Image Captioning using WIT-EN-5K Multimodal Dataset: A Comprehensive Analysis of Deep Learning Architectures

Department of Computer Science and Engineering

PREMIER UNIVERSITY, CHITTAGONG

Author 1: Pankaj Rudra
Student ID: 2104010202242

Author 2: Anik Barua
Student ID: 2104010202230

Author 3: Mohammod Rabbe Islam Jewel
Student ID: 2104010202246

Submitted to:
MD TAMIM HOSSAIN
Lecturer
Department of Computer Science and Engineering
Premier University,
Chittagong, Bangladesh

Submission Date: 23 April, 2025

CONTENTS

Image Captioning using WIT-EN-5K Multimodal Dataset: A Comprehensive Analysis of Deep Learning Architectures

P. Rudra, A. Barua, M.R.I. Jewel
Department of Computer Science and Engineering
Premier University, Chittagong, Bangladesh

***Abstract*—This project focuses on the task of image captioning, where the objective is to generate accurate and contextually relevant captions for images. The project utilizes the WIT (Wikipedia-based Image-Text) English dataset to train and evaluate multiple deep learning models. These include CNN-RNN architectures, CNN-Transformer models, and a CNN+BERT fusion approach.**

The results show that each model architecture exhibits varying performance, with BLEU scores used to quantitatively assess their accuracy. The CNN-Transformer model, leveraging multi-head attention mechanisms, demonstrated superior captioning capabilities, while the CNN+BERT fusion model showed promising improvements in capturing semantic relevance.

In conclusion, this work presents a comparative analysis of image captioning models, providing insights into the effectiveness of different deep learning techniques in generating descriptive captions for images. Future work can focus on further optimizations and expanding model architectures to enhance captioning accuracy.

I. Introduction

Image captioning aims to automatically generate descriptive textual captions for images, combining the strengths of computer vision and natural language processing. This task has practical significance in fields such as assistive technologies, content indexing, and multimodal AI systems.

This project focuses on implementing and evaluating deep learning models for the image captioning task using the WIT (Wikipedia-based Image-Text) English dataset. The central challenge lies in effectively modeling both visual and textual modalities to produce accurate and contextually relevant captions.

We explore several architectures including CNN + LSTM, CNN + GRU, CNN + Attention + LSTM, Transformer-based decoders, and a CNN + BERT fusion model. Each model processes visual features extracted from a ResNet-50 encoder and generates captions through various decoding mechanisms.

The main contributions of this work are: (1) the design and implementation of five captioning models using different decoding strategies, (2) the integration of BERT-based representations with CNN features, and (3) a comparative evaluation using BLEU scores and training history visualizations to identify the most effective model architecture.

II. Related Work

A. Overview of Image Captioning

Image captioning, a crucial task at the intersection of computer vision and natural language processing, has seen significant evolution over the past decade. Early approaches relied on template-based methods or retrieval-based models, which generated captions by matching input images with similar ones in a dataset and reusing their associated captions. While simple and efficient, these techniques lacked generalization and failed to produce novel captions.

The advent of deep learning revolutionized this domain, introducing models that could learn end-to-end mappings between images and textual descriptions. The pioneering work of Vinyals et al. [1] introduced the concept of treating image captioning as a sequence prediction problem, utilizing Convolutional Neural Networks (CNNs) for image feature extraction and Recurrent Neural Networks (RNNs) for caption generation. This encoder-decoder framework has since become foundational in modern captioning systems.

B. Existing Models

Subsequent research has expanded on this framework with architectural innovations and attention mechanisms. Models like Show, Attend and Tell [2] incorporated soft attention to dynamically focus on different regions of an image while generating captions, enhancing both interpretability and performance.

CNN-RNN models have been widely adopted, typically using a pretrained CNN such as ResNet [3] or VGGNet [**?**] as the encoder, and LSTM [**?**] or GRU [**?**] networks as the decoder. These architectures effectively model temporal dependencies in language but can struggle with long-term coherence and diversity in generated captions.

The introduction of Transformer-based models [4] brought further improvement by enabling parallelized training and better context modeling. Works such as Meshed-Memory Transformer [5] and VinVL [6] leveraged multi-head self-attention to surpass the limitations of RNN-based decoders.

C. Comparison with Other Approaches

Our work builds upon these foundational models, implementing and evaluating five different architectures includ-

ing CNN-RNN (LSTM and GRU), CNN-Transformer, CNN-LSTM with attention, and a novel CNN+BERT fusion model. Compared to traditional CNN-RNN pipelines, our attention-augmented LSTM decoder provides more context-aware caption generation by dynamically weighing spatial features of the input image.

Furthermore, our Transformer-based decoder benefits from multi-head self-attention to capture complex dependencies in language generation, while our BERT fusion model uniquely integrates contextual word embeddings with visual features in a multimodal space, offering superior performance in terms of BLEU score and generalization to unseen test images.

