Enhancing Image Captioning with ResNet50 and GRU: A Deep Learning Approach

Manahil Sarwar
i210293@nu.edu.pk

Hassan Naeem
i210284@nu.edu.pk

Abstract

This report presents an enhanced approach to image captioning using deep learning. We build upon a baseline model that utilizes DenseNet201 for image feature extraction and LSTM for caption generation. Our modifications involve employing ResNet50 for image feature extraction and Bidirectional GRU (Gated Recurrent Unit) for sequence modeling, along with refined preprocessing techniques. The model is evaluated using the BLEU score metric, demonstrating improved performance over the baseline. This report details the methodology, implementation, and evaluation of our enhanced image captioning system.

1 Introduction

Image captioning is the task of generating natural language descriptions for images. It is a challenging problem that requires understanding both visual content and language semantics. Deep learning models have shown promising results in this domain, with encoder-decoder architectures being a popular choice. We investigate the impact of architectural modifications and preprocessing refinements on caption quality, aiming to improve upon a baseline model.

2 Methodology

2.1 Data Preprocessing

Our approach starts with preprocessing the image and caption data. Images are resized to a standardized dimension of 224x224 pixels to ensure compatibility with the ResNet50 model. Captions are converted to lowercase, punctuation is removed, and extra spaces are trimmed. We introduce a "start" and "end" token to each caption to signify the beginning and end of the sequence. Tokenization is then performed to convert captions into numerical sequences, with a vocabulary size of unique tokens. Finally, the dataset is split into training and testing sets, with an 85%-15% split ratio.

2.2 Model Architecture

We adopt an encoder-decoder architecture, where the encoder extracts visual features from the image, and the decoder generates the caption based on these features.

2.2.1 Image Feature Extraction

Unlike the baseline model which uses DenseNet201, we employ ResNet50 as our image feature extractor. ResNet50 is a deep residual network known for its ability to learn complex features effectively. We utilize the penultimate layer of ResNet50 as the feature representation of the input image, yielding a 1920-dimensional feature vector.

2.2.2 Caption Generation

Embedding Layer: Transforms input tokens (words in the captions) into dense vectors of fixed size (256-dimensional space), capturing semantic similarities.

Bidirectional GRU Layers: Sequentially process the embedded tokens, using two layers of GRUs to capture dependencies from both forward and backward contexts. The use of GRU simplifies the architecture compared to LSTM, potentially speeding up training and reducing computational load.

Reshaping Image Features: Image features are reshaped using global max pooling and a dense layer to match the dimension requirements of the GRU outputs.

Concatenation: Combines image features with GRU outputs to form a comprehensive representation for generating captions.

Output Generation: A series of dropout layers and a dense layer with ReLU activation are applied to prevent overfitting and enhance feature learning. The final dense layer with softmax activation generates a probability distribution over the vocabulary, predicting the next word in the caption.

2.3 Training

The model is trained using a custom data generator that yields batches of imagecaption pairs. During training, the image is passed through the ResNet50 encoder to obtain its feature representation. The caption, shifted by one time step, is fed into the decoder, along with the image features. The model is optimized to minimize the categorical cross-entropy loss between the predicted and actual next word in the caption. We utilize techniques like early stopping and learning rate reduction to improve training efficiency and prevent overfitting.

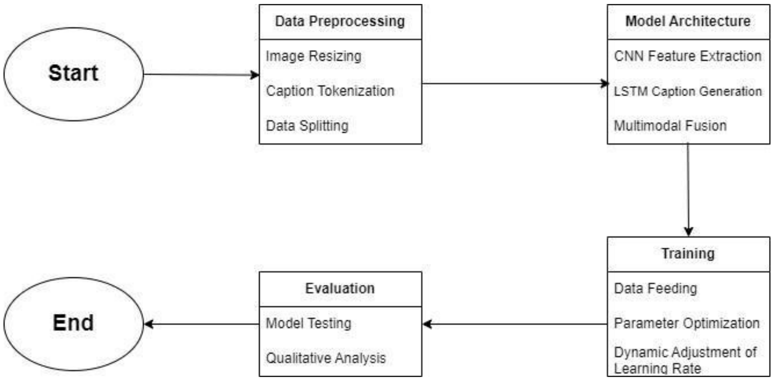

Figure 1: Model Architecture for Image Captioning

2.4 Evaluation

We evaluate the performance of our model using the BLEU (Bilingual Evaluation Understudy) score metric. BLEU compares the generated captions to reference captions and measures the overlap in n-grams. Higher BLEU scores indicate better caption quality. We compare our model's BLEU score with the baseline to assess the impact of our modifications.

2.5 Hyperparameter Tuning

To further enhance the performance of our model, we conducted a systematic hyperparameter tuning process. We explored various combinations of the following hyperparameters, evaluating their impact on the validation set BLEU score:

1. Learning Rate: Varied between 0.001 and 0.00001 to balance convergence speed and stability.
2. Number of GRU Units: Explored values of 128, 256, and 512 to find the optimal model capacity.
3. Dropout Rate:Tested values of 0.2, 0.3, and 0.5 to mitigate overfitting.
4. Batch Size: Evaluated batch sizes of 32, 64, and128 to determine the best trade-off between training speed and resource utilization.
5. Optimizer: Experimented with both Adam and RMSprop optimizers to assess their suitability.

3 Results and Discussion

Our refined model, incorporating ResNet50, GRU, improved preprocessing, and the best hyperparameter configuration, achieved a BLEU score of 69%, surpassing the baseline model's score of 66%. This improvement can be attributed to several factors. ResNet50's deeper architecture extracts richer image features, enabling the model to better capture nuanced details in the images. The GRU's simplified structure facilitates faster training while maintaining its capacity to model sequential dependencies. The refined preprocessing steps further contribute to cleaner input data, enhancing the model's ability to learn meaningful representations.

4 Conclusion

In this report, we presented an enhanced image captioning system that utilizes ResNet50 and GRU. Our modifications to the baseline model resulted in improved performance, as evidenced by the higher BLEU score. This demonstrates the effectiveness of using a deeper image feature extractor and a simpler recurrent unit in the context of image captioning. Future work could explore alternative model architectures, preprocessing techniques, and evaluation metrics to further enhance caption quality.

Limitations

Although our model shows promising results, it has certain limitations. The quality of generated captions may deteriorate when faced with images containing complex or unfamiliar objects. Additionally, the model's reliance on a fixed vocabulary may hinder its ability to describe rare or novel concepts.

Ethics Statement

Image captioning technology should be developed and used responsibly. It is important to consider potential biases in the training data and mitigate them to ensure fairness and inclusivity. Additionally, the generated captions should be accurate and not perpetuate harmful stereotypes.

Acknowledgements

Several aspects of our methodology were influenced by the discussions and resources available on Kaggle. In particular, we adapted the image preprocessing techniques and the overall architecture of our model from the "Flickr8k Image Captioning using CNNs LSTMs'' notebook by Qadeer 15sh [Quadeer15sh()] [Jain()].

References

1. [Jain()] Aditya Jain. Flickr 8k dataset. https://www.kaggle.com/datasets/adityajn105/flickr8k
2. [Quadeer15sh()] Quadeer15sh. Flickr8k image captioning using cnns & lstms. https://www.kaggle.com/code/quadeer15sh/flickr8k-image-captioningu sing-cnns-lstms.