International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Automatic Image Caption Generation with Deep Learning

Mr. Vudumula Govinda Rao[1], G. Sriya Sahithi[2], G. Shyamala Deepika [3], B. Parameswara Rao [4] , B. Ravi Kanth[5]

[1] Assistant Professor, Department of Computer Science
[2-5] B.Tech Student, Department of Computer Science
[1-5] Raghu Engineering College, Visakhapatnam

---***---

Abstract - — Artificial intelligence research has long focused on the task of automatically defining what is included in a picture or image. This paper presents the CNN and RNN-LSTM models used in the Automatic Caption Generator implementation. It takes into account recent developments in image processing and performing research on translation through automation. The dataset utilized was Flickr8k. We have employed BLEU scores to assess the efficiency of the system that has been described. Based on the scores, one could classify the resulting captions as excellent or awful. The primary uses of this paradigm are virtual assistants, picture indexing, social media, assistive technology for the blind, altering app suggestions, and many more areas.

***Key Words*:** Automatic Caption Generator, CNN, RNN-LSTM, Computer vision, Machine translation, Flickr8k dataset, BLEU scores, Performance evaluation, Virtual assistants, Picture indexing, Social media, Assistive technology for the blind, App suggestions.

1. INTRODUCTION

The task of creating descriptive text for photographs is known as image captioning in artificial intelligence. It interprets the information of a picture and generates a coherent, accurate caption by utilizing computer vision and natural language processing. This method comprises identifying objects and situations in an image, comprehending how they relate to one another, and coming up with a sentence that captures these features. Often, large datasets of photos with handwritten labels are used to train image captioning algorithms. Computers can acquire human-like visual perception skills. Images can be captioned with computer-generated text, such as in this example: "A fluffy white cat is sitting on a windowsill." This approach effectively teaches computers to understand visuals and speak in human-like language.

The encoder-decoder architecture upon which Image Caption Generator models are based allows them to produce appropriate and legitimate captions using input vectors. This paradigm combines image recognition and NLP. The task is to recognize and comprehend the scenario that the picture depicts, then read the explanation in a natural language like English that comes next.

Fig 1. A fluffy white cat is sitting on a windowsill

Two fundamental models serve as the foundation for our approach: CNN and RNN-LSTM. The final application employs RNN-LSTM as a decoder and CNN as an encoding algorithm to retrieve characteristics from images in order to arrange and produce captions. Among the main uses for the program are self-driving cars, which can explain the environment in which they are operating and help the blind by translating scenes into subtitles, audio, CCTV cameras—which have the ability to sound warnings in the event that any criminal activity is discovered—while describing the scene, revising suggestions, social network updates, and much more. Computers can explain images in a manner similar to that of humans thanks to image captioning. This enhances computers' capacity for visual comprehension and enables them to communicate using descriptive language.

2. LITERATURE SURVEY

“S. Ren, K. He, R. Girshick, and J. Sun, ‘Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks’: [1]”

In order to attain both high detection accuracy and computational economy in object detection, this study introduces Faster R-CNN. Employing RPNs, Faster R-CNN advances on the R-CNN and Fast R-CNN approaches by generating possible areas of interest for object detection in an image. Real-time object detection performance is achieved by the suggested approach, which has had a significant influence on deep learning and computer vision research.

"K, He, G. Gkioxari, P. Dollar, and R. Girshick 'Mask R-CNN': [2] "

This paper introduces Mask R-CNN, a modification of Faster R-CNN, to predict segments for each region of interest (RoI). Mask R-CNN achieves state-of-the-art performance in object detection and grouping of instances. Computer vision has advanced significantly as a result of the effort, especially in areas like object identification and picture segmentation.

"Ben-David and Shalev-Shwartz's 'Understanding Machine Learning: From Theory to Algorithms': In [3] "

Through an examination of several learning paradigms and techniques, the theoretical underpinnings of machine learning are comprehensively explored in this book. For academics and business professionals, it provides an organized method for comprehending the fundamental ideas of machine learning, making it an invaluable tool.

" 'Large-scale learning of common visual patterns' is a publication by Wu, Donahue, and Fei-Fei, J : [4]"

This study discusses methods for large-scale visual pattern learning with a focus on the detection and representation of frequent patterns in visual data. This approach advances computer vision and machine learning applications by using large datasets to train models that identify and categorize visual patterns.

