ISSN: 2755-6190

Research Article

Open Access Journal of Artificial Intelligence and Technology

Deep Learning for Amharic Image Captioning: Enhancing Ethiopian Cultural Heritage Accessibility with Optimized Models

Simachew Alamneh[1*] and Mohammed Abebe[2]

[1]*Department of Computer Science, Bonga university, Bonga, Ethiopia*
[2]*Arba Minch Institute of Technology, Arba Minch University, Arba Minch, Ethiopia*

[*]**Corresponding author**
Simachew Alamneh, Department of Computer Science, Bonga University, Bonga, Ethiopia.

Received: February 04, 2026; **Accepted:** February 12, 2026; **Published:** February 19, 2026

ABSTRACT

Visual comprehension in Artificial Intelligence (AI) enables machines to describe images in natural language. However, image captioning research has largely overlooked low-resource languages such as Amharic due to limited domain-specific datasets and linguistic complexity. To address these gaps, this study developed a deep learning-based Amharic image captioning model focused on Ethiopian cultural heritage to promote cultural preservation. A dataset of 4,258 cultural heritage images, each annotated with five expert-verified Amharic captions (21,290 captions in total), compiled from reputable sources, including UNESCO's World Heritage List, Awaze Tours, and Visit Ethiopia. Text preprocessing handled Amharic's morphological complexity through tokenization, stop-word removal, character normalization, and abbreviation expansion. The dataset is divided using a 70:15:15 training, validation, and testing split for balanced model evaluation. The proposed model employs a pre-trained ResNet50 encoder with a GRU decoder as the baseline architecture. Performance is compared using attention and Transformer-based variations, evaluated with BLEU, ROUGE, METEOR, and CIDEr metrics. The ResNet50–GRU baseline with beam search (beam width = 3) achieved the best overall balance between accuracy and efficiency (BLEU-1: 0.5096, BLEU-4: 0.1196, METEOR: 0.2093, CIDEr: 0.2692) among the evaluated models. While the Transformer decoder generated richer captions, its higher computational cost makes the baseline model more suitable for mobile and resource-limited applications. This research demonstrates the potential of deep learning for Amharic image captioning and emphasizes the importance of high-quality datasets and efficient architectures for low-resource languages and cultural heritage preservation.

Keywords: Amharic Image Captioning, Ethiopian Cultural Heritage, Deep Learning, Transformer, Computer Vision, Natural Language Processing

Introduction

The proliferation of visual content, driven by technological advancements and widespread smartphone use, has created both opportunities and challenges in managing and interpreting large-scale image data. Image captioning, integrating computer vision (CV) and natural language processing (NLP), enables machines to generate meaningful textual descriptions of images, a task humans perform intuitively [1]. However, machines primarily extract low-level features, resulting in a persistent "semantic gap" between human perception and machine interpretation [2]. Addressing this gap through automatic captioning improves accessibility, facilitates organization and retrieval of visual data, and supports applications in search, accessibility, and education [3]. It is particularly beneficial for individuals with visual impairments, enhancing accessibility, independence, and social inclusion [4]. Ethiopia, home to over 80 ethnic groups and languages, predominantly speaks Amharic (አማርኛ), with 37.1 million native speakers and 20-25 million second-language users [5,6]. As the country's official language and the second most spoken Semitic language globally, following Arabic, written in the Ge'ez Fidel script as shown in the script in Figure 1, Amharic reflects Ethiopia's rich cultural heritage [7,8]. The nation also hosts 12 UNESCO World Heritage sites [9]. Generating accurate Amharic captions for culturally rich images is challenging due to linguistic complexity and nuanced cultural contexts.

Image captioning seeks to generate accurate textual descriptions of images. Despite substantial progress in this domain, significant challenges remain, particularly for low-resource languages such as Amharic. A primary limitation lies in dataset preparation, as effective captioning models require large, high-quality annotated corpora, which are scarce for underrepresented

Citation: Simachew Alamneh, Mohammed Abebe. Deep Learning for Amharic Image Captioning: Enhancing Ethiopian Cultural Heritage Accessibility with Optimized Models. Open Access J Artif Intel Tech. 2026. 2(1): 1-9. DOI: doi.org/10.61440/OAJAIT.2026.v2.22

languages. Ethiopian cultural heritage is underrepresented digitally due to limited policies, funding, infrastructure, and institutional support [10]. This scarcity hinders the ability to create robust models capable of capturing linguistic nuances and generating meaningful captions [11]. The lack of such datasets constrains progress in cultural heritage research and perpetuates the marginalization of low-resource languages in AI-driven image captioning. A second challenge lies in semantic inconsistency, where captions fail to capture the true essence of images. Despite advances in image captioning for languages like English, many models struggle with semantic accuracy, cultural sensitivity, and generalization to morphologically complex languages such as Amharic [12]. These models rely heavily on large-scale annotated datasets, which are scarce for under-resourced languages (ULs) and often lack standardized orthography, complicating data collection [13]. This study aims to bridge gaps in Amharic image captioning by developing a deep learning encoder-decoder model tailored to Amharic's script and morphology, alongside creating a foundational image-captioning corpus to advance linguistic inclusivity in AI. A third challenge involves generating contextually accurate captions from images, which remains difficult for morphologically rich, low-resource languages such as Amharic. Traditional CNN-RNN architectures struggle with long-range dependencies, complex relationships, attention mechanisms, multi-modal data processing, and scalability [14]. In contrast, Attention-based models like Transformers outperform traditional architectures by dynamically weighting image regions and generalizing across languages. This study addresses challenges in low-resource languages by developing a context-aware Amharic image captioning model, enhancing visual accessibility, and advancing AI for underrepresented languages.

