Caption Generation for Images: PicInfo

Anirudh Prashant Kalghatkar, Sachin Kashinath Rathod
Department of Computer Science
University of Colorado Boulder

Abstract

Image captioning involves generating human-readable descriptions or sentences that accurately depict the content of an image. In this study, we propose an image caption generation utilizing two NLP techniques Greedy Search(argmax) and Beam Search. Convolutional Neural Networks (CNN), specifically the InceptionV3 model is used for image feature extraction. The combination of CV and NLP techniques are applied to a Facebook public multimodal dataset (PMD) comprising 566,747 training images and 25,010 test images. From this dataset, 16,000 training images are utilized for feature extraction and training the Natural Language Processing (NLP) model, while 1,600 test images are employed for feature extraction and testing the NLP model. The image caption generation techniques (Greedy Search and Beam Search) are compared based on BLEU (Bilingual Evaluation Understudy), ROUGE-L and METEOR score. Through comprehensive experimentation and evaluation, this study aims to provide insights into the comparative performance of different caption generation approaches.

1 Introduction

Image captioning, also known as photo captioning, refers to the automated process of generating textual descriptions for an image. This innovative technique encapsulates the content of an image in textual form, enabling a deeper understanding and interpretation of visual data. Leveraging a combination of natural language processing techniques such as Greedy Search and Beam Search, alongside computer vision model like InceptionV3 for feature extraction, image captioning algorithms strive to accurately depict the visual content of images through descriptive text.

The task of image captioning holds significant importance across various domains due to its wide range of applications. In intelligent transportation, image captions can aid in analyzing traffic patterns, identifying road hazards, and enhancing navigation systems. In network image analysis, captions can facilitate the understanding of complex visual data for cybersecurity purposes. Moreover, in the medical field, image captions can provide guidance to practitioners during diagnosis, treatment planning, and medical research. Additionally, image captions play a crucial role in assisting visually impaired individuals by providing them with a textual description of their surroundings, thus enabling greater independence and access to visual information. Overall, image captioning has far-reaching implications in improving safety, accessibility, and efficiency across various sectors of society.

2 Related Work

The paper [1] proposes a model built upon the encoder-decoder framework, introducing enhancements in both feature extraction and decoding processes. Specifically, the encoder utilizes a ResNest network architecture augmented with a Squeeze-and-Excitation module to extract more informative image features. The decoder incorporates a two-layer long short-term memory (LSTM) architecture with multi-head attention mechanisms for improved understanding of feature relationships and generation of accurate text descriptions. Based on the findings of this study, the ResNest network architecture utilized in the proposed model demands significant computational resources. Considering resource constraints, we opted to integrate InceptionV3. This decision was motivated by the renowned capability of InceptionV3 to balance computational power while still effectively extracting informative image features, aligning with our project's objectives and constraints.

This paper [2] introduces a novel approach to image captioning, employing an LSTM-based language model alongside a pre-trained CNN and semantic keywords extraction module for feature extraction. This methodology significantly enhances caption efficiency by accurately describing objects and integrating semantic labels. Moreover, a facial recognition system is integrated to identify and recognize celebrity faces, enabling the generation of personalized captions by replacing instances of individuals with their names. Evaluation metrics such as BLEU and METEOR scores are utilized to assess caption precision. Building upon these findings, our study proposes the integration of GRU in the decoder architecture to address the complexities associated with LSTM architecture and the vanishing gradient problem inherent in RNNs.

The proposed model [3] leverages a combination of convolutional neural networks (CNNs), specifically InceptionV3, for feature extraction, and recurrent neural networks (RNNs), particularly the GRU architecture, for generating text from these features using Greedy search (argmax). Additionally, the model incorporates an attention mechanism during caption generation to improve contextual relevance. Evaluation of the model is conducted on the MSCOCO database, demonstrating its efficacy in generating natural language descriptions of images. However, it is worth noting that the evaluation of the model on the MSCOCO database lacks the utilization of metrics such as BLEU score to assess the quality of generated captions. Additionally, the study does not compare the performance of the proposed Greedy search approach with other text generation techniques, such as Beam search. Incorporating these evaluation metrics and comparative analyses could provide a more comprehensive assessment of the model's effectiveness and contribute to further insights into image captioning methodologies.

3 Methodology

3.1 Work Expectations

In this study, we compare the two NLP techniques that is Greedy Search and Beam Search and expect that Beam Search will outperform the Greedy Search as Beam Search offers more diverse and higher-quality results, because it selects top-k predictions iteratively, and ultimately selects the caption with the highest probability based on the model's output.

3.2 Dataset Preprocessing

The Public Multimodal Dataset (PMD) utilized in our study comprises publicly available image-text pair datasets, with a subset derived from COCO containing 566,747 training images and 25,010 test images. From this subset, 16,000 training images and 1,600 test images were selected for analysis. Preprocessing of the dataset involved downloading images using the Python library urllib from the provided image URLs. To standardize the input for the InceptionV3 model, which accepts images of size 299 x 299, the downloaded images were resized accordingly. This preprocessing step ensured consistency in input dimensions for subsequent feature extraction processes.

