https://doi.org/10.65091/icicset.v2i1.14

# Image Captioning using Transformer Model

Anisha Adhikari
*Department of Software Engineering*
*Nepal College of Information Technology*
*Lalitpur, Nepal*
anisha.221609@ncit.edu.np

Mahigya Dahal
*Department of Software Engineering*
*Nepal College of Information Technology*
*Lalitpur, Nepal*
mahigya.221624@ncit.edu.np

Rudra Nepal
*Department of Software Engineering*
*Nepal College of Information Technology*
*Lalitpur, Nepal*
rudra.nepal@ncit.edu.np

Priya Shilpakar
*Department of Software Engineering*
*Nepal College of Information Technology*
*Lalitpur, Nepal*
priya.221633@ncit.edu.np

***Abstract*—This paper presents a deep learning approach for generating descriptive captions for images through the combination of deep learning driven image understanding and text generation. The rapid progress in deep learning has greatly enhanced machines' capacity to interpret visual information and generate natural-language descriptions. Among these advancements, Transformer architectures have proven especially effective due to their self-attention mechanisms, which enable models to better capture global image, text relationships. Traditional image captioning models that use CNNs for visual encoding and RNNs for sentence generation often struggle to model long-range dependencies and produce contextually rich descriptions.**

**To address these challenges, this study introduces an intelligent image captioning framework that utilizes InceptionV3 for visual feature extraction and a custom-built Transformer-based encoder decoder architecture for generating textual descriptions. Unlike conventional systems that depend on pre-trained language models, the decoder in this work is implemented and optimized exclusively for image captioning. Comprehensive training and evaluation are conducted using a large-scale benchmark dataset, MS COCO, to validate the system's effectiveness across diverse image domains. This paper aims to integrate image understanding and linguistic reasoning by providing a practical framework with applications in accessibility, digital asset management, and automated content generation.**

***Index Terms*—Image Captioning, Transformers, Inception V3, Natural Language Processing, Deep Learning, Computer Vision, Encoder-Decoder Architecture.**

## I. Introduction

Artificial intelligence has made significant progress in both computer vision and natural language processing (NLP), enabling systems to interpret visual content and generate meaningful text. Image captioning, which involves producing natural-language descriptions of visual content, stands as the intersection of these two domains. This task evaluates not only a model's ability to recognize visual elements but also its capability to express them in coherent and contextually relevant sentences. Earlier image captioning systems typically used Convolutional Neural Networks (CNNs) for feature extraction and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) units, to generate sentences. While these approaches achieved early success, they often struggled with long-range dependency modeling, limited parallelization, and inconsistent semantic coherence. With the emergence of Transformers, supported by self-attention mechanisms, captioning architectures became more efficient and accurate. Models such as ViT, CLIP, and BLIP demonstrate the potential of attention-based designs for multimodal tasks. Motivated by these advancements, this study develops an end-to-end image captioning system using InceptionV3 as a pretrained visual encoder and a custom Transformer-based decoder trained specifically for caption generation. The system aims to produce accurate, contextually rich descriptions and is deployed through an interactive web interface, allowing users to upload images and receive real-time captions. This practical integration enhances accessibility and demonstrates, while highlighting real-world applicability of Transformer-based captioning models.

### A. Problem Statement

The rapid growth of visual content across social media, digital publications, and online platforms has created a demand for systems that can automatically describe images in meaningful and accessible ways. Accurate image captioning plays an essential role in information retrieval, content organization, and accessibility—particularly for visually impaired users who rely on descriptive text to interpret visual information. Traditional image captioning methods, primarily based on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) such as LSTMs, often struggle to capture long-range linguistic dependencies and complex visual–semantic relationships. As a result, these models frequently produce captions that are repetitive, generic, or semantically incomplete.