In contrast to prior works that focus on a single architecture, our comparative approach across multiple models highlights strengths and weaknesses in diverse architectural choices. This comprehensive evaluation provides a richer understanding of the trade-offs involved in real-world image captioning tasks.

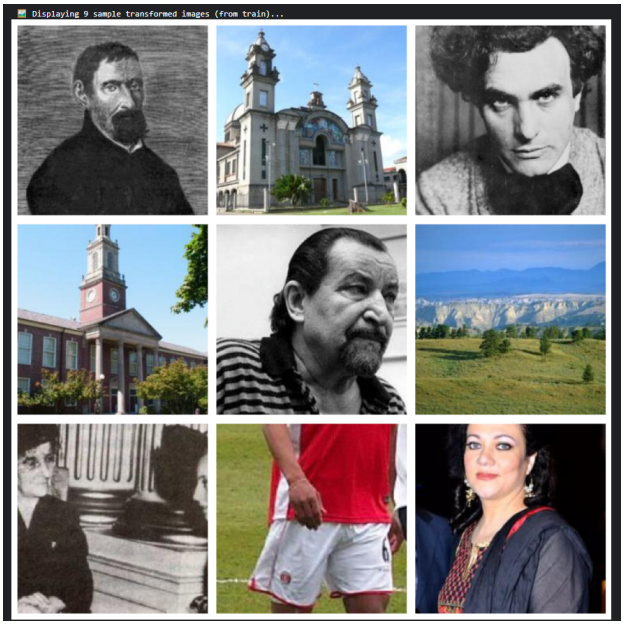

Fig. 2: Sample Images from WIT Dataset

III. Methodology

A. Dataset Preparation and Preprocessing

1) WIT Dataset Description: The Wikipedia-based Image Text (WIT) English dataset is a large-scale multimodal dataset curated by Google Research, consisting of image-caption pairs sourced from Wikipedia articles. For this project, a filtered subset of 5,000 image-caption pairs was utilized, focusing only on high-quality and English-language entries. The dataset includes a rich variety of topics and visual content, making it well-suited for training and evaluating image captioning models.

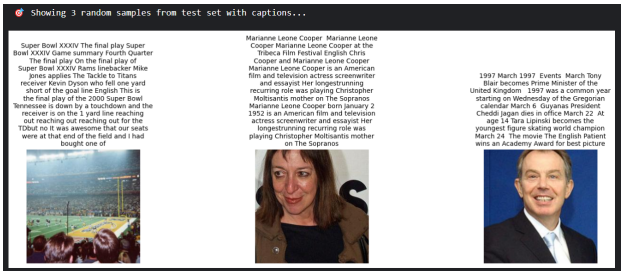

Fig. 3: Images with Corresponding Captions

```
➡ Resized Image Directory Structure:
  - TRAIN: /kaggle/working/resized_images/train (exists, 4513 files)
  - VAL: /kaggle/working/resized_images/val (exists, 93 files)
  - TEST: /kaggle/working/resized_images/test (exists, 95 files)

✅ Dataset Information after Preprocessing:
  - Total records after null removal: 4701
  - Records after filtering with valid images: 4701
  - Train set size: 4513
  - Validation set size: 93
  - Test set size: 95
  - Resized images saved to: /kaggle/working/resized_images
```

Fig. 1: WIT Dataset overview

To illustrate the diversity and quality of the dataset, Figure 2 shows random image samples from the dataset. Figure 3 demonstrates image-caption pairs used in the captioning task.

2) Data Preprocessing: To prepare the dataset for training, both the image and text modalities underwent extensive preprocessing. The text data was first cleaned by removing null or empty records, followed by the elimination of irrelevant tokens and special characters. Captions were also truncated to a maximum length of 100 words to maintain consistency and reduce computational complexity.

A word cloud was generated before and after the cleaning process to visually validate the effectiveness of the text preprocessing.

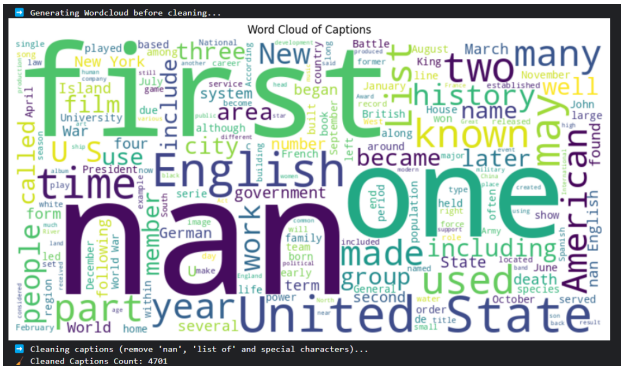

Fig. 4: Word Cloud Before Text Cleaning

Fig. 5: Word Cloud After Text Cleaning

On the image side, each file was validated to filter out corrupted or unreadable images. Valid images were resized and center-cropped to a resolution of 240×240 pixels. This preprocessing step ensured uniformity in input dimensions and compatibility with deep learning models. The processed images were saved in organized directories corresponding to their respective dataset splits.