3. MECHANISM

Image captioning has drawn a lot of attention lately, especially from the natural language community. Recent developments within the domains of language processing, computer vision, and neural networks have made it possible to accurately describe images by representing their visually grounded meaning in the simplest way possible. CNNs are examples of contemporary techniques. To achieve the same result, relevant image and anthropocentric definition of datasets are merged using RNNs. We demonstrate the application of our alignment approach in retrieval research on datasets like Flicker.

A. Image Captioning Methods

There are several methods for creating image captions; some are no longer commonly used, but it is still vital to have a basic understanding of them. We classify the prevailing technique for generating aesthetic captions under three main categories:

(1) Captioning with Templates
(2) Captioning via Retrieval
(3) Crafting Unique Captions

The majority of deep learning-based techniques for creating image captions fall under the heading of creative caption creation. Consequently, we limit our attention to deep learning-based novel caption generation. Deep machine learning and visual space are the primary methods used in this novel caption generation-based picture caption system. Using learning strategies is another approach to categorize deep learning-based picture captioning:

(1) Exploration of visual spatial relationships
(2) Integration of multiple modes of information
(3) Utilization of guidance through learning
(4) Discovery without explicit supervision
(5) Emphasis on detailed captioning
(6) Analysis of the complete scene context
(7) Application of specific architectural frameworks
(8) Designing architectures with a focus on composition
(9) Incorporation of LSTM structures for language modelling
(10) Implementation of mechanisms for selective attention
(11) Foundation on underlying semantic principles
(12) Specialization in describing unfamiliar objects

B. Retrieval-based Approaches

Captions can be produced in multimodal and visual contexts. Captions for retrieval-based systems are retrieved from collections of previously published captions. Retrieval-based approaches produce syntactically correct generic captions, but they are not able to produce semantically valid image-specific captions. From the training data set, they initially retrieve hyperlinked photos and descriptions. We designate these captions as "prospective descriptions."

C. Novel Caption Generation

Creative captions may be made with both aesthetic and multifaceted contexts. The fundamental approach used in this category analyzes the image visually before utilizing a linguistic mode to create titles for the images. Compared to earlier techniques, these algorithms can produce semantically superior captions for every image. Deep learning algorithms are the foundation of most novel approaches to caption creation. As such, our primary focus is on unique picture captioning techniques that rely on deep learning.

D. Techniques for neural network-based picture captioning

1) Multifunctional versus Pictorial Realm Captioning systems powered by learning algorithms may be used to produce captions for pictures in both multimodal and visual spaces. In the methods based on visual space, the language decoder receives the picture characteristics and their corresponding captions in separate packets. In contrast, the visuals and accompanying descriptive text in a multimodal area lead to the discovery of a communal multimodal area.

The language decoder is subsequently granted access to this multimodal representation.

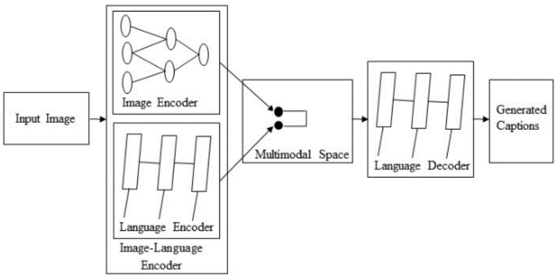


Fig 2. A multimodal space-based image captioning block diagram

2) Multimodal Space

A conventional multifunctional spatial-based approach consists of four components: a language encoder, a visual element, a multimodal space element, and a language decoder. A general design of multifaceted spatial photo labeling methods is shown in Fig. 2. The visual section uses an extensive as an extraction tool to extract the image characteristics. Using character extraction, the language's encoding portion develops an extensive characteristic integration for each word. The repetitive layers receive the semantic temporal context after that. In the multidimensional segment, the text characteristics and the pictorial characteristics are mapped onto a single space. Subsequently, the language decoder receives the created map and utilizes it to decode the map and predict captions. The following stages are taken by the techniques in this category:

(1) Text and images are simultaneously learned in a multimodal environment through the use of combining computational models of speech with advanced cognitive networks.

