IMAGE CAPTION GENERATOR USING DEEP LEARNING

TEAM:

G.DHAKSHAYANI

K.VENKATESH

CH.SAI KUMAR

SHUBHAM PANDEY

I. ANU LIKITHA

CONTENT

01 PROBLEM STATEMNT

02 INTRODUCTION

03 APPLICATIONS OF IMAGE CAPTION GENERATOR

04 KEY IDEA IMPLEMENTATION

05 MODEL

06 RESULTS

07 FUTURE PROSPECTS

08 CONCLUSION

PROBLEM STATEMENT

The task is to develop an image caption generator using deep learning techniques. The goal of this model is to automatically generate a textual description that accurately describes the content of an image.

The input to the model is an image, and the output is a natural language sentence that describes the image. This task requires the model to have a good understanding of both visual and textual content.

The model needs to be trained on a large dataset of images and their corresponding captions. The model should be able to learn the relationship between the image and the corresponding caption, and generate accurate captions for new images that it has never seen before.

INTRODUCTION

Image caption generator is a technology combining computer vision, natural language processing, and machine learning techniques. Computer vision is used to identify an image's objects, people, and actions. Natural language processing is used to generate the accurate textual description that best describes the image. And finally, machine learning is used helps to improve the accuracy of the generated captions by training the model on large datasets of annotated images.

A convolutional neural network (CNN) is used to extract visual features from the image, then the images are fed into a long short-term memory (LSTM) network to generate a sequence of words that describe the image as the image caption. This model has revolutionized the way we process and interpret visual information, allowing us to automate the tedious and time-consuming task of manually generating captions for images.

DATA SET

- Flickr 8K dataset, which comprises a total of 8000 images, and contains the images and their respective captions.

- We used 6000 images for training and 2000 images for validation and testing

KEY IDEA IMPLEMENTATION

Stage 1

PRE-PROCESSING

To prepare the data for the image captioning system, a two-part preprocessing approach was adopted.

Stage 2

CNN

It detects the objects in the image

Stage 3

LSTM

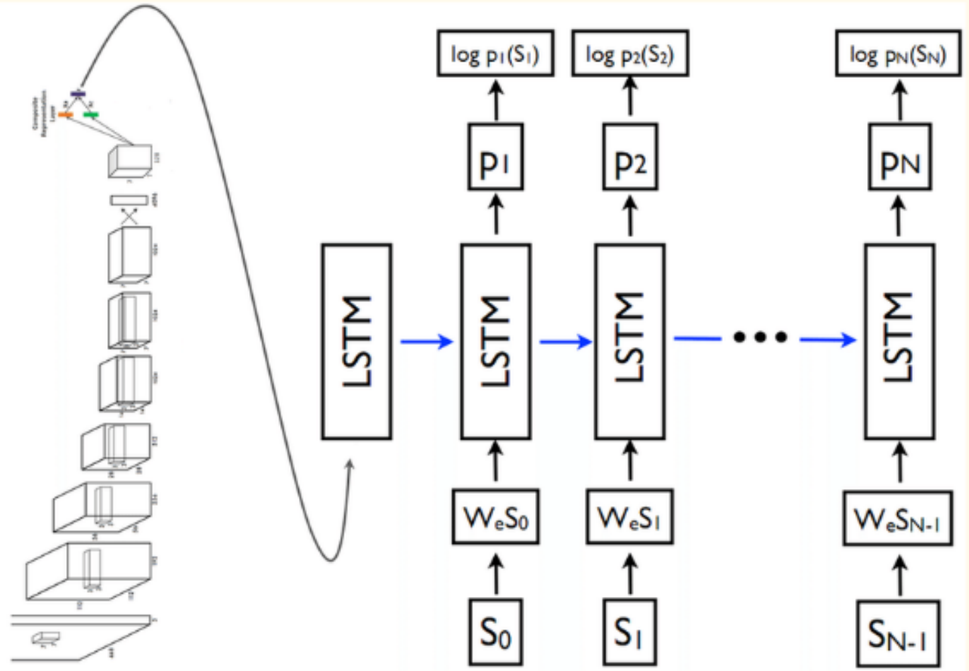
It make captions considering the objects in the image

