



Image Captioning Report

Under the Guidance of:

Dr. Amol Bhopale

Table Of Contents

Abstract	3
Introduction	3
Literature Review	4
Proposed Methodology	7
Experimental Setup	12
Results	14
Result Analysis & discussion	17
Conclusion & Future Scope	17
References	18

Image Captioning Using LSTM And Transformer

1. Abstract

Image captioning is a challenging task in the field of computer vision and natural language processing, where the goal is to generate a descriptive caption for an input image automatically. In this project, we propose a novel approach to address the image captioning task using a combination of convolutional neural networks (CNNs) for image feature extraction and recurrent neural networks (RNNs) for sequence generation. Our methodology leverages the power of pre-trained CNNs, specifically the popular ResNet architecture, to extract high-level features from input images. These features are then fed into an LSTM-based RNN model, which generates a sequence of words to form a coherent caption. To further enhance the performance of our captioning system, we later integrate a Transformer-based model, leveraging its advanced architecture and self-attention mechanisms to capture long-range dependencies and contextual nuances in images. To train our models, we utilize a large-scale dataset of images paired with human-annotated captions. We evaluate the performance of both the LSTM-based and Transformer-based models using standard evaluation metrics such as BLEU score, METEOR score. Experimental results demonstrate the effectiveness of our approach in generating accurate and contextually relevant captions for a wide range of images. Additionally, we conduct extensive ablation studies to analyse the impact of different components of our models on their performance. Our findings not only contribute to the advancement of image captioning research but also hold promise for practical applications in areas such as image retrieval, assistive technologies, and content generation. By leveraging both LSTM and Transformer architectures, we aim to push the boundaries of AI-driven image understanding and interpretation, fostering further research in the intersection of computer vision and natural language processing.

2. Introduction

In recent years, the convergence of computer vision and natural language processing has led to significant advancements in tasks such as image understanding and captioning. Image captioning, in particular, has emerged as a challenging yet crucial problem in the field, with applications ranging from assistive technologies to content generation in social media platforms. The fundamental goal of image captioning is to automatically generate a descriptive and contextually relevant caption for a given input image, thereby bridging the semantic gap between visual and textual modalities.

Motivated by the growing interest in this interdisciplinary domain, our project aims to explore novel methodologies for tackling the image captioning task. By leveraging state-of-the-art

deep learning techniques, we seek to develop a robust and efficient model capable of generating coherent captions that accurately describe the content of diverse images. Our approach draws inspiration from the success of convolutional neural networks (CNNs) in image feature extraction and recurrent neural networks (RNNs) in sequence modeling, building upon the strengths of each architecture to achieve superior performance in caption generation.

Through this project, we endeavor to address several key challenges inherent to image captioning, including the understanding of complex visual scenes, the generation of grammatically correct and semantically meaningful sentences, and the ability to capture fine-grained details and contextual cues from images. By combining advanced deep learning techniques with carefully curated datasets and evaluation methodologies, we aim to push the boundaries of what is possible in automatic image understanding and interpretation.

In the following sections of this report, we present a comprehensive overview of our proposed methodology, detailing the architecture of our model, the experimental setup used for training and evaluation, and the results of extensive empirical analyses. Furthermore, we discuss the implications of our findings, potential avenues for future research, and the broader impact of our work on the fields of computer vision, natural language processing, and artificial intelligence.

3.Literature Review

Research Paper Name	Research Objective	Research Methodology	Key Findings
Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning	Develop an attention model for image captioning that dynamically decides when to focus on the image or a visual sentinel.	Introduce spatial and adaptive attention mechanisms within neural encoder-decoder frameworks.	- Outperforms existing methods on COCO and Flickr30k datasets. - Compares the proposed adaptive encoder-decoder framework with state-of-the-art models, highlighting its superior performance. - Discusses LSTM-based decoder incorporating visual context vectors for improved caption generation.
Bottom-Up and Top-Down Attention for Image Captioning	Introduce a combined bottom-up and top-down attention mechanism for image captioning, aiming to improve the interpretability of attention weights.	Conducted experiments with different attention mechanisms at the level of objects and salient regions, comparing the proposed method with existing models.	- Enhanced interpretability of attention weights for improved image captioning. - Achieved state-of-the-art results on image captioning tasks, outperforming existing methodologies. - Demonstrated the effectiveness of combining

