Tilburg University

IMPROVING IMAGE CAPTIONING WITH TRANSFORMER EMBEDDINGS

A NOVEL HYBRID APPROACH COMBINING CNN-LSTM WITH BERT OR GPT FOR ENHANCED ACCURACY AND CONTEXTUAL UNDERSTANDING

KARIM DRAFAT

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

382198

COMMITTEE

dr. Elisabeth Huis in 't Veld, MBA
prof. dr. Travis Wiltshire

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

August 26th, 2024

WORD COUNT

8613

ACKNOWLEDGMENTS

IMPROVING IMAGE CAPTIONING WITH TRANSFORMER EMBEDDINGS

A NOVEL HYBRID APPROACH COMBINING CNN-LSTM WITH BERT OR GPT FOR ENHANCED ACCURACY AND CONTEXTUAL UNDERSTANDING

KARIM DRAFAT

Abstract

Image captioning (IC) refers to the process of generating descriptive textual representations of visual content. Transformer-based models represent the state-of-the-art in various NLP tasks, including IC (Ondeng et al., 2023). They have revolutionized IC by leveraging self-attention mechanisms to efficiently capture contextual information in parallel, surpassing traditional CNN-LSTM systems that process information sequentially. However, despite recent advancements in deep learning and NLP, IC encounters challenges in handling diverse visual and linguistic contexts effectively. These challenges include accurately interpreting complex scenes, understanding nuanced linguistic expressions, and generating contextually coherent captions across a wide range of images. This research introduced a novel architecture for IC. The goal is to combine the traditional encoder-decoder (CNN+LSTM) models with transformer-based models such as BERT or GPT. Instead of using only a pre-trained CNN for image feature extraction, the structure incorporates BERT as a second encoder for captions. The decoder part remained LSTM, resulting in a multimodal/hybrid architecture. Additionally, the study intended to evaluate and compare the use of GPT in place of BERT within the same architecture, and experiment with fine-tuning either the entire architecture or just the LSTM component. As a baseline reference to compare with our research approach, the study utilized a soft attention model integrated between the CNN and LSTM layers similar to Xu et al. (2015) approach. Evaluation on the Flickr30k dataset showed promising results in improving caption accuracy and semantic richness, highlighting advancements in multimodal image understanding and description. The models, particularly those utilizing GPT embeddings with fine-tuning, demonstrated competitive performance compared to recent architectures (Alqahtani et al., 2024; Patel & Varier, 2020; Verma et al., 2024), and outperformed the baseline model. Moreover, we evaluated the best and least scored predicted captions using

BLEU, METEOR, and CIDEr scores. The analysis revealed that common errors included repetition, missing information, incorrect details, and grammar issues. A detailed evaluation of the lowest scoring captions revealed that errors in missing information and incorrect details were more prevalent, while repetition and grammar issues were less frequent but still notable. These findings highlight areas for improvement in IC models.

1 INTRODUCTION

This study explored and compared two novel models that combine pretrained convolutional neural networks (CNNs) trained on large-scale datasets with embeddings from BERT or GPT, integrated with LSTM networks. By using fine-tuning and transfer learning methods, the goal of this study is to determine whether the integration of transformer-based models embeddings (BERT or GPT) with traditional CNN-LSTM architectures for image captioning (IC) affects the quality and performance of the generated captions.

1.1 *Problem statement*

Image captioning (IC) combines computer vision and natural language processing (NLP) to enable machines to understand and communicate visual information in a human-like manner. IC is crucial in various applications, such as assisting visually impaired individuals, enhancing multimedia search engines, and enriching content creation for social media platforms

IC systems face several challenges despite advancements in deep learning and NLP fields. One of the primary challenges is the need to train models on sufficiently diverse and large-scale datasets to generalize effectively across different visual and linguistic contexts.Traditionally, datasets like MSCOCO have been pivotal but often lack sufficient variation in visual content and linguistic complexity. Current state-of-the-art latent variable models only capture a limited variety based on the paired annotations that are supplied for each image (Aneja et al., 2019). The diversity for a particular image is therefore restricted to the annotated data (Mahajan & Roth, 2020).

Furthermore, IC systems must address the complex many-to-many relationship between images and captions (Aneja et al., 2019). This relationship implies that a single image may correspond to multiple plausible captions, and conversely, a single caption may describe various images depending on context and interpretation. For instance, an image depicting a crowded city street could inspire captions focusing on different aspects such as traffic congestion, urban diversity, or architectural styles. Effective models should therefore integrate mechanisms to align visual and textual modalities across varying levels of detail, ensuring accurate and informative descriptions.

Both traditional CNN-LSTM models and more recent transformer-based approaches face distinct limitations in the context of IC. Traditional CNN-LSTM models, while effective in capturing spatial features from images and sequential dependencies in text, are constrained by their sequential nature which hinders their ability in capture spatio-temporal features (Fang et al., 2021). On the other hand, transformers are computationally intensive, requiring substantial resources and large-scale datasets for effective training. This dependency on large-scale data can be a significant limitation in scenarios where data availability and diversity are limited (Bai et al., 2021). The computational and memory complexity of transformers is quadratic with respect to the sequence length, which is the trade-off for its capacity. As a result, training large transformers is highly costly and time-consuming (Fournier et al., 2023). Additionally, transformers are prone to overfitting when trained on smaller datasets, which can reduce generalization and performance on unseen data (Bai et al., 2021).

Addressing these limitations requires innovative approaches that combine the strengths of existing models while mitigating their weaknesses. Pre-trained models, developed on large datasets, provide a promising solution by capturing diverse linguistic and visual features. Hybrid models integrating CNNs or pre-trained feature extractors with transformer-based language models like BERT or GPT can improve the contextual understanding of visual content and enhance caption generation across diverse scenarios.

1.2 *Research questions*

Therefore, to address the primary purpose of this study, we posed the following research question:

To what extent do BERT and GPT enhance the contextual understanding and semantic coherence of image captions within a hybrid CNN-Transformer-LSTM architecture?

> *We expanded on the main research question with the following sub-research questions:*

The sub-questions can be listed separately as follows::

RQ1 *What is the accuracy and performance of the proposed hybrid models?*

Fine-tuning and freezing weights of pre-trained models have become common practice in determining the performance of the hybrid architecture (Howard & Ruder, 2018). Fine-tuning allows the models like BERT and GPT to adapt to the specific task of caption generation, potentially improving contextual understanding and coherence (Zhou & Srikumar, 2021). Conversely, freezing weights preserves the pre-trained models' integrity but may restrict their ability to adapt to the unique aspects of IC (Zhuang et al., 2020). Investigating these approaches will clarify how adaptation strategies influence the overall quality and efficacy of the hybrid architecture.

RQ2 *What are the differences in captioning accuracy and error type when using BERT or GPT versus baseline model*

Comparing the integration of BERT or GPT embedding with the baseline model in the context of captioning accuracy and error types offers insights into their respective strengths and weaknesses. BERT, with its bidirectional context understanding, may excel at capturing intricate semantic details within captions (Devlin et al., 2018). In contrast, GPT's autoregressive nature might be better suited for the coherence and flow of longer captions (Black et al., 2022). Understanding these differences can help determine which model is more appropriate for different scenarios of image content and caption complexity.

RQ3 *What are the most frequent captioning errors, and what are the underlying reasons for these recurring error patterns?*

Identifying and analyzing frequent captioning errors highlights the challenges faced current models and architectures. These errors could range from semantic inaccuracies to syntactical inconsistencies.

1.3 *Social and scientific impact*

This work has both a societal and a scientific impact. Primarily, the aim is to develop a deep learning model that provides significant social benefits across various domains. This hybrid approach enhances accessibility for individuals with visual impairments by generating detailed and nuanced captions, thereby empowering them to better understand and engage with visual information. Additionally, it improves image organization and search capabilities, facilitating more accurate categorization and retrieval of visual content. Moreover, the model could aid in surveillance systems by automatically generating descriptive captions for security footage, enhancing threat detection and response capabilities.

The scientific impact of this thesis is multifaceted. Firstly, it contributes to the ongoing development of more sophisticated IC models by demonstrating the benefits of combining powerful pre-trained models with LSTM networks. Secondly, the comparative analysis of fine-tuning versus transfer learning offers valuable insights into optimizing training strategies for different resource constraints. Finally, the findings could lead to more accessible and cost-effective methods for training high-performing IC models, potentially broadening their application in various fields such as assistive technologies, multimedia search engines, and content generation.

2 RELATED WORK

2.1 *Encoder decoder models*

IC in deep learning has seen significant progress with the development of various architectures, particularly through the evolution of encoder-decoder models, transformer-based models, and the integration of word embedding techniques. One prominent architecture is the encoder-decoder, as described by Vinyals et al. (2015) and Xu et al. (2015). The encoder, typically based on CNN and pooling layers, extracts visual features from input images and encodes them into a fixed-length vector. The decoder, often based on LSTM networks, generates captions sequentially by utilizing the encoded representation, with LSTM's memory cells enabling the capture of long-range dependencies.

In addition to these foundational models, attention mechanisms have significantly enhanced captioning performance. The "Show, Attend and Tell" model by Xu et al. (2015) introduced attention mechanisms to address the limitations of fixed-length vector representations in traditional models. This approach employs both soft and hard attention mechanisms. Soft attention computes a weighted sum of all visual features, allowing the model to focus on different parts of the image by assigning varying weights to each feature. This method is computationally efficient and differentiable, enabling end-to-end training. Conversely, hard attention involves selecting a specific part of the image to focus on, which introduces discrete choices and requires techniques like reinforcement learning for training, making it less straightforward to optimize (Shen et al., 2018).

Recent encoder-decoder approaches, such as the one proposed by Saeidimesineh et al. (2023), introduce a parallel encoder-decoder framework for IC that utilizes diverse module types within both the encoder and decoder to capture complex semantic relationships. Instead of relying on a single type of module, the proposed method employs augmented parallel blocks, such as graph convolutional networks and multi-head attention blocks in the encoder, and LSTM with an Attention over Attention (AoA) structure in the decoder. This parallelization allows for the simultaneous processing of different feature representations, which are then integrated through concatenation to model higher-level semantic concepts effectively.

Other researchers such as Zhu et al. (2023) have explored an automatic-questioning method in IC. The paper introduces ChatCaptioner, an innovative method that applies automatic questioning to IC. Unlike traditional models that focus solely on generating captions, ChatCaptioner uses ChatGPT to pose a series of insightful questions about an image to BLIP-2, a vision question-answering model.

2.2 *Transformer-Based models*

Transformer-based models have emerged as a new architecture and are considered the state-of-the-art for IC due to their ability to capture long-range dependencies and semantic relationships within images and text (Y. Li et al., 2022; Vasireddy et al., 2023). Unlike traditional CNN-LSTM architectures, which rely on CNNs to extract visual features and LSTM networks to generate captions sequentially, transformer models utilize self-attention mechanisms to weigh the importance of different parts of the input. This allows for parallel processing and more effective capture of complex contextual information (Vasireddy et al., 2023). This parallel processing capability enables transformers to outperform CNN-LSTM models by effectively modeling global context and capturing intricate image-text interactions more effectively.

Transformers in IC architecture are often used as decoders, and have achieved competitive scores (Alqahtani et al., 2024). Recent studies have also explored utilizing transformers not only as decoders but also as encoders, wherein extracting image features using Vision transformer and generating contextually descriptions through models like GPT (Patel & Varier, 2020; Vasireddy et al., 2023). However, CNN-LSTM architectures have their strengths, particularly in extracting spatial features from images and generating sequential outputs, which can capture temporal dependencies (Patel & Varier, 2020).

Both encoder-decoder models and transformer-based models in IC face challenges in capturing the complex many-to-many relationship between images and captions. Encoder-decoder models often produce captions that lack diversity and specificity, failing to fully reflect the wide range of possible descriptions an image can evoke (Wang & Chan, 2019). Similarly, transformer-based models, despite their advanced self-attention mechanisms, can sometimes miss nuanced interpretations, leading to captions that do not adequately represent the multiple plausible descriptions that different contexts might elicit (Wang & Chan, 2019).

