e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

# VISUAL SEMANTIC EXTRACTION FOR TEXTUAL DESCRIPTION USING CNN AND LSTM

**Sahadev Bhaganagare[*1], U. Mukesh Gopinandh[*2], P. Nithin Kumar[*3], Dr. K. Suresh[*4], K. Chandusha[*5]**

[*1,2,3]Final Year Student, Department Of Information Technology Malla Reddy College Of Engineering And Technology Hyderabad, Telangana, India.

[*4]Associate Professor, Department Of Information Technology Malla Reddy College Of Engineering And Technology Hyderabad, Telangana, India.

[*5]Assistant Professor, Department Of Information Technology Malla Reddy College Of Engineering And Technology Hyderabad, Telangana, India.

## ABSTRACT

The process of generating a textual description for images is known as image captioning. Nowadays it is one of the recent and growing research problems. Day by day various solutions are being introduced for solving the problem. Even though many solutions are already available, a lot of attention is still required for getting better and precise results. So, we came up with the idea of developing an image captioning model using different combinations of Convolutional Neural Network architecture along with Long Short Term Memory in order to get better results. This project focuses on the realm of image captioning with the help of machine learning, employing advanced neural network architectures to unite the semantics of visual content and natural language descriptions. The project "Visual semantic extraction for textual description using CNN and LSTM" is a deep learning based project that leverages the power of convolutional neural networks (CNNs) for image feature extraction and long short term memory (LSTM) for sequential language generation, our model endeavours to independently generate descriptive captions for diverse image. There are multiple use cases and applications of this model like accessibility, social media content sharing, human-computer interaction and robotics. In conclusion, image captioning, driven by advancements in machine learning, has emerged as a powerful tool for bridging the gap between visual content and natural language. The continuous improvement in accuracy and versatility signifies its growing significance in diverse applications, promising a more accessible and enriched user experience across various domains.

## I. INTRODUCTION

Visual Semantic Extraction for Textual Description is a fascinating field at the intersection of computer vision and natural language processing (NLP), aiming to generate textual descriptions for images automatically. This documentation provides an overview of using Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) for image captioning, outlining their architectures and how they work together to achieve this task. There are multiple use cases and applications of this model like accessibility, social media content sharing, human-computer interaction and robotics. Labelling the satellite picture with atmospheric conditions and various captions of land cover or land use is challenging. The results of used algorithms will enable the worldwide community for a better understanding of what, how, and why deforestation is happening everywhere over the globe - and the ultimate way to reply. Furthermore, existing methods generally can‟t differentiate between main causes of forest loss and natural ones. Higher resolution imagery has already been shown to be exceptionally good at this, but robust methods haven't yet been developed for Planet imagery. To overcome this problem our aim is developing a combination of CNN and RNN algorithm encoder decoder architecture to caption these satellite images. The data images were carried out from Earth‟s full frame analytic scene products using 4 class satellites in sun synchronous orbit and International artificial satellite orbit. Each contains a few bands of information: green, red, blue, infrared and therefore the set of chips for this project uses an actual pattern. The precise spectral responses of the satellites used for images are found within the Planet documentation. Each of those channels is in a 16-bit digital number format that meets the specification of the world. An inventory of training file names and their labels, the labels are space-delimited. High resolution

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

of images have already shown the proof of exceptionally better performance at this, but the robust methodologies haven't yet been developed for earth imagery. Overcoming this problem our aim is developing a combination of algorithm, encoder decoder architecture to caption these satellite images. Specifically, we trained deep convolutional neural networks (CNNs) to find out image features and used multiple classification frameworks including long short-term memory (LSTM) label captioning and binary cross entropy to predict multi-class, multi-label images. The purpose Visual semantic extraction for textual description is to bridge the gap between visual information and textual understanding, making images more accessible, interpretable, and useful in a wide range of applications. Image caption generators leverage the capabilities of modern machine learning models to automatically generate descriptive text, contributing to a more comprehensive understanding of visual content. Visual semantic extraction for textual description involves developing a system that can automatically generate descriptive captions for images. This area is at the intersection of computer vision and natural language processing, and it has several applications and potential scopes. The objectives for an visual semantic extraction for textual description consists of Generate Descriptive Captions which enable to develop a system capable of automatically generating descriptive and accurate captions for a wide variety of images. It can also improve accessibility: Enhance accessibility for visually impaired individuals by providing textual descriptions of images in digital content. Enhance Content which improves content understanding and user engagement by supplementing images with informative textual descriptions. Facilitate Image Retrieval: Enable efficient image retrieval by associating images with relevant textual descriptions, enhancing search capabilities in image-heavy databases or applications. Support Multimodal Applications: Enable integration with multimodal applications by providing a natural language interface for interacting with images. Explore Advanced Techniques for Investigating and implementing advanced techniques such as attention mechanisms, reinforcement learning, or multimodal fusion to improve captioning performance.

