**IMAGE CAPTIONING SYSTEM**

**A PROJECT REPORT**

**for**

**Introduction To AI (AI101B)**

**Session (2024-25)**

**Submitted by**

**Priyanshu Asthana**

**202410116100154**

**Sakshi Khandelwal**

**202410116100175**

**Sanya Dwivedi**

**202410116100184**

**Submitted in partial fulfilment of the**

**Requirements for the Degree of**

**MASTER OF COMPUTER APPLICATION**

**Under the Supervision of**

**Ms. Komal Salgotra**

### Assistant Professor



**Submitted to**

**Department Of Computer Applications**

**KIET Group of Institutions, Ghaziabad**

**Uttar Pradesh-201206**

**(Month 2025)**

**CERTIFICATE**

Certified that **Priyanshu Asthana 202410116100154, Sakshi Khandelwal 202410116100175, Sanya Dwivedi 202410116100184** has/ have carried out the project work having “**IMAGE CAPTIONING SYSTEM**” (INTRODUCTION TO AI) (AI101B)) for **Master of Computer Application** from Dr. A.P.J. Abdul Kalam Technical University (AKTU**)** (formerly UPTU), Lucknow under my supervision. The project report embodies original work, and studies are carried out by the student himself/herself and the contents of the project report do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

**Ms. Komal Salgotra Dr. Akash Rajak**

**Assistant Professor Dean**

**Department of Computer Applications Department of Computer Applications**

**KIET Group of Institutions, Ghaziabad KIET Group of Institutions, Ghaziabad**

**An Autonomous Institutions An Autonomous Institutions**

**ABSTRACT**

This project explores the development of an image captioning system that automatically generates natural language descriptions for images. The system integrates advanced techniques from computer vision and natural language processing (NLP) to enable machines to interpret visual content and generate corresponding textual descriptions. The primary focus of the project is to leverage the Flickr8k dataset, which contains a variety of images along with human-generated captions, and apply deep learning models like Convolutional Neural Networks (CNN) for visual feature extraction and Long Short-Term Memory (LSTM) networks for sequence generation. The model uses a combination of the InceptionV3 architecture for feature extraction and an LSTM for generating coherent captions based on the extracted features.

**ACKNOWLEDGEMENTS**

Success in life is never attained single-handedly. My deepest gratitude goes to my project supervisor, **Komal Salgotra** for his/ her guidance, help, and encouragement throughout my project work. Their enlightening ideas, comments, and suggestions.

Words are not enough to express my gratitude to Dr. Akash Rajak, Professor and Dean, Department of Computer Applications, for his insightful comments and administrative help on various occasions.

Fortunately, I have many understanding friends, who have helped me a lot on many critical conditions.

Finally, my sincere thanks go to my family members and all those who have directly and indirectly provided me with moral support and other kind of help. Without their support, completion of this work would not have been possible in time. They keep my life filled with enjoyment and happiness.

**Priyanshu Asthana**

**Sakshi Khandelwal**

**Sanya Dwivedi**

**TABLE OF CONTENTS**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Certificate i. | |  |
|  | Abstract ii. | |  |
|  | Acknowledgements iii. | |  |
|  | Table of Contents iv. | |  |
|  | Introduction 5 | |  |
|  | Objective 6-7 | |  |
|  | Tools and Technologies used 8 | |  |
|  | Dataset Description 9-11  Methodology 12-13  Evaluation 13-15  Results 16  Challenges Faced 17  Conclusion 18-19  References 20 | |  |
|  |  |  |  |
|  |  |  |  |
|  |  | |  |
|  |  | |  |

**INTRODUCTION**

Image captioning is a sophisticated and interdisciplinary domain at the intersection of **computer vision** and **natural language processing (NLP)**. Its primary goal is to develop systems that can **interpret visual content** and express it in **coherent natural language**. This requires a deep understanding of both visual elements (objects, scenes, actions) and the ability to convert this understanding into grammatically and semantically correct sentences.

