iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.61513

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

Captionify: Bridging the Gap Between Vision and Language with Neural Networks

Mrs. Bindu K P[1], M Rohini[2], Prajwal Gowda M[3]
[1]*Assistant Professor,* [2, 3]*Student, Department of Computer Science, K S School of Engineering and Management, Bangalore, Karnataka*

Abstract: ***This research presents a unique approach that combines Long Short-Term Memory (LSTM) networks with Convolutional Neural Networks (CNNs) to generate picture captions. The model makes use of the CNNs' ability to extract complex spatial features from pictures and the LSTM's ability to create and expand on logical textual descriptions. This combination improves the resilience and efficiency of the captioning system by successfully addressing the two difficulties of linguistic description and visual understanding. Comparative tests show that the model performs better than existing approaches in generating accurate and contextually relevant captions. This development highlights the promise of the CNN-LSTM architecture in smoothly integrating visual input with textual interpretation in addition to pushing the envelope of image captioning systems.***
Keywords: Image Captioning, Machine Learning, Neural Networks, CNN, LSTM.

I. INTRODUCTION

The field of computer vision has seen remarkable growth recently, largely due to advancements in deep learning techniques. Image captioning, a particularly complex application, demands an effective blend of visual perception and linguistic production skills. Our study introduces a cutting-edge image caption generator that harnesses the synergistic potentials of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. CNNs are adept at distilling hierarchical visual features, whereas LSTMs excel in recognizing and following the sequence of textual elements. The integration of these technologies aims to overcome the complexities found in merging image understanding with verbal articulation. This paper details the development, execution, and testing of our model, highlighting its ability to enhance the accuracy and coherence of the captions it generates. This research adds to the growing field of multimodal artificial intelligence, offering a powerful approach to addressing the sophisticated challenge of producing meaningful and context-aware captions for images.

II. SIGNIFICANCE OF THE SYSTEM

The main goal of this project is to create a dependable and precise automatic image captioning system through the use of sophisticated deep-learning strategies. This includes developing a multimodal model that successfully combines the visual features identified by the ResNet50 convolutional neural network with the sequential data managed by long short-term memory networks (LSTMs). Key tasks involve preprocessing and cleaning the caption dataset, establishing word-to-index and index-to-word mappings, acquiring word embeddings from GloVe, and training the model to produce coherent and contextually appropriate captions. The performance of the model will be evaluated using recognized metrics such as BLEU scores. Moreover, this project aims to advance the field by exploring the complexities and developments in image captioning, demonstrating its practical uses, and potentially setting the stage for future research in multimodal AI systems.

III. LITERATURE SURVEY

1) *M. Sailaja; K. Harika; B. Sridhar; Rajan Singh, Image Caption Generator using Deep Learning: 2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*

Over the last few years, deep neural networks made image captioning conceivable. The image caption generator provides an appropriate title for an applied input image based on the dataset. The present work proposes a model based on deep learning and utilizes it to generate a caption for the input image. The model takes an image as input and frames the sentence related to the given input image by using some algorithms like CNN and LSTM.

©IJRASET: All Rights are Reserved

This CNN model is used to identify the objects that are present in the image and the Long Short-Term Memory (LSTM) model will not only generate the sentence but also summarize the text and generate the caption that is suitable for the project. So, the proposed model mainly focuses on identifying the objects and generating the most appropriate title for the input images.

2) *C. S. Kanimozhiselvi; Karthika V; Kalaivani S P; Krithika S, Image Captioning Using Deep Learning, 2022 International Conference on Computer Communication and Informatics (ICCCI).*

The process of generating a textual description for images is known as image captioning. Nowadays it is one of the recent and growing research problems. Day by day various solutions are being introduced for solving the problem. Even though many solutions are already available, a lot of attention is still required to get better and precise results. So, we came up with the idea of developing an image captioning model using different combinations of Convolutional Neural Network architecture along with Long Short-Term Memory to get better results. We have used three combinations of CNN and LSTM for developing the model. The proposed model is trained with three Convolutional Neural Network architectures such as Inception-v3, Xception, and ResNet50 for feature extraction from the image and long short-term memory for generating the relevant captions. Among the three combinations of CNN and LSTM, the best combination is selected based on the accuracy of the model. The model is trained using the Flicker8k dataset.

