International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

An Improved Image Captioning Approach

Raghad Al-Misned[1], Mohammed Al-Hagery[2]

[1]Qassim University, College of Computer, Department of Computer Science, Buraydah, Saudi Arabia
Email: *raghadabdulaziz123[at]gmail.com*

[2]Qassim University, College of Computer, Department of Computer Science, Buraydah, Saudi Arabia
Email: *drmalhagery[at]gmail.com*

Abstract: ***Image captioning and the task of automatically generating descriptive captions for images has gained significant attention in recent years due to its wide-ranging applications. These applications include; content accessibility, content indexing, and automated content generation. This proposal will help explore the intersection of Language-Image Pre-training approaches and image captioning, aiming to develop a robust model capable of generating accurate and contextually relevant captions for diverse images. This is achieved by leveraging the power of a state-of-the-art vision-language pre-training model. The proposed approach aims to set new benchmarks in image captioning by leveraging innovative techniques and advancing the integration of vision and language models. We propose a novel architecture and methodology, including advanced attention mechanisms and multimodal fusion techniques, to enhance captioning performance and improve the understanding of visual content by machines. This can be achieved through comprehensive experimentation and evaluation on benchmark datasets, to demonstrate the effectiveness and practical utility of our proposed approach. Our findings will not only contribute to the advancement of image captioning technology by combining a pre-trained vision-language model with innovative strategies like Parameter-Efficient Fine-Tuning (PEFT), and it will hold significant implications for various real-world applications, including assistive technology, content indexing, and automated content generation in a number of different domains.***

Keywords: Natural Language Processing, Computer Vision, Image Captioning, Deep Learning.

1. Introduction

The advent of multimodal machine learning, integrating vision and language, has opened new frontiers in artificial intelligence. Among the key challenges in this domain is image captioning—the task of generating coherent and contextually accurate textual descriptions of images. Image captioning has gained substantial attention due to its wide-ranging applications, including enhancing content accessibility for visually impaired individuals, indexing multimedia content for efficient retrieval, and automating content generation in various industries [1], [2]. Despite the advances, achieving robust and context-aware captioning remains challenging, primarily due to the semantic gap between visual and textual modalities.

The emergence of vision-language pre-training models, such as Bootstrapping Language-Image Pre-training (BLIP), has significantly improved the performance of multimodal tasks by leveraging large-scale paired image-text data [3]. These models provide a robust foundation for addressing challenges in vision-language integration by enhancing the alignment between visual features and textual representations. However, further research is required to optimize their performance for specific tasks like image captioning. Challenges such as generating captions for complex or ambiguous scenes, managing diverse linguistic styles, and improving multimodal fusion require innovative approaches.

This study proposes advancing image captioning through enhanced language-image pre-training approaches, focusing on refining BLIP-based methodologies. By incorporating advanced attention mechanisms and multimodal fusion strategies, the research aims to improve captioning accuracy and contextual relevance. Comprehensive experimentation on benchmark datasets will validate the effectiveness of the proposed approach, contributing significantly to the broader fields of computer vision (CV) and natural language processing (NLP). Additionally, the findings will have practical implications for real-world applications, including assistive technologies, content indexing, and automated content generation.

2. Related Work

The field of image captioning has evolved rapidly, with numerous studies contributing to its progress. This section reviews key contributions, focusing on vision-language pre-training, multimodal integration, and advanced architectures.

In early advancements, Karpathy and Fei-Fei [4] introduced deep visual-semantic alignments for image captioning, employing a model that aligns image regions with sentence fragments. This work laid the foundation for neural approaches to image captioning. Xu et al. [5] extended this idea by introducing attention mechanisms in their “Show, Attend, and Tell” model, enabling the generation of more contextually relevant captions by focusing on salient regions of an image.

Anderson et al. [6] proposed bottom-up and top-down attention mechanisms, combining object detection with top-down contextual reasoning. This approach significantly improved the interpretability and performance of captioning models. Following this, Chen and Dolan [7] emphasized the importance of high-quality datasets, which are critical for training robust models.

