Information Dynamics and Applications
https://www.acadlore.com/journals/IDA

Enhancing Image Captioning and Auto-Tagging Through a FCLN with Faster R-CNN Integration

Shalaka Prasad Deore[1*], Taibah Sohail Bagwan[1], Prachiti Sunil Bhukan[1], Harsheen Tejindersingh Rajpal[1], Shantanu Bharat Gade[2]

[1] Computer Engineering, MES College of Engineering, S. P. Pune University, 411001 Pune, India
[2] School of Architecture, Computing and Engineering, University of East London, E16 2RD London, UK

* Correspondence: Shalaka Prasad Deore (shalakasonawane25@gmail.com)

Received: 10-27-2023 **Revised:** 12-17-2023 **Accepted:** 01-25-2024

Citation: S. P. Deore, T. S. Bagwan, P. S. Bhukan, H. T. Rajpal, and S. B. Gade, "Enhancing image captioning and auto-tagging through a FCLN with faster R-CNN integration," *Inf. Dyn. Appl.*, vol. 3, no. 1, pp. 12–20, 2024. https://doi.org/10.56578/ida030102.

Abstract: In the realm of automated image captioning, which entails generating descriptive text for images, the fusion of Natural Language Processing (NLP) and computer vision techniques is paramount. This study introduces the Fully Convolutional Localization Network (FCLN), a novel approach that concurrently addresses localization and description tasks within a singular forward pass. It maintains spatial information and avoids detail loss, streamlining the training process with consistent optimization. The foundation of FCLN is laid by a Convolutional Neural Network (CNN), adept at extracting salient image features. Central to this architecture is a Localization Layer, pivotal in precise object detection and caption generation. The FCLN architecture amalgamates a region detection network, reminiscent of Faster Region-CNN (R-CNN), with a captioning network. This synergy enables the production of contextually meaningful image captions. The incorporation of the Faster R-CNN framework facilitates region-based object detection, offering precise contextual understanding and inter-object relationships. Concurrently, a Long Short-Term Memory (LSTM) network is employed for generating captions. This integration yields superior performance in caption accuracy, particularly in complex scenes. Evaluations conducted on the Microsoft Common Objects in Context (MS COCO) test server affirm the model's superiority over existing benchmarks, underscoring its efficacy in generating precise and context-rich image captions.

Keywords: Faster Region Convolutional Neural Network (R-CNN); Long Short-Term Memory (LSTM); Image captioning; Object detection

1 Introduction

The domains of object detection and image captioning stand as pivotal elements in the expansive field of computer vision. These tasks, central to numerous applications including robotics, medical imaging, content moderation, and assisting visually impaired individuals, involve the identification and descriptive articulation of objects within images. Such capabilities are essential for enabling machines to comprehend visual data and interact effectively with their surroundings. Faster R-CNN has emerged as a preeminent method in object detection. It is distinguished by its dual-stage framework, which combines deep CNNs with region proposal networks (RPNs). This approach has garnered significant attention due to its high levels of accuracy and efficiency in localizing and classifying objects within images.

Concurrently, the task of image captioning requires the generation of descriptive textual captions that accurately convey the visual content of images. This necessitates an intricate understanding of both the visual and semantic dimensions of the images, bridging the gap between visual perception and textual representation. The fusion of computer vision and NLP enables these models to effectively translate visual data into coherent textual descriptions. In this study, the integration of Faster R-CNN within the realms of object detection and image captioning is critically examined. The objective is to harness the advanced detection capabilities of Faster R-CNN to augment the accuracy and efficacy of image captioning systems. By leveraging the rich spatial information and detailed object features provided by Faster R-CNN, the precision of generated captions can be significantly enhanced, yielding more accurate depictions of objects within images.

In this research, a comprehensive examination of the Faster R-CNN framework is conducted, with a focus on delineating its crucial components and underlying mechanisms. The challenges inherent in integrating Faster R-CNN into image captioning frameworks are explored, along with the pertinent techniques devised to surmount these obstacles. The study also encompasses a discussion on the dataset requisites, training methodologies, and evaluation metrics deployed in the experimental validation of the proposed approach. This investigation contributes significantly in two distinct aspects. Firstly, it elucidates the proficiency of Faster R-CNN in object detection, evidenced by its exemplary performance on established benchmark datasets. Secondly, the research elucidates how the incorporation of Faster R-CNN into image captioning systems augments their capability, resulting in textual descriptions of images that are both more accurate and informative.