2.3 *Transformers embeddings*

Word embedding, a technique in NLP, maps words or phrases to real-number vectors, facilitating the integration of textual information with visual features in IC tasks. However, researchers raised concerns about using pre-trained embeddings like GloVe or Word2Vec due to their limitation of assigning a single vector to each word, regardless of contextual understanding (Devlin et al., 2018).

As introduced in the original BERT paper by Devlin et al. (2018), the embeddings are derived from the hidden state of the final transformer encoder stack for the special [CLS] token in BERT and similar models. This token attends to all other tokens in the input, thereby capturing rich contextual information from the entire input. However, there has been limited research specifically on transformer embedding, yet related studies on contextual embedding have demonstrated their utility in tasks such as document classification and ranking (MacAvaney et al., 2020; Reimers & Gurevych, 2019).

BERT and GPT, employing transformer architectures, offer bidirectional contextualized word embedding. BERT with transformer encoders utilizes masked language modeling for unsupervised training, considering both left and right context, whereas GPT models, such as GPT-3, leverage large-scale pre-training on diverse text to generate context-aware representations, providing effective alternatives for tasks like IC (Ghojogh & Ghodsi, 2020).

2.4 *Proposed approach*

The proposed approach aimed to address the previously mentioned limitations by integrating the strengths of both encoder-decoder and transformer-based architectures while mitigating their weaknesses. By employing pre-trained CNNs, which have been trained on extensive datasets, augmented with embeddings from BERT or GPT,this hybrid approach bridges the gap in the literature by incorporating both spatial and semantic information.

3 METHODOLOGY

In accordance with the identified research gaps in the literature, we employed model comparison, error analysis, and out-of-sample generalization techniques to ensure the effectiveness of the data science approach. The primary dataset used for this study is Flickr30k.

As shown in Figure 1, two models fine-tune both the CNN and embedding, while the other two use transfer learning by freezing these components and only training the LSTM. The study also employed cross-validation and evaluation using BLEU, METEOR, and CIDEr metrics to compare model performance. The analysis explored the effectiveness of fine-tuning versus transfer learning, and the benefits of BERT versus GPT embedding.

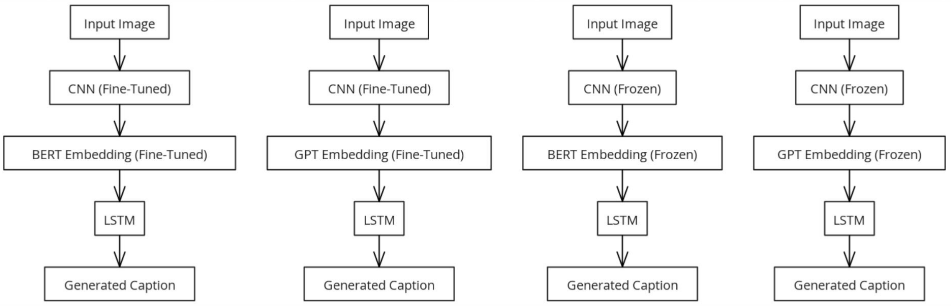


Figure 1: IC proposed models

We implemented a baseline model to compare with our approach. The model integrates a soft attention mechanism between the CNN and LSTM layers similar to Xu et al. (2015) method. After extracting feature maps from the CNN, the attention mechanism computes a weighted sum of these feature maps instead of directly passing the entire feature vector to the LSTM. This is done by calculating attention weights, which compare the current hidden state of the LSTM with feature vectors from different parts of the image. The resulting context vector, highlighting the most relevant parts of the image, is then combined with the LSTM's hidden state to predict the next word in the caption. This soft attention approach was chosen over the hard attention mechanism discussed in Xu et al. (2015) due to its smoother, differentiable nature, making it easier to train. To ensure a fair comparison and accommodate our limited resources, we used the same hyperparameters for all models.

3.1 *Data description*

For our study, we utilized the Flickr30k dataset, a well-established resource in the field of IC. The original Flickr30k dataset comprises 31,783 images collected from the Flickr photo-sharing website. Each image in this dataset is accompanied by five descriptive captions, provided by human annotators. These captions are crafted to describe the key elements and objects within the images, offering a rich open source of data for training and evaluating IC models.

The images in the Flickr30k dataset cover a wide range of scenes and activities, from everyday situations to more complex, multi-object scenarios. This diversity ensures that the models trained on this dataset are robust and capable of handling various contexts and visual elements.

The choice of the Flickr30k dataset is motivated by its widespread use in IC benchmarks, providing a reliable and comparable basis for evaluating our novel models.

3.2 *Prepossessing*

Figure 2 summarizes the main prepossessing steps we followed for our approach.

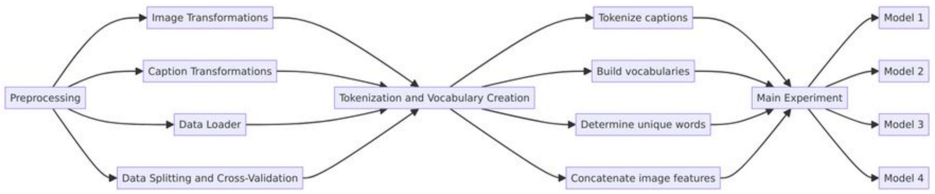


Figure 2: Preprocessing steps

3.2.1 ***Image transformations***

In our IC model, images undergo a series of transformations to ensure they are in the optimal format for neural network processing. First, each image is resized to 224x224 pixels, a standard input size for many deep learning models (Nafi'iyah & Setyati, 2021; Rochmawanti & Utaminingrum, 2021).

The images are then converted to PyTorch tensors, enabling efficient manipulation and processing. Following this, a normalization step is applied using pre-defined mean and standard deviation values: [0.485, 0.456, 0.406] for the mean and [0.229, 0.224, 0.225] for the standard deviation. These specific values are derived from the ImageNet dataset, which contains a vast collection of natural images and serves as a benchmark for many pre-trained models, including ResNet50 used in our model (C. Li et al., 2021; Nakamura & Harada, 2019). The rationale for using these values includes standardizing the pixel values to have a mean of zero and a standard deviation of one, which accelerates training and improves model performance.

3.2.2 *Caption transformations*

To enhance the usability of captions in our IC model, we standardized the original caption data, maintaining the structured format but ensuring consistency and compatibility with our system. The captions now follow a streamlined format, preserving the image name, comment number, and comment text. This transformation facilitates efficient loading, processing, and tokenization, contributing to more accurate and reliable caption generation.

3.2.3 *Data loader*

We created a data loader that prepares images for the model by reading image files, applying transformations, and batching images with their captions. The data loader enables parallel processing, which significantly speeds up the training process, ensures that the GPU is utilized effectively, and improves the model's generalization by shuffling and batching the data.

3.2.4 *Data splitting and cross-validation*

The Flickr30k dataset was split into training, validation, and test sets, with 70% allocated for training and 30% for validation/testing. The training set was used to fit the model parameters, the validation set for tuning hyperparameters, and the test set for final model evaluation. Captions were tokenized and processed to build a vocabulary for effective learning. To enhance model evaluation, 5-Fold Cross-Validation was used during hyperparameter tuning by dividing the dataset into five equal parts and using each fold for validation once. The choice of 5-fold cross-validation was influenced by limited computational resources. More folds could potentially lead to better performance by providing a more refined evaluation.

However, it also increases the total training time linearly with the number of folds. Therefore, a balance was struck between model evaluation granularity and computational feasibility.

3.3 *Tokenization and vocabulary creation*

Tokenization and vocabulary creation are essential for preparing textual data in training IC models. For our approach, we used separate tokenization processes tailored to models using BERT or GPT embeddings. The BERT-based model utilized the BERT tokenizer to split captions into tokens, building a vocabulary from these tokens to encode captions into numerical representations. This vocabulary is crucial for encoding the captions into numerical representations during model training and inference, enabling the model to learn meaningful patterns and relationships within the text (Hu et al., 2021). Similarly, the GPT-based model employed the GPT tokenizer, creating a vocabulary that allows for effective caption encoding. This approach enhances the model's ability to learn and generate diverse, coherent captions by capturing a wide range of words and concepts.

To train the model on consistent set of words and to use memory more efficiently , we had to determine the number of different words used in our dataset's captions. We collected all the captions for each image and applied tokenization, where we broke down these captions into individual, standardized, and lowercase words. We found that our dataset contains approximately 30,522 unique words.

3.4 *Main experiment*

Inspired by the work of Vinyals et al. (2015) , Xu et al. (2015), and Satti et al. (2023), we utilized four distinct models for IC, employing transfer learning techniques. The first model combined a CNN with BERT embeddings and an LSTM, where both the CNN and BERT embeddings were fixed, and only the LSTM was trained. The second model also used a CNN with BERT embedding and an LSTM, but all components were fine-tuned. Similarly, the third model combined a CNN with GPT embeddings and an LSTM, with only the LSTM being trained while the CNN and GPT embedding remained fixed. The fourth model fine-tuned all components: the CNN, GPT embedding, and the LSTM. This approach allowed us to explore the effectiveness of fixed versus fine-tuned pre-trained embedding in conjunction with sequential modeling using transfer learning.

3.5 *Proposed method*

We used the CNN pre-trained ResNet-50 model to extract high-level features from images. The final classification layer is excluded, and BERT embeddings are generated using a pre-trained BERT model to process the captions. The LSTM network then processes the concatenated features and embeddings to generate the output sequence. To prevent overfitting, early stopping and dropout regularization were implemented. Lastly, a fully connected layer maps the LSTM outputs to the vocabulary space to generate captions.

The hyperparameters used for this model are as follows: the vocabulary size is set to 30,522, which was determined from the dataset captions. The maximum caption length is 256, and the CNN feature size is 2,048, representing the ResNet-50 model's final convolutional layer before flattening. Both BERT hidden size 768 and LSTM hidden size 256 were selected to balance model complexity and computational resources. Dropout regularization with a probability of 0.5 was added to randomly deactivate neurons during training and enhance adaptability. The model was trained using cross-validation. During each fold of cross-validation, the data are split into training and validation sets. The training dataset and validation dataset are created from these splits, and data loaders are generated with a batch size of 32. We also tried a batch size of 64 with the aim that the gradients computed during backpropagation would be averaged over more samples. However, it significantly increased memory usage.

The model is trained using the Adam optimizer with a learning rate of 0.001, and the criterion used for training is CrossEntropyLoss with an ignore_index of zero to handle padding. During training, the model goes through 16 epochs per fold. In each epoch, the model processes the images and captions, tokenizes the captions using the BERT tokenizer, and generates outputs. These outputs are compared to the target captions to compute the loss, which is then used to update the model weights. The training loss is tracked, and the performance of the model is monitored to select the best model based on the validation loss.

Similar to the first model, we fine-tuned all components while replacing the BERT embedding with GPT. We utilized the pre-trained ResNet-50 model for extracting high-level features from images and used LSTM networks to handle sequential data processing. The same training approach was employed for both models, including cross-validation, the Adam optimizer with a consistent learning rate, and common hyperparameters such as batch size.

3.6 *Transfer learning*

In the final approach, we employed transfer learning by fine-tuning only the LSTM component while keeping the CNN feature extractor (ResNet-50), BERT embeddings, and GPT embeddings frozen. This strategy leverages pre-trained weights of the CNN and embedding layers, optimizing only the LSTM to adapt to specific captioning tasks. By freezing the CNN and embeddings, the aim was to retain their learned features and contextual understanding while focusing computational resources on refining the sequential processing capabilities of the LSTM. This method is designed to balance training efficiency and model performance, allowing us to explore the benefits of pre-trained representations in IC tasks without retraining entire networks from the beginning.