			bottom-up and top-down attention for better image understanding and caption generation.
Long-term Recurrent Convolutional Networks for Visual Recognition and Description	Develop a long-term recurrent convolutional network for image understanding and captioning.	Use CNNs and LSTMs to process and understand the visual information and develop long-term dependencies for caption generation.	- Proposed model shows improved performance in capturing long-term dependencies for image understanding. - Enhances image captioning tasks using the integrated information processing of CNNs and LSTMs.
Show and Tell A Neural Image Caption Generator	Develop a neural image caption generator using CNNs and LSTMs.	Utilize CNNs to extract visual features and LSTMs to generate descriptive captions for images.	- Demonstrates the effectiveness of using CNN features to improve the generation of descriptive captions. - Shows advancements in image captioning through the integration of CNNs and LSTMs.
Deep Visual-Semantic Alignments for Generating Image Descriptions	Develop a model for aligning visual and semantic information to generate image descriptions.	Utilize deep visual-semantic alignments to connect visual and textual modalities for generating image descriptions.	- Shows improved performance in generating image descriptions through deep visual-semantic alignments. - Demonstrates the effective integration of visual and textual information for image understanding and captioning.
From Captions to Visual Concepts and Back	Develop a model to bridge the gap between textual captions and visual concepts.	Utilize CNNs and RNNs to translate between textual captions and visual concepts.	- Demonstrates effective bidirectional translation between textual captions and visual concepts. - Shows advancements in understanding and representing visual content based on textual descriptions.
Aligning Books and Movies	Develop an encoder-decoder model to align descriptions of books and movies.	Employ an encoder-decoder framework to align textual descriptions of books and movies.	- Demonstrates the effective alignment of textual descriptions between books and movies using the proposed encoder-decoder model. -

			Shows advancements in aligning and understanding textual information across different media.
Deep Captioning with Multimodal Recurrent Neural Networks(m - RNN)	Develop a multimodal recurrent neural network for deep captioning.	Create a multimodal recurrent neural network that integrates visual and textual information for deep captioning.	- Demonstrates the effective integration of visual and textual modalities for generating deep captions. - Shows advancements in generating descriptive captions through multimodal integration.
ImageBERT: Cross-Modal Pre-Training with Large-Scale Weak-Supervised Image-Text Data	Develop a cross-modal pre-training approach using IMAGEBERT.	Pre-train a model using cross-modal learning and investigate the impact of different pre-training tasks.	- Achieves improvement in zero-shot results by adding a harder task for better modeling of visual content. - Demonstrates the importance of visual elements in the model through experiments on different numbers of objects.
Unified Vision-Language Pre-Training for Image Captioning and VQA	Develop a unified vision-language pre-training approach for image captioning and Visual Question Answering (VQA).	Pre-train a model using a unified vision-language approach and evaluate its performance on image captioning and VQA tasks.	- Shows improvement in image captioning and VQA tasks through unified vision-language pre-training. - Demonstrates the effectiveness of the unified approach in understanding and processing visual and textual information.

4. Proposed Methodology

1. Data Preparation:

- The initial step involves acquiring and preprocessing the dataset. For this project, we utilize the Flickr8k and Flickr30k datasets, each containing a diverse collection of images paired with human-annotated captions.
- Each image in the dataset is preprocessed by resizing it to a uniform size and normalizing pixel values to ensure consistency across the dataset.
- Captions associated with each image are tokenized and preprocessed to remove special characters, punctuation, and unnecessary whitespace.