The primary objective of this paper is to shed light on the potential of employing Faster R-CNN as a dual-purpose framework, serving both object detection and image captioning tasks. By capitalizing on the robustness of this model, a considerable advancement in computer vision systems is anticipated, enhancing their proficiency in understanding and interpreting visual data, which holds immense potential for real-world application enhancement. The intersection of computer vision and NLP has made notable strides in generating image captions, thereby bridging the visual-textual comprehension divide. This advancement is pivotal, as it equips machines with the capability to interpret and articulate the content of images, marking a significant stride towards realizing artificial intelligence. Nonetheless, current image captioning methodologies often falter in accurately capturing the intricate context of images, consequently restricting their effectiveness in generating insightful descriptions. To address this gap, the present study proposes an innovative image captioning system that capitalizes on contextual information, thereby markedly improving the precision of image descriptions.

The core aim of this study is the development and demonstration of a visual feature extraction component, leveraging the state-of-the-art Faster R-CNN for the extraction of pertinent visual features from input images. Subsequently, captions are generated using advanced neural language models, notably CNN and Artificial Neural Networks (ANN). Building upon previous research, a transition from Recurrent Neural Networks (RNN) to CNN is undertaken in the neural network architecture, aiming to enhance performance. Furthermore, this research diverges from the traditional focus on individual words, opting instead to explore the use of phrases as fundamental units in order to augment both the semantic and syntactic quality of the generated captions.

This paper presents a thorough analysis of the proposed novel approach to image captioning, highlighting its architectural enhancements and the integration of contextual information. This is pursued with the objective of surmounting the limitations inherent in existing methodologies and providing captions that more accurately reflect the relationships among entities within images. An exploration of the data set requirements, training protocols, and evaluation metrics employed in assessing the system's performance is also included. A comparative analysis of the results obtained from the modified neural network, juxtaposed against existing models, is conducted to underscore its superiority in generating more precise and meaningful image captions. The implications of this research are far-reaching, potentially advancing machine-based image comprehension and supporting a spectrum of applications spanning from image retrieval to autonomous systems and human-machine interaction. Subsequent sections encompass an overview of related works in image captioning, an exposition of the proposed model's architecture, a delineation of the data set and training methodologies, a presentation of experimental results, and a conclusion that offers perspectives on future enhancements in image captioning systems.

2 Related Work

The domain of automated image captioning, particularly when employing the Faster R-CNN model, has received considerable focus in recent scholarly research. This literature review section aims to encapsulate key contributions that have significantly influenced the evolution of automated image captioning methodologies utilizing Faster R-CNN. Anderson et al. [1] introduced a novel approach incorporating both bottom-up and top-down attention mechanisms in the context of image captioning and Visual Question Answering. This research underscored the pivotal role of attention mechanisms in enhancing image comprehension and the accuracy of caption generation. The integration of Faster R-CNN for object detection, combined with a top-down attention mechanism, resulted in marked improvements in caption generation, as evidenced by their findings. Lu et al. [2] explored an adaptive attention mechanism employing a visual sentinel for image captioning. This model was distinguished by its dynamic learning capability, which enabled it to focus on varying regions of an image during the process of caption generation. The enhanced relevance and contextual understanding of the captions generated by this model illustrated the efficacy of adaptive attention mechanisms in identifying salient features within images. The adoption of Faster R-CNN for object detection, coupled with the integration of semantic knowledge, significantly elevated the performance of their image captioning approach.