(2) Descriptions are generated by the programming acquisition mechanism using the information from Step 1.

3) Comparison of different neural network approaches with supervision

Training information and intended results, or labels, are mixed in supervised learning. Conversely, unsupervised learning makes use of unlabeled data. Unsupervised learning methods are one type of Generative Adversarial Networks (GANs). An agent's objective in a specific type of machine learning approach known as reinforcement learning is to use incentive signals and exploration to find data and/or labels. Numerous techniques for captioning photos make use of GAN-based algorithms and reinforcement learning.

4) Structural layout versus Representation-transformation architecture

Certain methods only generate captions using a standard encoder and decoder. However, some approaches do so by utilizing many networks.

i) Image captioning using encoder-decoder architecture-based method

The operation of neural network-based image captioning systems is straightforward from beginning to end. These methods bear some resemblance to neural machine translation as they both use the encoder-decoder architecture. In this network, an LSTM receives global visual data from CNN's hidden activations and uses them to produce a word sequence. The following are the general phases of a typical approach in this field:

1. A standard CNN is used to identify the components and their relationships, as well as to ascertain the sort of scenario.

2. The result of the first step is fed into a model that processes it into either words or rearranged phrases, constituting an image description.

To obtain the final caption, a deep multimodal similarity algorithm is used to re-rank these potential captions.

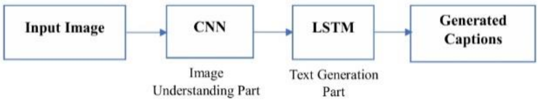


Fig 3. An illustration of a basic encoder-decoder architecture used for picture captioning

ii) Image captioning based on compositional architecture

Techniques containing distinct, useful construction components that are influenced by modular design: To extract significant concepts from the picture, a CNN is first employed. Next, a list of possible descriptions is generated by using a syntax analysis.

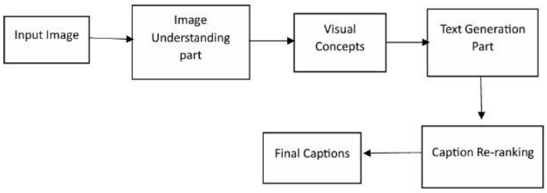


Fig 4. A compositional network-based captioning system's block diagram

A common method in this category adheres to the following steps:

(1) To extract picture characteristics, CNN is utilized.

(2) Perceptual notions (such as characteristics) originate from visual features.

(3) Using the information from Steps 1 and 2, a language model produces multiple labels.

(4) To find high-quality picture descriptions, the resulting captions are reranked using a deep multifaceted similarity model.

Think of image captioning as a collaboration between two essential components of the computer's brain:

a) CNNs (The Eye):

Convolutional Neural Networks (CNNs) retrieve information through pictures by processing input in a two-dimensional matrix format. The output, pooling, fully-connected, convolutional, SoftMax, and input layers are among its many layers. A 3D matrix containing the picture data is supplied to the input layer. The feature extractor, often referred to as the convolutional layer, performs convolutions and computes dot products. In this layer, ReLU lowers negative values to zero. Pooling layers lower the picture volume after convolution. Fully Connected layers establish connections between neurons throughout layers to incorporate weights and biases. SoftMax allows for multi-classification by assigning a probability to objects using a formula. The output layer encodes the final findings before sending them to an LSTM model.

b) RNNs (The Mouth):

The RNN-LSTM system uses CNN output to generate captions with descriptions. Sequential data analysis is done using recurrent neural networks, or RNNs. RNNs are assisted in forecasting sequences by long short-term memory (LSTM), which stores and uses prior knowledge. LSTM is very effective at understanding and anticipating sequences because it uses input processing techniques like forget gates and remembers previous steps. This setup places the RNN as the computer's "mouth," fusing language skills with CNN-generated embeddings to provide complex image descriptions.

c) VGG16:

VGG16 is a popular tool since it is good at extracting both basic and complex visual attributes. Using 3x3 filter layers of concision with a stride of 1 and constantly using the identical linings and max pool layer of 2x2 filter stroke 2 are the key features of VGG16. This allows us to infer details about the cat's silky fur, pointed ears, and even the window sill it is resting on.