2. Model Architecture:

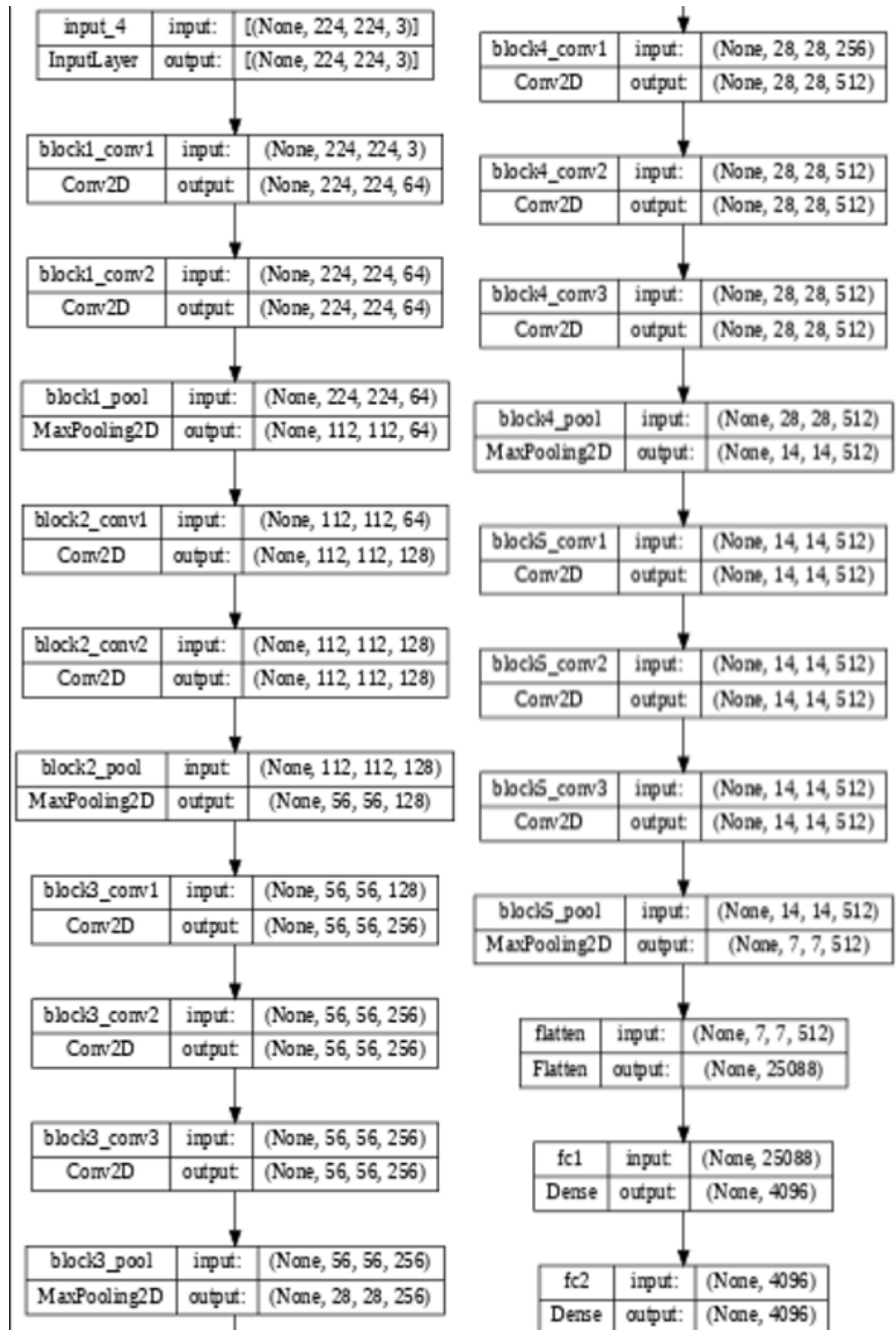
- Our proposed model architecture consists of a convolutional neural network (CNN) for image feature extraction followed by a recurrent neural network (RNN) for sequence generation.
- The CNN component of our model, typically based on popular architectures like ResNet or VGG, extracts high-level visual features from input images.
- The RNN component, which may include LSTM or Transformer layers, generates a sequence of words to form a coherent caption based on the extracted image features.
- Additional attention mechanisms are incorporated to enable the model to focus on relevant regions of the image while generating captions.