A seminal contribution in the field was made by Xu et al. [3] through their work on neural image caption generation with visual attention. They introduced an attention-based model that was adept at focusing on pertinent regions of an image while formulating captions. The integration of a CNN for feature extraction, along with a LSTM

network-based decoder, significantly bolstered the quality and pertinence of the image captions generated by their model. The introduction of the Faster R-CNN model by Ren et al. [4] marked a significant advancement in the field of object detection. This framework, which integrates RPNs with a CNN for feature extraction, has facilitated real-time object detection. The implementation of Faster R-CNN in automated image captioning models has proven to be instrumental in enhancing object localization and recognition accuracy, a development corroborated by numerous scholarly studies. Xiao et al. [5] presented novel LSTM attention based model (ALSTM) for tagging images. In existing models, ALSTM learns to refine input vector via sequential context information and network hidden states, instead of typical LSTM. As a result, ALSTM is able to react to more pertinent characteristics, including spatial attention, visual relations, and a greater focus on the most pertinent context terms. Additionally, Chen and Hu [6] presented a novel text-based visual attention model for image captioning that employed self-attention mechanisms, eschewing recurrent structures. This simplification of the captioning process, while maintaining competitive performance, illustrated an alternative methodology in image captioning that relied solely on attention mechanisms. It gives previously generated text, automatically eliminates unnecessary information to focus on a single salient object. Ren et al. [7] delved into deep reinforcement learning-based image captioning, employing an embedding reward model. This approach incorporated Faster R-CNN for object detection and harnessed LSTM-based caption generation. By applying reinforcement learning techniques, their model demonstrated marked improvements in the quality of generated image captions. Zhan et al. [8], presented model which will improve both picture representation and caption generation by better utilizing the semantics found in captions. Initially, this model uses supervised multi-instance learning weak model to build graph which explains caption-guided visual relationship. Next, contextual and nearby nodes with their textual and visual characteristics are added to improve the representation. Using this author achieved promissing results. Many researchers used faster R-CNN model for image captioning and achieved promising results [9–11].

Zhou et al. [12] introduced an innovative approach VQA by embedding external knowledge and attribute-based reasoning into their model. Capsul Network is implemented which uses dynamic routing to achieve the attention output. This method computes coupling coefficients between the underlying and output capsules in order to update the attention weights. Omri et al. [13] and Zhu and Yan [14] proposed deep learning method to improve results of automated image captioning. In the study of Thangave et al. [15], provides a heterogeneous data fusion-based deep learning model for image captioning. The descriptive text is created, the long short-term memory is used for decoding, and mask recurrent neural networks i.e faster R-CNN are used in the coding layer. In the study of Rehab and Hahn [16] also porved that deep leaning approache is very effective for improving results of image captioning. In the study of Abdelrahman ert al. [17] suggested and investigated two approaches to the picture captioning problem utilizing two distinct self-supervised learning models. The comprehensive study done in several studies [18–20] suggested to use deep learning models to get promissing results in image captioning especially R-CNN proved the best model.

3 Proposed Approach

The methodology adopted for automated image captioning in this study integrates the capabilities of Faster R-CNN and LSTM, utilizing a combination of bottom-up and top-down attention mechanisms, as depicted in Figure 1. The integration of these mechanisms leverages the object detection capabilities of Faster R-CNN and the FCLN to pinpoint salient visual features within an image. This approach encompasses the use of a pre-trained Faster R-CNN model for the extraction of object features, a LSTM network for caption generation, and a bottom-up/top-down attention mechanism to focus on significant object regions.

The method employs Faster R-CNN to detect objects by generating region proposals, which serve as inputs to the FCLN. The FCLN performs fine-grained localization in the images, mapping them to their respective class labels. Subsequently, the features extracted from the images are processed by the LSTM model, which is responsible for generating a caption that describes the image.

The proposed methodology contains the following steps.

3.1 Dataset Preparation

For image captioning, a dataset comprising images paired with corresponding captions is imperative. The MS COCO dataset, known for its diversity in image content and complexity, was utilized in this study. It features annotations for 80 different object categories. The dataset is segmented into various splits, including training, validation, and test sets. For the purposes of this study, a selection of 5000 images was made for testing and another 5000 for validation. Each image in the MS COCO dataset is accompanied by five captions and labels for object classes, providing a comprehensive set for training and evaluating the model.

