Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 2 (2024)

Automatic Image Captioning Using Deep Learning

Adnan Khan[a], Gulshanovar Chauhan[b], Adil Khan[c,] Harshit Chaudhary[d], Md Shahid

Department of CSE, Meerut Institute of Engineering & Technology, Meerut 250001,India

Abstract: This study employs sophisticated Deep Learning techniques to develop a robust Automatic Image Captioning model, integrating Convolutional Neural Networks (CNNs) for intricate feature extraction and Long Short-Term Memory networks (LSTMs) for nuanced sequence generation. Aimed at addressing the surge in online visual content, the technology facilitates effective image interpretation with applications spanning accessibility enhancement for the visually impaired, improved content indexing for search capabilities, and heightened social media engagement through contextually relevant image captions. The research contributes valuable insights to computer vision, tackling challenges in generating coherent image descriptions. The meticulously tuned model undergoes both quantitative and qualitative evaluation, showcasing promising outcomes for innovative applications in content retrieval and human-computer interaction. Ultimately, this research aspires to advance automatic image understanding, promoting enhanced accessibility to visual information and propelling progress in artificial intelligence.

***keywords*:** Deep Learning, CNN's, LSTM's, Image Understanding, Computer Vision, Sequence Generation.

Introduction

1.1 INTRODUCTION

This Paper delves into **Automatic Image Captioning**, employing advanced **Deep Learning techniques** to craft a model proficient in autonomously generating coherent and contextually relevant image captions. Utilizing **Convolutional Neural Networks (CNNs)** for robust feature extraction and **Long Short-Term Memory networks (LSTMs)** for intricate sequence generation, the primary goal is to grasp nuanced relationships within diverse images.

1.2 HISTORY & BACKGROUND

Automatic Image Captioning has witnessed significant evolution, notably propelled by seminal work in computer vision. The integration of **Convolutional Neural Networks (CNNs)** gained momentum in the **mid-2010s**, enhancing robust feature extraction. Concurrently, the development of **Long Short-Term Memory networks (LSTMs)** further advanced sequence generation capabilities. Pioneering research in image recognition dates back to the early **2000s**, laying the groundwork for subsequent breakthroughs. This paper builds upon this historical trajectory, aiming to push the boundaries of image understanding.

1.3 MAIN COMPONENTS & TECHNOLOGIES

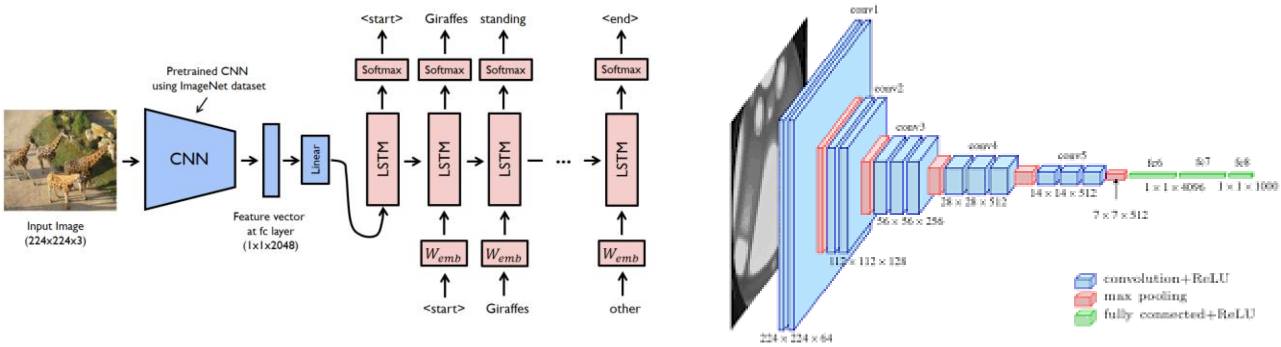
This model utilizes a robust algorithm for Automatic Image Captioning, integrating **Convolutional Neural Networks (CNNs)** with **Long Short-Term Memory networks (LSTMs).** CNNs employ multiple convolutional layers for intricate feature extraction, while LSTMs handle sequential information, capturing contextual relationships within images. The model undergoes training on a diverse dataset, leveraging optimization techniques and fine-tuning hyperparameters for enhanced performance. Evaluation **metrics encompass accuracy, BLEU scores, and qualitative assessment through sample image-caption pairs,** ensuring a comprehensive and technically sound solution for advanced image understanding and captioning