## II. EARLIER WORK

### 2.1 Bag of Words

The "bag of words" approach refers to a simple but effective method for representing textual descriptions of images. Instead of considering the sequence or order of words in a caption, the bag of words approach treats each word as an independent entity and counts its occurrence in the entire caption. This creates a "bag" of words, disregarding their order or relationship to each other. Similarly, in image captioning, the bag of words approach disregards the spatial arrangement or relationship of objects within the image. Instead, it focuses solely on identifying objects or concepts present in the image and counting their occurrences. This can be achieved through techniques like object detection or feature extraction, which generate a list of objects or features present in the image. Once the objects or features are identified, the bag of words approach constructs a representation of the image caption by counting the occurrences of each object or feature in the caption. This representation can then be used as input to train machine learning models for generating captions based on the visual content of the image. While the bag of words approach may lack the ability to capture the sequential nature of language or the spatial relationships between objects in the image, it provides a simple and effective way to represent textual descriptions in image captioning tasks, particularly when paired with other techniques for capturing sequential or spatial information.

### 2.2 Nearest Neighbour

The nearest neighbour approach refers to a method of generating captions for an image based on the similarity of other images in a dataset. This approach relies on the assumption that images with similar visual content tend to have similar captions. To implement nearest neighbour in image captioning, a dataset of images paired with their corresponding captions is typically used. When a new image is presented for captioning, its visual features are extracted, often using techniques like convolutional neural networks (CNNs) to represent the image in a high-dimensional feature space. Next, the similarity between the features of the new image and those of images in the dataset is computed. This similarity can be measured using various distance metrics, such as Euclidean distance or cosine similarity. The nearest neighbour approach selects the caption associated with the image that is most similar to the new image based on these computed similarities. The caption from the nearest neighbour image is then used as the generated caption for the new image. While the nearest neighbour can

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

provide captions for images that are visually similar to those in the dataset, it may not always capture the nuances or context-specific information present in the new image. Additionally, its performance heavily relies on the quality and diversity of the dataset used for training.

## III. PROPOSED DEEP LEARNING MODEL

The proposed system uses two models, Convolutional Neural Network(CNN) and Long Short Term Memory(LSTM). Image captioning using a combination of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks involves a multi-step process. Initially, the input image undergoes preprocessing to ensure it meets the requirements of the subsequent CNN model. This typically includes resizing and normalising pixel values. Following preprocessing, a pre-trained CNN model is employed to extract pertinent features from the image. These features capture various visual information embedded within the image and are represented as feature vectors. Once the visual features are extracted, they are passed to an LSTM network for sequential language generation. The LSTM model is initialised with a special "start" token, signalling the beginning of the caption generation process. Subsequently, the LSTM generates words one at a time, based on the input visual features and the previously generated words. At each time step, the LSTM predicts the next word in the sequence, utilising its internal state and the information provided by the preceding words. During the training phase, the model is trained on a dataset consisting of pairs of images and their corresponding ground truth captions. The objective is to minimise the disparity between the generated captions and the ground truth captions. This is achieved by optimising the parameters of the LSTM network using techniques such as backpropagation through time (BPTT) and gradient descent. Once trained, the model can be evaluated on a separate validation or test dataset to assess its performance. Evaluation metrics such as BLEU, METEOR, or CIDEr are commonly used to compare the generated captions against human-annotated captions. By combining the capabilities of CNNs for visual feature extraction and LSTMs for sequential language generation, this architecture enables the generation of descriptive captions that accurately depict the content of input images.

## IV. METHODOLOGY

### 4.1 Convolutional Neural Network

Convolutional Neural Network (CNN) is a Deep Learning algorithm which takes in an input image and assigns importance (learnable weights and biases) to various aspects/objects in the image, which helps it differentiate one image from the other. One of the most popular applications of this architecture is image classification. The neural network consists of several convolutional layers mixed with nonlinear and pooling layers. When the image is passed through one convolution layer, the output of the first layer becomes the input for the second layer. This process continues for all subsequent layers. After a series of convolutional, nonlinear and pooling layers, it is necessary to attach a fully connected layer. This layer takes the output information from convolutional networks. Attaching a fully connected layer to the end of the network results in an N dimensional vector, where N is the number of classes from which the model selects the desired class.

### 4.2 Long Short Term Memory

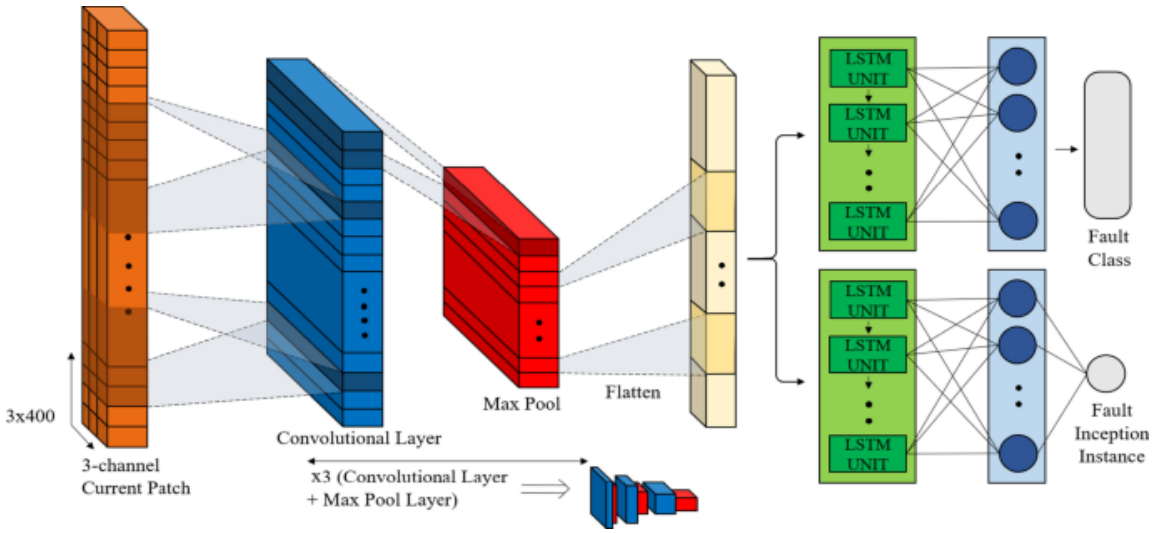
Long Short-Term Memory (LSTM) networks are a type of Recurrent Neural Network (RNN) capable of learning order dependence in sequence prediction problems. This is most used in complex problems like Machine Translation, Speech Recognition, and many more. The reason behind developing LSTM was, when we go deeper into a neural network if the gradients are very small or zero, then little to no training can take place, leading to poor predictive performance and this problem was encountered when training traditional RNNs. LSTM networks are well- suited for classifying, processing, and making predictions based on time series data since there can be lags of unknown duration between important events in a time series. LSTM is way more effective and better compared to the traditional RNN as it overcomes the short term memory limitations of the RNN. LSTM can carry out relevant information throughout the processing of inputs and discards non-relevant information with a forget gate.

### 4.3 CNN and LSTM Architecture

Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks are commonly used together to generate textual descriptions of images. CNNs are employed to extract features from the input

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

image. These networks are well-suited for image processing tasks due to their ability to automatically learn hierarchical representations of visual features. In the context of image captioning, a CNN is typically pretrained on a large dataset for image classification or object detection tasks. The learned convolutional layers capture low-level features such as edges and textures, while deeper layers capture more abstract and high-level features related to objects, shapes, and scenes within the image. Once the features are extracted using the CNN, they are passed to an LSTM network for sequential language generation. LSTM networks are a type of recurrent neural network (RNN) designed to capture dependencies and long-range dependencies in sequential data. In image captioning, the LSTM network takes the extracted visual features from the CNN as input and generates a sequence of words to form the caption. During training, the LSTM network learns to predict the next word in the sequence based on the input features and the previously generated words. This process continues until an end-of-sentence token is generated or a maximum sequence length is reached. The parameters of both the CNN and LSTM are optimised jointly through backpropagation and gradient descent to minimise the discrepancy between the generated captions and the ground truth captions in the training dataset. By combining CNNs for visual feature extraction and LSTMs for sequential language generation, the architecture effectively bridges the semantic gap between visual content and natural language, enabling the model to generate descriptive captions that accurately describe the content of the input image.




### 4.3.1 Image Preprocessing

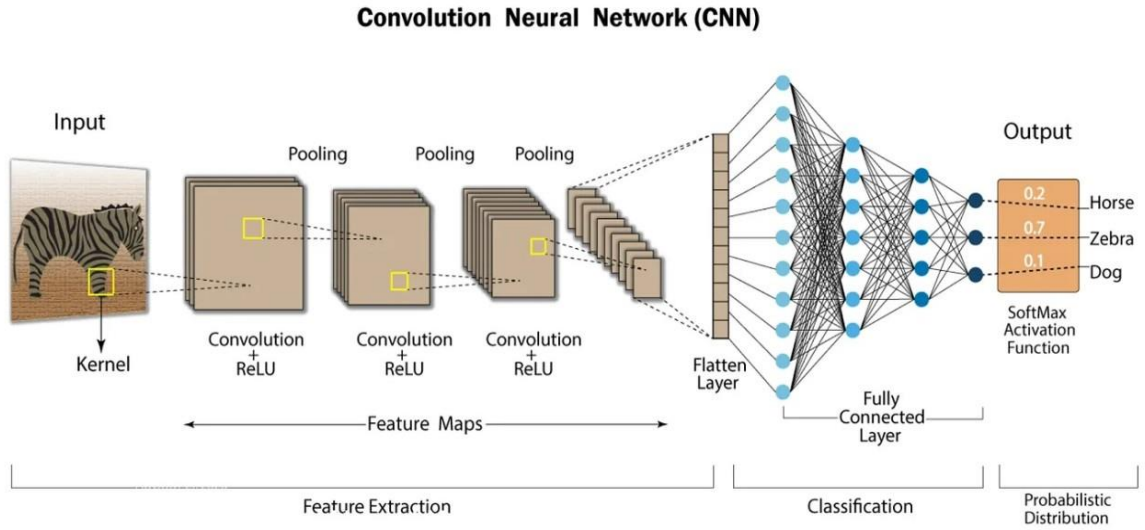
In image captioning, preprocessing plays a crucial role in preparing both the images and their associated captions for effective analysis by the captioning model. Initially, the images are resized to a standardised dimension. This step ensures uniformity in image size, facilitating consistent processing across the dataset. Additionally, normalisation techniques are applied to standardise the pixel values of the images. By scaling pixel values to a common range, such as [0, 1], variations in lighting conditions and colour distributions among images are mitigated, ensuring that the model's performance is not adversely affected by differences in image characteristics. Furthermore, data augmentation techniques are often employed to enrich the training dataset and improve the robustness of the captioning model. Augmentation involves applying transformations such as rotation, flipping, or adding noise to the images, thereby increasing the diversity of the dataset. This augmentation helps the model generalise better to unseen images by exposing it to a wider range of variations during training. In parallel, textual information in the form of image captions undergoes preprocessing. This involves tokenization, where each caption is broken down into individual words or tokens. Additionally, a vocabulary is created, mapping each unique word to an index. This vocabulary is essential for converting the textual input into numerical representations that can be understood by the captioning model. Tokenization and vocabulary creation ensure that the model can effectively process and generate captions based on textual input. Moreover, in the feature extraction phase, a pre-trained Convolutional Neural Network (CNN) is utilised to extract high-level visual features from the preprocessed images. By removing the fully connected layers of the CNN and utilising the output from one of the convolutional or pooling layers, the model captures meaningful

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

visual information. These extracted features serve as input to the captioning model, enabling it to generate descriptive captions based on the visual content of the images.

#### 4.3.2 Feature Extraction with CNN

Feature extraction with Convolutional Neural Networks (CNNs) in the context of image captioning involves several steps. Firstly, CNNs are utilised to extract meaningful features from the input image. These features capture various visual aspects such as edges, textures, and object shapes. These features are crucial for understanding the content of the image. Once the CNN processes the image, it creates a representation of the image in the form of feature maps. These feature maps contain information about different aspects of the image at various spatial locations. The deeper layers of the CNN typically capture higher-level features, such as object parts and their spatial arrangements. In image captioning, these extracted features serve as the foundation for generating descriptive captions. After the CNN extracts features from the image, these features are fed into a decoder network, often a Recurrent Neural Network (RNN) or Transformer architecture. This decoder generates a sequence of words that forms the caption. The decoder network is conditioned on the extracted features, ensuring that the generated captions are semantically relevant to the content of the image. By leveraging both visual information from the CNN and linguistic context from the decoder, the model can produce coherent and accurate captions for a wide range of images. Training such a system involves optimising both the CNN and the decoder jointly. This training process aims to minimise the discrepancy between the predicted captions and the ground truth captions in a large dataset of image-caption pairs. Through this optimization process, the model learns to associate visual features with corresponding textual descriptions, enabling it to generate meaningful captions for unseen images.




#### 4.3.3 Sequence Initialization

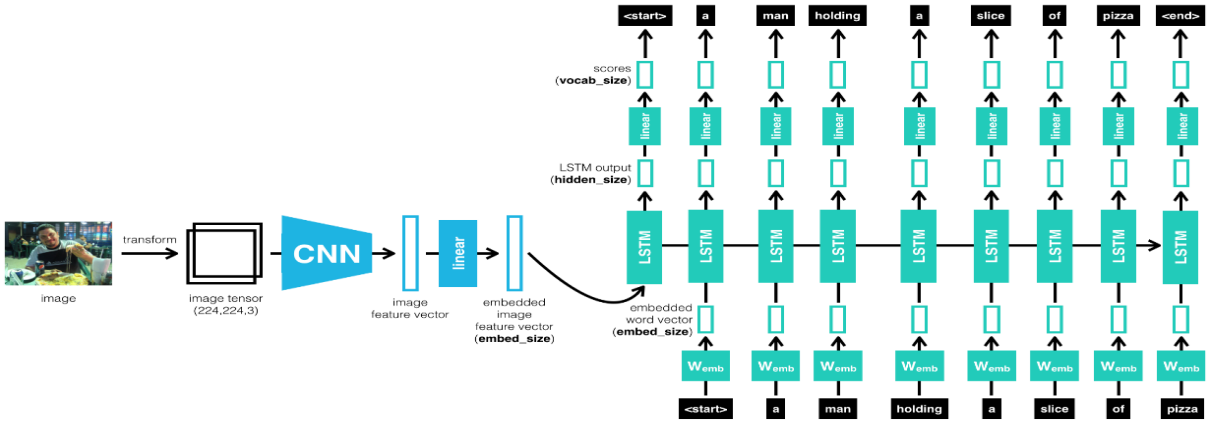
Sequence initialization plays a crucial role in various tasks such as natural language processing (NLP) and sequence generation. In the context of tasks like language modelling, machine translation, or image captioning, sequence initialization refers to the process of initialising the hidden state of the recurrent neural network (RNN) or the transformer model before generating the sequence. Before generating the first token or word of the sequence, the model needs to have an initial state that provides context for the generation process. This initial state serves as a starting point for the model to begin generating the sequence. In RNNs, this initial state often consists of a vector representation that captures some contextual information about the input sequence or the task at hand. For example, in language modelling, the initial state of the RNN might be initialised using a special token or a zero vector, indicating the beginning of a new sequence. In machine translation, the initial state might be initialised based on an encoding of the source sentence, providing context for generating the target translation. In transformer models, sequence initialization involves a similar concept but is often more sophisticated. Transformers use positional encodings to provide information about the position of tokens in the sequence. These positional encodings are added to the token embeddings before feeding them into the

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

transformer layers. Additionally, transformers may also use special tokens such as <bos> (beginning of sequence) to initialise the generation process. Regardless of the specific method used for sequence initialization, the goal is to provide the model with sufficient context to start generating the sequence. This context helps guide the generation process, ensuring that the model produces coherent and relevant output. Through training, the model learns to use this initial context effectively to generate sequences that are appropriate for the given task.

### 4.3.4 Language Generation using LSTM

Language generation with Long Short-Term Memory (LSTM) networks is a fundamental task in natural language processing (NLP). LSTMs are a type of recurrent neural network (RNN) specifically designed to capture long-range dependencies in sequential data, making them well-suited for tasks like language modelling and text generation. At the core of language generation with LSTMs lies the ability of these networks to learn patterns and structures from sequences of words. The process typically begins with tokenizing the input text into a sequence of discrete tokens (words or subwords). Each token is then represented as a vector through techniques like word embeddings, which capture semantic relationships between words. During training, the LSTM model learns to predict the next word in a sequence given the previous words. This is achieved by feeding the input sequence word by word into the LSTM network, and at each step, the LSTM updates its hidden state based on the current input word and the previous hidden state. The hidden state serves as a memory that retains information about the context of the sequence seen so far. Once the LSTM processes the entire input sequence, it generates an output sequence by predicting the next word at each step based on the current input word and the hidden state. This process continues recursively until a predefined stopping criterion is met, such as reaching a maximum sequence length or generating an end-of-sentence token. During generation, the LSTM's hidden state serves as a context vector that captures the information learned from the input sequence. This context vector guides the generation process, influencing the choice of words at each step based on the learned patterns and structures in the training data. Language generation with LSTMs involves training the network to minimise the discrepancy between the predicted sequence and the ground truth sequence using techniques like teacher forcing or reinforcement learning. Through this training process, the LSTM learns to generate coherent and contextually relevant sequences of text, capturing the stylistic and semantic characteristics of the training data.




## V. TRAINING

Training a neural network for image captioning is a complex process that involves teaching the model to generate descriptive captions for images. Initially, the training data, comprising pairs of images and corresponding captions, is prepared. Each image is passed through a Convolutional Neural Network (CNN) to extract meaningful visual features, capturing important aspects like objects, shapes, and textures. Concurrently, the captions are tokenized into word sequences. The process then transitions into caption generation, where these visual features from the CNN, along with an initial state typically initialised with special tokens, are combined to initiate caption generation. Here, a recurrent neural network (RNN), such as an LSTM or Transformer, is often used. The model predicts the next word in the caption sequence based on the context provided by the image features and the previously generated words. Throughout this process, at each step of

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

caption generation, the model's output is compared against the actual words in the caption using a loss function. This quantifies the disparity between the predicted and true words, allowing the network to learn from its mistakes. Gradients of the loss function with respect to the model's parameters are computed through backpropagation, indicating how parameters should be adjusted to minimise the loss. Optimization algorithms like stochastic gradient descent (SGD) or Adam are then employed to update the model's parameters iteratively, reducing the loss and improving caption generation accuracy. This process continues over multiple epochs, with the model gradually refining its understanding of image-caption relationships. Validation is crucial throughout training, where the model's performance is evaluated on a separate validation set to ensure it generalises well to unseen data. Adjustments to hyperparameters may be made based on validation results to optimise performance further.

