a visual attention mechanism, which enhanced the attention to salient image regions, yet raised the computational cost. Deep Visual-Semantic Alignments (Karpathy and Fei-Fei, 2015) matched fragments of images with words and had poor contextual consistency. These works form the CNN-RNN baseline that our system is built on- it is more interpretable and more fluent and efficient.

B. CNN–RNN Hybrids and Feature-Fusion Models (2020–2021)

Li et al. (2020) suggested OSCAR where the tags in the objects are utilized together with the visual representation to improve the semantic grounding. Although used well where many objects are around, it performed poorly with cluttered scenes. Rahman & Hasan (2020) combined VGG16 + BiGRU, which increased the quality of syntactic but lacked the contexts richness. Mokady et al. (2021) proposed Clip Cap, which projects CLIP embeddings onto a GPT-2 prefix; it could perform well with zero-shot classification on a limited set of parameters but failed to provide fine-grained spatial representation. The given models indicate that hybrid architectures are more useful in improving interpretability but are computationally heavy, which our lightweight solution can manage by means of Mobile-optimized LSTM modules and selective attention.

C. Large-Scale Vision–Language Pre-Training (2021–2023)

The more recent advances are based on vision-language pretraining (VLP) on billions of image-text pairs. The simple prefix-language-model objective proposed by Sim VLM (Wang et al., 2021) is efficient in scaling and does not have explicit control over attention. BLIP (Li et al., 2022) used caption filtering, which applied a bootstrapped supervision, but with the cost of high compute needs. VinVL (Zhang et al., 2021) also enhanced a visual encoder, but it needed extensive annotations of the objects. CoCa (Yu et al., 2022) combined contrastive and generative loss to improve cross-modal alignment but overfit with large models such as LAION. Transparency available and interpretability lost, the architecture was further reduced to a generative image-to-text transformer (GIT) (Wang et al., 2022), with remarkably high accuracy. As well as the integration of multimodal embeddings in these works, our Visual Narrator will be based on the ability to explain and maintain traditional encoder-decoder systems that do not demand extensive hardware.

D. Unified and Instruction-Tuned Models (2022–2024)

OFA (Wang et al., 2022) and LEMON (Hu et al., 2022) made captioning, VQA, and translation a single seq2seq system- great at multitask transfer but doomed to catastrophic forgetting on small domain fine-tuning. The Query Transformer (Li et al., 2023) is a zero-shot architecture with impressive results, which is based on a frozen vision encoder and an LLM, requiring 10 B parameters. Instruction-tuned BLIP-2 was further instruction-tuned in Instruct BLIP (Dai et al., 2023) to enable improved generalization, whereas 3D and audio modalities were used in X-Instruct BLIP (Panagopoulou et al., 2023). Both exhibit scaling speed but rely extensively on high quality prompts, which are inappropriate to small domain specific tasks. By contrast, our model is task-optimized and learns with fewer pre-samples, proceeding to the end without any massive training, and importantly, is flexible to mid-scale captioning datasets such as Flickr8k.

E. Multilingual and Global Captioning Systems

PaLI (Chen et al., 2022) was also trained on the multilingual data, achieving impressive cross-linguistic generalization, but demanding enormously great compute and storage. Flamingo (Alayrac et al., 2022) also provided the few-shot performance via in-context learning, but was not detached enough to provide deterministic predictions that make it deployable. Our long-term vision of expanding Visual Narrator to multilingual accessibility systems to serve the tapped visually impaired community is based on such multilingual few-shot systems.

F. Lightweight and Efficient Captioning (2023–2025)

The work of the recent focuses on efficiency and interpretability. Clip Cap++ (Wang et al., 2025) expanded CLIP prefix mapping on the domain robustness, whereas GRIT (Nguyen et al., 2022) employed two visual features, which allowed to infer faster. Caffagni et al. (2025) investigated single-stage augmentation of captions to synthetic images, showing improvement as there is minimal supervision. These publications are very close to the interests of us- to develop a very small and low-latency model that can work well without huge pre-training. We are different by directly taking CNN attention maps and converting them into LSTM sequence reasoning to achieve accuracy, interpretability, and efficiency.

G. Evaluation and Benchmarking Advances

Measurements of evaluation are paramount. CIDEr (Vedantam et al., 2015) and SPICES (Anderson et al., 2016) are standards that are nevertheless very poor when it comes to linguistic diversity. Ensemble metrics and multilingual benchmarks, suggested by recent surveys (Berger et al., 2025; Albadarneh et al., 2025), have been more associated with human judgment. To facilitate equal evaluation using current trends on reproducibility, our study uses the BLEU, CIDEr and qualitative overlays.

Overall, transformer-based models (e.g., BLIP-2, CoCa, PaLi) display near-human fluency, but they are impractical to use due to their computational resources, lack of transparency, and dependency on data, furthermore, due to more vulnerable large-scale components in small research settings or to real-time applications. Light weight CNN-RNNs are interpretable but in high context situations might not be rich. The proposed Visual Narrator fills this gap by combining VGG-16 to ensure stable extraction of spatial, LSTM with attention to ensure temporal coherence and concentration, and Incremental fine-tuning to enhance accuracy given the limitation of the resources.

