International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(18s) 2026 pp. 265-281
https://doi.org/10.70917/ijcisim-2026-4865
Article

Novel Method for Caption Based Description in Image using Machine Learning

Rachita Dubey[1], Vivek Shukla[2]

[1] Department of Computer Science and Engineering, Dr. C. V. Raman University, Bilaspur, Chhattisgarh, India.
Email: rachitadubey1991@gmail.com
[2] Department of Computer Science and Engineering, Dr. C. V. Raman University, Bilaspur, Chhattisgarh, India.
Email: rohit@cvru.ac.in

Abstract: The challenging research area of image captioning aims to produce written descriptions that faithfully capture the atmosphere and events captured in a photograph. In order to recognize objects interpret changing conditions and understand the semantic connections within the image advanced machine learning algorithms are required. This study builds an image captioning system using CNN, RNN and ResNet architectures. Here the MSCOCO benchmark dataset is used to enable precise scenario inference with CNN acting as the encoder and RNN as the decoder. By utilizing skip connections and avoiding multiple convolutional layers the system uses ResNet which efficiently utilizes its layers and reduces computation time. This approach resolves the gradient explosion problem and enhances model performance. The proposed model performs better than previous implementations in several evaluation criteria such as BLEU, METEOR, CIDEr and ROUGE. Using the Pillow library the system preprocesses images to enhance brightness and minimize size for optimal training. Additionally the models predicted accuracy is raised by utilizing the TorchVision library. With these enhancements the system can now generate high-quality captions with precise semantic understanding and contextual value.
Keywords: Image Captioning, CNN, RNN, ResNet, DNN, LSTM, MSCOCO, Flickr30k, Flickr8k

1. Introduction

Understanding an image is the process of determining its meaning and content. Usually this procedure is connected to the pertinent area of computer vision research which involves automatically replicating real-world features from visual input. This information can however be obtained from other sources in some applications such as references that contain related texts or meta-data. To enable computers to meaningfully interpret and analyze visual information in a way similar to human perception is the aim of image understanding. In order to recognize objects scenes actions and relationships within image techniques such as computer vision deep learning and artificial intelligence must be applied. Through tasks like object detection image segmentation feature extraction and semantic analysis this process leverages visual data to glean insights. Image comprehension is crucial for applications like driverless cars medical imaging facial recognition and image captioning where accurate image interpretation is necessary for automation and decision-making [1]. Computer vision is a branch of artificial intelligence (AI) that enables computers to understand and interpret visual data from their environment. Methods for collecting processing analyzing and extrapolating important information from images or videos are necessary to mimic human visual perception. Key computer vision tasks include motion tracking object detection image classification facial recognition and scene comprehension. Many applications use this technology such as autonomous cars medical imaging security surveillance augmented reality and industrial automation which require intelligent machine analysis and responsiveness to visual inputs [2].

Fig. 1. Image Description Generation

To enable precise classification, detection or analysis, feature recognition is a process used in computer vision and pattern recognition that entails locating and extracting significant attributes from objects or images. Edges textures shapes, colors and other essential elements that set one object or pattern apart from another are examples of features. Modern deep learning-based approaches use convolutional neural networks (CNNs) to automatically learn and recognize complex features from large datasets while traditional methods use techniques like Scale-Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG) and Speeded-Up Robust Features (SURF) to extract meaningful visual features. Applications like facial recognition, biometric authentication, industrial automation and medical image analysis all make extensive use of feature recognition. It aids in object manipulation and navigation in robotics and in the detection of road signs and obstacles in autonomous vehicles. Additionally it is essential for content-based image retrieval handwriting recognition and manufacturing defect detection. Robust feature extraction techniques are crucial since the accuracy of feature recognition is influenced by variables such as lighting occlusions and changes in object appearance. As artificial intelligence and machine learning have developed feature recognition has become more effective and versatile enabling greater accuracy in a variety of real-world applications [8].

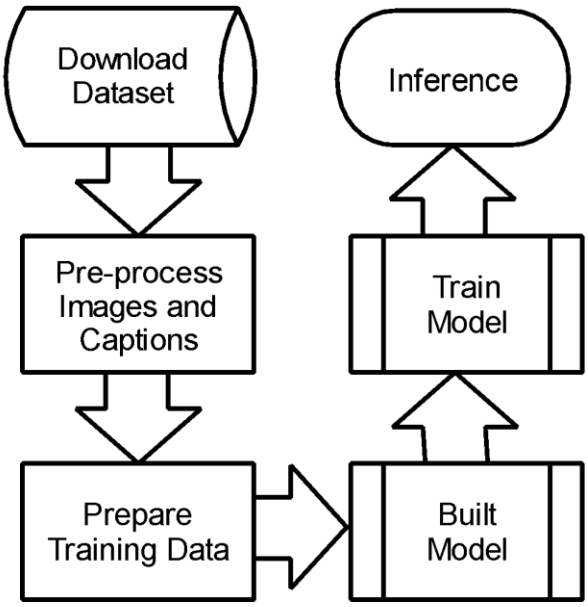

Fig. 2. Image Description Block Diagram in Machine Learning

As an encoder Convolutional Neural Networks (CNNs) have been used to extract features from input images. The decoder is linked to the hidden states of the CNN layers and network of recurrent neurons. Textual extraction from pertained features is started by the RNN which functions as a language processing model. The extraction of an accurate or precise caption remains a challenge in the fields of artificial intelligence and image processing. Both the grammar and the meaning of the extracted caption should be accurate. A better prediction model that can produce an accurate caption is what many researchers are searching for. Facebooks image captioning project will allow users to post their photos with automatically generated captions eliminating the need for users to manually compose the caption. Facebook suggested a package for natural language processing called Torch Vision. Torch vision is specifically made to recognize objects and analyze the situation in an image. The basic process model for captioning images is depicted in Figure 2 where the training data is prepared by first downloading the image dataset and pre-processing the images and captions. The system then finished training for inferences and constructed the model. Systems in earlier works are a little weak in certain areas making it difficult to accurately describe the image and resulting in a low BLEU score. A method for comparing and calculating accuracy between the reference and candidate captions is called BLEU. The most widely used application facilitates interaction between humans and computers and can be used to add subtitles to videos. Figure 3 displays the various captions that were anticipated for the provided image.

