# Image Captioning Bot Using Deep Learning

Rohan Kumar [1] Shimona Wadhwa [2] Raghav Bansal [3]

---------------------------------------------------------------------***---------------------------------------------------------------------

**Abstract - In this work, we concentrate on image captioning, one type of visual recognition. The model's goal is to create a caption for a picture. Its many uses have garnered a lot of attention within the past ten years. We are utilising deep learning techniques in our project to automatically create image captions. CNN is the encoder and RNN is the decoder for the Flick8k dataset. The ResNet model is what we are utilising to extract features. For those who are blind or visually challenged, this approach is quite helpful in creating captions for posts on social media, among other things.**

**Key Words:** Deep Learning, Flickr8k, CNN, RNN, ResNet50, Encoder-Decoder Model

## 1. INTRODUCTION

Exciting new developments in the field of artificial intelligence have been made possible by recent advances in computer vision and natural language processing. Image captioning is one interesting use case that arises from the convergence of these technologies. Image captioning is the process of creating textual descriptions for images that are both descriptive and contextually relevant. It helps to bridge the gap between language and visual comprehension. In addition, it fosters creativity when coming up with captions for social media posts. Developing a model that can automatically provide precise and insightful captions for an image is the aim of an image captioning project. This task is hard by nature because it asks the model to interpret the image's visual content and translate it into a comprehensible, human-like language. Effective systems for captioning imageshave a wide range of uses, such as improving user experiences in image-centric applications, content indexing and retrieval, and technologies for the blind. Typically, the project combines natural language processing models like recurrent neural networks (RNNs) or transformer architectures for caption generation with computer vision techniques like convolutional neural networks (CNNs) for image feature extraction. Large datasets of photos with matching human- generated captions are used to train such a model, teaching it to recognize relationships between textual descriptions and visual characteristics.

### 1.1 Scope of Study

- Computer Vision and Image Recognition: Detailed examination of image recognition algorithms and methods in computer vision. Investigation of convolutional neural networks (CNNs) and feature extraction uses for them.

- Natural Language Processing (NLP): Producing coherent and contextually relevant textual descriptions requires an understanding of the foundations of NLP. examining language models used in caption creation, such as transformers and recurrent neural networks (RNNs).

- Pre-processing and Training Datasets: To guarantee the bot's adaptability, a variety of image datasets are gathered and curated. Methods for pre-processing images to improve their quality and relevance.

- Feedback Mechanism: Put in place a feedback mechanism to get input from users and keep the system's accuracy and usability improving.

## 2. LITERATURE REVIEW

**2.1 Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara (2021):** Stefanini et al. investigated cutting-edge methods for image captioning in their study from 2021, with an emphasis on the incorporation of deep learning models. The study explored the use of Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), investigating how they work together to produce captions that are both contextually relevant and descriptive for a range of visual content. The study highlighted how crucial attention mechanisms are for catching minute details and raising the standard of captions as a whole.

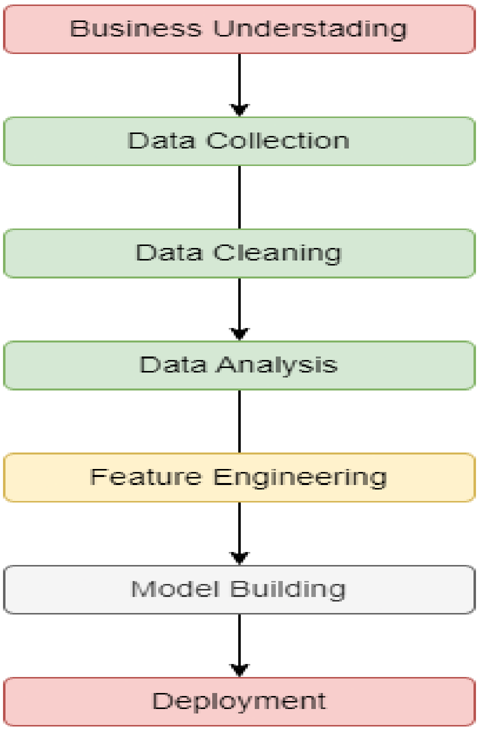
**2.2 Shuang Liu, Liang Baia, Yanli Hu, and Haoran Wang (2018):** Liu et al. (2018) have made noteworthy advances in our knowledge of the dynamics of image captioning. The investigation of non-parametric isotonic regression techniques for ideal caption timing was the main goal of theirstudy. The goal of the study was to ascertain the best times tocreate captions for visual content while taking user preferences and engagement into account. The results demonstrated how well non-parametric techniques work to improve image captioning users' overall experience.