Our Key Contributions are:

- We propose a deep learning–based Amharic image captioning model for Ethiopian cultural heritage, supported by a curated dataset, tailored annotations, and preprocessing.
- We release an open cultural heritage image–caption dataset and integrate Explainable AI through visual attention maps for interpretability.
- We highlight research gaps in low-resource image captioning, offering insights for future ML research on underrepresented languages.

Figure 1: Example of an Image and its five Amharic Captions

Literature Review

This study reviewed existing literature to identify gaps and to develop optimized deep learning models for Amharic image captioning, enhancing accessibility to Ethiopian cultural heritage. Image captioning has evolved from ancient visual narratives, such as Ancient Ethiopian visual narratives, which are renowned for their integration of vivid paintings and inscriptions in Ge'ez or, later, Amharic, to modern AI systems that link images with text. As shown in Figure 2, this progression highlights the enduring human need to connect visuals with language, while current research addresses challenges such as cultural context, dynamic settings, and abstract concepts [15,16].

Image captioning methods have progressed from retrieval-based to template-based approaches. Retrieval-based methods generate captions by matching images with similar ones in a database, but rely on large datasets to cover diverse queries [17]. Template-based methods, using predefined structures such as triplets or quadruplets, produce grammatically correct captions but lack flexibility and natural variability, making them unsuitable for Amharic's morphological richness [18]. Both approaches struggle to produce detailed, context-rich descriptions, highlighting the shift to DNN-based encoder-decoder architectures, which were originally used in machine translation for more state-of-the-art solutions [1].

Figure 2: An Ethiopian meal depicted through painting as a work of art The Amharic text at the top of the painting, "ምግብ ሲመገቡ," translates to "while eating food" [19]. Early efforts in image captioning combined Convolutional Neural Networks (CNNs) for visual feature extraction with Recurrent Neural Networks (RNNs) for sequence generation. For instance, the CNN-RNN framework demonstrated automated captioning and speech generation using Flickr8k, highlighting the potential for assistive applications [20,21]. Despite bridging visual-to-text translation, such models struggled to capture context-rich, human-like descriptions, revealing the need for advanced architectures like attention mechanisms and transformers.

The seminal work Show and Tell introduced CNN-LSTM-based captioning, effective for English datasets [22]. However, its direct encoder-decoder linkage limited selective attention to image regions, constraining semantic richness. Similarly, a Hindi CNN-LSTM model improved BLEU scores through multilayered architectures but remained restricted by translation limitations [23]. These studies collectively underscore the limitations of traditional CNN-RNN/LSTM models in generating contextually nuanced captions.

Recurrent alternatives like Gated Recurrent Units (GRUs) and bidirectional GRUs have addressed efficiency and contextual

understanding the CNN-GRU hybrid incorporated a semantic validator for alignment between captions and images, achieving high BLEU, METEOR, ROUGE-L, and CIDEr-D scores on the MS COCO Dataset [24-26]. Yet, it still lacked attention or transformer integration, highlighting the trade-off between complexity and performance. Attention mechanisms revolutionized captioning, exemplified by Show, Attend, and Tell Allowing dynamic focus on relevant image regions and establishing a foundation for modern attention-based models [27]. Domain-specific applications extended these principles. For example, Bangladeshi cultural heritage captioning adopted CNN-LSTM architectures but underutilized attention mechanisms, while ARTalk for Chinese ceramics combined EfficientNet, CBAM, YOLOv3, and Knowledge Graphs to enhance cultural relevance [28,29]. Although ARTalk achieved high BLEU and CIDEr scores, challenges persisted in abstract pattern recognition and metaphor understanding. Similarly, captioning for Egyptian and Chinese art revealed that attention mechanisms may underperform in cultural heritage contexts, indicating the need for specialized architectural adaptations [30].

Region-specific studies for tourism Applied Efficient Net and Transformer-based architectures, demonstrating promising BLEU and METEOR results [31,32]. However, Direct translations in dataset preparation resulted in a lack of cultural richness and nuance. Language- specific works, including Assamese, Bangla and Nepali confirmed the utility of attention and transformer-based methods in low-resource settings but pointed to challenges in dataset size, annotation diversity, and cross-lingual translation [33-35].

Focusing on Amharic, the hybrid attention-based CNN-Bi-GRU model achieved a 21% 4G- BLEU improvement over baseline methods, showing the significance of domain-specific datasets and visual attention. Yet, reliance on automated translation may introduce cultural bias, necessitating larger and more diverse datasets. Transformer-based models for other low-resource languages, such as Urdu and Bengali End-to-end attention models capture long-range dependencies but are limited by small datasets and narrow multilingual coverage [34,36].

Studies reveal a clear progression from basic CNN-RNN/LSTM models to attention-enhanced and transformer-based architectures, which are particularly effective in low-resource or culturally specific contexts. Despite these advances, three critical gaps remain: existing research rarely addresses Amharic or other Semitic low-resource languages; cultural heritage datasets for languages such as Bangla or Chinese art exist, yet Amharic heritage is largely underrepresented; and although transformer-based models achieve strong performance, their computational efficiency is crucial for deployment in Ethiopia's resource-constrained settings. These challenges highlight the need for developing optimized Amharic image captioning models tailored to Ethiopian cultural heritage.

