Python for AI

23AIE232M

Semester V

Submitted by:

Arya Jayasakar – BL.EN.U4EAC22003

Ishaan Shokeen – BL.EN.U4EAC22022

Krishnendu M Uday – BL.EN.U4EAC22027

Submitted to:

Dr. Nidhin Prabhakar T V
Assistant Professor
Department of Computer Science and Engineering
Amrita Vishwa Vidyapeetham, Bengaluru

ACKNOWLEDGEMENTS

The satisfaction that accompanies successful completion of any task would be incomplete without mention of people who made it possible, and whose constant encouragement and guidance have been source of inspiration throughout the course of this project work.

We offer our sincere pranams at the lotus feet of **"AMMA", MATA AMRITANANDAMAYI DEVI** who showered her blessing upon us throughout the course of this project work.

We owe our gratitude to **Prof. Manoj P.**, Director, Amrita Vishwa Vidyapeetham Bengaluru Campus. We would like to place our heartfelt gratitude to **Dr. Gopalakrishnan E.A.,** Principal, Amrita School of Computing and Amrita School of Artificial Intelligence, Bengaluru for his valuable support and inspiration.

It is a great pleasure to express our gratitude and indebtedness to our project guide **GUIDE NAME AND DESIGNATION**, Department of Computer Science and Engineering, Amrita School of Computing, Bengaluru for her/his valuable guidance, encouragement, moral support, and affection throughout the project work.

We would like to thank express our gratitude to project panel members for their suggestions, encouragement, and moral support during the process of project work and all faculty members for their academic support. Finally, we are forever grateful to our parents, who have loved, supported and encouraged us in all our endeavors.

ABSTRACT

The "Image Caption Generator" is a user-friendly desktop application that integrates advanced deep learning technologies to generate meaningful captions for uploaded images. This application utilizes the BLIP (Bootstrapped Language-Image Pre-training) model, a state-of-the-art framework for image captioning. Through a graphical user interface (GUI) built with Python's Tkinter library, users can upload images, view them, and generate descriptive captions that encapsulate the visual content of the image. The application combines the power of the Hugging Face Transformers library with PyTorch to process the uploaded images and produce natural language captions.

The project aims to simplify the process of understanding and describing visual data, making it accessible to users across various domains such as accessibility tools for visually impaired individuals, content creation, and automated image documentation. By leveraging pre-trained models, the application eliminates the need for extensive training data, ensuring efficient and accurate caption generation. This document presents the application architecture, implementation details, and practical use cases, highlighting its potential to bridge the gap between visual and textual data.

INTRODUCTION

In today's digital age, the ability to interpret and describe visual content has become increasingly significant across a wide range of applications, from accessibility tools for visually impaired individuals to automated content generation for social media and e-commerce. With the advancements in artificial intelligence and deep learning, image captioning has emerged as a transformative technology that bridges the gap between visual and textual data.

The "Image Caption Generator" project is designed to simplify and automate the process of generating captions for images by combining a user-friendly graphical interface with cutting-edge AI models. The application is built using Python's Tkinter library for the GUI and leverages the BLIP (Bootstrapped Language-Image Pre-training) model, a state-of-the-art framework for image captioning. By processing an uploaded image, the application generates a concise, descriptive caption that summarizes the visual content in natural language.

This tool has vast applications in various domains, including assisting visually impaired users to understand visual content, enhancing content creation workflows, improving photo organization, and automating metadata generation for images. Furthermore, the use of pre-trained models, such as those provided by Hugging Face Transformers, ensures high accuracy and performance without the need for extensive computational resources.

The "Image Caption Generator" project is an example of how AI technologies can be applied to solve real-world problems, making image interpretation more accessible and efficient for diverse user groups. This document explores the design, implementation, and potential applications of the project, showcasing its ability to transform how we interact with visual data.

LITERATURE REVIEW

[1] The field of image captioning has witnessed substantial progress, particularly with the advent of deep learning methodologies. This section reviews key advancements and approaches in image captioning, with insights derived from the work of Preeti Voditel, Aparna Gurjar, Aakansha Pandey, Akrati Jain, Nandita Sharma, and Nisha Dubey, as well as related studies.

A. Human-Like Image Understanding

One of the primary objectives of image captioning is to replicate the human brain's ability to perceive and describe visual content accurately and efficiently. This task involves recognizing objects, actions, and relationships within an image and generating coherent textual descriptions. Recent advancements in deep learning, particularly in computer vision, have facilitated the training of models to mimic this capability with reasonable accuracy.

B. Challenges in Image Caption Generation

Despite technological progress, image captioning faces challenges in generating accurate and grammatically coherent captions. The process requires:

1. Identifying objects, attributes, and relationships within an image.
2. Selecting appropriate words and structuring them into meaningful sentences. These tasks demand a combination of visual perception and linguistic proficiency, making it a complex undertaking.

C. Integration of Computer Vision and Natural Language Processing (NLP)

The task of image captioning necessitates a dual approach that integrates computer vision and NLP:

- Computer Vision: To analyze and extract features from images.
- Natural Language Processing: To articulate visual information in a natural and human-readable format.
  This integration underpins the development of effective captioning systems capable of bridging the gap between visual and textual modalities.