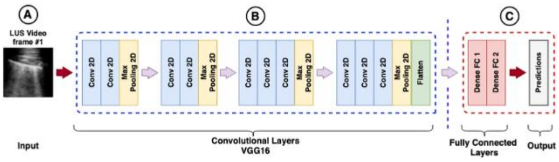


Fig 5. Example of feature extraction using a VGG16

The spatial filtering and max pool layers are positioned in the same positions in the design. Two FC (completely connected layers) and a softmax are included in the final step. But since we just need the characteristics of the image, we will eliminate the final layer. The sixteen layers in VGG16, each with a distinct weight, are represented by the number 16. This network is very large, with an estimated 138 million parameters. The system was trained using the 14 million images from 1000 categories that make up the ImageNet dataset.

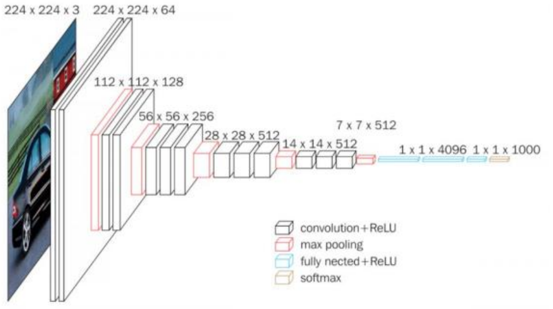


Fig 6. Architecture of VGG16

d) LSTM:

A sequential model with LSTM units is particularly good at sequence prediction tasks. The word that will appear next can be predicted based on the preceding paragraph. In terms of getting over the short-term memory constraints of standard RNNs, it has done better than them. The LSTM processes useful information and discards irrelevant information using a forget gate. Compared to traditional RNNs, LSTMs aim to prevent vanishing gradients and retain information for longer periods of time. LSTMs can learn by backpropagation through time and layers while maintaining a constant error rate.

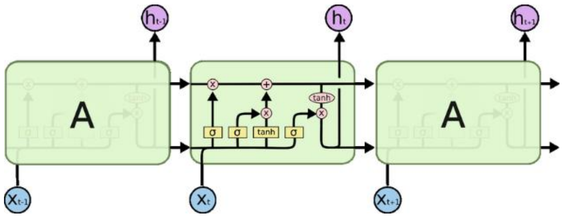


Fig 7 .The repeating module in an LSTM containing four interacting layers

e) LSTM Model Training:

We are prepared to give our captions life now that we have our image qualities. Our LSTM model's training is covered in this section. The true magic happens here: we train our model to describe our photos in a descriptive manner. In the CNN-LSTM architecture, CNN layers take features out of the input data and mix them with LSTMs in order to anticipate sequences. The term "CNN LSTM" in this course refers to LSTMs that have a CNN front end; they were formerly referred to as the LRCN model, or Long-term Recurrent Convolutional Network. This architecture tackles the problem of writing explanations for images. A CNN that was trained for categorization of images was converted to a feature extractor for captioning.

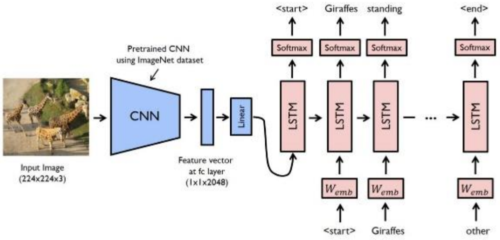


Fig 5. Example of feature extraction using a VGG16

4. IMPLEMENTATION METHODOLOGY

This project requires a dataset containing both images and captions. The image captioning model should be trainable using the provided dataset.

A. Flickr 8k Dataset:

The Flickr 8k dataset is a freely accessible standard for visual-to-text conversion. This compilation consists of 8091 photos, each with five captions. These images were grabbed from multiple Flickr groups. A detailed explanation of all the objects and activities in the picture is given in each caption. The collection is more general in nature because it does not include images of famous people or places, although covering a wide range of events and scenarios. The following qualities make the dataset ideal for this study:

- The image captioning model's ability to handle various image types is improved by using a diverse range of training photos.
- The generality of the model is increased and overfitting is decreased when many captions are mapped to a single image.

