Image Caption Generator

Rohit Dhaipule
115848739
rdhaipule@cs.stonybrook.edu

Nithin Katla
115788093
nkatla@cs.stonybrook.edu

Nitin Gopala Krishna Sontineni
115743470
nsontineni@cs.stonybrook.edu

Vishnutej
115934115
vmuddasani@cs.stonybrook.edu

***Abstract*— Our project tries to solve an image captioning problem, bridging computer vision and natural language processing. This model accommodates an extensive array of input images, from diverse subjects like people in various activities to animals, objects, and more. The primary objective of this project is to assess the model's performance between datasets like Flickr8k and Flickr30k. Utilizing a comprehensive dataset, we utilize the VGG Convolutional Neural Network for image feature extraction and LSTM for encoding the generated caption. We, then use a decoder network to generate the next word of the caption. We found that model trained on Flickr30k dataset generated better captions than that of Flickr8k**

I. Introduction

In an era marked by the rapid advancement of technology, the field of computer vision has witnessed a remarkable transformation. One of the most intriguing challenges within this domain is the problem of image captioning. Being able to describe the contents of an image using well-formed English sentences is a huge challenge, but it could make a big difference. For example, by helping people with visual impairments understand the contents of images on the internet. This task is much more difficult than, say, the well-documented image classifications or object recognition tasks that have been at the forefront of computer vision. A description must capture more than just the objects in an image. It must describe how those objects interact with each other, their attributes, and the activities they engage in.

Traditional image captioning models typically use a convolutional neural network (CNN) to extract features from the image, followed by a recurrent neural network (RNN) to generate the caption. CNNs are well-suited for extracting visual features from images, while RNNs are able to model the sequential nature of language. However, these models have several limitations. First, they are often trained on small datasets, which can lead to overfitting. Second, they can struggle to generate accurate and informative captions for complex images. Third, they may generate captions that are grammatically incorrect or semantically inconsistent.

In the past decade, the application of Long Short-Term Memory (LSTM) for image captioning has garnered increasing attention in the field. LSTM, as a standalone solution, is utilized for the crucial task of generating captions, offering the advantage of tailored development from scratch without relying on pre-trained models. Our chosen approach involves leveraging the VGG16 model specifically for extracting visual features from images. VGG16, known for its effectiveness in image processing, contributes by capturing intricate details and representations of the visual content. Simultaneously, the LSTM model is employed to handle the sequential and contextual aspects of language generation. By combining these two components, our methodology aims to create a holistic image captioning model that excels in both visual understanding and language contextualization.

To assess the effectiveness of our proposed model, we conduct a comprehensive evaluation across diverse datasets, including popular benchmarks like Flickr8k and Flickr30k. This evaluation allows us to thoroughly analyze and understand the model's performance under various data scenarios, providing valuable insights.

II. Literature Review

Numerous scientific endeavors have delved into the analysis of visual data, particularly within the realm of videos, employing intricate systems. Traditionally, these methods heavily leaned on object recognition, often resorting to templates or rule-based approaches for text generation to describe images. However, a notable drawback of these methods was their inherent lack of flexibility and tendency towards rigidity.

In response to these limitations, innovative approaches surfaced, combining deep convolutional nets for image classification with recurrent networks for sequence modeling. This convergence resulted in unified networks capable of generating descriptive captions for images. Notably, this departure from rigid rule-based systems introduced a newfound flexibility [4].

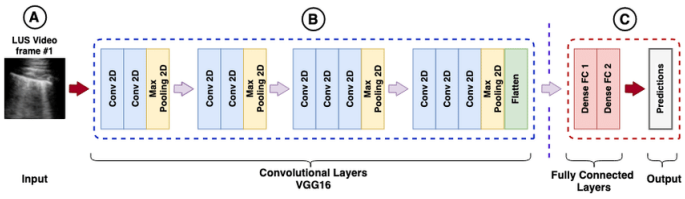

Fig. 1. Features of image extracted using VGG16

Furthermore, certain methodologies evolved to address the challenge of producing segmentation masks based on complex and implicit textual queries. To facilitate the evaluation of these evolving methods, a benchmark comprising over a thousand image-instruction pairs was established. These pairs intricately integrated reasoning and leveraged world knowledge, providing a robust testing ground for the effectiveness of these novel approaches [5].