Actual Output : A dog is running through the snow
Predicted Output : dog is running through the snow

Fig. 3. Example of Image Description Generation

A. Application & Requirement of Image Captioning

Image captioning can be used for a number of purposes including automatic caption creation for Facebook or Instagram posts visual assistance image indexing editing application recommendations and other natural language processing tasks. With its wide range of applications and specifications image captioning is a challenging field. With audio facilities this model might be more helpful for blind people. Figure 4 shows the example of application of image captioning.

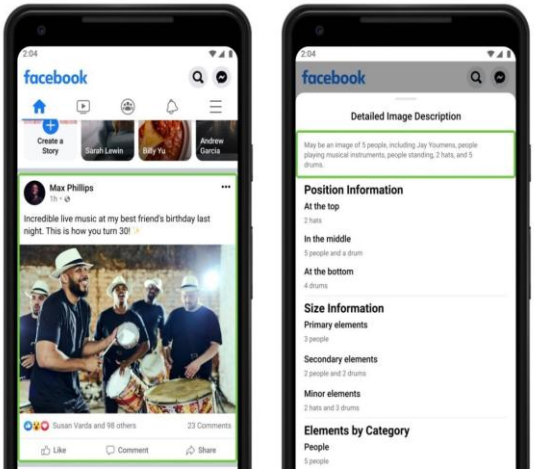

Fig. 4. Application of Image Captioning

2. Related Works

One popular method for producing descriptive language in a natural way is through image captions. The task of analyzing images and creating captions based on the results using computer vision techniques is a difficult one in the field of artificial intelligence. A system is necessary to create coherent sentences by relating specific elements that must be both syntactic and semantic. In order to get a better caption you must get the objects that are shown in the picture and explain how they relate to one another or how the activity is related to the situations in the picture. Numerous studies have been conducted using CNN RNN DNN LSTM and many other machine learning techniques to accomplish the goal of image captioning. The majority of the researchers compared their system to a number of benchmarks including COCO captioning Flickr8K Flickr30K and many others. However Flickr8K is the most widely used dataset which includes 8092 photos or challenges to test the systems functionality. K. Vijay et al. [63] presented a technique that is associated with the stemming algorithm. A natural language processing algorithm called the stemming algorithm is used to extract the image caption from various image types. It is a traditional approach unrelated to the machine learning methodology. It is predicated on the morphological variant approach which establishes a connection between words of similar types. This template-matching-based method is ineffective at achieving higher levels of accuracy with precise captions. Xinru Wei et al. [64] presented a technique related to DenseNet. The system makes use of the 10000-image CBIR benchmark. According to its name DenseNet is a highly densely populated network with layers that are extremely high in range. This makes the system more complex which lowers computation efficiency and necessitates a lot more memory to store the trained weight model. The training phase takes a very long time to complete. Because of its intricate layers DenseNet is incredibly slow. In comparison to ResNet VGG-16 RNN and many others the network should be lighter and employ effective pre-processing techniques to enable faster compilation. Yu Niange et al. [65] presented a technique that is associated with the Order Embedding algorithm. The system was designed to extract the image features and keywords using this type of algorithm then embed the keywords based on the field-related topic. It groups the keywords according to the category to which they belong. Both the upper and lower bound elements are included. A two-way approach is used. Iteration has been tainted in both directions to obtain a higher degree of accuracy along with more precise captions from top to bottom and bottom to top. Here the

system illustrates the outcome of the suggested system using two datasets: MSCOCO and Flickr30K. Min Yang et al. [66] presented a technique that is connected to the multitasking learning algorithm known as MLADIC. An algorithm consisting of two phases was employed to relate the information from the source image to the target information. Flickr 30K and Oxford 102 are used for testing and MSCOCO is used for model training. The source domain is thus the first one and the target domain is the second. Sharma Grishma et al. [67] presented a technique that is connected to CNN and RNN models. As with the other models CNN is used here to extract image features for keyword analysis which is then sent to the RNN to refine the meaningful sentences. They asserted a superior model over the GRU after comparing their results with the GRU (Gated Recurrent Unit) model. Additionally the system was compared to the VGG16 model and it was determined that CNN+RNN+LSTM could achieve better results than the VGG16 models. Wadhwa Tarun et al. [69] introduced a technique that uses LSTM to finally generate the image caption and a CNN-RNN hybrid model to extract the image features. The authors used the Flickr 30K benchmark which has 31783 photos with five captions each. A. Krishnakumar et al. [70] presented a technique related to deep learning methodology. Additionally they employed CNN and RNN learning models to analyze the features of the images and BLEU was used to obtain scores and compare them to the ground truth values. They employed datasets for image captioning that comprised 8091 images however the system was trained using 6000 images and tested using only 1000 images which is somewhat less. It is necessary to test the system with a minimum of 6000 images in order to verify its authenticity. Yinago Hidekazu et al. [72] presented a technique that has to do with the attention mechanism. The caption flow is managed by a multiple perceptual generator. LSTM is used for feature decoding while Attention is a weight model used for the encoding phase. The system employs VGG16 since it can be used to create multiple perceptive models and employs 16 layers to assess system performance. et al. Megha J. Panicker et al. [74] presented a technique that is associated with the Convolutional Neural Network (CNN) and the transfer learning algorithm. For increased accuracy this system combines CNN with Long Short Term Memory (LSTM). They conducted the testing phase and obtained the appropriate results using Flickr8K. Bilingual Evaluation The results are compared to the extracted text using Understudy. This system combined two different methods for creating captions to create a hybrid model.

