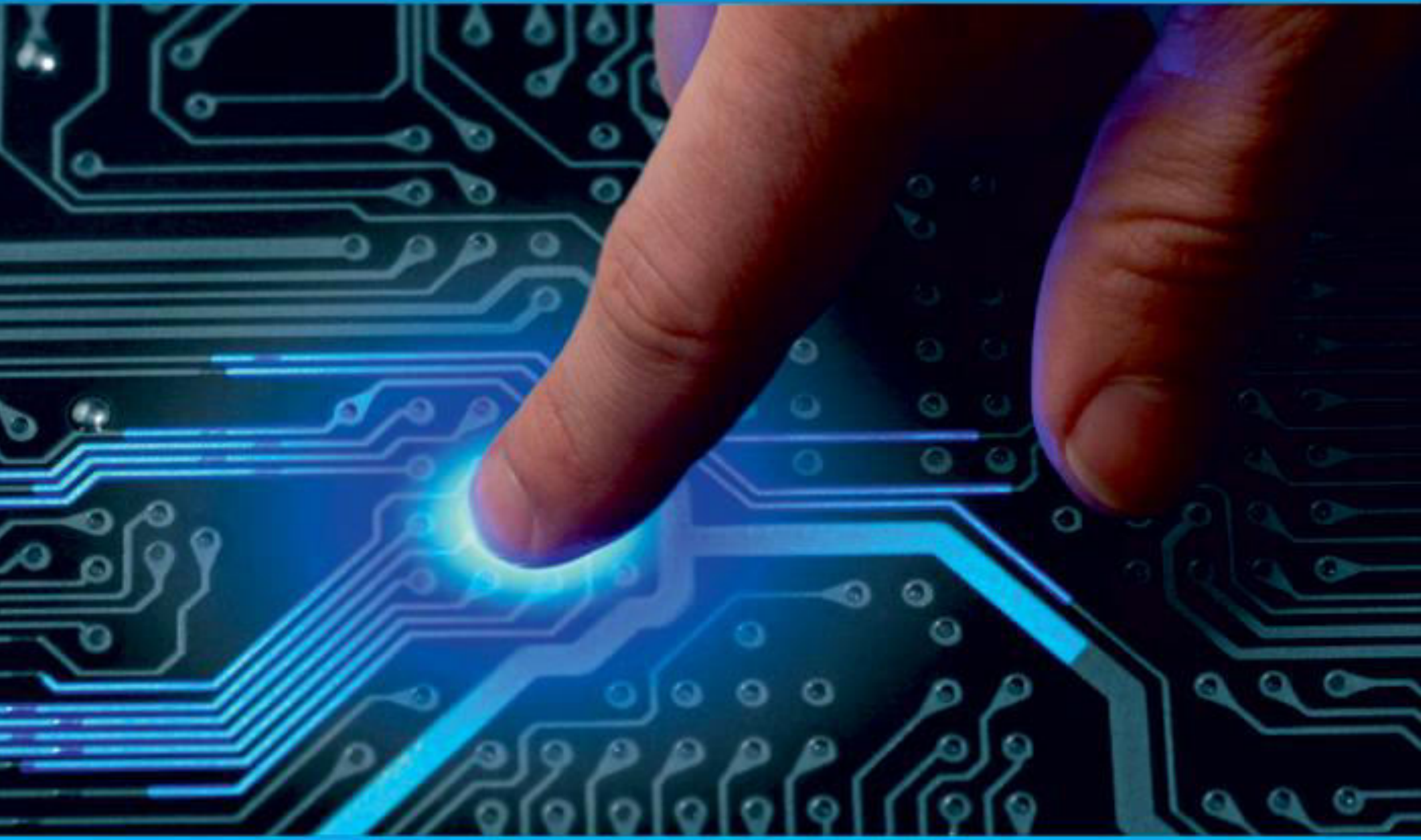
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 3, March 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

| DOI: 10.15680/IJIRCCE.2024.1203171 |

CAPTION CRAFT:Contextual Image Captioning

Surendra Tedla[1], Nasar Hussain Shaik[2], Sathwik Vemula[3], Suma Latha K[4]

Department of Computer Science and Engineering, KKR & KSR Institute of Technology and Sciences

(Affiliated to JNTUK), Guntur, India[1,2,3]

Associate Professor, Department of Computer Science and Engineering, KKR & KSR Institute of Technology and Sciences (Affiliated to JNTUK), Guntur, India[4]

ABSTRACT: In today's world of social media, almost everyone is a part of social platform and actively interacting with each other through internet. People on social media upload many pictures on their social media accounts with different captions. Thinking about the appropriate caption is a tedious process. Caption is important to effectively describe the content and meaning of a picture.