D. Existing Image Captioning Methods

The literature classifies image captioning techniques into three broad categories:

1. Template-Based Captioning:
   - Utilizes predefined templates with placeholders for objects, attributes, and actions.
   - Advantages: Simplicity and ease of implementation.
   - Limitations: Lack of flexibility and inability to generate novel captions.
2. Retrieval-Based Captioning:
   - Generates captions by selecting from a repository of existing captions in the training dataset.
   - Advantages: Faster generation and adherence to existing linguistic patterns.
   - Limitations: Limited creativity and adaptability to new contexts.
3. Novel Caption Generation:
   - Employs deep learning techniques, such as convolutional neural networks (CNNs) and long short-term memory networks (LSTMs).
   - Advantages: Ability to create innovative and contextually relevant captions.
   - Limitations: Computationally intensive and reliant on extensive training datasets.

E. Approaches to Caption Generation

Image captioning can be approached from two distinct perspectives:

1. Top-Down Approach:
   - Generates captions based on an overall summary or abstract representation of the image.
2. Bottom-Up Approach:
   - Focuses on identifying specific elements within the image and combining them into coherent sentences.
   - This method ensures more granular and detailed captions.

F. Context Sequence Memory Network (CSMN)

A significant advancement in image captioning is the introduction of the Context Sequence Memory Network (CSMN). This architecture leverages long-term memory to enhance the captioning process, addressing the limitations of traditional recurrent neural networks (RNNs), such as the vanishing gradient problem. CSMN enables models to retain and utilize contextual information over extended sequences, improving caption quality and relevance.

G. Visual Relationship Graphs

Recent research highlights the potential of visual relationship graphs in improving caption generation. These graphs encode relationships between objects and their attributes, providing richer contextual information. By incorporating these graphs, captioning models can generate more accurate and contextually nuanced descriptions.

[2] This paper focuses on generating image captions using deep learning models, primarily utilizing Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). The system employs an encoder-decoder framework, where CNNs extract image features and Long Short-Term Memory (LSTM) units create contextually accurate and coherent captions.

Techniques: Attention mechanisms are incorporated to enhance the contextual alignment between visual elements and the corresponding text. Transformer-based models like ViLBERT and LXMERT are highlighted for their strong multimodal processing, while hybrid models such as Oscar and UNITER effectively combine CNN and transformer architectures.

Dataset: Flickr8k dataset.

Evaluation Metrics: BLEU and CIDEr scores are used to assess the quality of generated captions, ensuring both semantic and grammatical precision. The findings emphasize the role of attention mechanisms and transfer learning in improving model performance.

This study outlines a robust approach combining conventional and innovative techniques for better image captioning. Future work suggests focusing on refining language fluency, adopting multimodal learning, and addressing potential ethical concerns.

[3] This paper investigates image caption generation using deep learning models, focusing on the integration of EfficientNet (CNN) for image feature extraction and GRU (Gated Recurrent Units) for caption generation. The system is built using an encoder-decoder architecture, where the CNN encoder processes input images and extracts features, while the GRU decoder generates descriptive captions.

Techniques: The study incorporates the Bahdanau Attention Mechanism to improve information flow and align image regions with corresponding words, addressing the challenge of information loss in fixed-length encoding.

Dataset: MSCOCO 2017 dataset.

Evaluation Metrics: The model's performance is evaluated using the BLEU score, which measures the similarity between generated and reference captions, ensuring the quality and relevance of the output.

This research highlights the advantages of using EfficientNet for efficient feature extraction and the Bahdanau attention mechanism for context-aware captioning. The study demonstrates significant improvements in the quality of generated captions, showcasing the effectiveness of combining CNNs, RNNs, and attention mechanisms. Future work suggests incorporating generative models like GANs or VAE with LSTM for enhanced performance and exploring region-based detection algorithms.

[4] This paper investigates the generation of natural language descriptions from visual data, focusing on still image captioning. Early systems used complex, hand-designed components like visual primitive recognizers, limited to specific domains such as traffic or sports.

Techniques: The study reviews template-based methods, graph-based approaches, and neural network integrations. Template-based methods use predefined structures, while graph-based models create complex detection graphs but still struggle with novel combinations. Neural networks co-embed images and text, improving expressivity but not generating descriptions for unseen objects.

Proposed Model: The paper introduces an approach combining deep convolutional networks (CNNs) for image feature extraction and recurrent networks (RNNs) for sequence modeling, enabling better tracking of objects in generated text and improved benchmark performance.

Evaluation Metrics: Emphasis is placed on generation metrics over ranking metrics, as traditional methods do not effectively capture the complexity of valid descriptions, especially with larger vocabularies.

This research highlights the shift from rigid early techniques to innovative CNN-RNN models. The approach improves flexibility and expressiveness, and future work could enhance generative capabilities and explore multimodal learning for better results.

[5] This paper discusses image caption generation, a key research area combining deep learning and natural language processing (NLP) to produce descriptive labels for images. This is vital for organizing digital content and improving accessibility for visually impaired individuals.

