Deep Learning Term Project Report

CNN-LSTM Image Captioning Model Documentation

1. Introduction

This report presents an in-depth overview of an image captioning model based on a convolutional neural network (CNN) for feature extraction and a recurrent neural network (RNN) with LSTM units for caption generation. The model is designed to generate descriptive captions for input images by leveraging both visual and linguistic information.

2. Model Architecture

2.1 Encoder (CNN) Definition

The `Encoder_CNN` class defines the CNN-based encoder responsible for extracting features from input images using a pre-trained ResNet-50 model.

- **Model Description:**
  - The encoder utilizes a pre-trained ResNet-50 architecture to extract high-level features from images.
  - The ResNet-50 model is adapted by removing the last fully connected layer (fc) and replacing it with a linear layer (embed) to obtain image feature embeddings of a specified size (`sz_embed`).

2.2 Decoder (RNN) Definition

The `Decoder_RNN` class defines the RNN-based decoder responsible for generating captions based on extracted image features.

- **Model Description:**
  - The decoder consists of an embedding layer followed by an LSTM (Long Short-Term Memory) layer.
  - The LSTM processes embedded caption tokens along with image features to generate sequential outputs.
  - A linear layer (fc) predicts the next word in the caption sequence.

3. Training Setup

The training process involves fine-tuning the encoder-decoder model using a dataset of images paired with corresponding captions.

- **Training Steps:**
    a. Initialize the encoder and decoder models.
    b. Define optimizer and loss function (CrossEntropyLoss).
    c. Iterate over batches of image-caption pairs:
        - Pass images through the encoder to obtain feature embeddings.
        - Feed feature embeddings and caption tokens into the decoder to generate captions.
        - Compute loss based on predicted captions and ground truth captions.
        - Update model parameters using backpropagation.

4. Evaluation Using ROUGE-L Metrics

After training, the model is evaluated using the ROUGE-L metric to assess the quality and similarity of generated captions compared to ground truth captions.

- **ROUGE-L Metrics:**
    - Precision: Measures the proportion of correct words in the generated caption.
    - Recall: Measures the proportion of correct words in the reference caption.
    - F-measure: Harmonic mean of precision and recall, indicating overall caption quality.

5. Evaluation Results and Interpretation

The ROUGE-L evaluation results provide insights into the performance of the image captioning model:

- ROUGE-L Precision: 0.35

This metric measures the accuracy of generated captions in terms of word selection compared to reference captions.

- ROUGE-L Recall: 0.189

Indicates the comprehensiveness of generated captions by capturing relevant words from reference captions.

- ROUGE-L F-measure: 0.246

Represents the overall effectiveness of the model in generating descriptive and accurate captions.

6. Conclusion

In summary, this report demonstrates the implementation, training, and evaluation of an image captioning model using CNN for feature extraction and RNN for caption generation. The ROUGE-L evaluation results validate the model's ability to generate meaningful captions for images, highlighting areas for potential improvement and future research.