Image captioning is one of the applications of Deep Learning which involves fusion of work done in computer vision and natural language processing, and it is typically performed using Encoder-Decoder architecture. In this work, the aim is to improve the performance and explain ability of generating captions for images of different types and resolutions in way a user can use it accordingly. CNN (convolution neural network) and LSTM (long short-term memory) are used using the concept of encoder-decoder along with attention network to build this model. A CNN is used for image feature extraction purpose where only the important features are extracted from the resultant image. LSTM is used to predict the next word in the sentence.

KEYWORDS: Image, Caption, CNN, LSTM, Neural Networks.

I. INTRODUCTION

In the past few years, computer vision in the image processing area has made significant progress, like image classification and object detection. Benefiting from the advances of image classification and object detection it becomes possible to automatically generate captions to understand visual content of an image. The goal of image captioning is to automatically generate descriptions for a given image.

Making a computer system detect objects and describe them using natural language processing (NLP) in an age-old problem of Artificial Intelligence. This was considered an impossible task by computer vision researchers till now. With the growing advancements in Deep learning techniques, availability of vast datasets, and computational power, models are often built which will generate captions for an image.

Image caption generation is a task that involves image processing and natural language processing concepts to recognize the context of an image and describe them in a natural language like English or any other language. It generates syntactically and semantically correct sentences. In this paper, we present a deep learning model to describe images and generate captions using computer vision and machine translation. Image caption generators can find applications in Image segmentation as used by Facebook and Google Photos, and even more so, its use can be extended to video frames. They will easily automate the job of a person who has to interpret images.

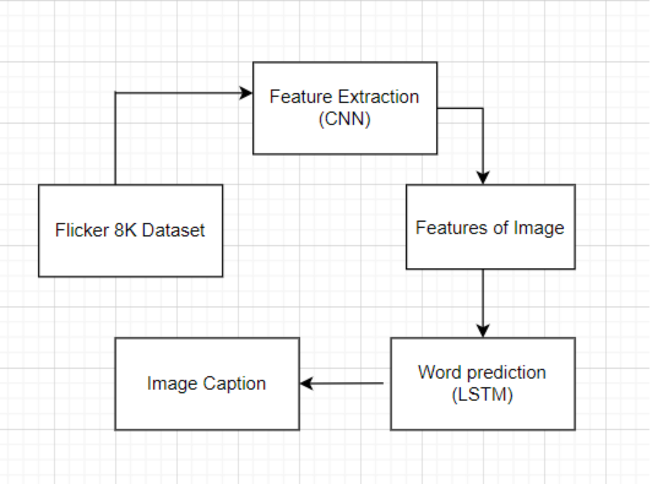


Fig 1: Graphical model of Image Captioning

II. FEATURE EXTRACTION

Convolutional Neural networks are specialized deep neural networks that can process the data that has input shape like a 2D matrix. Images can be easily represented as a 2D matrix. CNN is crucial in working with images. It takes as input an image, assigns importance (weights and biases) to various aspects/objects in the image, and differentiates one from the other. The below Figure 3.3 demonstrates the architecture of CNN for object classification. The CNN makes use of filters which help in feature learning much the same as a human brain identifying objects in time and space. The architecture performs a better fitting to the image dataset due to the reduction in the number of parameters involved (from 2048 to 256) and the reusability of weights.

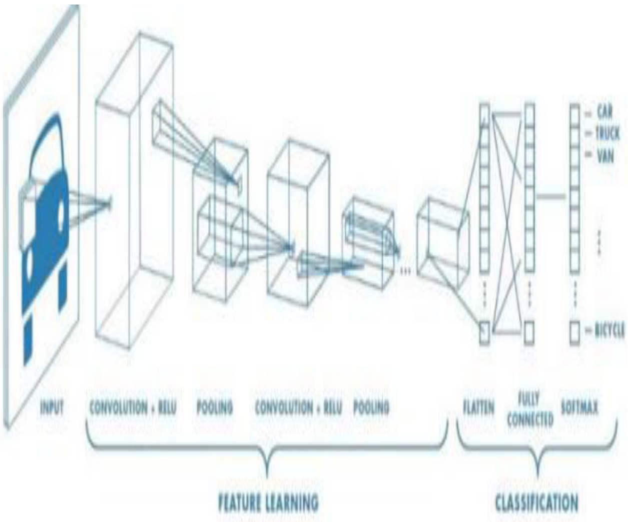


Fig 2: Working of Feature extraction from a image

III. CONTEXT ENCODING

Long Short-Term Memory networks – usually just called "LSTMs" – are a special kind of RNN, capable of learning long-term dependencies. Remembering information for long periods is practically their default behaviour, and this behaviour is controlled with the help of "gates". While RNNs process single data points, LSTMs can process entire sequences. Not only that, but they can also learn which point in the data holds importance, and which can be thrown away.

