# ConvNeXt-Driven Image Captioning in Hindi with Adaptive Attention and Transformer Decoder

Anjali Sharma* and Mayank Aggarwal

*Department of Computer Science and Engineering, FET, Gurukul Kangri Deemed to be University Haridwar, India*
*E-mail: 23631001@gkv.ac.in; mayank@gkv.ac.in*
**Corresponding Author*

Received 12 June 2026; Accepted 23 July 2026

## Abstract

This work presents a unified deep learning framework for automatic Hindi image caption generation in low-resource settings. The framework combines a ConvNeXt encoder, an SE-inspired adaptive attention module, and a Transformer decoder to improve visual representation learning and generate contextually relevant Hindi descriptions. The proposed framework unifies a ConvNeXt-based hierarchical visual encoder, an adaptive attention mechanism enabling dynamic saliency modulation, and a transformer decoder optimized for high-fidelity caption synthesis via multi-head self-attention. This synergy facilitates effective handling of Hindi's morphological richness, enabling precise cross-modal alignment and robust non-sequential dependency modeling. Empirical evaluations conducted on benchmark datasets demonstrate competitive performance. The model achieves a training accuracy of 77.80%, a validation accuracy of 77.56%, and stable optimization with losses near 1.26. Caption quality metrics shows the effectiveness of the proposed framework, attaining BLEU-1/2/3/4 means of 0.8646, 0.6282, 0.5401, and 0.4429, respectively, alongside a CIDEr score of 0.8158 and METEOR of 0.6622. Additionally, low WER (0.2535) and CER (0.2607)

*Journal of Graphic Era University, Vol. 14_2,* 607–624.
doi: 10.13052/jgeu0975-1416.14210

values, coupled with an F1-Score of 0.8313, affirm the robustness and linguistic coherence of generated captions. The study contributes a scalable paradigm for multilingual captioning and establishes methodological foundations applicable to broader multimodal research within low-resource linguistic domains.

**Keywords:** Hindi Image Captioning, ConvNeXt-based Image Encoder, Adaptive Attention Mechanism, Transformer Decoder.

## 1 Introduction

In recent years, image captioning, which creates descriptive text from visual inputs, has drawn much interest because of its uses in automated content creation, accessibility, and human-computer interaction. While English picture captioning has been the subject of much research, low-resource languages like Hindi have received less attention. For sequence creation, early picture captioning systems used Convolutional Neural Networks (CNNs) [1] in conjunction with Recurrent Neural Networks (RNNs) [2], such as Gated Recurrent Units (GRU) [3] and Long Short-Term Memory (LSTM) [4]. By adding attention mechanisms, models could better align textual descriptions with picture attributes and concentrate on pertinent visual areas while creating captions. With parallelized decoding and improved contextual representation, Transformers have become strong substitutes for RNNs, significantly improving performance across various language creation tasks. Despite these developments, current methods for captioning Hindi images are frequently used on antiquated architectures like ResNet with LSTMs, making it challenging to capture hierarchical feature representations efficiently [5]. Furthermore, conventional attention processes are less flexible, making them less suited to deal with Hindi's morphological and syntactic plasticity. In order to overcome these drawbacks, we provide a brand-new deep learning system that creates excellent Hindi captions by combining ConvNeXt, a sophisticated CNN architecture, with an adaptive attention mechanism and a Transformer decoder. Our method uses ConvNeXt's hierarchical feature extraction capabilities, which enable the model to have linguistic coherence while preserving fine-grained visual features. The key contribution of this work are as follows:

- **Integration of ConvNeXt** as the image encoder for high-quality feature extraction.

- **Implementation of adaptive attention** to dynamically align linguistic and visual contexts.
- **Optimization of a Hindi-specific Transformer decoder** to enhance fluency and grammatical correctness in caption generation.
- **Extensive evaluations** on benchmark datasets demonstrating the effectiveness and competitive performance of the proposed framework.

