Image Caption Generator using Deep Learning Algorithm – VGG16 and LSTM

B. Bhaskar Rao[1], Kalki Chaitanya Lade[2], Avinash Pasumarthy[3], Parasuram Swamy Katreddy[4], Garapati Chaitanya Nagendra Kumar [5]

[1] Professor, Dept of CSE, GITAM (Deemed to be University), Visakhapatnam, Andhra Pradesh, sIndia
[2,3,4,5] Student, GITAM (Deemed to be University), Visakhapatnam, Andhra Pradesh, India.

---***---

Abstract - *Deep learning Image Caption Generator (ICG) with VGG16 architecture. The aim of this study is to create a model that can produce captions for input photographs that are both descriptive and pertinent to the context. Our strategy leverages the rich feature representations that VGG16 has acquired from its pre-trained convolutional layers in order to close the semantic gap that exists between textual descriptions and visual material. The encoder in the suggested model extracts features from images using a convolutional neural network (CNN) based on VGG16, while the decoder uses recurrent neural networks (RNNs) to generate captions. The encoder efficiently encodes visual information by utilizing VGG16 to extract high-level features from images. This allows the decoder to provide captions that accurately match the content and context of the input images. The attention techniques incorporated by the decoder enable the model to concentrate on pertinent image regions during the generation of each caption word, hence augmenting the diversity and informativeness of generated captions. Furthermore, methods like beam search and vocabulary augmentation are used to encourage coherence and diversity in generated captions. Test results on reference datasets like Flickr8k show how well the suggested method works to produce insightful captions for a variety of photos. Both qualitative and quantitative assessments demonstrate how the model can generate linguistically and contextually relevant captions, highlighting its potential for a range of uses such as picture understanding, retrieval, and accessibility for those with visual impairments.*

Key Words: Artificial Intelligence, Convolutional Neural Network, Caption Generator, VGG16, LSTM

1.INTRODUCTION

The Image Caption Generator creates insightful captions for them. It analyzes visual content and produces logical written descriptions by using sophisticated neural network topologies. The model uses a combination of recurrent neural networks (RNNs) and convolutional neural networks (CNNs) to produce sequential captions based on extracted information from images. A sizable dataset of matched photos and captions is used in the training of the Image Caption Generator. The model picks up the ability to link particular visual cues with relevant textual descriptions throughout the training phase. It can now generalize and create captions for brand-new, undiscovered photos as a result. After being trained, the Image Caption Generator may provide a relevant and appropriate caption for an input image. The model's produced captions seek to highlight the important elements and details in the picture, offering a useful aid for automatic description and comprehension of images. The student uses a variety of datasets with matched photos and captions to do this. The deep learning model is trained using this dataset as its basis. The algorithm picks up complex relationships and patterns between written descriptions and visual elements during training. Effective feature extraction from images is facilitated by the use of pretrained CNNs. RNNs, on the other hand, aid in the sequential production of captions, guaranteeing a coherent and suitably contextual story. The project showcases the student's expertise in machine learning and deep learning techniques by implementing and optimizing cutting-edge neural network topologies. The resulting Image Caption Generator seems promising for a number of uses, such as helping people with visual impairments access visual content, improving search engine indexing for content, and making contributions to the larger artificial intelligence community. Because image caption generators can automatically provide descriptive captions or written summaries for photographs, they have become indispensable tools for a variety of applications. These tools are essential for improving accessibility since they offer descriptions for those who are blind or visually impaired. as well as supporting others who might find it difficult to understand visual stuff. Additionally, they make it easier for search engines and databases to index and retrieve content since they make it possible for them to recognize and classify images according to their content. By automatically creating descriptions for visual information, image caption generators on social media help users share it more quickly and efficiently while also saving time and effort. Furthermore, these generators help with the fast generation of pertinent captions for photos used in articles, blogs, or presentations in workflows involving content creation, such as journalism or blogging. By giving descriptions for the visuals used in learning materials, they aid in accessibility and comprehension in educational environments. Additionally, in research and analysis contexts, picture captioning technology can automate activities like sentiment analysis and image categorization. Image captions improve user experiences in augmented reality (AR) and virtual reality (VR) applications by offering textual descriptions or context for virtual environments. Lastly, they can be included into assistive technologies to provide those who are visually impaired with real-time explanations of their environment. All things considered, picture caption generators improve the usability, accessibility, and searchability of visual content in a variety of contexts and applications.

