IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, May 2024

# Automatic Image Captioning using Deep Learning

**A. T. V Sai Sony[1], N. Sailaja[2], B. Roshini[3], K. Prasad[4],** Mr. S. Syam Kumar[5]
Students, Department of Computer Science and Engineering (Data Science)[1,2,3,4]
Assistant Professor, Department of Computer Science and Engineering (Data Science)[1,2,3,4]
Dadi Institute of Engineering & Technology (Autonomous), Anakapalle, India

**Abstract:** *Image captioning aims to detect this information by describing the image content through image and text processing techniques. Now-a-days image captioning is one of the recent and growing research problem. 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 for getting better and precise results. So, we came up with the idea of developing a 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 combination of CNN and LSTM for developing the model. The proposed model is trained with three Convolutional Neural Network architecture such as Inception-v3,Xception, ResNet50 for feature extraction and Long ShortTerm Memory for generating the relevant captions. Among these models, the best combination based on the accuracy of the model. It is trained using the Flicker8k dataset.*

**Keywords:** Object detection, Image captioning, Deep neural networks, Semantic-instance segmentation

## I. INTRODUCTION

Automatic image captioning using deep learning has emerged as a compelling research area with wide-ranging applications in computer vision and natural language processing. The task involves generating descriptive textual captions for images, enabling machines to understand and communicate visual content in human-like language. This technology holds immense potential for assisting visually impaired individuals, enhancing image retrieval systems, and enabling better content understanding in multimedia applications. The advent of deep learning techniques, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), has significantly advanced image captioning capabilities by enabling end-to-end learning from raw image pixels to natural language descriptions. By leveraging large-scale image-caption datasets and sophisticated model architectures, recent research has achieved remarkable progress in generating accurate and contextually relevant captions for diverse images. However, challenges persist in capturing fine-grained details, handling complex scenes, and ensuring model robustness across different domains and languages. In this literature survey, we explore seminal works and recent advancements in automatic image captioning, highlighting key techniques, datasets, evaluation metrics, and future research directions in this rapidly evolving field.

Furthermore, automatic image captioning serves as a fundamental task in multimodal artificial intelligence, bridging the gap between visual perception and natural language understanding. By enabling machines to comprehend and articulate visual content, image captioning facilitates applications such as image indexing, content-based image retrieval, and personalized content recommendation systems. Recent advancements in deep learning, coupled with the availability of large-scale annotated datasets, have propelled image captioning research forward, leading to increasingly sophisticated models capable of generating detailed and contextually coherent descriptions for a wide variety of images. Despite significant progress, ongoing challenges include handling ambiguity, generating diverse and creative captions, and understanding complex visual scenes with multiple objects and interactions. Addressing these challenges requires continued exploration of novel model architectures, multimodal fusion techniques, and robust evaluation methodologies, paving the way for more intelligent and expressive image captioning systems.

## II. LITERATURE SURVEY

The literature on automatic image captioning using deep learning encompasses a diverse range of approaches, techniques, and advancements. Seminal works such as "Show and Tell" by Vinyals et al. introduced the concept of

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, May 2024

directly mapping images to natural language descriptions using convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Subsequent research has focused on improving model architectures, incorporating attention mechanisms, and leveraging multimodal information for more informative captions. Works like "Show, Attend and Tell" by Xu et al. introduced attention mechanisms to focus on relevant image regions during caption generation, while "Bottom-Up and Top-Down Attention" by Anderson et al. combined bottom-up image features with top-down attention mechanisms for improved captioning performance. Additionally, transformer-based models, such as the vision transformer (ViT) and image transformer (DeiT), have shown promise in directly processing image pixels for caption generation. Benchmark datasets like COCO and Flickr30k have facilitated standardized evaluation, while novel training techniques like self-critical sequence training have improved caption quality metrics. Challenges remain in capturing fine-grained details, handling diverse image types and domains, and ensuring model generalization. Ongoing research aims to address these challenges through innovations in model architectures, data augmentation strategies, and multimodal fusion techniques, driving forward the capabilities and applicability of automatic image captioning systems.
Continuing the literature survey, recent advancements in automatic image captioning have explored novel model architectures and training methodologies to further improve caption quality and model generalization. Transformer-based architectures, originally designed for natural language processing tasks, have been adapted to handle visual input and have shown promise in capturing global contextual information for generating coherent captions. Works such as the Visual Transformer (ViT) and Dense ViT have demonstrated competitive performance in image captioning tasks, highlighting the effectiveness of transformer-based models in processing visual data. Additionally, research efforts have focused on leveraging pre-trained visual features from large-scale image classification models, such as ResNet and EfficientNet, to initialize captioning models and facilitate faster convergence and better performance. Domain adaptation techniques have been explored to enhance model robustness across diverse image domains and improve caption quality for specific application scenarios. Furthermore, advancements in reinforcement learning-based approaches, such as actor-critic methods and policy gradient algorithms, have enabled more effective optimization of captioning metrics directly during training, leading to improved caption quality and coherence. Multimodal fusion strategies, including late fusion, early fusion, and cross-modal attention mechanisms, continue to be a focal point of research, with efforts aimed at effectively integrating visual and textual information to generate more informative and contextually relevant captions. Overall, the literature survey highlights the breadth of research in automatic image captioning, spanning from model architectures and training techniques to multimodal fusion strategies and domain adaptation methods, with ongoing efforts focused on advancing the state-of-the-art in captioning quality and applicability across various domains and applications.