Pre processing

Original Caption	Cleaned Caption
A football player , his helmet half off , look at the ground .	football player his helmet half off look at the ground
An ice speed skater be skate on the ice wear a red , blue , and gray uniform .	an ice speed skater be skate on the ice wear red blue and gray uniform
Number five of the University of Miami man 's basketball team have the ball .	number five of the university of miami man basketball team have the ball

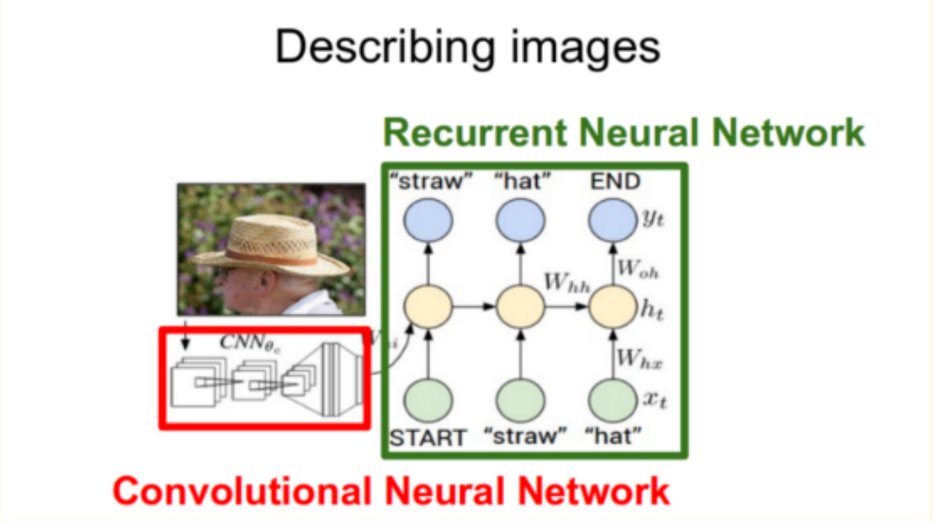
- First, the images and their corresponding captions were cleaned and preprocessed independently.
- The image data was processed by utilizing the Exception application from the Keras API, which is pre-trained on the ImageNet dataset.
- The captions were cleaned using the tokenizer class available in Keras, which enabled vectorization of the text corpus and the storage of the resulting vectors in a separate dictionary. Each word of the captions was then processed further to obtain a representation suitable for use in the image captioning model.