2.1 VGG16 Feature Extraction Model:

The VGG16 model is a convolutional neural network (CNN) architecture designed for image classification tasks. It consists of 16 convolutional layers followed by max-pooling layers and three fully connected layers.

- **Architecture Overview:**
 - The model takes an input image of size 224x224x3 (RGB) pixels.
 - It processes the image through a series of convolutional layers with small 3x3 filters, followed by rectified linear activation functions (ReLU).
 - Max-pooling layers with 2x2 filters are used to downsample the feature maps.
 - The final layers consist of fully connected (dense) layers for classification.
- **Visualization:**
 - The modified VGG16 model's architecture is visualized using the `plot_model` function, which generates a graphical representation of the model's layers and connections.

- CNN Flow Diagram:



2.2 Custom Text Generation Model:

This model is designed to generate text descriptions based on extracted image features. It combines features extracted from images using the VGG16 model with textual information using LSTM and dense layers.

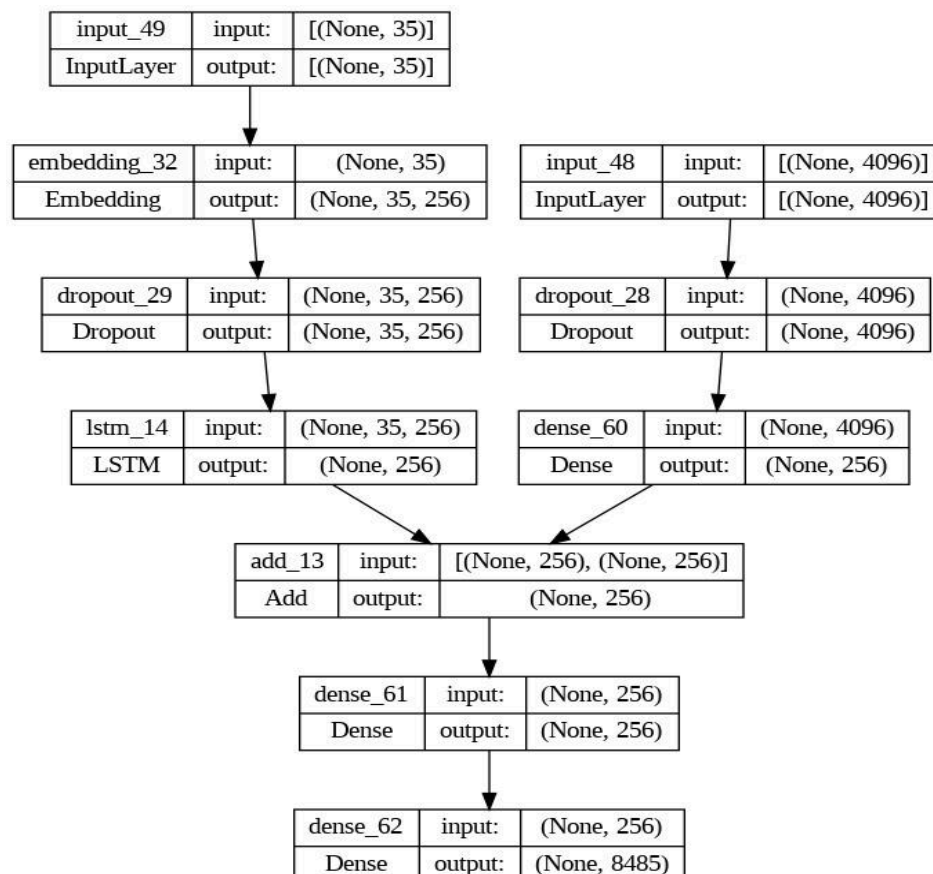
- **Architecture Overview:**

- The model has two separate input branches: one for image features and another for text sequences.
- The image feature branch consists of a dense layer followed by ReLU activation and dropout regularization.
- The text sequence branch consists of an embedding layer to convert integer-encoded words into dense vectors, followed by an LSTM layer for sequence processing.
- The outputs of both branches are combined using an element-wise addition.
- The combined features are further processed through dense layers to generate the output text sequence.