1.1 Convolutional Neural Network

A specific kind of deep learning model called a Convolutional Neural Network (CNN) is used to process and analyze visual input, especially photos and videos. It takes its cues from how the visual cortex of an animal is structured, which enables it to automatically learn from and extract information from unprocessed input data. For several computer vision applications, including object identification, picture segmentation, and image classification, CNNs have emerged as the cutting-edge method. Convolutional layers, pooling layers, and fully linked layers are the main parts of a CNN. In order to extract features and produce feature maps, convolutional layers apply filters to the input data. By reducing the spatial dimensions of the data, pooling layers efficiently lower the computational load and manage overfitting. The final predictions or classifications based on the learnt features are handled by fully connected layers. CNNs can identify objects and features in a picture independent of their position because of their capacity to learn local patterns and spatial hierarchies. Because of this, they are very good at jobs requiring the comprehension of intricate visual patterns. CNNs' hierarchical architecture makes it possible for them to reliably and accurately recognize objects and shapes from low-level features like edges and textures to high-level features like shapes.

1.2 VGG-16

A convolutional neural network (CNN) architecture called VGG16 was put forth by the University of Oxford's Visual Geometry Group (VGG). It is well known for being both easy to use and efficient at classifying images. This is a quick synopsis of VGG16: 1. Architecture: There are 16 layers in VGG16, comprising 3 fully connected layers and 13 convolutional layers. The convolutional layers have small 3x3 pixel receptive fields and are stacked one on top of the other. The feature maps' spatial dimensions are down sampled by the usage of max-pooling layers. 2. Filter Size: VGG16 employs 3x3 filters with a 1-pixel stride in all of the convolutional layers. This tiny filter size keeps the number of parameters reasonable while enabling a deeper network design 3. Depth: Compared to earlier CNN architectures like Alex-Net, which had just 5 convolutional layers, VGG16 has 13 convolutional layers, making it deeper. With higher depth, VGG16 can extract more intricate features from photos. 4. Fully Connected Layers: The VGG16 architecture consists of three fully connected layers with 4096 neurons each, which come after the convolutional layers. A final soft-max layer is used for classification. Learning high-level features and generating predictions based on the features collected are made easier by these completely connected layers. 5. Pre-training: For a variety of computer vision tasks, VGG16 is frequently utilized as a feature extractor or pre-trained model. VGG16 may be trained for specific tasks using smaller datasets by pre-training on large-scale picture datasets such as ImageNet. This helps VGG16 understand generic features.

2. EXISTING SYSTEM

The current systems demonstrate how picture captioning models have developed over time, starting with simple encoder-decoder architectures and moving on to more complex attention processes and transformer-based techniques. Research endeavors are ongoing to tackle problems like producing varied captions, managing uncommon ideas, and enhancing the comprehensibility of produced content. By addressing certain shortcomings in the current systems or investigating novel designs, the student project might add to this environment. The suggested Image Caption Generator system is a creative development in the field, enhancing the advantages of current models and adding new capabilities to solve certain problems. The following is a summary of the main elements and attributes of the suggested system: Enhanced Attention processes: To enhance the model's capacity to concentrate on pertinent image regions during caption production, the suggested method integrates sophisticated attention processes. In order to capture more complex interactions between visual and textual features, experiments are being conducted with variations of attention processes, such as self-attention and multi head attention. Multimodal Fusion Strategies: The suggested system investigates multimodal fusion strategies to improve the integration of textual and visual information. The model looks into methods such as cross-modal attention, early fusion, and late fusion in an attempt to develop a more sophisticated understanding of the connections between images and captions. Fine-tuning with Contrastive Learning: The suggested approach investigates pretraining on sizable datasets with a contrastive loss function, taking advantage of contrastive learning's advantages. Context-Aware Caption Generation: Using contextual data to enhance the captioning process is the main goal of the suggested system. Through the examination of contextual clues present in the photos and the generation of a word sequence, the model endeavors to generate captions that are both contextually consistent and descriptive. Managing unusual Concepts and Biases: The suggested approach includes techniques to manage unusual concepts in images and reduce biases in training data, addressing shortcomings in current systems. More inclusive and objective captions can be produced by utilizing strategies like data augmentation, domain adaptation, and bias correction. Enhancing User Feedback through Interactive Learning: The suggested system investigates interactive learning strategies, enabling user comments on captions that are generated. The model is trained using user feedback, which allows it to adjust and get better based on actual user interactions. Comprehensive Evaluation measures: The evaluation approach combines more human-centric measures, such as coherence, inventiveness, and relevance of created captions, with more conventional metrics, such as BLEU, METEOR, and CIDEr. Interpretability and Explanability: The goal of the suggested solution is to improve the output captions' interpretability by offering a glimpse into the model's decision-making process. Attention maps and visualization approaches are used to identify the areas of the image that impact particular words in the generated captions.