Hence, the only relevant information is passed on to the next layer. Their ability to remember information for extended periods of time which makes them widely used in various research areas.

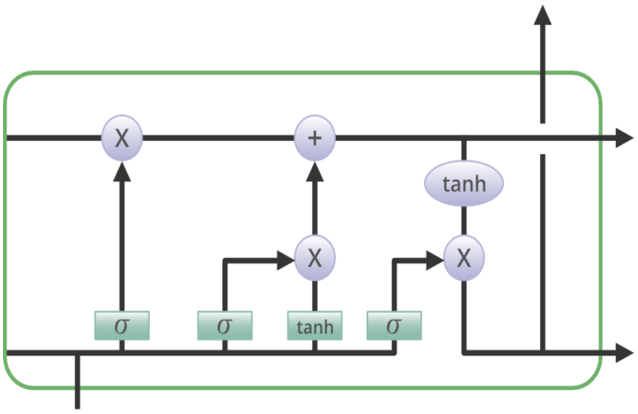


Fig 3: Graphical model of Context Encoding

IV. EXISTING SYSTEM

The current landscape of image captioning technology is marked by limitations in accuracy and inclusivity. Existing systems primarily rely on image analysis algorithms, which often fail to capture the nuanced details and deeper meanings

within visual content, resulting in captions lacking depth and resonance with users. Moreover, the absence of user-centric features exacerbates the issue of impersonalized captions, as they may feel generic and detached from the diverse perspectives of users, leading to a lack of meaningful engagement. Furthermore, the existing systems lack accessibility features, such as audio support for the visually impaired, which limits their usability and inclusivity. Without alternative means of engagement, individuals with visual impairments are effectively excluded from accessing and comprehending the content conveyed through image captions.

V. PROPOSED SYSTEM

The proposed deep learning and AI-based model aim to address this gap by considering user preferences and interpretations. By integrating user-centric features, the model seeks to generate captions that resonate more closely with individual preferences and contextual understanding. This user-centric approach ensures that the image captions not only reflect the visual content accurately but also align with the diverse perspectives and expectations of users. This enhancement is crucial for delivering a more personalized and meaningful image captioning experience. Moreover we also added an audio library which can read out the caption. This specific feature helps the blind to listen the Context in the image.

VI. SYSTEM ARCHITECTURE

We utilize a CNN + LSTM to take an image as input and output a caption. An "encoder" LSTM maps the source sentence (which is of variable length) and transforms it into a fixed-length vector representation, which in turn is used as the initial hidden state of a "decoder" LSTM which ultimately generates the final meaningful sentence as a prediction. RNN's have become very powerful. Especially for sequential data modelling. It encode images to a high level representation then decodes this using a language generation model.

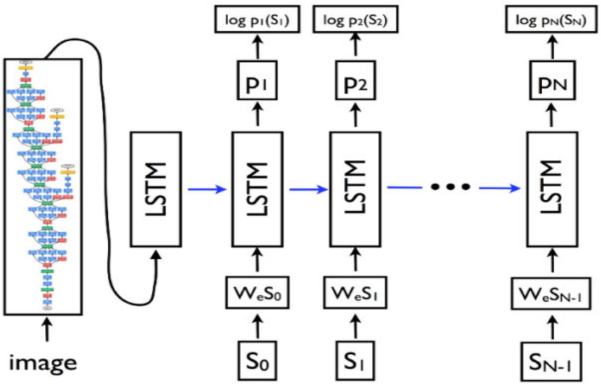


Fig 4: System Architecture of Algothims

VII. LITERATURE

In related work the relevant information is enhanced on prior study on image caption generation and attention. Several approaches for getting image descriptions have recently been presented. Automatic image captioning generation has emerged as a promising research area in recent years, because to advances in deep neural network models for Computer Vision (CV) and Natural Language Processing (NLP). In general, there are three types of image captioning modeling techniques: neural-based approaches, attention-based strategies, and RL-based methods framework. A popular captioning technique involves integrating CNN and LSTM, with CNN extracting image features and LSTM framework generating the language model. An attention mechanism can be used to improve the contextual aspect of natural language sequences. The use of attention to describe image content is consistent with human understanding.