3. Problem Identification

Automatic image caption generation is a contemporary engineering technique that identifies objects in the frames and provides a caption for the image based on the situation depicted. Numerous studies have been examined and certain shortcomings have been found. Numerous studies are based on CN-RNN techniques. Although RNN can be used to decode or terminate features it cannot be used as the primary model for benchmark image training. RNNs computational efficiency is a little bit lacking. The same vanishing and exploding structuring issue affects RNN. In comparison to other machine learning techniques the networks structure is complex and the training process is slow. Compared to other machine learning techniques some studies that rely on CNN+LSTM require more time to train the model. Additionally the large network necessitates more memory. Dropout is more difficult to execute. Order embedding and template augmentation are the foundations of a few. Methods based on templates and order embeddings are very traditional and do not use machine learning. Returning the result based on pre-defined templates is the conventional method. The best uses for DenseNet are in disease diagnosis and medical image classification. The perfect system must be able to work with any type of dataset achieve high accuracy and provide accurate captions. A lightweight network will help the system move toward a faster model.

4. Proposed Work

Image captioning is the process of describing an image using natural language. If you want to caption an image you must first identify the main objects their attributes and their relationships. The goal is to produce syntactically and semantically sound sentences that describe what the image shows. The concept of ground truth is not negligible in picture captioning. Five ground truth captions are included for every image in the COCO training and validation dataset. Figure 5 illustrates while some of the objects in the picture might be mentioned in the ground truth captions others might not. It seems that the authors of the ground truth captions disagreed on the primary concept of this image. All five authors make reference to the man in the foreground and the vehicle he is traveling in. The vehicles identity—

is it a bike or a scooter?—as well as the man's features like his shirtlessness and backpack camouflage pants—are already up for debate. It should be noted that the word camouflage is spelled incorrectly in the ground truth caption. Two captions identify the street as the location. They bring up the man in the wheelchair. The bench is not mentioned in the caption even though it is visible in the image and is even recognized as one of the items in the data set. The automated image captioning task is made more difficult by this variation in ground truth captions. Examples of useful applications for image captioning include content-based image indexing and retrieval visual intelligence in chatbots as well as assisting the blind and visually impaired by offering natural-language descriptions of an image or scene as speech. In the context of computer vision image captioning is categorized as a scene understanding task. Semantic understanding of visual scenes also known as scene understanding is a high-level task area that includes a variety of tasks such as object recognition object localization within the scene object attributes object characterization and the production of a semantic description of the scene. Image captioning is a crucial high-level task in the semantic understanding task family. It primarily relies on object detection and requires knowledge of the images higher-level context as well as the interdependencies of different object instances. The metadata is expanded with different language styles from two datasets FlickrStyle10k and Personality-Captions. With localized narratives captions are more explicitly anchored in the image. A subset of the Flickr30k dataset is combined with extra captions in various language styles such as romantic or humorous to create the FlickrStyle10k dataset [144]. FlickrStyle10k seeks to train models that can output language in particular styles thereby expanding the usage contexts of models in contrast to standard captioning datasets such as COCO and Flickr30k which offer factual captions.

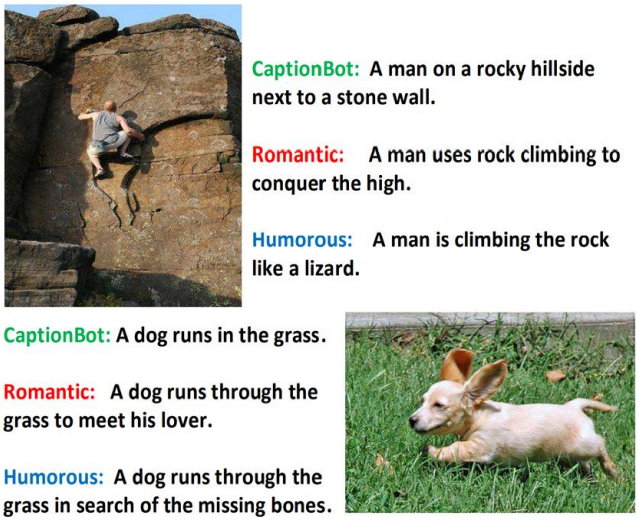

Fig 5. Captions with different styles based on the FlickrStyle10k dataset

The suggested system trains the model for creating image captions using ResNet. Skip connections allow ResNet to make efficient use of the layers. Even when training a model with hundreds or thousands of layers ResNet can still outperform the other models. Regression classification and object recognition are all excellent ways to represent the model. To train the model and produce a better model by omitting some layers the suggested system employs 152 layers. The system is lighter than the other models because it uses tiny filters just like the VGG-19. The suggested architecture is depicted in figure 6; an image acquisition is sustained encoded using a CNN model and trained using ResNet. With the help of the trainable layer a multilayer perceptron—a fully connected feed forward neural network—can generate candidate captions. The RNN then begins decoding the captions using the softmax layer.