3.3 Feature Extraction

Feature extraction was conducted utilizing the InceptionV3 model, selected due to its efficiency in computation and resource utilization. Given the constraints posed by limited resources, the pretrained InceptionV3 was employed for feature extraction. This choice was driven by its ability to extract informative features from images while demanding less computational power, thus ensuring optimal performance within the available resource constraints.

3.4 Tokenization

Tokenization of the training and test image captions was accomplished using the TensorFlow tokenizer, a robust tool for text preprocessing. This process involved converting the raw textual data into a sequence of tokens, allowing for efficient handling and analysis of the caption texts. Following tokenization, two mappings were created: id2word, which maps token indices to corresponding words, and word2id, which performs the reverse mapping. These mappings facilitated the conversion between token indices and their corresponding words, enabling seamless integration of textual data into subsequent stages of the image captioning pipeline.

3.5 Model Architecture

In our model architecture, the encoder employs a CNN architecture with a Dense layer as a fully connected network, complemented by a Rectified Linear Unit (RELU) activation function.

This configuration efficiently extracts features from the features extracted by InceptionV3. Transitioning to the decoder, we integrate a Bahdanau attention mechanism, facilitating alignment of relevant image regions with corresponding words in the caption. This attention mechanism enhances the model's ability to generate accurate and contextually relevant descriptions. The decoder further incorporates an embedding layer to process decoder inputs, propagating generated embeddings to a GRU layer. The output from the GRU layer is then passed through a fully connected dense layer, with input size corresponding to the GRU units. Subsequently, a second fully connected dense network, with the vocabulary size as input, generates the predicted caption. This process iterates recursively, with the predicted caption serving as input to the decoder until the highest probable sentence is generated. Figure 1 depicts the above architecture.

3.6 Image Caption Generation

For image caption generation, our model employs two distinct methods: Greedy Search and Beam Search. In Greedy Search, the caption is predicted by iteratively selecting the token with the maximum probability from the predicted vector until either the "<end>" token is encountered or the maximum caption length is reached. Conversely, in Beam Search, the model generates multiple candidate captions (with a beam width of 3), refines them by iteratively selecting the top-k predictions, and ultimately selects the caption with the highest probability based on the model's output. From the output obtained image caption generation process continues until either the "<end>" token is encountered or the maximum caption length is reached.

3.7 Evaluation

In evaluating the performance of our image captioning model, we employed three key metrics: BLEU score, METEOR score, and ROUGE_L. These metrics provided quantitative measures to assess the quality and similarity of generated captions to reference captions. Additionally, manual inspection was conducted to qualitatively evaluate the captions, focusing on their resemblance to human-generated descriptions. By combining quantitative metrics with qualitative assessment, we obtained a comprehensive understanding of our model's performance, ensuring robust evaluation and validation of the

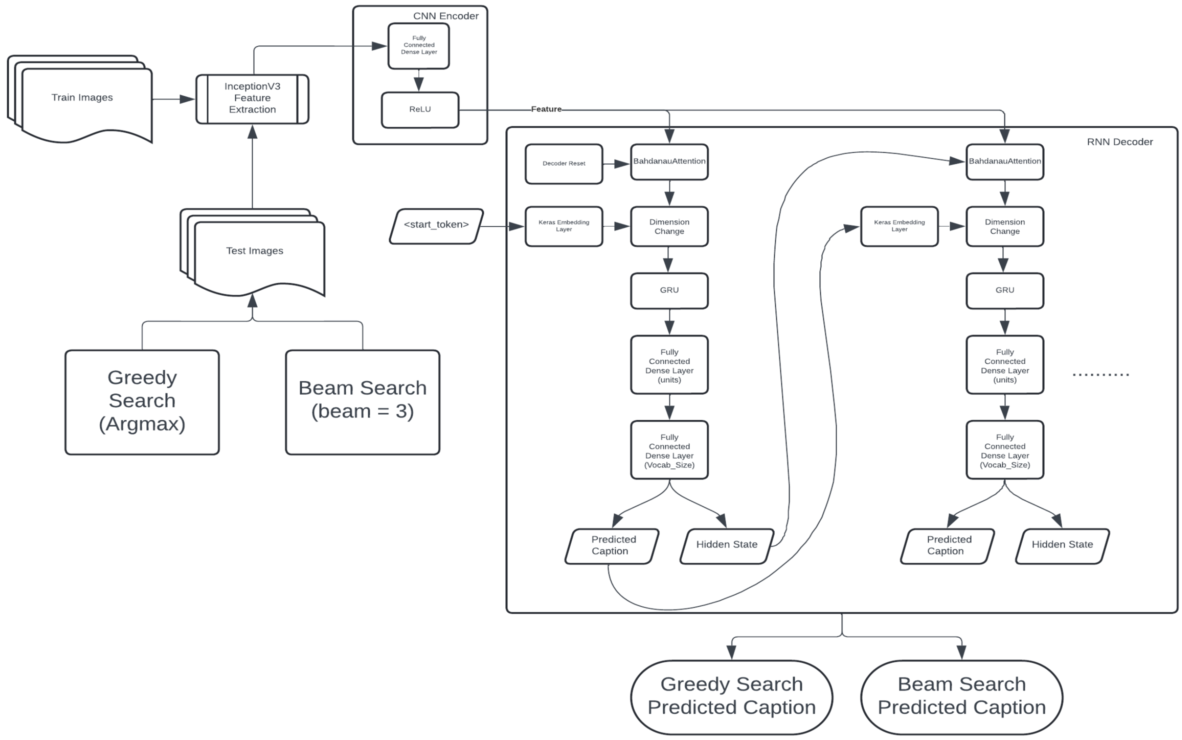

Figure 1: Model Architecture

generated captions in terms of both linguistic accuracy and semantic coherence.