Recent advances in Transformer-based architectures have notable gains in performance across both vision and natural language processing tasks through multi-head attention mechanisms and ability to model global contextual relationships. Despite their advantages, there is still limited exploration of

lightweight and practical CNN–Transformer hybrid models for image captioning, particularly in real-time user environments. This research addresses this gap by implementing and evaluating a Transformer-based image captioning system that integrates a CNN encoder with a Transformer decoder. The study investigates the system's ability to generate accurate, context-rich captions and examines its effectiveness compared to traditional RNN-based approaches.

### *B. Objectives*

- To present and evaluate a Transformer-based image captioning model that generates accurate and context-aware textual descriptions of images.
- To demonstrate the practical applicability of the model through a web-based interface that provides real-time caption generation for uploaded images.

### *C. Research Gaps and Motivation*

Although image captioning systems have advanced significantly, several gaps persist:

- **Task-specific decoder limitations:** Many Transformer-based models depend on pre-trained language decoders that are not optimized for visual grounding, often resulting in generic or partially relevant captions.
- **Limited evaluation on medium-scale datasets:** Most research focuses on large datasets such as MS-COCO, with fewer studies analyzing Transformer performance on smaller datasets suited for academic or resource-restricted environments.
- **Contextual accuracy issues:** Existing models often identify objects but struggle to generate captions that consistently reflect relationships, actions, and scene context.

Motivated by these gaps, this study develops and evaluates an end-to-end system using InceptionV3 for visual feature extraction and a Transformer encoder-decoder to generate text captions that are precise, well-structured, and relevant.

## II. Literature Review

### *A. Evolution of Image Captioning: From CNN–RNN Models to Transformer Architectures*

Image captioning has been a vibrant research area uniting domains of computer vision and natural language processing, evolving significantly over the past decade. Early models primarily used CNN architectures to obtain image representations and RNN-based models, especially LSTM networks, to produce captions in sequence. However, these models often struggled with capturing complex relationships and contextual dependencies in both the visual and textual domains. One foundational work by Vinyals et al. introduced the Show and Tell model, which combined CNNs and LSTMs to generate image captions [1]. This approach demonstrated the feasibility of end-to-end trainable image captioning systems but faced challenges with long-range dependencies and generating diverse, contextually relevant captions.

More recent research has shifted towards Transformer-based architectures, which can capture relationships more effectively due to their self-attention mechanisms. For example, "Attention is All You Need" by Vaswani et al. revolutionized sequence modeling by replacing recurrent structures with Transformers, enabling improved parallelism and contextual understanding [2]. This innovation laid the groundwork for applying Transformer models to vision-language tasks.

Building on this, Dosovitskiy et al. proposed the Vision Transformer (ViT), which directly applies Transformers to image patches, outperforming traditional CNNs on various image recognition benchmarks [3]. Subsequent works such as CLIP by Radford et al. leveraged large-scale contrastive learning to align images and text embeddings, demonstrating powerful zero-shot capabilities in image captioning and retrieval tasks [4].

### *B. Background Study*

Generating precise text to describe visual content is a complex task in the field of image captioning. It involves effectively modeling long-range dependencies, attending to important image regions, and aligning visual and linguistic features. Several studies have proposed methods to address these challenges in different ways. Xu et al introduced visual attention to image captioning, allowing models to focus on salient regions of an image while generating each word [5]. Lu et al introduced adaptive attention, allowing models to dynamically decide when to rely on visual versus linguistic context, improving caption relevance [6]. Anderson et al showed that region-level attention enhances captioning accuracy by focusing on object-specific features [7]. Karpathy and Fei-Fei proposed visual-semantic alignment, pairing image regions with textual phrases to improve caption grounding [8]. Building on these advances, our work employs a Transformer-based architecture to generate captions by integrating both visual and textual features effectively, aiming to improve caption quality and relevance. By studying these various attention-based and visual-semantic models, we developed a customized encoder–decoder framework that achieves efficient caption generation tailored to our dataset.