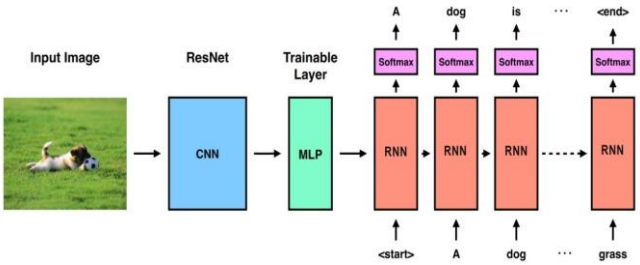

Fig 6. Proposed Architecture

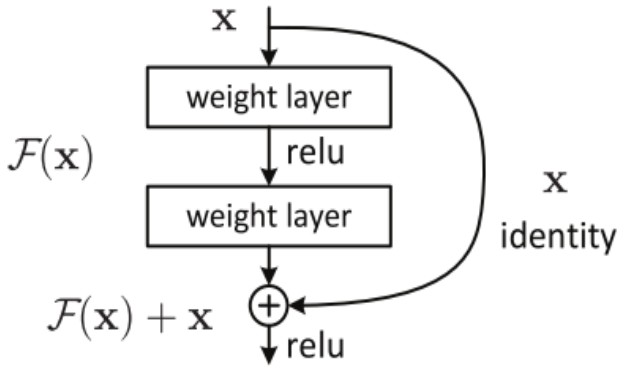

Fig 7. Residual Connection Network

where f(x) is the activation function f(x)+x is the forward identity function and x is regarded as identity. In this case the system uses the residual blocks to calibrate the model for training and its input passes through both the weight connections and the skip connections also referred to as identity shortcuts.

A. Encoder & Decoder

CNN is used in the proposed system to encode the data for machine translation. In this case the input is a sequence but both the input and output sequences may differ in length. One at a time the machine encodes and decodes each vector. Starting with a vector the machine begins encoding and decoding calculating all the way to the final conditional probability distribution.

$P(y_t = j|y_1^{t-1}, I)$ is the probability distribution

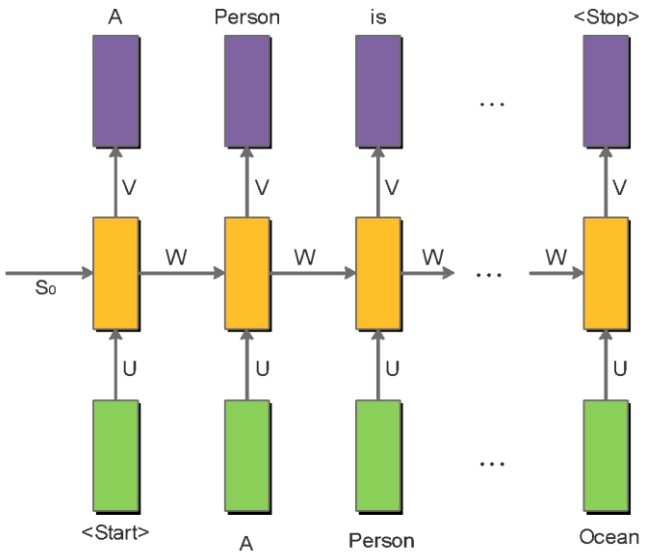

Fig 8. Encoder-Decoder Model

The Torch Vision library is also used to improve performance. The system trains and tests the model using the MSCOCO dataset which consists of 82783 photos with ground truth captions. The system is tested with 39622 images each with five captions and 40504 images are available for the validation process. As illustrated in Fig. 10 the training flowchart requires that the dataset be loaded with ground truth captions first and then the images are resized to 224 by 224 pixels before encoding. CNN converts the image into feature vectors after the data is ready and the ResNet-152 model can then be used for training and validation.

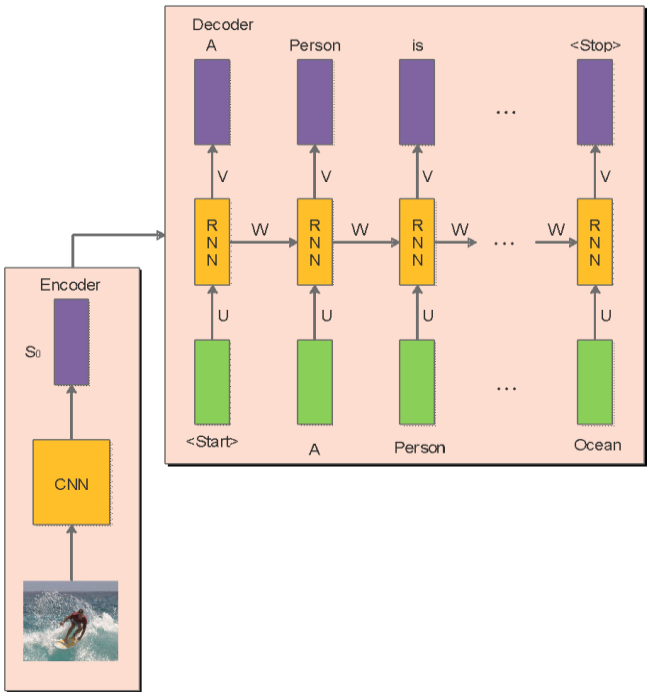

Fig 9. Encoder (CNN)-Decoder (RNN) Model

B. ResNet

In 2012 a deep learning model called Residual Network (ResNet) overtook ImageNet in other words ResNet supplanted ImageNet [48]. More layers in the model are necessary to achieve accurate predictions and accuracies but this is a common issue in deep learning since it increases the models weight and causes the system to experience the vanishing gradient problem. Because it employs the skip connection method ResNet differs significantly from other models. Skip connections allow hidden layers output to bypass some layers and go straight to the output. ResNet resolves the issue of vanishing or exploding gradients and enhances system performance. Tensorflow and the Keras API are necessary for ResNet to be implemented or for creating the models architecture. ResNet has been trained using a variety of datasets particularly CIFAR which has over 60000 photos of different types of objects including cars trucks horses cats dogs ships frogs and many more.