By addressing the fundamental challenges in Hindi image captioning, this study contributes to advancing multimodal AI applications in low-resource linguistic contexts. These applications have potential applications in education, assistive technologies, and automated media annotation.

## 2 Related Work

This section discusses the growth of image captioning techniques is essential to contextualize the need for improved approaches, especially for linguistically complex and low-resource languages like Hindi.

### 2.1 CNN + RNN-Based Methods

Early developments in image captioning leveraged the powerful feature extraction capabilities of Convolutional Neural Networks (CNNs) alongside the sequence modeling ability of Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks. A prominent example is the Show and Tell model proposed by [6], which utilized a CNN (Inception) to encode image features and an LSTM decoder to generate natural language descriptions. This approach demonstrated that deep neural networks could jointly learn to extract visual features and generate meaningful captions, effectively bridging computer vision and natural language processing.

### 2.2 Attention Mechanisms

The integration of attention mechanisms marked a significant advancement in image captioning. [7] introduced the Show, Attend and Tell model, which employed soft visual attention to dynamically focus on different regions of the image while generating each word of the caption. This innovation allowed the decoder to selectively attend to relevant image regions, improving both performance and interpretability. Subsequent work expanded upon this idea with hard attention (stochastic sampling) and adaptive attention mechanisms, which further enhanced captioning quality by refining how models balance between visual and linguistic cues during decoding [8, 9].

### 2.3 Transformers for Vision and Language

With the success of Transformer architectures in natural language processing, they were soon extended to multimodal tasks. [10] proposed the Vision Transformer (ViT), which treated image patches as input tokens to a standard Transformer, achieving competitive results with convolution-based architectures. Inspired by this, several models such as ViLBERT [11], Oscar [12], and UNITER [13] introduced Transformer-based architectures to jointly process vision and language modalities. These models utilize self-attention to align visual and textual information more effectively and have set new benchmarks in vision-language tasks including image captioning, visual question answering, and image-text retrieval.

### 2.4 ConvNeXt: Modernizing CNNs

While Transformers gained traction, [14] proposed ConvNeXt, a pure convolutional network that reimagined the CNN architecture by integrating design principles from Transformers, such as large kernel sizes, GELU activations, and layer normalization. ConvNeXt achieved performance on par with Transformer-based models like Swin Transformer [15] on standard image recognition benchmarks, reaffirming the competitiveness of convolutional approaches. Building upon this foundation, our work incorporates ConvNeXt as the image encoder, combined with a multilingual decoder enhanced by adaptive attention mechanisms and mBART-based translation capabilities [16]. This hybrid design aims to leverage the strengths of convolutional architectures in local feature extraction and Transformers in sequence modeling and contextualization, particularly for multilingual image captioning tasks.

Recent advancements in vision-language learning have further improved image captioning through large-scale pre-training and multimodal representation learning. Models such as BLIP [17] and BLIP-2 [18] leverage vision-language pre-training to generate high-quality captions with strong visual-text alignment, while GIT [19] employs a unified Transformer architecture for image understanding and text generation. These developments demonstrate the effectiveness of Transformer-based multimodal learning and motivate the integration of efficient visual encoders with attention-based language decoders for multilingual image captioning.

Despite significant progress, most image captioning models remain optimized for English or other high-resource languages. Hindi poses distinct challenges – rich morphology, flexible word order, compound word formation,

**Table 1** Summary of research gaps in image captioning

| Approach | Strengths | Research Gaps |
|---|---|---|
| CNN + RNN | Strong feature extraction and sequence generation | Limited context modeling and multilingual support |
| Attention Mechanisms | Improved focus and interpretability | Generalization and fine-grained alignment remain challenging |
| Transformers | Effective vision-language alignment via self-attention | High compute cost; limited multilingual capabilities |
| ConvNeXt | Modernized CNNs with strong performance | Underused in captioning and multilingual applications |

and limited annotated data – that traditional CNN–RNN pipelines and standard Transformer models fail to handle effectively. Legacy architectures used in Hindi captioning, such as ResNet + LSTM, also struggle to capture hierarchical visual cues and the language's syntactic complexity. These gaps highlight the need for a more adaptive and linguistically aware architecture. The Table 1 Summarizes the research gaps identified from the existing work.