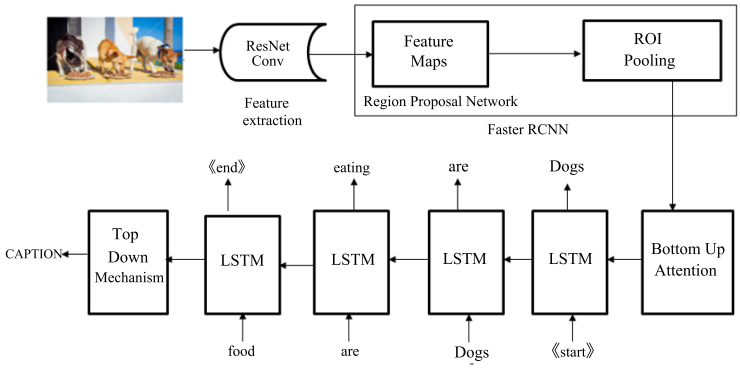

Figure 1. Architecture of the proposed system

3.2 Object Detection Using Faster R-CNN

The pre-trained Faster R-CNN model, based on ResNet-101, was employed to extract features and perform object detection on the MS COCO dataset. The model underwent fine-tuning and was trained on the training data. It generated region proposals using the RPN, identifying potential object bounding boxes to accurately detect objects within the dataset.

3.3 Bottom-Up Attention

Bottom-up features were extracted from the Faster R-CNN model, representing individual object regions and their associated feature vectors. These object regions were then ranked based on their importance scores or confidence values. The top-k object regions, determined by their importance scores (where k is a predefined parameter), were processed by the FCLN.

3.4 LSTM for Caption Generation

The LSTM network is initialized for the generation of captions. Inputs to the LSTM include the visual features derived from the bottom-up attention mechanism and the words generated previously. Words are represented as dense vectors via an embedding layer. The LSTM network undergoes training with caption pairs from the dataset, predicting subsequent words based on earlier words and visual features. During training, ground truth words are inputted into the LSTM network at each timestep. A softmax layer calculates the probability distribution over the vocabulary for each word.

3.5 Top-Down Attention

The hidden state of the LSTM serves as a query to focus on the visual features obtained from the bottom-up attention mechanism. Techniques such as dot product are employed to compute attention weights between the LSTM hidden state and visual features. These attention weights are crucial for computing the weighted sum of visual features, resulting in the formation of a context vector.

3.6 Caption Generation

This context vector, obtained from the top-down attention mechanism, is amalgamated with the current timestep's LSTM hidden state. The amalgamated vector is then inputted into the LSTM for the generation of the subsequent word. This procedure is iteratively executed until an end-of-sentence token is produced or the maximum caption length is reached.

3.7 Training and Evaluation

The model undergoes end-to-end training through backpropagation and optimization techniques. The discrepancy between generated captions and provided captions is measured using cross-entropy loss during training. For evaluating the model's performance, the Bilingual Evaluation Understudy (BLEU) score metrics are utilized. The BLEU score, a prominent metric in various NLP tasks, assesses the similarity between generated text and the captions in the dataset.

4 Implementation

4.1 Faster R-CNN Model

In the employed approach, the Faster R-CNN model, pre-trained on the ResNet-101 backbone network, was utilized. ResNet-101, characterized by its 101-layer architecture, is responsible for comprehending the input image

through high-level visual feature extraction. These features are represented via a feature map, which is subsequently inputted into the RPN. The RPN, integrated with the backbone network, generates candidate object proposals, essentially potential bounding boxes encapsulating objects of interest. The generation of these proposals is influenced by anchor boxes of predefined shapes, scales, and aspect ratios. The bounding boxes proposed by the RPN are forwarded to the Region of Interest (RoI) pooling layer, which extracts fixed-size feature maps for each region from the output feature maps of the backbone network. Subsequently, a fully connected layer, attached to the RoI pooling layer, undertakes object classification within the proposed regions and bounding box regression to refine their coordinates. The process concludes with the application of non-maximum suppression, filtering out redundant bounding box proposals based on confidence scores and overlapping regions. The highest-scoring proposals are retained, while others are discarded.

4.2 Bottom-Up Attention Mechanism

The Bottom-up Attention Mechanism synergizes the functionalities of Faster R-CNN and FCLN in the realm of image caption generation. This mechanism commences with Faster R-CNN extracting visual features and identifying regions of interest within the image. These regions are then assigned importance scores, facilitating the attention mechanism during caption generation. Integrated with FCLN, the attention mechanism at each timestep dynamically focuses on the most relevant regions, considering the current context and the evolving captions. FCLN aids in further refining the localization of these selected regions. Crucially, attention weights are computed, factoring in both the importance scores and the detailed localization provided by FCLN. This ensures that the captions generated are not only descriptive but also centered on the most significant visual aspects of the image, thereby elevating the quality of the captions.