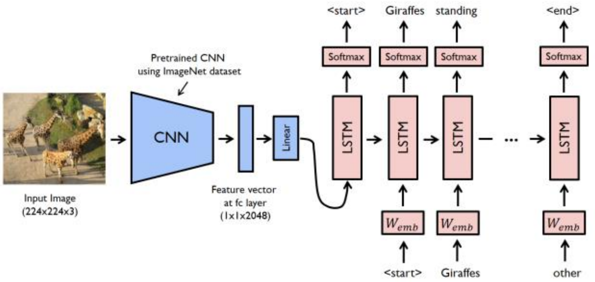


Fig -2.1: existing system of caption generator

3. PROPOSED SYSTEM

The 16-layer Visual Geometry Group network is known as VGG16. The Oxford University's Visual Geometry Group proposed the convolutional neural network (CNN) architecture. Karen Simonyan and Andrew Zisserman's 2014 publication "Very Deep Convolutional Networks for Large-Scale Image Recognition" introduced VGG16. Convolutional Layers: VGG16 has five max-pooling layers and thirteen convolutional layers, each of which is followed by a ReLU activation function. Fully Connected Layers: the final layer classifies data using a softmax activation function after each of the three fully connected layers, which are each followed by a ReLU activation function. Depth and Filter Size: The term VGG16 comes from its fixed architecture, which consists of 16 layers. Max-pooling layers have 2x2 filters with stride 2 and all convolutional layers utilize 3x3 filters. Number of Filters: The number of filters increases by two after each max-pooling layer until it reaches 512 filters, with 64 filters in the first layer. Easy to Understand design: VGG16's simple design makes it a popular choice for instructional reasons and as a starting point architecture for a variety of computer vision jobs. Efficient Feature Extraction: VGG16 is well-known for its capacity to extract rich hierarchical features from images. This capability has been shown to be useful in a variety of applications, including object identification, image classification, and feature extraction for further tasks. Transfer Learning: VGG16 is frequently used for transfer learning because of its efficiency and the availability of pre-trained weights on large-scale datasets like ImageNet. Training time and resources can be saved by fine-tuning the pre-trained weights on smaller datasets for certain tasks. Stability and Robustness: The stability and robustness of VGG16 have been demonstrated in a variety of tasks and datasets through considerable research and testing.