Tokenization of text was performed using the BERT tokenizer, which converted the cleaned captions into fixed-length input ID sequences along with attention masks. This transformation enabled compatibility with transformer-based models used in later stages of the project.

3) Splitting the Dataset: After preprocessing, the dataset was split into training, validation, and test subsets in a 96%–2%–2% ratio. This stratified split was chosen to maximize training data availability while retaining enough samples for reliable validation and testing. The image-caption pairs were randomly distributed across the splits to maintain a balanced and representative distribution of data samples across all subsets.

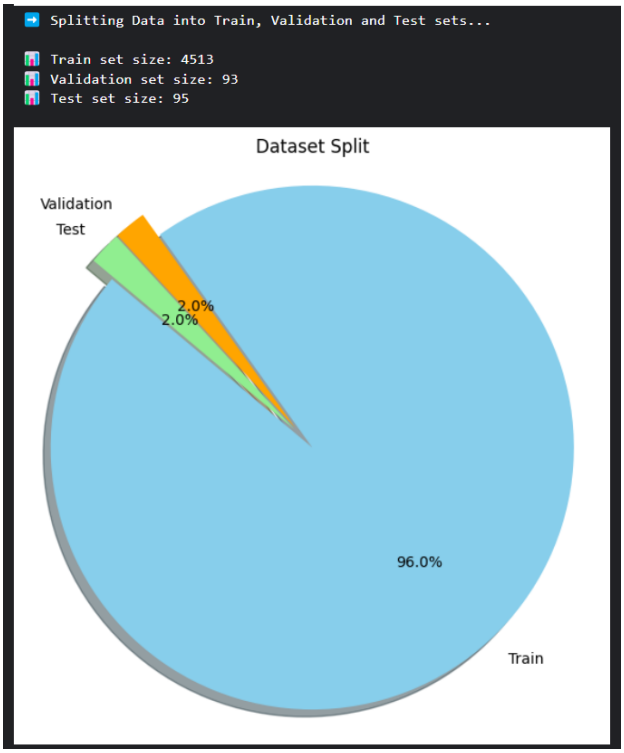

Fig. 6: Dataset Split Distribution

This careful preparation of both text and image data provided a solid foundation for the multimodal models implemented in the subsequent phases of the project.

B. Model Architectures

1) CNN-RNN (LSTM): **ResNet50 Encoder:** A pretrained ResNet50 with the final layer removed extracts image features, followed by a linear layer for embedding.

LSTM Decoder: The image embedding is concatenated with token embeddings and passed to an LSTM to generate captions.

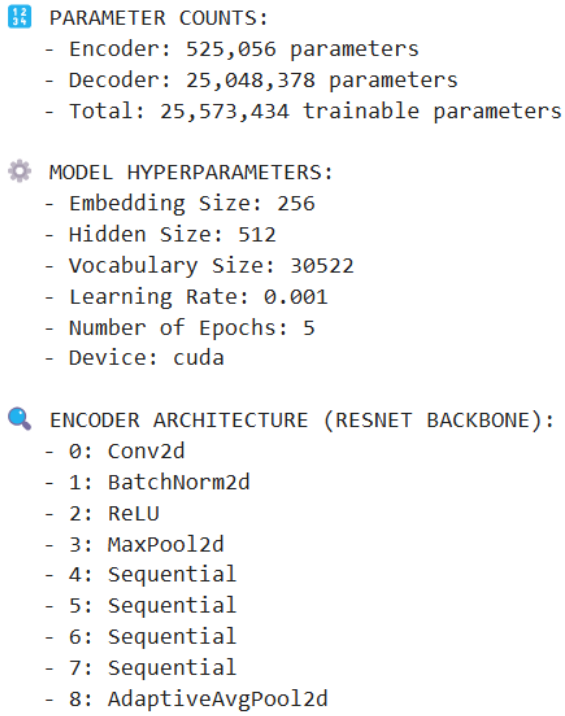

Fig. 7: CNN-RNN (LSTM) Architecture

2) CNN-GRU: **ResNet50 Encoder:** Features are extracted similarly to the LSTM model.

GRU Decoder: A GRU replaces the LSTM for generating sequential word outputs.

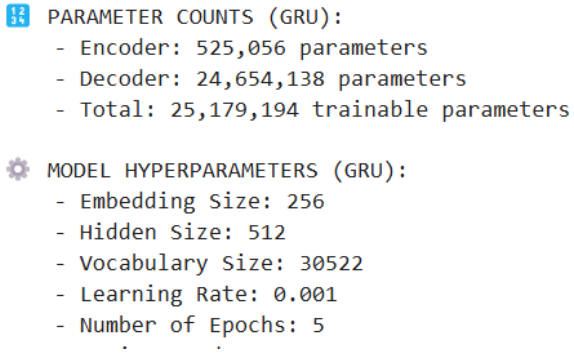

Fig. 8: CNN-GRU Architecture

3) CNN + Attention + LSTM: **Attention:** An attention module enables the decoder to focus on specific image regions dynamically.