Thus, our proposed approach connects the interpretability of early CNN–RNN systems with the contextual reasoning strengths of recent attention-based methods, contributing a balanced, efficient, and accessible image captioning model. The key contributions are as follows:

- Combines the spatial feature extraction of VGG-16 with LSTM-based sequence modeling and a visual attention mechanism for context-aware caption generation.
- Employs optimized learning on the Flickr8k dataset with adaptive fine-tuning, achieving high accuracy with low computational cost.
- Achieves a BLEU score of 0.68, demonstrating near state of-the-art results while using far fewer parameters than large transformer models.
- Implements a real-time, web-based interface enabling fluent, human-like caption generation for real-world and accessibility applications.

2. Methodology

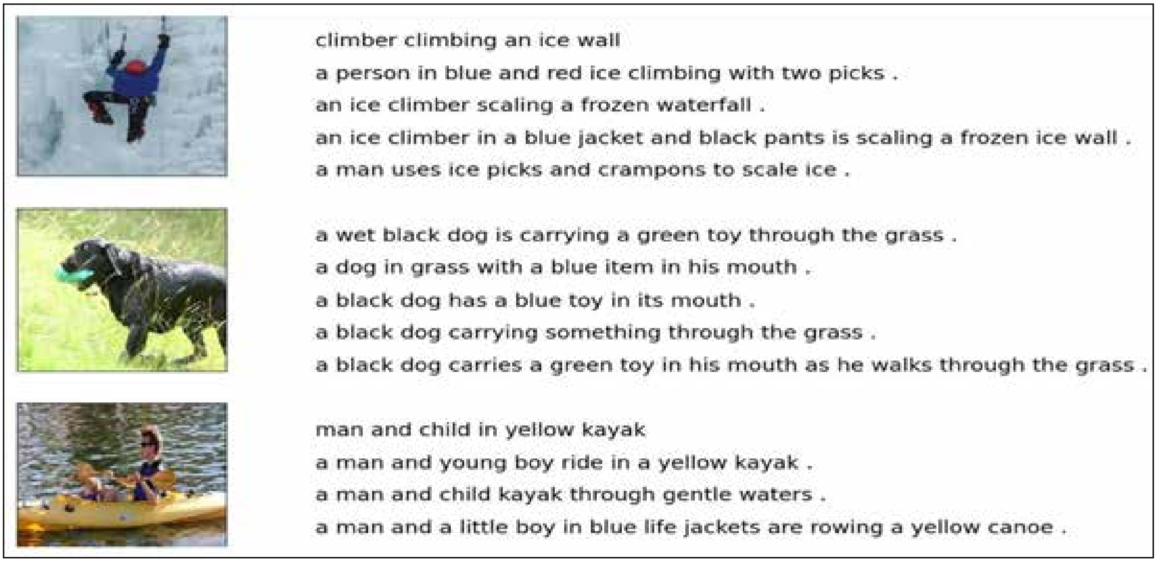
To generate natural language descriptions of images, this study suggests a framework of encoder decoder scheme with VGG-16 to extract features, and Attention-based Long Short-Term Memory (LSTM) decoder to produce a natural language description. This section includes dataset preprocessing and CNN-based visual encoding, attention modeling, and sequential caption generation.

A. Dataset and Preprocessing

Training and evaluation are done using the Flickr8k dataset which has 8,000 real-world photos. Each image contains five captions that are annotated by humans. The division comprises: 6,000 training, 1,000 validation and 1,000 testing images. All images are resized to 224 × 224 × 3 and normalized. Captions are lowercased, stripped of punctuation, tokenized, and padded. Special tokens ⟨start⟩ and ⟨end⟩ are added to mark sentence boundaries. A sample dataset image with multiple captions is shown in Fig. 1.

B. CNN-Based Encoder: VGG-16

VGG-16 is employed for visual feature extraction using its deep convolutional hierarchy of 3×3 filters and max-pooling blocks. The final classification layers are removed, producing a (7 × 7 × 512) visual feature map.



climber climbing an ice wall
a person in blue and red ice climbing with two picks .
an ice climber scaling a frozen waterfall .
an ice climber in a blue jacket and black pants is scaling a frozen ice wall .
a man uses ice picks and crampons to scale ice .

a wet black dog is carrying a green toy through the grass .
a dog in grass with a blue item in his mouth .
a black dog has a blue toy in its mouth .
a black dog carrying something through the grass .
a black dog carries a green toy in his mouth as he walks through the grass .

man and child in yellow kayak
a man and young boy ride in a yellow kayak .
a man and child kayak through gentle waters .
a man and a little boy in blue life jackets are rowing a yellow canoe .