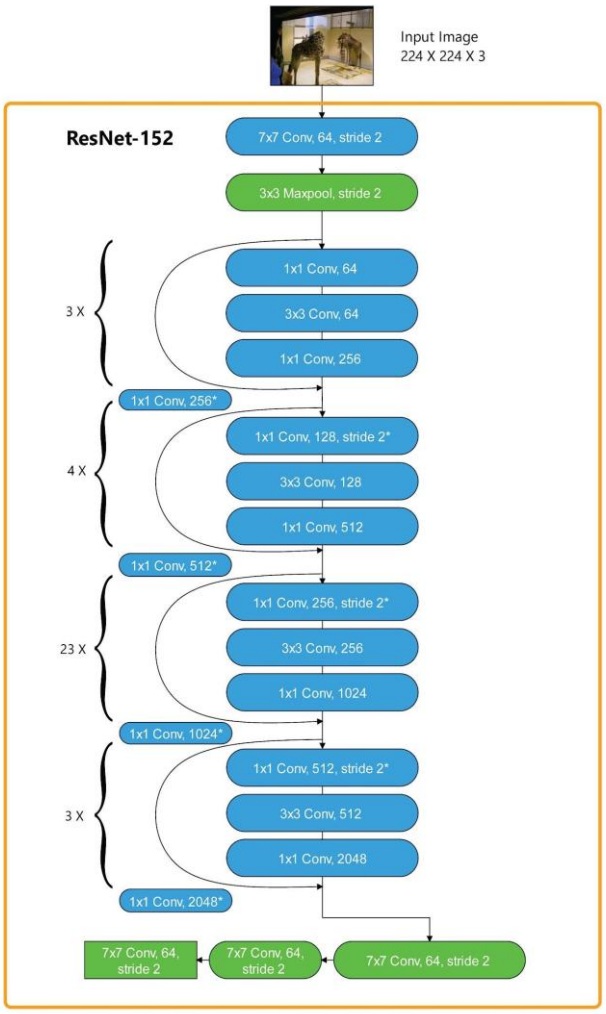

Fig 10. ResNet Architecture

C. Flow Chart

The training module flowchart is depicted in Figure 11 where the dataset and its ground truth captions are loaded first. After the system normalizes the data for CNN encoding the ResNet model is initialized and trained using the features that were extracted. Then caption can be decoded by using RNN and system pertained certain words tagged with start and end word. N is the total number of words in the candidate caption the system completes the extraction if the final word is found. The model has been trained using reference and candidate captions and if a word is not in the vocabulary it can be appended and a caption can be generated.

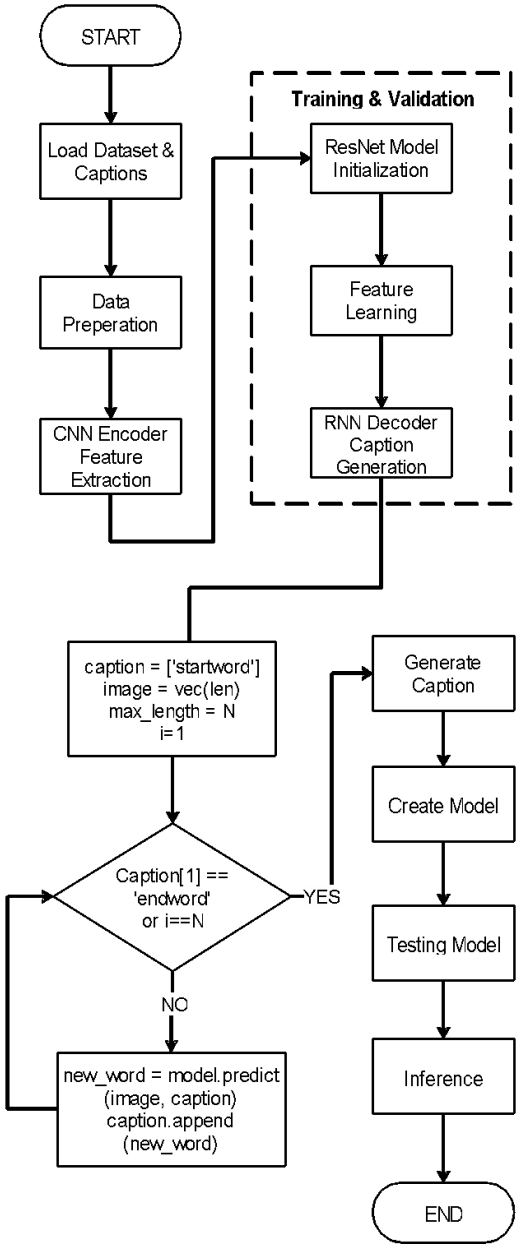

Fig 11. Flowchart of Training Module

Figure 12 illustrates the testing model flowchart in which the encoder and decoder models are loaded first followed by the ResNet model and input data for caption extraction. If there havent been any decode errors inference can be done from the first word to the last. After the final word is reached decoded words can be obtained and a caption can be created by employing an attention mechanism to sequentially arrange the words in a meaningful way. Once the candidate caption is complete it can be compared to the reference caption to calculate metrics like Rouge BLEU Meteor and CIDEr scores. These metrics can be used to determine how accurately the system determines how much the reference caption and the candidate caption differ or are similar. The system maintained higher scores than the template augmentation model that came before it. The ability of ResNet to quickly and accurately learn complex functions is very promising. Degradation problems occur when a models performance declines during training causing the system to deteriorate. Since skip connection resolved the degradation issue it can be regarded as the superior model

when compared to the others. ResNet can minimize the models load by skipping some layers and transferring the output of one layer straight to the next layer.