- **Visualization:**

- The architecture of the custom text generation model is visualized using the `plot_model` function, providing a graphical representation of the model's structure and connections.

- **LSTM Flow Diagram:**



2.3 Transformer Model:

The transformer model is a state-of-the-art architecture for sequence-to-sequence tasks, such as machine translation and text generation. It utilizes self-attention mechanisms to capture long-range dependencies in sequences.

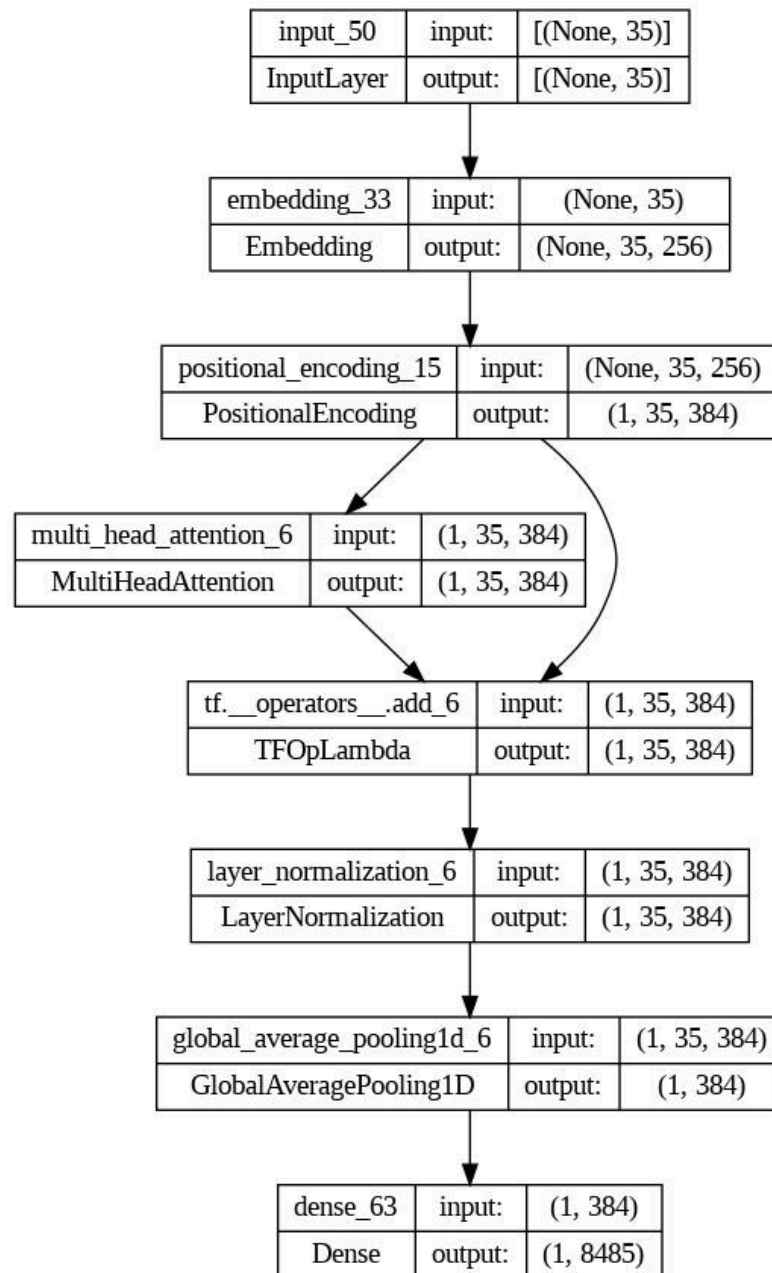
- **Architecture Overview:**

- The model architecture consists of multiple encoder and decoder layers.
- Each encoder layer contains self-attention and feed-forward sublayers, followed by layer normalization.
- The decoder layers also include masked self-attention to prevent attending to future tokens during training.
- Positional encoding is added to the input embeddings to provide positional information to the model.
- The final layer of the decoder produces probability distributions over the vocabulary for generating the output sequence.