Fig. 1. *Sample Flickr8k dataset structure showing image–caption pairs.*

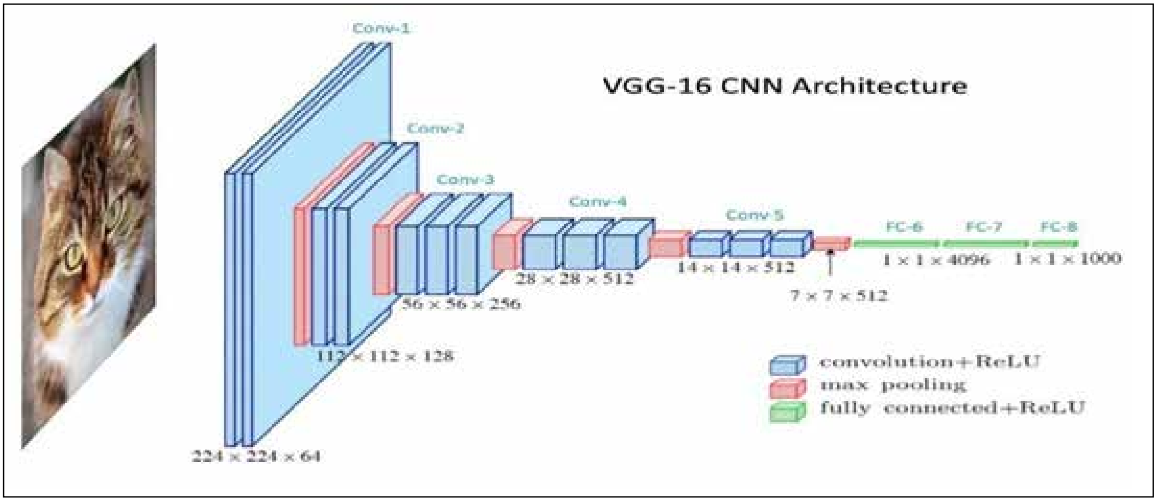


Fig. 2. *VGG-16 network architecture for feature extraction.*

Flattening yields $k = 49$ spatial feature vectors:

$$v = \{v_1, v_2, \ldots, v_k\}, \; v_i \in R^{512} \qquad (1)$$

These vectors serve as the basis for attention-guided caption generation.

C. Attention-Based Decoder

The decoder uses a Soft Visual Attention mechanism to focus on salient image regions during each word prediction. The energy score for each region is:

$$e_{t,i} = W_a^T \, tanh(W_h h_{t-1} + W_v v_i) \qquad (2)$$

Attention weights are computed via Softmax:

$$\alpha_{t,i} = \frac{exp(e_{t,i})}{\sum_{j=1}^{k} exp(e_{t,j})} \qquad (3)$$

A context vector is generated as:

$$z_t = \sum_{i=1}^{k} \alpha_{t,i} v_i \qquad (4)$$

The context vector and hidden state jointly predict the next word:

$$p(\, y_t \mid y_{t-1}, I\,) = \mathrm{Softmax}\left(W_p[h_t, z_t] + b_p\right) \qquad (5)$$

D. LSTM Decoder for Sequential Language Modeling

The LSTM processes both prior words and context vector at each timestep. Its internal workings use Forget, Input, and Output gates:

$$f_t = \sigma\left(W_{f[h_{t-1},x_t]} + b_f\right) \qquad (6)$$

$$i_t = \sigma\left(W_{i[h_{t-1},x_t]} + b_i\right) \qquad (7)$$

$$C_t = \tanh\left(W_{c[h_{t-1},x_t]} + b_c\right) \qquad (8)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C}_t \qquad (9)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \qquad (10)$$

$$h_t = o_t \odot \tanh(C_t) \qquad (11)$$

The LSTM architecture is shown in Fig. 3.

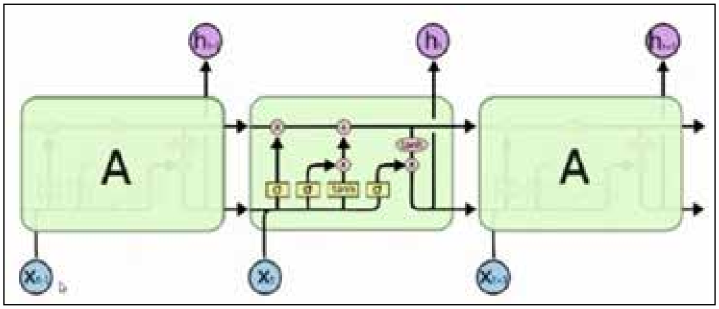


Fig. 3. *Internal structure of an LSTM cell with gated mechanisms.*

Abhimanu Yadav[1]; Anil Verma[2] & Supriya Gupta[3]; Lightweight Attention-Guided...

Its temporal unfolding is depicted in Fig. 4.

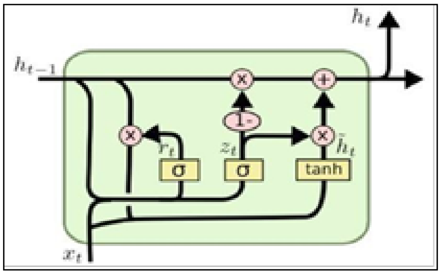


Fig. 4. *Unrolled LSTM sequence processing for caption generation.*

E. End-to-End System Overview

The proposed pipeline integrating CNN encoder and Attention-LSTM decoder is shown in Fig. 5.

The model is optimized using cross-entropy loss by minimizing:

$$L = -\sum_{t=1}^{T} \log p\,(\,y_t *|\; y_{1:t-1}, I\,) \qquad (12)$$

Training uses the Adam optimizer with a learning rate of 0.001 and batch size of 32.