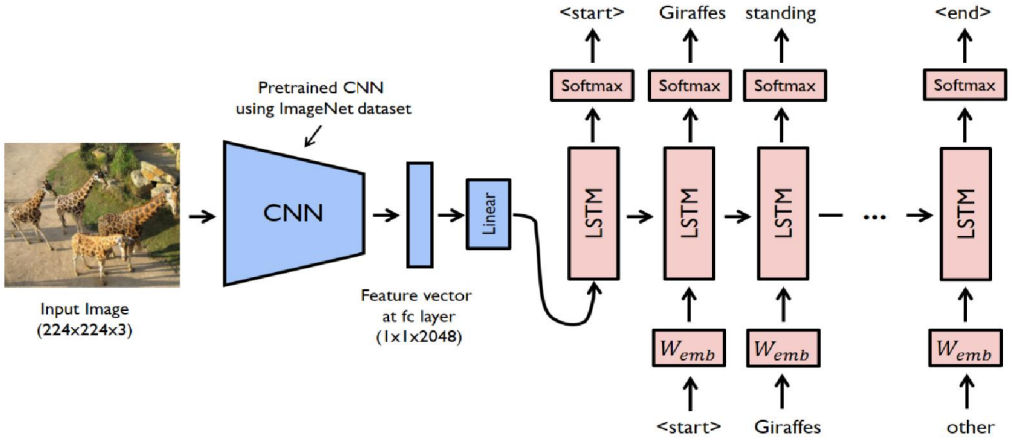
## III. METHODOLOGY

The research methodology architectural analysis is examined in Figure .The image captioning generates description for a image which is stuctured by gathering a large dataset of images paired with corresponding captions. Common datasets include MS COCO, Flickr30k, and Pascal VOC.Preprocess the images (resize, normalize, etc.) and tokenize the captions (convert text into numerical representations).Choosing a suitable deep learning architecture for image captioning, such as CNN-RNN architectures (e.g., CNN + LSTM).Explore pre-trained models for both image feature extraction (e.g., VGG, ResNet, or Inception) and language modeling and then Model Training by splitting the dataset into training, validation, and test sets.Train the deep learning model using the training set. During training evaluating the performance of the trained model using various metrics such as BLEU, METEOR, CIDEr, ROUGE, and SPICE Monitor the model's performance over time and update it as necessary to adapt to changing requirements or datasets.
The methodology for developing an automatic image captioning system using deep learning entails a systematic approach to leverage deep neural networks for generating descriptive captions from images. It begins with the collection of a substantial dataset comprising images paired with corresponding captions, sourced from repositories like MSCOCO or Flickr30k. Following data acquisition, preprocessing steps involve standardizing image sizes, tokenizing captions, and encoding data into numerical representations. Next, a suitable deep learning architecture is selected, considering models like CNN-RNN hybrids or transformer-based architectures, with image feature extraction typically handled by pre-trained convolutional neural networks such as InceptionV3 or ResNet. Building the model involves implementing the chosen architecture using frameworks like TensorFlow or PyTorch, followed by training on a split

