Image Captioning System Using ResNet-50 and LSTM

Harsh Chinchakar[1], Gaurav Ratnaparkhi[2], Atharva Tanawade[3], Saloni Raj Singh[4], Divyansh Modi[5], Amaan Shaikh[6], Vidya Patil[7]

[1] Dr. Vishwanath Karad MIT World Peace University, Pune, harshchinchakar33@gmail.com
[2] Dr. Vishwanath Karad MIT World Peace University, Pune, gauravratnaparkhi063@gmail.com
[3] Dr. Vishwanath Karad MIT World Peace University, Pune, 1032210471@mitwpu.edu.in
[4] Dr. Vishwanath Karad MIT World Peace University, Pune, 1032210991@mitwpu.edu.in
[5] Dr. Vishwanath Karad MIT World Peace University, Pune, 1032211377@mitwpu.edu.in
[6] Dr. Vishwanath Karad MIT World Peace University, Pune, amaan.sk03@gmail.com
[7] Dr. Vishwanath Karad MIT World Peace University, Pune, vidya.patil@mitwpu.edu.in

Abstract. Generating informative descriptions for photographs automatically has become an interesting but challenging endeavor. This paper introduces AI Image Captioning using ResNet-50 and LSTM (AICRL), a unified model that uses LSTM with ResNet50 for automatic image captioning. AICRL combines an encoder that uses ResNet50 to build a full image representation with a decoder that uses LSTM and a soft attention mechanism to predict the next phrase while highlighting particular features of the image. AICRL is evaluated using metrics like BLEU. It was trained on the Flickr8k and COCO 2017 dataset individually and optimizes the likelihood of the target description sentence given training photos. The results highlight the effectiveness of AICRL in creating image descriptions. Additionally, the PERSONALITY-CAPTIONS challenge is presented, to generate engaging captions through the integration of configurable characteristics of personality and style. A large dataset is applied to the models created by combining state-of-the-art methods in sentence and image representations. The proposed models demonstrate both robust performance on the unique PERSONALITY-CAPTIONS task and state-of-the-art performance on well-known datasets such as Flickr8k and COCO 2017. Online assessments confirm the best performance of the proposed model close to already existing models.

Keywords: Image captioning, AICRL, ResNet50, LSTM, Soft attention

1 Introduction

The increasing availability of digital photographs with written descriptions has sparked interest in automatic image captioning. For humans, this is an easy task, but it is quite challenging for robots to do. This is because the task necessitates not just the identification of objects but also an intricate understanding of the connections and features found in pictures that can be articulated using natural language. Three different kinds of picture captioning research have emerged: template-based, retrieval-based, and unique techniques to creating captions. Although the previous two categories guarantee syntactic accuracy, whereas the latter— notably, unique caption generation—uses deep learning and machine learning approaches to aim for semantic accuracy. Using long short-term memory (LSTM) with a soft attention mechanism and a convolutional neural network, ResNet50, the primary goal of this study is to create a single-joint model known as AICRL. An LSTM-based decoder designed to target certain visual features yields accurate phrase prediction, while the AICRL encoder ResNet50 offers a full image representation. Our empirical investigation shows that AICRL is effective in providing correct picture captions through model design and hyperparameter fine-tuning. An extensive review of relevant literature, the proposed AICRL model, experimental evaluations, and concluding remarks are presented in the next sections. In order to provide robots, the same degree of perceptual capacity as humans in terms of analyzing visual information and articulating contextual descriptions, this study illustrates how the areas of computer vision and natural language processing come together. In an attempt to fill the perceived gap between written descriptions and visual content, this article explores the intersection of computer vision and natural language processing. Combining ResNet50 and LSTM has become a powerful framework for handling the complex process of captioning images. Our research in this paradigm focuses on evaluating the relative performance of several encoder designs in the ResNet50-LSTM framework, specifically ResNet50.

2 Related Work

2.1 Template-based approach:

Imagine this like a fill-in-the-blank mad lib for images. These methods rely on predefined templates that have slots for objects, actions, and attributes. The system first identifies these elements within the image and then plugs them into the blanks in the template. This ensures captions are grammatically correct, but it also makes them inflexible. The length and overall structure of the caption are predetermined by the template, limiting creativity and variation.

2.2 Retrieval-based approach:

This approach acts like a resourceful librarian searching for relevant text. It retrieves captions from a large database by finding images with similar visual content or sentences that best describe what the image shows. While this method can generate captions that are generally correct and grammatically sound, it often struggles with specificity. The retrieved captions might be generic descriptions that could apply to many similar images, lacking the details unique to the specific picture at hand.