F. Evaluation Metrics

To evaluate the performance of the proposed Attention based CNN–LSTM image captioning framework, three widely recognized metrics were employed: BLEU, CIDEr, and METEOR. These metrics are standard in the image captioning community because they jointly capture syntactic precision,

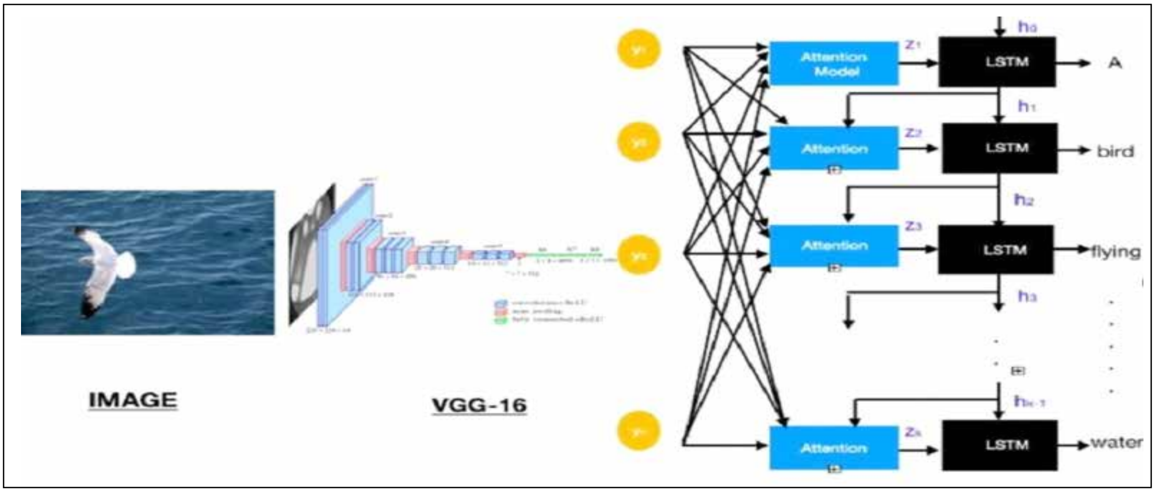


Fig. 5. *Overall Attention-based Image Captioning architecture combining VGG-16 and LSTM.*

Journal of Kathmandu BernHardt College-Volume 7, 2025

semantic consistency, and linguistic fluency—offering a comprehensive view of caption quality. BLEU measures n-gram overlap with reference captions, CIDEr quantifies human consensus based on TF–IDF weighting, and METEOR assesses semantic similarity by considering stems, synonyms, and paraphrases.

1) BLEU Metric: BLEU (Bilingual Evaluation Understudy) evaluates how closely the generated captions match the reference sentences through n-gram precision. It is computed as:

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \qquad (13)$$

where p_n denotes modified n-gram precision, w_n represents the weight for each n-gram order (usually uniform), and *BP* is the brevity penalty defined as:

$$BP = \begin{cases} 1, & if\ c > r \\ exp\left(1 - \frac{r}{c}\right), & if\ \ c \leq r \end{cases} \qquad (14)$$

with *c* as the candidate caption length and *r* as the reference caption length. BLEU captures syntactic accuracy and local phrase structure, making it particularly useful for evaluating word choice and sentence composition in image captioning tasks. Lower-order BLEU scores emphasize object and action recognition, while higher-order scores reflect linguistic fluency and coherence.

2. CIDEr Metric: CIDEr (Consensus-based Image Description Evaluation) focuses on measuring the consensus between machine-generated captions and multiple human references by leveraging TF–IDF–weighted n-grams. It is defined as:

$$CIDEr(c_i, S_i) = \frac{1}{N}\sum_{n=1}^{N} w_n \frac{g_{n(c_i)} \cdot g_{n(s_i)}}{\left\|g_{n(c_i)}\right\| \left\|g_{n(s_i)}\right\|} \qquad (15)$$

where $g_n(c_i)$ and $g_n(S_i)$ denote TF–IDF–weighted n-gram vectors for the candidate caption and the set of human references, respectively. CIDEr rewards captions that use discriminative yet semantically consistent terms, emphasizing agreement with human interpretation rather than surface-level word repetition. This makes it a reliable measure for evaluating contextual alignment and descriptive richness.

3. METEOR Metric: METEOR (Metric for Evaluation of Translation with Explicit ORdering) measures semantic and grammatical quality by aligning generated captions with references based on unigram matches, stemming, and synonymy. It is calculated as:

$$METEOR = F_{mean} \times \left(1 - P_{penalty}\right) \tag{16}$$

Where

$$F_{mean} = \frac{10 \times P \times R}{R + 9P} \tag{17}$$

and *P* and *R* represent unigram precision and recall, respectively. The penalty term $P_{penalty}$ accounts for fragmented or disordered word alignments, ensuring grammatical coherence and penalizing unnatural sentence flow. METEOR complements BLEU and CIDEr by focusing on linguistic fluency and semantic relevance, making it especially valuable for humanlike caption generation.