Datasets Utilized: The **COCO (Common Objects and Contexts)** dataset, containing 100,000 images with five captions each, is commonly used for training due to its diverse and rich content.

Technological Framework: Image captioning models use **Convolutional Neural Networks (CNNs)** to extract image features and **Long Short Term Memory (LSTM) networks** for generating sequential text.

Role of LSTM: LSTMs are effective at capturing long-term dependencies, essential for coherent and relevant captions.

NLP Integration: NLP techniques help convert visual information into human-readable text, improving machine communication.

Applications and Impact: Image captioning aids in organizing content and enhances accessibility, making it valuable for web development and users relying on screen readers.

This research highlights the intersection of deep learning and NLP, emphasizing continuous advancements that improve the accuracy and applicability of image captioning systems

[6] This paper reviews advancements in image caption generation, emphasizing the integration of deep learning techniques to create accurate and contextually relevant descriptions. The approach commonly involves Convolutional Neural Networks (CNNs) for image feature extraction and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, for generating sentences based on extracted features.

Techniques: CNNs, such as the Visual Group Geometry (VGG) network, provide effective image features for captioning. The combination of CNNs and LSTMs has become standard practice. Recent models also incorporate attention mechanisms to focus on specific parts of the image while generating captions, enhancing the relevance and coherence of the output.

Dataset: Datasets like MSCOCO, which pair images with multiple descriptive sentences, are commonly used. These large-scale datasets improve model accuracy and reduce training losses, facilitating more effective captioning.

Evaluation Metrics: Common metrics include BLEU, METEOR, and CIDEr, which compare generated captions to reference captions to assess quality. Results are further categorized based on accuracy, such as error-free, minor errors, or unrelated descriptions, to better understand model performance.

This review highlights the significant role of CNNs, RNNs, attention mechanisms, diverse datasets, and robust evaluation metrics in advancing image captioning. Future research may focus on refining models for better language fluency, multimodal integration, and addressing ethical considerations in caption generation.

[7] This paper presents a robust approach to generating image captions using advanced deep learning techniques.

Techniques: The study highlights the use of an attention-based encoder-decoder model. **Inception V3**, a pre-trained CNN on ImageNet, is used for feature extraction due to its effectiveness. **Gated Recurrent Units (GRUs)** are chosen over LSTMs for processing sequential data, and an **attention mechanism** is incorporated to focus on relevant image parts, improving caption quality.

Dataset: The **MS-COCO** dataset, with over 300,000 images and multiple human-generated captions per image, is used for training. It provides comprehensive real-world scene data, outperforming datasets like **Flickr 8k** and **Flickr 30k** in contextual depth.

Evaluation Metrics: While not explicitly detailed, the paper likely employs standard metrics like **BLEU**, **METEOR**, or **CIDEr** for evaluating caption quality against human-written descriptions.

This study showcases a strong image captioning framework using deep learning, effective datasets, and evaluation metrics. Future research could focus on refining attention mechanisms or applying the model to diverse datasets.

[8] This paper presents an advanced approach to image caption generation using deep learning techniques, with a focus on combining CNNs and LSTMs.

Techniques: The study highlights the use of a recurrent convolutional architecture and introduces the **Neural Image Captioning (NIC)** model, which combines **GoogLeNet** and **LSTM** to optimize accurate caption generation. The **LRCN model** is also discussed for its versatility but noted for performance limitations. **Region-based Convolutional Networks (R-CNN)** are employed to align visual features with image regions, providing better precision than earlier models. Additionally, a visual concepts-based approach incorporates nouns, verbs, and adjectives through multiple instance learning to enhance caption quality.

Dataset: The **Flickr8k dataset**, consisting of 8,000 images each paired with five descriptions, is used for training. It helps reduce overfitting and improves model generalization, with 6,000 images specifically used in the training set.

Evaluation Metrics: The paper evaluates model performance using **BLEU scores**, reporting a score of 69.8 to demonstrate effectiveness in generating precise captions. A qualitative analysis is also conducted, showcasing example captions and discussing challenges in accurately capturing visual details.

This study underscores a strong image captioning framework combining CNNs, LSTMs, and evaluation metrics like BLEU scores for robust performance. Future research could focus on refining model architectures or incorporating additional datasets for broader applicability.

[9] This paper analyzes automatic image caption generation, focusing on deep learning techniques, datasets, and evaluation metrics.

Techniques: The study highlights a **multimodal deep learning framework** that integrates **CNNs** with unlabeled data for feature learning. **CNN and RNN integration** is used for image processing and sentence generation, improving object tracking in captions. The **"Show, Attend and Tell" model** introduced attention mechanisms (hard and soft) to enhance caption quality.

Dataset: All models were trained on the same dataset for fair comparison, though specific details are not provided.

Evaluation Metrics: Performance is evaluated against human-generated captions, noting that training epochs vary by model complexity. Measuring average metric values across models and epochs helps assess overall model reliability.

This review emphasizes the use of deep learning, attention mechanisms, and standardized evaluation in image captioning, suggesting future work could focus on optimizing attention or testing with diverse datasets.

[10] This paper explores deep learning techniques for image caption generation, focusing on model architectures, datasets, and evaluation metrics.