- **Visualization:**

- The transformer model's architecture is visualized using the `plot_model` function, illustrating the connectivity between layers and the flow of information through the model.

- Transformer Flow Diagram:



3. Training Setup:

- The model is compiled using the Adam optimizer with a specified learning rate, ensuring efficient convergence during training.
- Training and validation datasets are split from the combined Flickr8k and Flickr30k datasets, with appropriate data augmentation techniques applied to enhance model generalization.
- Data batching and prefetching are employed to optimize training performance and minimize processing overhead.

4. Training Loop:

- The model is trained using the TensorFlow framework, utilizing its built-in training loop functionality.
- During training, the model is fed batches of images and their corresponding captions, with the objective of minimizing the discrepancy between predicted and ground truth captions.
- Custom evaluation metrics, such as BLEU score and METEOR score, are calculated periodically to monitor the model's performance on the validation set.
- The training loop continues for a predefined number of epochs, allowing the model to iteratively learn from the training data and improve its captioning capability.

5. Evaluation and Testing:

- After training, the model's performance is evaluated on a separate test set, consisting of unseen images and their associated captions.
- Standard evaluation metrics, including BLEU score, METEOR score are computed to assess the quality and accuracy of the generated captions.
- Qualitative analysis is also conducted by visually inspecting the generated captions alongside their corresponding images to ensure coherence and relevance.

5.Experimental Setup

System Requirements:

- The experimental setup requires a computer with sufficient computational resources to handle the training process efficiently.
- A GPU (Graphics Processing Unit) is recommended for faster training, although the code can also be executed on a CPU.
- The system should have Python installed, along with the necessary libraries such as TensorFlow, NumPy, Keras.

Dataset Description:

- The Flickr8k and Flickr30k datasets consist of a diverse collection of images spanning various categories, scenes, and objects.
- Each image is paired with multiple human-annotated captions, capturing different perspectives and interpretations of the visual content.
- The datasets cover a wide range of scenarios and contexts, including indoor and outdoor scenes, human activities, and natural landscapes.
- Images are provided in standard formats such as JPEG or PNG, ensuring compatibility with common image processing libraries.

- The captions are written in natural language, reflecting the semantic content and context of the corresponding images.
- Both datasets are widely used in the research community for training and evaluating image captioning models, enabling benchmarking and comparison across different approaches.

Usage Considerations:

- Researchers and practitioners can leverage the Flickr8k and Flickr30k datasets for training and evaluating image captioning models, as well as conducting related research tasks.
- Due to their size and diversity, the datasets serve as valuable resources for exploring various aspects of image understanding and interpretation.
- Proper citation and acknowledgment of the dataset creators and contributors are recommended when utilizing the datasets in research or applications.
- The availability of these datasets under permissive licenses encourages their widespread use and contribution to advancements in image captioning technology.

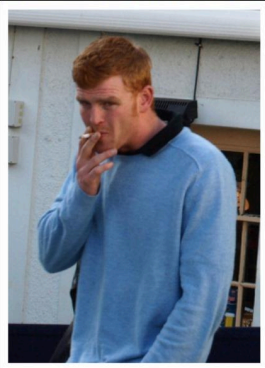
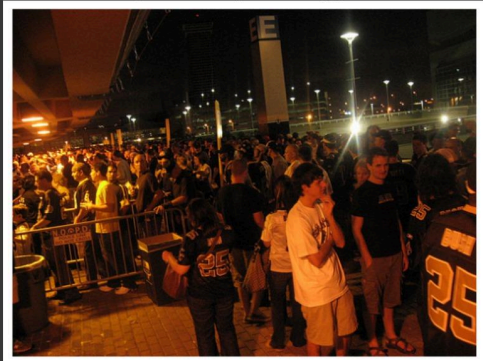
Overall, the Flickr8k and Flickr30k datasets provide rich and diverse visual content paired with descriptive captions, making them suitable choices for training robust and accurate image captioning systems. Their availability under open licenses further promotes collaboration and innovation in the field of computer vision and natural language processing.