3) *Chetan Amritkar; Vaishali Jabade, Image Caption Generation Using Deep Learning Technique, 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*

In Artificial Intelligence (AI), the contents of an image are generated automatically which involves computer vision and NLP (Natural Language Processing). The neural model which is regenerative, is created. It depends on computer vision and machine translation. This model is used to generate natural sentences which eventually describe the image. This model consists of a Convolutional Neural Network (CNN) as well as a Recurrent Neural Network (RNN). The CNN is used for feature extraction from images and RNN is used for sentence generation. The model is trained in such a way that if an input image is given to the model, it generates captions that nearly describe the image. The accuracy of the model and smoothness or command of the language model learned from image descriptions are tested on different datasets. These experiments show that the model frequently gives accurate descriptions for an input image.

4) *Varsha Kesavan; Vaidehi Muley; Megha Kolhekar: Deep Learning based Automatic Image Caption Generation. 2019 Global Conference for Advancement in Technology (GCAT)*

The paper aims to generate automated captions by learning the contents of the image. At present images are annotated with human intervention and it becomes a nearly impossible task for huge commercial databases. The image database is given as input to a deep neural network (Convolutional Neural Network (CNN)) encoder for generating a "thought vector" which extracts the features and nuances out of our image and the RNN (Recurrent Neural Network) decoder is used to translate the features and objects given by our image to obtain a sequential, meaningful description of the image. This paper systematically analyzes different deep neural network-based image caption generation approaches and pre-trained models to conclude on the most efficient model with fine-tuning. The analyzed models contain both with and without `attention' concepts to optimize the caption-generating ability of the model. All the models are trained on the same dataset for concrete comparison.

IV. METHODOLOGY

A. Resnet-50

ResNet (Residual Network) is a type of deep neural network architecture designed to address the vanishing gradient problem during training of deep convolutional neural networks (CNNs). ResNet introduces skip connections, also known as residual connections, which allow the network to learn residual functions. These skip connections pass the input directly to the output of deeper layers, enabling the model to skip over certain layers. This helps in mitigating the vanishing gradient problem, making it easier to train very deep networks.

In the context of an image caption generator, ResNet can play a crucial role in feature extraction from images. The encoder part of an image captioning model typically uses a pre-trained CNN, such as ResNet, to extract meaningful features from the input images. The idea is to leverage the knowledge learned by the pre-trained ResNet model on a large dataset (e.g., ImageNet) to capture high-level features in images.

Here's how the ResNet model can be integrated into an image caption generator:

B. Pre-trained ResNet as Image Encoder

The ResNet model is used as a feature extractor for images.
The model is typically pre-trained on a large dataset for image classification tasks (e.g., ImageNet).
The weights learned during pre-training capture hierarchical and abstract features in images.

C. Feature Extraction

Given an input image, the pre-trained ResNet model is used to extract features from intermediate layers.
The features represent high-level visual information present in the image.

D. Integration with Captioning Model

The extracted image features are then passed to the decoder part of the image captioning model.
The decoder, often implemented as a recurrent neural network (RNN) or transformer, generates a textual description of the image based on the input features.

E. LSTM

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture designed to capture and learn long-term dependencies in sequential data. While LSTMs are commonly used for natural language processing tasks, they can also play a crucial role in image caption generators.
Here's how LSTMs are typically incorporated into an image caption generator:

F. Sequence Modeling

In the context of image captioning, LSTMs are used to model sequential information, such as generating a sequence of words in a sentence.
The LSTM is employed as the decoder part of the image captioning model, taking as input the features extracted from the image.

G. Image Feature Input

The LSTM receives the image features (extracted by a pre-trained CNN like ResNet) as an initial input.
These features serve as the context or starting point for generating the image caption.

H. Word Generation

The LSTM generates words one at a time, considering the context provided by the image features and the previously generated words.
At each time step, the LSTM produces a probability distribution over the vocabulary, and a word is sampled from this distribution.

I. Recurrent Connections:

LSTMs have recurrent connections that allow them to maintain and update an internal memory state, which helps capture long-term dependencies in the sequence.
The internal state is updated at each time step based on the input features and the previously generated word.

J. Training

During training, the model is optimized to minimize the difference between the predicted caption and the ground truth caption.
The loss is computed based on the generated word probabilities at each time step.

K. Word Embeddings

To deal with the discrete nature of words, word embeddings are often used to represent words as continuous vectors.
The LSTM generates these embeddings, which are then used to predict the next word.

By using an LSTM as the decoder in an image caption generator, the model can effectively capture the dependencies between words in a sentence and generate coherent and contextually relevant captions for images.
The combination of a pre-trained image encoder (e.g., ResNet) and an LSTM-based decoder allows the model to leverage both visual information from the image and linguistic context to produce meaningful and descriptive captions.