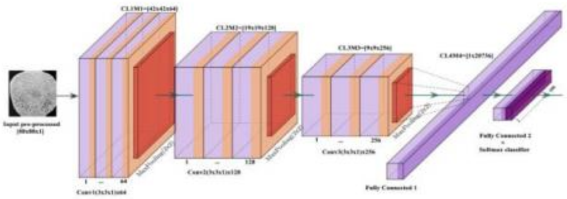


Fig -3.1: CNN

4. METHODOLOGY

An artificial neural network type that works especially well for interpreting visual data is called a convolutional neural network (CNN). In applications including object identification, picture classification, and image recognition, CNNs have demonstrated remarkable efficacy. Many layers, including convolution layers, pooling layers, and fully linked layers, make up a convolutional neural network. • The main purpose of a convolutional neural network (CNN), a deep learning architecture, is to interpret structured grid data, including pictures and videos. When it comes to computer vision applications, object identification, facial recognition, and picture categorization, CNNs have shown to be incredibly efficient. Layers for convolution: • They are made up of a collection of adjustable filters that glide over the input data to carry out convolution operations. • These filters compute dot products between the filter and small portions of the input data in order to capture various patterns, such as edges, textures, or more sophisticated characteristics. Layers of Pooling: • It divides the input feature map into rectangular, non-overlapping sections (often 2 by 2 or 3 by 3), keeping just the highest value in each sector. Fully Connected Layers: These layers enable high-level feature abstraction by connecting each neuron to every other neuron in the layers above and below.When generating final judgments or predictions, such categorizing an image, fully connected layers are frequently utilized. Flattening: Feature maps are usually flattened into a 1D vector before input is passed through fully connected layers. The data is made simpler for conventional neural network layers by this change. Batch normalization is a technique that normalizes each layer's input in order to speed up and stabilize training. It can lessen overfitting and aid in the network's increased learning efficiency. Output Layer: Depending on the architecture and goal of the network, the output of the CNN, which is normally a fully connected layer, is utilized for tasks like regression and classification.

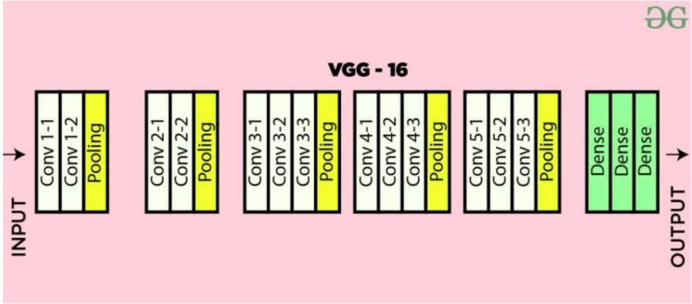
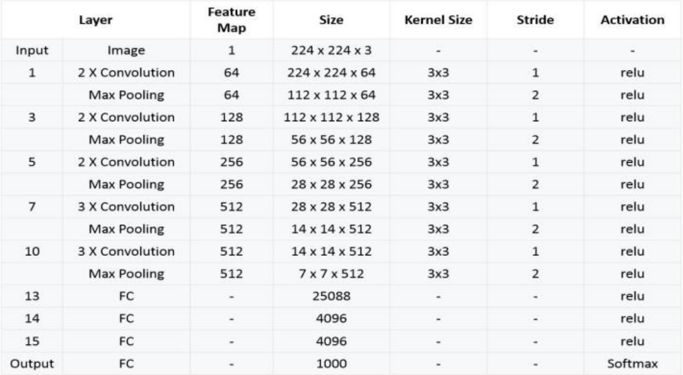


Fig-4.1: VGG-16 architecture



Layer		Feature Map	Size	Kernel Size	Stride	Activation
Input	Image	1	224 x 224 x 3	-	-	-
1	2 X Convolution	64	224 x 224 x 64	3x3	1	relu
	Max Pooling	64	112 x 112 x 64	3x3	2	relu
3	2 X Convolution	128	112 x 112 x 128	3x3	1	relu
	Max Pooling	128	56 x 56 x 128	3x3	2	relu
5	2 X Convolution	256	56 x 56 x 256	3x3	1	relu
	Max Pooling	256	28 x 28 x 256	3x3	2	relu
7	3 X Convolution	512	28 x 28 x 512	3x3	1	relu
	Max Pooling	512	14 x 14 x 512	3x3	2	relu
10	3 X Convolution	512	14 x 14 x 512	3x3	1	relu
	Max Pooling	512	7 x 7 x 512	3x3	2	relu
13	FC	-	25088	-	-	relu
14	FC	-	4096	-	-	relu
15	FC	-	4096	-	-	relu
Output	FC	-	1000	-	-	Softmax

Fig-4.2: sizes of images in different layers

5. PROPOSED MODULE AND ALGORITHM

5.1 LSTM (LONG AND SHORT TERM MEMORY)

A kind of recurrent neural network (RNN) called Long Short-Term Memory (LSTM) is intended to better recognize and understand long-term dependencies in sequential input. It uses a more intricate memory cell layout with input, forget, and output gates to solve the vanishing gradient issue. In applications involving sequential data, such time series analysis and natural language processing, LSTMs are frequently employed. Recurrent neural network (RNN) architecture known as Long Short-