In the context of feature extraction from images, VGG16, or Visual Geometry Group 16, emerged as a prominent convolutional neural network (CNN) architecture. Characterized

Gray haired man in black suit and yellow tie working in a financial environment.
A graying man in a suit is perplexed at a business meeting.
A businessman in a yellow tie gives a frustrated look.
A man in a yellow tie is rubbing the back of his neck.
A man with a yellow tie looks concerned.

A butcher cutting an animal to sell.
A green-shirted man with a butcher's apron uses a knife to carve out the hanging carcass of a cow.
A man at work, butchering a cow.
A man in a green t-shirt and long tan apron hacks apart the carcass of a cow while another man hoses away the blood.
Two men work in a butcher shop; one cuts the meat from a butchered cow, while the other hoses the floor.

Fig. 2. Two images from Flickr data sets and their five captions

by its straightforward design philosophy, VGG16 employs small 3x3 convolutional filters throughout its 16 layers, including 13 convolutional and three fully connected layers. This design contributes to its deep yet easily interpretable structure, enabling the model to learn hierarchical features of increasing complexity. Leveraged widely as a feature extractor in transfer learning scenarios, VGG16's pre-trained weights on large image datasets facilitate the extraction of meaningful features from images.

In the pursuit of understanding the impact of training data on image captioning models, our research takes a distinctive approach by focusing on the utilization of segmentation models for feature extraction. This strategy involves breaking down an image into distinct segments, allowing the model to concentrate on specific regions of interest and enhancing its ability to comprehend intricate details within complex visual scenes. Complementing this, our research integrates LSTM for caption generation, acknowledging the instrumental role LSTM networks play in seamlessly collaborating with CNNs. LSTM's architecture, equipped with memory cells and gates (input, output, and forget), effectively addresses challenges like the vanishing gradient problem encountered in training deep networks over long sequences.

The proposed literature review will extend its scrutiny to existing studies on diverse datasets, probing their impact on a model's capacity to comprehend and articulate visual scenes. A particular focus will be placed on establishing the correlation between dataset size and the quality of captions generated. Our methodology outlines a comprehensive fine-tuning process for both segmentation and language models, incorporating diverse datasets for experimentation. Rigorous evaluation metrics will be employed to assess the performance of the model in both image segmentation and caption generation. To navigate challenges such as overfitting and ethical considerations tied to dataset choices, a holistic approach will be taken.

III. Datasets

Flickr8k and Flickr30k are two popular datasets for image captioning. They are both subsets of the Flickr image hosting website, and they contain images with multiple human-written captions. Some examples are given in Fig.2.

1) ***Flickr8k*** *[1]:* This dataset consists of 8,000 images extracted from Flickr, each paired with five different captions. Each caption describes the content of the image, providing a textual description for machine learning algorithms to learn from. The grammar of the annotations for this dataset is simpler and the images cover a wide range of everyday scenes and objects. We adopt the standard separation of training, validation and testing set which is provided by the dataset. There are 6,000 images for training, 1,000 images for validation and 1,000 images for testing.

2) ***Flickr30k*** *[2]:* The Flickr30k dataset is an extension of the Flickr8k dataset and contains 30,000 images. Like the Flickr8k dataset, each image is paired with 5 human-generated captions. Unlike the Flickr8k dataset, which focuses on relatively simple scenes, the Flickr30k dataset includes more complex and diverse images with a broader range of objects, scenes, and activities. Here, we plan to use 28,000 images for training, 1,000 images for validation and 1,000 images for testing.

Both Flickr8k and Flickr30k are widely used in image captioning research and development. They are also used in other computer vision tasks, such as object detection, scene segmentation, and image retrieval.

The statistics of the datasets are as follows:

Dataset Name / Size	Train	Validation	Test
Flickr8k	7281	834	810
Flickr30k	28604	3472	3179

IV. Proposed Approach

A. Data Collection and Preparation

The success of generating suitable captions for input images relies heavily on the quality and suitability of the data used in the training process. In this section, we describe the data cleaning and preparation procedures to be followed for our project on generating captions for images. These procedures are crucial for ensuring the reliability and performance of our