V. PROPOSED SYSTEM

Our proposed system employs a combination of ResNet-50 and LSTM ensuring a seamless fusion of visual and linguistic information. The ResNet-50 feature vector serves as a foundation for the LSTM to generate contextually relevant captions, effectively marrying the strengths of both modalities. The proposed architecture aims to overcome challenges associated with understanding complex visual scenes and maintaining linguistic context, ultimately leading to improved image captioning performance. The use of ResNet-50 as a feature extractor and LSTM for sequence modeling represents a state-of-the-art approach in the field, aligning with contemporary advancements in deep learning for multimodal tasks.

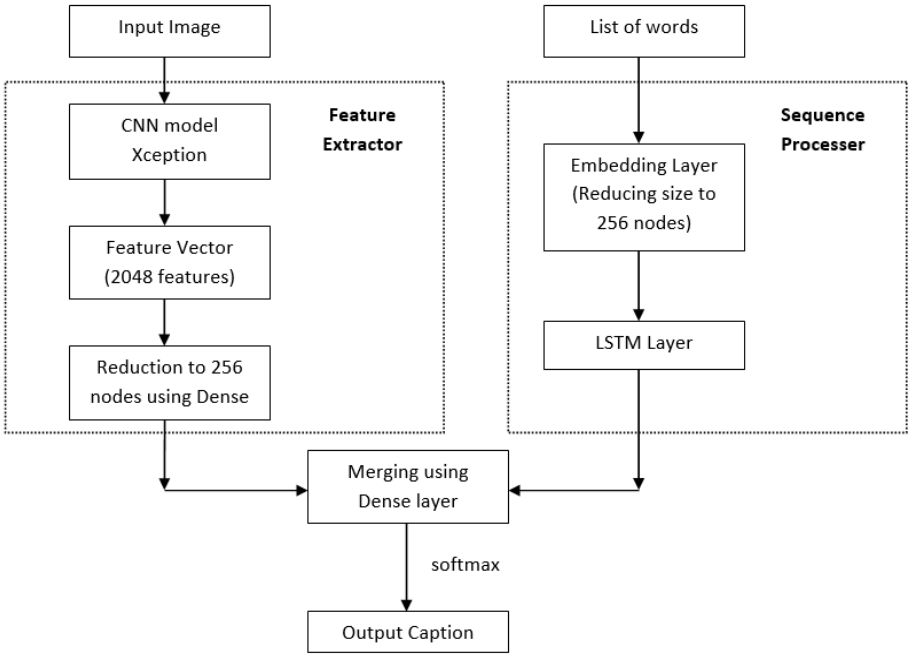


Fig 1. Proposed Architecture

A. *Introduction of Input Design*

In an information system, input is the raw data that is processed to produce output. During the input design, the developers must consider the input devices such as PC, MICR, OMR, etc.
Therefore, the quality of the system input determines the quality of the system output. Well-designed input forms and screens have the following properties −

1) It should serve specific purposes effectively such as storing, recording, and retrieving the information.
2) It ensures proper completion with accuracy.
3) It should be easy to fill and straightforward.
4) It should focus on the user's attention, consistency, and simplicity.
5) All these objectives are obtained using the knowledge of basic design principles regarding –

- What are the inputs needed for the system?
- How end users respond to different elements of forms and screens.

B. *Objectives for Input Design*

The objectives of input design are −

1) To design data entry and input procedures

2) To reduce input volume
3) To design source documents for data capture or devise other data capture methods
4) To design input data records, data entry screens, user interface screens, etc.
5) To use validation checks and develop effective input controls.

C. Output Design

The design of output is the most important task of any system. During output design, developers identify the type of outputs needed and consider the necessary output controls and prototype report layouts.

D. Objectives of Output Design

The objectives of input design are:

1) To develop an output design that serves the intended purpose and eliminates the production of unwanted output.
2) To develop the output design that meets the end user's requirements.
3) To deliver the appropriate quantity of output.
4) To form the output in an appropriate format and direct it to the right person.
5) To make the output available on time for making good decisions.

VI. RESULTS

The performance of the image captioning system using the CNN and LSTM model showcases its ability to generate precise and context-appropriate captions across various images. The model's efficacy is measured using recognized metrics like BLEU, METEOR, and CIDEr. The BLEU scores reveal a high linguistic resemblance between the captions created by the model and the reference captions, confirming the model's capability to grasp the subtleties of different scenes. METEOR scores gauge the fluency and coherence of the captions, whereas CIDEr scores emphasize the model's success in producing varied and descriptive captions that resonate with human judgment. Comparisons with existing methods demonstrate that the CNN-LSTM model excels in terms of caption quality and contextual comprehension. Moreover, qualitative reviews through human evaluations confirm the model's strength in producing captions that are both linguistically sound and semantically substantial. Overall, these results highlight the effectiveness of the CNN-LSTM hybrid approach in the field of image captioning, underlining its potential for broad applications in computer vision and natural language processing.