Methodology

Dataset Collection, Description, and Annotation Quality

The study employed a domain-specific Amharic dataset curated from authentic sources (e.g., UNESCO, Visit Ethiopia) to capture Ethiopia's cultural heritage. Unlike general datasets (e.g., MS COCO, which provides broad coverage, domain-specific data ensures culturally relevant annotations, but also poses challenges in bias, precision, and generalizability, particularly in low-resource language contexts [36]. The dataset comprises 4,258 Ethiopian cultural heritage images annotated with 21,290 Amharic captions, covering attire, food, festivals, landmarks, daily life, handicrafts, performing arts, and natural landscapes. Expert-guided annotations ensured cultural authenticity and linguistic accuracy, making the dataset a balanced and representative resource for Amharic image captioning. The dataset is filtered to retain only high-resolution, culturally authentic images, excluding blurred or watermarked content. Annotation involved 4,258 images labeled by Amharic speakers and cultural experts using a structured framework that ensured both descriptive accuracy and cultural relevance. Each image is annotated with five captions to capture diverse perspectives and enrich semantic depth, following best practices [37].

Data Preprocessing

Image Processing

For Amharic ECH captioning, preprocessing ensures images are suitable for feature extraction through resizing, noise removal, and enhancement. Images were resized to match model-specific input sizes (e.g., ResNet/VGG: 224×224, Xception: 299×299). Enhancement techniques, including histogram equalization and sharpening, improved contrast and fine details, ensuring clearer cultural features for caption generation. Feature extraction transforms preprocessed images into rich visual representations for captioning, in which pre-trained CNNs (VGG16, ResNet50, EfficientNet, Xception) are employed. CNN layers captured hierarchical features (textures, structures, semantics), while Transformers modeled global dependencies. Hybrid CNN-Transformer approaches further enhanced feature richness [38]. These extracted features serve as robust inputs for generating accurate and context-aware Amharic captions.

Text Preprocessing

Text preprocessing is essential for Amharic captioning due to its complex morphology [39]. Key steps include: (i) short form expansion, where abbreviations are simplified and numbers expressed in words for clarity; (ii) character normalization, unifying orthographic variants such as ሀ/ኀ and ጸ/ፀ to reduce redundancy; (iii) punctuation handling, standardizing or removing symbols that add noise; (iv) tokenization and padding, converting captions into word sequences and aligning lengths for model input; and (v) vocabulary optimization, applying a frequency threshold to balance richness and efficiency. Analysis shows most captions fall between 7–13 words, ensuring concise yet expressive descriptions, with a frequency threshold of 2 yielding the best trade-off between vocabulary size and cultural representation.

Experimental Design

The study proposes a deep learning approach for accurate, culturally relevant Amharic captioning of Ethiopian heritage images. Formally, given an input image I, the model aims to generate a caption $W = (W_0, W_1, \ldots, W_n)$, where each word W_t is predicted based on the image features and previously generated words, maximizing the conditional probability $P(W_t \mid I, W_0, \ldots, W_{t-1})$. The architecture employs CNNs (ResNet50, VGG16,

InceptionV3, EfficientNet) for robust visual feature extraction, feeding RNN decoders (LSTM, GRU) enhanced with Bahdanau and Luong attention to generate contextually aligned Amharic captions. To enhance fluency and coherence, a Transformer-based decoder is integrated, aligns them with caption tokens, and RNNs with attention refine sequence generation, yielding culturally accurate, semantically rich Amharic captions. Like this flow: Image → CNN/ → RNN + Attention/Transformer Decoder → Amharic Caption as depicted in in Figure 3.

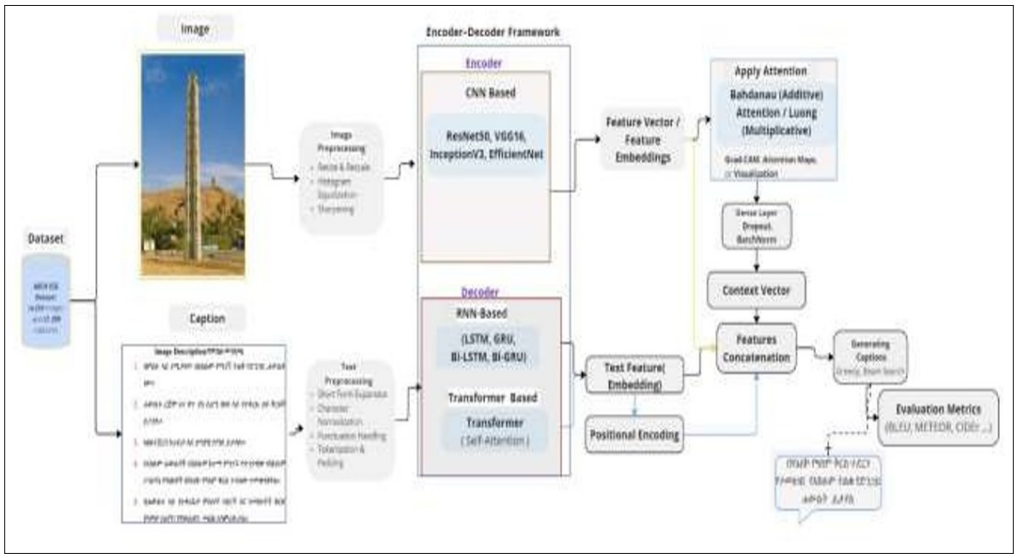


Figure 3: Amharic ECH Image Caption Generation Model

Experimental Results and Discussion

The proposed Amharic ECH image captioning model is developed and evaluated using 4,258 culturally rich image-caption pairs on Google Colab with GPU support. Images were preprocessed through resizing and normalization, while Amharic captions underwent tokenization, character-level normalization, and subword encoding to handle linguistic complexity. Three deep learning architectures, CNN+LSTM, CNN+GRU with attention, and CNN+Transformer, are trained and compared using multiple train-test splits to assess performance. Standard evaluation metrics such as BLEU, METEOR, ROUGE-L, and CIDEr were employed to measure the models' ability to generate accurate and contextually meaningful captions.