### *C. Dataset Selection and Preparation*

For initial development and experimentation, the Flickr8k dataset was used to validate the model architecture and training pipeline. Due to its smaller size, Flickr8k enabled rapid prototyping, faster experimentation, and early debugging of the captioning system. Results from Flickr8k are not reported , as the focus of this study is on evaluating model performance on large-scale benchmarks.

After successful validation on Flickr8k, the model was trained and evaluated on the MS COCO dataset, which provides a large-scale and diverse collection of images suitable for assessing model performance and generalization. All quantitative results presented in this paper, including BLEU and METEOR scores, are obtained from evaluation on the MS COCO dataset.

### D. Model Selection: The Transformer

The Transformer model is central to our image captioning system due to its superior ability to model complex dependencies between inputs and outputs. The model was chosen for its architecture, which offers significant advantages over traditional CNN–RNN approaches. While Meshed-Memory Transformers introduce gated multi-level feature interactions that improve the model's ability to capture both local and global visual context, ultimately producing more coherent and detailed captions [9], implementing such a model requires modifying the attention mechanism and handling multiple visual feature levels, which makes it much more complex than a standard ViT+Transformer decoder setup. This additional complexity can be challenging for smaller projects or medium-sized datasets. The model was chosen for its architecture, which offers significant advantages over other approaches:

- **Self-Attention:** Transformers employ self-attention mechanisms to process all elements of input data simultaneously, enhancing efficiency and enabling better contextual understanding compared to sequential models.
- **Suitability for Image Captioning:** The architecture effectively captures relationships among multiple areas in an image, allowing the creation of precise and context-aware textual descriptions
- **Parallel Computation and Long-Range Dependencies:** Transformers can handle long-range dependencies and support parallel computation during training, providing substantial advantages in scalability and training efficiency.

Overall, the Transformer's efficiency, scalability, and strong performance in modeling complex multimodal relationships make it an ideal choice for generating coherent and contextually accurate image captions.

### E. Evaluation Metrics

To evaluate the quality and relevance of the generated image captions, two widely recognized automatic evaluation metrics were employed: BLEU [10] and METEOR [11]. These metrics provide quantitative measures by comparing the generated captions with human-written reference captions.

*1) BLEU (Bilingual Evaluation Understudy Score):* BLEU assesses similarity by comparing sequences of *n*-words in the predicted captions with the reference captions, emphasizing precision. It examines how many terms or expressions in the generated text correspond to those in human-created annotations.

- **Purpose:** BLEU is used to assess the accuracy of word choice and syntactic correctness in generated captions.
- **Rationale:** It is particularly effective for evaluating short, concise, and grammatically precise outputs.

*2) METEOR (Metric for Evaluation of Translation with Explicit Ordering):* METEOR evaluates caption quality by analyzing both precision and recall while taking into account semantic aspects such as related words, word forms, and sequence. This offers a more linguistically informed caption that matches its reference.

- **Purpose:** METEOR is used to assess how well automatically produced captions capture the intended meaning of the reference annotations.
- **Rationale:** It is especially helpful for evaluating fluent, human-like captions that may use different words while preserving the original semantics.

### F. Conclusion of Literature Review

The literature supports the adoption of Transformer-based architectures combined with CNN feature extractors as an effective framework for image captioning. Key factors for success include:

- Effective modeling of long-range dependencies and global image–text relationships.
- Decoder optimization specifically trained for image captioning rather than relying solely on pre-trained language models.
- Evaluation on both medium-scale and large-scale datasets to ensure robustness and generalizability.

This foundation informs the design of the proposed system, enabling the generation of accurate, contextually relevant, and coherent captions across diverse images.

## III. METHODOLOGY

This study presents a Transformer-based image captioning framework capable of generating natural-language descriptions for input images. Unlike approaches that rely heavily on generic pretrained language models for decoding, the proposed method trains the Transformer decoder specifically for the captioning task. Meanwhile, a pretrained InceptionV3 network is used only for visual feature extraction. This task-focused training enables the model to generate captions with more contextually accurate and better reflect the visual information.