Recent advancements in vision-language pre-training have further revolutionized image captioning. Radford et al. [8] introduced CLIP, which leverages natural language supervision for learning transferable visual representations. Building on this, Li et al. [3] proposed BLIP, a unified framework for vision-language tasks that demonstrated state-

DOI: https://dx.doi.org/10.21275/MS25429090655

of-the-art performance across multiple benchmarks. Zhou et al. [9] introduced UNITER, which employs pre-training on large-scale datasets for universal image-text representations.

Transformers have also played a pivotal role in advancing image captioning. Dosovitskiy et al. [10] demonstrated the effectiveness of transformers in visual tasks, inspiring their adoption in multimodal applications.

3. Methodology

The methodology for enhancing image captioning through the proposed VLP (Vision-Language Pre-training) approach integrates advanced techniques to optimize the model's ability to generate accurate and contextually relevant captions for images. We refine the model's image-text alignment by combining a pre-trained vision-language model with innovative strategies like Parameter-Efficient Fine-Tuning (PEFT). This enables the model to better understand and generate captions that are closely aligned with the visual content. The overall objective is to enhance the model's performance on the well-established Flickr30K dataset, utilizing cutting-edge fine-tuning techniques to bridge the gap between visual and textual information and improve the accuracy of automated caption generation.

The methodology can be explained in different steps and sub steps, as follows:

1) **Dataset Selection & Preprocessing:** Choose a widely used image captioning dataset Flickr30K. Images are scaled, normalized, and transformed to RGB format, whereas captions are converted into tokens and padded.
2) **Pretraining Language-Image Model:** Use a vision-language model as the base. The architecture includes an image encoder, a text decoder, and a caption decoder to fuse image-text embeddings and generate relevant captions.
3) **Enhanced Pretraining Technique:** Improve how the model relates images to their captions by using Parameter-Efficient FineTuning (PEFT) techniques.
4) **Model Fine-Tuning:** The enhanced BLIP model is fine-tuned in experimental setup.
5) **Caption Generation:** Automatically creating descriptive text based on the content of an image.
6) **Evaluation Metrics:** Assess the model's performance through evaluation metrics. Figure 1 depicts the process flow of the methodology.

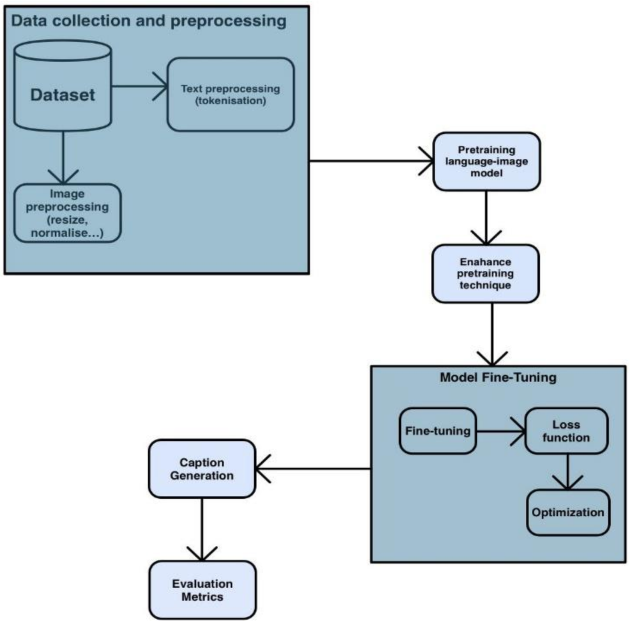
Data collection and preprocessing
Dataset
Text preprocessing (tokenisation)
Image preprocessing (resize, normalise...)
Pretraining language-image model
Enahance pretraining technique
Model Fine-Tuning
Fine-tuning
Loss function
Optimization
Caption Generation
Evaluation Metrics