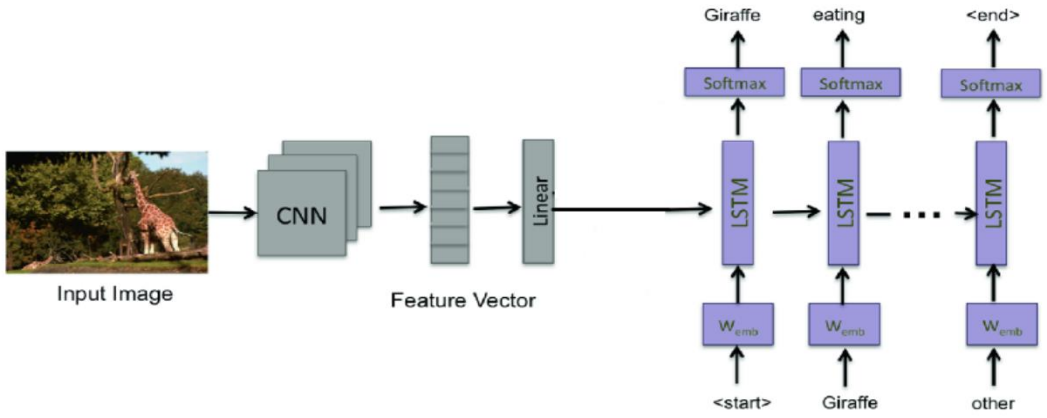
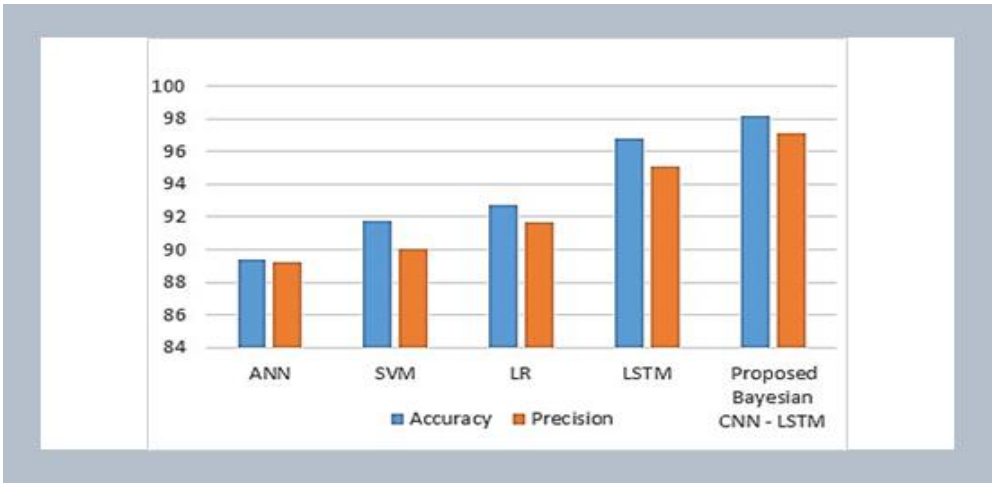



Figure 2: Baseline encoder-decoder architecture using a pretrained ResNet-50 convolutional neural network encoder and long short term memory model decoder

## VI. EVALUATION

In the evaluation phase of an image captioning model using Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, several crucial steps are taken to assess the model's performance and the quality of generated captions: Initially, the trained model is evaluated on a separate validation or test dataset, distinct from the data used for training. This dataset comprises images and their corresponding ground truth captions, serving as a benchmark to measure the model's captioning accuracy. During evaluation, the model generates captions for each image in the validation or test dataset based on its visual features extracted by the CNN and the language model learned by the LSTM. These generated captions are then compared to the ground truth captions provided in the dataset. Evaluation metrics such as BLEU (Bilingual Evaluation Understudy), METEOR (Metric for Evaluation of Translation with Explicit Ordering), or CIDEr (Consensus-based Image Description Evaluation) are commonly employed to assess the quality of the generated captions. These metrics quantify the similarity between the generated and ground truth captions, considering factors like precision, recall, and semantic relevance. BLEU measures the overlap between n-grams (contiguous sequences of n words) in the generated and reference captions, rewarding precision and brevity. METEOR incorporates more complex linguistic and semantic factors, including synonyms and paraphrases, to compute a similarity score between the generated and reference captions. CIDEr evaluates the consensus between the generated captions and multiple reference captions, accounting for variations in human annotations. It considers not only individual word matches but also semantic similarity and diversity in language usage. Additionally, qualitative evaluation methods such as human judgement or user studies may be employed to assess the perceptual quality and contextual relevance of the generated captions. Human evaluators may rate the generated captions based on criteria such as fluency, relevance, and coherence compared to the ground truth captions. The evaluation results provide insights into the model's strengths and weaknesses, helping researchers fine-tune hyperparameters, optimise architectures, and improve training strategies to enhance captioning performance. Ultimately, the goal of evaluation is to ensure that the image captioning model produces accurate, relevant, and contextually coherent captions that effectively describe the visual content of images.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com