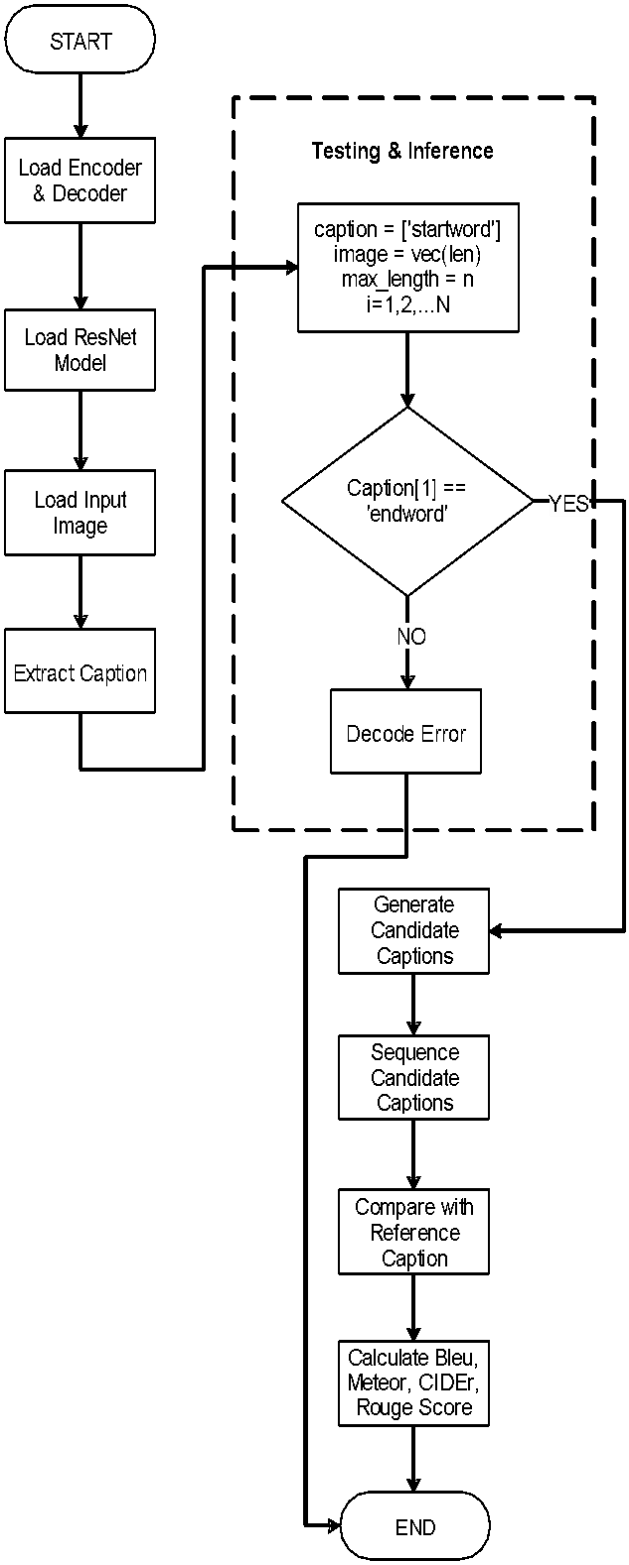

Fig 12. Flowchart of Testing Module

D. Algorithm

Table 1 ResNet-152 Algorithm

ResNet Image Captioning Algorithm

Initialization

Input: Set of Image A=(a1, a2, a3 ,…… aN) with annotation B=(b1, b2, b3 ,…… bN)

Output: Image Description

Step 1: Input frame

Step 2: Standardize frame by rescale

$$(wid_{new}, hei_{new}) = \frac{K}{max(wid, hei)}(wid, hei)$$

Where (wid_{new}, hei_{new}): the new width as well as height of the frame, K is considered as the switching value, $max(wid, hei)$is the maximum value

Step 3: Declare the identity block

f(a)+a

Step 4: Configure beam size, batch size and number of epochs, *Weih* as weight model for hidden layers, *Weib* as weight for output layer and *Bh, Ba* as bias for hidden as well as output layer respectively.

Step 5: Evaluate output for each convolutional layer

$$E_l = ReLU\left(b_l * s_l(E_{l-1}) + id(E_{l-1})\right)$$

If bl =1; it is treated as normal ReLU but if bl = 0 then it would be;

$$E_l = ReLU\left(id(E_{l-1})\right)$$

Step 6: Compute survival probability

$$P_l = 1 - \frac{l}{N}(1 - P_L)$$

Where PL is the survival probability and N is the total blocks in the model.

Step 7: Evaluate probability distribution

$$P_d(y_t = j | y_1^{t-1}, I)$$

Step 8: Built the model

Step 9: Load built model

Step 10: Decode knowledge

$$P_t = RNN\left(P_{t-1}, e\left(\hat{h}_{t-1}\right)\right)$$

h1, h2, h3, h4,……,ht-1 are hidden decoding states

Step 11: Create description using attention

Step 12: Correlate candidate with reference caption

Step 13: Evaluate BLEU, CIDEr, METEOR and ROUGE

Step 14: End

5. Result Analysis

We consider metrics such as BLEU METEOR and CIDEr scores to evaluate the efficacy of different image captioning techniques. These parameters assess the quality of generated captions by contrasting them with human-generated references. To assess the quality of the generated captions several metrics have been developed each focusing on a different aspect of the output.

