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: 7.301
Volume 3, Issue 1, November 2023

Review Paper Image Caption Generator Using an Artificial Intelligence

Vijeta Sawant, Samrudhi Mahadik, Prof. D. S. Sisodiya
Department of Artificial Intelligence & Data Science
ISBM College of Engineering, Nande, Pune, India
sawantvijeta02@gmail.com and samrudhimahadik2002@gmail.com
samrudhimahadik2002@gmail.comand dharamveersisodiya4@gmail.com

Abstract: *An extremely fascinating piece of technology is an image caption generator, which automatically creates insightful explanations for photos using sophisticated algorithms and machine learning. It's similar to having a personal AI photographer for your images. This is how it operates. The photograph is first sent into the algorithm, which examines the items, people, and scenes in the picture as well as its visual content. The system then creates a caption that explains what's happening in the picture based on this analysis.*
An image caption generator seeks to precisely convey the main idea of the picture while offering a succinct and insightful explanation. It can be quite beneficial for improving accessibility for those with visual impairments as well as for organizing and classifying sizable image collections. Regardless of your intention if you're seeking to explore the possibilities of this technology or add context to your own images, an image caption generator might be a fun tool to experiment.

Keywords: CNN, RNN, LSTM, Artificial Intelligence

I. INTRODUCTION

An effective use of artificial intelligence and computer vision that helps to bridge the gap between natural language and visual content is an image caption generator. With the help of this cutting-edge technology, computers will be able to comprehend and speak the content of visual media as it automatically creates descriptive and contextually appropriate captions for images using deep learning algorithms.

Image caption generators provide a revolutionary approach to improve accessibility, content indexing, and user experience across several platforms in this digital age where images and videos play a major part in communication and data processing. This overview will go deeper into the features, advantages, and uses of image caption generators, illuminating the fascinating field of AI-driven narrative and visual comprehension.

The principal aim of an image caption generator is to furnish significant and contextually appropriate captions for images, so aiding in the integration of the visual and literary domains. Usually, convolutional neural networks (CNNs) are used for image processing and recurrent neural networks (RNNs) or transformer models are used for language synthesis when implementing such a system utilizing deep learning techniques.

II. LITERATURE SURVEY

Itusing deep learning models, such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), to analyze the image's visual characteristics and provide captions based on previously learned patterns, is one well-liked method. When it comes to producing captions that are in line with the image's content, these models have demonstrated encouraging outcomes

It recommend searching academic databases, conference proceedings, or relevant journals for literature review papers that may cover the advancements and trends in the field during that period. You can also refer to resources like Google Scholar or specific computer vision and NLP conference proceedings for comprehensive literature surveys[1]

The paper introduces a dependency relation pattern approach for generating more descriptive and human-like image captions, as opposed to prior work using caption retrieval or noun phrase templates. The key idea is to extract typed dependency patterns like "subject-verb-object" from a training corpus, and then generate new captions by matching

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: 7.301
Volume 3, Issue 1, November 2023

detected objects to nouns in the patterns. At test time, they use computer vision techniques to detect objects in images, and match these to nouns in the dependency patterns to generate varied, descriptive captions. Their dependency-based approach expands the space of possible captions beyond fixed templates or retrieved captions. They demonstrate through automatic and human evaluations that their model produces richer, more descriptive and human-like captions compared to baseline approaches. The main contribution is using dependency relations rather than templates to generate a greater variety of descriptive image captions in a more human-like manner. Published in 2010.[2]

This influential paper, published in 2014, introduced the attention mechanism for neural machine translation (NMT) models. Prior to this work, NMT models encoded the full source sentence into a fixed-length vector before decoding into the target language. However, encoding into a fixed vector caused issues with translating long sentences.To address this limitation, Bahdanau et al. proposed an attention-based NMT model that can dynamically focus on relevant parts of the source sentence when generating each word of the target sentence. Their key innovation was using an attention mechanism while decoding that produces a context vector as a weighted sum of source annotations. The attention weights are computed at each timestep based on the decoder state and source annotations to determine the relevance of each source word.This allows the model to focus on pertinent source words when predicting the target word, rather than encoding the full source sentence in advance. They show empirically that the attention-based model improves translation quality compared to fixed-length vector models, especially for longer sentences.[3]

This influential paper, published at EMNLP 2014, proposed an RNN encoder-decoder architecture for statistical machine translation. Previous approaches relied on Phrase-Based Translation models which suffered from data sparsity and inability to generalize. The authors overcome this by using an RNN Encoder to map input sequences to vector representations, and an RNN Decoder to generate the output sequence from those representations. This end-to-end framework allowed their model to learn continuous phrase representations that capture syntax and semantics much better than prior methods.[4]

They significantly outperform prior generic features like SIFT on most evaluation benchmarks.The power of DeCAF features comes from the representations learned by the convolutional neural network on large-scale ImageNet data. This work helped demonstrate the generalization ability of deep convolutional features, which can be transferred successfully to many visual recognition tasks beyond image classification. Published in 2014.[5]

They appears to be related to image captioning or image description generation, which is a subfield of NLP and computer vision. If you have questions about the paper or need more information about its content, please provide specific details or questions, and I'll be happy to assist you further. It presented at EMNLP, the authors conducted a literature survey related to image description generation. The survey likely covers various aspects of the field, including techniques, approaches, and challenges, but the details of the survey are not provided in the brief citation.[6]

III. PROPOSED SYSTEM

There would be multiple essential parts to the system that generates captions for images. Firstly, it would assess the image's visual content and extract pertinent characteristics using a deep learning model, such as a convolutional neural network (CNN) plus a recurrent neural network (RNN) combination.