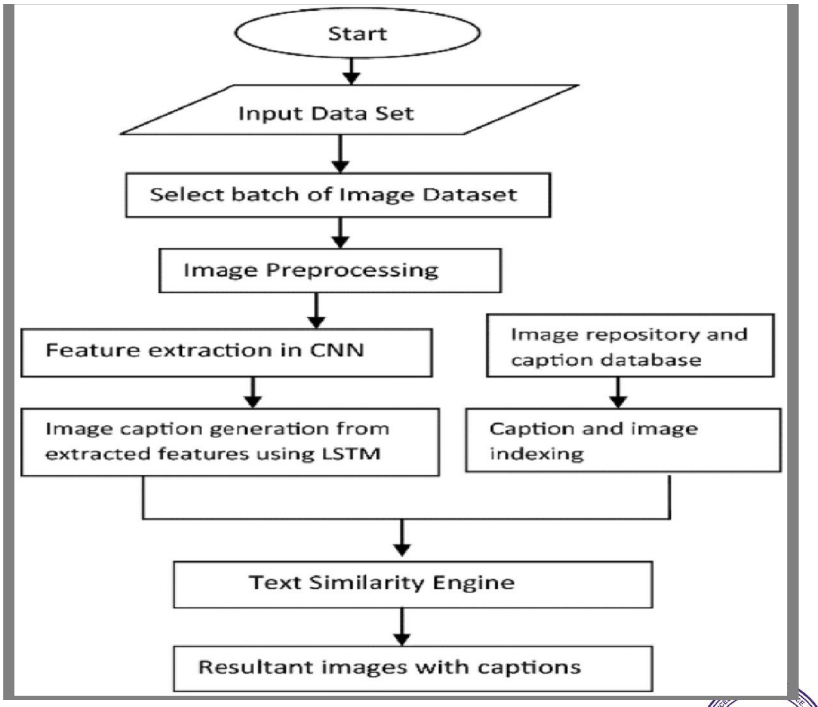
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, May 2024

dataset. Hyperparameter tuning then optimizes model performance, employing techniques such as grid search or random search. Evaluation metrics like BLEU, METEOR, and CIDEr assess model performance, guiding fine-tuning and preventing overfitting. Testing on a separate test set gauges the model's generalization ability, with qualitative assessment of generated captions. Deployment into a production environment follows, integrating the model with a user interface for practical use. Continuous monitoring and maintenance ensure sustained performance through periodic retraining and addressing any drift or performance issues that arise. This structured methodology facilitates the development of robust and effective automatic image captioning systems capable of generating accurate and contextually relevant captions for diverse images.

## IV. PROPOSED SYSTEM




## V. FLOWCHART



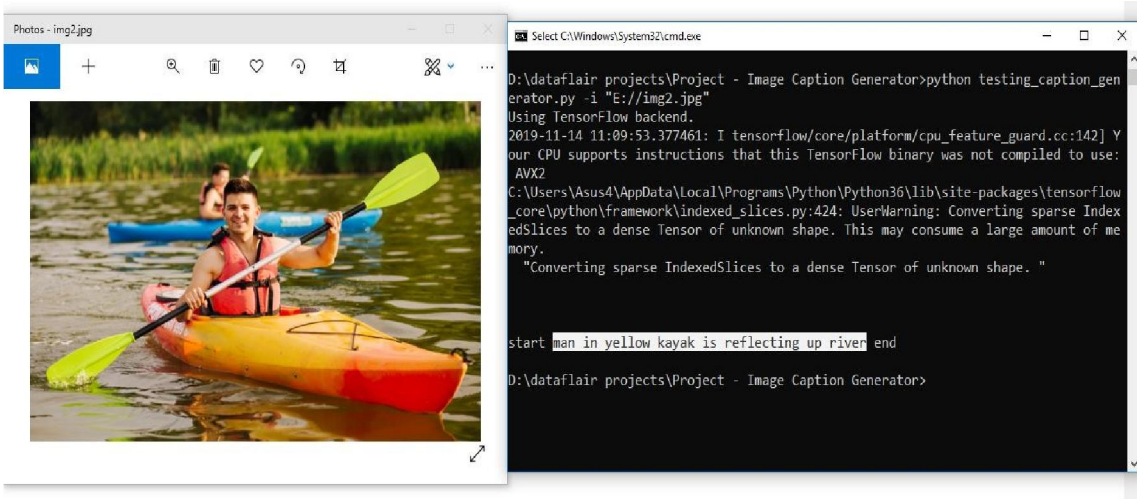
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, May 2024

## VI. IMPORT DATASET

```
In [3]: import os
import pickle
import numpy as np
from tqdm.notebook import tqdm

from tensorflow.keras.applications.vgg16 import VGG16, preprocess_input
from tensorflow.keras.preprocessing.image import load_img, img_to_array
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
from tensorflow.keras.models import Model
from tensorflow.keras.utils import to_categorical, plot_model
from tensorflow.keras.layers import Input, Dense, LSTM, Embedding, Dropout, add

WARNING:tensorflow:From C:\ProgramData\anaconda3\Lib\site-packages\keras\src\losses.py:2976: The name tf.losses.sparse_softmax_
cross_entropy is deprecated. Please use tf.compat.v1.losses.sparse_softmax_cross_entropy instead.

In [4]: BASE_DIR = 'C:\AUTOMATIC'
WORKING_DIR = 'C:\AUTOMATIC'

In [5]: # load vgg16 model
model = VGG16()
# restructure the model
model = Model(inputs=model.inputs, outputs=model.layers[-2].output)
# summarize
print(model.summary())

WARNING:tensorflow:From C:\ProgramData\anaconda3\Lib\site-packages\keras\src\backend.py:1398: The name tf.executing_eagerly_out
side_functions is deprecated. Please use tf.compat.v1.executing_eagerly_outside_functions instead.
```