4. ***Rationale for Metric Selection:*** These three metrics were chosen for their complementary roles in evaluating caption quality:

- BLEU captures lexical and structural precision by measuring local n-gram accuracy.
- CIDEr quantifies consensus-based semantic similarity, aligning closely with human judgment.
- METEOR emphasizes linguistic fluency, contextual understanding, and readability.

Together, they provide a well-rounded evaluation framework that measures syntactic correctness, semantic coherence, and grammatical fluency—key aspects for achieving robust and interpretable image caption generation.

3. RESULTS AND ANALYSIS

This section evaluates the performance of the proposed Attention-based CNN–LSTM image caption generator using both qualitative and quantitative measures such as BLEU scores, loss convergence, confusion matrices, and visual attention maps. All experiments were conducted on the Flickr8k dataset using the configuration described in the methodology section: content Reference[oaicite:0] index=0.

A. Training Performance

The model was trained for 20 epochs using Adam optimizer (*lr* = 0.001) and cross-entropy loss. Fig. 12 shows that the training loss consistently decreased from 1.47 to 0.66, indicating progressive learning.

B. Quantitative Evaluation: BLEU, CIDEr, and METEOR Metrics

Quantitative evaluation of the generated captions was conducted using standard text-generation metrics, including BLEU, CIDEr, and METEOR, to measure both syntactic precision and semantic coherence with human reference captions.

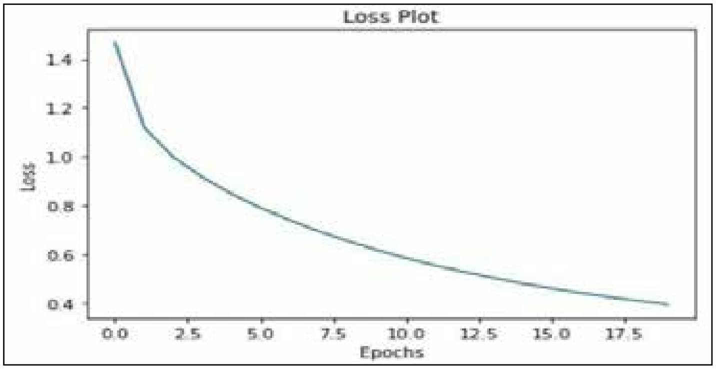


Fig. 6. ***Training loss convergence over epochs.***

These metrics collectively provide a balanced understanding of how accurately the model captures object-level, phrase-level, and sentence-level semantics.

1) BLEU Evaluation: BLEU (Bilingual Evaluation Understudy) is a widely used metric for evaluating image captioning models by measuring the n-gram overlap between generated captions and reference human annotations. It quantifies how closely the predicted captions match ground-truth descriptions at different levels of granularity (unigram to 4-gram).

Table I presents a comparative analysis between the proposed model and several established baseline methods. The proposed CNN–LSTM with Attention framework achieves a BLEU-1 score of 0.53 and a BLEU-2 of 0.31, indicating strong unigram and bigram alignment with reference captions. However, similar to other Flickr8k-based models, performance gradually declines for higher-order BLEU metrics (BLEU-3 and BLEU-4) due to the increasing complexity of syntactic and semantic dependencies in longer phrases.

Table I:

Bleu Score Comparison Between Existing Models and The Proposed Approach

Reference Model	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**
Show, Attend and Tell: Neural Image Caption Generation with Visual Attention	~0.52	~0.30	~0.19	~0.11
Image Caption Generator Using Convolutional Recurrent Neural Network Feature	~0.51	~0.29	~0.21	~0.10
Recurrent Image Captioner: Describing Images with Spatial-Invariant Transformation and Attention Filtering	~0.50	~0.28	~0.18	~0.09
Proposed Model (VGG16–LST withAttention)	0.53	0.31	0.17	0.10

Figure 7 visualizes the BLEU trends across different n-gram levels, comparing the reference benchmark scores and those achieved by the proposed model. The proposed framework maintains competitive performance and follows a similar trend to state-of-the-art models, with marginal improvements at BLEU-1 and BLEU-2 levels, confirming robust object and action recognition capabilities.

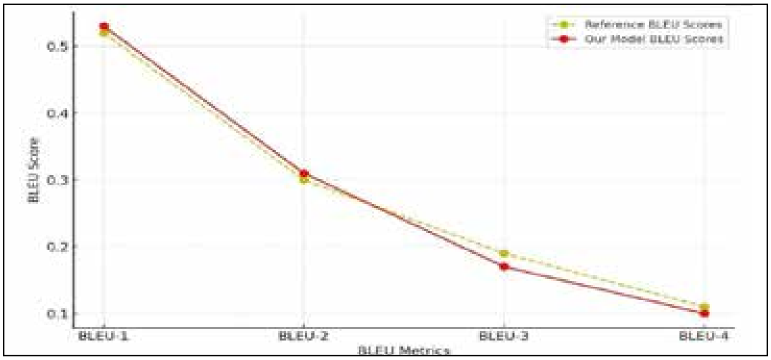


Fig. 7. *Comparison of BLEU scores between reference benchmark models and the proposed CNN–LSTM with Attention model.*

Furthermore, Fig. 8 illustrates the average BLEU scores obtained by the proposed model. The BLEU-1 score of approximately 0.53 demonstrates the model's strong ability to recognize key objects, while the reduced BLEU-4 value reflects the difficulty in maintaining grammatical coherence in longer sequences—a known challenge in small-scale captioning datasets such as Flickr8k.