A. BLEU (Bilingual Evaluation Understudy)

For tasks like image captioning and machine translation BLEU (Bilingual Evaluation Understudy) is one of the most widely used metrics for evaluating the quality of text generated by machines. BLEU was developed by Papineni and colleagues. n-grams which are continuous sequences of n items from a given text are used in 2002 to compute the overlap between the produced text and one or more reference texts.

$$Precision_n = \frac{Number\ of\ matched\ n\ grams}{Total\ number\ of\ n\ grams\ in\ generated\ text}$$

$$Brevity\ Penalty\ (BP) = \{1\ \ e^{\left(1-\frac{Lenth\ of\ reference\ text}{Lenth\ of\ generated\ text}\right)}$$

$$if\ length\ of\ generated\ text > lenth\ of\ reference\ text\ then\ 1$$

$$if\ length\ of\ generated\ text \leq lenth\ of\ reference\ text\ then\ e^{\left(1-\frac{Lenth\ of\ reference\ text}{Lenth\ of\ generated\ text}\right)}$$

$$BLUE = BP.exp\left(\sum_{n-1}^{N} \quad w_n\ log\ log\ (Precision_n)\right)$$

B. ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

A collection of measures called ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is mainly used to assess how well automatic text summarization algorithms summarize texts. ROUGE, created by Lin in 2004, is notably well-liked for challenges involving the comparison of produced texts with references authored by humans in the domains of machine learning and natural language processing.

$$Precision_L = \frac{Lowest\ Common\ Subsquence}{Total\ number\ of\ words\ generated}$$

$$Recall_L = \frac{Lowest\ Common\ Subsquence}{Total\ number\ of\ words\ in\ reference}$$

$$F1\ Socre = \frac{2.Precision_L.Recall_L}{Precision_L + Recall_L}$$

C. METEOR (Metric for Evaluation of Translation with Explicit ORdering)

METEOR is an evaluation metric designed to gauge the quality of text production-related jobs like machine-generated translations and image captioning. Developed in 2007 METEOR was intended to provide a more

comprehensive and nuanced evaluation than traditional metrics like BLEU by considering the generated texts grammatical and semantic properties.

$$Precision_L = \frac{Number\ of\ matches}{Total\ number\ of\ words\ generated}$$

$$Recall_L = \frac{Number\ of\ matches}{Total\ number\ of\ words\ in\ reference}$$

$$F1\ Socre = \frac{2.Precision_L.Recall_L}{Precision_L + Recall_L}$$

METEOR= BP.F1

D. CIDEr (Consensus-based Image Description Evaluation)

A measure called CIDEr (Consensus-based image Description Evaluation) was created expressly to assess how well machine learning models produce image descriptions. CIDEr, which was first presented by Vedantam et al. in 2015, focuses on the consensus across several human-generated captions in order to overcome some of the drawbacks of more conventional metrics like BLEU and ROUGE.

$$TF - IDF(w) = TF(w) \times IDF(w)$$

$$IDF(w) = log\ log\ \left(\frac{Total\ number\ of\ captions}{1 + number\ of\ caption\ containing\ w}\right)$$

$$CIDEr(n) = \frac{\sum_{i=1}^{N} \quad TF - IDF(w_i).Count_{match}(w_i)}{\sum_{i=1}^{N} \quad TF - IDF(w_i).Count_{total}(w_i)}$$

$$CIDEr = \frac{1}{N}\sum_{n=1}^{N} \quad CIDEr(n)$$

E. Result Obtained

Figure 13 shows the screenshot of the result obtained by the proposed model.

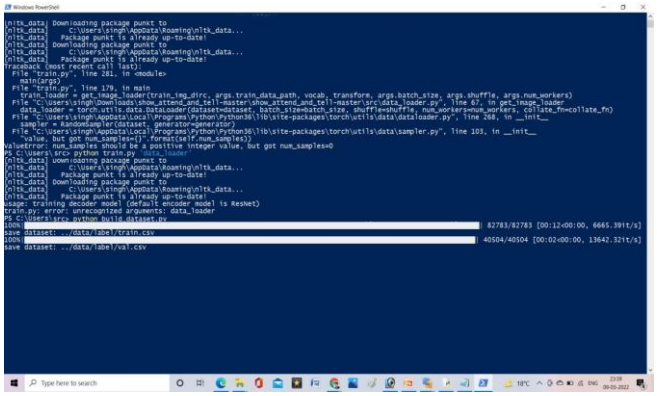

Fig 13. Screenshot of Recorded Result

F. Datasets

MSCOCO also known as Microsoft COCO [159] is a very large dataset that researchers can use for image segmentation image recognition and image captioning. It includes over 40000 images for the validation process and over 82000 images for training. A well-liked benchmark dataset for computer vision and natural language processing is Flickr30k also referred to as Flickr30k Entities. Flickr a photo-sharing website provides 31783 images [6] with five human-written captions each. Objects or areas of interest in the photos are surrounded by bounding boxes that offer accurate localization information. Flickr30k also known as Flickr30k Entities is a popular benchmark dataset for computer vision and natural language processing.

Table 2 Comparison between MSCOCO and Flickr30K

Feature	MSCOCO	Flickr30k
Full Name	Microsoft Common Objects in Context	Flickr30k
Total Images	328,000	31,783
Number of Captions	Over 1.5 million	158,915
Captions per Image	5	5
Image Source	Various sources, including Flickr	Flickr
Image Content	Diverse scenes with everyday objects	Images more focused on people and actions
Annotation Quality	High-quality, crowd-sourced annotations	High-quality, crowd-sourced annotations
Dataset Split	Training: 118,287 images	Training: 29,783 images
	Validation: 5,000 images	Validation: 1,000 images
	Test: 5,000 images	Test: 1,000 images
Use in Research	Widely used for benchmarking image captioning models	Commonly used but less diverse than MSCOCO
Challenges	Large dataset, diverse object categories	Smaller dataset, often used for initial testing
Benchmarks and Metrics	Evaluated with BLEU, METEOR, ROUGE, CIDEr	Evaluated with BLEU, METEOR, ROUGE
Accessibility	Freely available	Freely available