The flickr8k dataset was selected because:
(1) It is small in size. Consequently, the model may be trained on low-end quickly.
(2) The data has been labelled correctly. Every picture has five captions.
(3) The dataset is available for free download.

B. Image Data Preparation:

The visual data should be transformed into pertinent features to develop a machine learning system. Any image that wants to be trained into a machine learning algorithm must first undergo the acquisition of characteristics. We partitioned the dataset into three parts: train_image, validation_image, and test_image, which comprise 4855, 1618, and 1618 photos, respectively. The features are retrieved using the VGG-16 model and CNN. In 2015, this model was victorious in the ImageNet Large Scale Visual Recognition Task, categorizing photos into 1000 categories. This method works well for picture captioning, where image identification is necessary.

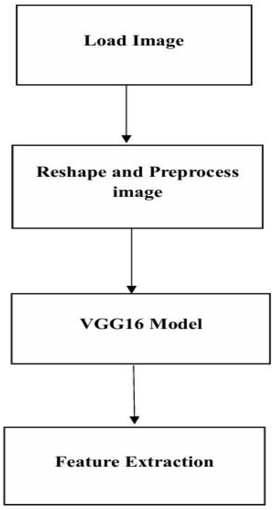


Fig 9. VGG-based feature extraction in images

There are 16 weight layers in the VGG-16 network; additional layers lead to a higher level of feature extraction from images. With a max pooling layer to minimize picture volume size and 3*3 convolutional layers, the VGG-16 network has a straightforward architecture. The last layer of the image, which makes categorization predictions, is eliminated. Next, a feature representing the internal representation right before classification is returned. This model captures characteristics from an input image with dimensions of 224*224*3 and produces a 4096-element vector in one dimension.

C. Caption Data Preparation:

Each shot in the Flickr 8k collection has five captions. We just used one caption per shot. During data preparation, image ids serve as keys, while a dictionary's values are retained for captions.

i) Data Cleaning
For data mining or neural networks to use unprocessed content, it must first be transformed into a legible format. Before using the text for the project, it needs to be cleaned as follows:

- Remove punctuation marks.
- Get rid of the numbers.
- Remove single-length words.
- Lowercase to uppercase character conversion. Eliminating stop words from text data might hinder the ability to construct a grammatically correct caption for this project.

5. PROPOSED ARCHITECTURE

A. Phases of the application:

i) Features Extraction
The features from the photos are being retrieved. It generates vector features, often known as embeddings. The CNN model extracts features from original images and compresses them to a smaller, RNN suitable feature vector. Thus, it is also known as Encoder.

ii) Tokenization
The application's next phase is RNN, which decodes feature vectors fed from CNN. Here, the word sequence is anticipated, but the captions are produced.

iii) Prediction
Following tokenization, the final step is prediction. After decoding the vectors, the predict_caption() function generates the final output.

6. DATA FLOW DIAGRAM

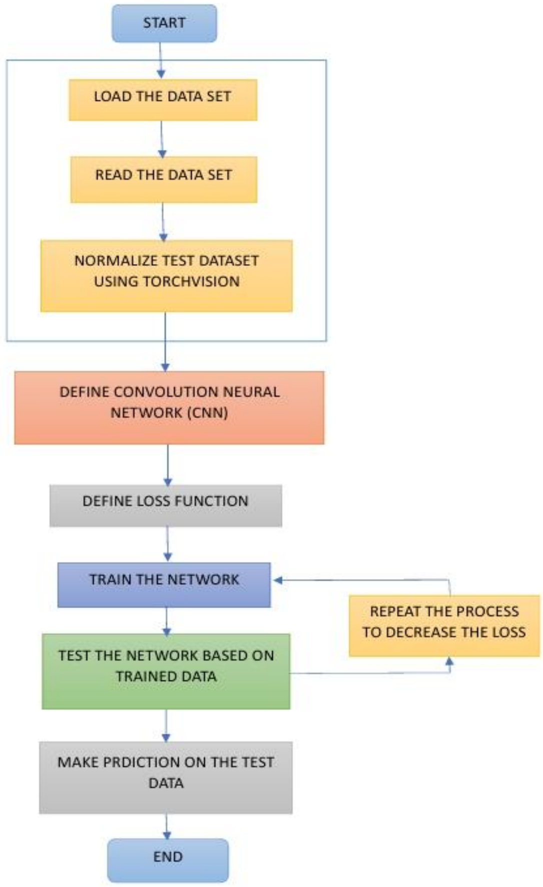
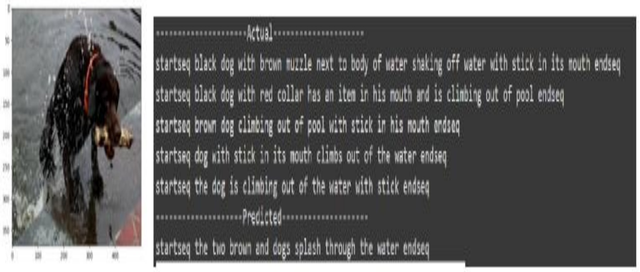