Techniques Used: The study uses **CNNs** like **VGG16** and **ResNet50** for feature extraction and **LSTMs** for generating captions, making them ideal for sequence prediction. The **"Show and Tell" model** is highlighted as a benchmark, combining LSTMs with advanced CNNs such as **Inception v3**.

Dataset: The **Flickr8k dataset** (8,000 images with multiple captions) is used for training, known for its diverse content that aids model generalization.

Evaluation Metrics: Accuracy is used for performance assessment, with **ResNet50** achieving 73% compared to 29% for **VGG16**. The captions are also processed with **Google's gTTS** for text-to-speech, showing real-world applicability.

This paper demonstrates advancements in deep learning for image captioning, emphasizing CNNs, LSTMs, and practical applications like text-to-speech. Future research could enhance models with attention mechanisms or larger datasets.

[11] This paper reviews techniques, datasets, and evaluation metrics in image caption generation using deep learning.

Techniques: The study uses **CNNs** for feature extraction and **LSTMs** for generating captions, enabling effective visual data encoding. **Attention mechanisms** improve context-aware captioning, while **beam search algorithms** enhance word sequence selection.

Dataset: The primary dataset is **Flicker 8K** (8,000 images with captions), suitable for training. **MSCOCO** offers more data but requires greater computational resources.

Evaluation Metrics: Common metrics like **BLEU, METEOR, and CIDEr** are used to assess n-gram overlap and semantic similarity, providing quantitative performance measures.

This review emphasizes using deep learning, effective datasets, and evaluation metrics to advance image captioning.

IMPLEMENTATION

This can be divided into four main sections: GUI Design and Layout, Methodologies Used, Evaluation Metrics, and Python Packages Used. These sections outline the technical and conceptual details of the project.

1. GUI Design and Layout

The graphical user interface (GUI) is designed to provide an intuitive experience for users to interact with the application.

- **Tools**: The GUI is built using the Tkinter library in Python, known for its simplicity and versatility in creating desktop applications.
- **Components**:
    1. **Image Display Area**:
        - A Label widget serves as a placeholder for displaying uploaded images. Images are resized using the Pillow library to fit within the application's dimensions while maintaining their aspect ratio.
    2. **Upload Button**:
        - A Button widget allows users to browse and upload an image file from their device. The file dialog restricts the selection to supported formats like .png, .jpg, .jpeg, and .bmp.
    3. **Caption Generation Button**:
        - This button is initially disabled and is enabled only after an image is uploaded. It triggers the caption generation process using the pre-trained BLIP model.
    4. **Caption Display Area**:
        - A Text widget is used to display the generated caption in a scrollable format. It supports multiline outputs for detailed captions.
- **Layout**:

    - The layout is structured vertically with proper spacing and padding to ensure clear separation between components. The GUI adjusts dynamically to fit images and text within the application window.

2. Methodologies Used

The methodologies used in the project integrate deep learning techniques with efficient user interface design.

- **Image Preprocessing**:
    - The uploaded image is opened using the Pillow library, resized to a maximum resolution of 700x500 pixels, and converted to the RGB color space.
    - Preprocessing ensures compatibility with the BLIP model, which requires images in a specific format for input.
- **Image Captioning**:
    - The BLIP (Bootstrapped Language-Image Pre-training) model is used for caption generation. It leverages a transformer-based architecture to generate natural-language descriptions for visual content.
    - The Hugging Face transformers library provides pre-trained weights and processors for seamless integration.
- **Model Inference**:
    - The processed image is passed through the BLIP model, and the output is decoded into human-readable text. The model uses attention mechanisms to understand visual content and generate contextually relevant captions.

3. Evaluation Metrics

The performance of the "Image Caption Generator" is evaluated qualitatively and quantitatively:

- **Qualitative Evaluation**:
    - Human feedback is gathered to assess the relevance and descriptiveness of the generated captions.

    - The application is tested with diverse images, including landscapes, objects, and human activities, to ensure robustness.

- **Quantitative Evaluation**:
    - **BLEU (Bilingual Evaluation Understudy)**: Measures the similarity between generated captions and ground truth captions.
    - **CIDEr (Consensus-based Image Description Evaluation)**: Evaluates how well the generated captions align with human-annotated references.
    - **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)**: Focuses on recall by comparing overlapping words between generated and reference captions.

4. Python Packages Used

Several Python libraries are utilized in the implementation, each serving a specific purpose:

1. **Tkinter**:
    - **Purpose**: To build the graphical user interface.
    - **Description**: A standard Python library for creating desktop applications, offering widgets like labels, buttons, and text areas for user interaction.
2. **Pillow**:
    - **Purpose**: For image processing.
    - **Description**: A fork of the Python Imaging Library (PIL), used to open, manipulate, and resize images to fit the GUI layout and model input requirements.
3. **Transformers**:
    - **Purpose**: For integrating the BLIP model.
    - **Description**: A library by Hugging Face that provides pre-trained models and tokenizers for natural language processing and image captioning tasks.
4. **Torch**:
    - **Purpose**: For running the BLIP model.

- **Description**: A deep learning framework that supports tensor operations and model inference on both CPUs and GPUs.