Table 3 Experimental Results

Dataset	MSCOCO
B1	0.579
B2	0.404
B3	0.279
B4	0.191
M	0.195

R	0.396
C	0.6

Table 4 Results Comparison

	Ren C. Luo [16]	**Proposed**
B1	0.238	0.57
B2	0.109	0.404
B3	0.05	0.279
B4	0.022	0.191
M	0.096	0.195
R	0.249	0.396
C	0.143	0.6

6. Conclusion & Future Scope

One difficult task in the field of artificial intelligence is captioning images. Image captioning has been the subject of numerous studies and acute and precise captioning still requires the highest level of precision. A system is necessary to create coherent sentences by relating specific elements that must be both syntactic and semantic. In order to get a better caption you must get the objects that are shown in the picture and explain how they relate to one another or how the activity is related to the situations in the picture. Many machine learning techniques can be used to accomplish the goal of image captioning and numerous studies have used CNN RNN DNN LSTM and many more. However increased precision and reduced processing time are necessary. The main training model for the suggested system is ResNet along with CNN as the encoder RNN as the decoder and Torch Vision as a library. ResNet can make use of the layers by using the skip connection technique. Compared to the previously suggested model such as the Template Augmentation Method the system maintained a high level of precision. The MSCOCO benchmark was retained by the system which trained a model for it and obtained a more effective model for captioning images. The system may eventually be evaluated using various benchmarks that comprise multiple pictures or data. Future network size reductions can minimize training time and correspondingly lower system costs. BLEU METEOR CIDEr and ROUGE which characterize the quality of machine translation can be improved in the future. A lot of researchers are now using hybrid models to improve system performance. By combining two different methods they are able to improve their prediction accuracy. ResNet can be used in the future to train various datasets and improve precision. Precision can be improved by using ResNet a little more in the future. Semantic concept-based techniques or any other method that can more effectively rearrange the generated words to improve the sentences meaning with pertinent information can take the place of CNN and RNN.

REFERENCES

1. Lee, S., & Kim, I. (2018). Multimodal feature learning for video captioning. Mathematical Problems in Engineering, 2018.
2. X. Wei, Y. Qi, J. Liu and F. Liu, "Image retrieval by dense caption reasoning," 2017 IEEE Visual Communications and Image Processing (VCIP), 2017, pp. 1-4, doi: 10.1109/VCIP.2017.8305157.
3. Feng et al., "Image Captioning: Transforming Objects into Sentences," IEEE Transactions on Image Processing, 27(6), 2875-2889, 2018.
4. Krahmer et al., "Natural Language Generation in Discourse Context: A Survey," Computational Linguistics, 28(4), 495-528, 2002.
5. Reiter, E., & Dale, R. (2000). "Building Natural Language Generation Systems." *Cambridge University Press*.
6. Liu, J., Chen, Y., & Zhang, J. (2018). "Automatic Image Captioning with Template-Based Systems." *Journal of Visual Communication and Image Representation, 54*, 140-150.
7. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). "Show and Tell: A Neural Image Caption Generator." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

8. Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). "Spice: Semantic Propositional Image Caption Evaluation." *European Conference on Computer Vision*.
9. Sharma, Himanshu & Srivastava, Swati. (2022). Multilevel attention and relation network based image captioning model. Multimedia Tools and Applications. 82. 1-23. 10.1007/s11042-022-13793-0.
10. Xu, K., Zhu, Y., & et al. (2015). "Show, Attend and Tell: Neural Image Caption Generation with Attention Mechanism." *Proceedings of the International Conference on Machine Learning*.
11. Radford, A., Kim, J. W., Hallacy, C., et al. (2021). "Learning Transferable Visual Models From Natural Language Supervision." *International Conference on Machine Learning*.
12. Dong, H., Yang, S., & Zhang, C. (2017). "Image Captioning with Reinforcement Learning." *Proceedings of the IEEE International Conference on Computer Vision.
13. Xu, K., Zhu, Y., & et al. (2015). "Show, Attend and Tell: Neural Image Caption Generation with Attention Mechanism." *Proceedings of the International Conference on Machine Learning.
14. Vaswani, A., Shard, N., & et al. (2017). "Attention is All You Need." *Advances in Neural Information Processing Systems.
15. Chen, J., Chen, K., & et al. (2020). "Self-Attention for Image Captioning." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
16. Li, Y., Wang, S., & et al. (2018). "Multimodal Attention for Image Captioning." *International Journal of Computer Vision.
17. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *arXiv preprint arXiv:2010.11929.
18. Gao, Y., Liu, J., & et al. (2021). "Image Transformer." *Proceedings of the International Conference on Machine Learning.
19. *Brown, T. B., Mann, B., Ryder, N., et al. (2020). "Language Models are Few-Shot Learners." *Advances in Neural Information Processing Systems.*
20. Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). "BLEU: a method for automatic evaluation of machine translation." *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 311-318.
21. Lin, C. Y. (2004). "ROUGE: A Package for Automatic Evaluation of Summaries." *Proceedings of the Workshop on Text Summarization Branches Out.
22. Lavie, A., & Agarwal, A. (2007). "Meteor: An Automatic Metric for MT Evaluation." *Proceedings of the Second Workshop on Statistical Machine Translation*, 1-4.
23. Vedantam, R., Zitnick, C. L., & Parikh, D. (2015). "CIDEr: Consensus-based Image Description Evaluation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4566-4575.
24. Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). "SPICE: Semantic Propositional Image Caption Evaluation." *European Conference on Computer Vision (ECCV)*.