Image Caption Generator Using CNN-LSTM Architecture

Abstract

Image captioning combines computer vision and natural language processing to generate textual descriptions for images, enabling applications in accessibility, content indexing, and social media. This project implements a CNN-LSTM encoder-decoder model using the Flickr8k dataset, focusing on feature extraction via DenseNet201 and sequence generation with LSTM. The system demonstrates foundational capabilities in mapping visual content to natural language, though challenges like overfitting and generic captions highlight opportunities for improvement.

Project Objectives

- Develop a CNN-LSTM framework for automated image captioning.
- Optimize feature extraction and sequence modeling for accuracy.
- Evaluate performance using quantitative metrics.
- Identify limitations and propose future enhancements.

Technology Stack

- Feature Extraction: DenseNet201 (pre-trained CNN)
- Sequence Modeling: LSTM, Embedding Layer
- Training Framework: TensorFlow/Keras
- Optimization: Adam optimizer, Categorical Cross-Entropy Loss
- Dataset: Flickr8k (8,000+ images, 5 captions per image)

Implementation Methodology

Step 1: Data Preprocessing

- Dataset: Flickr8k for training/validation.
- Caption Cleaning: Lowercasing, special character removal, and tokenization.
- Image Processing: Resizing, feature extraction via DenseNet201's Global Average Pooling layer.

Step 2: Model Architecture

- Encoder: DenseNet201 extracts high-level feature vectors from images.
- Decoder:
  - Embedding layer converts tokenized captions into vector representations.
  - LSTM processes sequences, generating word probabilities.
- Training:

  - Up to 50 epochs with early stopping to prevent overfitting.
  - Batch processing via data generators.

Step 3: Inference Pipeline

- Caption Generation: Predicts words sequentially using beam search, starting from "startseq" until "endseq" or max length.

Key Findings & Outcomes

- Performance: The model successfully generates captions that capture primary objects and actions in images.
- Strengths: Demonstrates the ability to learn visual-to-text mappings and produce relevant descriptions.
- Limitations:
  - Overfitting observed due to limited dataset size.
  - Occasional generic or incomplete captions in complex scenes.

Future Work

- Dataset Expansion: Train on larger and more diverse datasets (e.g., MS COCO, LAION-COCO) to improve generalization.
- Architecture Enhancements: Integrate attention mechanisms or transformer-based decoders for better context modeling.
- Evaluation: Incorporate human-in-the-loop assessments and advanced evaluation metrics.

References

- Kapuriya, Monali & Lakkad, Zemi & Shah, Satwi. (2024). Image Caption Generator Using CNN and LSTM. International Journal of Innovative Science and Research Technology(IJISRT). 1375-1382. 10.38124/ijisrt/IJISRT24AUG851.

Team Members

Himanshu Dahiya (e22cseu1144)