START
LOAD THE DATA SET
READ THE DATA SET
NORMALIZE TEST DATASET USING TORCHVISION
DEFINE CONVOLUTION NEURAL NETWORK (CNN)
DEFINE LOSS FUNCTION
TRAIN THE NETWORK
REPEAT THE PROCESS TO DECREASE THE LOSS
TEST THE NETWORK BASED ON TRAINED DATA
MAKE PRDICTION ON THE TEST DATA
END

Fig 10. Diagram of the data flow

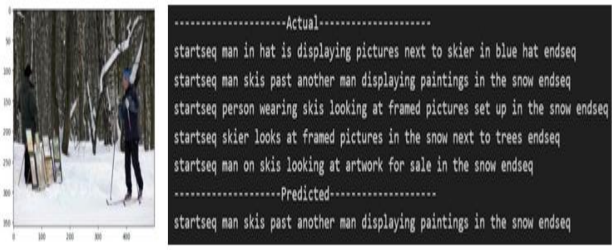
7. RESULT AND ANALYSIS

Generated Captions:



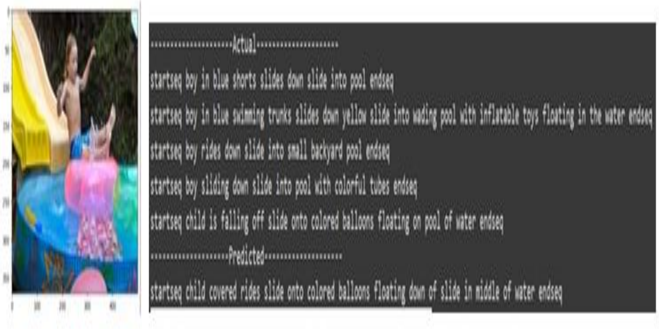
```
--------------------Actual--------------------
startseq black dog with brown muzzle next to body of water shaking off water with stick in its mouth endseq
startseq black dog with red collar has an item in his mouth and is climbing out of pool endseq
startseq brown dog climbing out of pool with stick in his mouth endseq
startseq dog with stick in its mouth climbs out of the water endseq
startseq the dog is climbing out of the water with stick endseq
--------------------Predicted--------------------
startseq the two brown and dogs splash through the water endseq
```

Fig 11. Dog with a stick in its mouth climbs out of the water



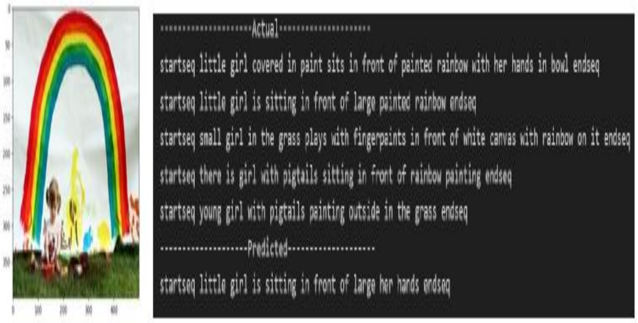
```
--------------------Actual--------------------
startseq man in hat is displaying pictures next to skier in blue hat endseq
startseq man skis past another man displaying paintings in the snow endseq
startseq person wearing skis looking at framed pictures set up in the snow endseq
startseq skier looks at framed pictures in the snow next to trees endseq
startseq man on skis looking at artwork for sale in the snow endseq
--------------------Predicted--------------------
startseq man skis past another man displaying paintings in the snow endseq
```