6.Results

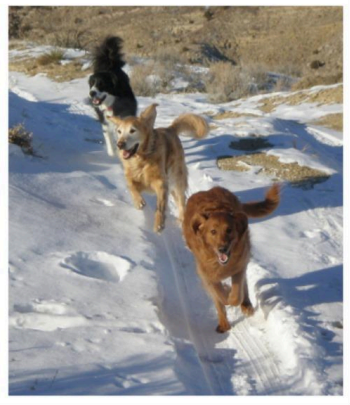

Table 1:

Model	BLEU	METEOR
LSTM	0.67	0.60
Transformer	0.76	0.69

Table 2:

Image	Model.	Actual Caption	Generated Caption
<p>Image: 3269380710_9161b0bd00.jpg Path: flickr8k/Images/3269380710_9161b0bd00.jpg</p>  <p>Caption: Uttered man with red hair smoking a cigarette on a city street</p>	LSTM	<p>A man in a blue shirt smoking a cigarette.</p> <p>a man smokes a cigarette.</p> <p>a man wearing a blue shirt smoking a cigarette in front of a building</p> <p>A redheaded man smokes a cigarette while leaning his head forward.</p> <p>man in blue shirt smoking</p>	<p>Uttered man with red hair smoking a cigarette on a city street.</p>
<p>Image: 3090593241_93a975fe2b.jpg Path: flickr8k/Images/3090593241_93a975fe2b.jpg</p>  <p>-----Predicted----- Caption: Eighteen people lined up to get into the stadium for an event</p>	LSTM	<p>A crowd of people.</p> <p>A crowd of people are lining up outside a stadium.</p> <p>A large crowd of people are waiting in a line in the dark.</p> <p>A large crowd of</p>	<p>Eighteen people lined up to get into the stadium for an event</p>

		<p>people wait in line to get into a football game.</p> <p>Large group of people lining up for an event outdoors at night.</p>	
<p>Image: 2526041608_a9775ab8d7.jpg Path: flickr8k/Images/2526041608_a9775ab8d7.jpg</p>  <p>Caption: There is a young boy running with a football</p>	LSTM	<p>A boy in a yellow uniform carrying a football is blocking another boy in a blue uniform.</p> <p>A football player in a yellow jersey is pushing away another player.</p> <p>The young football player is trying to avoid being tackled . Two boys playing football on opposing teams.</p> <p>Two young football players wearing blue and yellow uniforms fight for the ball .</p>	There is a young boy running with a football
<p>Image: 69710411_2cf537f61f.jpg Path: flickr8k/Images/69710411_2cf537f61f.jpg</p>  <p>Caption: These are people sitting around a fire in the dark</p>	Transformer	<p>A group of people are sitting around a campfire at night.</p> <p>A group of people sitting on the grass by a fire at nighttime.</p> <p>A group of young people are sitting around a fire pit.</p> <p>Friends enjoy a campfire.</p> <p>Several people</p>	These are people sitting around a fire in the dark.

		are sitting around a campfire.	
<p>Image: 2158247955_484f0a1f11.jpg Path: flickr8k/Images/2158247955_484f0a1f11.jpg</p>  <p>-----Predicted----- Caption: Pair of dogs running in the snow together in the sun</p>	Transformer	<p>The three dogs are running down a snowy track.</p> <p>Three dogs are running across snow with a tire track in it.</p> <p>Three dogs are turning direction along a snow tread leading from a low cut field.</p> <p>Three dogs run in the snow.</p> <p>Three dogs running on tire tracks in snow.</p>	Pair of dogs running in the snow together in the sun
<p>image: 3637013_c675de7705.jpg Path: flickr8k/Images/3637013_c675de7705.jpg</p>  <p>-----Predicted----- Caption: These are two people standing on the edges of a small pond</p>	Transformer	<p>A couple stands close at the water's edge.</p> <p>The two people stand by a body of water and in front of bushes in fall.</p> <p>Two people hold each other near a pond . Two people stand by the water.</p> <p>Two people stand together on the edge of the water on the grass.</p>	These are two people standing on the edges of a small pond