4 Experiments

4.1 Quantitative Results

The model was trained over 10 epochs, converging to a final loss of 0.007312, as illustrated in Figure 2. Employing the PMD COCO dataset, encompassing 16,000 training images, and 1,600 test images, image captioning was executed using both Greedy Search and Beam Search methodologies. Evaluation metrics such as BLEU score, METEOR score, and ROUGE_L were utilized to gauge the efficacy of these NLP techniques. The findings, outlined in Table 1, underscore the proficiency of the NLP techniques in generating image captions, elucidating their performance across diverse evaluation criteria.

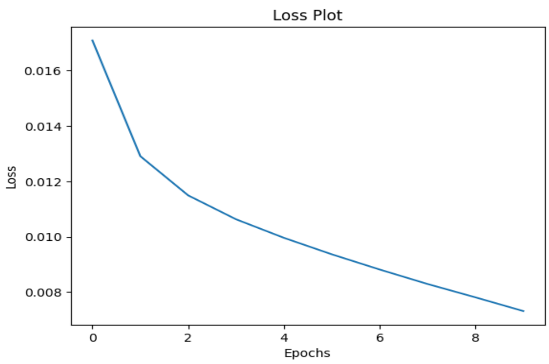

Figure 2: Epoch v/s Loss Plot

Technique	BLEU	METEOR	ROUGE_L
GREEDY ARGMAX	0.068	0.288	0.857
BEAM SEARCH	0.053	0.236	0.777

Table 1: Evaluation Metric

4.2 Qualitative Results

In Figure 3, an illustration from the PMD dataset is presented. Here, a comparative analysis between Greedy Search and Beam Search is depicted. While Greedy Search accurately identifies the scene as "a man riding a skateboard down a road," Beam Search predicts it as "a skateboard down a street." This exemplifies a limitation of Beam Search, indicating its tendency to produce sentences that may lack human-like fluency and context.

Figure 3:

Real Caption: A young man riding a skateboard down a street.

4.3 Findings

For the evaluation conducted on the 1,600 test images, Greedy Search exhibited a notably faster processing time, completing caption generation within 5 minutes, compared to Beam Search which took 65 minutes. In terms of evaluation metrics, Greedy Search yielded a BLEU score of 0.068, METEOR score of 0.288, and ROUGE_L score of 0.857. In contrast, Beam Search obtained a lower BLEU score of 0.053, METEOR score of 0.236, and ROUGE_L score of 0.777. These results underscore the trade-off between computational efficiency and caption quality, with Greedy Search demonstrating relatively better performance across all metrics compared to Beam Search.

4.4 Limitations

It's worth noting that our study has certain limitations that should be acknowledged. Firstly, we restricted the dataset to a subset consisting of 16,000 train images and 1,600 test images from the larger dataset comprising 566,747 train images and 25,010 test images. Additionally, to mitigate computational demands, we limited the training epochs to 10 and constrained the Beam search width to 3. While these measures were necessary for practical reasons, they may have impacted the model's overall performance and generalizability.

5 Conclusion

In conclusion, our study reveals that Greedy Search (Argmax) outperformed Beam Search in generating image captions, exhibiting both faster processing times and more human-like outputs. While Beam Search theoretically offers the potential for more diverse and higher-quality results, our findings suggest that Greedy Search yielded superior outcomes in this context. This unexpected result may be attributed to various factors, including training the model on a smaller dataset, limiting epochs to 10 due to resource constraints, and employing a lower beam width of 3. Moving forward, further investigation is warranted to better understand the interplay between search algorithms, dataset size, and training parameters in image captioning tasks.

6 Future Work

In future work, we aim to enhance the performance of our image captioning model by exploring several avenues. Firstly, training the model on a larger dataset could lead to improved generalization and reduced overfitting, thereby enhancing the model's ability to generate accurate and diverse captions. Additionally, increasing the number of epochs beyond the current limitation of 10 would allow for more comprehensive learning, potentially resulting in further improvements in model performance. Moreover, we plan to experiment with higher beam width values, such as 7 or 10, in Beam search to explore its impact on caption quality and diversity. By systematically investigating these avenues, we anticipate gaining deeper insights into the factors influencing image captioning model performance and refining our approach to achieve even better results.

7 References

1. Rongrong et al. An Image Captioning Model Based on SE-ResNest and EMSA

2. Abisha et al Semantic Driven CNN -LSTM Architecture for Personalised Image Caption Generation

3. Ansar Hani et al. Image Caption Generation Using a Deep Architecture