**Importance of Image Captioning:**

Image captioning has emerged as a vital area of research due to its vast potential in various domains:

* **Search Engines:** By generating descriptive tags and summaries, search engines can improve image retrieval accuracy.
* **Accessibility:** Visually impaired users benefit from screen readers that can narrate image content using generated captions.
* **Social Media & Content Creation:** Platforms can auto-caption images, improving engagement and accessibility.
* **Surveillance & Law Enforcement:** Helps in automated reporting and documentation from surveillance footage.
* **Medical Imaging:** Can assist in summarizing X-rays or scans, aiding in preliminary diagnostics.

This report explores the development and implementation of an image captioning system using **deep learning**, focusing on the integration of a **CNN.**

**OBJECTIVE**

The primary objective of this project is to design, develop, and evaluate an **image captioning system** that can generate accurate, fluent, and contextually appropriate textual descriptions for images. This involves bridging the gap between **visual perception** and **linguistic expression**, leveraging the strengths of deep learning architectures in both computer vision and natural language processing domains.

The system is built and trained using the **Flickr8k dataset**, which provides a diverse set of images and multiple human-generated captions per image. The combination of a **Convolutional Neural Network (CNN)** for extracting visual features and a **Recurrent Neural Network (RNN)**—specifically Long Short-Term Memory (LSTM)—for generating language allows the model to process complex visual data and translate it into meaningful language output.

**Specific Goals of the Project:**

1. **Develop an End-to-End Image Captioning Pipeline:**  
   Create a complete workflow starting from data preprocessing, feature extraction, model training, to caption generation and evaluation.
2. **Fusion of Vision and Language Models:**  
   Implement an architecture that effectively combines CNN-based visual feature extraction with RNN-based text generation to produce human-like captions.
3. **Learn Semantics and Syntax from Captions:**  
   Enable the model to learn how to map visual elements (such as objects, actions, and settings) to structured linguistic outputs, capturing both semantic meaning and grammatical correctness.
4. **Handle Sequence Generation:**  
   Design a system capable of generating full sentences word-by-word, predicting each word based on both the image and the preceding words in the caption.
5. **Utilize Pre-trained Models for Transfer Learning:**  
   Integrate a pre-trained InceptionV3 model to extract deep and rich visual features, thereby reducing the need for large amounts of training data and accelerating convergence.
6. **Evaluate Caption Quality with Quantitative Metrics:**  
   Use standard metrics such as **BLEU score** to quantitatively assess the accuracy and relevance of generated captions compared to reference human annotations.
7. **Address Challenges in Data and Model Performance:**  
   Explore methods to handle challenges such as limited vocabulary size, rare word occurrence, overfitting on small datasets, and computational resource constraints.
8. **Lay Foundation for Future Research:**  
   Build a baseline model that can be expanded upon with more advanced techniques such as attention mechanisms, transformers, and multilingual captioning in future iterations.

**TOOLS AND TECHNOLOGIES USED**

| **Tool/Technology** | **Purpose** |
| --- | --- |
| **Python** | Scripting and development language |
| **TensorFlow/Keras** | Deep learning frameworks used to build and train the CNN-LSTM model |
| **Google Colab** | Cloud environment providing GPU acceleration |
| **Flickr8k Dataset** | Dataset with 8,000 images and five captions each |
| **InceptionV3** | Pretrained CNN for extracting deep image features |
| **LSTM** | RNN variant used for generating textual sequences |
| **Tokenizer** | Preprocessing tool for converting words into integers |

**DATASET DESCRIPTION**

The **Flickr8k dataset** is a widely used benchmark dataset for image captioning and visual description generation tasks. It serves as an ideal starting point for developing and evaluating deep learning models that aim to bridge the gap between computer vision and natural language processing.