2.3 Creative caption creation for images:

This creative approach uses machine learning, particularly for in-depth instruction, Convolutional Neural Networks (CNNs), are one kind of deep learning model, function similarly to highly skilled picture analysts. They analyze the picture, highlighting important characteristics and details. Next, another model—typically a long short-term memory (LSTM) network—receives this knowledge and uses it to deliver stories like a professional storyteller. The LSTM network provides a new caption that explains the particular content of the image based on the retrieved attributes. Extremely accurate captions tailored to individual images are possible with this method. But training these deep learning models may be computationally expensive, and for them to perform well, they frequently need enormous volumes of data.

3 Model

In the proposed image captioning system, the power of ResNet-50, a pre-trained deep convolutional neural network (CNN) architecture as shown in Fig. 1, is leveraged for feature extraction. ResNet-50 excels at learning robust and informative visual representations from images. Its deep convolutional layers automatically learn hierarchies of features, progressively capturing from minute elements like textures and edges to sophisticated semantics concepts like objects and their relationships. This rich feature extraction serves as a critical first step in the image captioning pipeline. The pre-trained weights of ResNet-50, are utilized to benefit from its ability to identify and encode a wide range of visual elements within an image, providing a strong foundation for the subsequent stages of the model where these features are used to generate accurate and descriptive captions.

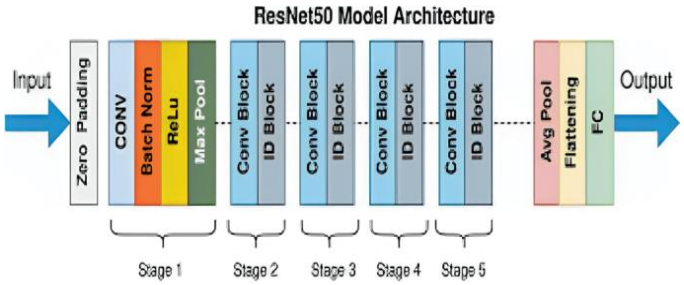

Fig. 1. ResNet50 Model Architecture (Chu, 2020)

4 Training

Preprocessing

The purpose of this study's pre-processing pipeline was to get the COCO 2017 dataset ready for use in the image captioning model training. Using the OpenCV (cv2) library, the images were initially downsized to 224x224 pixels to provide equal input

dimensions, which is essential for the CNN-based feature extraction procedure. After resizing, the image pixel values were normalized to the [0, 1] range by dividing by 255. This ensures that the inputs are within a tolerable numerical range, which facilitates effective training. To tokenize the captions from the COCO dataset and create lists of individual words, the textual data was taken. Tokenization was developed to map words to integer indices for further model processing, and it is a necessary step in creating a vocabulary of unique terms. Using this language, the captions were then encoded as a series of integers. To determine the ideal configurations for model training, hyperparameter tweaking was also carried out by adjusting the learning rate, batch size, and optimizer type. Together with methodical hyperparameter tuning, these pre-processing procedures—which included image scaling, normalization, tokenization, and encoding—ensure the dataset's readiness for efficient training and assessment of the image captioning model.

5 Experiments and Analysis

In this study, an image captioning system, a pre-trained ResNet-50 architecture is employed for feature extraction. ResNet-50 is a deep convolutional neural network (CNN) known for its ability to learn robust visual representations from images. The model is trained and tested using Flickr8k and COCO 2017 individually to fine-tune its feature extraction capabilities for the task of image captioning. The extracted features effectively captured low-level details to high-level semantic concepts, providing a strong foundation for the subsequent stages of the model. This approach achieved promising results, demonstrating the effectiveness of ResNet-50 in the proposed image captioning system.

The learning rate, batch size, and optimizer were the hyperparameters that were adjusted in this fine-tuning procedure. The batch size ranged from [32, 64, 128], while the learning rate was chosen from the set [0.001, 0.01, 0.1]. Two optimizers, Adam and SGD, were taken into account.

Several issues surfaced during our experimenting process, each of which needed to be carefully considered and resolved. The risk of overfitting, in which the model learns to commit the training data to memory instead of applying it to new samples, was a significant problem. To counteract this, a variety of regularization strategies were used, including early halting and dropout, which effectively stopped overfitting by pushing the model to acquire more reliable features. In addition, the computational requirements presented a noteworthy obstacle, particularly considering the intricacy of neural network training. We experimented with batch sizes and learning rates, streamlined our code, and achieved the highest possible performance within our restrictions despite having restricted computational resources, especially access to GPUs such as the P100.

Using pre-trained weights from a ResNet architecture, the model was first fine-tuned. Then, the fully connected layers were changed for caption creation while the convolutional layers remained the same. To better capture image attributes essential to captioning, the approach updated the model's parameters selectively based on task-specific data. This tactic resulted in increased performance as well as faster convergence.

The outcomes showed notable improvements in measures including BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores as shown in Table 1, demonstrating how well the fine-tuning strategy worked to improve the capabilities of the picture captioning model. The BLEU score with Flicker8k using beam search when k=3 is 0.493.