### A. Transformer-Based Model Architecture

The central component of the captioning framework is a Transformer-based architecture in which separate encoder and decoder modules cooperate to generate textual descriptions. Instead of processing information sequentially, the model analyzes long-range interactions through layered attention mechanisms, enabling it to understand broader associations between visual representations and the produced sentences.

The overall process within this architecture is summarized below:

1) Every token in the input sentence is first mapped into a vector of dimension $d_{model}$- forming its learned embedding.
2) To preserve word order, each embedding vector is combined with a positional encoding vector of the same dimension through element-wise addition
3) These enriched vectors enter the encoder stack, which is composed of two major sublayers. Through multi-head attention, the encoder considers all tokens in the

sequence simultaneously, enabling it to build a context-aware representation.

4) At each time step $t$, the decoder receives the token predicted during the previous step $t-1$, allowing it to generate output text one word at a time.
5) Similar to the encoder, positional encodings are added to the decoder's input tokens to preserve ordering during generation.
6) The enhanced decoder input passes through three internal sublayers.
   - The first sublayer applies *masked self-attention*, ensuring that the model cannot access future tokens during prediction.
   - The second sublayer incorporates the encoder's output, allowing the decoder to reference the encoded source representations.
   - The third sublayer applies a *feed-forward neural network* to further refine the intermediate representations.
7) After the decoder stack, the resulting vector goes through a fully connected projection layer and then a softmax classifier, which selects the most probable next token in the caption.

### B. Encoder and Decoder

*1) Encoder:* The encoder is composed of a stack of six identical layers ($N = 6$), where each layer contains two sublayers:

1) **Multi-Head Self-Attention:**The first sublayer uses multi-head self-attention, where each head processes projected queries, keys, and values independently. The outputs from all heads are merged and passed through a concluding linear projection to generate the attention output.
2) **Feed-Forward Network (FFN):** The second sublayer uses a position-wise feed-forward network, where the same transformation is applied independently at every sequence location:

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2$$

Each encoder layer has unique weights ($W$) and biases ($b$). To stabilize optimization, every sublayer is wrapped with a residual connection followed by layer normalization:

$$\text{LayerNorm}(x + \text{Sublayer}(x))$$

This allows the encoder to learn progressively abstract representations while maintaining stable training.

*2) Decoder:* The decoder is formed by stacking $N = 6$ identical layers, with each layer consisting of three sequential submodules.

1) **Masked Self-Attention Module:** The first component processes the decoder's previous output sequence, enriched with positional cues, and performs attention over earlier time steps only. The masking mechanism restricts each token to rely solely on information from preceding positions, preventing access to future predictions.
2) **Encoder–Decoder Attention Stage:** The next component carries out an attention operation where the decoder's intermediate representations act as the query signals, while the encoder's output provides the reference information. This step enables the decoder to selectively draw from the visual features extracted by the encoder.
3) **Feed-Forward Network (FFN):** The third component applies the same two-layer nonlinear transformation to every location in the sequence independently, enabling richer representation learning.

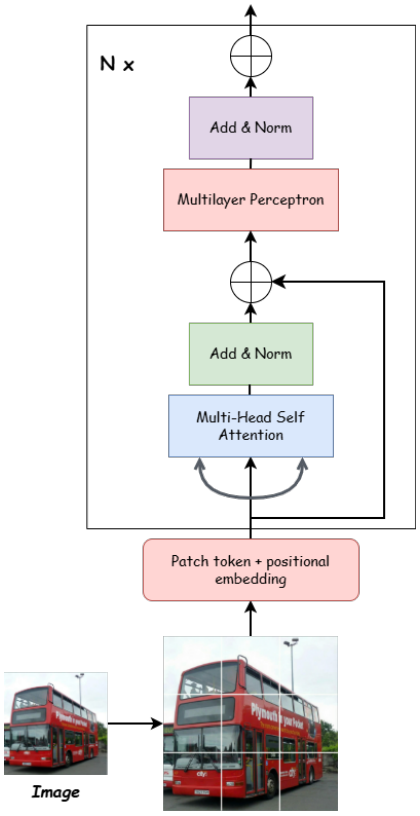


Fig. 1. Encoder

Each component is wrapped with residual pathways and layer normalization to maintain stable optimization. Positional information is injected into the decoder's input vectors to retain awareness of sequence order.