3.7 *Evaluation metrics*

To evaluate the models, we tracked loss during training for optimization and used validation loss to select the best-performing model. To assess the quality of generated captions, we employed BLEU, METEOR, and CIDEr metrics, which offer deeper insights into how well the captions match human language and understanding.

BLEU measures the precision of n-grams (sequences of words), ensuring the generated captions closely match the reference captions in terms of word choice and order. METEOR goes further by considering synonyms, word order, and stemming, focusing on capturing the overall meaning of the captions. CIDEr evaluates the frequency of n-grams in the generated caption relative to their frequency in the reference captions. It evaluates how well the captions align with human-written descriptions, rewarding those that capture the most salient aspects of the image (González-Chávez et al., 2024).

The metrics were chosen because they comprehensively assess the precision, fluency, and semantic accuracy of the generated captions, providing a robust indication of their performance in real-world scenarios.

3.8 *Out-of-Sample generalization*

To assess out-of-sample generalization, we evaluated the models on unseen test data after all training and validation phases. The test data was never exposed during training or hyperparameter tuning, ensuring a true test of the models' ability to generalize to new inputs.

In evaluating the test set, we compared the models based on the evaluation metrics, which provided insights into how well the generated captions align with human descriptions. This methodical approach ensures that our evaluation on unseen test data was rigorous and aligned with best practices in machine learning for IC.

3.9 *Error analysis*

Error analysis for the IC model was conducted by evaluating the best and lowest-scoring predicted captions. The analysis involved several systematic steps to identify recurring patterns of mistakes. First, the model was run to generate captions for a test set, and both the predicted captions and the ground truth captions were collected for comparison. The predicted captions were then evaluated using the metrics scores to determine their accuracy. After obtaining these scores, we summed them for each caption to get a comprehensive performance metric. This cumulative score allowed us to classify the captions effectively. Finally, a manual inspection of selected captions was conducted to categorize the types of errors.

3.10 *Root cause analysis*

We conducted a detailed evaluation to understand the nature and frequency of common mistakes made by each model. Initially, we identified that our dataset comprised 31,783 images, with 15% (4,767 images) allocated for testing. We then focused on the bottom 5% of predictions with the lowest scores, amounting to 238 captions, to analyze the common errors.

The analysis involved manually categorizing errors into four types: repetition, missing information, incorrect details, and grammar issues. The data was then distributed to reflect the performance trends of each model.

4 RESULTS

4.1 *Models performance*

To answer the main research question of this study: To what extent do BERT and GPT enhance the contextual understanding and semantic coherence of image captions within the proposed hybrid architecture?

Initially, we observed that our approach showed promising results during the training phase even though the models were trained for only 16 epochs. As shown in Figure 3, the performance trends indicated that additional epochs could have further improved the models' effectiveness, suggesting that the models were still learning and had not yet reached their full potential.

All models displayed a positive trend with decreasing training and validation losses. The most significant improvements were observed in the models with full fine-tuning, indicating that comprehensive fine-tuning generally led to better performance and generalization.

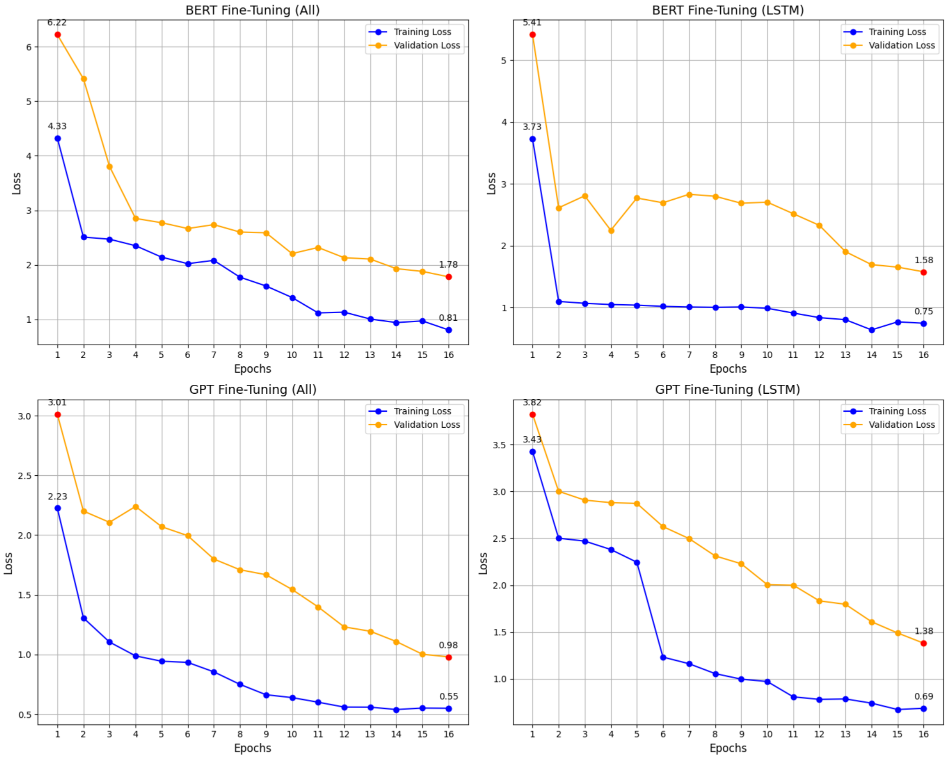


Figure 3: Training and validation loss over epochs

We assessed the models using the three primary metrics to evaluate their accuracy and performance within each epoch and answer the first sub-question. Afterward, we selected the best metric performances of the proposed hybrid models. A summary of the results is listed in Table 1. Our analysis reveals significant insights into how different approaches to fine-tuning and pre-training impact model performance.

Model Name	BLEU	METEOR	CIDEr
ResNet-50 + BERT+ LSTM (fine tuning for ALL)	42.5	20.11	55.14
ResNet-50 + GPT + LSTM (fine tuning for ALL)	67.8	48.54	42.18
ResNet-50 + BERT+ LSTM (fine tune only LSTM)	48.6	47.2	44.71
ResNet-50 + GPT + LSTM (fine tune only LSTM)	62,33	41.12	48.69

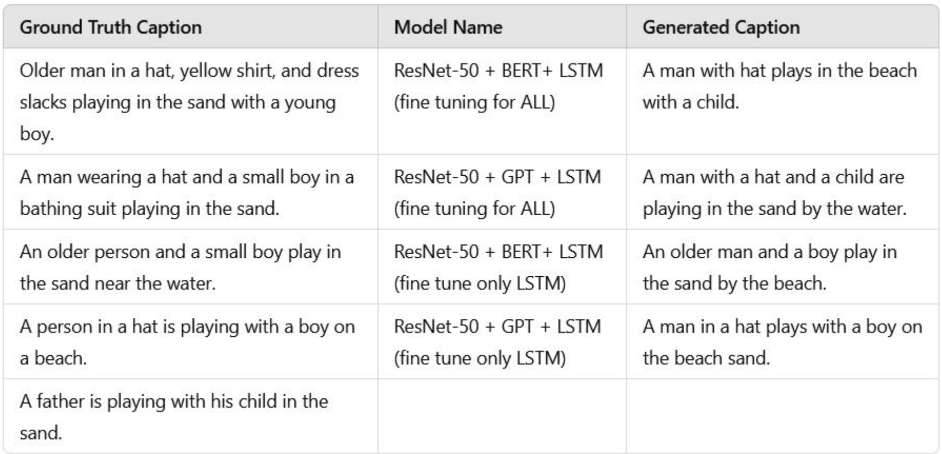
Table 1: Best metrics performances of the proposed hybrid models

The BLEU score, which measures the precision of n-grams in generated captions compared to reference captions, reflects the accuracy of the model in producing text sequences that align closely with human-written captions. The "ResNet-50 + GPT + LSTM (fine-tuning for ALL)" model achieved the highest BLEU score of 67.8. This high score indicates that this model excels at generating captions with n-grams that closely match the reference captions, demonstrating the effectiveness of fine-tuning all components of the model. The significant improvement in BLEU scores when fine-tuning is applied highlights the importance of optimizing all components to enhance precision in language generation.

The METEOR score considers synonyms, stemming, and word order, to provide a more nuanced evaluation of semantic accuracy and fluency. The "ResNet-50 + GPT + LSTM (fine-tuning for ALL)" model also achieved the highest METEOR score of 48.54. This result underscores the models' superior ability to capture semantic meaning and fluency in the generated captions. The substantial gain in METEOR scores, especially with full fine-tuning, emphasizes how fine-tuning all components of the model enhances both the contextual understanding and the naturalness of the generated text.

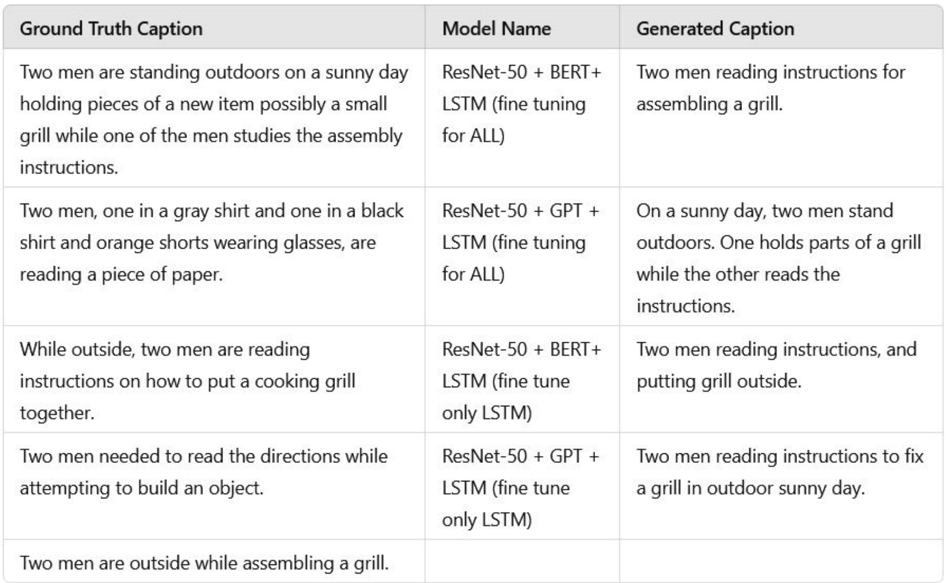
The CIDEr score evaluates how well the generated captions align with human consensus by measuring the frequency of n-grams in the generated captions relative to reference captions. The 'ResNet-50 + GPT + LSTM (fine-tuning only LSTM)' model achieved the highest CIDEr score of 48.69. This model's strong CIDEr performance demonstrates that it effectively captures and reflects common human descriptions of the image, highlighting the advantage of using deep, pre-trained models like GPT in capturing nuanced and contextually relevant descriptions.

During the inference process, the models were capable of generating accurate captions for most images, as illustrated in Figures 4 and 5. The captions demonstrate a high level of naturalness and fluency, resembling human speech and conveying meaningful details. However, for more challenging images—such as those requiring identification of ethnicity, age, or symbols, or those that are ambiguous—the quality of the generated captions may be lacking, as shown in Figures 6 and 7. We noticed that models with BERT embeddings usually limit the length of generated captions, while those with GPT tend to create longer texts.