Experiment Results

The experiments are organized sequentially to isolate the contribution of each component. We first examined the effect of dataset partitioning, testing multiple splits (80:15:5, 80:10:10, 70:15:15, and 60:20:20) using a ResNet50–LSTM as a baseline model. The 70:15:15 configuration consistently outperformed the alternatives, achieving the strongest BLEU, ROUGE, and CIDEr scores with balanced METEOR performance. Although the 80:10:10 split gave a slightly higher BLEU-3, the difference was negligible, while the 60:20:20 configuration performed the weakest due to reduced training data. Consequently, a 70:15:15 data split ratio is adopted for subsequent experiments.

We then developed a benchmark using custom CNNs trained from scratch with varying feature dimensions (128, 256, 512) fused with an LSTM decoder. All models performed poorly, confirming that CNNs trained on limited datasets cannot capture robust representations, leading to weak caption quality. This result reinforced the necessity of pretrained CNNs for effective captioning in low-resource contexts, motivating the adoption of VGG16, ResNet50, and InceptionV3 as encoder backbones.

Next, we explored whether classical image enhancement methods, including histogram equalization, CLAHE, and sharpening, could improve captioning quality. Despite producing visually sharper images, these methods yielded minimal or negative gains, indicating that pretrained CNNs already incorporate operations akin to contrast normalization and noise reduction. We therefore proceeded with raw or minimally processed images. At this stage, several pretrained CNNs were compared. VGG16 emerged as the best overall trade-off between quality and efficiency, ResNet50 provided superior long-sequence fluency, and ResNet101 delivered semantically rich captions at higher computational cost. MobileNetV2 and EfficientNetB0 proved more efficient but with modest accuracy, suggesting their suitability for resource-constrained environments.

Finally, we evaluated decoder architectures. LSTM, GRU, Bi-LSTM, and Bi-GRU models were paired with VGG16 and ResNet50 encoders. GRU- and LSTM-based models consistently outperformed their bidirectional counterparts, which underperformed due to incompatibility with autoregressive captioning. Among the configurations, ResNet50–GRU delivered the strongest BLEU-4, METEOR, and CIDEr scores and converged more efficiently than the other alternatives, while VGG16–LSTM performed competitively in ROUGE metrics. These findings highlighted the importance of encoder–decoder compatibility, leading us to select ResNet50–GRU as the preferred configuration for subsequent experiments.

Continuing from the experimental analysis, we performed hyperparameter tuning to further refine the ResNet50-GRU architecture, targeting improvements in BLEU, METEOR, ROUGE-L, and CIDEr. An initial search explored embedding size, GRU units, dense layer size, dropout, L2 regularization, learning rate, batch size, and optimizer choice. Bayesian optimization with Optuna was employed for efficiency, leveraging pruning and early stopping across 19 trials [40]. The

best candidate (embedding_dim = 512, GRU = 512, dense_units = 256, dropout = 0.3, L2 = 2.43e-5, learning rate = 0.001, batch size = 16, Adam optimizer) achieved strong performance but showed signs of overfitting beyond epoch 8 and plateaued despite extended training. A refined configuration, derived through trial-and-error, reduced complexity and improved stability, with embedding_dim = 128, GRU = 128, dense_units = 128, dropout = 0.5, L2 = 1e-4, learning rate = 0.0001, batch size = 32, and Adam optimizer. This final setup yielded more robust generalization and consistent convergence, demonstrating the importance of balancing model capacity with regularization in low-resource Amharic captioning tasks.

Results of Baseline Optimized Model (ResNet50-GRU)

The baseline optimized ResNet50-GRU model (vocabulary size: 8,736; max caption length: 23) using Greedy and Beam Search decoding strategies. Beam Search with a width of 3 slightly outperformed Greedy in BLEU (e.g., B1: 0.5096 vs. 0.5016) and METEOR (0.2093 vs. 0.2009), indicating modest improvements in caption quality. However, Beam Search incurred substantially higher inference times (Val+BLEU: 1,387.97s vs. 187.49s). Increasing the beam width to 5 yielded negligible gains, while further increasing the computation time. These results highlight a trade-off between accuracy and efficiency, suggesting Beam Search with beam width (BW=3) for quality-focused tasks and Greedy Search for scenarios prioritizing speed.

Table 1: Baseline Model Performance Comparison of Greedy Search and Beam Search (Beam Width= 3)

Method	B1	B2	B3	B4	R1	R2	MET	CIDEr
Greedy	0.5016	0.2961	0.1868	0.1139	1.0000	1.0000	0.2009	0.2760
(BW=3)	0.5096	0.2993	0.1917	0.1196	1.0000	1.0000	0.2093	0.2692

For Table 1, B1-B4 represent BLEU-1 to BLEU-4 scores, R1–R2 represent ROUGE-1 and ROUGE-2 scores, MET is the METEOR score, and CIDEr is the Consensus-based Image Description Evaluation metric.