## 3 Methodology

### 3.1 Proposed Methodology

This study proposes a hybrid encoder–decoder architecture that integrates ConvNeXt-based hierarchical visual encoding with a Transformer decoder, connected through a custom Adaptive Attention module to enhance multimodal alignment. Figure 1 illustrates the overall flow of the proposed image captioning system.

#### 3.1.1 Visual Encoder Using ConvNeXt

The encoder employs ConvNeXt, a modern convolutional neural network that incorporates design principles from Vision Transformers while preserving convolutional efficiency. Given an input image

$$I \in \mathbb{R}^{3\times H\times W}, \tag{1}$$

ConvNeXt extracts a hierarchical feature representation

$$X \in \mathbb{R}^{C\times H'\times W'}. \tag{2}$$

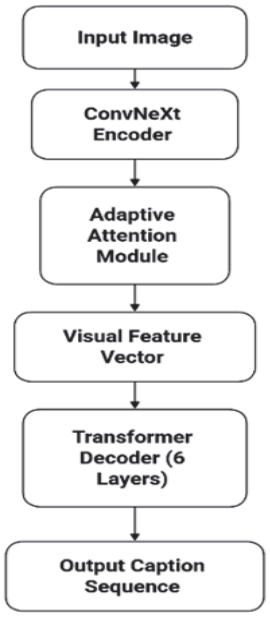



**Figure 1** Proposed image captioning system flowchart.

### 3.1.2 Adaptive Attention Module

To refine global context and recalibrate channel-wise feature responses, an adaptive attention module inspired by the Squeeze-and-Excitation (SE) mechanism is applied after ConvNeXt feature extraction. [20]. The module combines channel-wise weighting with gated feature recalibration, defined as:

$$y = \sigma\big(W_2\,\mathrm{ReLU}(W_1\,\mathrm{GAP}(X))\big) \odot X, \tag{3}$$

where $\mathrm{GAP}(\cdot)$ denotes global average pooling, $W_1$ and $W_2$ are learnable projection matrices, $\sigma(\cdot)$ is the sigmoid activation function, and $\odot$ represents element-wise multiplication. The resulting feature map $y$ emphasizes visually informative channels before decoding.

### 3.1.3 Transformer Decoder

The decoder consists of six stacked Transformer layers. Each layer includes multi-head self-attention to model intra-caption dependencies, cross-attention to align visual features with textual tokens, and a feed-forward network with a hidden dimension of 2048. Layer normalization and dropout with probability $p = 0.1$ are applied throughout.

Caption tokens are embedded into a 512-dimensional space, and positional information is incorporated using sinusoidal positional encoding:

$$\mathrm{PE}(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right), \tag{4}$$

$$\mathrm{PE}(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right). \tag{5}$$

A linear projection followed by a softmax layer produces the final token probability distribution [21].

### 3.1.4 Image Caption Generation Algorithm

This architecture effectively combines ConvNeXt's strong visual representation, the Adaptive Attention module's dynamic saliency modulation, and the Transformer decoder's contextual generation capability.

**Table 2** Image caption generation using convnext and transformer decoder

| Step | Operation |
|---|---|
| 1 | Input image $I$ and caption tokens $Y = \{w_1, \ldots, w_T\}$. |
| 2 | Extract visual features: $f \leftarrow \text{AdaptiveAttention}(\text{ConvNeXt}(I))$. |
| 3 | Generate token embeddings with positional encoding: $E(Y) \leftarrow \text{Embedding}(Y) + \text{PositionalEncoding}$. |
| 4 | For each Transformer decoder layer, compute masked self-attention: $Z_1 \leftarrow \text{LayerNorm}(E(Y) + \text{MHA}(E(Y), E(Y), E(Y)))$. |
| 5 | Compute cross-attention between textual and visual features: $Z_2 \leftarrow \text{LayerNorm}(Z_1 + \text{MHA}(Z_1, f, f))$. |
| 6 | Apply the feed-forward network: $Z_3 \leftarrow \text{LayerNorm}(Z_2 + \text{FFN}(Z_2))$. |
| 7 | Repeat Steps 4–6 for each decoder layer. |
| 8 | Generate the predicted caption: $\hat{Y} \leftarrow \text{Softmax}(\text{Linear}(Z_3))$. |