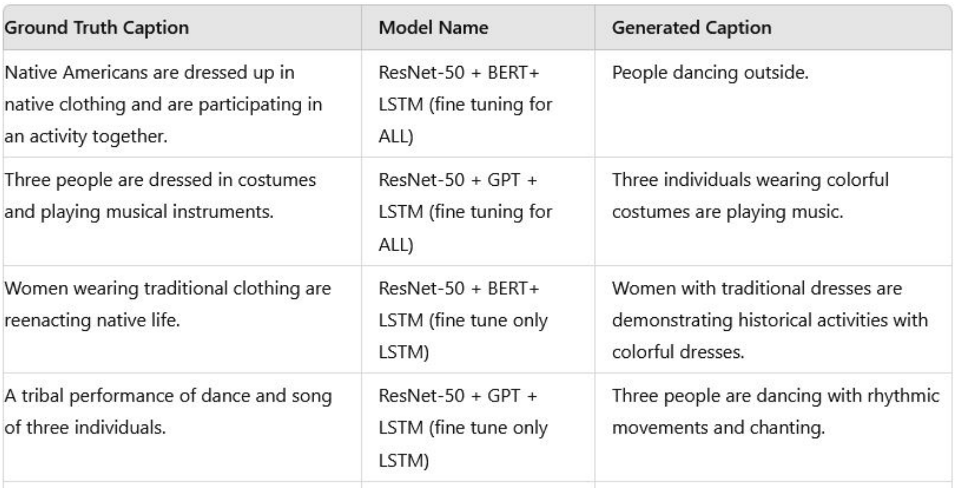
Ground Truth Caption	Model Name	Generated Caption
Older man in a hat, yellow shirt, and dress slacks playing in the sand with a young boy.	ResNet-50 + BERT+ LSTM (fine tuning for ALL)	A man with hat plays in the beach with a child.
A man wearing a hat and a small boy in a bathing suit playing in the sand.	ResNet-50 + GPT + LSTM (fine tuning for ALL)	A man with a hat and a child are playing in the sand by the water.
An older person and a small boy play in the sand near the water.	ResNet-50 + BERT+ LSTM (fine tune only LSTM)	An older man and a boy play in the sand by the beach.
A person in a hat is playing with a boy on a beach.	ResNet-50 + GPT + LSTM (fine tune only LSTM)	A man in a hat plays with a boy on the beach sand.
A father is playing with his child in the sand.		

Figure 4: Example 1 of generated captions



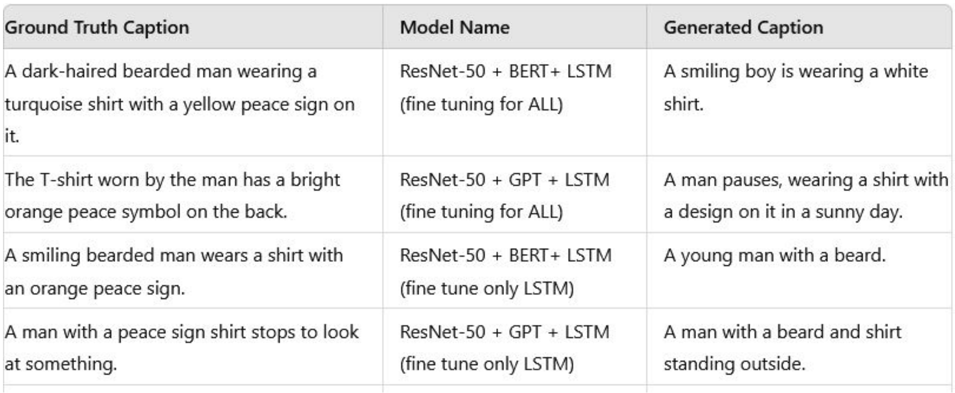
Ground Truth Caption	Model Name	Generated Caption
Two men are standing outdoors on a sunny day holding pieces of a new item possibly a small grill while one of the men studies the assembly instructions.	ResNet-50 + BERT+ LSTM (fine tuning for ALL)	Two men reading instructions for assembling a grill.
Two men, one in a gray shirt and one in a black shirt and orange shorts wearing glasses, are reading a piece of paper.	ResNet-50 + GPT + LSTM (fine tuning for ALL)	On a sunny day, two men stand outdoors. One holds parts of a grill while the other reads the instructions.
While outside, two men are reading instructions on how to put a cooking grill together.	ResNet-50 + BERT+ LSTM (fine tune only LSTM)	Two men reading instructions, and putting grill outside.
Two men needed to read the directions while attempting to build an object.	ResNet-50 + GPT + LSTM (fine tune only LSTM)	Two men reading instructions to fix a grill in outdoor sunny day.
Two men are outside while assembling a grill.		

Figure 5: Example 2 of generated captions



Ground Truth Caption	Model Name	Generated Caption
Native Americans are dressed up in native clothing and are participating in an activity together.	ResNet-50 + BERT+ LSTM (fine tuning for ALL)	People dancing outside.
Three people are dressed in costumes and playing musical instruments.	ResNet-50 + GPT + LSTM (fine tuning for ALL)	Three individuals wearing colorful costumes are playing music.
Women wearing traditional clothing are reenacting native life.	ResNet-50 + BERT+ LSTM (fine tune only LSTM)	Women with traditional dresses are demonstrating historical activities with colorful dresses.
A tribal performance of dance and song of three individuals.	ResNet-50 + GPT + LSTM (fine tune only LSTM)	Three people are dancing with rhythmic movements and chanting.

Figure 6: Example 3 of generated captions



Ground Truth Caption	Model Name	Generated Caption
A dark-haired bearded man wearing a turquoise shirt with a yellow peace sign on it.	ResNet-50 + BERT+ LSTM (fine tuning for ALL)	A smiling boy is wearing a white shirt.
The T-shirt worn by the man has a bright orange peace symbol on the back.	ResNet-50 + GPT + LSTM (fine tuning for ALL)	A man pauses, wearing a shirt with a design on it in a sunny day.
A smiling bearded man wears a shirt with an orange peace sign.	ResNet-50 + BERT+ LSTM (fine tune only LSTM)	A young man with a beard.
A man with a peace sign shirt stops to look at something.	ResNet-50 + GPT + LSTM (fine tune only LSTM)	A man with a beard and shirt standing outside.

Figure 7: Example 4 of generated captions

4.2 *Baseline comparison*

The subsequent analysis addressed the second sub-question regarding the differences in captioning accuracy and error types when using BERT or GPT versus the baseline model.

The baseline model showed lower scores, indicating less precision and coherence in the generated captions, as presented in Figures 8, 9, and 10. In contrast, models using BERT or GPT embeddings demonstrated better performance. For example, BERT embeddings, even with selective fine-tuning, led to higher METEOR and CIDEr scores, enhancing semantic accuracy and reducing errors. GPT-based models, particularly when fully fine-tuned, exhibited superior performance across all metrics, demonstrating better contextual understanding and fewer inaccuracies.

It is important to note that we reduced the batch size of the baseline model from 64 to 32 to match those used in our models due to limited computational resources. This adjustment deviates from the original training conditions described by Xu et al. (2015), where a batch size of 64 was found to significantly improve convergence speed without diminishing performance. The modification we made might explain why the baseline model showed lower scores compared to our models, indicating that pre-trained embeddings from BERT and GPT, even with a smaller batch size, provide a notable advantage in generating more accurate and meaningful captions.

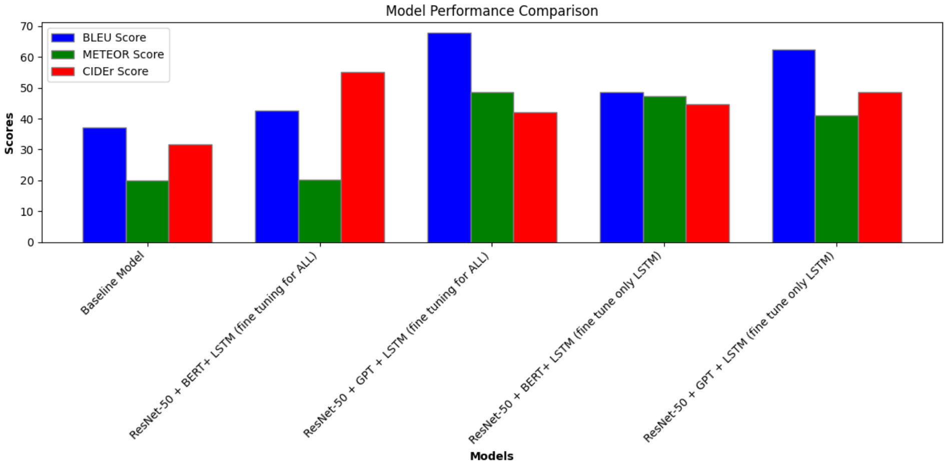
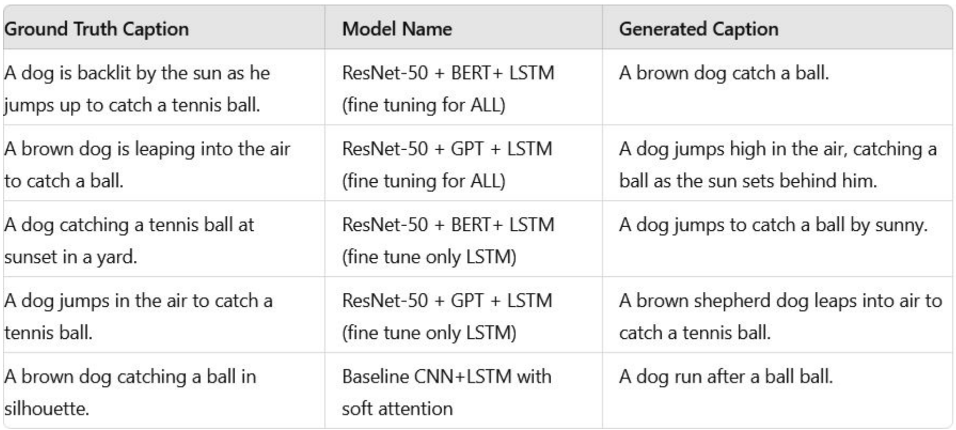
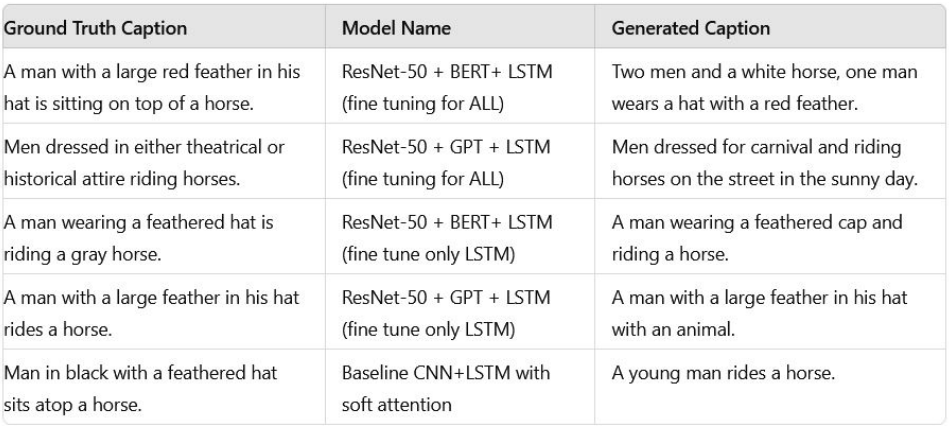


Figure 8: Comparison of baseline model performance and our proposed models.



Ground Truth Caption	Model Name	Generated Caption
A dog is backlit by the sun as he jumps up to catch a tennis ball.	ResNet-50 + BERT+ LSTM (fine tuning for ALL)	A brown dog catch a ball.
A brown dog is leaping into the air to catch a ball.	ResNet-50 + GPT + LSTM (fine tuning for ALL)	A dog jumps high in the air, catching a ball as the sun sets behind him.
A dog catching a tennis ball at sunset in a yard.	ResNet-50 + BERT+ LSTM (fine tune only LSTM)	A dog jumps to catch a ball by sunny.
A dog jumps in the air to catch a tennis ball.	ResNet-50 + GPT + LSTM (fine tune only LSTM)	A brown shepherd dog leaps into air to catch a tennis ball.
A brown dog catching a ball in silhouette.	Baseline CNN+LSTM with soft attention	A dog run after a ball ball.