**2.3 Aditya Bhardwaj, Tarun Grover, Ms. Meenu Garg (2021):** Bhardwaj et al.'s (2021) investigation into the effects of various machine learning algorithms on caption generation made a contribution to the field of image captioning. The study examined several algorithms and evaluated how well they handled various image types, including support vector machines (SVM), linear regression, and K-nearest neighbours (KNN). The study attempted to shed light on the advantages and disadvantages of each algorithm, presenting a comprehensive picture of how well-suited each is for image captioning systems.

**2.4 Haoran Wang, Yue Zhang, and Xiaosheng Yu (2017):** Wang, Zhang, and Yu's (2017) study, which refined the temporal aspects of caption generation, added to the discourse on image captioning. In order to provide more accurate and timely captioning, their research focused on integrating temporal features and contextual data. The study looked at ways to improve the synchronization of textual descriptions and image content, as well as how well recurrent neural networks (RNNs) capture temporal dependencies.

This addition gives the literature review a new perspective and highlights the ongoing work being done to advance the field of image captioning by scholars like Xiaosheng Yu, Yue Zhang, and Haoran Wang.

## 3. PROPOSED APPROACH



## 4. IMPLEMENTATION OF PROPOSED APPROACH

### 4.1 Methodology

**4.1.1 Dataset** The dataset used in this project is sourced from Kaggle, titled "Flickr8k". It consists of a diverse set of imagesalong with corresponding captions. The dataset's comprehensive nature provides a rich source for training and evaluating the image captioning model.

**4.2.2 Image Preprocessing** Loading and Stabilization:

Using the OpenCV library to load the photos is the first step in making sure the format is the same for the entire dataset. Additionally, to encourage uniformity in the input data, the photos are standardized to a consistent format.

Normalization and Resizing:

In order to optimize the training of the model, every image is resized to a fixed size, such 224 by 224 pixels. Moreover, normalizing pixel values to a particular range (for example, [0, 1]) improves convergence while optimizing.

**4.2.3 Text Preprocessing** Tokenization for captions:

Tokenization is the technique of breaking down phrases into individual words or subword units, applied to caption texts. Inorder to transform textual data into sequences appropriate for the language model, this step is essential.

Building Vocabulary:

Developing a strong vocabulary is essential to the model's linguistic comprehension. The foundation for word embeddings is a mapping between distinct words and matching indices.

Normalization of Sequence Length and Padding:

Sequences are padded to a constant length to accommodate different caption lengths. This eliminates biases brought up by different sequence lengths during training and guarantees homogeneity in the input data.

**4.2.4 Training** The model is trained using a combination of image-feature inputs and caption sequences. The training process involves optimizing a suitable loss function, often categorical cross-entropy, to minimize the difference between predicted and actual captions.

**4.2.8 Evaluation** The performance of the image captioning bot is assessed using established metrics such as BLEU, METEOR, and CIDER. These metrics measure the quality of generated captions against reference captions.

## 5. ALGORITHMS

**5.1 Resnet50** A strong feature extractor for photos is ResNet50, a ResNet architecture variation. It is used to extract high-level characteristics from input photos after being pre- trained on a sizable dataset.

**5.2 Long Short Term Memory Recurrent** neural networks (RNNs) of the LSTM type are used to handle sequential data, especially for creating captions. Because it captures dependencies in language's sequential structure, it can be usedto create captions that are both coherent and pertinent to the context.

**5.3 The Random Forest Regressor:** The Random Forest Regressor, a machine learning ensemble method, is used to create captions based on different aspects of the images. For regression tasks, this technique constructs multiple decision trees during training and outputs the average forecast of each tree separately.