| Layer (type) | Output Shape | Param # |
|---|---|---|
| input_1 (InputLayer) | [(None, 224, 224, 3)] | 0 |
| block1_conv1 (Conv2D) | (None, 224, 224, 64) | 1792 |
| block1_conv2 (Conv2D) | (None, 224, 224, 64) | 36928 |
| block1_pool (MaxPooling2D) | (None, 112, 112, 64) | 0 |
| block2_conv1 (Conv2D) | (None, 112, 112, 128) | 73856 |
| block2_conv2 (Conv2D) | (None, 112, 112, 128) | 147584 |
| block2_pool (MaxPooling2D) | (None, 56, 56, 128) | 0 |
| block3_conv1 (Conv2D) | (None, 56, 56, 256) | 295168 |
| block3_conv2 (Conv2D) | (None, 56, 56, 256) | 590080 |
| block3_conv3 (Conv2D) | (None, 56, 56, 256) | 590080 |
| block3_pool (MaxPooling2D) | (None, 28, 28, 256) | 0 |
| block4_conv1 (Conv2D) | (None, 28, 28, 512) | 1180160 |
| block4_conv2 (Conv2D) | (None, 28, 28, 512) | 2359808 |
| block4_conv3 (Conv2D) | (None, 28, 28, 512) | 2359808 |
| block4_pool (MaxPooling2D) | (None, 14, 14, 512) | 0 |

## VII. RESULT



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, May 2024

## VII. CONCLUSION

In conclusion, the development of an automatic image captioning system using deep learning represents a significant advancement in the intersection of computer vision and natural language processing. Through the systematic methodology outlined, leveraging large-scale datasets, sophisticated model architectures, and rigorous training and evaluation processes, we can create robust and effective systems capable of generating descriptive captions for diverse images. By harnessing the power of deep neural networks, such systems can understand and articulate visual content in human-like language, unlocking a myriad of applications across various domains. The methodology's iterative nature allows for continual refinement and improvement, ensuring that the system remains adaptive to evolving data trends and user needs. Additionally, the deployment of such systems into production environments enables practical use cases, including assistive technologies for visually impaired individuals, content retrieval systems, and multimedia content understanding platforms. Looking ahead, ongoing research and innovation in automatic image captioning will continue to push the boundaries of what is possible, driving towards more accurate, contextually relevant, and universally applicable captioning solutions that enhance human-computer interaction and enrich user experiences in the digital landscape.The systematic methodology presented for developing an automatic image captioning system underscores the transformative potential of deep learning in bridging the gap between visual perception and natural language understanding. By following this structured approach, we can harness the power of large-scale datasets, state-of-the-art model architectures, and advanced training techniques to create highly accurate and contextually relevant captioning systems. These systems have far-reaching implications across diverse fields, including accessibility, content retrieval, and multimedia analysis. Furthermore, the iterative nature of the methodology ensures adaptability to evolving data trends and user needs, facilitating continuous improvement and refinement of the captioning system. As these systems are deployed into production environments, they offer tangible benefits such as improved accessibility for visually impaired individuals, enhanced search capabilities in image databases, and enriched multimedia content understanding. Looking ahead, ongoing research efforts will further drive innovation in automatic image captioning, pushing the boundaries of captioning quality, domain adaptability, and multimodal understanding. Ultimately, these advancements will empower users to interact with visual content more intuitively and effectively, ushering in a new era of human-computer collaboration and communication.

## REFERENCES

**[1].** Abhaya Agarwal and Alon Lavie. 2008. Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with Human rankings of machine translate on output. In Proceedings of the Third Work shop on Statistical Machine Translation.AssociationforComputationalLinguistics,115–118.

**[2].** Ahmet Aker and Robert Gaizauskas. 2010. Generating image descriptions using dependency relational patterns. In Proceedings of the 48th annual meeting of the association for computational linguistics Association for ComputationalLinguistics,1250–1258.

**[3].** Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice : Semantic propositional image Caption evaluation. In European Conference on Computer Vision. Springer,382–398.

**[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. ar Xiv preprint arXiv:1707.07998 (2017).

**[5].** Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, Trevor Darrell, Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, et al.2016.Deep compositional captioning: Describing novel object categories without paired training data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.