

# **MACHINE LEARNING TECHNIQUES LAB**

## **PROJECT REPORT**

**MIRRA. G**

**22011101067**

### **AIM :**

The aim of the project is to design and implement an image captioning system capable of automatically generating descriptive and contextually relevant captions for images.

### **DATASET DESCRIPTION :**

The Flickr8K dataset is a collection of 8,000 high-quality images sourced from the Flickr photo-sharing platform. Each image in the dataset is paired with five descriptive captions, providing different textual descriptions of the visual content. The dataset is commonly split into training, validation, and test sets, facilitating model training, evaluation, and benchmarking. Researchers use this dataset to develop and assess image captioning algorithms, aiming to automatically generate informative and contextually relevant captions for images.

### **TYPE OF MACHINE LEARNING USED :**

The image captioning project primarily uses supervised learning techniques.

In supervised learning:

- **Training Data:** The project relies on a dataset where each image is paired with multiple human-generated captions. These captions serve as the ground truth or target output for the corresponding images during model training.

- ***Model Training:*** Deep learning models, such as convolutional neural networks (CNNs) for image feature extraction and recurrent neural networks (RNNs) for generating captions, are trained using the paired image-caption data. The models learn to map input images to their corresponding captions based on the supervision provided by the labelled training data.
- ***Evaluation:*** The performance of the trained models is evaluated using metrics like BLEU score, which compares the generated captions with the human-written references. This evaluation process requires access to ground truth captions, indicating the supervised nature of the learning task.

## **MACHINE LEARNING TECHNIQUES AND DEEP LEARNING MODELS USED :**

### **Machine learning techniques :**

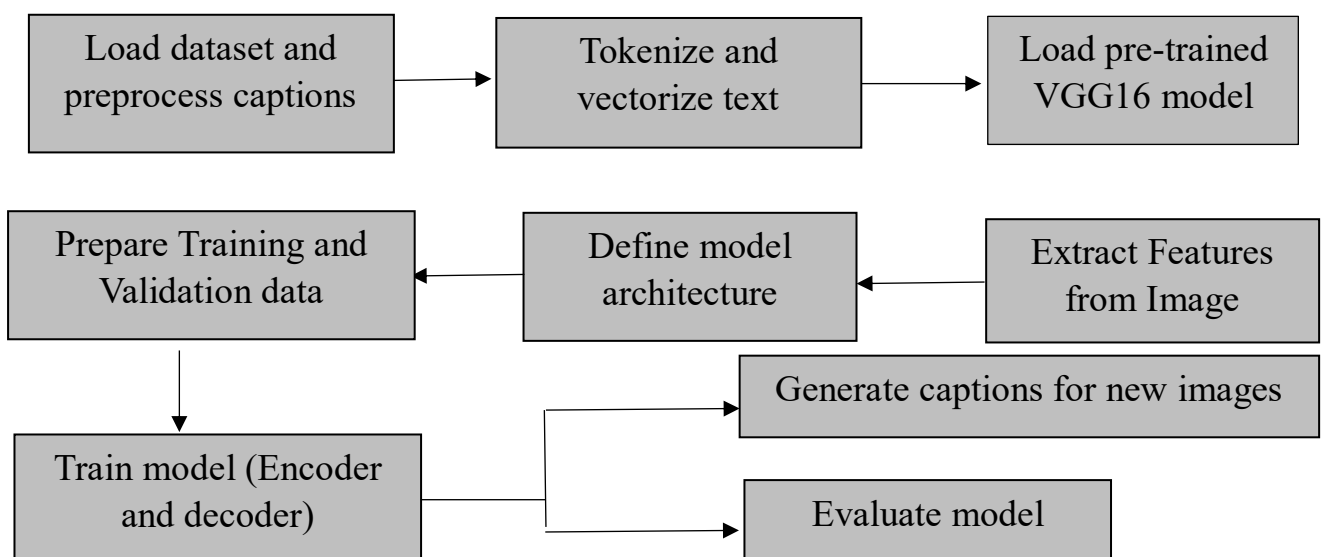
- ***Tokenization:*** Tokenization is used to convert text into numerical sequences. In this project, the captions are tokenized using the Tokenizer class from Keras, which assigns a unique integer to each word in the vocabulary.
- ***Padding:*** Since captions may have varying lengths, they are padded to ensure uniform input size for the LSTM model. The `pad_sequences` function from Keras is employed for this purpose.
- ***One-Hot Encoding:*** Caption words are one-hot encoded to represent them as binary vectors. This encoding scheme is used to transform the output words into a format suitable for categorical cross-entropy loss during model training.

- **BLEU Score:** The Bilingual Evaluation Understudy (BLEU) score is a metric used to evaluate the quality of generated captions. It compares the predicted captions with reference captions provided in the dataset. The NLTK library is utilized to compute the BLEU score in this project.

### Deep learning models :

- **VGG16:** The VGG16 model is used for feature extraction from images. It is a convolutional neural network (CNN) architecture pre-trained on the ImageNet dataset. The model is truncated after the penultimate layer to obtain a 4,096-dimensional feature vector representing the image content.
- **LSTM:** Long Short-Term Memory (LSTM) networks are used for sequence modelling and generation of captions. LSTM is a type of recurrent neural network (RNN) capable of capturing long-term dependencies in sequential data. It processes the encoded image features along with input text sequences to predict the next word in the caption.

### DIAGRAMMATIC REPRESENTATION OF THE WORKING OF THE SYSTEM :

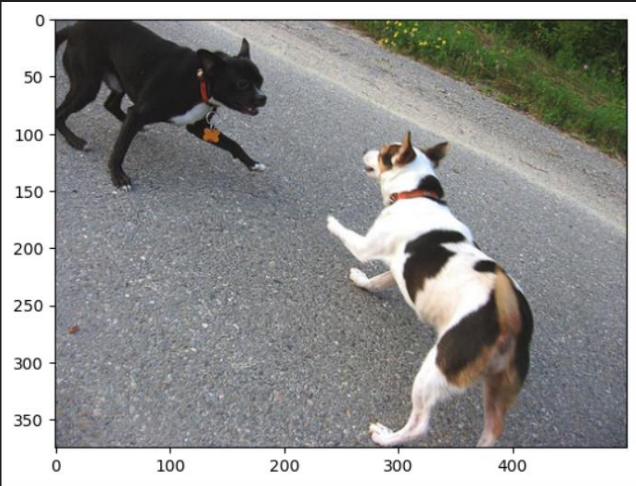


## CAPTIONS GENERATED FOR IMAGES IN THE DATASET:

```
generate_caption("1001773457_577c3a7d70.jpg")
```

[29] ✓ 1.1s

-----Actual-----  
startseq black dog and spotted dog are fighting endseq  
startseq black dog and tri-colored dog playing with each other on the road endseq  
startseq black dog and white dog with brown spots are staring at each other in the street endseq  
startseq two dogs of different breeds looking at each other on the road endseq  
startseq two dogs on pavement moving toward each other endseq  
-----Predicted-----  
startseq two dogs are playing with plastic toys in the grass endseq



```
generate_caption("101669240_b2d3e7f17b.jpg")
```

[31] ✓ 0.9s

-----Actual-----  
startseq man in hat is displaying pictures next to skier in blue hat endseq  
startseq man skis past another man displaying paintings in the snow endseq  
startseq person wearing skis looking at framed pictures set up in the snow endseq  
startseq skier looks at framed pictures in the snow next to trees endseq  
startseq man on skis looking at artwork for sale in the snow endseq  
-----Predicted-----  
startseq skier in red coat is displaying pictures of framed pictures endseq