Beam Search (beam width = 3) yielded slightly better BLEU and METEOR scores than Greedy Search but at a much higher computational cost. Given its speed efficiency, Greedy Search is more suitable for real-time or resource-constrained applications despite the minor performance trade-off. As shown in Figure 6, over 50 epochs, the model shows a sharp early loss reduction (epochs 0-10) followed by stable convergence, with a small train-validation gap indicating strong generalization. Accuracy improves consistently, with training slightly ahead of validation, suggesting further gains with longer training. The baseline model trained efficiently, converging in ~4,824 s (~1.34 h).

Apply An Attention Mechanism to the Proposed Model

We integrated Bahdanau soft attention into the ResNet50-GRU model, reshaping ResNet50's conv5_block3_out features into 49×2048 patches for fine-grained focus. Caption tokens (128-dim embeddings, GRU-128) were combined with attention-weighted context vectors and processed via ReLU, batch normalization, and dense layers before softmax prediction. Trained with Adam (1e-4), the model achieved BLEU-1/4 scores of 0.4792/0.0902, ROUGE-1/2 of 0.9891/0.9673, METEOR 0.1802, and CIDEr 0.1912, reflecting strong semantic capture but limited higher-order precision. Attention visualizations further improved interpretability by highlighting image regions influencing each word prediction.

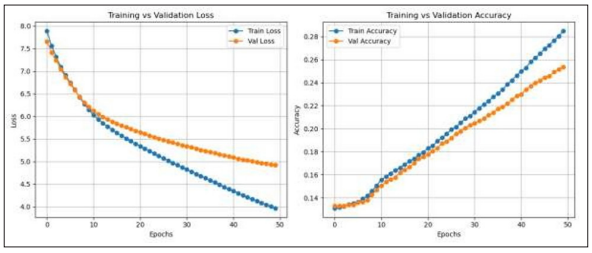


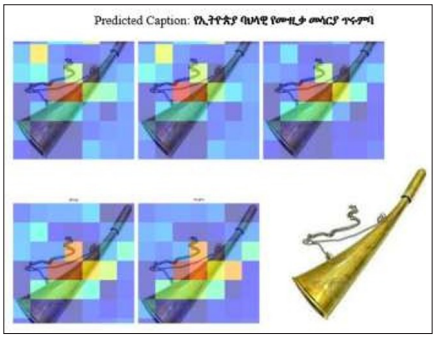
Figure 4: Baseline Model Training Performance: Loss and Accuracy plot.

Table 2: Experimental Result of Bahdanau (Additive) Attention in Image Captioning

BLEU- 1		BLEU- 2	BLEU- 3	BLEU- 4	ROUGE- 1	ROUGE- 2	METEOR
0.4792	0.2547	0.1549	0.0902	0.9891	0.9673	0.1802	0.1912

The results show that the model captures meaningful structure and semantics in captions, though higher-order n-gram precision (BLEU-3/4) remains limited, an expected outcome with attention-based models on small datasets. The added attention visualization (as shown in Figure 5) further strengthens interpretability by revealing how image regions influence each predicted word, thereby clarifying the link between visual and linguistic features.

Figure 5 illustrates the attention-based captioning model applied to two images: a brass horn, where the model focuses on key parts like the bell and chain to generate the Amharic caption "የኢትዮጵያ ባህላዊ የሙዚቃ መሳርያ ጥሩምባ" ("Ethiopian traditional musical instrument, the trumpet"), and an elderly man in traditional attire, with attention on his face, hat, and clothing producing "የገዳ ስርአት በባህላዊ የፀጉር አሰራር" ("Gada system in traditional hair styling"). These visualizations demonstrate how the model selectively attends to relevant regions during word prediction. Training and validation metrics show steadily increasing accuracy and decreasing loss, with a slight gap between training and validation accuracy indicative of generalization rather than overfitting, suggesting that continued training or hyperparameter tuning could further enhance performance.



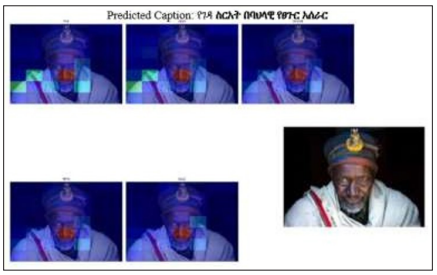


Figure 5: Predicted Captions with Attention Visualization

Figure 6 depicts the model's training progress over 40 epochs, showing a steady decline in both training and validation loss alongside a consistent rise in training and validation accuracy, indicating effective learning and improved performance, with convergence patterns suggesting minimal overfitting.

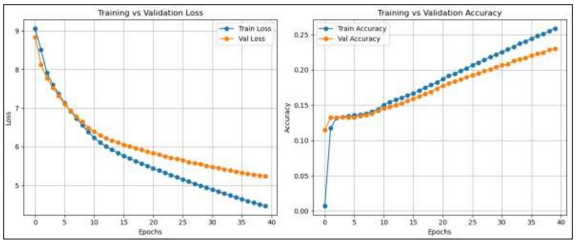


Figure 6: Model Performance: Training and Validation Loss and Accuracy Curves

Apply the Transformer Decoder on the Proposed Model

In this experiment, we built an image captioning model by combining a ResNet-50 encoder and a Transformer decoder to balance visual feature extraction with long-range sequence modeling. The encoder is pretrained on ImageNet, modified with a projection and Layer Normalization, and adapted for Grad-CAM explainability. The transformer model is configured with an embedding size of 256, a hidden size of 512, 4 attention heads, 3 layers, a maximum sequence length of 23, trained for 28 epochs with a learning rate of 1e-4. The Transformer decoder employed multi-head attention, positional encoding, and causal masking for autoregressive generation. Trained end-to-end with a frozen encoder, the model achieved BLEU-1 to BLEU-4 scores of 0.4078, 0.2768, 0.2052, and 0.1634, respectively, reflecting strong word-level accuracy with natural decline at higher n-grams. Captions were generally fluent and coherent, with occasional semantic drift, and attention visualizations confirmed effective grounding of language in visual regions. Overall, the results highlight robust generalization and contextual relevance despite dataset limitations as presented in Figure 7.