Fig 12. Man on skis looking at artwork for sale in the snow



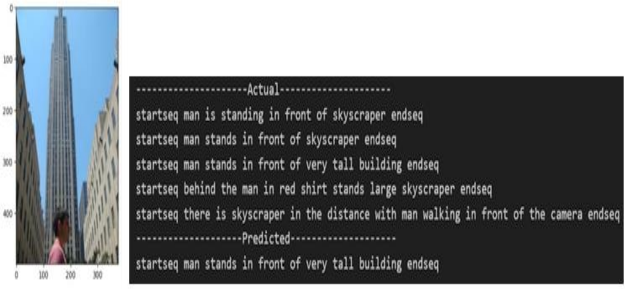
```
--------------------Actual--------------------
startseq boy in blue shorts slides down slide into pool endseq
startseq boy in blue swimming trunks slides down yellow slide into wading pool with inflatable toys floating in the water endseq
startseq boy rides down slide into small backyard pool endseq
startseq boy sliding down slide into pool with colorful tubes endseq
startseq child is falling off slide onto colored balloons floating on pool of water endseq
--------------------Predicted--------------------
startseq child covered rides slide onto colored balloons floating down of slide in middle of water endseq
```

Fig 13. Boy sliding into the pool with colourful tubes



```
--------------------Actual--------------------
startseq little girl covered in paint sits in front of painted rainbow with her hands in bowl endseq
startseq little girl is sitting in front of large painted rainbow endseq
startseq small girl in the grass plays with fingerpaints in front of white canvas with rainbow on it endseq
startseq there is girl with pigtails sitting in front of rainbow painting endseq
startseq young girl with pigtails painting outside in the grass endseq
--------------------Predicted--------------------
startseq little girl is sitting in front of large her hands endseq
```

Fig 14. Little girl is sitting in front of large painted rainbow



```
--------------------Actual--------------------
startseq man is standing in front of skyscraper endseq
startseq man stands in front of skyscraper endseq
startseq man stands in front of very tall building endseq
startseq behind the man in red shirt stands large skyscraper endseq
startseq there is skyscraper in the distance with man walking in front of the camera endseq
--------------------Predicted--------------------
startseq man stands in front of very tall building endseq
```

Fig 15. Man stands in front of skyscraper

. The predicted captions demonstrate a good understanding of the scene, capturing the key elements. The predicted captions are grammatically correct and effectively convey the intended message. The predicted captions are consistent with the visual content of the image and provide a coherent description that matches the actual scene depictedIn this project we have taken 2 categorical features like spending score, income. But we can even use age, race, nationality, profession, work experience, locality and perform segmentation based on 2 or more feature

6. CONCLUSION

The capacity to produce captions for photos automatically is still under development; current work is focused on enhancing phrase generation and image extraction. We used a smaller dataset (Flickr8k) to effectively complete the project proposal because to restricted computational resources. Our model produces semantically correct output in user-selected languages by drawing on a lexicon of words from existing captions in the training dataset. This method provides insights as the interpreter and a CNN as the source of information.The goal of the model is to make it more likely that a given image will result in a statement. Using Flickr8k training images, we tested the model and got good results. We assess the built-in model's accuracy using a metric called Bilingual Evaluation Understudy. A similar model's accuracy can be increased with a larger dataset that contains more images. It would also be helpful to comprehend how unsupervised data from isolated settings and images could enhance techniques for explaining visuals.

REFERENCES

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3156–3164.

[2] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3128–3137.

[3] S. Kiros, R. Salakhutdinov, and R. Zemel, "A Simple Framework for Automatic Image Captioning," in arXiv preprint, 2014, arXiv:1412.4464. [Online]. Available: http://arxiv.org/abs/1412.4464

[4] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in Proceedings of the International Conference on Machine Learning (ICML), 2015, pp. 2048–2057.

[5] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A Dataset for Movie Description," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3202–3212.

[6] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear Attention Networks for Image Captioning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10971–10980.