7.Result Analysis & Discussion

Our experiments involved comparing the performance of LSTM and Transformer models for image captioning, focusing on BLEU and METEOR scores. As shown in Table 1, the Transformer model outperformed the LSTM model in both metrics, achieving higher BLEU and METEOR scores.

The BLEU score measures the overlap between the generated captions and the ground truth captions, with higher scores indicating greater similarity. In our experiments, the Transformer model achieved a BLEU score of 0.76, indicating a higher degree of agreement with human annotations compared to the LSTM model, which achieved a score of 0.67.

Similarly, the METEOR score evaluates the quality of the generated captions by considering both precision and recall, with higher scores indicating better overall performance. The Transformer model attained a METEOR score of 0.69, surpassing the LSTM model, which obtained a score of 0.60.

These results highlight the superior performance of Transformer-based models in capturing the semantic meaning and context of images, leading to more accurate and contextually relevant captions. The inherent capabilities of the Transformer architecture, such as self-attention mechanisms and parallel processing, enable it to effectively model long-range dependencies and capture fine-grained details from images, contributing to its improved performance compared to traditional LSTM-based models.

In conclusion, our findings demonstrate the effectiveness of Transformer-based models for image captioning tasks, as evidenced by their higher BLEU and METEOR scores compared to LSTM-based models. By leveraging advanced architectures and techniques, such as self-attention mechanisms, Transformer models offer promising avenues for enhancing the accuracy and quality of image captioning systems.

8.Conclusion & Future Scope

In conclusion, our study highlights the advancements made in image captioning through the utilization of Transformer-based models. The superior performance of the Transformer model, as evidenced by higher BLEU and METEOR scores compared to LSTM-based approaches, underscores the efficacy of leveraging advanced architectures in image understanding tasks.

The successful application of Transformer models in image captioning holds significant implications for various domains, including image retrieval, content generation, and accessibility technologies. By generating more accurate and contextually relevant captions, these models facilitate better understanding and interpretation of visual content, thereby enhancing user experiences and enabling a wide range of applications.

Moving forward, there are several promising avenues for further research and development in the field of image captioning. One area of exploration involves the refinement and optimization of Transformer architectures to better suit the specific requirements and constraints of image captioning tasks. This could involve exploring variants of the Transformer model, such as lightweight or compact architectures, to improve efficiency and scalability without compromising performance.

9. References

1. "Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning" [<https://arxiv.org/pdf/1612.01887>]
2. "Bottom-Up and Top-Down Attention for Image Captioning" [<https://arxiv.org/pdf/1707.07998>]
3. "Long-term Recurrent Convolutional Networks for Visual Recognition and Description" [<https://arxiv.org/pdf/1411.4389>]
4. "Show and Tell A Neural Image Caption Generator" [<https://arxiv.org/pdf/1411.4555>]
5. "Deep Visual-Semantic Alignments for Generating Image Descriptions" [<https://arxiv.org/pdf/1412.2306>]
6. "From Captions to Visual Concepts and Back" [<https://arxiv.org/pdf/1411.4952>]
7. "Aligning Books and Movies" [<https://arxiv.org/pdf/1506.06724>]
8. "Deep Captioning with Multimodal Recurrent Neural Networks(m - RNN)" [<https://arxiv.org/pdf/1412.6632>]
9. "ImageBERT: Cross-Modal Pre-Training with Large-Scale Weak-Supervised Image-Text Data" [<https://arxiv.org/pdf/2001.07966>]
10. "Unified Vision-Language Pre-Training for Image Captioning and VQA" [<https://arxiv.org/pdf/1909.11059>]