Figure 7 shows the strengths and limitations of the Transformer-based model. Image A was accurately captioned, reflecting traditional Ethiopian cotton spinning. Image B's caption was generic and failed to identify the Oromo Gada elder. Image C recognized a work of art but missed the specific woven basket (mesob), and Image D broadly described Ethiopia's cultural heritage without specifying the Lalibela rock-hewn churches. Overall, captions were well- structured and contextually relevant but sometimes lacked specificity, highlighting the model's data demands and the dataset's heterogeneity.

Figure 7: Sample images with captions generated by the Transformer decoder model.

Result Discussion and Comparative Analysis of all Models

In this study, we compared three image captioning encoder-decoder architectures: (1) ResNet50–GRU baseline, (2) ResNet50–GRU with additive visual attention, and (3) ResNet50–Transformer Decoder. The BLEU scores for each model are summarized in Table 3 below.

Table 3: BLEU Score Comparison of Image Captioning Models

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
ResNet50–GRU Baseline	0.5096	0.2993	0.1917	0.1196
ResNet50–GRU + Additive Attention	0.4792	0.2547	0.1549	0.0902
ResNet50–Transformer Decoder	0.4078	0.2768	0.2052	0.1634

The evaluation highlighted trade-offs among the models. ResNet50–GRU achieved the highest BLEU-1 and BLEU-2 scores but lagged on BLEU-3 and BLEU-4. GRU with additive attention performed slightly worse, likely due to overfitting. The Transformer excelled in BLEU-3 and BLEU-4, capturing long-range dependencies, but had lower BLEU-1 and BLEU-2 scores and higher computational cost. Beam search slightly improved results at the expense of speed. Overall, the Transformer generated richer captions, while ResNet50–GRU was more efficient for real-time deployment.

Comparison Of the Caption Quality of Our Model and A Large Language Model (Llm)

It is worth noting that recent advances in Large Language Models (LLMs) offer new opportunities for image captioning, including for LRLs such as Amharic. LLMs can generate contextually rich and semantically coherent captions by leveraging extensive pretraining on large text corpora. To explore this, we experimented with Google Gemini by uploading an image (Figure 5.7) and using the prompt, "Generate a caption for this image." The model produced captions that emphasized general features such as color, mood, or setting, for example, "A vibrant

orange outfit brings joy and highlights the charm of its traditional setting."

In contrast, our proposed model generates concise, culturally specific captions, such as "የጉራጌ ባህላዊ ልብስ (Gurage traditional dress)," accurately identifying the attire worn by the Gurage community. This highlights our model's ability to capture domain-specific cultural knowledge rather than focusing on general descriptive or emotional aspects. While LLMs hold promise for richer, more elaborate captioning, their high computational requirements can limit practicality for mobile or resource-constrained applications.

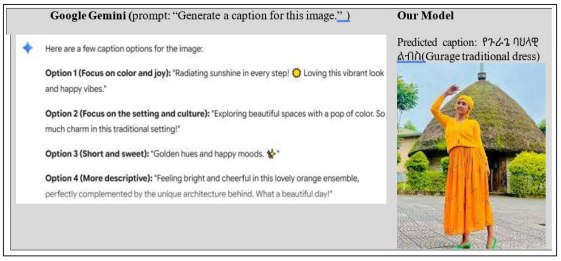


Figure 8: Comparison of generated captions: our model vs. Google Gemini LLM

Comparative Analysis with Other Studies

Our model shows promising performance on the ECH dataset (4258 images) spanning diverse cultural themes. Despite a BLEU-4 of 0.1196, lower than the Local SOTA Hybridized Attention model on Flickr8k (0.388), this result is notable given the high vocabulary size (8,736 tokens), long captions (up to 23 tokens), and the morphological complexity of Amharic as a low-resource language [1]. As Table shows, specialized datasets like Ancient Artworks Captioning (BLEU-4 = 0.25) and low-resource Arabic models (BLEU-4 ≤ 0.092) also struggle, highlighting the challenge of generating culturally and semantically appropriate captions. Specific datasets often fail to generalize to new domains, as they rely on patterns unique to their training data, leading to performance drops in different contexts [41]. Our model benefits from multiple captions per image, capturing diverse cultural perspectives and producing contextually meaningful outputs. While there remains a gap with global SOTA in generic image captioning (BLEU-4 > 0.4), our low-resource, low-computation approach demonstrates the feasibility of heritage-focused captioning. This work provides a foundational benchmark, and future improvements could leverage attention mechanisms, bidirectional layers, or transformer-based architectures while preserving domain specificity.