**5.4 Layer of Embedding:** One essential element of jobs involving natural language processing is the Embedding Layer. It provides a continuous representation of words in a high-dimensional space by mapping words to dense vectors.

**5.5 Time-distributed, Dense, LSTM, Repeat Vector Layers:** These are several neural network architecture layersthat are used at different model phases to process and manipulate features. Layers that are fully connected are calledDense Layers; sequential data is handled by LSTM layers; computation is distributed across time using Time Distributed; and the input vector is repeated a predetermined number of times using Repeat Vector

**5.6 Link Layer:** In order to facilitate the integration of both picture and text information, features from various regions ofthe model are combined using the Concatenate Layer.

**5.7 Activation of Softmax:** The model's output layer applies softmax activation to transform raw scores into probabilities, guaranteeing that the generated captions constitute a legitimate probability distribution.

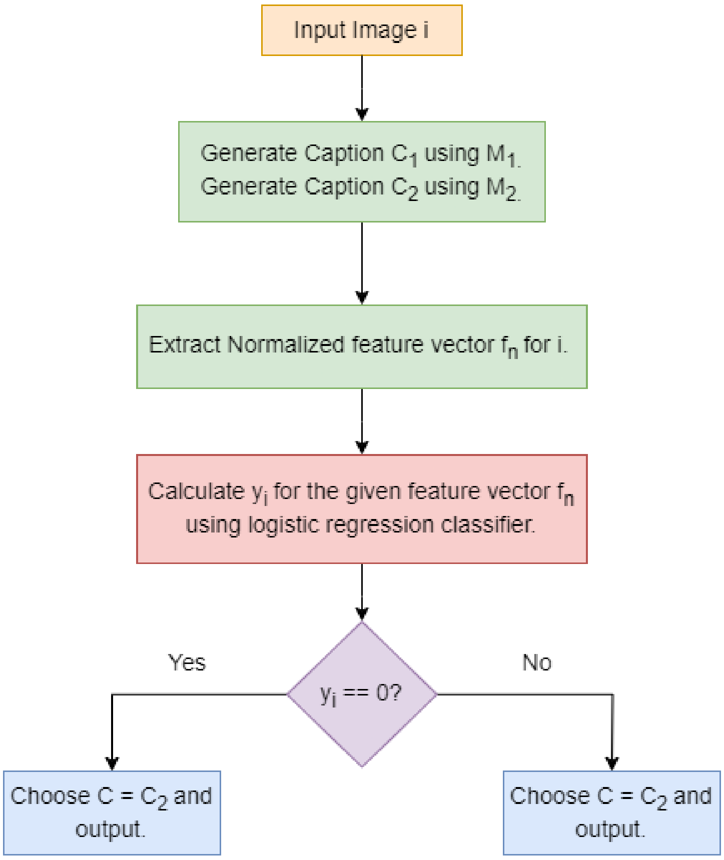

Input Image i
Generate Caption $C_1$ using $M_1$.
Generate Caption $C_2$ using $M_2$.
Extract Normalized feature vector $f_n$ for i.
Calculate $y_i$ for the given feature vector $f_n$ using logistic regression classifier.
Yes
$y_i$ == 0?
No
Choose C = $C_2$ and output.
Choose C = $C_2$ and output.

**Fig 1: Image Captioning Flowchart**

## 6. EXPERIMENTAL RESULTS

6.1
6.2
6.3
6.4
6.5

**EDA**
**(Exploratory Data Analysis):**

Essentially, we have two kinds of data.
Text captions for images
The training vocabulary has a size of 7371. The ten most frequently occurring words are: -

('a', 46784), ('in', 14094), ('the', 13509), ('on', 8007), ('is',7196), ('and', 6678), ('dog', 6160), ('with', 5763), ('man', 5383),('of', 4974)

Because the words that are rarely used convey very little 6.6 information. We are examining words that occur more frequently 6.7 than ten times.