Term Memory (LSTM) networks was created expressly to solve the vanishing gradient problem, which arises frequently while training conventional RNNs on data sequences. Sequential data modeling applications including time series prediction, image caption generation, and natural language processing (NLP) are good fits for LSTM networks. LSTM networks are commonly employed in conjunction with convolutional neural networks (CNNs) for the purpose of generating captions for images. The general procedure is as follows: 1. Feature Extraction with CNNs: To begin, the input image's features are extracted using a pre-trained CNN model (such as VGG, ResNet, or Inception). CNN output can be thought of as a high-level representation of the picture information. CNNs are very good at capturing spatial hierarchies of characteristics inside images. 2. Sequence Modeling using LSTMs: An LSTM network is given the features that were taken out of the CNN. The LSTM's job is to produce a logical string of words that makes up the image caption by learning the sequential dependencies between these attributes. Training Procedure: The model is fed pairs of photos together with the captions that go with them while it is being trained. After the CNN extracts the image features, the ground truth caption and the image features are sent to the LSTM. By forecasting the subsequent word in the sequence using the previously created words and the image attributes, the LSTM has the ability to create captions. 4. Caption Generation: The trained model uses the CNN to extract features from a new image at inference time. After receiving these features, the LSTM creates a caption word by word until either a certain maximum caption length is reached or a unique end-of-sequence token is generated.

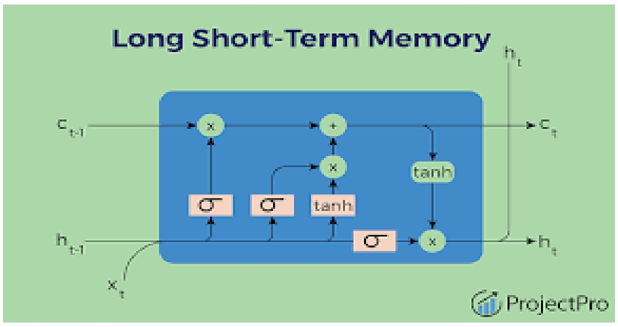
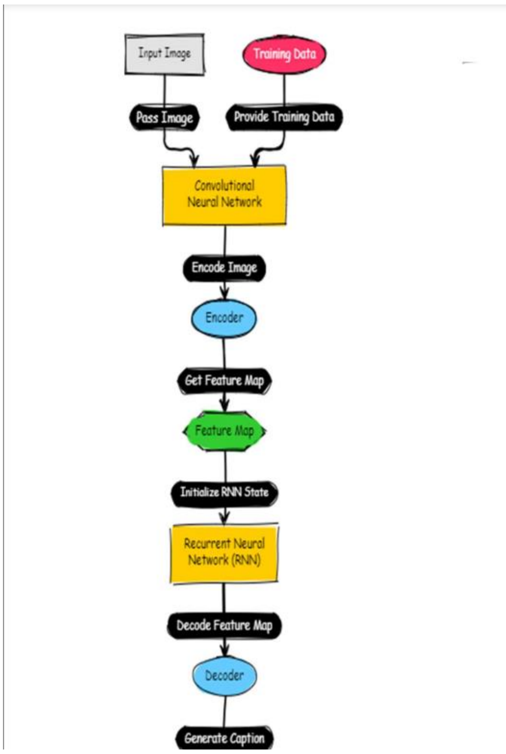