Table 4: Comparative Analysis with Other Studies

Study	Dataset	Dataset Type	Dataset Size	Captions per	Performance (BLUE Score)				Notes / Remarks
				Image	B-1	B-2	B-3	B-4	
[42]	Ancient Artworks	Specific	17,940 images	Up to 5	0.42	0.33	0.28	0.25	English; LSTM-MC-OUT; focuses on historical and symbolic meanings.
[43]	Arabic- COCO +Arabic Flickr8k	General	88,783 images	5 (COCO) , 3	0.39	0.24	0.15	0.09	Arabic; OSCAR with AraBERT; uses object tags; needs native Arabic datasets.
[44]	Arabic Flickr8k	General	8,000 images	(Flickr8 k) 3	0.36	0.21	0.12	0.06	VGG16 + GRU; Preprocessing emphasis; small dataset limitation.
Local SOTA [1]	Flickr8k dataset	General	8,000+mages	5	0.61	0.50	0.43	0.38	CNN + visual attention + Bi- GRU trained on translated English captions.
Our model	Amharic ECH	Specific	4258 images	5	0.509	0.299	0.191	0.119	low-resource baselines; cultural heritage focus.

Abbreviations: B1–B4 = BLEU-1 to BLEU-4

Conclusions and Future Work

This study addressed the challenge of generating Amharic image captions for Ethiopian cultural heritage using deep learning, focusing on dataset creation, model evaluation, and practical deployment for heritage and tourism applications. A dedicated dataset of 4,258 images with 21,290 culturally relevant captions was developed, mitigating the scarcity of labeled data and addressing the linguistic complexity of Amharic. Evaluation using BLEU, ROUGE, METEOR, and CIDEr revealed that no single metric fully captured cultural accuracy. Among CNN encoders, ResNet50 provided the best balance of feature quality, efficiency, and performance. The ResNet50–GRU baseline achieved the highest BLEU scores and was practical for real-time deployment, while the Transformer produced richer captions at higher computational cost, highlighting the trade-off between descriptiveness and efficiency.

For future work, expanding and diversifying datasets is essential to improve generalization and reduce bias. Developing bilingual or multilingual captioning systems in Amharic, English, and other local languages could broaden accessibility and tourism impact. Integrating speech synthesis with text captioning may enhance inclusivity for visually impaired or low-literacy users. Additionally, exploring fully Transformer-based end-to-end architectures holds potential for improved performance by jointly learning visual and linguistic representations. These directions can foster more accurate, accessible, and culturally rich captioning systems, supporting the preservation and promotion of Ethiopian cultural heritage through technology.

References

1. Solomon R, Abebe M. "Amharic Language Image Captions Generation Using Hybridized Attention-Based Deep Neural Networks," Applied Computational Intelligence and Soft Computing. 2023. 2023: 1-11.
2. Henderson M, Tarr M, Wehbe L. "Interpretable mid-level encoding models of human visual cortex reveal associations