### C. Attention Mechanism

**Self-Attention:** Computes a representation of each token by attending to all other tokens:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

**Multi-Head Attention:** To extend this mechanism, several attention transformations are computed independently using

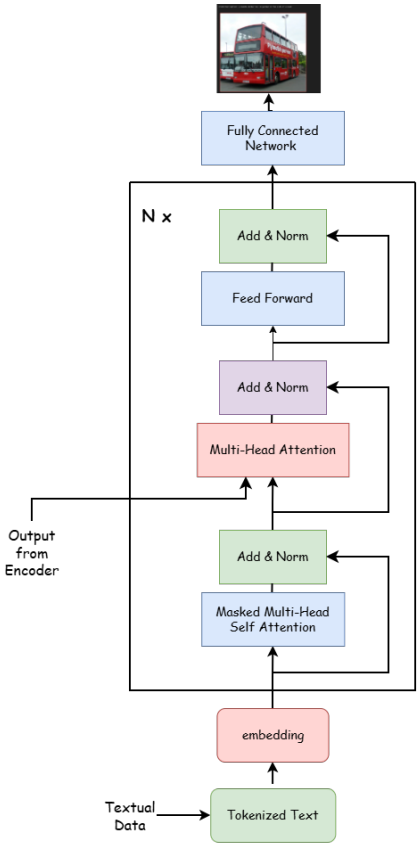


Fig. 2. Decoder

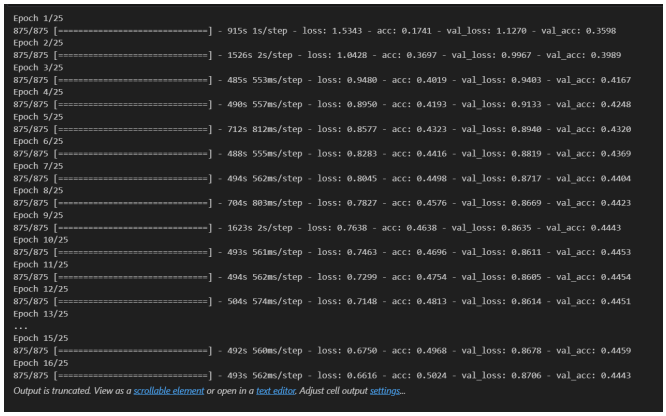


Fig. 3. Screenshot showing model training over multiple epochs.

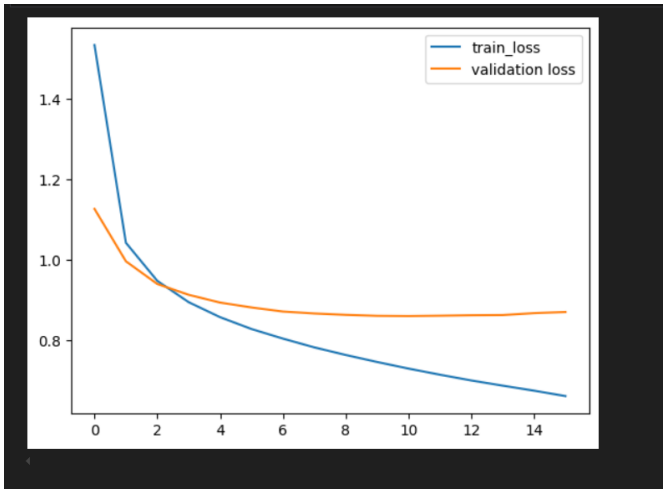


Fig. 4. Training and validation loss curves over epochs.

different learned projections. Their outputs are then merged and linearly mapped to form the final representation:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \quad i = 1, ..., h$$

Using multiple attention heads enables the model to analyze input features from various learned perspectives, allowing it to aggregate complementary relational patterns.