A. Authors and Affiliations

I. SiZhen Li, School of Electronic and Information Engineering Beijing Jiao Tong University Beijing China.
II. Linlin Huang School o f Electronic and Information Engineering Beijing Jiao Tong University Beijing China.

B. Identify the Headings

In this work, the caption generation method employs the neural framework proposed where instead of translating text from one language to another, an image is translated into a caption or sentence that describes it. Beam Search is used which selects the word sequence which has the highest cumulative score of all words in its sequence as the caption. In our work the beam search algorithm is implemented with various beam width values to get much efficient captions. Visual attention has been shown to be an efficient approach for image captioning generation. When developing the target language, these attention-based captioning models may learn where to focus in the image. They may learn the distribution of spatial attention during the last convolutional layer of the CNN, or they learn the distribution of semantic attention from visual characteristics learned from social media images. Whereas these methodologies demonstrate the efficiency of the attention mechanism, they do not investigate the contextual information in the encoding sequence. Our attention layer is distinct in that it is structured in a sequential order, with each hidden state of an encoding stage contributing to the formation of decoding words.

C. Figures and Tables

a) Positioning Figures and Tables: Here we have used flickr 8k dataset.

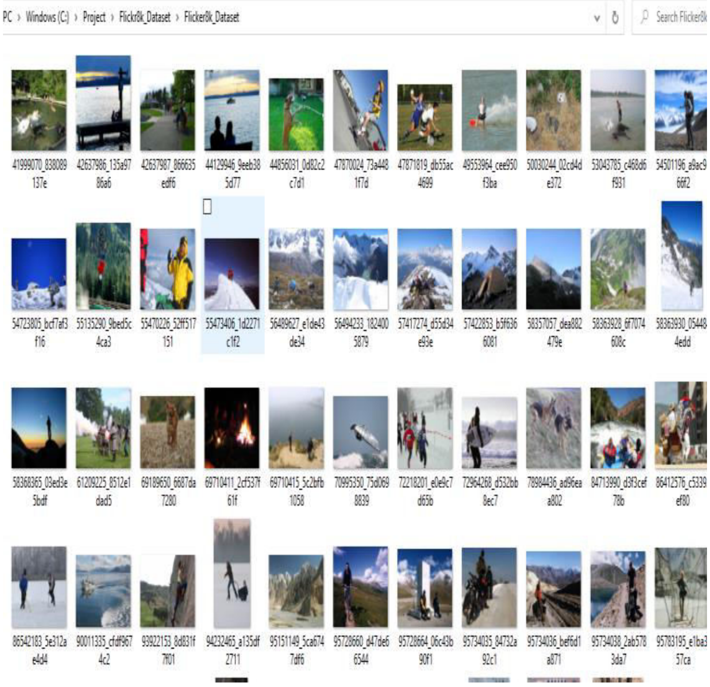
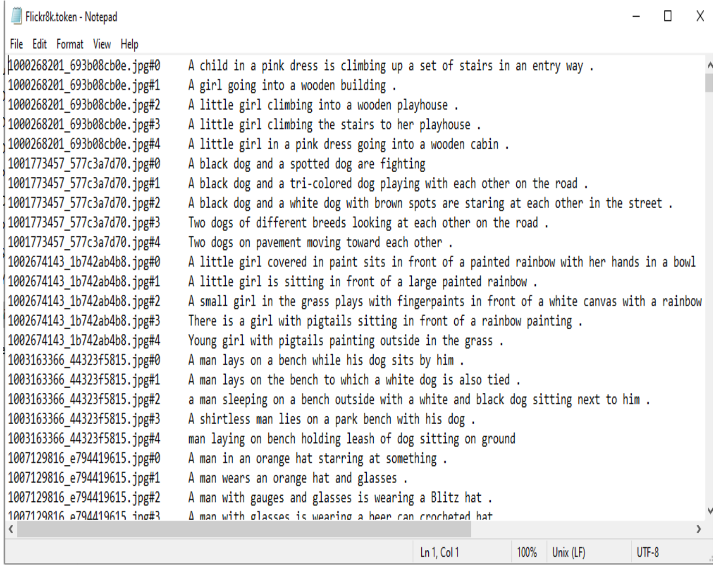