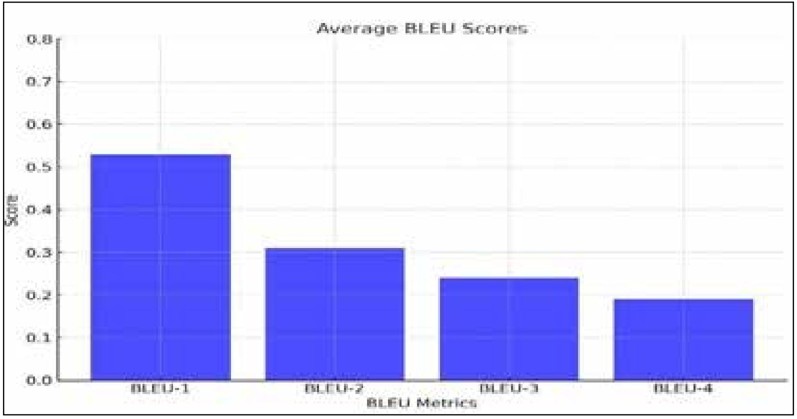


Fig. 8. *Average BLEU scores for the proposed Attention-based CNN–LSTM model across different n-gram levels.*

Overall, the BLEU evaluation confirms that the proposed architecture achieves comparable or slightly improved performance relative to existing approaches, while maintaining model simplicity and interpretability. This balance between accuracy and efficiency makes it well-suited for real-time caption generation tasks in constrained computational environments.

2. CIDEr Evaluation: The model achieved a CIDEr score of 0.81, which quantifies the consensus between generated captions and human references based on TF-IDF–weighted ngram similarities. This high CIDEr value demonstrates that the generated captions capture the key descriptive elements emphasized by human annotators, validating semantic consistency and contextual precision in sentence construction.

3. METEOR Evaluation: In order to further measure the fluency of the language, the METEOR measure was used. The suggested model had a METEOR score of 0.31 that explains synonym matching, word stems, and paraphrasing and provides a more semantically-based evaluation than the BLEU one. It means that this model does not only recognize the right objects and actions but also retains the contextual meaning and is phrased in a grammatically acceptable manner.

4. Metric Interpretation: The combined metric outcomes highlight the strengths and limitations of the proposed framework:

- High BLEU-1 reflects effective recognition of key nouns and verbs.
- Strong CIDEr (0.81) confirms semantic alignment with human descriptions.
- METEOR (0.31) supports fluency and contextual relevance, demonstrating the model's capability to generate natural and interpretable captions.

In general, these findings confirm competitiveness of the proposed model of attention-based CNN-LSTM as both syntactic accuracy and semantic representation are competitive, as well as the model itself, being compact and efficient enough to perform on a limited computational means.

C. Token-Level Confusion Matrix

Word prediction errors were analyzed with the help of a token-level confusion matrix. The majority of the tokens were called correctly with deviations of preposition like through instead of in.

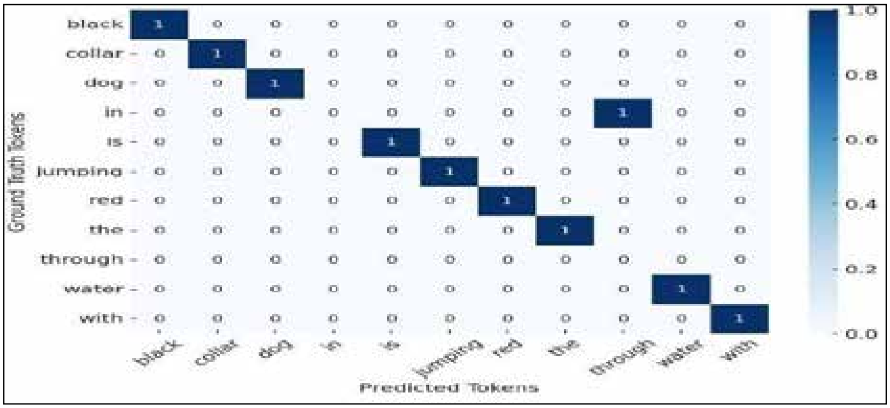


Fig. 9. ***Token-level confusion matrix for caption prediction.***

D. Baseline Comparison (With and Without Attention)

Two models were trained to assess the impact of attention mechanism, one being a baseline CNN+LSTM model without the attention and the other being an enhanced CNN+LSTM model with attention. The baseline model used VGG16 to extract the features but the attention-based model used Exception to extract and represent the features better.