FIG-5.1: Working of LSTM

5.2 ARCHITECTURE/WORKFLOW



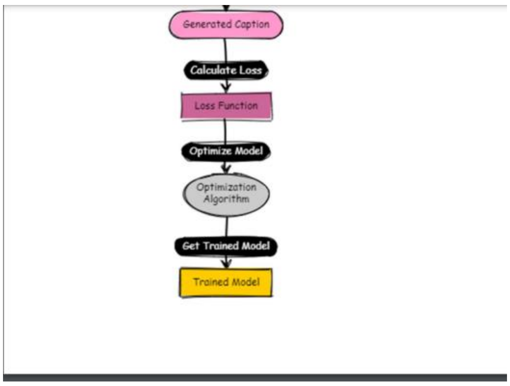


FIG-5.2: Architecture of the project

5.3 UML Diagrams

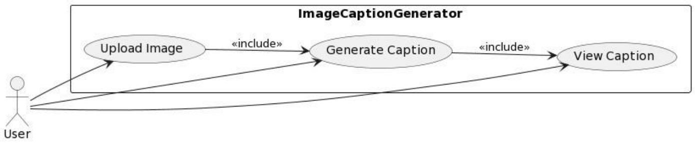


Fig5.3: Use case diagram

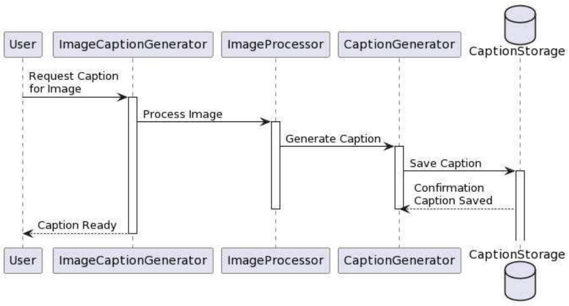


Fig-5.4: sequential diagram

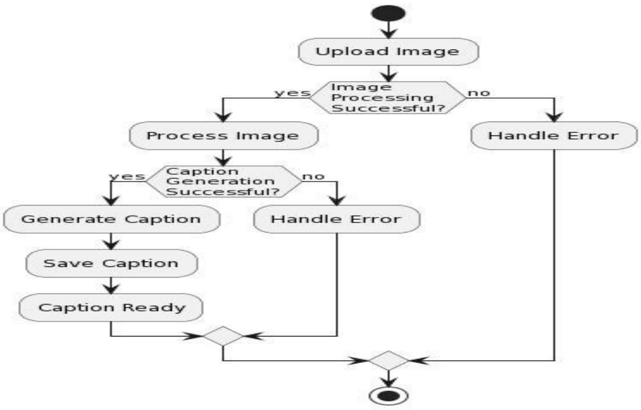
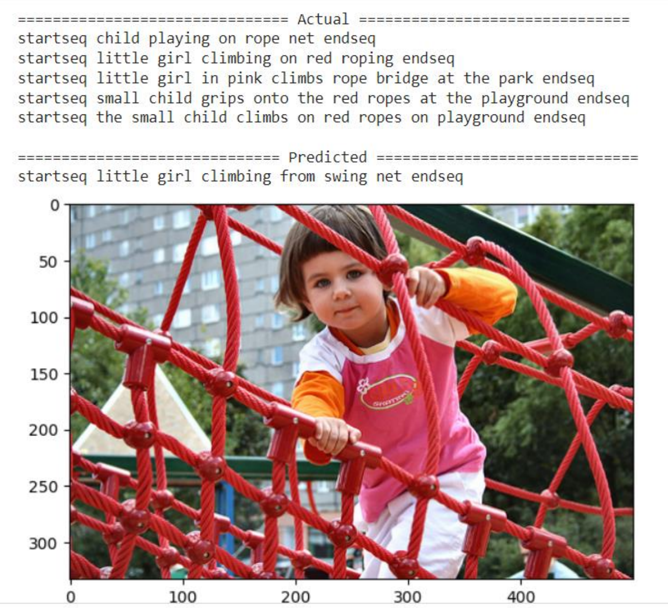


