Image Caption Generation Using VGG16 as Encoder and LSTM as Decoder

Abstract:

The AI-based automatic image captioning project using deep learning aims to develop a system capable of generating meaningful and contextually relevant captions for images. It integrates computer vision and natural language processing, utilizing VGG16 for feature extraction and LSTMs for sequence generation. The model is trained on the Flickr8K dataset, consisting of 8,000 images with associated textual descriptions. By analysing key visual elements, the system converts them into coherent and human-like captions, improving machine understanding of visual content. The project's performance is evaluated using BLEU and ROUGE scores, ensuring accuracy and linguistic coherence. This technology has significant applications in AI accessibility, digital content indexing, and intelligent image search, making images more searchable and interpretable. Potential real-world implementations include assisting visually impaired users, enhancing automated photo organization, and improving multimedia content tagging. Future advancements will focus on refining contextual accuracy using attention mechanisms, expanding datasets for greater diversity, and optimizing model efficiency for real-time applications.

Methodology:

Dataset Selection: The Flickr8K dataset is used for model training and evaluation.

Feature Extraction: VGG16 extracts high-level visual features from input images.

Caption Generation: An LSTM-based decoder processes extracted features to generate captions.

Evaluation: BLEU and ROUGE scores assess caption accuracy and semantic relevance.

Stepwise Solution Approach:

Step 1: Preprocess the dataset, including text normalization and image resizing.

Step 2: Train VGG16 for feature extraction and LSTM for sequence modelling.

Step 3: Optimize model parameters and evaluate performance using BLEU and ROUGE metrics.

Step 4: Refine model architecture for improved caption quality and coherence.

Key Findings:

The model successfully generates relevant captions with moderate linguistic coherence.

BLEU and ROUGE scores indicate semantic alignment but highlight challenges in capturing complex linguistic structures.

Reference:

Fatima, Shan & Gupta, Kratika & Goyal, Deepti & Mishra, Suman. (2024). Image Caption Generation Using Deep Learning Algorithm. Educational Administration Theory and Practices. 30. 10.53555/kuey.v30i5.4311.

Team:

Anurag Munjal	E22CSEU1181
Sanchit Mishra	E22CSEU1183