5. **OS (Optional)**:
    - **Purpose**: For handling file paths.
    - **Description**: A standard library used to manage file directories and paths during image upload.

IMPLEMENTATION

The "Image Caption Generator" successfully bridges the gap between visual and textual data using state-of-the-art AI technologies and an intuitive GUI.

By leveraging the BLIP model and Hugging Face Transformers, the application generates concise and meaningful captions for a wide range of images.

The modular design and use of Python libraries ensure efficiency, scalability, and ease of maintenance.

This project demonstrates the potential of AI in automating tasks like image annotation and accessibility, making it useful for various applications such as assisting visually impaired individuals, content creation, and image documentation.

Overall, the project provides a functional and robust tool for understanding and describing visual content, showcasing the capabilities of modern AI in solving real-world problems.

Result

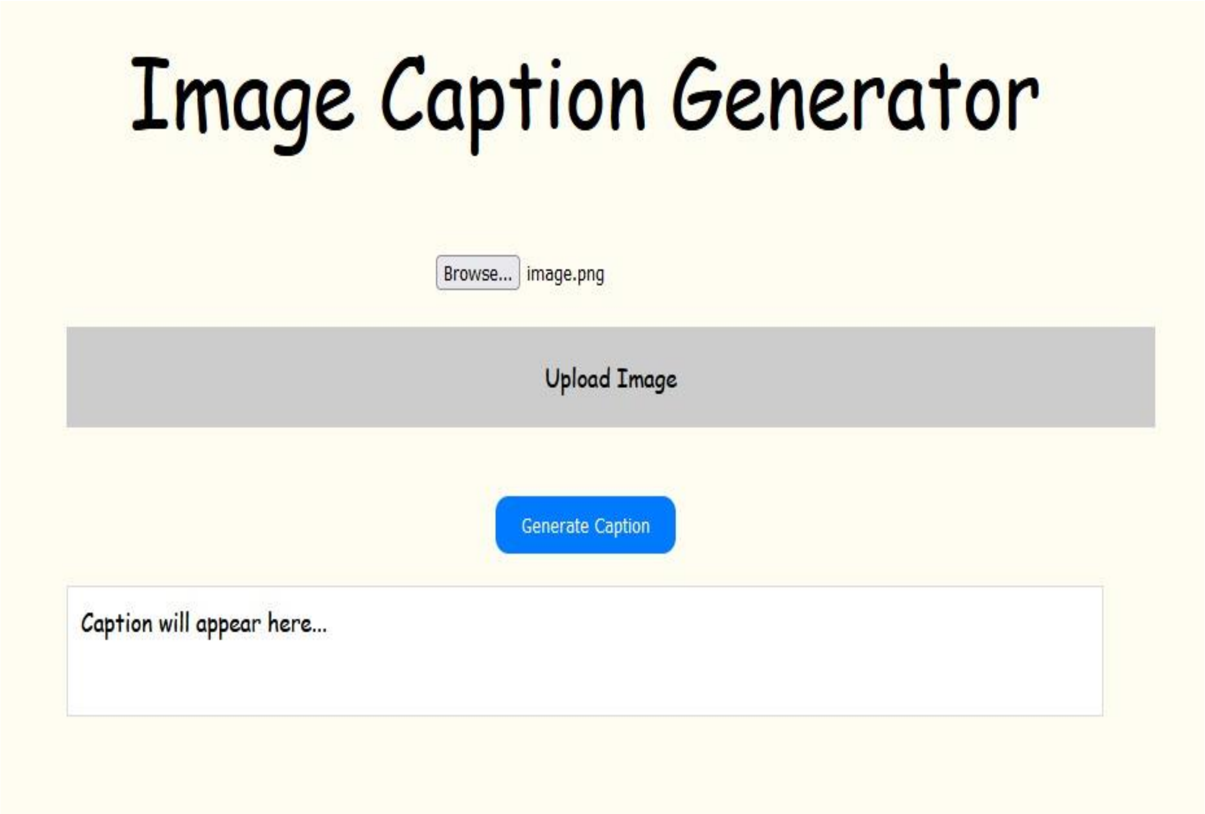
Image Caption Generator
Browse... image.png
Upload Image
Generate Caption
Caption will appear here...