## IV. Results and Discussion

### A. Training Performance

The CNN–Transformer model was trained over multiple epochs with EarlyStopping to prevent overfitting.

Figure 3 shows the training logs, and Figure 4 presents the training and validation loss curves.

- The training loss consistently declined throughout the epochs, ending at 0.15 in the last epoch, which demonstrates successful learning.
- Validation loss closely followed the training loss and stabilized at 0.17. suggesting good generalization.

In general, the convergence pattern shows that the network successfully captured the relationship between visual representations and their corresponding textual descriptions.

*1) **Loss Behavior Across Epochs**:* The training loss started at approximately 1.59 in the first epoch and consistently decreased to around 0.67 by epoch 16. This steady decline indicates that the model was effectively learning visual–text associations over time.

The validation loss also decreased sharply in the early stages, from around 1.18 to roughly 0.87, and then gradually stabilized. After epoch 10, the validation loss maintained a narrow range ($\approx$ 0.87–0.89), showing no signs of sudden increase or divergence.

*2) **Generalization and Overfitting Analysis**:*

- The training and validation loss curves follow a similar downward trend, which indicates healthy learning behavior.
- The difference between the training and validation loss stayed minimal during the entire training process.
- Although the training loss continues to decline, the validation loss plateaus after epoch 10, suggesting that the model has reached an optimal point in terms of generalization.
- The absence of sharp increases in validation loss suggests

| Metric | Score |
|---|---|
| BLEU | 0.1606 |
| METEOR | 0.3615 |

no significant overfitting.

*3) **Overall Observation**:* The stability of the validation loss and the consistent reduction of training loss confirm that the model successfully learned meaningful feature caption mappings while maintaining good generalization capability. This reflects the effectiveness of using a Transformer decoder combined with CNN visual features for the image captioning task.

### *B. Quantitative Evaluation*

The BLEU metric, first proposed by Papineni et al remains one of the most commonly used for automatically generated image descriptions. Both metrics were calculated to measure how well the model's captions correspond to the reference annotations.

The BLEU score of 0.1606 is lower than those reported by state-of-the-art models on the MS COCO dataset, which typically achieve scores between 0.30 and 0.40. This is primarily due to the limited training duration, as the model was not trained for enough epochs to fully converge on the large-scale dataset. Despite this, the METEOR score of 0.3615 suggests that the generated captions capture meaningful semantic content, even when exact n-gram overlap is limited. These results indicate that extended training and further optimization could substantially improve performance.

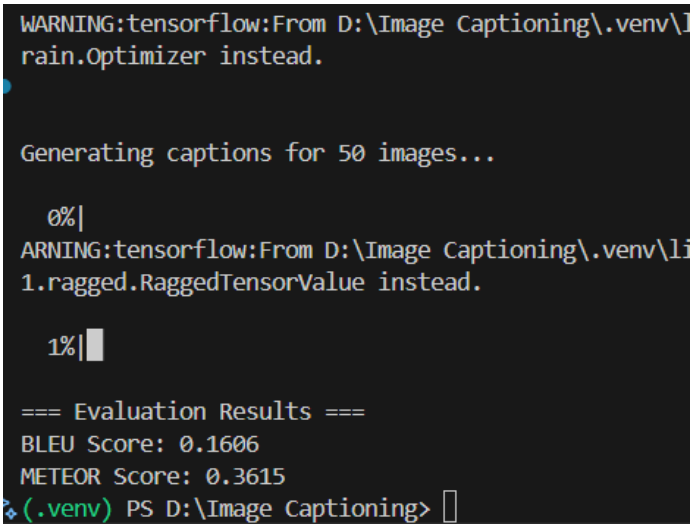


Fig. 5. Screenshot showing BLEU and METEOR scores

### *C. Qualitative Evaluation*

Selected test images were passed through the model to examine caption quality . Observations include:

Sample 1: The model correctly identified the main objects and generated a coherent description.