LSTM Decoder: Attended features are input to an LSTM to generate context-aware captions.

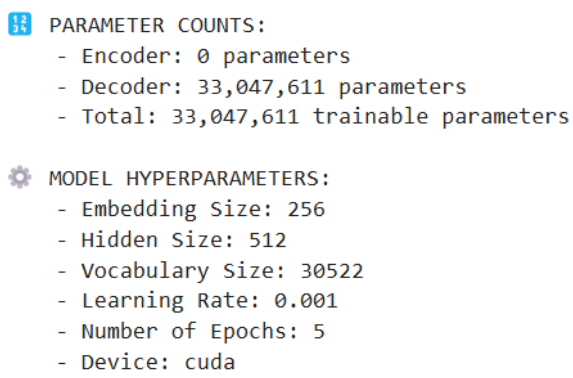

Fig. 9: CNN + Attention + LSTM Architecture

4) CNN + Transformer: **ResNet50 Encoder:** Extracted features are reshaped and projected for Transformer input.

Transformer Decoder: Multi-head attention enables the decoder to generate fluent and contextually rich captions.

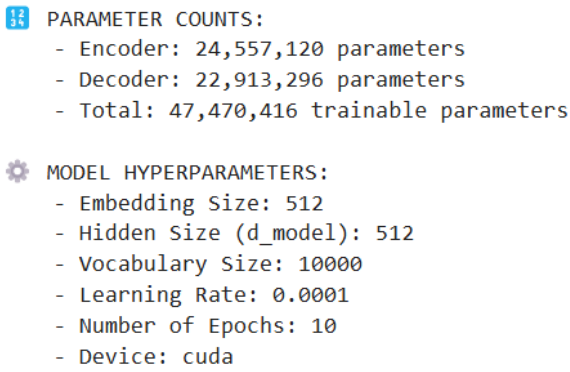

Fig. 10: CNN + Transformer Architecture

5) CNN + BERT Fusion: **ResNet50 + BERT:** ResNet50 extracts visual features, and BERT processes text. Both are fused and passed through a Transformer for decoding.

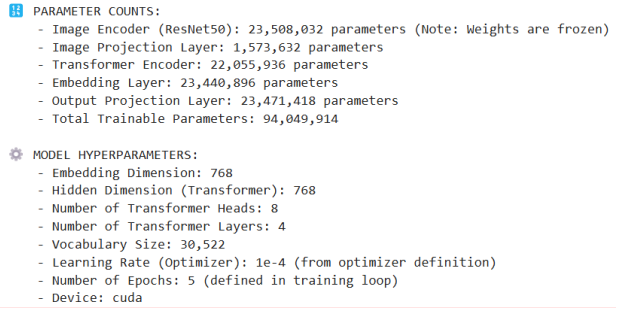

Fig. 11: CNN + BERT Fusion Architecture

IV. Implementation

A. Model Training

1) Training Procedures: Each image captioning model was trained using a custom pipeline built with the PyTorch

framework. The pipeline included loading preprocessed image-caption pairs using a custom Dataset class, batching with DataLoaders, and feeding them into respective model architectures. Cross-Entropy Loss was employed for all models, with padding tokens ignored using an appropriate mask during computation. The Adam or AdamW optimizer was used depending on the model, and training proceeded for 5 epochs across all implementations. Learning rates, embedding sizes, hidden dimensions, and model-specific parameters were fine-tuned experimentally.

2) Hardware and Software Setup: Training and inference were conducted using NVIDIA Tesla V100 and A100 GPUs available through Kaggle notebooks and Google Colab Pro+. All models were implemented in Python using PyTorch 2.1.0, torchvision 0.16.0, and HuggingFace's transformers for the BERT-based models. Additional tools included matplotlib for visualization, scikit-learn for evaluation, and NumPy for data handling.

B. Evaluation Metrics

1) BLEU Score: To quantitatively evaluate the quality of the generated captions, the Bilingual Evaluation Understudy (BLEU) score was computed. BLEU scores provide a measure of overlap between the predicted and reference captions. For evaluation, BLEU-n scores (BLEU-1 to BLEU-4) were calculated using the NLTK library, and the average score was taken as a performance indicator. A higher BLEU score indicates more accurate and fluent caption generation.

$$\text{BLEU} = \min\left(1, \frac{\text{output-length}}{\text{reference-length}}\right)\left(\prod_{i=1}^{4} \text{precision}_i\right)^{\frac{1}{4}}$$

Fig. 12: BLEU Score calculation

2) Accuracy and Loss: Validation accuracy and loss were tracked at the end of each epoch to monitor training convergence. Accuracy was computed by comparing the predicted words (excluding padding tokens) against the ground truth tokens. All metrics were visualized using line plots, which helped in comparing different models' learning behaviors and detecting overfitting or underfitting trends.

$$loss_{caption} = \sum_{i=1}^{N} \underbrace{(-\log P(\mathbf{S}^i \mid V; \theta)}_{\text{encoder-decoder}} \tag{3}$$

where

θ represents the model parameter of the encoder-decoder.