LSTM



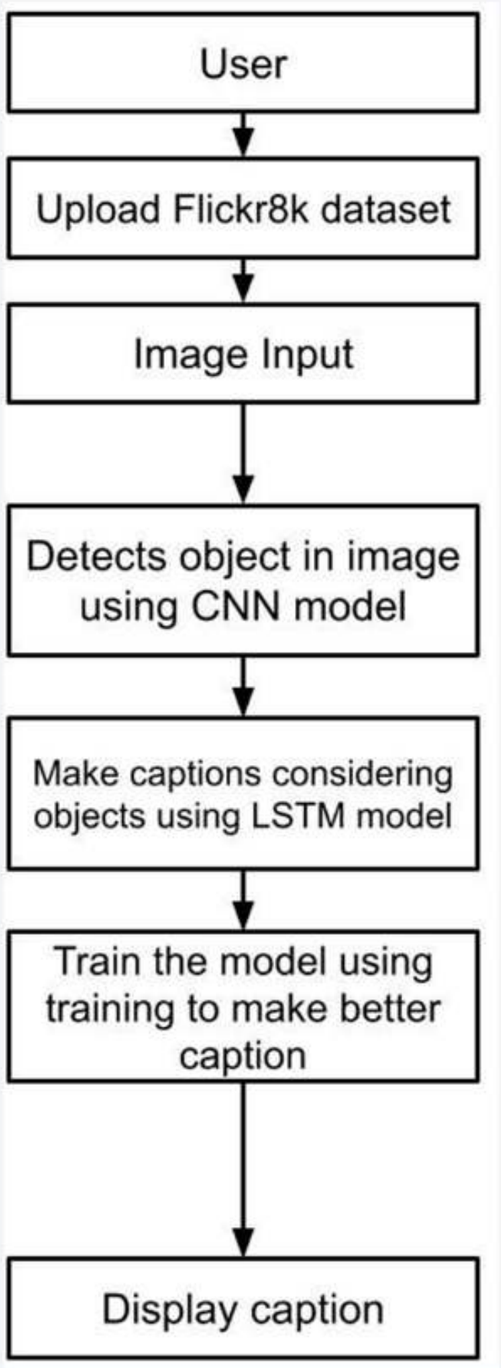
LSTM the potential for long-term dependence. Their behavior is to remember the knowledge for the long term, and this behavior is governed by "GATES." LSTM can handle whole data sets, whereas RNN can only process single data sets.It also determines which data is retained and which is moved to the next layer. The three main gates are INPUT, OUTPUT, and FORGET.These gates determinewhether to forget the current cell value, read it, or output it. The hidden states are crucial since the previous ones are transmitted to the next layer. Hidden layers serve as the neural network's memory, storing data observed by the neural network. As a result, it enables the neural network to function as if it were a human brain