## 4 Dataset Preparation

The experiments were conducted on the MS COCO 2017 dataset, using 30,000 images for training and 10,000 images for validation. The dataset was selected because it contains a wide variety of everyday scenes with multiple descriptive captions for each image.

As MS COCO provides only English captions, the five reference captions associated with each image were translated into Hindi using Google Translate and manually reviewed for basic grammatical and semantic consistency. Text preprocessing included Unicode normalization, punctuation removal,

vocabulary creation, and the addition of $\langle \mathrm{BOS} \rangle$,$\langle \mathrm{EOS} \rangle$, and $\langle \mathrm{PAD} \rangle$ tokens. Captions were padded or truncated to a fixed length. Images were resized to 224 × 224 pixels and normalized using ImageNet statistics. During training, random horizontal flipping and random cropping were applied for data augmentation, while center cropping was used during validation.

## 5 Training & Experimental Setup

The training and experimental setup is designed to evaluate the proposed ConvNeXt + Transformer-based caption generation model under realistic conditions. The model is trained on the MS COCO 2017 dataset [22], using 30,000 images for training and 10,000 for validation. A CrossEntropy loss function with label smoothing (0.1) is employed to improve generalization [23]. Optimization is handled by the AdamW optimizer with distinct learning rates for the encoder and decoder layers, governed by a cosine annealing learning rate schedule [24]. The model is trained over 5 epochs on a GPU-enabled environment, and batch size is determined dynamically by the DataLoader. Training includes enhancements such as model checkpointing (saving the best model based on validation loss), tracking multiple performance metrics (accuracy, Top-1/Top-5 accuracy, F1-score, precision, recall), and logging GPU memory usage and epoch-wise execution time for performance analysis. Table 3 shows the training and experimental setup.

**Table 3** Training and experimental setup

| Aspect | Value |
|---|---|
| Dataset | MS COCO 2017 (Train: 30,000 / Val: 10,000) |
| Loss Function | CrossEntropy with Label Smoothing factor of 0.1 |
| Optimizer | AdamW with differential learning rates for encoder and decoder modules |
| Learning Rate Schedule | Cosine Annealing strategy for gradual decay |
| Training Epochs | 5 total passes through the training data |
| Batch Size | Dynamically defined by the PyTorch DataLoader configuration |
| Device | GPU-enabled system with memory usage monitoring |
| Checkpointing | Best model weights saved based on minimum validation loss |
| Evaluation Metrics | Accuracy, Top-1, Top-5 Accuracy, F1-Score, Precision, Recall |
| Efficiency Tracking | GPU memory and time logged per training epoch |

## 6 Results and Discussion

The proposed vision–language image captioning model was trained for five epochs using an encoder–decoder architecture with attention mechanisms. The training and validation curves exhibit steady convergence, where both loss values decrease consistently to final levels of 1.2776 (training) and 1.2635 (validation). Correspondingly, the accuracy trends in Figure 2 illustrate continuous improvement, achieving 77.80% training accuracy and 77.56% validation accuracy, indicating effective learning and minimal overfitting. The training and validation loss curves are also shown over five epochs. Both curves decrease steadily and remain closely aligned, indicating stable optimization and good generalization with minimal overfitting throughout the training process.

The model's overall performance metrics are summarized in Figure 3. These results confirm the model's suitability for real-time or resource-constrained environments.

