Image Caption Generator with Batch Normalization

Abstract:

We develop an image caption generator by combining Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), we develop a model capable of generating descriptive captions for images. Utilizing the Flickr8k dataset, our methodology involves data preprocessing, feature extraction use of using a modified ResNet-50 CNN, caption generation with LSTM-based RNNs and the use of a layer which carries out batch normalization.

Introduction:

Automatic image captioning is a challenging task in computer vision and natural language processing, requiring the generation of descriptive and contextually relevant captions for images. The task involves not only understanding the content of the image but also generating coherent natural language descriptions.

This project focuses on developing an advanced image caption generator system using deep learning techniques. The system aims to automatically generate descriptive captions for images, enhancing the understanding of visual content by associating it with textual descriptions.

Methodology:

This project involves a series of steps which include:

- **Data Loading:** Flickr8k image dataset which contains about 8000 images, with 5 captions per image is loaded
- **Tokenization**: Items such as punctuation marks are removed, sentence is converted to lower case and unique words are extracted which characterize the image
- **Vocabulary Building:** A dictionary is constructed which maps the images with a list of 5 captions based on the frequency threshold of 5
- **Numericalization**: Assign each word of the vocabulary with a unique index value
- **Image Transformation:** Images are resized to a size of 256 x 256 and then a random crop is performed which randomly chooses a location to crop images down to a target size of 224 x 224. The images are then normalized.
- **Feature Extraction:** Features are extracted from the images by using ResNet-50, pre-trained on ImageNet images
- **Batch Normalization:** We have removed the last fully connected layer and instead added a linear layer with batch normalization to enhance model convergence

- **RNN-Decoder:** Features will be passed onto LSTM to retain relevant details while regulating the flow of information. An embedding layer will then convert captions to vectors

- **Loss Function:** Cross-entropy will measure the difference between the predicted captions generated by the RNN and the ground truth captions

- **Optimizer Function:** Adam optimizer will update the parameters of the RNN based on the gradients computed from the loss function

- **Fine Tuning and Improvements**: The model is trained and tested on multiple batches of inputs. Parameters are then fine tuned to improve accuracy of predictions and to improve the overall performance of the model

Results:

The following table depicts the model both with and without batch normalization implemented and their BLEU scores on multiple epochs:

Model/Epochs	25	30	50
With Batch Normalization	0.28	0.33	0.379
Without Batch Normalization	0.24	0.27	0.32