Figure 9: Example 5 of generated captions

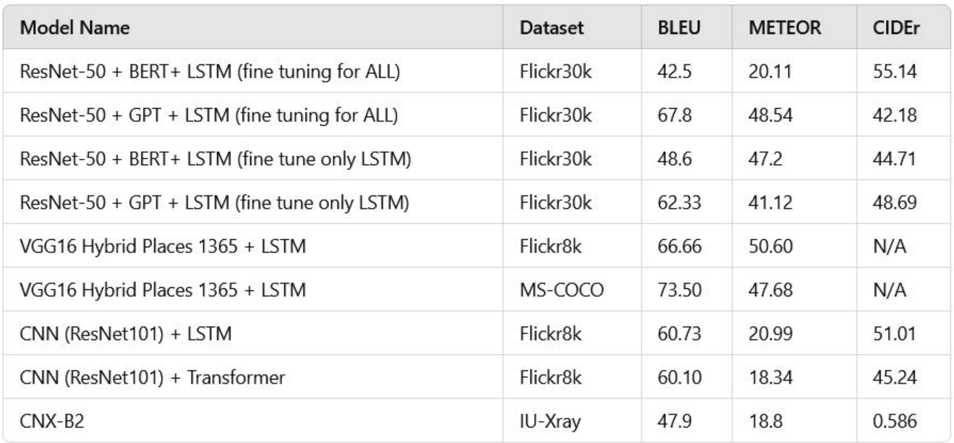


Ground Truth Caption	Model Name	Generated Caption
A man with a large red feather in his hat is sitting on top of a horse.	ResNet-50 + BERT+ LSTM (fine tuning for ALL)	Two men and a white horse, one man wears a hat with a red feather.
Men dressed in either theatrical or historical attire riding horses.	ResNet-50 + GPT + LSTM (fine tuning for ALL)	Men dressed for carnival and riding horses on the street in the sunny day.
A man wearing a feathered hat is riding a gray horse.	ResNet-50 + BERT+ LSTM (fine tune only LSTM)	A man wearing a feathered cap and riding a horse.
A man with a large feather in his hat rides a horse.	ResNet-50 + GPT + LSTM (fine tune only LSTM)	A man with a large feather in his hat with an animal.
Man in black with a feathered hat sits atop a horse.	Baseline CNN+LSTM with soft attention	A young man rides a horse.

Figure 10: Example 6 of generated captions

The improvements observed in the BERT and GPT models underscore the importance of fine-tuning and using pre-trained models. Fine-tuning allows the models to adapt specifically to the IC task, enhancing their ability to generate relevant and accurate captions. Pre-trained models like BERT and GPT, which have been trained on extensive datasets, bring rich linguistic and contextual knowledge, significantly improving performance over the baseline. This pre-existing knowledge helps generate captions that are not only more accurate but also more contextually appropriate and semantically coherent.

We then compared the performance of recent IC architectures with our models, as shown in Figure 11. The comparison reveals that both the choice of architecture and dataset have a significant impact on the results.



Model Name	Dataset	BLEU	METEOR	CIDEr
ResNet-50 + BERT+ LSTM (fine tuning for ALL)	Flickr30k	42.5	20.11	55.14
ResNet-50 + GPT + LSTM (fine tuning for ALL)	Flickr30k	67.8	48.54	42.18
ResNet-50 + BERT+ LSTM (fine tune only LSTM)	Flickr30k	48.6	47.2	44.71
ResNet-50 + GPT + LSTM (fine tune only LSTM)	Flickr30k	62.33	41.12	48.69
VGG16 Hybrid Places 1365 + LSTM	Flickr8k	66.66	50.60	N/A
VGG16 Hybrid Places 1365 + LSTM	MS-COCO	73.50	47.68	N/A
CNN (ResNet101) + LSTM	Flickr8k	60.73	20.99	51.01
CNN (ResNet101) + Transformer	Flickr8k	60.10	18.34	45.24
CNX-B2	IU-Xray	47.9	18.8	0.586

Figure 11: Comparison of recent IC architectures performances and our proposed models.

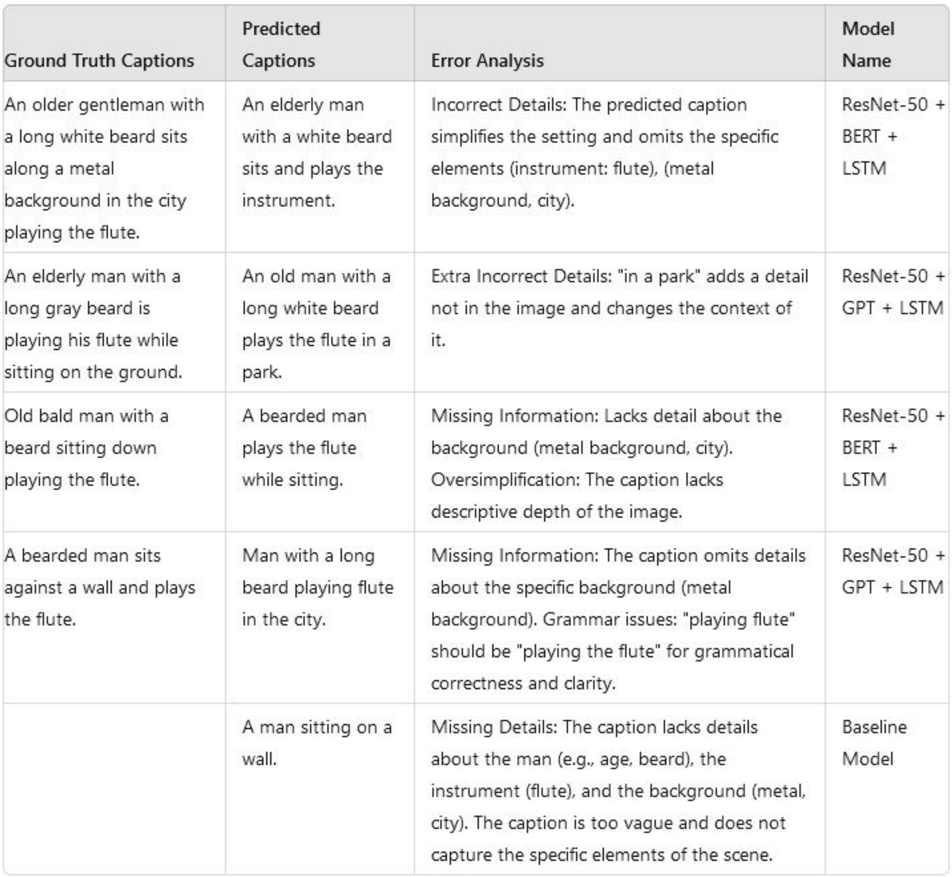
Our models show competitive performance, particularly the ones utilizing GPT and fine-tuning approaches. Fine-tuning for all components, especially in models like ResNet-50 + GPT + LSTM, has proven beneficial, yielding high scores across multiple metrics.

While the models trained on Flickr30k may not always surpass those trained on smaller datasets like Flickr8k, the results are commendable given the training limitations. The performance highlights the effectiveness of advanced architectures and fine-tuning strategies in achieving high-quality image captioning.

4.3 *Error Analysis*

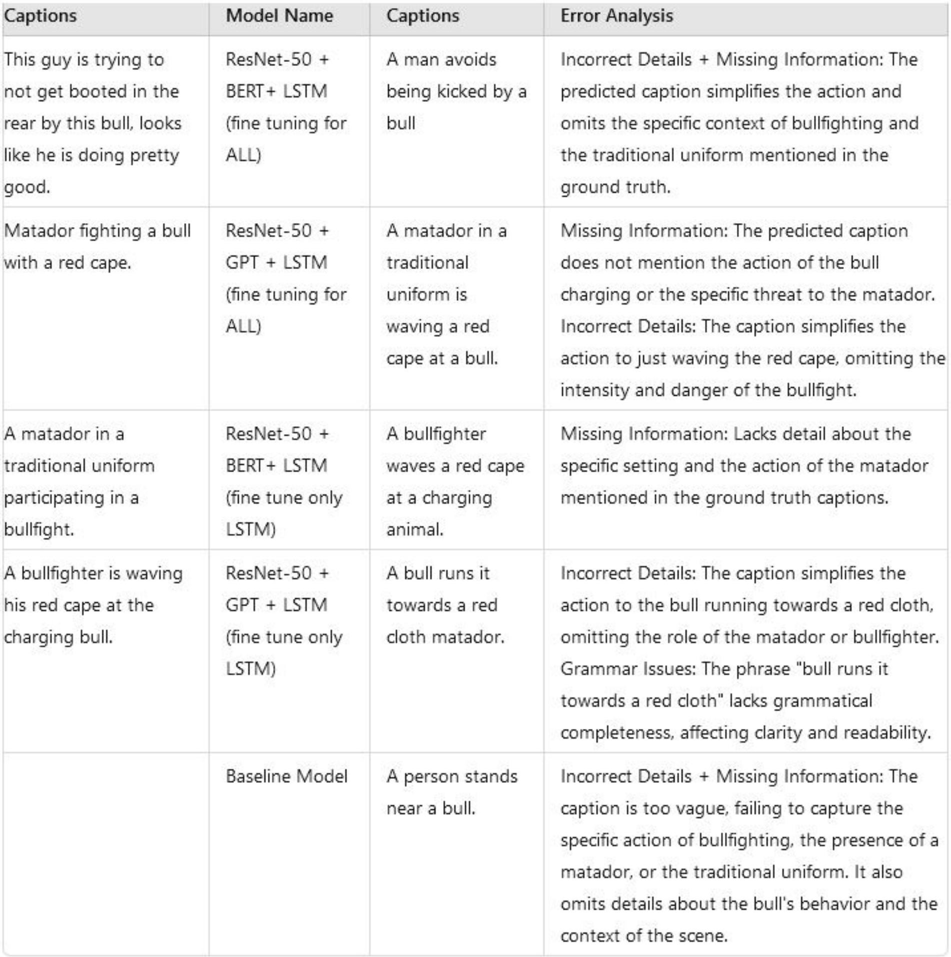
As described in section 3.9 and 3.10, we conducted an evaluation of the best and lowest-scored predicted captions. The analysis aimed to answer the last sub-question about which captioning errors occur most frequently and the underlying reasons behind recurring error patterns.

Our inspection of selected captions allowed us to identify and categorize different types of errors. These categories included repetition, such as repeated words or phrases; missing information, where important objects or actions were omitted; incorrect details, involving inaccurate descriptions of objects or actions; grammar issues that affected readability; and irrelevant information, where unnecessary or unrelated details were included. Examples of generated captions with detailed error analysis are presented in Figures 12 and 13.