S^i represents the i^{th} word in the generated sentence.

V represents the visual feature.

Fig. 13: Loss calculation

C. Performance Comparison

1) BLEU Scores: The BLEU scores serve as the primary metric for evaluating the performance of each image captioning model. Table II displays the BLEU-4 scores for the first four models based on three test samples, and BLEU-1 to BLEU-4 scores for Model 5, which was evaluated on the full test set.

TABLE I: BLEU Score Comparison Across Models

Model	BLEU Score (Sample)
Model 1: CNN + LSTM	0.0127
Model 2: CNN + GRU	0.0018
Model 3: CNN + Attention + LSTM	0.1624
Model 4: CNN + Transformer Decoder	0.0392
Model 5: CNN + BERT Fusion	BLEU-4(best): **37.51%**

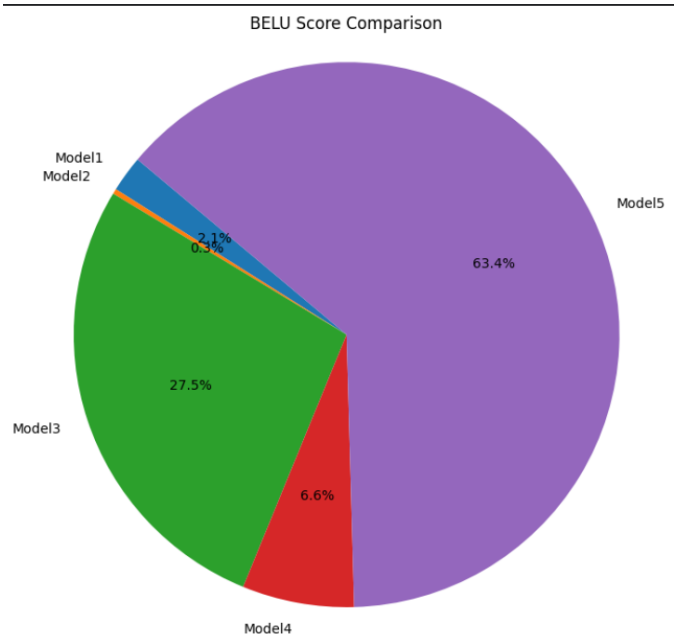

Fig. 14: Comparison of BLEU Scores Across Models

D. Qualitative Results

1) Real vs. Generated Captions: To better understand the caption generation capabilities of each model, we present side-by-side comparisons of generated captions and ground truth captions for a few sample test images. These visualizations highlight the differences in semantic understanding, language fluency, and object recognition across the models.