Quantitative evaluation using standard captioning metrics yielded strong outcomes: BLEU-1 to BLEU-4 [25] scores averaged 0.8646, 0.6282, 0.5401, and 0.4429, respectively, with a CIDEr score [26] of 0.8158 and METEOR of 0.6622 [27]. The F1-score (0.8313) and WER/CER (0.2535/0.2607) further confirm the linguistic coherence and accuracy of generated captions. Final model performance indicators are summarized in Figure 3 with the test

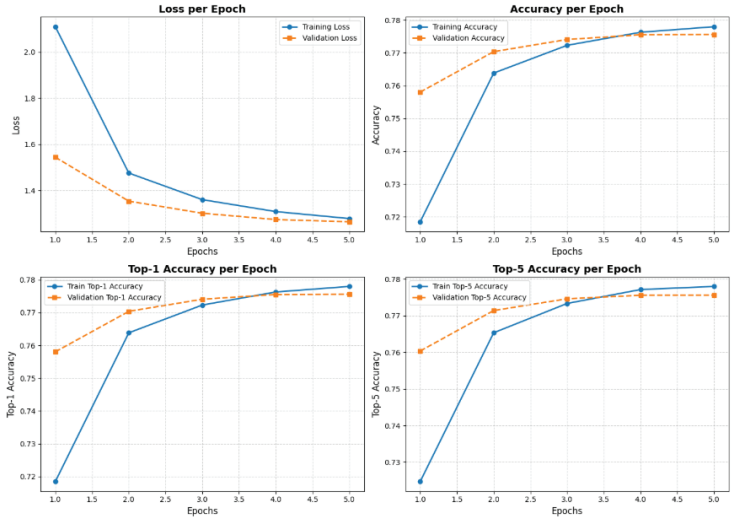



**Figure 2** Training and validation loss per epoch.

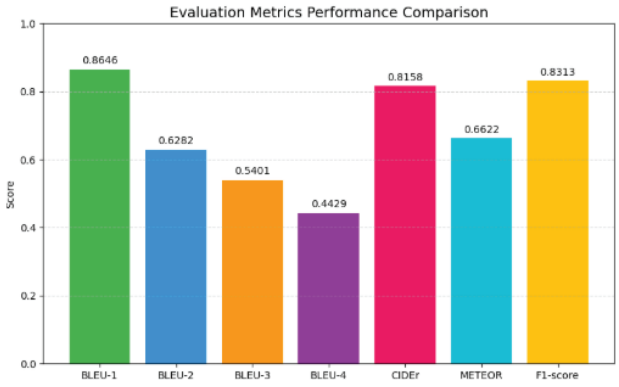



**Figure 3** Overall performance summary of the proposed model.

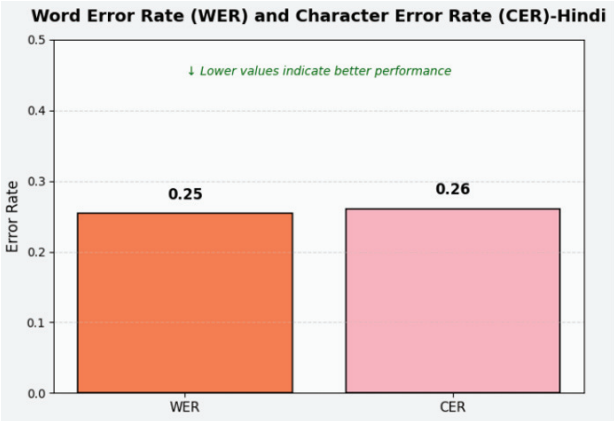



**Figure 4** WER/CER rates.

accuracy recorded at 76.08%, validating robust generalization. The high F1-score and caption evaluation metrics, together with low inference time and moderate computational requirements, demonstrate that the framework provides an effective balance between caption quality and computational efficiency. Figure 4 presents the Word Error Rate (WER) and Character Error Rate (CER). The relatively low values indicate that the generated captions are linguistically accurate and closely match the reference captions, confirming the effectiveness of the proposed framework for Hindi caption generation.