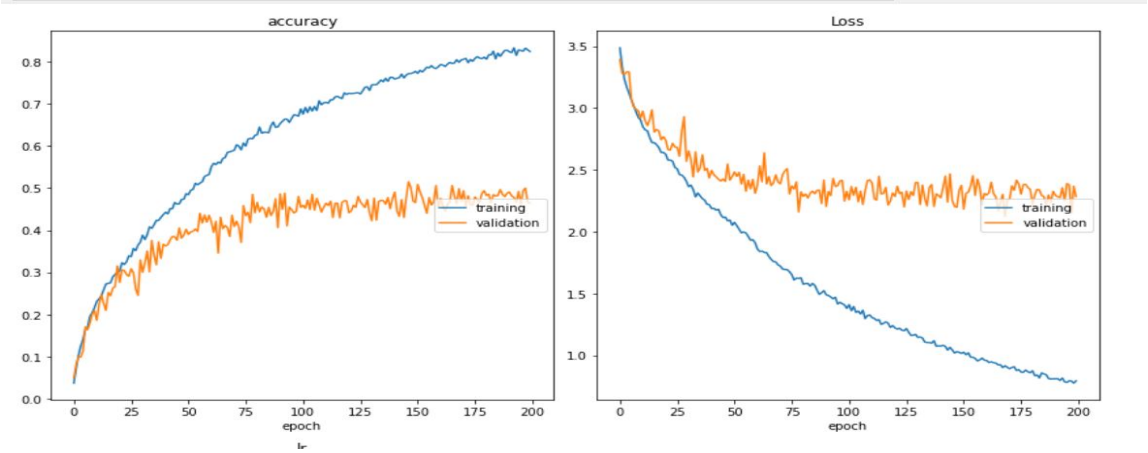


Fig 2. Graph of accuracy and loss of the model

VII. CONCLUSION AND FUTURE WORK

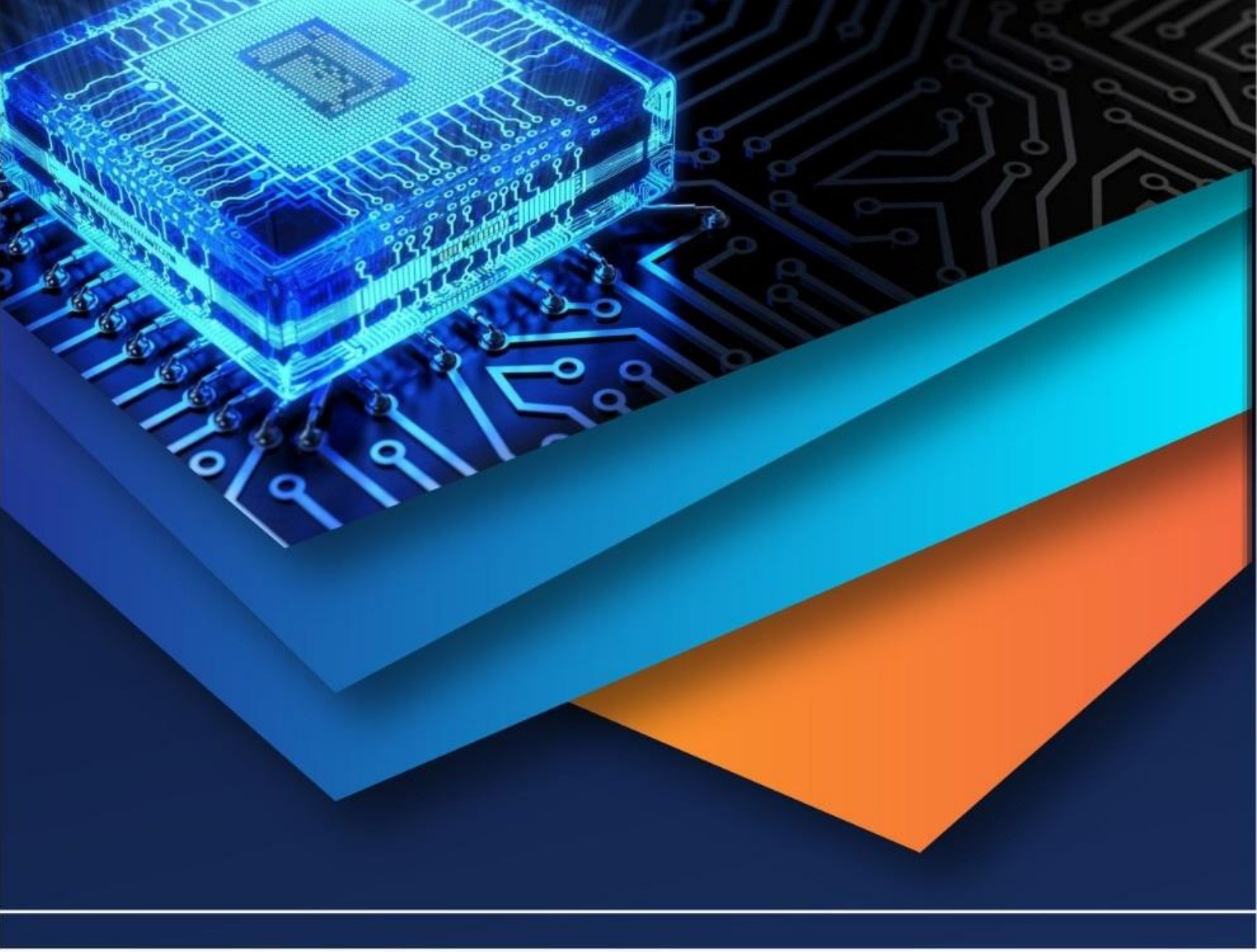
The image caption generator that combines Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks has demonstrated significant effectiveness and efficiency in creating descriptive captions for images. The CNN-LSTM model effectively harnesses the CNN layers to extract pertinent features and spatial information from images, and the LSTM layers to sequence this information and generate coherent, contextually appropriate captions. The merging of these two architectures successfully overcomes the hurdles associated with image comprehension and natural language creation, illustrating the effective synergy between visual processing and sequential data handling.