**Overview:**

The Flickr8k dataset consists of **8,000 natural images** collected from the Flickr website. Each image in the dataset is accompanied by **five unique human-generated captions**, making a total of **40,000 captioned sentences**. These captions provide a rich linguistic context describing the visual elements, actions, and scenarios represented in the images. The captions vary in phrasing and level of detail, allowing the model to learn how to generate flexible and diverse descriptions.

**Data Splits:**

To ensure a robust evaluation process and to avoid overfitting, the dataset is divided into three distinct subsets:

* **Training Set**: 6,000 images used for training the model and learning the visual-linguistic mappings.
* **Validation Set**: 1,000 images used during training to tune hyperparameters and avoid overfitting.
* **Test Set**: 1,000 images used for final evaluation and testing the performance of the trained model on unseen data.

Each split ensures that images are not shared across different sets, maintaining the integrity of the evaluation process.

**Image Content:**

The images in the dataset feature a wide range of real-world scenes, including:

* People engaged in various activities (e.g., playing sports, riding bicycles, walking dogs)
* Animals (dogs, cats, horses) in natural and urban settings
* Everyday objects such as cars, food items, buildings, and playground equipment
* Group and individual shots, providing diversity in terms of subject focus and background complexity

This visual variety helps train the model to generalize across different object categories and action types.

**Caption Characteristics:**

Each image is associated with five captions that:

* Are written by different annotators, providing multiple perspectives and phrasing variations.
* Typically range from **5 to 15 words**, capturing concise yet informative descriptions.
* Include both **simple subject-verb-object structures** ("A man is riding a bike") and more complex linguistic constructions ("Children are playing together in a field under the sun").

The captions are stored in a text file named Flickr8k.token.txt, formatted as image identifiers followed by their respective captions. This file is essential for linking images to their descriptions during preprocessing.

**Preprocessing of Captions:**

Before feeding the captions into the model, several preprocessing steps are applied:

* **Lowercasing** all text to maintain consistency.
* **Removing punctuation and special characters** to reduce noise.
* **Adding special tokens** such as 'startseq' and 'endseq' to mark the beginning and end of each caption. These tokens help the LSTM model learn when to start and stop the sentence generation process.
* **Tokenizing and indexing**: All words are converted into unique integer values using a tokenizer, and a vocabulary is built from the entire corpus.

**Relevance to the Project:**

The Flickr8k dataset is particularly well-suited for this project due to its manageable size, high-quality annotations, and diversity of visual and textual content. Although relatively small compared to datasets like MS-COCO, it is sufficient for building a proof-of-concept image captioning system. Its human-generated captions serve as a strong reference for training the model to produce natural and meaningful language descriptions based on image content.

The dataset's well-organized structure and extensive use in academic research also make it a valuable benchmark for comparing the performance of different image captioning architectures and evaluating the effectiveness of various enhancements.

**METHODOLOGY**

The methodology for building the image captioning system follows a structured deep learning pipeline:

**Step 1: Data Preprocessing**

* Cleaning captions: Lowercasing, punctuation removal, filtering non-alphabetic tokens.
* Tokenization: Each word is converted to an integer using Keras' Tokenizer.
* Vocabulary filtering: Words appearing less frequently can be filtered to reduce complexity.
* Sequence formatting: Each caption is framed between 'startseq' and 'endseq'.

**Step 2: Image Feature Extraction**

* Images resized to **299x299 pixels** for InceptionV3 compatibility.
* InceptionV3 used with the top layer removed to output **feature vectors**.
* Features stored in a dictionary mapping image names to 2048-dimensional vectors.

**Step 3: Text Preprocessing**

* Tokenized captions are padded to a **fixed length** for consistent input size.
* Vocabulary size is determined to define the output size of the softmax layer.
* Caption sequences are split into input-output pairs for training.

**Step 4: Model Architecture**

* **CNN (InceptionV3)** → Dense layer → Feature vector.
* **Caption sequence** → Embedding layer → LSTM.
* The two streams are **merged** using a concatenation layer.
* Output: Softmax layer over vocabulary to predict next word in the sequence.

