Image Caption Generator

1. Introduction

Image captioning is a challenging artificial intelligence task at the intersection of computer vision and natural language processing. The goal is to automatically generate a coherent and contextually relevant textual description for a given image. This capability enables machines to interpret visual content and express it in human language, facilitating applications in accessibility, content indexing, image retrieval, social media, and more[1][2][3]. The typical approach to image captioning leverages deep learning, particularly encoder-decoder architectures, where visual features extracted from images are translated into descriptive sentences.

2. Related Survey

Early image captioning methods relied on template-based or retrieval-based approaches, which were limited in flexibility and expressiveness[4][3]. The advent of deep learning introduced encoder-decoder frameworks, typically combining convolutional neural networks (CNNs) for image feature extraction and recurrent neural networks (RNNs), especially Long Short-Term Memory (LSTM) networks, for sequence generation[1][4][5][6][7][8]. Recent research has explored attention mechanisms, transformer-based models, and multimodal pretraining to further enhance caption quality and diversity[9][10][11]. Evaluation of image captioning models commonly employs automatic metrics such as BLEU, METEOR, ROUGE, CIDEr, and SPICE, though studies show these metrics correlate only weakly with human judgment, prompting interest in more comprehensive or ensemble-based evaluation methods[12][13].

3. Datasets

Several benchmark datasets are widely used for training and evaluating image captioning models:

- **Flickr8k**: Contains 8,000+ images, each annotated with five captions[14][15][5].
- **Flickr30k**: Comprises 31,000 images with five captions per image, offering greater diversity[14][15].
- **MS COCO (Common Objects in Context)**: The most popular dataset, with over 120,000 images and five human-annotated captions per image, covering 80 object categories[14][15].
- **Others**: Datasets like LAION-COCO, VizWiz-VQA, and Polaris provide large-scale or specialized benchmarks for advanced evaluation[14][12].

In this project, the Flickr8k dataset was used, with images and corresponding captions provided in structured files for training and validation.

4. Methodology

The project follows a standard deep learning pipeline for image captioning using a CNN-LSTM encoder-decoder architecture:

a) Data Preprocessing

- Captions are lowercased, cleaned of special characters and numbers, and wrapped with "startseq" and "endseq" tokens to mark sentence boundaries.
- Tokenization is applied to convert words into integer sequences suitable for input to neural networks.

b) Feature Extraction

- A pre-trained DenseNet201 CNN is used to extract high-level feature embeddings from each image. The output from the penultimate layer (Global Average Pooling) yields a fixed-length vector representation for each image.

c) Sequence Modeling

- The textual component uses an Embedding layer followed by an LSTM. At each time step, the model receives the image features and the current sequence of words to predict the next word in the caption.

d) Model Architecture

- The image feature vector is projected and concatenated with the embedded text sequence.
- The merged sequence is processed by an LSTM, followed by dense layers to generate a probability distribution over the vocabulary for the next word.
- The model is trained using categorical cross-entropy loss and optimized with Adam. Early stopping and learning rate reduction callbacks are used to prevent overfitting.

e) Training and Validation

- Data generators are used to efficiently feed image-caption pairs in batches.
- The model is trained for up to 50 epochs, with validation loss monitored to select the best model.

f) Inference

- To generate captions for new images, features are extracted, and the model predicts words sequentially, starting from "startseq" until "endseq" is produced or the maximum length is reached.

5. Results and Analysis

- The model was successfully trained on the Flickr8k dataset. Loss curves indicate some overfitting, likely due to the limited dataset size.
- Generated captions for test images generally align with the main objects and actions in the images, demonstrating the model's ability to learn visual-to-text mappings.
- Evaluation metrics such as BLEU, METEOR, and CIDEr can be used to quantitatively assess performance. In similar studies, BLEU scores for Flickr8k typically range around 0.53, with higher scores on larger datasets like MS COCO[15][13].
- Qualitative analysis shows that while the model captures the gist of most images, it occasionally produces generic or incomplete captions, highlighting the limitations imposed by dataset size and model complexity.

6. Conclusions and Future Work

This project demonstrates the feasibility of generating image captions using a CNN-LSTM encoder-decoder framework. The model effectively learns to map visual features to natural language descriptions, with reasonable performance on the Flickr8k dataset. However, challenges such as overfitting, limited vocabulary, and generic captions persist.

Future work may include:

- Training on larger and more diverse datasets (e.g., MS COCO, LAION-COCO) to improve generalization.
- Incorporating attention mechanisms or transformer-based decoders to enhance context modeling and caption diversity[9][10][11].
- Exploring data augmentation, transfer learning, and reinforcement learning strategies for better robustness and adaptability[13][9].
- Applying advanced evaluation metrics and human-in-the-loop assessments for more accurate performance measurement[12][13][2].
- Extending to multilingual or domain-specific captioning, and addressing ethical considerations such as bias in generated captions[2][10].

By advancing these directions, image captioning systems can become more accurate, context-aware, and broadly applicable across real-world scenarios.

1. https://builtin.com/articles/image-captioning
2. https://www.comet.com/site/blog/image-captioning-bridging-computer-vision-and-natural-language-processing/
3. https://bssspublications.com/PublishedPaper/Publish_412.pdf
4. https://pure.port.ac.uk/ws/files/58080375/A_Comprehensive_Review_on_Automatic_Image_Captioning.pdf
5. https://ijcrt.org/papers/IJCRT_196552.pdf
6. https://statusneo.com/image-captioning-with-resnet-and-lstm-a-powerful-deep-learning-approach/
7. https://developers.arcgis.com/python/latest/guide/how-image-captioning-works/
8. https://thesai.org/Publications/ViewPaper?Volume=11&Issue=5&Code=IJACSA&SerialNo=37
9. https://slogix.in/machine-learning/research-topics-in-image-captioning-using-deep-learning/
10. https://www.techscience.com/CMES/v142n3/59756
11. https://arxiv.org/pdf/2308.02833.pdf
12. https://arxiv.org/html/2408.04909v1
13. https://www.labellerr.com/blog/image-captioning-evaluation-and-fine-tuning/
14. https://paperswithcode.com/datasets?task=image-captioning
15. https://www.irjet.net/archives/V8/i5/IRJET-V8I5325.pdf

X-X-X-X-X