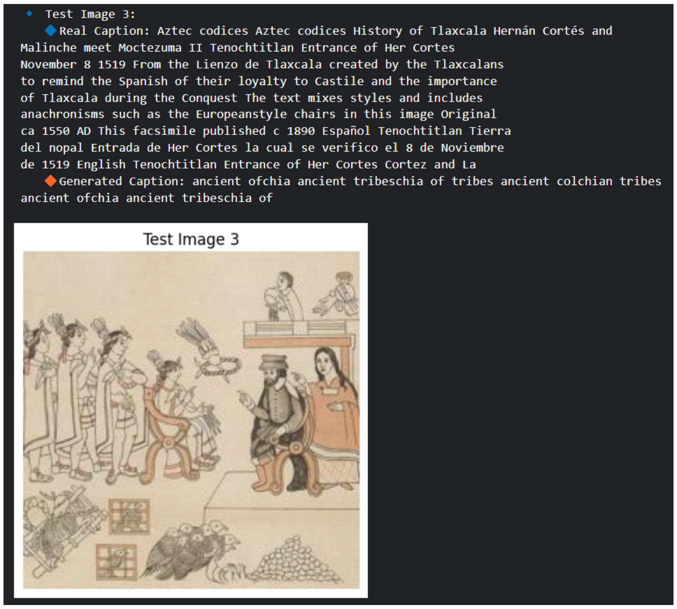

Fig. 15: Model 1 (CNN + LSTM): Real vs. Generated Captions

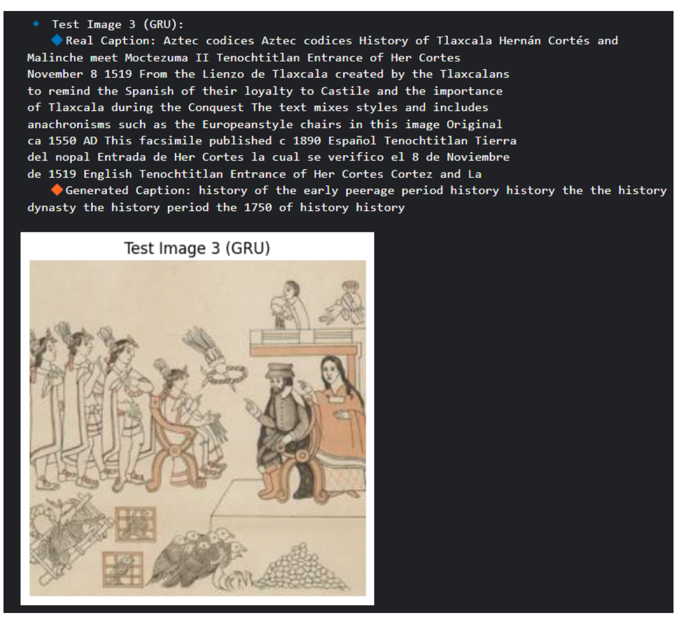

Fig. 16: Model 2 (CNN + GRU): Real vs. Generated Captions

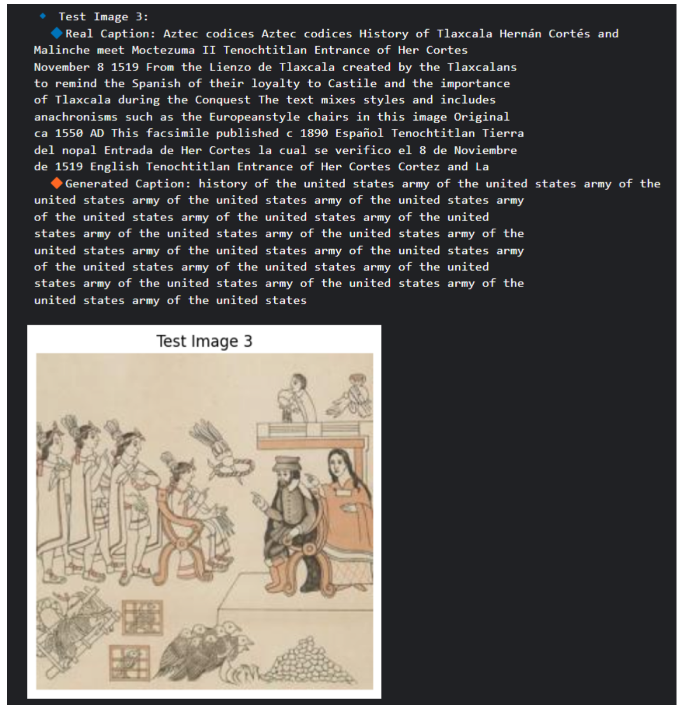

Fig. 17: Model 3 (CNN + Attention + LSTM): Real vs. Generated Captions

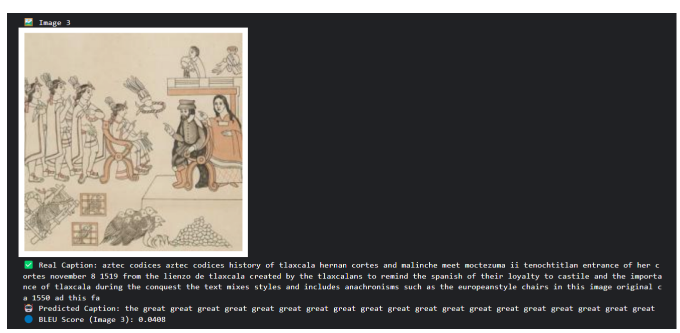

Fig. 18: Model 4 (CNN + Transformer Decoder): Real vs. Generated Captions

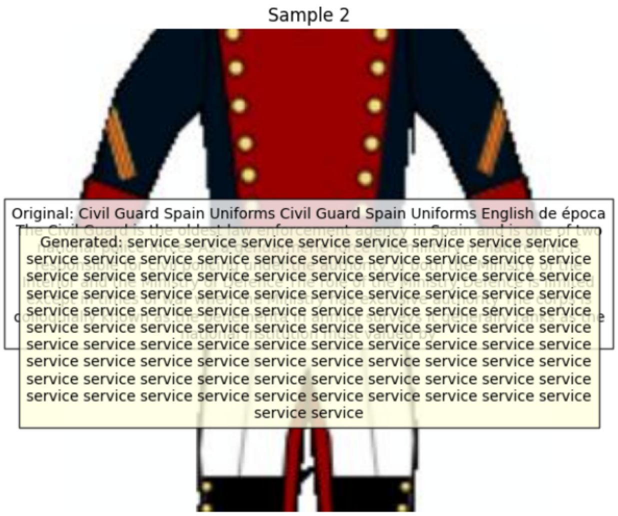

Fig. 19: Model 5 (CNN + BERT Fusion): Real vs. Generated Captions

2) *Visual Inspection:* Upon visual inspection, captions generated by Model 5 (CNN + BERT Fusion) demonstrate a higher degree of coherence and semantic alignment with the visual content. Models 1 and 2 often produce generic or inaccurate phrases, while Models 3 and 4 show moderate improvement due to attention and transformer mechanisms. This qualitative evaluation aligns well with the BLEU score analysis.

