Image Caption Generator

Project Overview:

The Image Caption Generator project aims to create a deep learning-based solution that automatically generates descriptive captions for images. This project will explore the intersection of Computer Vision and Natural Language Processing (NLP) to build a model capable of understanding the content of an image and articulating it in natural language. The project will leverage powerful machine learning frameworks like TensorFlow or PyTorch and explore different architectures, including CNN-LSTM and Transformers, to achieve this goal.

A key focus of this project is accessibility, particularly for visually impaired users. By integrating text-to-speech functionality, the system will not only generate descriptive captions but also convert them into audio output, making visual content accessible to users who cannot see. This accessibility feature transforms the image captioning system into a comprehensive assistive technology solution.

Team Members:

Team Leader: Abdelwhab Mohamed Mahmoud Saad (Phone: +201050711997, Email: abdelwhabmohammed0@gmail.com)

Name	Email	Depi_ID	Role
Abdelwhab Mohamed Mahmoud Saad	abdelwhabmohammed0@gmail.com	21076307	MileSone1: extracted image features using inception V3

			MileStone2: Trained: LSTM-Transformer Hyprid model with the Inception V3 features MileStone3: FineTuned BLIB2 with different training function (Progressive unfreezing) MileStone4: MLOps Implementation MileStone5: Final Project Report
Yousef Abdulati Abdelalim Mohammed	yosefabdulati@gmail.com	21086755	MileSone1: extracted image features using VGG19 MileStone2: prepared outputs from past milestone to be ready for NLP Training MileStone3: Developed a web application using Flask framework MileStone4: MLOps Implementation MileStone5: Final Project Report
Nadine khaled Abelfattah	2202013@student.eelu.edu.eg	21085422	MileSone1: extracted image features using VGG19 MileStone2: Evaluated our model performance with other pretrained models using BLEU & other methods MileStone3: FineTuned BLIB2 MileStone4: MLOps Implementation MileStone5: Future Improvement Recommendations

Shahd Ezzat Farag	shahdezzat510@gmail.com	21096709	MileSone1: Loaded & prepred Flickr 30K Data Set and EDA MileStone2: Evaluated our model performance with other pretrained models using BLEU & other methods MileStone3: Deployed the models using azure MileStone4: Model Monitoring MileStone5: Future Improvement Recommendations
Nouran Ahmed Fouad Eid	2202002@student.eelu.edu.eg	21086544	MileSone1: extracted image features using Resnet50 MileStone2: prepared outputs from past milestone to be ready for NLP Training MileStone3: Integrated text-to-speech (TTS) functionality to convert generated captions into audio MileStone4: Model Monitoring MileStone5: Final Presentation
Mahmoud Saad Ahmed	medordad@gmail.com	21070019	MileSone1: Loaded & prepared MS COCO Data Set and EDA MileStone2: Trained VGG19 Transformer-LSTM model MileStone3: FineTuned BLIB2 with different training function(Weight-Decay) MileStone4: Model Monitoring MileStone5: Final Presentation

Milestone1 (16/09/2025): Data Collection, Preprocessing, and Exploration

Objectives:

Collect, preprocess, and explore a suitable dataset for training an image captioning model.

Tasks:

. Data Collection:

Gather a labeled dataset for image captioning, such as Flickr k30, Flickr k10, or MS COCO.

Ensure the dataset contains a diverse range of images and corresponding captions to support robust model training.

. Data Preprocessing:

Image Preprocessing:

Resize and normalize images to a consistent format for the deep

1. learning model.
2. Extract features from the images using a pre-trained Convolutional
3. Neural Network (CNN) like Xception, VGG, ResNet, or Inception.

Text Preprocessing:

Tokenize the captions, creating a vocabulary of all unique words.

1. Convert the text captions into numerical sequences that the model can process.
2. Determine the maximum caption length to handle variable-length descriptions.

. Exploratory Data Analysis (EDA):

Visualize sample images and their corresponding captions.

Analyze the distribution of caption lengths and vocabulary size.

Investigate any potential biases or challenges in the dataset.

Deliverables:

1. Cleaned and Preprocessed Dataset: A fully processed dataset ready for model
2. development.
3. Preprocessing Pipeline Documentation: A detailed description of the image
4. and text preprocessing techniques applied.
5. EDA Report: A comprehensive exploration of the dataset, including visualizations
6. and identified challenges.

Milestone2 (30/09/2025): Model Development and Training

Objectives:

Develop and train a deep learning model for image caption generation.

Tasks:

1. **Model Selection:**
   1. CNN-LSTM Model: Implement a model that uses a CNN to extract image features and an LSTM (Long Short-Term Memory) network to generate the caption sequence.
   2. Transformer Model: Alternatively, implement a Transformer-based model, which uses attention mechanisms for both image feature extraction and language generation.

2. **Model Training:**

   1. Train the selected model on the preprocessed dataset.
   2. Use a generator to feed data in batches to the model, especially for large datasets.
   3. Split the data into training, validation, and test sets to evaluate model performance.