This project not only showcases the capabilities of deep learning in handling multimodal tasks but also highlights the importance of integrating specialized neural network architectures to enhance performance in complex applications like image captioning.

Several potential enhancements and extensions could further develop the image caption generator that utilizes CNN and LSTM technologies. First, introducing advanced architectures like attention mechanisms, transformer models, or pre-trained language models such as BERT could enhance the model's ability to discern complex visual-textual relationships. Expanding the training dataset with a wider variety of images could also improve the model's ability to generalize and describe a broader spectrum of visuals. Tailoring the model to specific domains or tasks, such as medical imaging or satellite imagery, may also prove beneficial, enabling specialized applications. Additionally, developing methods to increase the model's interpretability and controllability could lead to a deeper understanding and more precise management of the captioning process. Finally, implementing the model in real-world scenarios and collecting user feedback would offer valuable insights into its practical effectiveness and highlight areas for further improvement.

REFERENCES

[1] Show and Tell: A Neural Image Caption Generator by Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan et al. CVPR 2015
[2] Neural Image Caption Generation with Visual Attention by Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan ICML 2015
[3]] Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering by Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen (2019)
[4] Image Captioning with Semantic Attention by Qi Wu, Chunhua Shen, Anton van den Hengel. (CVPR 2017)
[5] DenseCap: Fully Convolutional Localization Networks for Dense Captioning by Vdovichenko et al. Justin Johnson, Andrej Karpathy, Li Fei-Fei, CVPR 2016
[6] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and Tell: A Neural Image Caption Generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[7] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., & Zemel, R. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In International Conference on Machine Learning (ICML).
[8] Mao, J., Xu, W., Yang, Y., Wang, J., & Huang, Z. (2014). Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). In International Conference on Learning Representations (ICLR).
[9] Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[10] Karpathy, A., & Fei-Fei, L. (2015). Deep Visual-Semantic Alignments for Generating Image Descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[11] Chen, X., & Lawrence, Zitnick, C. L. (2015). Mind's Eye: A Recurrent Visual Representation for Image Caption Generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[12] Chen, X., & Lawrence Zitnick, C. (2017). Learning to See by Moving. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).
[13] Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). DenseCap: Fully Convolutional Localization Networks for Dense Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[14] Wu, Q., Shen, C., & Dick, A. (2016). Image Captioning and Visual Question Answering Based on Attributes and External Knowledge. IEEE Transactions on Pattern Analysis and Machine Intelligence.
[15] 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 (NeurIPS).
[16] Xu, J., Mei, T., Yao, T., & Rui, Y. (2015). MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[17] Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollár, P., ... & He, K. (2015). From captions to visual concepts and back. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[18] Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., & Chua, T. S. (2017). SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[19] Wang, J., Yang, Y., Mao, J., Huang, Z., & Yuille, A. L. (2016). Cnn-rnn: A unified framework for multi-label image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[20] Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). Stacked attention networks for image question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)