Table 1. The performance Metrics using COCO 2017 dataset

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
AICRL-ResNet50	0.672	0.511	0.421	0.330

Table 2. Comparison with other Metrics **(Chu, 2020)**

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
AICRL-ResNet50	0.731	0.562	0.41	0.362

6 Project Implementation

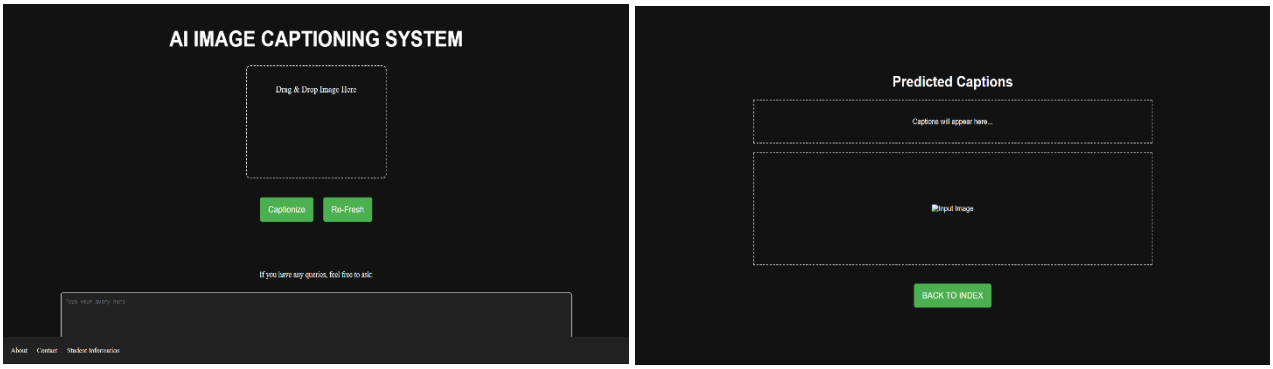

Fig. 2. The homepage of proposed Image Captioning System using ResNet-50 and LSTM.

Homepage Features of the proposed Image Captioning System are:

- Image insertion
- Captionize button
- Re-fresh button
- Query tab

- Info tab

Image Upload: When user click the "Input Image" button, he'll likely be prompted to select an image file from the device's storage. This could be done through a pop-up window allowing user to navigate through folders or by dragging and dropping the image directly onto the designated area.

Image Processing: Once the image is selected, the website will upload it. Behind the scenes, the website uses image recognition techniques to analyze the content of the image. This will involve extracting features like objects, colors, and scene composition.

Caption Generation: The website employs a pre-trained image captioning model. This model has been trained on a massive dataset of images and their corresponding captions. Based on the extracted features from the image, the model predicts a caption that describes the content.

Result Display: Once the caption is generated, the website displays it in the section that says "Captions will appear here..." This could be a single caption or maybe even a few options for the user to choose from.

Reference Captions Examples:

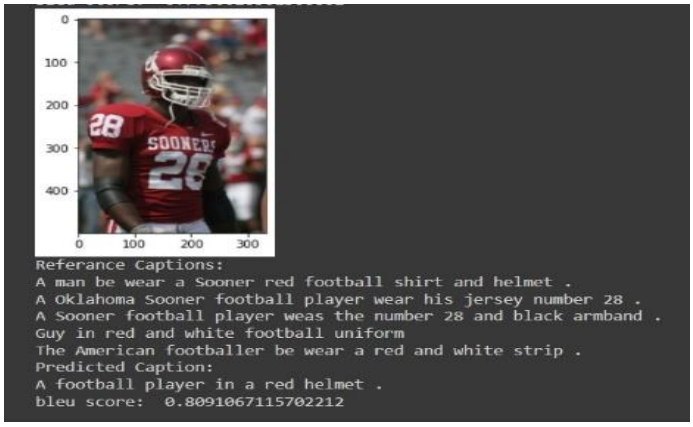

Fig.3 System Generated captions using the reference captions

Predicted Caption:
A football player in a red helmet.
bleu score: 0.809

7 Future Scope

The creation of automated content generation for educational materials is one less common but significant use case that can improve accessibility and engagement for a variety of learning audiences. Furthermore, there are intriguing prospects for combining our image captioning approach with Internet of Things (IoT) technologies. For example, using the proposed captioning model and smart eyewear with cameras, we will be able to help people who are blind or visually challenged. By giving the wearer audio descriptions of their surroundings in real time, this technology will greatly improve their independence and situational awareness. For these applications to ensure low latency and high dependability in real-world circumstances, more research into optimizing model efficiency and deployment on edge devices is essential. Our goal in future is to enhance the effect and wider use of image captioning technology in daily life by deepening the research in these areas.