Figure 1: Implementation Process

4. Experiments Application

In this section, we outline the experimental application used to evaluate the performance of the enhanced BLIP model with LoRA integration. The experiments are designed to test the impact of progressive fine-tuning across different subsets of the Flickr8K and Flickr30K datasets. The approach follows a staged fine-tuning process, where the model is iteratively exposed to different datasets to assess its learning capability and improvement. Each experiment explores the effect of training the model on a combination of images and captions from both Flickr8K and Flickr30K, with the fine-tuning process occurring in multiple stages. We first present the initial model incorporating LoRA. This is followed by additional fine-tuning stages on new and repeated data from the Flickr30K dataset. The goal of these experiments is to determine how the model adapts and improves with each fine-tuning iteration, ultimately contributing to better image caption generation.

a) Data Preprocessing

In the context of this study, data preparation is a crucial step to ensure the model receives high-quality inputs compatible with the pretrained BLIP model. For image preprocessing, all images are resized to a fixed size of 384×384 pixels, converted to RGB format if necessary, rescaled to a [0, 1] range using a factor of 1/255, and normalized by subtracting the mean and dividing by the standard deviation to match the BLIP model’s expected input format. For text preprocessing, captions are tokenized by breaking them into smaller units such as words or subwords, which are then converted into numerical IDs to allow the model to process the textual input effectively.

b) Experimental Application

In the following table, we provide an overview of the experimental configurations for each model, detailing the fine-tuning process and the datasets used at each stage. The table outlines the step-by-step fine-tuning procedure, where each model is initially trained on a subset of the Flickr8K dataset, then saved, loaded, and subsequently fine-tuned on additional subsets of the Flickr30K dataset. This staged approach allows for a detailed comparison of how each fine-tuning step impacts the model's performance. Each row represents a different model experiment, highlighting the specific datasets and the order in which they were applied during the fine-tuning process.

Table 1: Fine-Tuning Stages and Dataset Configuration for Each Model

Model Name	Fine-Tuning Stage 1	Fine-Tuning Stage 2
Model 1	1400 images (7000 captions) from Flickr8K	3000 images (15000 captions) from Flickr30K
Model 2	(Model 1 Loaded)	3000 new images (15000 captions) from Flickr30K
Model 3	(Model 2 Loaded)	3000 same images from Flickr30K (as Stage 2)

All models in the experiments were fine-tuned using a learning rate of 1e-5 for two epochs at each stage. The fine-tuning was performed with a consistent Low-Rank Adaptation (LoRA) configuration to enable parameter-efficient updates. The LoRA settings used across all models were: r = 16, lora_alpha = 32, lora_dropout = 0.05, bias = "none", and target_modules = ["query", "value"]. This setup was selected to effectively adapt the pre-trained BLIP model while minimizing computational overhead and maintaining model efficiency.

5. Results Analysis and Comparison

The evaluation metrics for the baseline model and the fine-tuned models show notable differences in performance, with each model showing improvements in various metrics. Table 2 contains the results for the Baseline Model, Model 1, Model 2, and Model 3 across the BLEU and CIDEr scores.

Table 2: Performance Comparison of Baseline and Fine-Tuned Models

Model	BLEU@1	BLEU@2	BLEU@3	BLEU@4	CIDEr
Baseline	0.53	0.41	0.30	0.21	0.43
Model 1	0.57	0.39	0.27	0.19	0.55
Model 2	0.70	0.52	0.38	0.26	0.57
Model 3	0.69	0.51	0.37	0.25	0.56

The Baseline Model achieved a BLEU@1 score of 0.53, BLEU@2 of 0.41, BLEU@3 of 0.30, BLEU@4 of 0.21, and a CIDEr score of 0.43. These values set a baseline for the captioning performance, reflecting a moderate ability of the model to generate relevant image captions.