Fig 5: Flickr Dataset

Figure Labels:



```
1000268201_693b08cb0e.jpg#0	A child in a pink dress is climbing up a set of stairs in an entry way .
1000268201_693b08cb0e.jpg#1	A girl going into a wooden building .
1000268201_693b08cb0e.jpg#2	A little girl climbing into a wooden playhouse .
1000268201_693b08cb0e.jpg#3	A little girl climbing the stairs to her playhouse .
1000268201_693b08cb0e.jpg#4	A little girl in a pink dress going into a wooden cabin .
1001773457_577c3a7d70.jpg#0	A black dog and a spotted dog are fighting
1001773457_577c3a7d70.jpg#1	A black dog and a tri-colored dog playing with each other on the road .
1001773457_577c3a7d70.jpg#2	A black dog and a white dog with brown spots are staring at each other in the street .
1001773457_577c3a7d70.jpg#3	Two dogs of different breeds looking at each other on the road .
1001773457_577c3a7d70.jpg#4	Two dogs on pavement moving toward each other .
1002674143_1b742ab4b8.jpg#0	A little girl covered in paint sits in front of a painted rainbow with her hands in a bowl
1002674143_1b742ab4b8.jpg#1	A little girl is sitting in front of a large painted rainbow .
1002674143_1b742ab4b8.jpg#2	A small girl in the grass plays with fingerpaints in front of a white canvas with a rainbow
1002674143_1b742ab4b8.jpg#3	There is a girl with pigtails sitting in front of a rainbow painting .
1002674143_1b742ab4b8.jpg#4	Young girl with pigtails painting outside in the grass .
1003163366_44323f5815.jpg#0	A man lays on a bench while his dog sits by him .
1003163366_44323f5815.jpg#1	A man lays on the bench to which a white dog is also tied .
1003163366_44323f5815.jpg#2	a man sleeping on a bench outside with a white and black dog sitting next to him .
1003163366_44323f5815.jpg#3	A shirtless man lies on a park bench with his dog .
1003163366_44323f5815.jpg#4	man laying on bench holding leash of dog sitting on ground
1007129816_e794419615.jpg#0	A man in an orange hat starring at something .
1007129816_e794419615.jpg#1	A man wears an orange hat and glasses .
1007129816_e794419615.jpg#2	A man with gauges and glasses is wearing a Blitz hat .
1007129816_e794419615.jpg#3	A man with glasses is wearing a beer can crocheted hat
```

Fig 6: Descriptions of images in the dataset.

VIII. CONCLUSION

In conclusion, our proposed deep learning and AI-based model represents a significant advancement in the field of image captioning. By prioritizing user preferences and interpretations, we bridge the gap between visual content and individual understanding. Through the integration of user-centric features, our model generates captions that resonate closely with diverse perspectives, ensuring a more personalized and meaningful experience for users. Moreover, the addition of an audio library further enhances accessibility, allowing the visually impaired to engage with image context through auditory means. This holistic approach not only improves caption accuracy but also fosters inclusivity, catering to a wider audience. As technology continues to evolve, it is imperative to consider the diverse needs and expectations of users. Our system embodies this ethos, striving to deliver a more enriching and inclusive image captioning experience for all.

ACKNOWLEDGMENT

This work has been supported by the Codegnan Destination solutions.

REFERENCES

[1] SiZhen Li, Linlin Huang, Context-based Image Caption using Deep Learning, IEEE Conference on Intellignet Computing and Signal Processing(ICSP 2021)

[2] Kulkarni G, Premraj V, Dhar S, et al. Baby talk: Understanding and generating simple image descriptions[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2011:1601-1608.

[3] Aneja J, Deshpande A, Schwing A. Convolutional Image Captioning[J]. 2017.

[4] Lu J, Xiong C, Parikh D, et al. Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning[J]. 2016:3242- 3250.

[5] Rennie S J, Marcheret E, Mroueh Y, et al. Self-Critical Sequence Training for Image Captioning[J]. 2016.

[6] Jie Hu , LiShen,Samuel Albanie,GangSun,Enhua Wu:Squeeze-andExcitation Networks.journal version of the CVPR 2018 paper,accepted by TPAMI.cs.CV.arXiv:1709.01507

[7] KARPATHY A, LI F-F.Deep visual-semantic alignments for gen-erating image descriptions[C]//Proceedings of the 2015 Interna-tional Conference on Computer Vision and Pattern Recognition. Washington,DC:IEEE Computer Society,2015:3128-3137.

[8] Xu K, Ba J, Kiros R, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention[J]. Computer Science, 2015:2048- 2057..

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

Scan to save the contact details

www.ijircce.com