1. Proposed Work Plan

2.1 ARCHITECTURE OF USED TECHNOLOGIES

- The architecture integrates Convolutional Neural Networks (CNNs) for image feature extraction and Long Short-Term Memory networks (LSTMs) for sequential caption generation. Raw images are processed through the CNN encoder and the resulting features are used by LSTM

2.2 DESCRIPTION OF MODULES



CNN-LSTM Model:

- **Encoder (CNN)**: Utilizes Convolutional Neural Networks to extract hierarchical features from input images. This encoder generates a **fixed-size vector representation.**
- **Decoder (LSTM)**: Employs Long Short-Term Memory networks to process the vector representation from the encoder and generate **sequential captions word by word.**

Transformer-Based Model:

- **Encoder (Transformer)**: Leverages the Transformer architecture for both image and text processing. This self-attention mechanism allows the model to capture **long-range dependencies** in both modalities efficiently.
- **Decoder (Transformer)**: Similar to the encoder, the decoder employs self-attention mechanisms to **generate captions sequentially** based on the input image features decoder to generate descriptive captions

2.3 ALGORITHM

Automatic Image Captioning Algorithm:

1. Image Feature Extraction (CNN):

- Utilize **Convolutional Neural Networks (CNNs)** to extract hierarchical features from input images.

2. Sequence Generation (LSTM):

- Implement **Long Short-Term Memory networks (LSTMs)** to process the features and generate sequential captions.

3. Training:

- Optimize the model's parameters using a suitable loss **function and backpropagation.**
- Iterate through the dataset for multiple epochs **to improve accuracy.**

4. Evaluation:

- Assess the model's performance on a **separate test dataset.**

- Measure metrics such **as accuracy, BLEU scores, and qualitative analysis** of generated captions.

5. Output:

- Generate descriptive captions for new input images using the **trained model.**

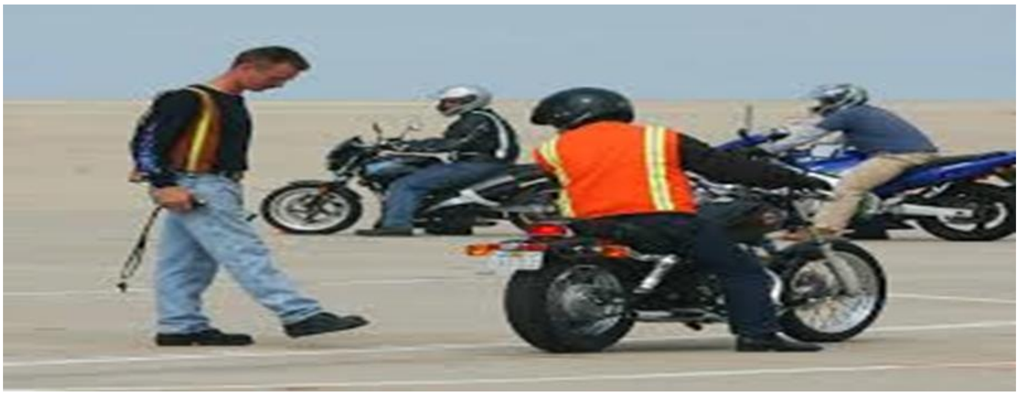
This algorithm encapsulates the core processes of the Automatic Image Captioning system, **combining the strengths of CNNs for image understanding and LSTMs** for sequential information processing.

3. Description of data set used

- In this project, **the Flickr8k dataset** played a pivotal role in training and evaluating an Automatic Image Captioning system. Leveraging the diversity of **8,000 images**, each paired with **five human-generated captions**, the dataset provided a rich and varied set of visual and textual data for model development.
- The Convolutional Neural Network (CNN) component of the project excelled at feature extraction from the images. Utilizing the **hierarchical features learned by the CNN**, the **Long Short-Term Memory networks (LSTMs) were adept at processing sequential information to generate coherent and contextually relevant captions.** The model underwent extensive training on the Flickr8k dataset, optimizing parameters through backpropagation and minimizing a suitable loss function.
- The diverse scenes and complex relationships present in the **Flickr8k images posed both challenges and opportunities.** The CNN-LSTM architecture demonstrated a capacity to understand and articulate the nuanced connections between visual elements and textual descriptions. **Evaluation metrics, including accuracy and BLEU scores, reflected the model's proficiency in generating captions that aligned closely with human-generated references**.
- The **Flickr8k dataset**, with its real-world images and diverse annotations, proved instrumental in honing the capabilities of the CNN-LSTM model. The project's success in image captioning was attributed to the **synergy between CNNs for image understanding and LSTMs for sequential context generation,** demonstrating the efficacy of this architecture in capturing intricate relationships within the visually rich **Flickr8k dataset.**