CNN

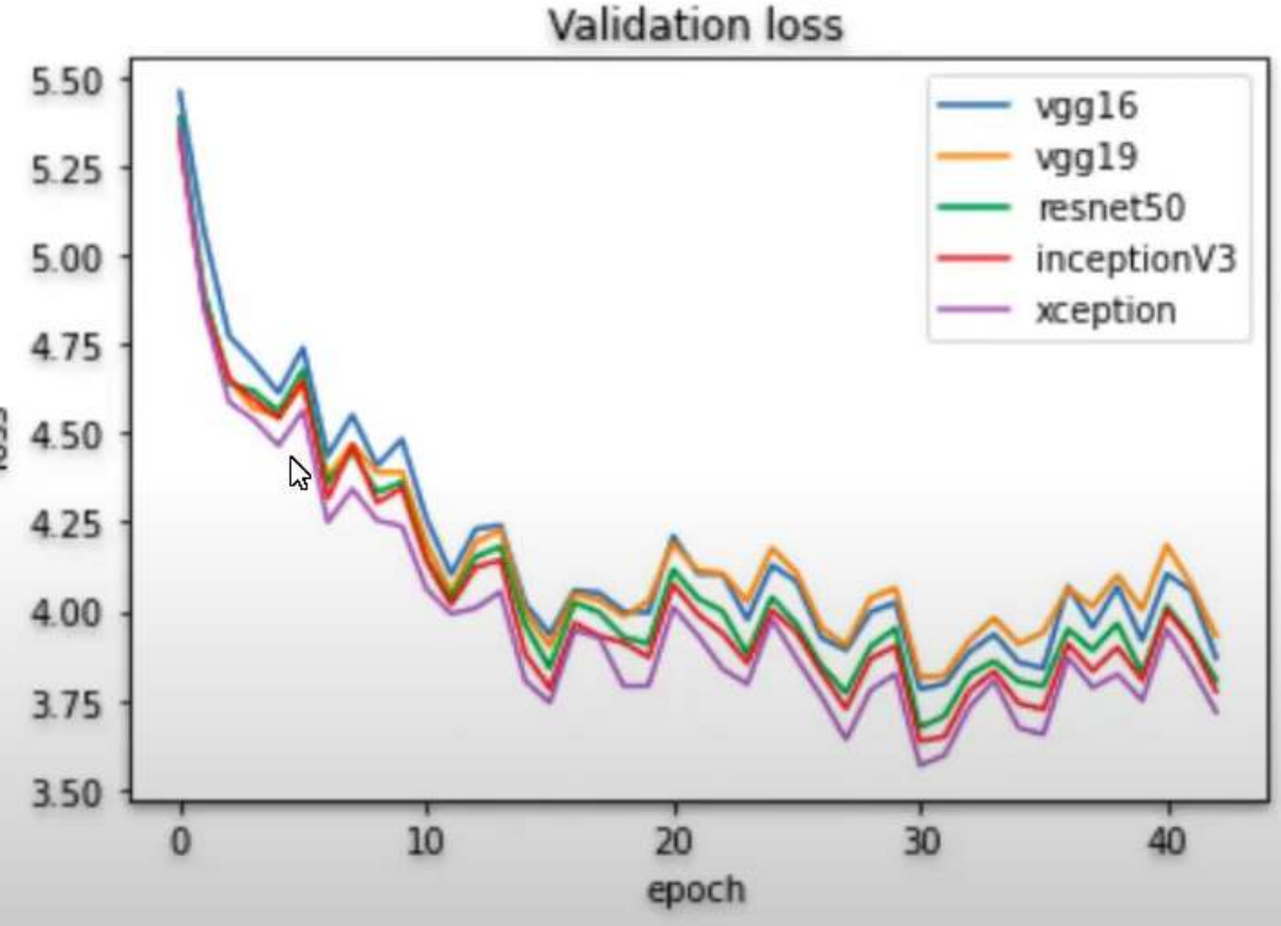


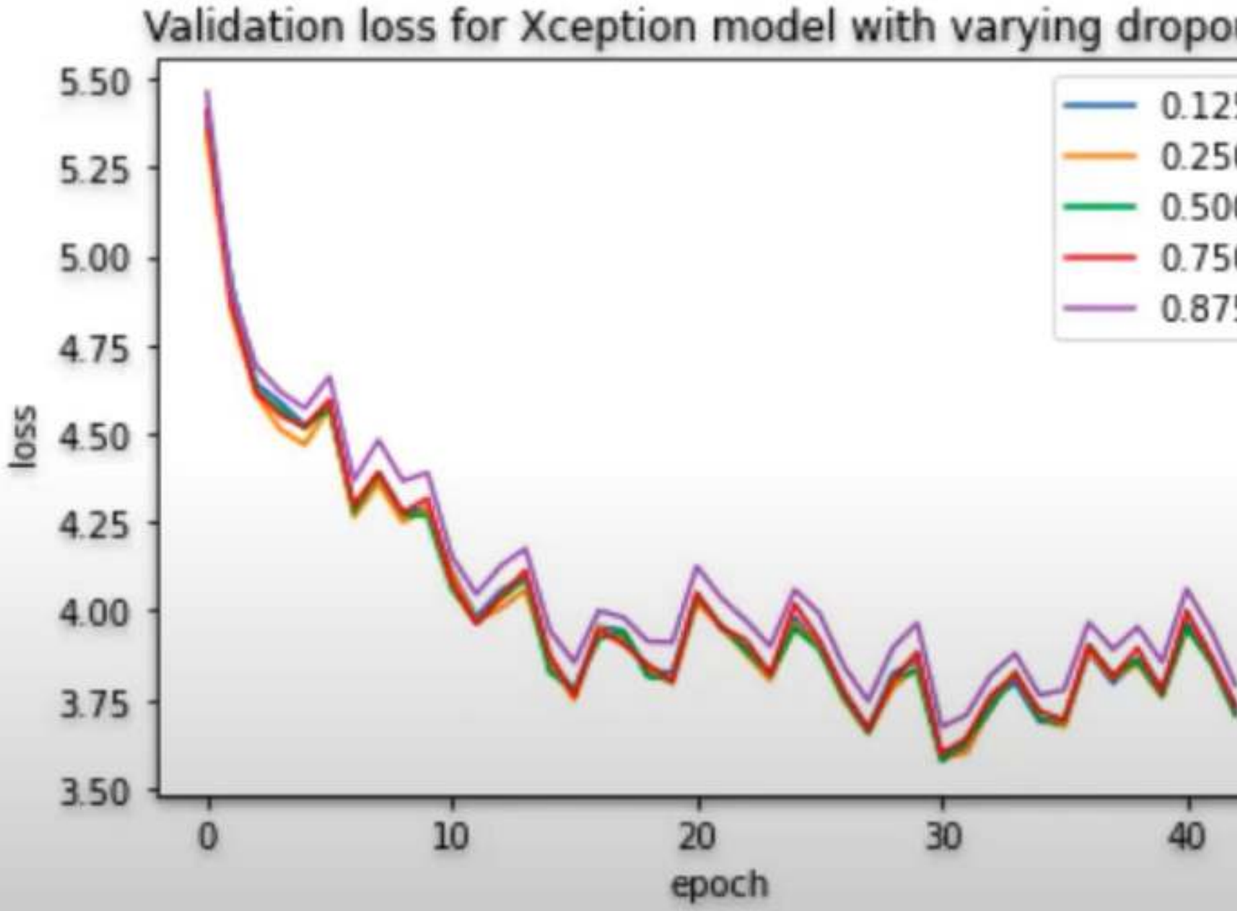
The "encoder" RNN maps the source sentence and transforms it into a fixed-length vector representation, which is then used as the hidden initial state of the "decoder" RNN, which then predicts the final meaningful sentence.
The "encoder" RNN maps the source sentence and transforms it into a fixed-length vector representation, which is then used as the hidden initial state of the "decoder" RNN, which then predicts the final meaningful sentence.
However, we replace this RNN with a deep CNN, which can create a rich representation of the input image by embedding it into a fixed-length vector, first training it for the image classification task, and then using the last hidden layer as the input. RNN. a decoder that generates sentences.

We used Keras Model from Functional API with three primary pieces to create a stacked model. To begin, we will utilize a feature extractor to lower the dimension of the input data from 2048 to 256, followed by a Dropout layer added to the CNN and LSTM to increase generalization. We pre-processed the photographs with the Xception model (minus the output layer) for the feature extractor and will utilize the retrieved features as input. Second, to process the textual input, we will employ an Embedding layer followed by an LSTM layer. Finally, we will use a Dense layer to combine the outputs of the feature extractor and sequence processor to create the final predictions.



Selection of Model





ADAM

optimizer

optimizer algorithm is used to update the weights of the neural networks during training. Adam stands for Adaptive moment estimation. it is combination of RMSProp and Momentum. i will adjust the learning rates of each parameter .

LOSS VALIDATION

VALIDATION LOSS INDICATES HOW WELL THE MODEL FITS NEW DATA.

CATEGORICAL CROSSENTROPY LOSS:

CATEGORICAL_CROSSENTROPY IS A COMMONLY USED LOSS FUNCTION FOR MULTI-CLASS CLASSIFICATION PROBLEMS WHERE EACH INPUT BELONGS TO ONE AND ONLY ONE CLASS.
IT MEASURES THE DISSIMILARITY BETWEEN THE TRUE DISTRIBUTION (GROUND TRUTH) AND THE PREDICTED DISTRIBUTION (OUTPUT OF THE MODEL).
FOR EACH EXAMPLE, IT COMPUTES THE CROSSENTROPY, WHICH IS A MEASURE OF HOW WELL THE PREDICTED PROBABILITIES MATCH THE TRUE DISTRIBUTION OF CLASS LABELS.
IT IS SUITABLE FOR PROBLEMS WHERE EACH SAMPLE CAN ONLY BELONG TO ONE CLASS, AND IT PENALIZES LARGE DEVIATIONS FROM THE TRUE CLASS PROBABILITIES.