Table 5 compares the proposed model with existing Hindi image captioning methods. The proposed framework achieves competitive performance owing to the hierarchical feature extraction of ConvNeXt, adaptive attention,

**Table 4** Model performance metrics

| Metric | Value |
|---|---|
| Model Size | 246.93 MB |
| Total Parameters | 61,682,051 |
| Trainable Parameters | 61,682,051 |
| FLOPs | 0.71 GFLOPs (Approx.) |
| Inference Time (per image) | 12.311 ms |
| GPU Memory Usage | 5681.19 MB |
| Throughput | 81.25 images/sec |

**Table 5** Comparison of the Proposed method with existing hindi image captioning approaches

| Authors | B1 | B2 | B3 | B4 |
|---|---|---|---|---|
| Mishra et al. | 62.9 | 43.3 | 29.1 | 19.0 |
| Singh et al. | 51.3 | 30.4 | 16.7 | 12.4 |
| Dhir | 57.0 | 39.0 | 26.4 | 17.3 |
| Rathi | 58.0 | 47.0 | 39.0 | 35.0 |
| Meghwal | 62.5 | 45.8 | 32.8 | 23.2 |
| Proposed Model | 86.46 | 62.82 | 54.01 | 44.29 |

and the Transformer decoder, which together improve caption quality and contextual understanding.

Table 6 shows the contribution of each component to the proposed framework. The complete model consistently outperforms all ablated variants, demonstrating the effectiveness of integrating ConvNeXt, Adaptive Attention, and the Transformer decoder.

Figure 5 demonstrates that the proposed model correctly identifies the primary object and its associated activity, generating a semantically meaningful Hindi caption. Figure 6 illustrates the model's ability to recognize multiple visual cues in a sports scene and produce a contextually appropriate caption.

**Table 6** Ablation study of the proposed framework

| Configuration | BLEU-4 | CIDEr | METEOR |
|---|---|---|---|
| ConvNeXt + Transformer (without Adaptive Attention) | 0.2321 | 0.3256 | 0.2847 |
| ConvNeXt + Adaptive Attention (without Transformer) | 0.3536 | 0.4859 | 0.3678 |
| CNN + Adaptive Attention + Transformer | 0.4111 | 0.6994 | 0.6259 |
| Proposed Model | 0.4429 | 0.8158 | 0.6622 |

Generated Caption: एक आदमी बॉल डव को मारने के लिए अपने टेनिस रैकेट तक पहुंचता है
Reference Caption: एक आदमी गेंद को हिट करने के लिए अपने टेनिस रैकेट तक पहुंचता है
BLEU-1: 0.7857, BLEU-2: 0.6954, BLEU-3: 0.6259, BLEU-4: 0.5758
CIDEr: 0.6735
WER: 0.2857, CER: 0.1774
F1-Score: 0.7857
METEOR: 0.5714

**Figure 5** Sample output 1: Hindi caption for a tennis image.

Generated Caption: बेसबॉल खिलाड़ियों का एक समूह काला बल्ला पकड़े हुए है
Reference Caption: काला बल्ला पकड़े बेसबॉल खिलाड़ियों का एक समूह
BLEU-1: 0.8000, BLEU-2: 0.7303, BLEU-3: 0.6465, BLEU-4: 0.5254
CIDEr: 1.0000
WER: 1.0000, CER: 0.9111
F1-Score: 0.8889
METEOR: 0.7927

**Figure 6** Sample output 2: Hindi caption for a baseball image.

Figure 7 shows that the framework effectively captures relationships among multiple objects in a complex indoor environment, resulting in a coherent and descriptive Hindi caption. Certain cases achieved perfect alignment with reference captions (BLEU-4 = 1.0000, METEOR = 0.9688), demonstrating the system's capacity to generate linguistically and culturally appropriate descriptions. Minor variations in domain-specific images suggest potential improvements through expanded multilingual datasets and fine-tuning.