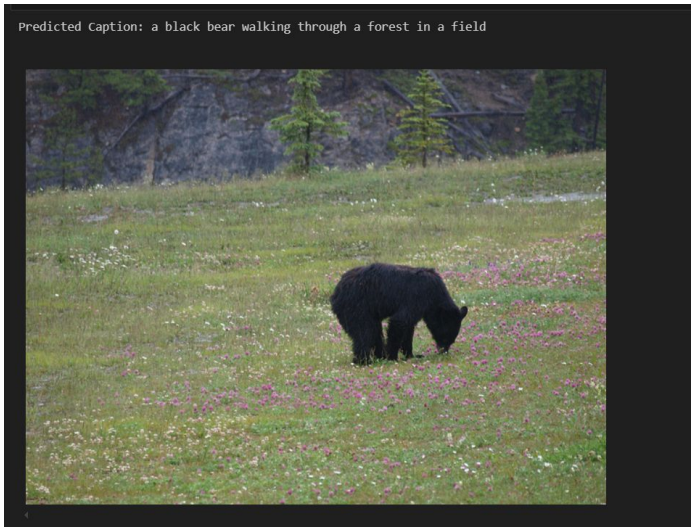


Fig. 6. Sample 1

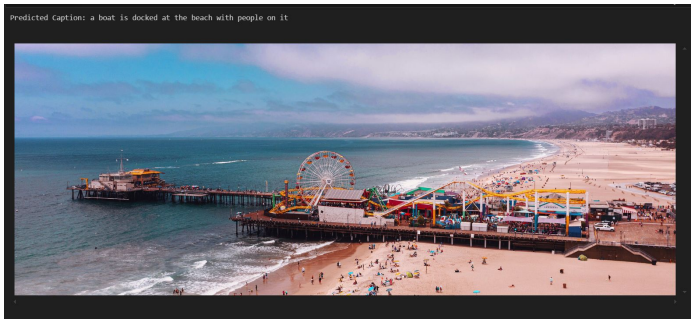


Fig. 7. Sample 2

Sample 2: Generated captions included minor inaccuracies in object relationships, highlighting challenges with complex scenes.

Sample 3: The image caption cityscape accurately recognized the cars and streets in the scene and produced a fluent, contextually appropriate description.

Sample 4: Caption was accurate but less descriptive, reflecting limitations of training data.

Sample 5: The generated caption correctly captured the key objects and the broader visual setting.

## V. Challenges

During the creation of the image captioning model, the team encountered a number of difficulties:

1) Dataset Limitations**:** The quantity and diversity of captions in the dataset impacted the system's performance in generating fine-grained or uncommon text.
2) Complex Scene Interpretation**:** Images containing multiple objects or complex interactions were more difficult for the model, often leading to partial or generic captions.
3) Training Time and Hardware Constraints**:** The Transformer decoder required significant computation, which