**Step 5: Training**

* **Teacher Forcing**: Feeding correct previous word during training improves learning.
* **Loss Function**: Categorical crossentropy for multi-class classification.
* **Optimizer**: Adam optimizer for adaptive learning rate and faster convergence.
* **Batching and Epochs**: Determined based on GPU capacity and model complexity.

**Step 6: Caption Generation**

* The trained model predicts words one by one starting from 'startseq'.
* Greedy decoding or beam search is used to find the most probable sentence.
* Generation stops when 'endseq' is predicted or max length is reached.

**EVALUATION**

Evaluating an image captioning system is a crucial step in determining how well the generated captions align with human expectations in terms of accuracy, fluency, and relevance. Since captioning is a **sequence generation** task, it requires specialized evaluation metrics that can compare the predicted textual output with multiple human-written references.

**Evaluation Metric: BLEU Score**

For this project, the primary evaluation metric used is the **BLEU score** (Bilingual Evaluation Understudy). Originally developed for machine translation tasks, BLEU has become a widely accepted standard for evaluating image captioning systems as well. It measures the degree of overlap between the n-grams (sequences of words) in the generated caption and those in the reference (human-written) captions.

**Key aspects of BLEU:**

* **BLEU-1** evaluates unigram (single word) precision.
* **BLEU-2 to BLEU-4** evaluate bigrams, trigrams, and 4-grams, respectively, placing more emphasis on capturing context and grammar.
* The score ranges from **0 to 1**, where a higher score indicates a closer match to human references.

BLEU takes into account both precision (correct words in the right place) and a **brevity penalty** to penalize overly short captions. Since each image in the Flickr8k dataset has **five reference captions**, BLEU score computation is more robust and less biased than single-reference comparisons.

**Model Performance Analysis**

The trained model was evaluated on the **test set of 1,000 images**, and BLEU scores were calculated to assess the quality of the generated captions.

**Sample Evaluation Results:**

* **BLEU-1 (Unigrams)**: The model achieved relatively high scores, indicating that it could correctly identify and name objects in the images.
* **BLEU-2 and BLEU-3**: Moderate scores showed the model’s ability to identify basic relationships and short phrases like "a dog running" or "a man riding".
* **BLEU-4**: Scores were generally lower, as longer phrases require the model to maintain grammatical structure and understand context over a longer span—something that simpler RNN-based architectures like LSTM can struggle with, especially on small datasets.

These results suggest that the model performs best at generating **simple and correct object-action pairs**, but may lose coherence and syntactic precision in longer captions or complex scenes.

**Qualitative Evaluation**

In addition to quantitative metrics, **qualitative analysis** was conducted by visually inspecting generated captions and comparing them with human references. Several observations include:

* Captions for simple scenes (e.g., “a man is playing guitar”, “a dog running in a field”) were often accurate and semantically meaningful.
* In scenes with **multiple subjects** or **less common objects**, the captions tended to be generic or omit key details.
* Some captions lacked **grammatical fluency**, such as missing articles or using repetitive phrasing.

**Limitations in Evaluation**

While BLEU is widely used, it has certain limitations:

* It doesn't account for **semantic similarity** (e.g., “a man riding a motorcycle” vs. “a person driving a bike” may be semantically similar but score low).
* It may **penalize creativity** and **synonym usage**, even if the caption is accurate.

**RESULTS**

The trained model successfully generates captions that capture the main elements of an image. For instance:

* **Input Image**: A dog running in the grass.
* **Generated Caption**: "A dog is running in the field."