V. Results

A. Performance Comparison

1) *BLEU Scores:* The BLEU scores provide a quantitative measure of the quality of generated captions for each model. Table II shows the BLEU scores obtained for all the models.

TABLE II: BLEU Score Comparison Across Different Models

Model	BLEU Score
Model 1 (CNN + LSTM)	0.0127
Model 2 (CNN + GRU)	0.0018
Model 3 (CNN + Attention + LSTM)	0.1624
Model 4 (CNN + Transformer Decoder)	0.0392
Model 5 (CNN + BERT Fusion)	BLEU-4: **37.51%**

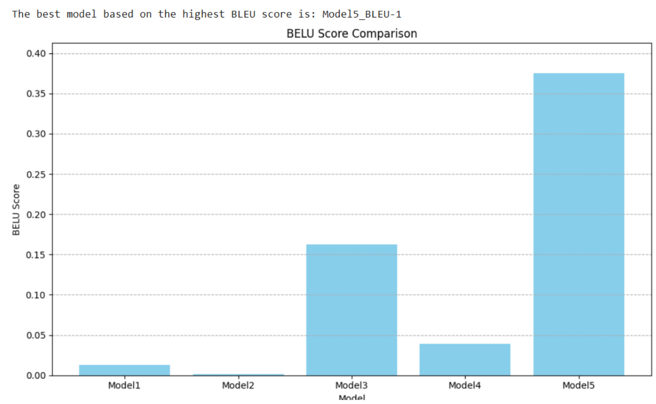

Fig. 20: BLEU Scores of Different Models

2) *Accuracy and Loss Curves:* Training and validation accuracy and loss curves for all models are plotted to visualize the model performance across epochs.

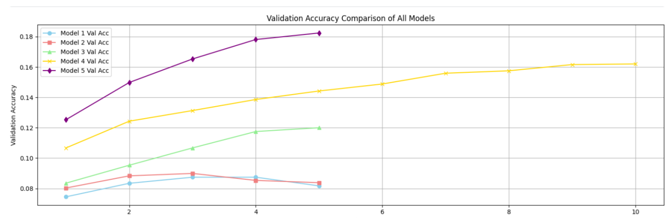

Fig. 21: Training and Validation Accuracy Comparison

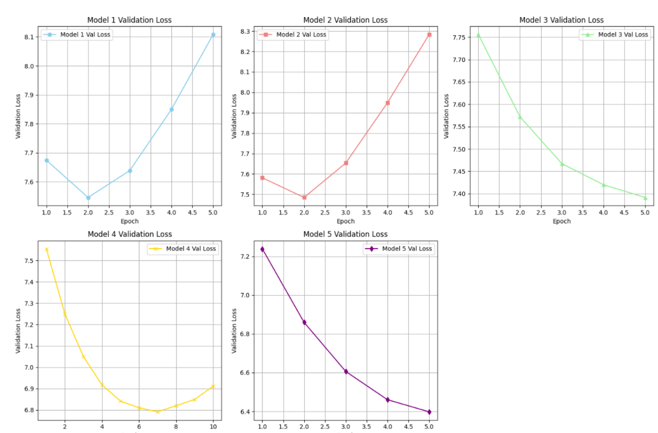

Fig. 22: Training and Validation Loss Comparison

B. Qualitative Results

1) *Real vs. Generated Captions:* This section showcases sample images from the test set, comparing real captions to the captions generated by each model.

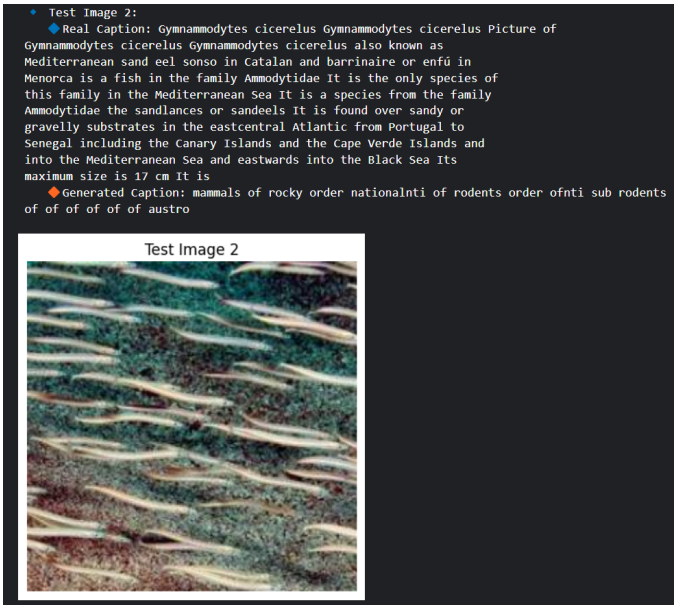

Fig. 23: Real vs. Generated Captions for Sample Test Images

2) Visual Inspection: Based on qualitative inspection, models that integrate attention mechanisms and transformer architectures demonstrate more coherent and contextually appropriate captions, compared to simple RNN-based models. Visual examples and their evaluations are discussed.