4.3 LSTM Model

In this study, the LSTM network is employed for the sequential generation of image captions. Image features, once extracted, are inputted into the LSTM network along with word embeddings. The LSTM model is subjected to training utilizing a dataset that pairs images with their respective captions. At each timestep, the model processes the current word embedding and the preceding hidden state to forecast the subsequent word in the caption. The optimization of the model during training involves updating the LSTM weights to enhance performance. Key hyperparameters are established, with the maximum caption length set to 25 words and the word embedding dimension fixed at 250. The LSTM model, comprising two layers, is designed to maintain hidden states and memory cells, vital for contextual comprehension and the generation of captions. The hidden states from the first layer of LSTM are channeled as input to the second layer, thereby enabling the model to discern and encapsulate more complex relationships inherent in the caption generation process. Throughout the process, the LSTM layers, at each timestep, are tasked with predicting the next word in the caption, based on the current input and historical hidden states. This iterative process persists until the predetermined maximum caption length is reached.

5 Dataset

For the evaluation of the proposed captioning model, the MS COCO 2014 captions dataset, a benchmark in the field, was employed. The Karpathy splits, previously utilized in various studies for result comparison, were adopted for model hyperparameter validation and offline testing. The MS COCO 2014 dataset encompasses 123,287 training images, each paired with five captions. In addition, separate sets, each comprising 5,000 images, were designated for validation and testing. These sets were instrumental in conjunction with the submissions to the MS COCO test server. The training of the model was conducted on a merged dataset of training and validation images, totaling 123,000 images.

Text pre-processing on the captions was executed in line with standard practices, entailing conversion of all sentences to lowercase, tokenization based on white spaces, and exclusion of words appearing fewer than five times. Consequently, the model's vocabulary comprised 10,010 words. For the assessment of caption accuracy, widely recognized automatic evaluation metrics, specifically Metric for Evaluation of Translation with Explicit ORdering (METEOR) and BLEU, were utilized. These metrics provide objective criteria for evaluating the model's performance against the ground truth captions. The selection of the MS COCO dataset and these standard evaluation metrics facilitates meaningful comparisons between the proposed model and existing methods in automated image captioning.

6 Result

In this study, a two-step training approach was adopted for the image caption generator employing Faster R-CNN and LSTM. Initially, the model underwent pretraining for 30 epochs, utilizing the Adaptive Moment Estimation (ADAM) optimizer alongside softmax cross-entropy loss. Performance monitoring occurred on the validation set, which comprised 5000 images. An early-stopping mechanism was implemented to optimize results. The training,

focused on cross-entropy loss, was executed on a single Graphic Processing Unit (GPU) and spanned approximately one day. The pretraining phase was critical in initializing the model and capturing essential visual features for the generation of accurate and meaningful captions. The optimization of model parameters via cross-entropy loss was aimed at enhancing the performance and generalization capabilities of the caption generator. Evaluations based on multiple metrics revealed significant results. As indicated in Table 1, the performance was assessed using BLEU, Consensus-based Image Description Evaluation (CIDEr), METEOR, and Semantic Propositional Image Caption Evaluation (SPICE) scores, which are standard metrics for evaluating the quality and accuracy of generated captions.

Table 1. Evaluation metrics

BLEU -1	70.3
BLEU -2	52.6
BLEU -3	38.1
BLEU -4	28
SPICE	23.7
METEOR	25.5
CIDEr	116.8

BLEU scores achieved were 70.3 for BLEU-1 and 28 for BLEU-4, as detailed in Table 1. BLEU metric, predominantly used in machine translation and image captioning, evaluates the quality and accuracy of machine-generated text. It measures the similarity between the generated text and a set of reference texts. The BLEU-1 score reflects the precision of unigram matches between the text produced by the model and the reference texts, while BLEU-4 score assesses the precision of matching n-grams (sequences of four words). Higher BLEU scores, as shown in Table 2, denote better alignment between the model-generated text and reference texts, underlining the efficacy and accuracy of the image captioning model.

