Implementation of AI-Powered Image Captioning Using Attention and Object Features

Abstract

Image captioning is a critical task in artificial intelligence that combines computer vision and natural language processing (NLP) to generate meaningful textual descriptions for images. This project focuses on implementing an advanced image captioning system by integrating convolutional neural networks (CNNs) for feature extraction, attention-based transformers for caption generation, and object detection models like YOLOv4 to enhance semantic understanding. The model is trained on large-scale datasets such as MS COCO and Flickr30k to improve caption accuracy and fluency. This implementation aims to develop a real-time image captioning application deployed via a web-based interface.

Project Objectives

- To develop a deep learning-based image captioning system that generates human-like captions.
- To utilize **attention mechanisms** to improve the contextual relevance of generated captions.
- To integrate **object detection (YOLOv4)** for enhanced image understanding.
- To create an interactive **web-based application** for real-time caption generation.

Technology Stack

- **Deep Learning Frameworks:** PyTorch
- **Image Processing:** OpenCV, Pillow
- **Feature Extraction Models:** ResNet-50, Xception
- **Object Detection Model:** YOLOv4
- **Attention Mechanism:** Transformer-based sequence model
- **Web Deployment:** Flask, React

Implementation Methodology

The implementation of the image captioning system follows a structured pipeline:

Step 1: Data Collection & Preprocessing

- **Dataset Selection:** Use large-scale datasets like MS COCO and Flickr30k for training.
- **Feature Extraction:**
  - Extract visual features using **ResNet-50/Xception** (CNN-based encoder).
  - Detect objects using **YOLOv4** for better scene understanding.
- **Preprocessing:**
  - Resize images to a standard size (e.g., 224x224 pixels).
  - Convert captions into numerical sequences using a tokenizer.

Step 2: Model Development & Training

- **Encoder:**
  - Use a CNN (ResNet-50/Xception) to extract visual embeddings.
- **Attention Mechanism:**
  - Implement a transformer-based attention module for enhanced focus on key image regions.
- **Decoder:**
  - Use an **LSTM/Transformer** to generate captions based on extracted image features.
- **Training:**
  - Train on MS COCO with **cross-entropy loss and BLEU score evaluation**.
  - Fine-tune using **reinforcement learning** for improved caption fluency.

Step 3: Model Integration & Web Deployment

- Develop a **Flask backend** for handling image uploads and caption generation.
- Create a **React-based frontend** for user interaction.
- Deploy on a **cloud-based server** (e.g., AWS, Google Cloud).

Key Findings & Expected Outcomes

- Attention-based models significantly improve caption relevance and coherence.
- Object detection (YOLOv4) enhances semantic understanding in complex images.
- Real-time image captioning is feasible with optimized deep learning architectures

References

Al-Malla, M.A., Jafar, A. & Ghneim, N. Image captioning model using attention and object features to mimic human image understanding. *J Big Data* **9**, 20 (2022). https://doi.org/10.1186/s40537-022-00571-w

Team Members

Himanshu Dahiya(e22cseu1144)