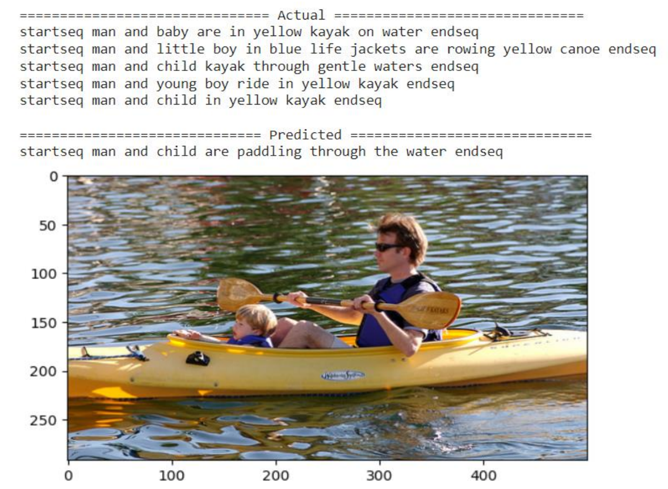
Fig-5.5: flowchart diagram

6. Results



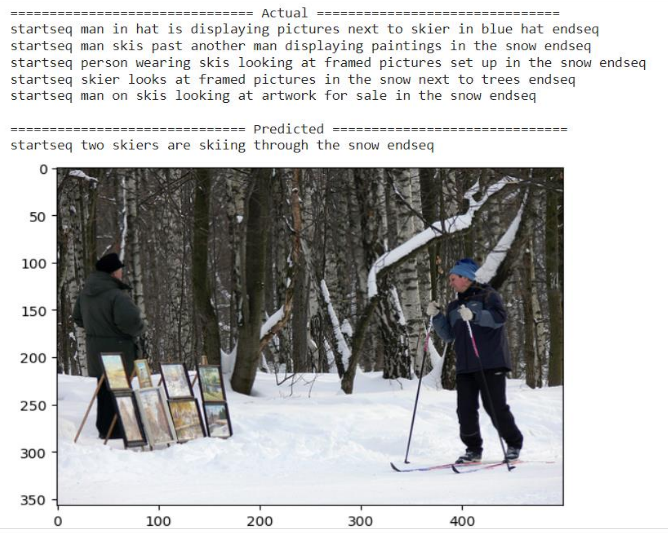
```
=============================== Actual ===============================
startseq child playing on rope net endseq
startseq little girl climbing on red roping endseq
startseq little girl in pink climbs rope bridge at the park endseq
startseq small child grips onto the red ropes at the playground endseq
startseq the small child climbs on red ropes on playground endseq

============================== Predicted ==============================
startseq little girl climbing from swing net endseq
```



```
=============================== Actual ===============================
startseq man and baby are in yellow kayak on water endseq
startseq man and little boy in blue life jackets are rowing yellow canoe endseq
startseq man and child kayak through gentle waters endseq
startseq man and young boy ride in yellow kayak endseq
startseq man and child in yellow kayak endseq

============================== Predicted ==============================
startseq man and child are paddling through the water endseq
```



```
=============================== 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 two skiers are skiing through the snow endseq
```

6. Conclusion

The Image Caption Generator (ICG) project, which uses deep learning and is based on the VGG16 architecture, has shown encouraging results in producing insightful and contextually appropriate captions for photos. The model has demonstrated the capacity to efficiently bridge the semantic gap between textual descriptions and visual material by using recurrent neural networks (RNNs) as a decoder for caption creation and VGG16 as an encoder for image feature extraction. Many approaches have been investigated and used during the project to improve the caliber and variety of generated captions. In order to increase variation in the generated captions and enhance caption coherence and fluency, attention mechanisms, beam search, and vocabulary enrichment have been used. Additionally, the model has been trained and evaluated on benchmark datasets such as Flickr8k, demonstrating its effectiveness in generating descriptive captions for diverse images.

6.1 future scope

Multimodal Understanding: Extend the model to include multimodal data. For example, combine textual and visual elements to provide captions that are richer in context. In order to improve the model's comprehension of textual context, this may entail integrating pre-trained language models like BERT or transformers. 2. minute-Grained Captioning: Learn how to create captions that effectively convey the relationships and minute details found in photographs, such as particular items, actions, and spatial configurations. This can entail adding modules for object identification or scene graph creation to the captioning process. 3. Interactive Captioning Systems: Create interactive captioning systems that let users edit or add to automatically generated captions. This way, users' involvement will help the model to continuously adjust and get better over time.

6.2 References

1. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention" by Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Richard Zemel, and Yoshua Bengio.

2. "Neural Machine Translation by Jointly Learning to Align and Translate" by Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio.

3. "Deep Visual-Semantic Alignments for Generating Image Descriptions" by Andrej Karpathy and Li Fei-Fei.

4. "Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models" by Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik