IMAGE CAPTION GENERATOR USING DEEP LEARNING MODEL

Namitha O
BTech CSE
Vellore Institute of Technology, Chennai
namitha1903@gmail.com

Kavitha D
School of Computer Science and Engineering
Vellore Institute of Technology,Chennai
kavitha.d@vit.ac.in

***Abstract*—Image captioning is a popular technique used to generate a description of an image. This approach involves identifying the essential objects and their features in an image, as well as the relationships between them. Recently, image captioning has become a valuable tool for a wide range of applications. To address this need, we propose a deep learning model that utilizes computer vision and machine translation to generate English sentences that describe the content of an image. The resulting captions can be particularly beneficial for visually impaired individuals, as they can use the descriptions to interpret visual information on the web. Additionally, the automated image interpretation provided by the model can replace the need for human interpretation. The descriptions generated by the model not only capture the content of the image but also the relationships between the objects and their respective attributes and functions.**

Keywords - Image, Caption, CNN, RNN, LSTM, Neural Networks

I. INTRODUCTION

Photo captions also referred to as captions, are utilized to communicate the visual information of an image in a localized language. The process of generating captions for photos is complex, as it requires the identification of objects in the image and their relationships, as well as the ability to fine-tune and express this information using natural language.Research in the field of image captioning has advanced significantly with recent developments in image classification, object recognition, and language modeling. These captions can be particularly beneficial for individuals with visual impairments as they can aid in interpreting web-based information. However, generating descriptive captions in well-structured English is a challenging task that requires the recognition of image content, relationships between objects, and language processing. Unlike traditional research on object recognition or image analysis, image captioning involves not only capturing the image's content but also conveying the objects' qualities and functions. This task requires the use of a natural language model such as English to provide semantic information. [1]The proposed Siamese Difference CaptioningModel (SDCM) accomplishes competitive performanceon the Spot-The-Diff baseline dataset, producing succinct,concise, meaningful, and readable textual interpretation with a commendable result.[2] A novel dual-attention image caption generation model has been proposed to exploit both visual attention andtextual attention, where visual attention enhances the understanding of image details and textual attention increases the integrity of the information.[3] A novel model is presented for image captioning,which is based on the original idea of GANs. It does notutilize intermediate algorithms such as policy gradient fortraining the model.[4] Their main attention is focused on neural network-based methods, which give state-of-the-art results. Because different frameworks are used in neural network-basedmethods. They divided them into subcategories and discussed each subcategory, respectively.[5]They proposed a novel image captioningmodel, called domain-specific image captioning generator, which generates a caption for a given image using visual and semantic attention, and produces a domain-specific caption with semantic ontologyby replacing the specific words in the general caption with domain-specific words.[6] They proposed *EnsCaption* which aims at enhancing a generation-retrieval ensemble model with a novel dual generator generative adversarial network, allowing for bothgeneration-based and retrieval-based image captioning methods to be mutually enhanced.Deep learning techniques use layered networks to simulate the human brain and identify interesting features from an image. Previous attempts to combine existing solutions have been made to move from the image to its definition. However, we propose a single integrated model that takes an image input and produces a direct sequence of words from a dictionary that adequately describes the image, increasing the likelihood of p (S | I). This research is motivated by recent developments in machine translation, which require increasing the function p(T|S). While traditional machine translation involved word-for-word translation, recent research has shown that using RNN can simplify the process and achieve modern functionality. The encoder reads the original sentence and turns it into a fixed-length vector. In recent years, CNN has been used to create a rich visual representation in a fixed-length vector that can be used for various visual functions. Therefore, it is natural to use a Convolutional Neural Network as an encoder by training it to split the image and using the hidden end layer as an RNN-based Long short-term memory layer of the decoder to generate captions. This model is known as Neural Image Captioning and is a fully trained stochastic gradient neural network. Additionally, our model integrates low-level networks with language models to better understand the image as a whole and extract details about each item and their relationships. In conclusion, the system can automatically construct a natural sentence to explain the given input image.

II. RELATED WORK

[1] Oluwasanmi, A., Aftab, M. U., Alabdulkreem, E., Kumeda, B., Baagyere, E. Y., & Qin, Z. (2019). Captionnet: Automatic end-to-end siamese difference captioning model with attention. *IEEE Access*, *7*, 106773-106783.CaptionNet: Automatic End-to-End SDCM With Attention, A. Oluwasanmi et al., 2019. A deep neural network is used in the proposed multi-modal end-to-end encoder-decoder design to create a natural sentence characterization for contrast within the image pairs. Their proposed model which is a supervised model incorporates many numbers of deep learning strategies in assessing the feasibility of photography, alignment, and computer-assisted variance between the two image elements, to create a wide range of coherent language model opportunities. A basic spot-the-difference baseline dataset with pairs of comparable images and descriptions is used to test model tests. To identify the contrast of two input images, the encoder uses a Siamese network that uses the same deep Convolutional Neural Network architecture that has a favorable effect on this data for 0.371 as Bleu1.[2]Liu, M., Li, L., Hu, H., Guan, W., & Tian, J. (2020). Image caption generation with a dual attention mechanism. *Information Processing & Management*, *57*(2), 102178. Their proposed model explores visual attention to deepen the understanding of the image, incorporating the image labels generated by a Fully Convolutional Network (FCN) into the generation of image captions. Furthermore, the model exploits textual attention to increasing the integrity of the information. Finally, the label generation, attached to the textual attention mechanism, and the image caption generation, have been merged to form an end-to-end trainable framework. The experimental results on the AIC-ICC image Chinese caption benchmark dataset show that their proposed model is effective and feasible.[3] Dehaqi, A. M., Seydi, V., & Madadi, Y. (2021). Adversarial Image Caption Generator Network. *SN Computer Science*, *2*(3), 1-14.

They proposed a novel model based on GAN networks where it generates the caption of the image through the representation of the image by utilizing the generator adversarial network and it does not need any secondary learning algorithm like policy gradient. Due to the complexity of benchmark datasets such as Flickr and Coco, in both volume and complexity, they introduced a new dataset and performed experiments on it. The experimental results show the effectiveness of the model compared to the state-of-the-art image captioning methods.[4] Bai, S., & An, S. (2019). A survey on automatic image caption generation. *Neurocomputing*, *311*, 291-304.Presented a survey on image captioning. Based on the technique adopted in each method, classified image captioning approaches are into different categories. Representative methods in each category were summarized, and the strengths and limitations of each type of work were talked about.Discussed early image captioning work which is mainly retrieval-based and template based. Then, the main attention was focused on neural network-based methods, which gave state-of-the-art results. Because different frameworks are used in neural network-basedmethods, further divided them into subcategories and discussed each subcategory, respectively. After that, state-of-the-art methods are compared on benchmark datasets. Finally, presented a discussion on future research directions of automatic image captioning.[5]Han, S. H., & Choi, H. J. (2020, February). Domain-specific image caption generator with semantic ontology. In *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)* (pp. 526-530). IEEE.Proposed a novel image captioning model, called domain-specific image captioning generator, which generates a caption for a given image using visual and semantic attention, and produces a domain-specific caption with semantic ontology by replacing the specific words in the general caption with domain-specific words. In the experiments, evaluated image caption generator qualitatively and quantitatively. The goalis to generate the domain-specific caption. In the domain-specific captions, the underline words are replaced words for general words in the general captions through the semantic ontology that is defined.[6]Yang, M., Liu, J., Shen, Y., Zhao, Z., Chen, X., Wu, Q., & Li, C. (2020). An ensemble of generation-and retrieval-based image captioning with dual generator generative adversarial network. *IEEE Transactions on Image Processing*, *29*, 9627-9640.

Proposed a novel EnsCaption model, which aims at enhancing an ensemble of retrieval-based and generation-based image captioning methods a caption generation model that synthesizes tailored captions for the query image, a caption re-ranking model that retrieves the best-matching caption from a candidate caption pool consisting of generated captions and pre-retrieved captions, and a discriminator that learns the multi-level difference between the generated/retrieved captions and the ground-truth captions. During the adversarial training process, the caption generation model and through a novel dual generator generative adversarial network. Specifically, EnsCaption was composed of the caption re-ranking model provide improved synthetic and retrieved candidate captions with high-ranking scores from the discriminator, while the discriminator based on the multi-level ranking was trained to assign low-ranking scores to the generated and retrieved image captions. The model absorbs the merits of both generation-based and retrieval-based approaches.

III. SYSTEM ARCHITECTURE

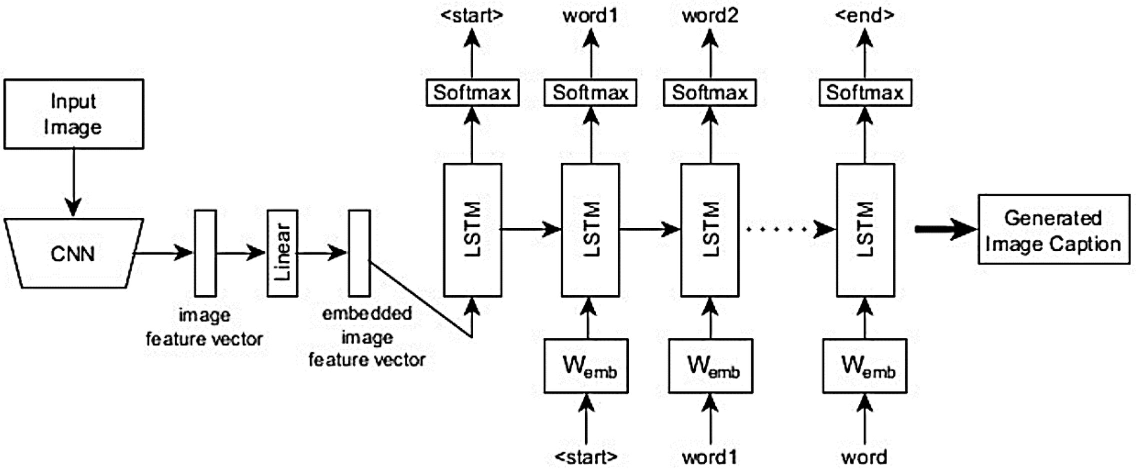


Fig1. Model Architecture

The system shown in Fig1 represented in this work employs a model consisting of a combination of a Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) to process one image input and produce a description of that image as output. In this model, an encoder is used to convert a sentence of varying length into a fixed-size vector representation, which is then sequentially taken as the initial hidden state by the decoder to generate an output in the form of a meaningful sentence. Instead of using a Recurrent Neural Network (RNN) as the decoder, we opt for a deep CNN as it is better suited for producing a high-quality representation of the input image through embedding with a fixed-size vector. Furthermore, we pre-train the CNN for image classification and extract the backend hidden layer of the network as the image representation. This image representation is then fed into the decoder, which is an LSTM, to generate a sentence description of the input image. By using this approach, we are able to produce a description of the input image that is both accurate and natural-sounding.

From the fig2, every cell knows how to measure those components for the input gate and how to adjust this input memory as an input modulator. Then this also knows the weight that removes a memory cell from the forget gate as well as the weight that handles the extraction of that memory (the output gate). The key behind the long-short-term memory structure is a horizontal, vertical line called a cell state. This cell status applies to every recurring section as well as being adjusted for all modules and is being taken care of by the gate. As a result, this information on the LSTM-based structure continues.

The module's output pt+1 contains the predicted word. The same long-short-term memory structure repeats till the last tokens (.) that were detected for a structure A sequence for those word predictions forms a picture explanation of the specific input.

This total train procedure is a combination model that is represented as a CNN encoder along with an RNN language generator.

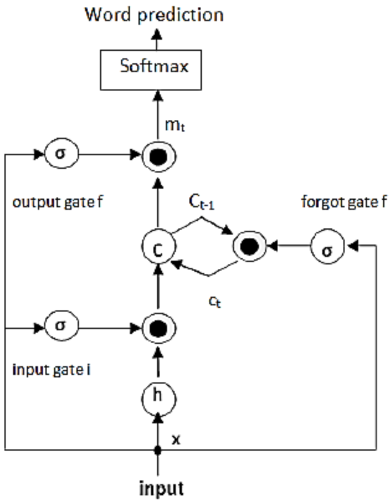


Fig2. LSTM Architecture for Language Generation

IV. METHODOLOGY

i. DATASET

MS COCO (Microsoft Common Objects in Context) is the largest dataset there are about 300,000 photos in total, and each with at least five captions comes with class labels, image segment labels, and a set of photo captions provided as annotations. Microsoft has presented a visual dataset having a large number of images depicting common objects in difficult daily forums. This makes MSCOCO different from other object detection datasets that possibly be specific areas of AI. This dataset is often used for training and benchmark detecting objects and is an example of each of the 80-item classification masks, pixel-class classification by 91 categories, algorithms for captions, and full panoptic group classification 80 thing categories.

ii. PREPROCESSING

There are several annotated picture collections that are available for this pre-processing purpose. Pre-processing involves loading the dataset and caching the output from the InceptionV3 model to the disk.

iii. INCEPTION V3

In the inceptionV3 architecture, we're now developing a tf.keras model with the exit layer as the conversion output layer. The output layer is 8x8x2048 in size. Each picture is transferred over the network, and the resultant vector represented as image name -> feature vector is recorded in the dictionary. After each image has been processed with InceptionV3, save the output to the disc.

iv. ALGORITHMS

a. ENCODER-DECODER FRAMEWORK

In the encoding section, at first, the model that identifies a context within an image that is given as input is represented, along with the convolutional neural networks for extracting the positions as it indicates an association of spatial relationships. A convolutional neural network is used for image processing and provides all the details of a close image, such as light, length, width, edges, etc. Annotation vectors, which are a collection of feature vectors, will be used to represent all data. The encoding section of the code generates L annotation vectors, each representing a D-dimensional corresponding object and its area of location, which is spatial in the image that is fed as input. Within the decoder section, a total of those annotation vectors were incorporated as a deep neural network model to produce the descriptive text. To generate meaning from images, we present a probabilistic and neural framework. Regular advancements within mathematical translation technology demonstrated that modern results can be achieved by directly increasing the chances of a precise translation in the form of an end-to-end method that takes within itself training as well as an inference based on an input sentence. These models employ RNN to combine a varying-size image as a fixed-size vector as well as "divide" the output sentence.

Image detection along with classification has been proven to be very beneficial for this network, They are widely used in object identification functions, owning drivable cars, writing photo sentences, and so on. This architecture takes a modern convolutional neural network in order to retrieve those features of input as well as convert it into the central vector with a fixed dimension. The initial condition for an RNN decoder is triggered by a received picture fragment. The proposed model would input one picture along with producing the y-encode caption for the 1-with-K word, which is encoded as a sequence.

$$y = \{y_1, \ldots, y_C\}, \; y_i \in R^K$$

c. DECODER: LSTM NETWORK-BASED SENTENCE GENERATOR

To generate authentic captions, LSTM employs recurrent neural networks. RNN is used to produce captions from the CNN output. To avoid gradient extinction, the RNN employs an LSTM layer, which generates captions by producing a single word for each step set in the context vector, previously hidden state, and produced words.

The probability of each word in the lexicon is the LSTM output, and to produce sentences, beam search is employed. Beam search is a heuristic search approach that looks for the most promising node in a graph with a limited number of possibilities. We employ k-best search in addition to beam search to produce phrases. Very similar to Viterbi's consistent timed search. The method repeatedly selects the best k sentence for all candidate sentences up to t and only retains the best effect of its k.

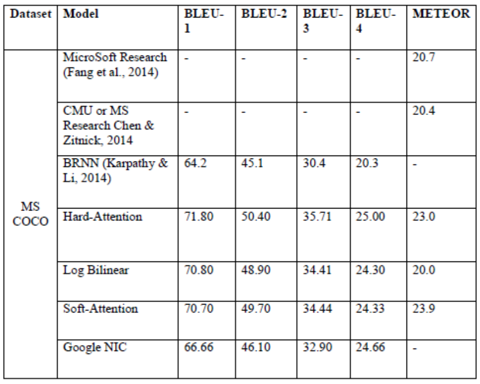
d. ATTENTION MECHANISM

When the Decoder creates each word in the output sequence, the attention module helps it to focus on the appropriate area of the image. When attention is absent, the decoder creates returning words by taking into account all areas of the image equally. This module takes the encoded picture as input and the previous stage's hidden Decoder state for each cycle. For instance, if the goal series is "Boy Eats a Banana," the boy's pixel inside the image is created as the word "Boy," and the banana's pixel is produced as the word "Banana" in pixels. Moreover, it produces the Attention score and assigns a weight to each pixel of the encoded image. The term that will appear in the following stage is related to this pixel weight when it is high. The RNN is then given the score together with the input name for the time step. It enables our RNN to focus on a photo's most crucial elements and generate the appropriate return phrases. By methodically focusing on important areas of the image, the attention modules help create a complete contextual context. Since it maintains the alignment of the crucial portions of the text image, this module considers the weight of each grid of the encoder vector feature. The relative rapidity of the annotation or grid in the creation of sequential word order is therefore appropriately indicated by the weight of attention.

b. ENCODER: CONVOLUTIONAL NEURAL NETWORKS FOR FEATURE EXTRACTION

IV. MODEL TRAINING

The Inception-V3 model is used to extract image features from the lowest convolutional layer, resulting in a vector of size (8, 8, 2048) which is then reshaped to (64, 2048). This reshaped vector is then fed into the CNN Encoder, which is responsible for understanding the content of the image by assigning weights to different image elements and separating them. To generate the image caption, an RNN is used to iterate over the image. The model first converts the image into a word vector, which is then passed to LSTM cells to generate a sentence word by word. This process involves using the context vector, previous hidden state, and pre-existing words to generate the next word in the sentence.



Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
MS COCO	MicroSoft Research (Fang et al., 2014)	-	-	-	-	20.7
	CMU or MS Research Chen & Zitnick, 2014	-	-	-	-	20.4
	BRNN (Karpathy & Li, 2014)	64.2	45.1	30.4	20.3	-
	Hard-Attention	71.80	50.40	35.71	25.00	23.0
	Log Bilinear	70.80	48.90	34.41	24.30	20.0
	Soft-Attention	70.70	49.70	34.44	24.33	23.9
	Google NIC	66.66	46.10	32.90	24.66	-

Table-1. With the MSCOCO dataset, BLEU-1,2,3,4 and METEOR measures were comparing with other approaches.

The implementation steps are as follows:

Choose a group of words that may be in the image's caption. In order to train the detectors iteratively, we use the weak monitoring method in multi-instance learning (MIL) to identify the words from the given vocabulary that correlate to the content of the related image.

We obtain a preliminary spatial response graph by applying a fully convolutional network to a picture. Each place on the response map represents a response that was discovered by applying the original CNN to the area of the input picture when the shift is applied (essentially scanning various areas of the image for potential items). We get a response map on the final fully connected layer by upsampling the picture, and we then apply the noisy-OR variant of MIL on the response map for each image. Every word generates a distinct probability.

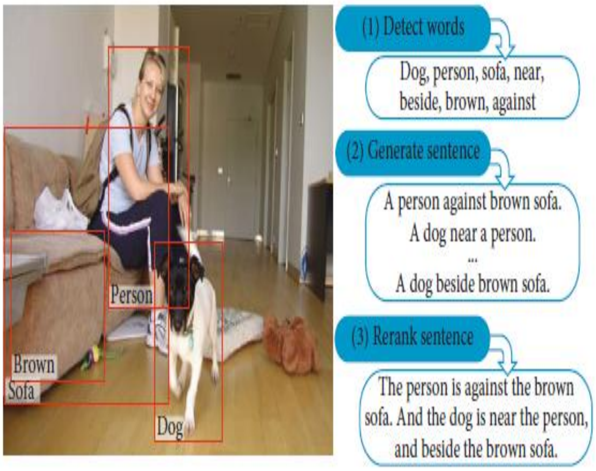

Person
Brown Sofa
Dog
(1) Detect words
Dog, person, sofa, near, beside, brown, against
(2) Generate sentence
A person against brown sofa.
A dog near a person.
...
A dog beside brown sofa.
(3) Rerank sentence
The person is against the brown sofa. And the dog is near the person, and beside the brown sofa.

Fig3. Method Based on the visual detector and language model.

In the process of creating captions, words are visually identified, and the most likely phrase is then found. The core of this procedure is the 0e language model, which specifies the probability distribution of a word sequence. Even though the maximum entropy language model (ME) is a statistical model, it can encapsulate data that is extremely meaningful.

For instance, the word "horse" is more likely to be followed by "running" than by "speaking." The right words can be distinguished using this information, and common-sense knowledge may be encoded.

V. CONCLUSION AND FUTURE WORK

The proposed model utilizes a convolutional neural network (CNN) encoder for feature extraction and an LSTM decoder for generating phrases based on the image content. The model aims to increase the probability of generating an accurate description for a given image and incorporates attention-based techniques for improved performance on the MS COCO dataset using BLEU metrics. The picture caption generator has potential applications for visually impaired individuals. Future research may investigate the impact of subsampling on CNNs and explore the use of auto-encoders for more accurate feature vectors. Additionally, the model can be extended to dense captioning and long description generation for other languages, as well as applied to tasks such as video captioning and visual question answering.

VI. REFERENCES

[1] Oluwasanmi, A., Aftab, M. U., Alabdulkreem, E., Kumeda, B., Baagyere, E. Y., & Qin, Z. (2019). Captionnet: Automatic end-to-end siamese difference captioning model with attention. *IEEE Access*, *7*, 106773-106783.

[2] Liu, M., Li, L., Hu, H., Guan, W., & Tian, J. (2020). Image caption generation with a dual attention mechanism. *Information Processing & Management*, *57*(2), 102178.

[3] Dehaqi, A. M., Seydi, V., & Madadi, Y. (2021). Adversarial Image Caption Generator Network. *SN Computer Science*, *2*(3), 1-14.

[4] Bai, S., & An, S. (2019). A survey on automatic image caption generation. *Neurocomputing*, *311*, 291-304.

[5]Han, S. H., & Choi, H. J. (2020, February). Domain-specific image caption generator with semantic ontology. In *2020 IEEE International*

Conference on Big Data and Smart Computing (BigComp) (pp. 526-530). IEEE.

[6]Yang, M., Liu, J., Shen, Y., Zhao, Z., Chen, X., Wu, Q., & Li, C. (2020). An ensemble of generation-and retrieval-based image captioning with dual generator generative adversarial network. *IEEE Transactions on Image Processing*, *29*, 9627-9640.

[7]Anu, M., & Divya, S. (2021, May). Building a voice-based image caption generator with deep learning. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 943-948). IEEE.

[8] Dehaqi, A. M., Seydi, V., & Madadi, Y. (2021). Adversarial Image Caption Generator Network. *SN Computer Science*, *2*(3), 1-14.

[9]Wang, Y., Zhang, C., Wang, Z., & Li, Z. (2022, July). Image Captioning According to User's Intention and Style. In *2022 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-9). IEEE.

[10]Verma, A., Saxena, H., Jaiswal, M., & Tanwar, P. (2021, July). Intelligence Embedded Image Caption Generator using LSTM-based RNN Model. In *2021 6th International Conference on Communication and Electronics Systems (ICCES)* (pp. 963-967). IEEE.