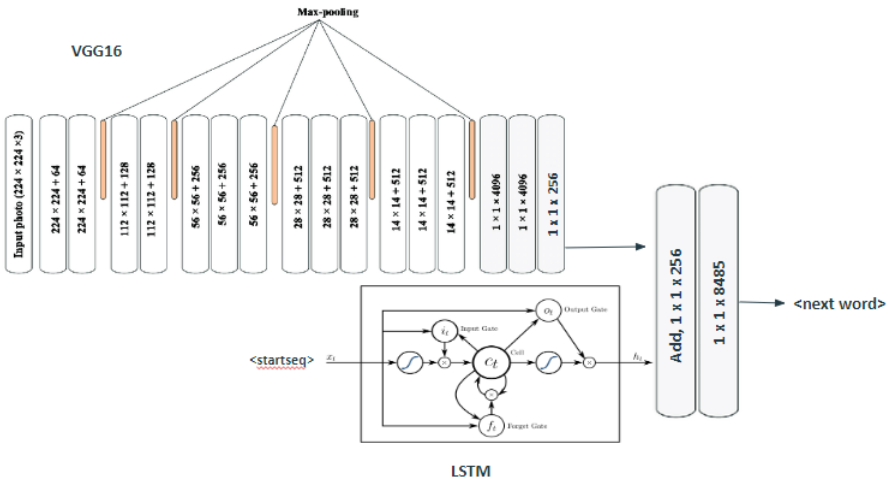

Fig. 3. Model Architecture

proposed model. We start with identifying sources of image and caption datasets. Common sources include image-sharing websites, such as Flickr, specialised datasets like MSCOCO. It is essential to ensure that the dataset is representative of the target application domain. Image preprocessing involves several steps, including resizing and normalizing images to create a consistent and standardized input for our model. Caption preprocessing involves removal of extra spaces, special characters and adding the startseq and endseq tags to denote the start and end of the caption.

B. Proposed Model

In this section, we present the architecture and methodology of our proposed model for generating capions for input images (See Fig. 3). Our approach leverages the strengths of both CNN and LSTM models to achieve a good captioning performance. The model consists of two key components: image feature extraction using VGG-16 model and caption generation using an LSTM model.

At the outset, an pre-processed input image is presented to the VGG16, which serves as the model's feature extractor. VGG16 analyzes the image, extracting hierarchical features that encapsulate spatial patterns and contextual information. This process forms a comprehensive representation of the visual content within the image, laying the foundation for subsequent stages.

Following feature extraction, the model transitions to the LSTM network for sequence encoding. The encoded image features obtained from VGG16 are used to initialize the hidden state and cell state of the LSTM. The LSTM, functioning as a sequence-to-sequence model, takes the generated caption sequence as input. This sequence is typically initialized with a start token and iteratively extended until an end token is generated or a predefined maximum caption length is reached.

At each step of the sequential generation process, the LSTM utilizes its hidden state and cell state from the previous step, along with the embedded representation of the current word in the sequence, to predict the next word. The embedding of each word is learned during the training process and represents the continuous vector space for words, aiding in capturing semantic relationships.

The decoder, then takes facilitates the generation of the caption by predicting one word at a time based on the context provided by the LSTM's hidden state and cell state. The sequential generation process continues until an end token is generated, signaling the completion of the caption. Its role is pivotal in predicting each subsequent word based on the context provided by both the LSTM and the initial visual features obtained from VGG16. This collaborative input allows the model to leverage both the spatial hierarchies captured by VGG16 and the sequential dependencies encoded by the LSTM, resulting in a more nuanced and contextually rich generation process.

Throughout the training process, the model is optimized to minimize the difference between the generated captions and ground truth captions using appropriate loss functions. This involves backpropagation and gradient descent to update the parameters of CNN, LSTM and decoder.

We train the model separately on all the data sets described in Section III and evaluate performance of the model across each dataset to understand how the model behaves with change in data.

C. Loss Function and Optimisation Algorithm

In our image captioning model, we employed the categorical cross-entropy loss function and ADAM optimizer. Cross-entropy loss is apt for this task as it measures the dissimilarity between predicted and actual probability distributions for multi-class classification, aligning with the nature of generating tokens in captioning. This loss function penalizes deviations from the ground truth, encouraging the model to produce more accurate captions.

The Adam optimizer, chosen for its efficiency in handling sparse gradients and adjusting learning rates for individual

parameters, aids in faster convergence. Its adaptive learning rate and momentum features enhance training stability and performance. A batch size of 32 ensures efficient memory utilization during training, and the selected learning rate of 0.001 strikes a balance between rapid convergence and avoids overshooting the optimal solution.