between feature and semantic tuning for natural scene images," J Vis. 2022. 22: 4118.
3. Mishra YV. "Image Captioning: A Comprehensive Review," Int J Res Appl Sci Eng Technol. 2024. 12: 144-148.
4. Kavitha R, Sandhya SS, Betes P, Rajalakshmi P, Sarubala E. "Deep learning-based image captioning for visually impaired people," in E3S Web of Conferences, EDP Sciences. 2023.
5. Yigezu M. "Ethiopia: Language Situation," in Encyclopedia of Language & Linguistics, Elsevier, 2006. 235-237.
6. "Amharic-Worldwide distribution." Accessed. 2024.
7. Asker L, Argaw AA, Gambäck B, Eyassu Asfeha S, Nigussie Habte L. "Classifying Amharic webnews," Inf Retr Boston. 2009. 12: 416-435.
8. Meyer R, Wakjira B. "Scripts and writing in Ethiopia," in The Oxford Handbook of Ethiopian Languages, Oxford University Press. 2023. 86-100.
9. "Ethiopia Laws - UNESCO World Heritage Convention." Accessed. 2025.
10. Adane A, Chekole A, Gedamu G. "Cultural Heritage Digitization: Challenges and Opportunities," Int J Comput Appl. 2019. 178: 1-5.
11. Hadi M, Safder I, Waheed H, Zaman F, Aljohani NR, et al. "A transformer-based Urdu image caption generation," J Ambient Intell Humaniz Comput. 2024. 15: 3441-3457.
12. Yeshambel T, Mothe J, Assabie Y. "Amharic Adhoc Information Retrieval System Based on Morphological Features," Applied Sciences. 2022. 12: 1294.
13. Leivada E, D'Alessandro R, Grohmann KK. "Eliciting Big Data from Small, Young, or Non-standard Languages: 10 Experimental Challenges," Front Psychol. 2019. 10.
14. Kim E, Bae J, Shim I. "Parallel Recurrent Module with Inter-Layer Attention for Capturing Long-Range Feature Relationships," IEEE Access. 2022. 10: 61960-61969.
15. Bosc-Tiessé C. "Le site rupestre de Qorqor (Gar'āltā, Éthiopie) entre littérature et peinture. Introduction à l'édition de la Vie et des miracles de saint Daniel de Qorqor et aux recherches en cours," Afriques. 2021.
16. Wang S, Zhuang J. "Advancements in Deep Learning-Based Image Captioning," Transactions on Computer Science and Intelligent Systems Research. 2024. 5: 464-469.
17. Ordonez V, Han X, Kuznetsova P, Kulkarni G, Mitchell M, et al. "Large Scale Retrieval and Generation of Image Descriptions," Int J Comput Vis. 2016. 119: 46-59.
18. Ushiku Y, Yamaguchi M, Mukuta Y, HaradaT. "Common Subspace for Model and Similarity: Phrase Learning for Caption Generation from Images," in 2015 IEEE International Conference on Computer Vision (ICCV), IEEE. 2015. 2668-2676.
19. "The Art of the Meal." Accessed. 2025.
20. Sudhakar J, Iyer VV, Sharmila ST. "Image Caption Generation using Deep Neural Networks," in 2022 International Conference for Advancement in Technology (ICONAT), IEEE. 2022. 1-3.
21. Hodosh M, Young P, Hockenmaier J. "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics," Journal of Artificial Intelligence Research. 2013. 47: 853-899.
22. Vinyals O, Toshev A, Bengio S, Erhan D. "Show and tell: A neural image caption generator," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. 2015. 3156-3164.
23. Poddar AK, Rani R. "Hybrid Architecture using CNN and LSTM for Image Captioning in Hindi Language," Procedia Comput Sci. 2023. 218: 686-696.
24. Zhou Z, Xu L, Wang C, Xie W, Wang S, et al. "An Image Captioning Model Based on Bidirectional Depth Residuals and its Application," IEEE Access. 2021. 9: 25360-25370.
25. Ahmad RA, Azhar M, Sattar H. "An Image captioning algorithm based on the Hybrid Deep Learning Technique (CNN+GRU)," in 2022 International Conference on Frontiers of Information Technology (FIT), IEEE. 2022. 124-129.
26. Lin TY, Maire M, Belongie S, Hays J, Perona P, et al. "Microsoft COCO: Common Objects in Context". 2014. 740-755.
27. Xu K, Ba J, Kiros R, Cho K, Courville A, et al. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in Proceedings of the 32nd International Conference on Machine Learning, F. Bach and D. Blei, Eds., in Proceedings of Machine Learning Research. 2015. 37: 2048-2057.
28. Alam S, Islam K, Sharmila N, Sovon ZR, Rahman RM. "Image Captioning-Bangladesh's Heritage Perspective Using Deep Learning," in 2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS). 2022. 1-8.
29. Zheng B, Liu F, Zhang M, Zhou T, Cui S, et al. "Image captioning for cultural artworks: a case study on ceramics," Multimed Syst. 2023. 29: 3223-3243.
30. Sheng S, Moens MF. "Generating Captions for Images of Ancient Artworks," in Proceedings of the 27th ACM International Conference on Multimedia, in MM '19. New York, NY, USA: Association for Computing Machinery. 2019. 2478-2486.
31. Fudholi DH, Windiatmoko Y, Afrianto N, Susanto PE, Suyuti M, et al. "Image Captioning with Attention for Smart Local Tourism using EfficientNet," IOP Conf Ser Mater Sci Eng. 2021. 1077: 012038.
32. Dittakan K, Prompitak K, Thungklang P, Wongwattanakit C. "Image caption generation using transformer learning methods: a case study on instagram image," Multimed Tools Appl. 2023. 83: 46397-46417.
33. Das R, Singh TD. "Assamese news image caption generation using attention mechanism,"
34. Multimed Tools Appl. 2022. 81: 10051-10069.
35. Palash AH, Al Nasim A, Saha S, Afrin F, Mallik R, et al. "Bangla Image Caption Generation Through CNN-Transformer Based Encoder-Decoder Network". 2022. 631-644.
36. Subedi B, Krishna Bal B. "CNN-Transformer based Encoder-Decoder Model for Nepali Image Captioning," in Proceedings of the 19th International Conference on Natural Language Processing (ICON), Md. S. Akhtar and Chakraborty T, Eds., New Delhi, India: Association for Computational Linguistics. 2022. 86-91.
37. Hadi M, Safder I, Waheed H, Zaman F, Aljohani NR, et al. "A transformer-based Urdu image caption generation," J Ambient Intell Humaniz Comput. 2024. 15: 3441-3457.
38. Kamel MM, Gil-Solla A, Guerrero-Vásquez LF, Blanco-Fernández Y, Pazos-Arias JJ, et al. "A Crowdsourcing Recommendation Model for Image Annotations in Cultural Heritage Platforms," Applied Sciences. 2023. 13: 10623.

39. Liang S, Hua Z, Li J. "Hybrid transformer-CNN networks using superpixel segmentation for remote sensing building change detection," Int J Remote Sens. 2023. 44: 2754-2780.
40. Belay TD, Ayele AA, Gelaye G, Yimam SM, Biemann C. "Impacts of Homophone Normalization on Semantic Models for Amharic," in 2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA), IEEE. 2021. 101-106.
41. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. "Optuna: A Next-generation Hyperparameter Optimization Framework," Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019.
42. Chan DM, Myers A, Vijayanarasimhan S, Ross DA, Seybold B, et al. "What's in a Caption? Dataset-Specific Linguistic Diversity and Its Effect on Visual Description Models and Metrics," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE. 2022. 4739-4748.
43. Cetinic E. "Iconographic Image Captioning for Artworks," in ICPR Workshops. 2021.
44. Emami J, Nugues P, Elnagar A, Afyouni I. "Arabic Image Captioning using Pre-training of Deep Bidirectional Transformers," in Proceedings of the 15th International Conference on Natural Language Generation, S. Shaikh, T. Ferreira, and A. Stent, Eds., Waterville, Maine, USA and virtual meeting: Association for Computational Linguistics. 2022. 40-51.
45. Hejazi H, Shaalan K. "Deep Learning for Arabic Image Captioning: A Comparative Study of Main Factors and Preprocessing Recommendations," International Journal of Advanced Computer Science and Applications. 2021. 12.