Ground Truth Captions	Predicted Captions	Error Analysis	Model Name
An older gentleman with a long white beard sits along a metal background in the city playing the flute.	An elderly man with a white beard sits and plays the instrument.	Incorrect Details: The predicted caption simplifies the setting and omits the specific elements (instrument: flute), (metal background, city).	ResNet-50 + BERT + LSTM
An elderly man with a long gray beard is playing his flute while sitting on the ground.	An old man with a long white beard plays the flute in a park.	Extra Incorrect Details: "in a park" adds a detail not in the image and changes the context of it.	ResNet-50 + GPT + LSTM
Old bald man with a beard sitting down playing the flute.	A bearded man plays the flute while sitting.	Missing Information: Lacks detail about the background (metal background, city). Oversimplification: The caption lacks descriptive depth of the image.	ResNet-50 + BERT + LSTM
A bearded man sits against a wall and plays the flute.	Man with a long beard playing flute in the city.	Missing Information: The caption omits details about the specific background (metal background). Grammar issues: "playing flute" should be "playing the flute" for grammatical correctness and clarity.	ResNet-50 + GPT + LSTM
	A man sitting on a wall.	Missing Details: The caption lacks details about the man (e.g., age, beard), the instrument (flute), and the background (metal, city). The caption is too vague and does not capture the specific elements of the scene.	Baseline Model

Figure 12: Example 1 of prediction errors



Captions	Model Name	Captions	Error Analysis
This guy is trying to not get booted in the rear by this bull, looks like he is doing pretty good.	ResNet-50 + BERT+ LSTM (fine tuning for ALL)	A man avoids being kicked by a bull	Incorrect Details + Missing Information: The predicted caption simplifies the action and omits the specific context of bullfighting and the traditional uniform mentioned in the ground truth.
Matador fighting a bull with a red cape.	ResNet-50 + GPT + LSTM (fine tuning for ALL)	A matador in a traditional uniform is waving a red cape at a bull.	Missing Information: The predicted caption does not mention the action of the bull charging or the specific threat to the matador. Incorrect Details: The caption simplifies the action to just waving the red cape, omitting the intensity and danger of the bullfight.
A matador in a traditional uniform participating in a bullfight.	ResNet-50 + BERT+ LSTM (fine tune only LSTM)	A bullfighter waves a red cape at a charging animal.	Missing Information: Lacks detail about the specific setting and the action of the matador mentioned in the ground truth captions.
A bullfighter is waving his red cape at the charging bull.	ResNet-50 + GPT + LSTM (fine tune only LSTM)	A bull runs it towards a red cloth matador.	Incorrect Details: The caption simplifies the action to the bull running towards a red cloth, omitting the role of the matador or bullfighter. Grammar Issues: The phrase "bull runs it towards a red cloth" lacks grammatical completeness, affecting clarity and readability.
	Baseline Model	A person stands near a bull.	Incorrect Details + Missing Information: The caption is too vague, failing to capture the specific action of bullfighting, the presence of a matador, or the traditional uniform. It also omits details about the bull's behavior and the context of the scene.

Figure 13: Example 2 of prediction errors

The ResNet-50 + GPT + LSTM (fine-tuning for all) model achieved the highest BLEU score (67.8) and METEOR score (48.54), but a lower CIDEr score (42.18). This discrepancy suggests that while the model excelled in generating syntactically and semantically correct phrases, it struggled with capturing the overall contextual richness as expected by human annotators. Common errors for this model included missing information and incorrect details. For example, captions often simplified actions or omitted critical context, such as the intensity of a bullfight or the specific setting as explained in Figure 13, which the CIDEr penalized heavily.

The ResNet-50 + BERT + LSTM (fine-tuning for all) model, on the other hand, scored lower on BLEU (42.5) and METEOR (20.11), but had a relatively higher CIDEr score (55.14). This model frequently produced captions with correct details but included grammatical issues and irrelevant information, which affected its precision and semantic alignment, as reflected by the lower BLEU and METEOR scores. For instance, it might describe an action accurately but fail in grammatical completeness or introduce unnecessary details, thereby reducing the overall clarity and readability

When fine-tuning only the LSTM, the ResNet-50 + BERT + LSTM model showed moderate improvements in BLEU (48.6) and METEOR (47.2) scores but a lower CIDEr score (44.71). This configuration often omitted significant details and context, leading to an oversimplification of the scenes. The absence of specific background elements or actions frequently resulted in a lack of descriptive depth, which negatively impacted the CIDEr score.

The ResNet-50 + GPT + LSTM (fine-tune only LSTM) model achieved a high BLEU score (62.33) and a reasonable METEOR score (41.12) with a moderate CIDEr score (48.69). This model generally produced grammatically correct and fluent captions but often omitted crucial elements, such as specific settings or objects, resulting in lower relevance and informativeness. These missing details and grammatical issues were common errors, leading to penalties in both METEOR and CIDEr scores.

Overall, the analysis revealed that the models exhibited distinct error patterns that affected their performance across different evaluation metrics. High BLEU scores were associated with precise phrasing, but often lacked contextual richness, impacting CIDEr scores. Higher METEOR scores indicated better semantic content but did not always correlate with overall relevance and informativeness, as measured by CIDEr. The most common errors included repetition, missing information, incorrect details, grammar issues, and irrelevant information, each influencing the scores differently, based on the nature of the evaluated metrics.

4.4 *Recurring error patterns*

As mentioned in Section 3.10, our analysis continues to address the last sub-question by conducting a detailed evaluation to understand the nature and frequency of common mistakes made by each model.

To simplify the process, we analyzed the bottom 5 percent of predictions with the lowest scores, totaling 238 captions, and categorized errors into four types only: repetition, missing information, incorrect details, and grammar issues. The general analysis showed that models made fewer errors in repetition and grammar issues compared to missing information and incorrect details. The final distribution of errors for each model is illustrated in Figure 14.

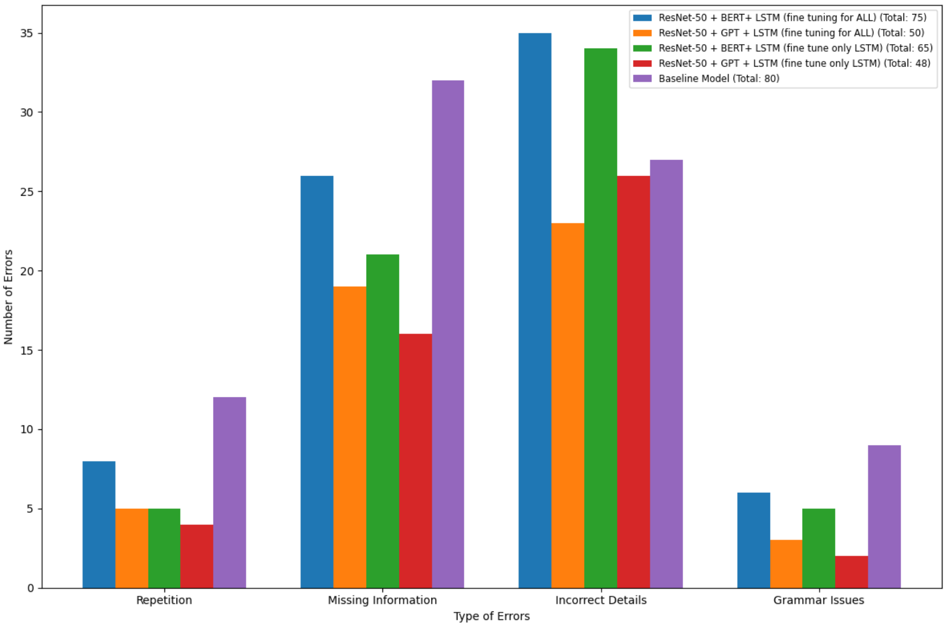
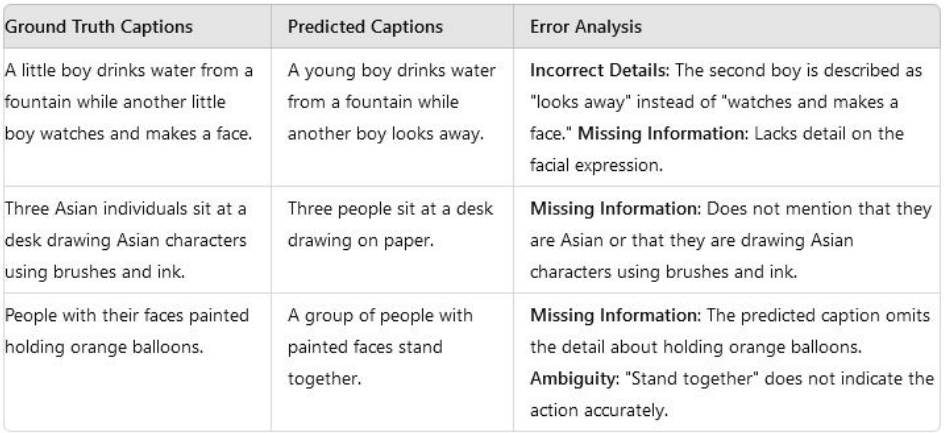


Figure 14: Distribution of error types per model

The distribution shown in Figure 14 reflects the overall performance of each model, with varying degrees of proficiency in generating accurate and coherent captions. Errors related to missing information and incorrect details were more prevalent, indicating areas where the models often struggled to capture the complete context or provided inaccurate descriptions, as seen in Figures 15 and 16. Repetition and grammar issues were less frequent but still notable, affecting the quality of the captions, as illustrated in Figures 17, 18, and 19.

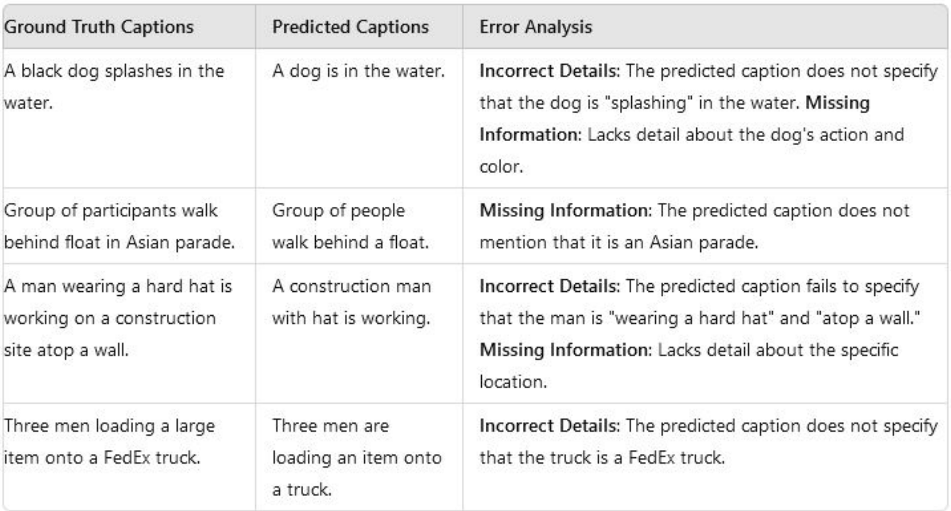
By examining Figure 16 closely, we noticed that incorrect details are prevalent when the predicted captions fail to specify critical elements of the scene, such as the action of a dog "splashing" in the water or the presence of a "hard hat", and the man being "atop a wall". These omissions result in captions that do not accurately reflect the image's key aspects.

Missing information frequently occurs when captions omit significant details, as seen in Figures 15 and 16. We can observe errors such as the context of an 'Asian parade' or specific elements like the color of a dog or the objects people are holding. The absence of these details leads to incomplete or ambiguous descriptions that do not fully convey the image's content.



Ground Truth Captions	Predicted Captions	Error Analysis
A little boy drinks water from a fountain while another little boy watches and makes a face.	A young boy drinks water from a fountain while another boy looks away.	**Incorrect Details:** The second boy is described as "looks away" instead of "watches and makes a face." **Missing Information:** Lacks detail on the facial expression.
Three Asian individuals sit at a desk drawing Asian characters using brushes and ink.	Three people sit at a desk drawing on paper.	**Missing Information:** Does not mention that they are Asian or that they are drawing Asian characters using brushes and ink.
People with their faces painted holding orange balloons.	A group of people with painted faces stand together.	**Missing Information:** The predicted caption omits the detail about holding orange balloons. **Ambiguity:** "Stand together" does not indicate the action accurately.