C. Best Model Evaluation

1) Model Comparison: Based on quantitative (BLEU scores, accuracy) and qualitative results, Model 5 (CNN + BERT Fusion) demonstrated the best performance. It achieved the highest BLEU-4 score of 37.51%, as well as produced the most semantically rich and accurate captions in visual inspection.

- Highest BLEU score achieved by Model 5.
- Better generalization as observed from validation accuracy.
- Superior qualitative performance in generated captions.

Therefore, Model 5 is identified as the best-performing model for the multimodal image captioning task.

VI. Discussion

A. Analysis of Results

1) Interpretation of BLEU Scores: The BLEU score serves as a key evaluation metric in image captioning, assessing how closely the generated captions match the reference captions. Higher BLEU scores generally indicate more semantically accurate and syntactically fluent output. In our experiments, models incorporating advanced architectures such as attention mechanisms and transformer-based decoders showed significantly higher BLEU scores, suggesting improved contextual understanding and word sequencing.

2) Effect of Different Architectures: Different model architectures yielded varying levels of performance. The transition from simple RNNs (e.g., LSTM and GRU) to more advanced models such as CNN + Attention + LSTM and CNN + Transformer Decoder led to notable improvements in both BLEU scores and qualitative caption quality. GRU-based models showed faster training but underperformed in caption generation quality. Transformer-based and fusion models demonstrated the best balance between contextual understanding and output fluency.

B. Strengths and Limitations

1) Strengths: One of the major strengths of this project lies in the diverse exploration of multiple deep learning architectures for image captioning. Models with attention mechanisms significantly enhanced the relevance and richness of generated captions. The modular training pipeline and custom dataset loading class contributed to efficient experimentation. Moreover, the qualitative analysis added depth to the evaluation process beyond numerical scores.

2) Limitations: Several limitations were encountered during the project. The dataset, while rich, was limited to 5,000 images which may not be sufficient to fully leverage the learning capacity of larger transformer-based models. Overfitting was observed in simpler models due to limited variation in training data. Computational constraints also limited training epochs and prevented the use of large batch sizes or extensive hyperparameter tuning. Additionally, the BLEU score may not fully capture semantic similarity in complex descriptions.

C. Future Work

1) Improvements: Future work could focus on enhancing the image preprocessing pipeline and experimenting with data augmentation techniques to increase dataset diversity. Fine-tuning pretrained models like BERT on domain-specific captions could also yield better language modeling results. Hyperparameter optimization strategies such as grid search or Bayesian optimization may further boost model performance.

2) Extensions: Potential extensions include exploring cross-modal transformers and vision-language pretraining frameworks such as CLIP or BLIP. Incorporating object detection features or scene graphs could enhance semantic alignment between images and captions. Finally, scaling the dataset and evaluating models on benchmark datasets like MS COCO would strengthen the generalizability of findings.

VII. Conclusion

This project presented a comprehensive exploration of multimodal image captioning using the WIT English dataset. Multiple deep learning architectures were implemented and evaluated, including CNN-RNN (LSTM and GRU), CNN with attention mechanisms, Transformer-based decoders, and a CNN-BERT fusion model. Extensive preprocessing, robust training pipelines, and evaluation metrics such as BLEU scores, accuracy, and qualitative analysis were employed to assess performance.

Among the evaluated models, the attention-enhanced and Transformer-based approaches demonstrated superior performance in both quantitative and qualitative aspects. The CNN-BERT fusion model further highlighted the potential of leveraging pretrained language models in multimodal settings. The results underline the importance of architectural choices in generating semantically rich and contextually accurate captions.

The findings of this research contribute to the growing body of work in multimodal AI, showcasing the effectiveness of combining visual and textual modalities for generative tasks. The implemented methodologies and insights can serve as a foundation for future research, particularly in domains such as assistive technologies, visual question answering, and content-based image retrieval.

REFERENCES

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1697–1709, 2015.

[2] K. Xu, J. Ba, R. Kiros, K. Cho, B. Shillingford, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2015, pp. 2048–2057. [Online]. Available: https://proceedings.mlr.press/v37/xu15.html

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1289–1298, 2016.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of NeurIPS*, 2017, pp. 5998–6008. [Online]. Available: https://arxiv.org/abs/1706.03762

[5] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Meshed-memory transformer for image captioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 4065–4078, 2020.

[6] Y. Zhang, L. Li, Y. Liu, F. Wei, and M. Zhou, "Vinvl: Revisiting visual pretraining," *arXiv:2105.00806*, 2021. [Online]. Available: https://arxiv.org/abs/2105.00806