Overall, the proposed model achieves a strong balance between accuracy, interpretability, and computational efficiency. The integration of ethical and inclusive design principles ensures that the generated captions align with linguistic diversity and responsible AI practices.


**Figure 7** Sample output 3: Hindi caption for a kitchen image.

## 7 Conclusion

This work presented an integrated deep-learning framework for automated Hindi image captioning, combining a ConvNeXt-based visual encoder, an adaptive attention mechanism, and a Transformer decoder. The architecture effectively addresses challenges posed by Hindi's morphological complexity by enhancing cross-modal alignment and enabling refined visual–semantic representation learning. Experimental evaluations demonstrate the robustness and competitiveness of the proposed system. The model achieved stable convergence with training and validation accuracies of 77.80% and 77.56%, respectively, and consistently strong caption quality scores, including BLEU-1/2/3/4 values of 0.8646, 0.6282, 0.5401, and 0.4429, a CIDEr score of 0.8158, METEOR of 0.6622, and low WER and CER values. An F1-score of 0.8313 further validates the linguistic coherence of the generated captions. Overall, the results confirm the effectiveness of integrating ConvNeXt with adaptive attention and Transformer-based decoding. The proposed approach offers a scalable foundation for future research in multilingual and low-resource image captioning systems. Despite its promising performance, the proposed framework has certain limitations. The Hindi captions are generated through translation of the MS COCO dataset, and the evaluation is limited to a single benchmark dataset. Future work will focus on developing large-scale manually annotated Hindi caption datasets, evaluating the model on additional multilingual benchmarks, and exploring recent vision-language foundation models to further improve caption quality and generalization.

## References

[1] G. Hoxha, F. Melgani, and J. Slaghenauffi, "A new CNN–RNN framework for remote sensing image captioning," in *Proc. Mediterranean and Middle-East Geoscience and Remote Sensing Symp. (M2GARSS)*, Tunis, Tunisia, 2020, pp. 1–4, doi: 10.1109/M2GARSS47143.2020.9105191.

[2] H. Wang, H. Wang, and K. Xu, "Evolutionary recurrent neural network for image captioning," *Neurocomputing*, vol. 401, pp. 249–256, 2020, doi: 10.1016/j.neucom.2020.03.087.

[3] S. Jaiswal, H. Pallthadka, R. P. Chinchewadi, and T. Jaiswal, "A deep learning model for automatic image captioning using GRU and attention mechanism," *Int. J. Comput. Eng. Res. Trends*, vol. 11, no. 1, pp. 28–36, Jan. 2024, doi: 10.22362/ijcert.v11i1.919.

[4] P. Singh, C. Kumar, and A. Kumar, "Next-LSTM: A novel LSTM-based image captioning technique," *Int. J. Syst. Assur. Eng. Manag.*, vol. 14, pp. 1492–1503, 2023, doi: 10.1007/s13198-023-01956-7.

[5] A. Jamil *et al.*, "Deep learning approaches for image captioning: Opportunities, challenges and future potential," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3365528.

[6] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015.

[7] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, Lille, France, Jul. 2015, pp. 2048–2057.

[8] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint* arXiv:1508.04025, 2015.

[9] P. Anderson, S. Gould, and M. Johnson, "Partially-supervised image captioning", Advances in Neural Information Processing Systems, vol. 31, 2018.

[10] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale", arXiv preprint arXiv:2010. 11929, 2020.

[11] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[12] X. Li *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 12375, Lecture Notes in Computer Science, Springer, Cham, 2020, doi: 10.1007/978-3-030-58577-8_8.

[13] Y.-C. Chen *et al.*, "UNITER: Universal image–text representation learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 12375, Lecture Notes in Computer Science, Springer, Cham, 2020, doi: 10.1007/978-3-030-58577-8_7.