V. Results

A. Metrics Used

To gauge the quality and effectiveness of our proposed model for image caption generation, we used BLEU. This allow us to quantitatively assess the model's performance and provide a comprehensive understanding of its strengths and weaknesses.

BLEU score quantifies the quality of generated text by comparing it to reference (human-generated) text. BLEU measures the overlap of n-grams (contiguous sequences of n words) between the generated and reference text, providing a score between 0 and 1, where a higher score indicates better performance. BLEU is valued for its simplicity and interpretability, though it may not capture all aspects of text quality.

B. Discussion

In this section, we compared the performance achieved by the model using different datasets. As described in Section 3, the datasets were split into training, validation, and test sets for evaluations purposes. To assess the effect of training data size on model performance, we training the model with both flickr8k and flickr30k datasets. After training, the best-performing checkpoint was selected based on its performance on the validation set and then evaluated on the test set. The resulting score on the test set are as follows:

Dataset	***BLEU-1***	***BLEU-2***	***BLEU-3***	***BLEU-4***
Flickr8k	0.525075	0.302179	0.188640	0.111971
Flickr30k	0.5832	0.3314	0.1895	0.1216

TABLE I

BLEU Score is between 0 to 1. High BLEU score indicates that the generated captions closely match the reference captions, showing better quality. Conversely, a low BLEU score suggests a significant mismatch, indicating poorer quality and less accurate caption generation.

As illustrated in Table I, the model reported the highest scores on the Flickr30k dataset. These superior results on a larger dataset can be attributed to the richness of data, offering a more comprehensive representation of diverse scenes, complexities, and their inherent contextual nuances.

1) Color Detection Ability: Both models exhibited impressive accuracy in identifying objects with consistent colors, highlighting their proficiency in learning from uniform input across all images. However, challenges surfaced for Flickr8k model when objects exhibited variability in color, leading to a decline in performance. For instance, Flickr8k model struggled to adapt to changing shirt color in the below image. In Fig.V-B1, we can see that the caption generated by the Flickr8k model includes references to a pink shirt and black pants, despite the absence of these elements in the visual content. In contrast, the caption produced by the Flickr30k model appears to be more accurate and aligned with the actual image content.

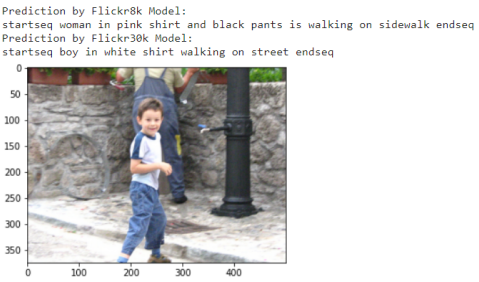

Fig. 4. Color Recognition

2) Ambiguity in Action Recognition: Flickr8k Model struggled with images depicting multiple actions, revealing a limitation in captioning complex scenes (See Fig 5). A larger and more diverse dataset like Flickr30k is crucial for training the model to better understand and describe dynamic visual scenarios, improving overall performance and generalization.

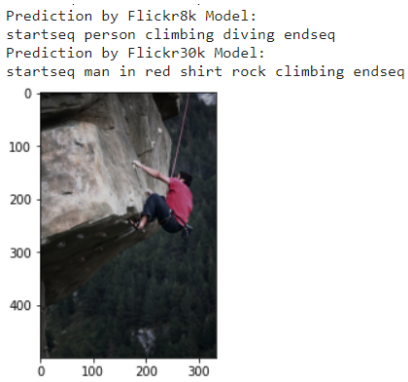

Fig. 5. Mismatch of action performed

3) Information Omission: The observed phenomenon of information omission in certain images by the Flickr8k model highlights a noteworthy limitation in the model's contextual understanding, contributing to a diminished capacity to provide accurate captions. One such example is the model's failure to recognize the presence of bikes (See Fig. 6, underscoring the necessity for a more nuanced comprehension of visual content.