Table 2. BLEU score comparison

MODEL	BLEU-1	BLEU-2	BLEU-3	BLEU-4
NIC	65.5	45.2	30.8	23.5
LRCN	63.68	42.18	29.31	20
Proposed Faster R-CNN model	70.3	52.6	38.1	28

A METEOR score of 25.5 was attained by the model, reflecting the quality of the generated captions in terms of fluency, grammar, and overall linguistic accuracy. The METEOR score, assessing both precision and recall, indicates that higher scores correlate with more accurate and linguistically sound captions. Furthermore, a SPICE score of 31.2 was obtained, evaluating the semantic content and relevance of the captions. This metric emphasizes not only the correctness of individual words but also the semantic relationships and coherence across the entire caption. An impressive CIDEr score of 116.8 was recorded, indicating a high degree of similarity between the generated captions and the reference captions in the dataset. This score is particularly indicative of the effectiveness of the proposed model in generating captions that closely resemble the ground truth provided in the dataset.

The enhanced performance of the proposed model underscores its capability to generate more accurate and contextually relevant captions for images. This achievement highlights the integration of Faster R-CNN's object detection capabilities into the image captioning process, enabling a deeper understanding of the visual content and the generation of more informative captions. In addition to metric evaluations, a qualitative analysis was conducted. An example, depicted in Figure 2, shows an image of a woman eating a piece of cake with a candle on top. The model successfully generated the caption, "A woman is eating a piece of cake with a candle," accurately capturing the main subject and activity in the image. This example illustrates the model's proficiency in identifying relevant objects, activities, and contextual details within the visual content.

Figure 3, in particular, displays an image featuring two puppies positioned adjacent to one another on a grassy field. The caption generated for this image, stating "Two puppies sitting on a grassy field next to each other," aligns precisely with the visual content. The model was effective in identifying the key elements: the subjects (puppies), their location (grassy field), and their relative positioning (sitting next to each other). This example illustrates the model's capability to provide comprehensive and descriptive captions that accurately reflect the visual content.

These instances serve as evidence of the model's ability to produce captions that are both accurate and contextually relevant. The model demonstrates its capacity to recognize vital visual cues, correctly identify objects, and succinctly describe their actions or attributes. The congruence between the generated captions and the corresponding images underscores the efficacy of the proposed approach in generating meaningful and precise captions across a diverse range of visual content.

```
session = Inference(args.model_path)
caption = session.prerdict(args.img_path, args.beam_search)
print(caption)
```

A woman is eating a piece of cake with a candle.

Figure 2. Generated caption from input image

```
session = Inference(args.model_path)
caption = session.predict(args.img_path, args.beam_search)
print(caption)
```

Two puppies sitting on a grassy field next to each other.

Figure 3. Generated caption from input image

7 Conclusion

This study has culminated in the development and training of a deep learning model for automated image captioning. The process entailed preprocessing of images and textual data, preparing them for the training phase, and enabling the model to generate captions for newly inputted images. The integration of visual attention mechanisms at both the bottom-up and top-down levels in this work facilitates a more targeted calculation of attention, focusing primarily on objects and other notable regions within the images. A potential extension of this model could involve its application in real-time image captioning, particularly in assistive technologies for the visually impaired, thereby enhancing their perception of the environment through live image descriptions.

Moreover, the research aims to bridge the gap between visual and linguistic comprehension by incorporating cutting-edge advancements in object detection. This approach presents several potential avenues for future exploration. However, even with the current methodology, substantial improvements are observed by replacing pre-trained CNN features with pre-trained bottom-up attention features. The proposed approach demonstrates competitive results across various images, signifying an improvement in the performance of the image captioning system.