8 Conclusions

This study highlights the effectiveness of the combination of ResNet-50 and LSTM models for automatic picture captioning, which represents a significant breakthrough in computer vision and natural language processing. The proposed system is capable of providing descriptive captions that effectively capture the meaning and context of a variety of photos through thorough testing and assessment. The system combines the strong feature extraction capabilities of ResNet-50 with the skillful sequence modeling of LSTM to produce a smooth combination of visual comprehension and linguistic expression. Our technique can overcome traditional limitations thanks to this harmonious integration, producing captions that are not only grammatically correct but also thematically and contextually rich. The proposed research has ramifications that go beyond academic domains, as it has potential applications across multiple sectors and domains. Our picture captioning system has the potential to transform how we view and interact with visual information, from helping visually impaired people access visual content to improving virtual assistants and recommendation systems. Furthermore, by using knowledge from computer vision, deep learning, and natural language processing to tackle a complex problem, The Proposed study emphasizes the value of multidisciplinary cooperation. Through the promotion of collaborations among these fields, the foundation is established for upcoming breakthroughs and developments in intelligent systems. Our findings pave the way for future investigation and improvement. Subsequent efforts could involve optimizing model topologies, investigating novel attention mechanisms, or incorporating extra modalities like text and audio to enhance the captioning process. The path toward improving image captioning with ResNet-50 and LSTM is a prime example of artificial intelligence's limitless potential and its revolutionary influence on how we perceive and engage with the visual world. Using AI for societal improvement and the improvement of human experiences we can push the boundaries of technology.

9 References

[1] Chu, Y., Yue, X., Yu, L., Sergei, M., & Wang, Z. (2020). Automatic image captioning based on ResNet50 and LSTM with soft attention. Wireless Communications and Mobile Computing, 2020, 1- 7. (Chu, 2020)
[2]] Shuster, K., Humeau, S., Hu, H., Bordes, A., & Weston, J. (2019). Engaging image captioning via personality. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12516-12526). (Shuster, 2019)
[3] Charu, S., Mishra, S. P., & Gandhi, T. (2020, January). Vision to Language: Captioning Images using Deep Learning. In 2020 International Conference on Artificial Intelligence and Signal Processing (AISP) (pp. 1-8). IEEE.(Charu, 2020)
[4] Dharsini, S. V., Razak, M. A., Modi, S., Reddy, P. K., & Bhatnagar, S. (2022, December). Captioning based image using Euclidean distance and resNet-50. In 2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI) (Vol. 1, pp. 1-5). IEEE..(Dharsini, 2023)
[5] Alam, M. S., Rahman, M. S., Hosen, M. I., Mubin, K. A., Hossen, S., & Mridha, M. F. (2021, October). Comparison of different CNN model used as encoders for image captioning. In 2021 International conference on data analytics for business and industry (ICDABI) (pp. 523-526). IEEE. (Alam, 2021)
[6] Raut, R., Patil, S., Borkar, P., & Zore, P. (2023). Image Captioning Using ResNet RS and Attention Mechanism. International Journal of Intelligent Systems and Applications in Engineering, 11(7s), 606-613(Raut, 2023)
[7] Al-Malla, M. A., Jafar, A., & Ghneim, N. (2022). Image captioning model using attention and object features to mimic human image understanding. Journal of Big Data, 9(1), 20(Al-Malla, 2022)
[8] Maru, H., Chandana, T. S. S., & Naik, D. (2021, April). Comparison of image encoder architectures for image captioning. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 740-744). IEEE. (Maru, 2021)
[9] Bhattacharya, A., Girishekar, E. S., & Deshpande, P. A. (2021). Empirical Analysis of Image Caption Generation using Deep Learning. arXiv preprint arXiv:2105.09906.(Bhattacharya, 2021)
[10] Mundargi, S., & Mohanty, H. (2020). Image Captioning using Attention Mechanism with ResNet VGG and Inception Models. International Research Journal of Engineering and Technology (IRJET), 7(09). (Mundargi, 2020)
[11] Suresh, K. R., Jarapala, A., & Sudeep, P. V. (2022). Image captioning encoder–decoder models using cnn-rnn architectures: A comparative study. Circuits, Systems, and Signal Processing, 41(10), 5719-5742.(Suresh, 2022)
[12] Patel, A., & Varier, A. (2020). Hyperparameter analysis for image captioning. arXiv preprint arXiv:2006.10923. (Patel, 2020)
[13] Sri Neha, V., Nikhila, B., Deepika, K., & Subetha, T. (2022). A Comparative Analysis on Image Caption Generator Using Deep Learning Architecture—ResNet and VGG16. In Computational Vision and Bio-Inspired Computing: Proceedings of ICCVBIC 2021 (pp. 209-218). Singapore: Springer Singapore.(Sri Neha, 2022)