3.1 Results & Observation

We conducted extensive testing on **approximately 300 images spanning diverse categories.** Our observations revealed that the system performed exceptionally well, generating perfect captions for **around 178 images.** Notably, the system excelled when dealing with **images containing a minimal number of objects, typically one or two**. However, challenges surfaced when the image featured individuals wearing multicolored shirts. In such instances, the system encountered difficulty accurately identifying colors and tended to designate **the brightest color, such as orange, as the predominant color in the image below**.



Caption: Men Riding Bikes Wearing orange Shirt.

And this also cannot determine moving and still objects and also doesn't

determine multiple same objects as shown in images below.

Caption: Dog is running through water.

Caption: Multiple Giraffe standing in grassy area

Our observations resulted in a **precision rate of 63%, surpassing the performance of previously utilized datasets.** This finding indicates an improved level of accuracy compared to those datasets used in earlier studies.

4. Conclusion

The process of generating image captions involves the utilization of **Convolutional Neural Networks and Long Short-Term Memory to identify objects and describe images**. This approach has numerous advantages, particularly when using convolutional techniques for **image caption generation**. Although automatically generating captions for images is a complex task, leveraging powerful deep learning models yields promising results. Looking ahead, there is potential to elevate our project by adapting the model to generate **captions for live videos**. Currently focused on **image captioning**, this represents a challenging task, and extending it to **live video captioning is even more intricate, demanding GPU-based processing**. Captioning videos holds great significance across various domains, automating tasks like video surveillance. Additionally, enhancing our model to produce **voice clips corresponding to generated captions can aid the visually impaired, providing them insights about the images.**

References

[1] **Fang, Hao**, et al. "*Transitioning between textual descriptions and visual concepts.*" In the proceedings of the **IEEE conference** on computer vision and pattern recognition, ***2015***

[2] **Xu, Kelvin**, et al. "*Demonstrate, focus, and articulate: Generating neural image captions with visual attention*." In the International conference on **Machine Learning**, ***2015***.

[3] **Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio**. ***2015.*** "*Neural machine translation through simultaneous learning of alignment and translation.*" In the **International Conference on Learning Representations (ICLR).**

[4] **[4]. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang**. ***2017***. *Bottom-up and top-down attention for image captioning and vqa.* arXiv preprint arXiv:1707.07998 (2017).

[5] **Socher, R., Karpathy, A., Le, Q.V., Manning, C.D., & Ng, A.Y. *(2014).*** "*Grounded Compositional Semantics for Discovering and Describing Images using Sentences.*" **Transactions of the Association for Computational Linguistics (TACL)**, ***2, 207-218.***

[6] **Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. *(2018)***. *Bottom-up and top-down attention for image captioning and VQA.* **Transactions on Pattern Analysis and Machine Intelligence (TPAMI).**

[7] **Lu, J., Xiong, C., Parikh, D., & Socher, R**. ***(2017)***. *Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In Proceedings of the IEEE* **Conference on Computer Vision and Pattern Recognition (CVPR).**

[8] **Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., ... & Bengio, Y.** ***(2015)***. *Show, attend and tell: Neural image caption generation with visual attention. In the International* **Conference on Machine Learning (ICML).**

[9] **Ren, S., He, K., Girshick, R., & Sun, J.** ***(2015).*** Faster R-CNN: *Towards real-time object detection with region proposal networks.* In Advances in **neural information processing systems (NIPS).**

[10] **Johnson, J., Karpathy, A., & Li, F. F.** ***(2016)***. DenseCap: *Fully convolutional localization networks for dense captioning.* In Proceedings of the IEEE Conference on **Computer Vision and Pattern Recognition (CVPR).**