Data Availability

The data used to support the research findings are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 6077–6086. https://doi.org/10.1109/CVPR.2018.00636

[2] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 3242–3250. https://doi.org/10.1109/CVPR.2017.345

[3] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *32nd International Conference on Machine Learning*, Lille, France, 2015, pp. 2048–2057.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017. https://doi.org/10.1109/TPAMI.2016.2577031

[5] F. Xiao, W. Xue, Y. Shen, and X. Gao, "A new attention-based LSTM for image captioning," *Neural Process. Lett.*, vol. 54, no. 4, pp. 3157–3171, 2022. https://doi.org/10.1007/s11063-022-10759-z

[6] H. Chen and H. Hu, "Image captioning with text-based visual attention," *Neural Process. Lett.*, vol. 49, pp. 177–185, 2019. https://doi.org/10.1007/s11063-018-9807-7

[7] Z. Ren, X. Wang, N. Zhang, X. Lv, and L. J. Li, "Deep reinforcement learning-based image captioning with embedding reward," in *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 290–298. https://doi.org/10.1109/CVPR.2017.128

[8] S. Zhan, X. Zhou, X. Qiu, and X. Zhu, "Improving image captioning with better use of captions," *arXiv preprint arXiv*, vol. 2006, p. 11807, 2020. https://doi.org/10.48550/arXiv.2006.11807

[9] A. Yadav, "Image captioning using R-CNN & LSTM deep learning model," *Image*, vol. 5, p. 8, 2021. https://ijisrt.com/assets/upload/files/IJISRT21MAY837.pdf

[10] Z. Yang, Y. J. Zhang, S. u. Rehman, and Y. Huang, "Image captioning with object detection and localization," in *9th International Conference on Image and Graphics, ICIG 2017*, Shanghai, China, 2017, pp. 109–118. https://doi.org/10.1007/978-3-319-71589-6_10

[11] G. Geetha, T. Kirthigadevi, G. Ponsam, T. Karthik, and M. Safa, "Image captioning using deep convolutional neural networks (CNNs)," *J. Phys.: Conf. Ser.*, vol. 1712, no. 1, p. 012015, 2020. https://doi.org/10.1088/1742-6596/1712/1/012015

[12] Y. Zhou, R. Ji, J. Su, X. Sun, and W. Chen, "Dynamic capsule attention for visual question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, Washington DC, USA, 2019, pp. 9324–9331. https://doi.org/10.1609/aaai.v33i01.33019324

[13] M. Omri, S. Abdel-Khalek, E. M. Khalil, J. Bouslimi, and G. P. Joshi, "Modeling of hyperparameter tuned deep learning model for automated image captioning," *Mathematics*, vol. 10, no. 3, p. 288, 2022. https://doi.org/10.3390/math10030288

[14] Y. Zhu and W. Yan, "Image-based storytelling using deep learning," in *5th International Conference on Control and Computer Vision, ICCCV 2022*, Xiamen, Fujian, China, 2022. https://doi.org/10.1145/3561613.3561641

[15] K. Thangavel, N. Palanisamy, S. Muthusamy, O. P. Mishra, S. C. M. Sundararajan, H. Panchal, A. K. Loganathan, and P. Ramamoorthi, "A novel method for image captioning using multimodal feature fusion employing mask RNN and LSTM models," *Soft Comput.*, vol. 27, pp. 14 205–14 218, 2023. https://doi.org/10.1007/s00500-023-08448-7

[16] A. Rehab and J. Hahn, "Improve image captioning by estimating the gazing patterns from the caption," in *IEEE/CVF Winter Conference on Applications of Computer Vision*. Waikoloa, HI, USA, 2022, pp. 1025–1034. https://doi.org/10.1109/WACV51458.2022.00251

[17] M. Abdelrahman, D. Dmitry, and Z. Sebaitre, "Image captioning through self-supervised learning," *Tech. Rep.*, 2022. https://doi.org/10.13140/RG.2.2.24293.88802

[18] Z. Zanyar and J. K. Kalita, "Neural attention for image captioning: Review of outstanding methods," *Artif. Intell. Rev.*, vol. 55, no. 5, pp. 3833–3862, 2022. https://doi.org/10.1007/s10462-021-10092-2

[19] A. Oluwasammi, M. U. Aftab, Z. Qin, S. T. Ngo, T. V. Doan, S. B. Nguyen, S. H. Nguyen, and G. H. Nguyen, "Features to text: A comprehensive survey of deep learning on semantic segmentation and image captioning," *Complexity*, vol. 2021, pp. 1–19, 2021. https://doi.org/10.1155/2021/5538927

[20] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Advances in Neural Information Processing Systems*, Montreal, Quebec, Canada, 2014, pp. 3104–3112.