## VII. CONCLUSION

In conclusion, the CNN-LSTM based image captioning model represents a significant advancement in bridging the semantic gap between visual content and natural language. By leveraging Convolutional Neural Networks (CNNs) for feature extraction and Long Short-Term Memory (LSTM) networks for sequential language generation, the model demonstrates promising capabilities in generating descriptive and contextually relevant captions for images. Through extensive training and evaluation, the model showcases its ability to learn from large datasets and produce captions that accurately describe the visual content depicted in the images. However, despite its successes, there are several avenues for further research and improvement. Firstly, enhancing the model's ability to capture finer details and nuances in both visual and textual information could lead to more accurate and nuanced captions. This may involve exploring more sophisticated feature extraction techniques, such as attention mechanisms, to focus on relevant image regions or incorporating semantic information into the captioning process. Moreover, improving the diversity and creativity of generated captions remains a challenge. While the model can generate accurate descriptions based on training data, it may struggle with generating diverse or imaginative captions for novel images. Addressing this issue could involve incorporating techniques from natural language generation and creativity research to encourage the model to produce more varied and imaginative outputs. Furthermore, extending the model's applicability to multimodal contexts, such as video captioning or generating captions for multimodal inputs (e.g., images paired with audio descriptions), could broaden its utility and impact. Integrating additional modalities and considering temporal information could enrich the captioning process and enable more comprehensive descriptions of complex visual scenes. Additionally, ensuring the model's robustness to diverse datasets, environments, and user scenarios is essential for real-world deployment. This may involve further validation and testing across a wide range of datasets and application domains to assess the model's generalisation capabilities and robustness to domain shifts. In summary, while the CNN-LSTM based image captioning model represents a significant step forward in automated image understanding and natural language generation, continued research and development are necessary to address existing challenges and unlock its full potential in diverse real-world applications. By advancing the model's capabilities in accuracy, diversity, robustness, and multimodal integration, we can further enhance its utility and impact across various domains, ultimately enriching human-computer interaction, accessibility, and multimedia content understanding.

## VIII. REFERENCES

[1] Farhadi A, Hejrati M, Sadeghi MA, Young P, Rashtchian C, Hockenmaier J, Forsyth D. Every picture tells a story: generating sentences from images. In: European conference on computer vision. Berlin: Springer; 2010. p. 15–29.

[2] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.

[3] Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation.

https://arxiv.org/abs/1406.1078. Accessed 3 Jun 2014.

[4] Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel R, Bengio Y. Show, attend and tell: neural image caption generation with visual attention. In: International conference on machine learning. New York: PMLR; 2015. p. 2048–57.

[5] Katiyar S, Borgohain SK. Image captioning using deep stacked LSTMs, contextual word embeddings and data augmentation. https://arxiv.org/abs/2102.11237. Accessed 22 Feb 2021.

[6] Redmon J, Farhadi A. Yolov3: an incremental improvement. https://arxiv.org/abs/1804.02767. Accessed 8 Apr 2018.

[7] Bochkovsky A, Wang CY, Liao HY. Yolov4: optimal speed and accuracy of object detection. https://arxiv.org/abs/2004.10934. Accessed 23 Apr 2020.

[8] Redmon J, Farhadi A. YOLO9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway: IEEE; 2017. p. 7263–71.

[9] Yin X, Ordonez V. Obj2text: generating visually descriptive language from object layouts. https://arxiv.org/abs/1707.07102. Accessed 22 Jul 2017.

[10] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. https://arxiv.org/abs/1409.1556. Accessed 4 Sep 2014.

[11] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Piscataway: IEEE; 2009. p. 248–55.

[12] Vo-Ho VK, Luong QA, Nguyen DT, Tran MK, Tran MT. A smart system for text-lifelog generation from wearable cameras in a smart environment using concept-augmented image captioning with modified beam search strategy. Appl Sci. 2019;9(9):1886.

[13] Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst. 2015;28:91–9.

[14] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway: IEEE; 2016. p. 770–8.

[15] Lanzendörfer L, Marcon S, der Maur LA, Pendulum T. YOLO-ing the visual question answering baseline. Austin: The University of Texas at Austin; 2018.

[16] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway: IEEE; 2016. p. 2818–26.

[17] Herdade S, Kappeler A, Boakye K, Soares J. Image captioning: transforming objects into words. https://arxiv.org/abs/1906.05963. Accessed 14 Jun 2019.

[18] Wang J, Madhyastha P, Specia L. Object counts! bringing explicit detections back into image captioning. https://arxiv.org/abs/1805.00314. Accessed 23 Apr 2018.

[19] Sharif N, Jalwana MA, Bennamoun M, Liu W, Shah SA. Leveraging Linguistically-aware object relations and NASNet for image captioning. In: 2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ). Piscataway: IEEE; 2020. p. 1–6.

[20] Variš D, Sudoh K, Nakamura S. Image captioning with visual object representations grounded in the textual modality. https://arxiv.org/abs/2010.09413. Accessed 19 Oct 2020.