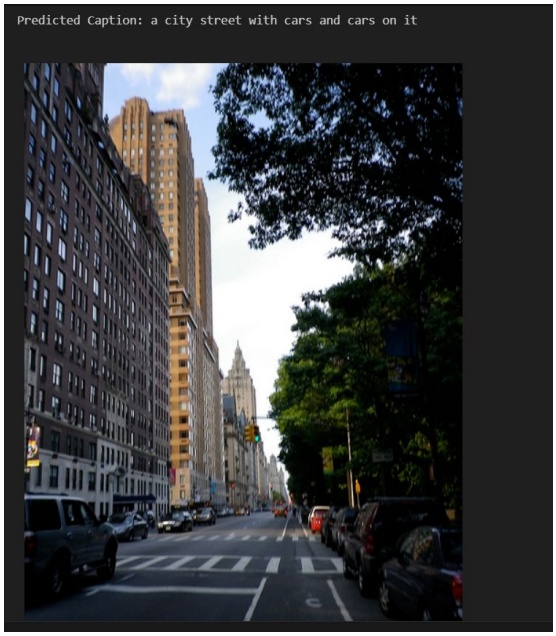


Fig. 8. Sample 3

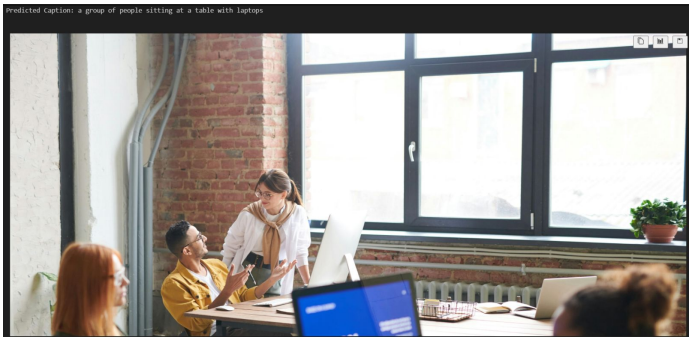


Fig. 9. Sample 4

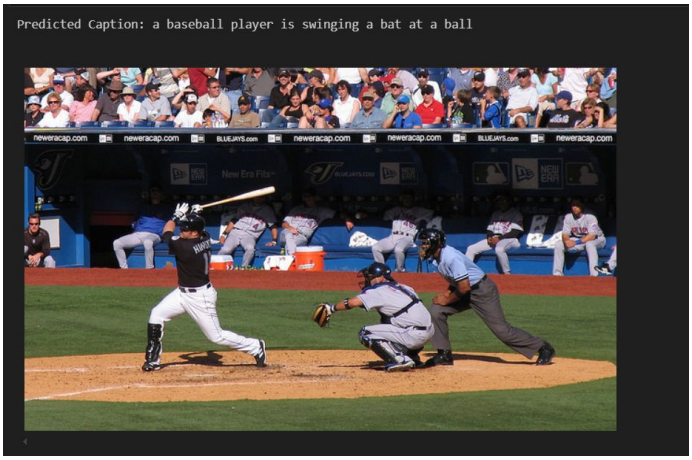


Fig. 10. Sample 5

increased training time and limited experimentation with larger architectures.

4) Vocabulary and Rare Words**:** The model occasionally struggled with rare words or objects that appeared infrequently in the training data.

## VI. Concluding Remarks and Directions

### A. Conclusion

This work presented an image captioning system that integrates CNN-based visual feature extraction with a Transformer decoder for sequence generation. The model effectively learned the mapping between visual inputs and textual descriptions, as demonstrated above.

The evaluation metrics, including BLEU and METEOR scores, indicate that the model is capable of generating meaningful and contextually relevant captions. Overall, the combination of deep visual representation and Transformer-based attention mechanisms proves to be a robust approach for vision-to-language tasks.

Despite certain limitations, such as moderate metric scores and occasional misinterpretation of fine-grained image details, the model provides a strong foundation for further research and improvement.

### B. Future Enhancements

Although the proposed model performs effectively, several further enhancements can be done to improve its accuracy, generalization, and real-world usability:

- **Adopting Multimodal Pretrained Models:** Integrating CLIP, BLIP-2, Flamingo, or LLaVA-style encoders could enhance image–text alignment and zero-shot performance.
- **Beam Search or Sampling-Based Decoding:** Using beam search, nucleus sampling, or contrastive decoding can produce captions that are more fluent and descriptive than greedy decoding.
- **Attention Visualization and Explainability:** Adding attention heatmaps may help analyze why the model produces certain captions, improving interpretability.
- **Supporting Multilingual Caption Generation:** Extending the model to generate captions in Nepali or multiple languages can increase its accessibility and real-world impact.

## References

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2020.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, 2021.

[5] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015, pp. 2048–2057.

[6] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3242–3250.

[7] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and vqa," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6077–6086.

[8] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3128–3137.

[9] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 575–10 584.

[10] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311–318.

[11] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72.