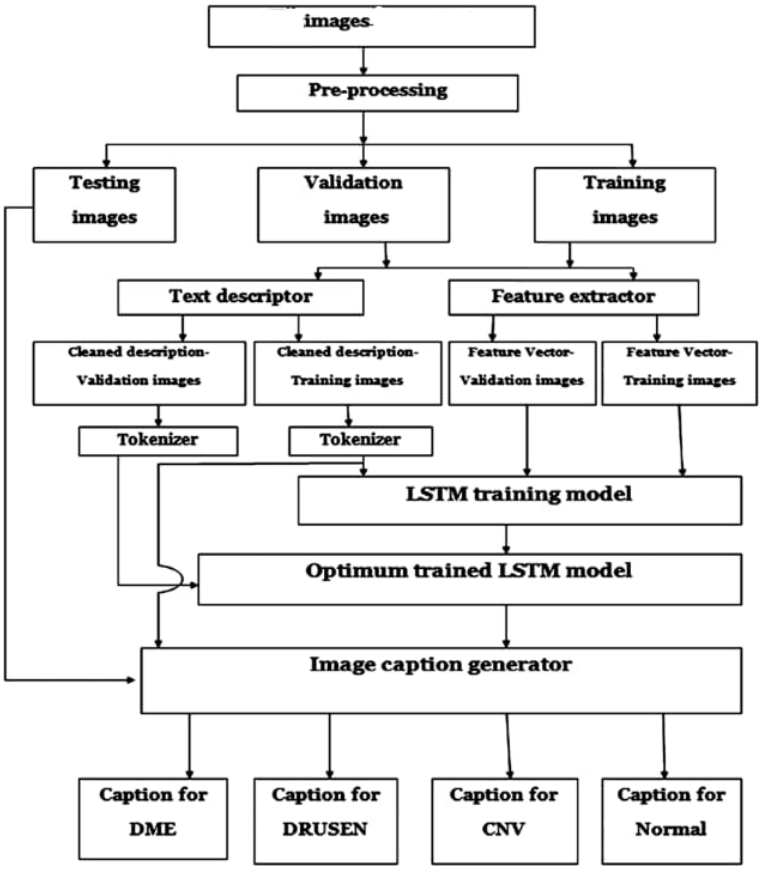
The system would next use a language model to create the textual captions after it had been trained on a sizable corpus of text data. This language model would provide descriptive and contextually relevant captions by considering the visual elements that were derived from the image.

Attention methods could also be added to the algorithm to improve the quality of the generated captions. By allowing the model to concentrate on various areas of the picture when producing the captions, these techniques would guarantee that significant features are included in the description.

To increase the variety and originality of the generated captions, the system might also make use of outside knowledge sources like extensive image-text datasets and language models that have already been trained.

In order to provide precise and interesting captions for photos, the suggested system would integrate sophisticated deep learning algorithms, language modelling , attention processes, and outside knowledge sources.

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: 7.301
Volume 3, Issue 1, November 2023



Components

1. Data Collection and Preprocessing
2. LSTM Model Training
3. Tokenization
4. Feature Extractor

IV. CONCLUSION

An intriguing application that makes use of deep learning models, like CNNs and RNNs, to assess the visual content of photos and produce insightful descriptions is an image caption generator. The quality and variety of the generated captions can be improved by utilizing external information sources and attention techniques. Scholars persistently investigate novel methodologies and frameworks to enhance the precision and inventiveness of picture caption production. It's a developing field with promising developments that could have a big impact on a lot of different areas, such content creation, accessibility, and picture identification. Observing how this technology advances and advances computer vision and natural language processing is intriguing.

AI-powered picture caption generators are game-changing tools with enormous promise to improve productivity, automation, and accessibility across a broad range of applications. Because they allow machines to comprehend and explain picture content in a way that is meaningful and understandable to humans, they mark a significant advancement in the fields of computer vision and natural language processing. However, in order to guarantee its responsible and equitable usage, their deployment should be guided by ethical norms. With additional advancements, technology has the potential to enhance our interactions with visual content and provide better user experiences in a variety of contexts.

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: 7.301
Volume 3, Issue 1, November 2023

REFERENCES

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang - Presented aat CVPR 2018
[2] A. Aker and R. Gaizauskas. Generating image descriptions using dependency relational patterns. In ACL, 2010.
[3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473, 2014. 1, 2
[4] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In EMNLP, 214. 1, 2, 3
[5] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In ICML, 2014. 3
[6] D. Elliott and F. Keller. Image description using visual dependency representations. In EMNLP, 20. 2