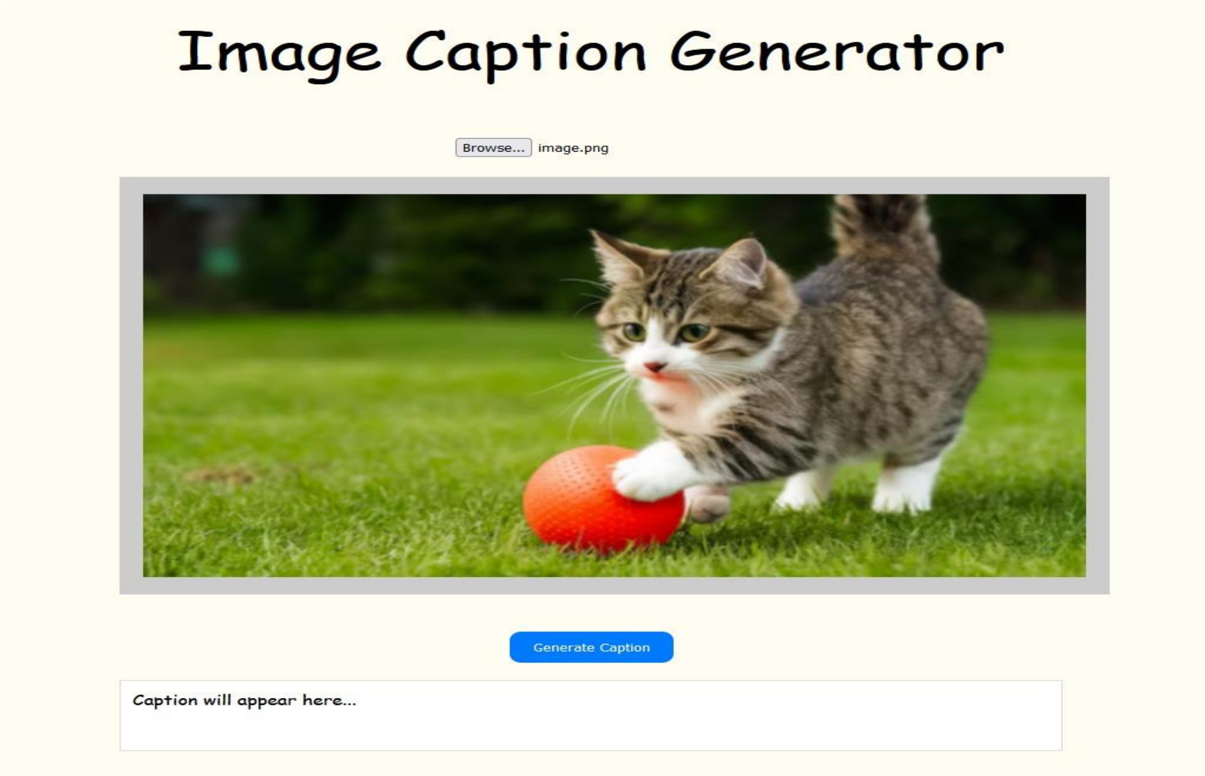
Image Caption Generator
Browse... image.png
Generate Caption
Caption will appear here...