Figure 15: CNN + BERT + LSTM fine tuning all



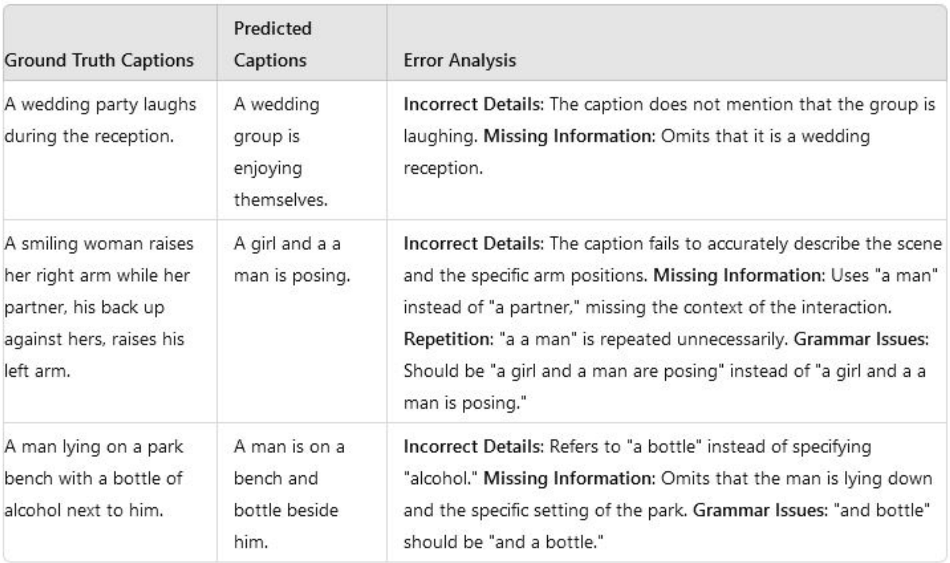
Ground Truth Captions	Predicted Captions	Error Analysis
A black dog splashes in the water.	A dog is in the water.	**Incorrect Details:** The predicted caption does not specify that the dog is "splashing" in the water. **Missing Information:** Lacks detail about the dog's action and color.
Group of participants walk behind float in Asian parade.	Group of people walk behind a float.	**Missing Information:** The predicted caption does not mention that it is an Asian parade.
A man wearing a hard hat is working on a construction site atop a wall.	A construction man with hat is working.	**Incorrect Details:** The predicted caption fails to specify that the man is "wearing a hard hat" and "atop a wall." **Missing Information:** Lacks detail about the specific location.
Three men loading a large item onto a FedEx truck.	Three men are loading an item onto a truck.	**Incorrect Details:** The predicted caption does not specify that the truck is a FedEx truck.

Figure 16: CNN + GPT + LSTM fine tuning LSTM

Figure 17 and 19 demonstrate how grammar issues also contribute to caption inaccuracies, particularly when captions use incorrect phrasing or lack necessary grammatical elements. For instance, phrases like "A girl and a a man is" should be corrected to "A girl and a man are" to properly convey the action. Similarly, "Beside bike" should be corrected to "beside the bike" to ensure grammatical correctness.

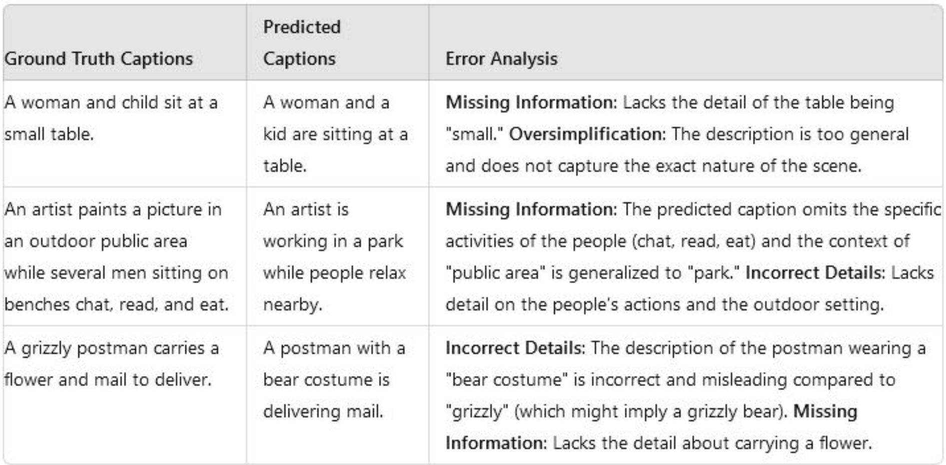
Notably, our models exhibit fewer errors in the generated captions compared to the baseline model, both in terms of the number of errors and their severity. This suggests that the advanced architectures and fine-tuning strategies used in our models not only reduce the frequency of errors but also mitigate their impact, leading to more accurate and reliable caption generation.

Addressing these issues involves improving the models' ability to capture and integrate detailed visual information, refining the alignment between visual and textual data, and enhancing the grammatical quality of the generated captions.



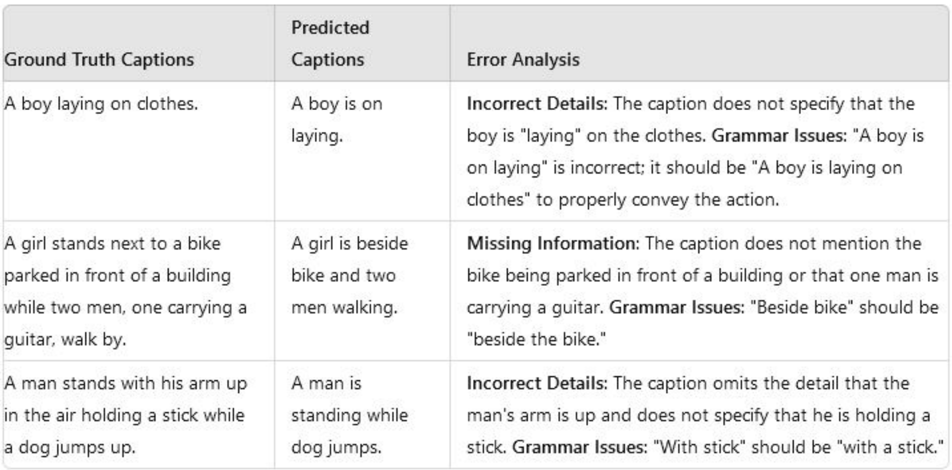
Ground Truth Captions	Predicted Captions	Error Analysis
A wedding party laughs during the reception.	A wedding group is enjoying themselves.	**Incorrect Details**: The caption does not mention that the group is laughing. **Missing Information**: Omits that it is a wedding reception.
A smiling woman raises her right arm while her partner, his back up against hers, raises his left arm.	A girl and a a man is posing.	**Incorrect Details**: The caption fails to accurately describe the scene and the specific arm positions. **Missing Information**: Uses "a man" instead of "a partner," missing the context of the interaction. **Repetition**: "a a man" is repeated unnecessarily. **Grammar Issues**: Should be "a girl and a man are posing" instead of "a girl and a a man is posing."
A man lying on a park bench with a bottle of alcohol next to him.	A man is on a bench and bottle beside him.	**Incorrect Details**: Refers to "a bottle" instead of specifying "alcohol." **Missing Information**: Omits that the man is lying down and the specific setting of the park. **Grammar Issues**: "and bottle" should be "and a bottle."

Figure 17: CNN + BERT + LSTM fine tuning LSTM



Ground Truth Captions	Predicted Captions	Error Analysis
A woman and child sit at a small table.	A woman and a kid are sitting at a table.	**Missing Information**: Lacks the detail of the table being "small." **Oversimplification**: The description is too general and does not capture the exact nature of the scene.
An artist paints a picture in an outdoor public area while several men sitting on benches chat, read, and eat.	An artist is working in a park while people relax nearby.	**Missing Information**: The predicted caption omits the specific activities of the people (chat, read, eat) and the context of "public area" is generalized to "park." **Incorrect Details**: Lacks detail on the people's actions and the outdoor setting.
A grizzly postman carries a flower and mail to deliver.	A postman with a bear costume is delivering mail.	**Incorrect Details**: The description of the postman wearing a "bear costume" is incorrect and misleading compared to "grizzly" (which might imply a grizzly bear). **Missing Information**: Lacks the detail about carrying a flower.

Figure 18: CNN + GPT + LSTM fine tuning all



Ground Truth Captions	Predicted Captions	Error Analysis
A boy laying on clothes.	A boy is on laying.	**Incorrect Details:** The caption does not specify that the boy is "laying" on the clothes. **Grammar Issues:** "A boy is on laying" is incorrect; it should be "A boy is laying on clothes" to properly convey the action.
A girl stands next to a bike parked in front of a building while two men, one carrying a guitar, walk by.	A girl is beside bike and two men walking.	**Missing Information:** The caption does not mention the bike being parked in front of a building or that one man is carrying a guitar. **Grammar Issues:** "Beside bike" should be "beside the bike."
A man stands with his arm up in the air holding a stick while a dog jumps up.	A man is standing while dog jumps.	**Incorrect Details:** The caption omits the detail that the man's arm is up and does not specify that he is holding a stick. **Grammar Issues:** "With stick" should be "with a stick."

Figure 19: Baseline model

5 DISCUSSION

The primary objective of this research was to assess whether BERT and GPT can enhance the contextual understanding and semantic coherence of image captions within a hybrid CNN-LSTM-transformer architecture. We also expanded on this by investigating three sub-questions.

The accuracy of the hybrid models compared to our baseline model and recent architectures, considering the effects of fine-tuning and freezing pre-trained model weights

The differences in captioning accuracy and error types between BERT and GPT versus the baseline model.

The most frequent captioning errors and their underlying causes, in order to identify challenges and areas for improvement.

5.1 *Result discussion*

To answer the main research question, we examined and compared the performance of four distinct models with the baseline model and analyzed their ability to generate accurate and contextually rich captions for images. Our findings provide comprehensive insights into the advantages and limitations of different approaches, particularly focusing on the impacts of fine-tuning versus freezing components and the comparative benefits of BERT or GPT embeddings versus the baseline model

Our results show that models using GPT embeddings with full fine-tuning significantly outperformed others across various metrics. The "ResNet-50 + GPT + LSTM (fine-tuning for ALL)" model achieved the highest BLEU score of 67.8 and a METEOR score of 48.54, demonstrating enhanced syntactic and semantic accuracy. However, its slightly lower CIDEr score suggests there is still room for improvement in contextual richness. Models with only LSTM fine-tuned, such as "ResNet-50 + BERT + LSTM (fine-tune only LSTM)," also performed well, indicating the benefits of focusing on LSTM refinement. GPT embeddings consistently outperformed BERT embeddings, as shown by the "ResNet-50 + GPT + LSTM (fine-tune only LSTM)" model's top CIDEr score of 48.69, highlighting GPT's advantage in handling longer text sequences and maintaining contextual coherence in IC tasks.

Furthermore, when compared to the baseline model or recent IC architectures, our models—particularly those that use GPT embeddings with fine-tuning—showed competitive performance. Although they did not always outperform models trained on smaller datasets like Flickr8k, the results remain commendable given the training constraints, underscoring the potential of advanced architectures and fine-tuning strategies in the high-quality IC.

To ensure a fair and accurate comparison, we used the same data and settings across all models. While better hyperparameters might exist for our specific data and architectures, the goal of this study was not to identify the optimal set. Instead, we aimed to use a robust set of hyperparameters to establish a strong baseline for comparing the embedding approaches. Fine-tuning these settings could potentially improve performance, suggesting that future research could explore further adjustments to fully leverage our approach.

It is important to note that the computation of evaluation metrics can vary significantly due to differences in the models' tokenization and embedding strategies. BLEU, which focuses on n-gram precision, can be influenced by the tokenization process, affecting how accurately n-grams are matched between generated and reference captions. METEOR, which considers synonyms and stemming, may yield different results based on the embedding models' capacity to capture semantic nuances. CIDEr, designed to reflect human judgment, can also vary depending on how well the models' embeddings represent the context and relevance of words in a caption.

Finally, our error analysis identified recurring patterns that impacted caption quality, such as missing information, incorrect details, repetition, and grammatical errors. Models with high BLEU scores often lacked contextual richness, impacting their CIDEr scores. Higher METEOR scores were indicative of better semantic content but did not always correlate with overall relevance. The detailed root cause analysis identified missing information and incorrect details as the most frequent errors, suggesting that while the models are proficient in generating syntactically correct phrases, they often struggle with capturing comprehensive contextual elements.

