Image Captioning using Attention Mechanism

Week 5: Understanding Attention Mechanism & Encoder-Decoder Architecture

We are at the **last part** of our project now. We will explore the role of attention mechanisms in sequence-to-sequence models, specifically focusing on the encoder-decoder architecture for image captioning tasks.

Introduction to Attention Mechanism:

1. **Concept of Attention:**
    - Understand the fundamental concept of attention mechanisms in deep learning, particularly in sequence-to-sequence tasks.

      https://www.youtube.com/live/bHfXYQgn0Cc?feature=shared
      Attention in transformers, visually explained | Chapter 6, Deep Learning

Attention in Sequence-to-Sequence Models:

2. **Implementing Attention:**
    - Implement attention mechanisms within the encoder-decoder architecture to improve the performance of tasks such as machine translation or image captioning.

      Self Attention in Transformer Neural Networks (with Code!)

Encoder-Decoder Architecture for Image Captioning:

3. **Building an Encoder-Decoder Model:**
    - Construct an encoder-decoder architecture for generating captions from images.

      Encoder Decoder Network - Computerphile
      Sequence-to-Sequence (seq2seq) Encoder-Decoder Neural Networks, …

Problem Statement

Generate descriptive and coherent textual captions for images.The Flickr_8K dataset, with its diverse collection of images and multiple human-provided captions, serves as an ideal dataset for developing and evaluating image captioning models.

Objective

The primary goal of this project is to develop a deep learning-based image captioning system that can automatically generate accurate and meaningful captions for images. You can use the ResNet50 model for image encoding and LSTM networks for sequence modeling of captions.

Dataset Description

- **Flickr_8K Dataset**: Contains 8,000 images, each annotated with five different captions provided by humans, capturing various aspects and perspectives of the images.
- **Training, Validation, and Test Sets**: The dataset is split into three sets, with the images for each set listed in `Flickr_8k.trainImages.txt`, `Flickr_8k.testImages.txt`, and `Flickr_8k.devImages.txt` respectively.
- **Captions**: All captions are stored in `Flickr8k.token.txt`, where each image is associated with five unique captions.

Methodology

1. **Image Feature Extraction**:

   Use the pre-trained ResNet50 model to extract high-level feature representations from each image in the dataset.The output of the last convolutional layer of ResNet50 will be used as the image features.

   Preprocess the captions by tokenizing the text and creating a vocabulary of words.Use Keras embedding layers to convert the tokenized captions into dense word vectors, capturing semantic relationships between words.

2. **Sequence Modeling with LSTM**:

   Implement an LSTM network to model the sequence of word embeddings, learning to generate the next word in the caption based on the previous words and the image features.Combine the image features with the text features at an appropriate stage in the LSTM to generate contextually relevant captions.

3. **Caption Generation**:

    Utilise two decoding strategies: Greedy Search and Beam Search (with k=3) to generate captions for the images.Greedy Search involves selecting the word with the highest probability at each step, while Beam Search considers multiple possible sequences to find the most likely caption.

Expected Outcomes

Caption Quality:

Generate captions that are not only accurate but also diverse and contextually appropriate, closely resembling human-provided captions.

GitHub - AmritK10/Image_Captioning: Image Captioning using LSTM and Deep Learning on Flickr8K dataset.

Submission Instructions

Submit the code and analysis on your respective GitHub and Google Drive link as stated in the Submission Form.