## CAPTIONS GENERATED FOR REAL-TIME DATA:

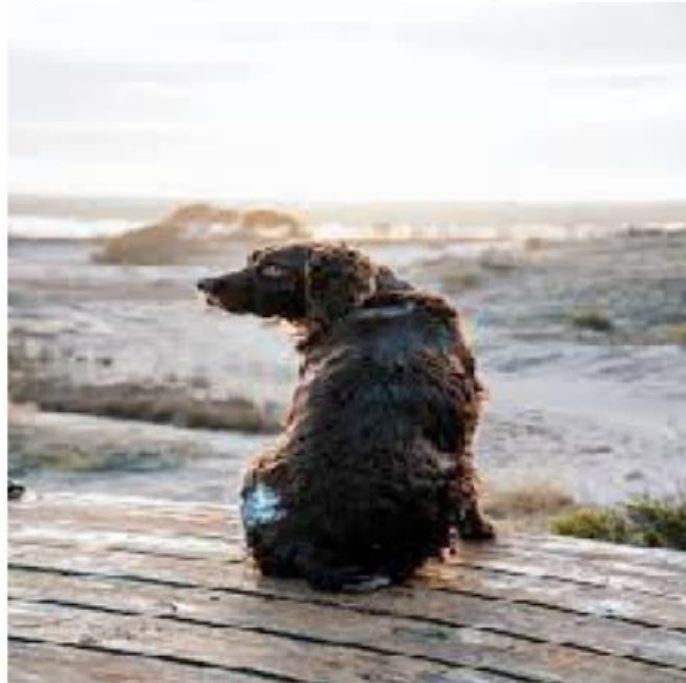
...

startseq little girl in pink shirt is stepping down to her toy toy endseq



...

startseq dog is running on the waters edge of the ocean endseq



## BLEU SCORE ACHIEVED :

```
▶ # calculate average BLEU score
avg_bleu1 = corpus_bleu(actual, predicted, weights=(1.0, 0, 0, 0))
avg_bleu2 = corpus_bleu(actual, predicted, weights=(0.5, 0.5, 0, 0))

print("Average BLEU-1: %f" % avg_bleu1)
print("Average BLEU-2: %f" % avg_bleu2)

[27] ✓ 0.2s

... Average BLEU-1: 0.532965
Average BLEU-2: 0.308947
```

BLEU – 1 (UNIGRAMS)	BLEU – 2 (BIGRAMS)
<b>0.532965</b>	<b>0.308947</b>

## RESULT ANALYSIS

When analyzing the achieved BLEU-1 and BLEU-2 scores of 0.532965 and 0.308947, respectively, the following points can be considered:

### ***BLEU-1 Score (0.532965):***

- BLEU-1 score measures the precision of individual words in the generated captions compared to the reference captions.
- A BLEU-1 score of 0.532965 indicates that around 53.3% of unigrams (single words) in the generated captions match those in the reference captions.
- This score suggests a moderate level of agreement between the generated and reference captions in terms of individual words.
- However, BLEU-1 does not account for word order or context, so a higher score doesn't necessarily mean the generated captions are semantically accurate.

### ***BLEU-2 Score (0.308947):***

- BLEU-2 extends the evaluation to bigrams (pairs of consecutive words) in addition to unigrams.
- A BLEU-2 score of 0.308947 indicates that around 30.9% of bigrams in the generated captions match those in the reference captions.
- This score provides an additional measure of agreement, capturing some level of word pair co-occurrence.
- However, BLEU-2 still has limitations in capturing higher-order dependencies or semantic similarity.

### ***Overall Analysis:***

- The achieved BLEU-1 and BLEU-2 scores suggest a reasonable level of alignment between the generated captions and the reference captions in terms of both individual words and word pairs.
- However, there is room for improvement, especially in capturing higher-order dependencies, semantic coherence, and contextual understanding.
- Further analysis may be needed to understand specific areas where the model struggles, such as handling complex semantics, rare words, or capturing nuanced relationships between objects in images.

## **CONCLUSION :**

Hence, we have developed an image captioning system using machine learning and deep learning techniques, with a moderate performance measure.

## **REFERENCE :**

[https://www.researchgate.net/publication/329037107\\_Image\\_Captioning\\_Based\\_on\\_Deep\\_Neural\\_Networks](https://www.researchgate.net/publication/329037107_Image_Captioning_Based_on_Deep_Neural_Networks)