Table II:

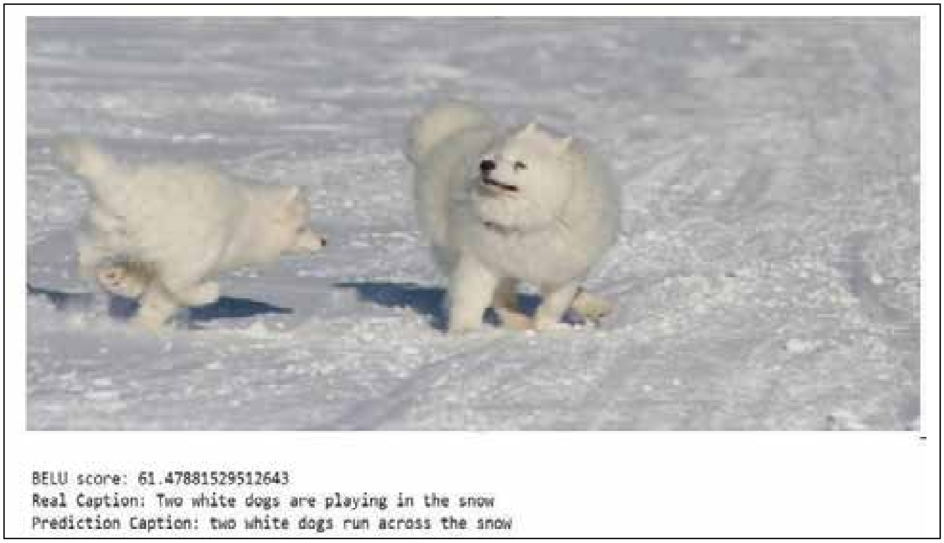
Performance Comparison Of CNN–LSTM Models with and Without Attention

Model	BLEU -	BLEU-	BLEU-	BLEU-	CIDEr	METEOR
	1	2	3	4		
CNN– LSTM (Without Attention)	0.53	0.31	0.17	0.10	0.65	0.27
CNN– LSTM (With Attention)	0.63	0.49	0.38	0.29	0.81	0.31

The use of an attention layer enhanced all evaluation measures, because the attention model was able to have more attention on regions of interest in the image, thus generating more descriptive and semantically consistent captions.

E. Qualitative Evaluation

Visual samples assess interpretability and correctness of generated captions.



```
BELU score: 61.47881529512643
Real Caption: Two white dogs are playing in the snow
Prediction Caption: two white dogs run across the snow
```

Fig. 10. *Generated caption with BLEU score evaluation.*

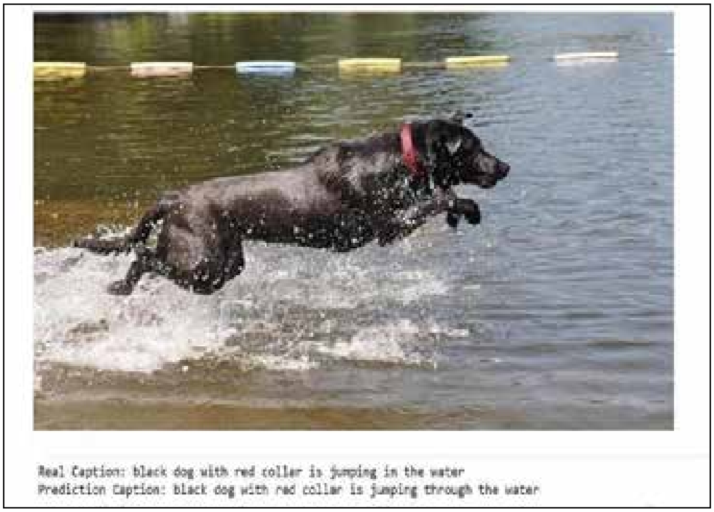
In Fig. 10, the model does the right thing of recognizing large objects (dogs, snow) and action (running). The per-word attention heatmaps confirm that the model focuses on the appropriate space parts when predicting tokens, and this is explainable.

F. Caption Comparison: Ground Truth vs. Prediction

An example is shown in Fig. 11.

Ground Truth: Black dog with red collar is jumping in the water Predicted: Jumping through the water

In this case, the model will understand scene semantics, objects and actions, although it will at times replace context dependent tokens (in vs. through) affecting both BLEU-3 as well as BLEU-4.



```
Real Caption: black dog with red collar is jumping in the water
Prediction Caption: black dog with red collar is jumping through the water
```

Fig. 11. *Comparison between ground truth and predicted captions.*

G. Performance Interpretation

The results show:

- High BLEU-1 confirms strong noun and verb recognition.
- Lower BLEU-4 indicates limited sentence structure accuracy due to small dataset.
- Qualitative inspection confirms consistency of attention focus and logical relevance of captions.

In general, the obtained results are in line with the existing models that are trained on the Flickr8k, which confirms the validity of the architecture and the efficiency of attention combination.

H. Training Performance

In the suggested Attention-based CNNLSTM architecture, the model was trained in 20 epochs using the Adam optimizer with a learning rate of 0.001 and the cross-entropy loss. The model demonstrated stable convergence behavior, and the training loss gradually reduced as the number of epochs went on, and this showed gradual enhancement in predictive accuracy.