Model 1, after fine-tuning with LoRA and the dataset from both Flickr8K and Flickr30K, demonstrated a slight improvement over the baseline in some of the BLEU scores. The BLEU@1 score increased to 0.57, while the BLEU@2 score dropped to 0.39. While BLEU@3 and BLEU@4 scores also decreased compared to the baseline, the CIDEr score improved to 0.55, suggesting a better match between the generated captions and the reference captions in terms of informativeness.

Model 2, which continued fine-tuning the previous model by adding additional training on another batch of Flickr30K images, showed more substantial improvements. The BLEU@1 score rose to 0.70, and BLEU@2 increased to 0.52. BLEU@3 and BLEU@4 also showed improvements, reaching 0.38 and 0.26, respectively. Additionally, the CIDEr score reached 0.57, indicating that the model has generated captions that are more semantically accurate and closer to human references.

Model 3, which was further fine-tuned on the same Flickr30K dataset used in Model 2, showed a slight drop in the BLEU@1 score (0.69) compared to Model 2. However, the performance in other metrics remained close to Model 2, with BLEU@2 at 0.51, BLEU@3 at 0.37, and BLEU@4 at 0.25. The CIDEr score was 0.56, which is slightly lower than Model 2, but still showed overall improvements compared to the baseline.

In conclusion, fine-tuning with the additional images from Flickr30K significantly enhanced the model's ability to generate accurate and descriptive captions. While the improvements in BLEU scores were more prominent in the first stages of fine-tuning (Model 2), further fine-tuning on the same dataset (Model 3) did not show substantial gains, suggesting that the model had reached its peak performance in terms of generating captions based on the given dataset. These results indicate that while fine-tuning provides substantial performance improvements, the model's ability to continue improving becomes less pronounced once the data has been fully utilized.

6. Conclusion

This study enhanced image captioning by integrating LoRA into the BLIP architecture, aiming to improve caption quality while maintaining computational efficiency. The model was fine-tuned on a subset of the Flickr8k and Flickr30k datasets, with systematic variations in fine-tuning stages. Results showed significant improvements, with Model 2 achieving notable increases in BLEU and CIDEr scores, including BLEU@1 0.70 and CIDEr 0.57. Using a learning rate of 1e-5 over two epochs and optimized LoRA settings, the model delivered better performance with minimal computational cost. This study underscores the practicality of employing efficient fine-tuning strategies for real-world image captioning systems, from accessibility solutions to automated media indexing.

7. Future Work

Future work can focus on experimenting with different vision-language pretraining (VLP) models to comprehensively evaluate various architectures. Expanding the model’s capability to generalize across diverse datasets will be essential to ensure robustness and scalability. Another promising direction is improving caption accuracy by experimenting with different fusion techniques to better integrate image and text representations. Further, modifications to the architecture, such as introducing novel attention mechanisms or integrating additional neural network layers, could potentially enhance both the model’s performance and the quality of generated captions.

References

[1] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," Proc.

Paper ID: MS25429090655 DOI: https://dx.doi.org/10.21275/MS25429090655

IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3128–3137, 2015.

[2] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," Proc. Int. Conf. Mach. Learn. (ICML), pp. 2048–2057, 2015.

[3] X. Li et al., "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 9839–9850, 2022.

[4] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3128–3137, 2015.

[5] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," Proc. Int. Conf. Mach. Learn. (ICML), pp. 2048–2057, 2015.

[6] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI), vol. 40, no. 4, pp. 1004–1016, 2018.

[7] X. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," *Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL)*, pp. 190–195, 2011.

[8] A. Radford et al., "Learning transferable visual models from natural language supervision," Proc. Int. Conf. Mach. Learn. (ICML), pp. 8748–8757, 2021.

[9] J. Zhou et al., "UNITER: Learning universal image-text representations," Proc. Eur. Conf. Comput. Vis. (ECCV), pp. 104–120, 2020.

[10] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," Proc. Int. Conf. Learn. Represent. (ICLR), 2020.