[14] S. Liu, S. Cao, X. Lu, J. Peng, L. Ping, X. Fan, F. Teng, and X. Liu, "Lightweight deep learning model, ConvNeXt-U: An improved U-Net network for extracting cropland in complex landscapes from Gaofen-2 images," *Sensors*, vol. 25, no. 1, Art. no. 261, 2025, doi: 10.3390/s25010261.

[15] Z. Zhou, Y. Yang, Z. Li, *et al.*, "Image captioning with residual Swin Transformer and actor–critic," *Neural Comput. Appl.*, vol. 37, pp. 8019–8031, 2025, doi: 10.1007/s00521-022-07848-4.

[16] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020, doi: 10.1162/tacla00343.

[17] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2022.

[18] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023.

[19] J. Wang, Z. Yang, X. Liu, Z. Wang, and L. Wang, "GIT: A Generative Image-to-Text Transformer for Vision and Language," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022.

[20] T. Xian, Z. Li, C. Zhang, and H. Ma, "Dual global enhanced transformer for image captioning," *Neural Networks*, vol. 148, pp. 129–141, 2022, doi: 10.1016/j.neunet.2022.01.011.

[21] Y. Shi, J. Xia, M. Zhou, and Z. Cao, "A dual-feature-based adaptive shared transformer network for image captioning," *IEEE Trans. Instrum. Meas.*, vol. 73, Art. no. 5009613, pp. 1–13, 2024, doi: 10.1109/TIM.2024.3353830.

[22] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Lecture Notes in Computer

Science, vol. 8693, Springer, Cham, 2014, doi: 10.1007/978-3-319-10602-1_48.

[23] A. Mao, M. Mohri, and Y. Zhong, "Cross-entropy loss functions: Theoretical analysis and applications," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, vol. 202, Proc. Mach. Learn. Res., Jul. 2023, pp. 23803–23828.

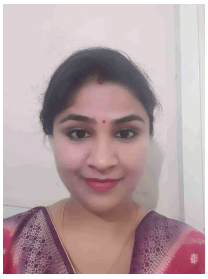
[24] R. Llugsi, S. E. Yacoubi, A. Fontaine, and P. Lupera, "Comparison between Adam, AdaMax and AdamW optimizers to implement a weather forecast based on neural networks for the Andean city of Quito," in *Proc. IEEE Fifth Ecuador Tech. Chapters Meeting (ETCM)*, Cuenca, Ecuador, 2021, pp. 1–6, doi: 10.1109/ETCM53643.2021.9590681.

[25] M. Ghassemiazghandi, "An evaluation of ChatGPT's translation accuracy using BLEU score," *Theory Pract. Lang. Stud.*, vol. 14, no. 4, pp. 985–994, Apr. 2024, doi: 10.17507/tpls.1404.07.

[26] G. Oliveira dos Santos, E. L. Colombini, and S. Avila, "CIDEr-R: Robust consensus-based image description evaluation," in *Proc. 7th Workshop Noisy User-generated Text (W-NUT)*, Nov. 2021, pp. 351–360, doi: 10.18653/v1/2021.wnut-1.39.

[27] H. Saadany and C. Orasan, "BLEU, METEOR, BERTScore: Evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text," *arXiv preprint* arXiv:2109.14250, 2021.

## Biography



**Anjali Sharma** is a dedicated academician with a strong background in Information Technology and Computer Science. She earned her B.Tech. in Information Technology and later completed her M.Tech. in Computer Science from Dr. A. P. J. Abdul Kalam Technical University. She is currently pursuing her Ph.D. from Gurukul Kangri (Deemed to be University), further advancing her research expertise and academic pursuits.

**Mayank Aggarwal** is holds a Ph.D. in Computer Science. He is professor and dean, Faculty of Engineering and Technology, Gurukula Kangri (Deemed to be University), Haridwar. He has 20+ years of experience. He was a recipient of the gold medal and the University Topper in B.Tech., and published several research papers and imparted several trainings in the field of cloud computing in collaboration with IBM for students and faculties throughout the country.