4) Caption length analysis: From the analysis in the previous section, there is information missing from the captions in case of Flickr8k data set. We analyzed the length of the input and output captions in Flickr8k and Flickr30k for the same reason. Our aim is to analyze if the larger information capture in output caption is because of larger average caption size in Flickr30k or because of better training data. The analysis unveiled a low positive correlation (Pearson Correlation Coef-

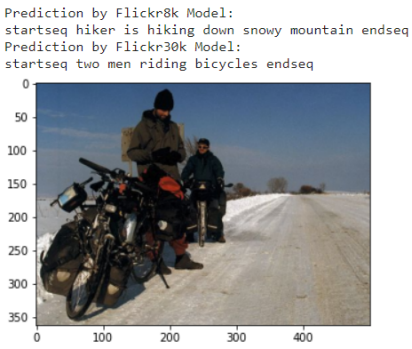

Fig. 6. Failure to identify the bikes

ficient: 0.0914) between the lengths of input captions and the predicted captions. Also, a low p-value of 0.0092 is observed.

In Fig. 7. we plotted a histogram of number of words in each caption of input and output in test data set for both Flickr8k and Flickr30k Data sets. The average number of output words in a caption is longer in case of Flickr30k. As the Correlation Coefficient is low, this means that in the context of Flickr30k, the model gets better trained on various scenarios and leverage the long-distance memory capabilities of LSTM to generate captions that are not only longer but also accurate.

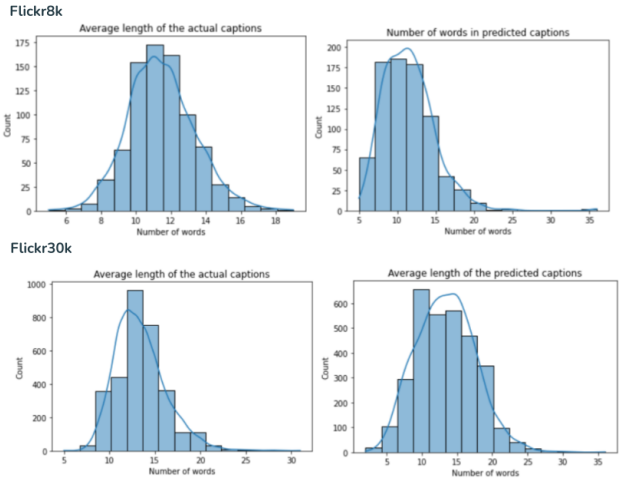

Fig. 7. Histogram representing lengths of captions

C. Challenges Faced

Initially, we considered leveraging pretrained models such as LISA and BERT to address the challenges posed by image captioning. However, we opted for the existing model due to several challenges that arose during the planning phase.

One significant challenge pertained to memory constraints, specifically limitations within GPU memory during model training. This constraint became particularly pronounced when dealing with extensive datasets or a high number of training parameters with an estimated training time of 35hrs for each epoch. The sheer size of the data or the complexity of the model parameters often surpassed the available GPU memory, making it impractical to proceed with the initially considered pretrained models.

VI. Future Research Directions

In the future, we can dig deeper into making image captions even better. One way is by focusing on capturing emotions in images, so captions can reflect feelings accurately. Additionally, exploring how captions can adapt to different cultural contexts will be important for more inclusive and diverse results. Investigating ways to handle images with multiple objects or complex scenes can improve the accuracy of captions in busy pictures. Moreover, we can use attention based models like transformers which capture the dependencies of multiple parts of the images better. These steps ensure a more modern approach and improvement of our image captioning model's performance across diverse situations.

VII. Conclusion

In conclusion, our project on image captioning has provided valuable insights into the intricate relationship between diverse datasets and the performance of our caption generation model. Through a meticulous analysis of datasets, notably Flickr8k and Flickr30k, we have acquired a nuanced understanding of the model's adaptability and generalization capabilities across varied content. The effective integration of the VGG16 for image feature extraction, coupled with the LSTM model for caption generation, has demonstrated a robust synergy. Not only has our exploration successfully addressed the challenges inherent in diverse dataset characteristics, but it has also laid a solid foundation for future advancements in the field. Moving forward, the incorporation of even more diverse datasets, fine-tuning for specific domains, and a sustained commitment to ethical considerations will undoubtedly propel the evolution of image captioning models. This project marks a meaningful contribution to the continuous refinement and responsible development of computer vision and natural language processing applications.

References

[1] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. In NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 139–147, 2010.

[2] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In ACL, 2014

[3] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft coco: Common objects in context. arXiv:1405.0312, 2014.

[4] Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and tell: A neural image caption generator." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156-3164. 2015.

[5] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu and Jiaya Jia. LISA: Reasoning Segmentation via Large Language Model." arXiv:2308.00692, 2023.