By analyzing these extremes, we gained deeper insights into the strengths and weaknesses of our models, highlighting where they excel and where improvements are needed. This classification process is crucial for understanding model behavior and guiding future refinements in IC tasks.

5.2 *Limitations*

Despite the promising results, this study has several limitations. One major constraint was the limited availability of computational resources, which significantly impacted our ability to experiment with larger models and datasets. Training deep learning models, particularly those involving advanced architectures, requires substantial computational power that was beyond our reach.

The choice of the Flickr30k dataset, although manageable in size, imposed limitations compared to larger datasets such as MS COCO, which consists of over 120,000 images and provides a more extensive benchmark for IC models. The volume of such data would have introduced significant bottlenecks in preprocessing images for training, making it infeasible within our constraints.

We also opted for ResNet-50 as our CNN model due to its balance between performance and computational efficiency. While more sophisticated models, such as ResNet-101 or ResNeXt, could enhance feature extraction and overall model performance, their higher computational and memory requirements were impractical given our constraints. Training such models with the limited computational resources available would have been extremely time-consuming and likely infeasible.

To manage resource constraints, we implemented a checkpoint training mechanism to save model progress after each epoch, which allowed training to resume after interruptions or hardware failures. While this approach mitigated some resource limitations, it did not completely overcome them. Despite these challenges, our study demonstrates the potential of hybrid CNN-LSTM-transformer architectures with BERT and GPT embeddings. However, future research with more powerful GPUs and larger datasets is needed for a more thorough evaluation and to further enhance image captioning performance.

5.3 *Implications and Future Directions*

The findings of this study highlight several important implications for both society and the field of image captioning. Integrating BERT and GPT embeddings into a hybrid CNN-LSTM-transformer architecture significantly enhances the contextual understanding and semantic coherence of image captions. This hybrid approach effectively bridges gaps in the literature by combining the strengths of spatial and semantic feature extraction, offering a robust solution for generating more accurate and meaningful captions.

In terms of societal implications, the ability to generate more accurate and contextual captions is crucial for improving accessibility, especially for visually impaired individuals, information retrieval systems, and automated content moderation systems. The recurring errors we found in captions can lead to misunderstandings about visual content. Improving contextual understanding would reduce these errors, making captions more reliable, informative, and helpful for users with impairments. Current issues such as repetition and irrelevant details are critical in information retrieval systems, as they reduce the precision of captions, making it difficult for search algorithms to match user queries with the most relevant content. Moreover, moderation systems depend on precise descriptions to detect inappropriate content. Errors such as omitting key details or providing incorrect information can lead to incorrect content moderation decisions, either flagging harmless content wrongly or failing to detect harmful material. Potential solutions, like those proposed by Qu et al. (2023) and Yan et al. (2016), include incorporating additional layers of semantic analysis to better understand and convey complex scenes. Implementing such approaches could address the errors revealed by our study. Future research could explore these solutions to further refine image captioning systems and mitigate the identified errors.

Despite training on a larger dataset like Flickr30k, our models did not always outperform those trained on smaller datasets, such as Flickr8k, partly due to the influence of hyperparameters. A larger dataset increases the complexity of the models' training, requiring more careful tuning of hyperparameters to effectively learn from the data. In our study, the primary aim was to establish a strong baseline rather than fine-tune hyperparameters for optimal performance. This approach likely resulted in some models not reaching their highest potential scores, highlighting the need for further hyperparameter optimization to fully leverage the advantages of training on a larger dataset. Future work could focus on refining these settings and potentially employ novel hyperparameter tuning methods, such as those proposed by Patel and Varier (2020) or Arasi et al. (2023), to enhance model performance. Tailoring hyperparameter adjustments could allow our models to surpass those trained on smaller datasets.

In summary, the study underscores the efficacy of hybrid models in image captioning and highlights the significant benefits of fine-tuning. It also emphasizes the importance of ongoing refinement to enhance the models' ability to generate contextually rich and semantically coherent captions, thereby improving their practical applicability and performance. Future research should continue to explore these directions, focusing on optimizing both the architecture and training processes to further advance the field.

6 CONCLUSION

Overall, this study contributes to the field of NLP and the image processing, as well as to the subfield of IC. It explores the impact of BERT and GPT embeddings within a hybrid CNN-LSTM-transformer architecture for IC by addressing the following questions:

Main RQ: To what extent do BERT and GPT enhance the contextual understanding and semantic coherence of image captions within a hybrid CNN-LSTM-transformer architecture? Our findings demonstrate that GPT embeddings significantly improve the contextual understanding and semantic coherence of image captions. The "ResNet-50 + GPT + LSTM (fine-tuning for ALL)" model achieved the highest performance metrics, including a BLEU score of 67.8 and a METEOR score of 48.54. These results highlight the effectiveness of GPT embeddings in capturing complex linguistic structures and context in IC tasks.

Sub-RQ 1: What is the accuracy and performance of the proposed hybrid models? The proposed hybrid models, particularly those utilizing GPT embeddings and comprehensive fine-tuning, outperformed the other models. The "ResNet-50 + GPT + LSTM (fine-tuning for ALL)" model demonstrated superior syntactic and semantic accuracy, underscoring the advantage of integrating advanced transformer embeddings in hybrid models.

Sub-RQ 2: What are the differences in captioning accuracy and error type when using BERT or GPT versus baseline model? Our experiments revealed that models using GPT embeddings generally achieved higher accuracy and better semantic coherence compared to those using BERT embeddings and the baseline model. While BERT-based models, such as "ResNet-50 + BERT + LSTM (fine-tune only LSTM)", showed notable performance improvements, they were less effective than GPT-based models. This difference is attributed to GPT's superior ability to handle complex contextual information, leading to more accurate and contextually relevant captions. Overall, our models outperformed the baseline model in most of the scores.

Sub-RQ 3: What captioning errors occur most frequently, and what are the underlying reasons behind the recurring error patterns? Error analysis identified recurring issues such as missing information and incorrect details, particularly in models that were not fully fine-tuned. These errors often stemmed from insufficient contextual understanding and semantic richness. The findings suggest that future models should focus on enhancing the contextual richness and comprehensive fine-tuning to minimize such errors and better align with human expectations.

7 SOURCE/CODE/ETHICS/TECHNOLOGY STATEMENT

There was no data collection from either human or animal subjects in this research. The authors acknowledge that they have no legal claim to the data they utilized, which is publicly accessible. QuillBot.com was used to check grammar and spelling

REFERENCES

Alqahtani, F. F., Mohsan, M. M., Alshamrani, K., Zeb, J., Alhamami, S., & Alqarni, D. (2024). Cnx-b2: A novel cnn-transformer approach for chest x-ray medical report generation. *IEEE Access*, *12*, 26626–26635.

Aneja, J., Agrawal, H., Batra, D., & Schwing, A. (2019). Sequential latent spaces for modeling the intention during diverse image captioning. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4261–4270.

Arasi, M. A., Alshahrani, H. M., Alruwais, N., Motwakel, A., Ahmed, N. A., & Mohamed, A. (2023). Automated image captioning using sparrow search algorithm with improved deep learning model. *IEEE Access*.

Bai, Y., Mei, J., Yuille, A. L., & Xie, C. (2021). Are transformers more robust than cnns? *Advances in neural information processing systems*, *34*, 26831–26843.

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., et al. (2022). Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fang, W., Chen, Y., & Xue, Q. (2021). Survey on research of rnn-based spatio-temporal sequence prediction algorithms. *Journal on Big Data*, *3*(3), 97.

Fournier, Q., Caron, G. M., & Aloise, D. (2023). A practical survey on faster and lighter transformers. *ACM Computing Surveys*, *55*(14s), 1–40.

Ghojogh, B., & Ghodsi, A. (2020). Attention mechanism, transformers, bert, and gpt: Tutorial and survey.

González-Chávez, O., Ruiz, G., Moctezuma, D., & Ramirez-delReal, T. (2024). Are metrics measuring what they should? an evaluation of image captioning task metrics. *Signal Processing: Image Communication*, *120*, 117071.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Hu, X., Yin, X., Lin, K., Zhang, L., Gao, J., Wang, L., & Liu, Z. (2021). Vivo: Visual vocabulary pre-training for novel object captioning. *proceedings of the AAAI conference on artificial intelligence*, *35*(2), 1575–1583.

Li, C., Yang, Y., Liang, H., & Wu, B. (2021). Transfer learning for establishment of recognition of covid-19 on ct imaging using small-sized training datasets. *Knowledge-Based Systems*, *218*, 106849.

Li, Y., Pan, Y., Yao, T., & Mei, T. (2022). Comprehending and ordering semantics for image captioning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17990–17999.

MacAvaney, S., Nardini, F. M., Perego, R., Tonellotto, N., Goharian, N., & Frieder, O. (2020). Efficient document re-ranking for transformers by precomputing term representations. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 49–58.

Mahajan, S., & Roth, S. (2020). Diverse image captioning with context-object split latent spaces. *Advances in Neural Information Processing Systems*, *33*, 3613–3624.

Nafi'iyah, N., & Setyati, E. (2021). Lung x-ray image enhancement to identify pneumonia with cnn. *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, 421–426.

Nakamura, A., & Harada, T. (2019). Revisiting fine-tuning for few-shot learning. *arXiv preprint arXiv:1910.00216*.

Ondeng, O., Ouma, H., & Akuon, P. (2023). A review of transformer-based approaches for image captioning. *Applied Sciences*, *13*(19), 11103.

Patel, A., & Varier, A. (2020). Hyperparameter analysis for image captioning. *arXiv preprint arXiv:2006.10923*.

Qu, X., Che, H., Huang, J., Xu, L., & Zheng, X. (2023). Multi-layered semantic representation network for multi-label image classification. *International Journal of Machine Learning and Cybernetics*, *14*(10), 3427–3435.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Rochmawanti, O., & Utaminingrum, F. (2021). Chest x-ray image to classify lung diseases in different resolution size using densenet-121 architectures. *Proceedings of the 6th International Conference on Sustainable Information Engineering and Technology*, 327–331.

Saeidimesineh, R., Adibi, P., Karshenas, H., & Darvishy, A. (2023). Parallel encoder–decoder framework for image captioning. *Knowledge-Based Systems*, *282*, 111056.

Satti, S. K., Rajareddy, G. N., Maddula, P., & Ravipati, N. V. (2023). Image caption generation using resnet-50 and lstm. *2023 IEEE Silchar Subsection Conference (SILCON)*, 1–6.

Shen, T., Zhou, T., Long, G., Jiang, J., Wang, S., & Zhang, C. (2018). Reinforced self-attention network: A hybrid of hard and soft attention for sequence modeling. *arXiv preprint arXiv:1801.10296*.

Vasireddy, I., HimaBindu, G., & Ratnamala, B. (2023). Transformative fusion: Vision transformers and gpt-2 unleashing new frontiers in image captioning within image processing. *International Journal of Innovative Research in Engineering & Management*, *10*(6), 55–59.

Verma, A., Yadav, A. K., Kumar, M., & Yadav, D. (2024). Automatic image caption generation using deep learning. *Multimedia Tools and Applications*, *83*(2), 5309–5325.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.

Wang, Q., & Chan, A. B. (2019). Describing like humans: On diversity in image captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4195–4203.

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

Yan, Z., Zhang, H., Jia, Y., Breuel, T., & Yu, Y. (2016). Combining the best of convolutional layers and recurrent layers: A hybrid network for semantic segmentation. *arXiv preprint arXiv:1603.04871*.

Zhou, Y., & Srikumar, V. (2021). A closer look at how fine-tuning changes bert. *arXiv preprint arXiv:2106.14282*.

Zhu, D., Chen, J., Haydarov, K., Shen, X., Zhang, W., & Elhoseiny, M. (2023). Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *arXiv preprint arXiv:2303.06594*.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, *109*(1), 43–76.