3. **Model Evaluation:**

   1. Evaluate the model's performance using metrics like BLEU (Bilingual Evaluation Understudy), METEOR, and ROUGE.

4. **Model Optimization:**

1. Fine-tune hyperparameters such as learning rate, batch size, and the number of layers to improve performance.
2. Explore techniques like beam search to generate higher-quality captions during inference.

Deliverables:

1. Trained Image Captioning Model: A deep learning model capable of generating captions for images.
2. Model Evaluation Report: A report detailing the performance of the model using relevant metrics.

Milestone3 (21/10/2025): Advanced Techniques and Deployment

Objectives:

Enhance the model and deploy it for real-time predictions.

Tasks:

1. **Transfer Learning and Fine-Tuning:** Fine-tune pre-trained models (e.g., from ImageNet) to improve accuracy and efficiency.
2. **Cloud Deployment:** Deploy the trained model to a cloud platform like Azure, AWS, or Google Cloud. Implement a RESTful API to allow for real-time caption generation from image inputs.
3. **Web Interface for Predictions:** Develop a web application using a framework like Flask or FastAPI to enable users to upload images and receive generated captions.

4. **Text-to-Speech Integration for Accessibility**:Integrate text-to-speech (TTS) functionality to convert generated captions into audio output. Implement this feature as an accessibility service for visually impaired users. Support multiple languages and voice options for enhanced user experience. Provide audio controls for playback speed, volume, and voice selection.

Deliverables:

1. Enhanced Model Using Transfer Learning: A fine-tuned model optimized for image captioning.
2. Deployed Model on Cloud: The image captioning model deployed for real-time predictions.
3. Deployed Model with Web Interface: A user-friendly web interface for real-time predictions.
4. Accessibility-Enhanced Application: A complete solution with text-to-speech capabilities for visually impaired users.

Milestone 4 (07/11/2025): MLOps, Monitoring, and Web Interface

Objectives:

1. Implement MLOps practices, develop a web interface for predictions, and establish model monitoring.

Tasks:

1. **MLOps Implementation:**
    1. Use tools like MLflow or DVC (Data Version Control) to track experiments, manage model versions and streamline deployment pipelines.

2. **Model Monitoring:**
    1. Implement monitoring tools to track model performance over time and detect issues like model drift.
    2. Set up alerting mechanisms to be notified when model performance degrades.

Deliverables:

1. **MLOps Pipeline Documentation:** Documentation detailing the MLOps practices used for tracking experiments and managing the deployment lifecycle.
2. **Model Monitoring Setup:** A continuous monitoring infrastructure with automated alerting to ensure the model's sustained performance.

Milestone5 (28/11/2025): Final Documentation and Presentation

Objectives:

1. Complete the final documentation and create a presentation to summarize the entire project.

Tasks:

1. **Final Report:**
   1. Document the complete project, from data collection to model deployment and monitoring.
   2. Address challenges encountered, solutions implemented, and the impact of the model on real-world applications.

3. **Final Presentation:**
   1. Develop an engaging presentation to showcase the project's workflow, results, and impact on potential use cases.
   2. Provide a demonstration of the deployed model in real-time (via the web interface or API).

3. **Future Improvements:**
   1. Suggest potential improvements for the model, such as incorporating more advanced techniques or extending the functionality of the web interface.
   2. Accessibility Testing and Validation: Conduct user testing with visually impaired individuals to validate the effectiveness of the text-to-speech feature.
   3. Gather feedback on voice quality, speech speed, and overall user experience. Document accessibility compliance with standards such as WCAG (Web Content Accessibility Guidelines).

Deliverables:

1. Final Project Report: A comprehensive summary of the project, from the initial problem statement to deployment.

2. Final Presentation: A polished presentation showcasing the model's functionality and impact.

3. Future Improvement Recommendations: Suggestions for future development and enhancements to the project.
4. Accessibility Validation Report: Documentation of user testing results and accessibility compliance assessment.

KPIs (Key Performance Indicators): Please specify the key metrics for measuring the success of your project based on the following aspects.

1. Data Quality
   - o Percentage of missing values handled: 0%
   - o Data accuracy after preprocessing: 100%
   - o Dataset diversity (representation of different categories): 95%
2. Model Performance
   - o Model accuracy (Accuracy/F1-Score): Fine-Tuned model (BLIP2) 97% our model (Inception v3 + Transformer-LSTM hyprid) 89%
   - o Model prediction speed (Latency): Fine-Tuned model: 1.90 it/s, Our Model: 572 ms/step
   - o Error rate (False Positive/False Negative Rate): Our model Bleu score: 7.85, Fine-Tuned model BLEU score: 17.99
3. Deployment & Scalability
   - o API uptime: 99%
   - o Response time per request: milliseconds
   - o (If applicable) Real-time processing speed (e.g., FPS for video models): 600 ms
4. Business Impact & Practical Use
   - o Reduction in manual effort: 99%
   - o Expected cost savings: 97%
   - o User satisfaction: 95%