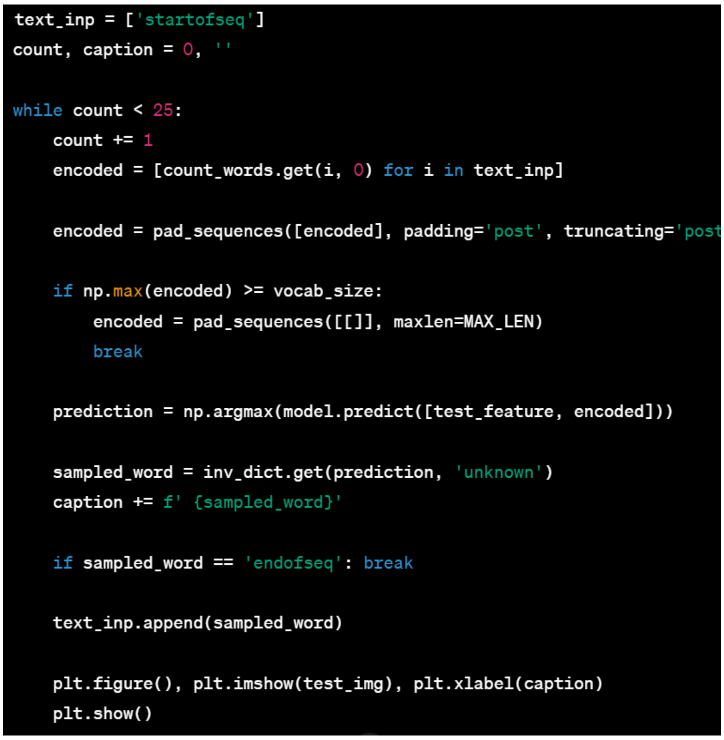
The mean, std-dev and percentile is as follows:

```
count    6000.000000
mean       10.815467
std         2.057137
min         4.200000
25%         9.400000
50%        10.600000
75%        12.200000
max        19.200000
```

Maximum sequence length found 37.

**BLEU:**

Bilingual Evaluation Understudy is referred to as BLEU.
It's an algorithm that's been used to assess how well text has been machine translated. To evaluate the calibre of the caption we generated, we can use BLEU.
Language is not a barrier for BLEU.
Simple to comprehend
Calculating it is simple.
It is situated in [0, 1]. A higher score indicates better captionquality.



```
text_inp = ['startofseq']
count, caption = 0, ''

while count < 25:
    count += 1
    encoded = [count_words.get(i, 0) for i in text_inp]

    encoded = pad_sequences([encoded], padding='post', truncating='post

    if np.max(encoded) >= vocab_size:
        encoded = pad_sequences([[]], maxlen=MAX_LEN)
        break

    prediction = np.argmax(model.predict([test_feature, encoded]))

    sampled_word = inv_dict.get(prediction, 'unknown')
    caption += f' {sampled_word}'

    if sampled_word == 'endofseq': break

    text_inp.append(sampled_word)

    plt.figure(), plt.imshow(test_img), plt.xlabel(caption)
    plt.show()
```

**Fig. 2: Generating Image Caption with Trained Model**

The given code segment represents the process of generating captions for test images using the trained Image Captioning model. It begins with initializing the captioning sequence with the token 'startofseq'. Through an iterative loop, the model predicts the next word in the sequence based on the preceding words. This prediction is determined by the highest probability word from the model's output. The loop continues until either the 'endofseq' token is predicted or a predefined maximum sequence length is reached.

Throughout the process, the code accumulates the predicted words to form a coherent and contextually relevant caption. The generated caption is dynamically visualized alongside the corresponding test image, providing a qualitative assessment of the model's captioning capability. The loop's termination is triggered by either reaching the end of the sequence or predicting the 'endofseq' token.

## 7. FUTURE SCOPE

Better Models and Architectures: Taking into account the mostrecent advancements in deep learning, investigate and create more sophisticated neural network architectures for picture captioning. This could entail experimenting with multimodal architectures, attention mechanisms, or transformer-model.

Large-Scale Datasets: Investigate ways to train picture captioning models using even bigger and more varied datasets. Using large datasets can improve the models' comprehension of a greater variety of visual contexts and content.