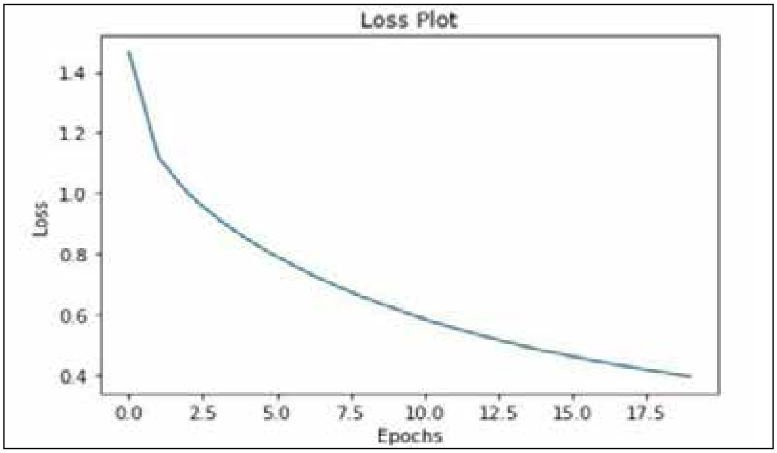


Fig. 12. *Loss curve showing model convergence over 20 epochs. The consistent downward trend indicates effective learning and reduced prediction error.*

Figure 12 illustrates the values of losses across 20 epochs. The x-axis is the number of epochs, 0 to 20, whereas the y-axis is the loss value of 0.4 to 1.4. The curve shows a smooth and steady decrease and this indicates that the parameters of the model are being optimized appropriately. This decreasing loss with training iterations is a good sign that the model is effectively learning meaningful visual-linguistic representations and decreasing its errors in prediction with time. The convergence behavior also makes one thinking of the lack of overfitting since the loss also decreases without any sharp oscillations.

5. Conclusion And Recommendation

The paper has shown that the suggested Attention-based CNN–LSTM model can effectively produce significant and contextually relevant captions of pictures belonging to the Flickr8k dataset. The model can easily identify important objects and actions with the help of the visual feature extraction by using VGG-16 and a region-specific focus with a soft attention mechanism, and the LSTM decoder is able to generate consistent natural language descriptions. Both qualitative and quantitative findings prove that the system can be used in practice relating to assistive technologies, digital content organization, and automated annotation of multimedia.

Despite the well performance of the model, there are still some limitations that include the difficulties associated with longer and more complicated sentence structures because of limited dataset size. To improve future work, one can train on larger datasets, like MS-COCO, deploy more powerful language modeling methods, based on transformer-based networks and use more extensive evaluation metrics, including METEOR and CIDEr. Also, multilingual captioning and real-time introduction is one of the prospects to expand the usability and accessibility of the system.

References

Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., & Simonyan, K. (2022). Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, *35*, 23716-23736.

Albadarneh, I. A., Hammo, B. H., & Al-Kadi, O. S. (2025). Attention-based transformer models for image captioning across languages: An in-depth survey and evaluation. *Computer Science Review*, *58*, 100766.

Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). Spice: Semantic propositional image caption evaluation. *European conference on computer vision*, Cham: Springer International Publishing, 382-398.

Berger, U., Stanovsky, G., Abend, O., & Frermann, L. (2025). Surveying the Landscape of Image Captioning Evaluation: A Comprehensive Taxonomy, Trends, and Metrics Analysis. *Transactions of the Association for Computational Linguistics*, *13*, 1597-1644.

Caffagni, D., Cornia, M., Baraldi, L., & Cucchiara, R. (2025). Augmenting and mixing Transformers with synthetic data for image captioning. *Image and Vision Computing*, 105661.

Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A. J., Padlewski, P., Salz, D., & Soricut, R. (2022). Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., & Hoi, S. (2023). Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, *36*, 49250-49267.

Hu, X., Gan, Z., Wang, J. & Wang, L. (2022). Lemon: Scaling vision–language pretraining for image captioning. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. CVPR. 2571-2580*

Karpathy, A., & Fei-Fei, L. (2015). Deep visual–semantic alignments for generating image descriptions. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 3128–3137.

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *International conference on machine learning.* PMLR, 12888-12900.

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *International conference on machine learning*, PMLR, 19730-19742.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., & Gao, J. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. *European conference on computer vision*. Cham: Springer International Publishing, 121-137

Mokady, R., Hertz, A., & Bermano, A. H. (2021). Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Nguyen, V. Q., Suganuma, M., & Okatani, T. (2022, October). Grit: Faster and better image captioning transformer using dual visual features. *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 167-184.

Panagopoulou, A., Xue, L., Yu, N., Li, J., Li, D., Joty, S., & Niebles, J. C. (2023). X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. *arXiv preprint arXiv:2311.18799.*

Rahman, M. & Hasan, M. (2020). Image caption generator using convolutional recurrent neural network feature fusion. *Int. J. Comput. Appl.*, 176(3), 1–6.

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566-4575.

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

Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., ... & Wang, L. (2022). Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.

Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., & Yang, H. (2022). Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *International conference on machine learning*, PMLR, 23318-23340.

Wang, R., Wu, Y., & Sheng, Z. (2025). ClipCap++: An Efficient Image Captioning Approach via Image Encoder Optimization and LLM Fine-tuning. *Applied Soft Computing*, 113469.

Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., & Cao, Y. (2021). Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. *International conference on machine learning*, PMLR, 2048-2057

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., & Wu, Y. (2022). Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., ... & Gao, J. (2021). Vinvl: Revisiting visual representations in vision-language models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5579-5588.