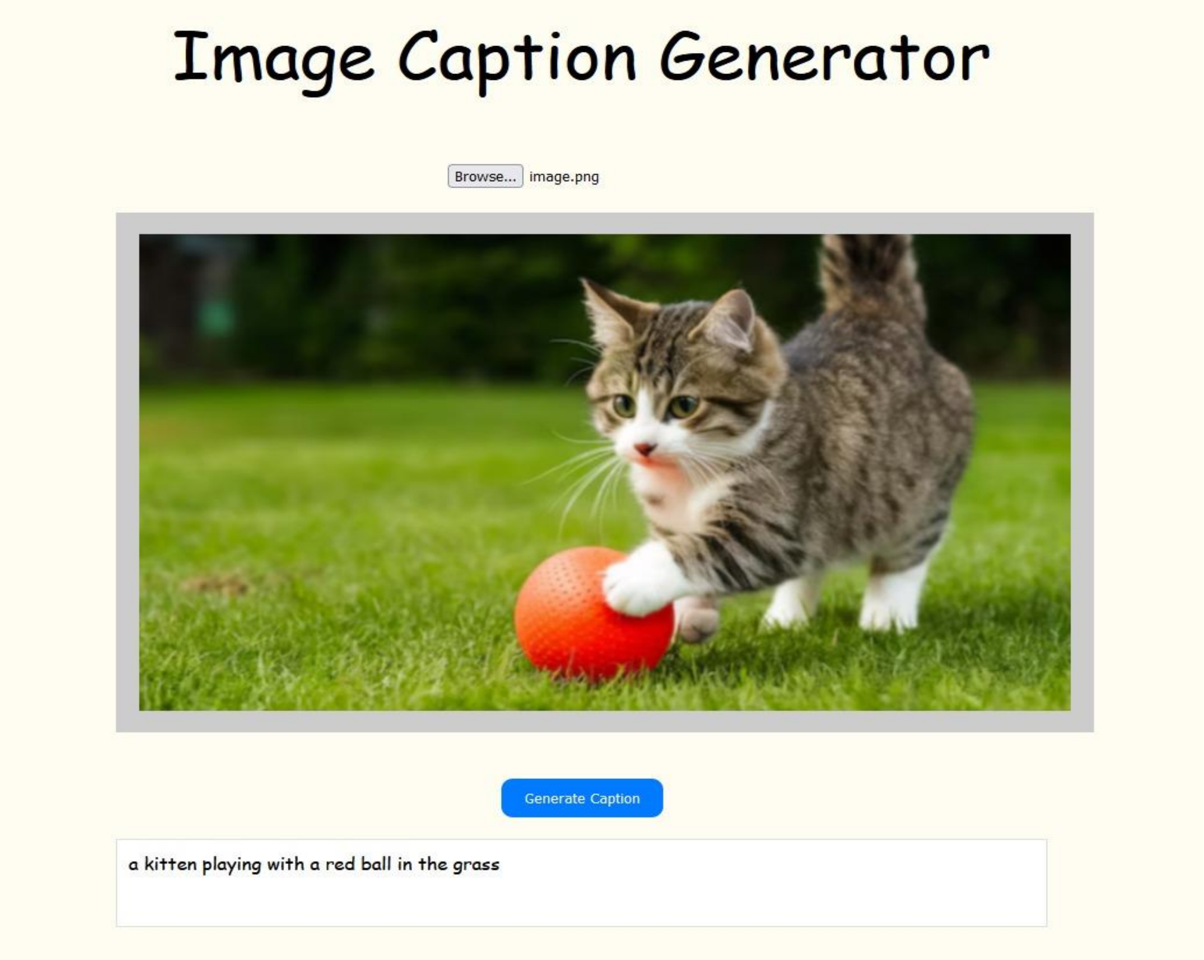

Image Caption Generator

image.png

a kitten playing with a red ball in the grass

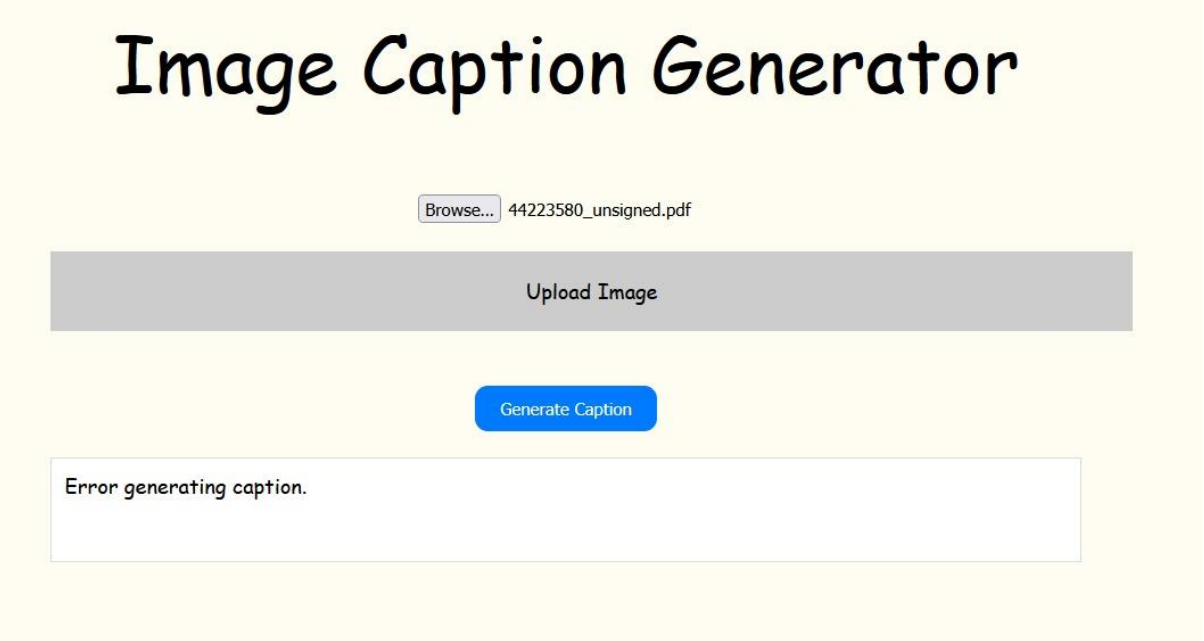

Image Caption Generator

44223580_unsigned.pdf

Error generating caption.

Conclusion

The "Image Caption Generator" successfully bridges the gap between visual and textual data using state-of-the-art AI technologies and an intuitive GUI. By leveraging the BLIP model and Hugging Face Transformers, the application generates concise and meaningful captions for a wide range of images. The modular design and use of Python libraries ensure efficiency, scalability, and ease of maintenance. This project demonstrates the potential of AI in automating tasks like image annotation and accessibility, making it useful for various applications such as assisting visually impaired individuals, content creation, and image documentation.

Overall, the project provides a functional and robust tool for understanding and describing visual content, showcasing the capabilities of modern AI in solving real-world problems.

FUTURE SCOPE

While the "Image Caption Generator" is a functional prototype, several enhancements can be implemented to improve its utility, scalability, and user experience:

1. **Enhanced Caption Accuracy**:
   - Integrate more advanced or fine-tuned models to generate captions that are contextually richer and more accurate.
   - Allow customization of captions, such as detailed or abstract descriptions.
2. **Support for Multiple Languages**:
   - Incorporate multilingual capabilities so users can generate captions in different languages, enhancing accessibility for non-English speakers.
3. **Batch Processing**:
   - Add functionality for batch processing, allowing users to upload and caption multiple images simultaneously.
4. **GPU Support**:

- Enhance performance by enabling automatic GPU detection and utilization, significantly reducing processing time for caption generation.