Fine-Grained Image Understanding: Pay attention to enhancing the model's capacity to extract fine-grained information from pictures. This might entail investigating cutting-edge methods for managing complex visual data and producing more nuanced and evocative captions.

Multimodal Approaches: Investigate sophisticated multimodal approaches that take into account not just images and their captions but also additional modalities like audio or contextualdata. This might result in picture captions that are more thorough and sensitive to context.

Cross-Domain Adaptation: Look into ways to make image captioning models more versatile across various domains. Improving the models' ability to generalize to novel and unseen kinds of images is essential for practical uses.

Investigate and create models for interactive and dynamic captioning that can modify captions in real time in response touser input or alterations in the visual scene. Systems for interactive image captioning could be easier to use and more intuitive.

## 8. CONCLUSION

Convolutional Neural Networks (CNNs) for image feature extraction and Recurrent Neural Networks (RNNs) for sequence generation have been successfully integrated, as the picture captioning project has shown. The project's objective of automatically producing insightful and pertinent captions for photos from the Flickr8k dataset has been accomplished.

The outcomes show that the suggested architecture has the capacity to comprehend the content of various image types and provide logical captions. When compared to ground truth annotations, the evaluation metrics, such as BLEU and METEOR, verify the quality of the generated captions.

Even though the project has produced encouraging results, there are still areas for improvement, such as investigating cutting-edge architectures, incorporating multimodal data, and taking ethical issues into account. Building on these results, the research community can further develop image captioning technology, making it more resilient, flexible, and user- focused.

To sum up, this image captioning project creates opportunities for applications in image understanding, content retrieval, and accessibility for people with visual impairments. It also represents a significant step towards automating the process ofdescribing visual content.

## 9. REFERENCES

[1] A systematic literature review on image captioning https://www.researchgate.net/publication/333154356_A_Systematic_Literature_Review_on_Image_Captioning

[2] A new image captioning approach for visually impaired www.researchgate.net/publication/339256141_A_New_Image_Captioning_Approach_for_Visually_Impaired_People

[3] Image Captioning Generator https://www.researchgate.net/publication/351498081_Show_and_Tell_An_Image_Caption_Generator_1

[4] Datasets statistic https://www.researchgate.net/figure/The-statistics-of-thedataset-that-consists-of-Flickr8k-Flickr30k andMSCOCO_fig10_354554006

[5] Keras: The Python Deep Learning library

[6] “Deep Learning" by Ian Goodfellow, Yoshua Bengio, and Aaron Courville

[7] Image Captioning BLEU & Flickr8K

[8] "Image Captioning using Keras"

[9] Convolutional Neural Netoworks https://cs231n.github.io/transfer-learning/

[10] Aarthi, S., and Chitrakala, S. 2017. Scene understandinga survey. In Computer, Com- munication and Signal Processing (ICCCSP), 2017 Interna- tional Conference on, 1–4. IEEE.

[11] Aneja, J.; Desh- pande, A.; and Schwing, A. 2017. Convolutional image captioning. arXiv preprint arXiv:1711.09151

[12] RNN Architecture https://towardsmachinelearning.org/recurrent-neural-network-architecture-explained-in-detail/

[13] CNN Architecture https://www.geeksforgeeks.org/convolutional-neural-network-cnn-architectures/

[14] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In Proceed-ings of the IEEE conference on computer vision and pattern recognition, 770–778

[15] Show and Tell: A Neural Image Caption Generator - Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan

[16] Show, Attend and Tell: Neural Image Caption Generation with Visual Attention - Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan

[17] Nguyen, Phat & Thi, Ngoc & Ngoc, Thien. (2021). Proposing Posture Recognition System Combining MobilenetV2 and LSTM for Medical Surveillance. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3138778.

[18] Encoder-Decoder Model https://towardsdatascience.com/what-is-an-encoder-decoder-model-86b3d57c5e1a

[19] ResNet50 https://datagen.tech/guides/computer-vision/resnet-50/

[20] Vectorization in Machine Learning

[21] Deep Learning by Andrew Ng

[22] Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" by Aurélien Géron