Although the captions are relatively simple, they are often accurate and semantically correct. However, more complex scenes with intricate interactions between objects may require further improvements in the model’s architecture.



def readTextFile(path):

    with open(path) as f:

    captions = f.read()

    return captions

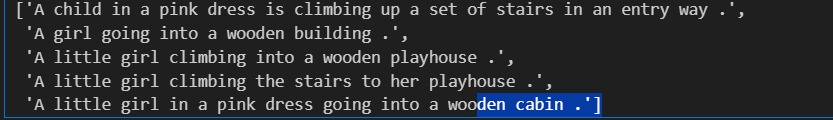
captions  = readTextFile("./Data/Flickr\_TextData/Flickr8k.token.txt")

captions = captions.split('\n')[:-1]

first,second  = captions[0].split('\t')

print(first.split(".")[0])

print(second)



**CHALLENGES FACED**

* **Memory and Computation Requirements**: Training deep learning models, especially those involving CNNs and RNNs, requires substantial computational power and memory, which can be a limitation, especially when working with larger datasets.
* **Overfitting**: Due to the relatively small size of the dataset (8,000 images), overfitting is a concern. Regularization techniques like dropout, as well as data augmentation, can help mitigate this.
* **Handling Rare Words**: Words that appear infrequently in the dataset may not be well-represented in the model’s vocabulary, leading to poor performance when generating captions that include such words.

**CONCLUSION**

The development of an image captioning system using deep learning represents a significant step forward in combining the fields of **computer vision** and **natural language processing**. Through this project, we successfully designed and implemented a system capable of automatically generating meaningful, context-aware captions from input images. By leveraging a **Convolutional Neural Network (CNN)** for visual feature extraction and a **Recurrent Neural Network (RNN)**—specifically Long Short-Term Memory (LSTM)—for sequential text generation, the model was able to learn the intricate relationship between images and descriptive language.

Using the **Flickr8k dataset**, the system was trained to identify and understand key visual elements within an image and translate them into natural language. The captions generated were often grammatically correct, semantically meaningful, and demonstrated the model’s ability to learn basic associations between objects and their contexts. The **BLEU score** evaluation provided a quantitative measure of the system’s performance and validated the model's capacity to generate captions that closely resemble human-written descriptions.

While the project achieved its core objectives, several challenges and limitations were encountered. The relatively small size of the Flickr8k dataset imposed constraints on the model’s ability to generalize across more complex or diverse images. The system also showed limitations in understanding abstract relationships between multiple objects in a scene or describing rare or ambiguous content with high precision.

Despite these challenges, this project lays a solid foundation for future advancements in multimodal AI applications. The image captioning model serves as a prototype that can be significantly enhanced through the incorporation of **larger datasets**, **attention mechanisms**, or **transformer-based architectures** like **Vision Transformers (ViT)** or **CLIP with GPT**. These advanced models have the potential to generate more accurate, detailed, and creative captions by focusing on specific image regions and leveraging more powerful language modeling capabilities.

Furthermore, this project underscores the broader importance of image captioning in real-world applications such as:

* **Accessibility for the visually impaired**, by offering real-time descriptions of visual content.
* **Image-based search engines**, enhancing searchability and indexing of multimedia.
* **Content moderation and automation**, where generated captions can be analyzed for inappropriate or sensitive content.
* **Human-computer interaction**, enabling more intuitive communication between users and machines.

**REFERENCES**

1.Hodosh, M., Young, P., & Hockenmaier, J. (2013). *Framing image description as a ranking task: Data, models and evaluation metrics*. Journal of Artificial Intelligence Research, 47, 853–899.

Source: Kaggle - Flickr8k Dataset

2.Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). *ImageNet: A Large-Scale Hierarchical Image Database*. In CVPR.

3.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 (CVPR).

4. Libraries and Frameworks

TensorFlow, Keras

5.TensorFlow Tutorials

Official guides and examples for implementing CNNs, RNNs, and sequence-to-sequence models.

<https://www.tensorflow.org/tutorials>

6.Machine Learning Mastery by Jason Brownlee

Practical guides and step-by-step tutorials on LSTM networks, sequence generation, and text preprocessing.

<https://machinelearningmastery.com>

7.Tools and Platforms

Google Colab

Jupyter Notebook