5. **Improved User Interface**:
    - Add drag-and-drop support for image upload and allow users to save generated captions directly to their device.
    - Incorporate a dark mode or other visual customizations for improved usability.
6. **Integration with Cloud Services**:
    - Enable cloud storage integration for uploading and storing images, making the tool more versatile for collaborative environments.
7. **Mobile and Web Versions**:
    - Develop mobile and web-based versions of the application to make it accessible to a broader audience.
8. **Custom Models**:
    - Allow users to upload or train custom models to meet domain-specific requirements, such as medical imaging or industrial applications.
9. **Real-Time Captioning**:
    - Extend the tool for real-time captioning through live video feeds, enabling applications in surveillance, video analysis, and more.
10. **Evaluation and Feedback Mechanism**:

- Implement a feedback loop where users can rate the generated captions, enabling continuous improvement of the underlying model.

REFERENCES

[1] Voditel, Preeti & Gurjar, Aparna & Pandey, Aakansha & Jain, Akrati & Sharma, Nandita & Dubey, Nisha. (2023). Image Captioning - A Deep Learning Approach Using CNN and LSTM Network. 343-348. 10.1109/ICPCSN58827.2023.00062.

[2] D. J. B. Saini, S. Kumar, K. Joshi, A. K. Pathak, S. Jain and A. Singh, "A Novel Approach of Image Caption Generator using Deep Learning," 2023 Third International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS), Gobichettipalayam, India, 2023, pp. 24-29, doi: 10.1109/ICUIS60567.2023.00012. keywords: {Deep learning;Training;Analytical models;Neural networks;Transfer learning;Reinforcement learning;Ubiquitous computing;Image;Caption;Xception;Recurrent neural network (RNN);Long short-term memory (LSTM);Convolutional neural networks (CNN);Deep learning;Computer vision (CV)},

[3] S. V. Patnaik, R. Mukka, R. Devpreyo and A. Wadhawan, "Image Caption Generator using EfficientNet," 2022 10th International Conference on Reliability,

Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2022, pp. 1-5, doi: 10.1109/ICRITO56286.2022.9964637. keywords: {Training;Vocabulary;Analytical models;Image coding;Predictive models;Planning;Noise measurement;Image Caption;CNN;RNN;LSTM;EfficientNet;Semantic Ontology;Attention based},

[4] Vinyals, Oriol & Toshev, Alexander & Bengio, Samy & Erhan, Dumitru. (2015). Show and tell: A neural image caption generator. 3156-3164. 10.1109/CVPR.2015.7298935.

[5] Sehgal, Smriti & Sharma, Jyoti & Chaudhary, Natasha. (2020). Generating Image Captions based on Deep Learning and Natural language Processing. 165-169. 10.1109/ICRITO48877.2020.9197977.

[6] C. Amritkar and V. Jabade, "Image Caption Generation Using Deep Learning Technique," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-4, doi: 10.1109/ICCUBEA.2018.8697360. keywords: {Visualization;Natural languages;Decoding;Training;Computer architecture;Microprocessors;Logic gates;Neural Network;Image;Caption;Description;Long Short Term memory(LSTM);Deep Learning},

[7] A. Hani, N. Tagougui and M. Kherallah, "Image Caption Generation Using A Deep Architecture," 2019 International Arab Conference on Information Technology (ACIT), Al Ain, United Arab Emirates, 2019, pp. 246-251, doi: 10.1109/ACIT47987.2019.8990998. keywords: {image captioning;convolutional neural networks;recurrent neural networks;attention mechanism},

[8] N. Indumathi, R. J. Divyalakshmi, J. Stalin, V. Ramachandran and P. Rajaram, "Apply Deep Learning-based CNN and LSTM for Visual Image

Caption Generator," 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2023, pp. 1586-1591, doi: 10.1109/ICACITE57410.2023.10183097. keywords: {Training;Deep learning;Visualization;Computer vision;Computational modeling;Transfer learning;Generators;Deep learning;CNN;RNN;LSTM},

[9] V. Kesavan, V. Muley and M. Kolhekar, "Deep Learning based Automatic Image Caption Generation," 2019 Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2019, pp. 1-6, doi: 10.1109/GCAT47503.2019.8978293. keywords: {automated captions;deep neural network;CNN;RNN;feature extraction;attention},

[10] J. Sudhakar, V. V. Iyer and S. T. Sharmila, "Image Caption Generation using Deep Neural Networks," 2022 International Conference for Advancement in Technology (ICONAT), Goa, India, 2022, pp. 1-3, doi: 10.1109/ICONAT53423.2022.9726074. keywords: {Deep learning;Recurrent neural networks;Image recognition;Machine vision;Natural languages;Multimedia Web sites;Object detection;Image Captioning;Deep Neural networks;CNN;RNN;Text-to-Speech},

[11] M. Sailaja, K. Harika, B. Sridhar, R. Singh, V. Charitha and K. S. Rao, "Image Caption Generator using Deep Learning," 2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC), Bhubaneswar, India, 2022, pp. 1-5, doi: 10.1109/ASSIC55218.2022.10088345. keywords: {Deep learning;Computational modeling;Neural networks;Generators;Convolutional neural networks;Object recognition;Long short term memory;Convolutional Neural Network (CNN);Long Short-Term Memory (LSTM);Machine Vision;Natural Language Processing (NLP);Input image;Framing the Sentence;Feature extraction},