FINAL MODEL

IMAGE FEATURE EXTRACTOR: XCEPTION

DROPOUT: 0.125

LSTM MEMORY UNIT: 256

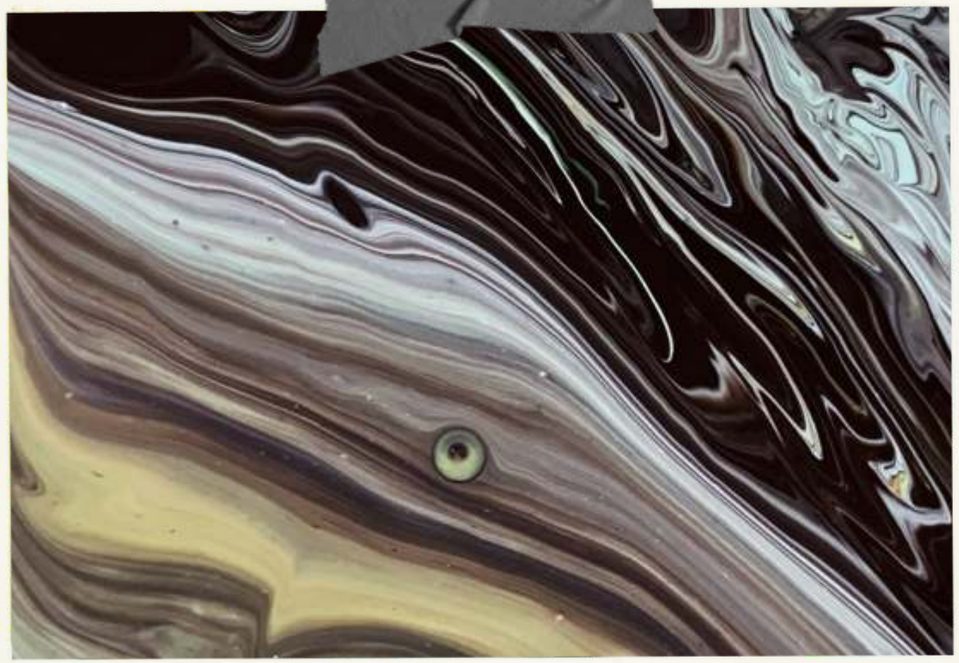
OPTIMIZER: ADAM (INITIAL LEARNING RATE = 0.001)

LOSS FUNCTION: CATEGORICAL_CROSSENTROPY

RESULT

OUR EXPERIMENT RESULTS DEMONSTRATE THAT OUR PROPOSED MODEL OUTPERFORMS MANY OTHER METHODS IN TERMS OF VARIOUS EVALUATION METRICS.

ADDITIONALLY, OUR QUALITATIVE EVALUATION CONFIRMS THAT THE GENERATED CAPTIONS ARE ACCURATE AND INFORMATIVE, INDICATING THE POTENTIAL OF OUR MODEL TO BE A VALUABLE TOOL IN REAL-WORLD APPLICATIONS.



FUTURE PROSPECTS

The performance of our model can be evaluated on additional benchmark datasets, such as the COCO and Flickr30k datasets. By using a larger database which will provide more insights into the advantages and disadvantages of our model and help identify areas for further improvement. In summary, our work has contributed to the growth of image captioning and provides a promising approach for generating informative and diverse captions for images. The future prospects of our proposed model are significant, and we believe that it can pave the way for future research in this field.

CONCLUSION

In conclusion, we have proposed an image caption generator that works on deep learning methods, which has demonstrated a great successful performance on the Flickr 8k dataset.

The model we proposed uses a combination of convolutional neural networks (CNN) and long short-term memory (LSTM) networks to generate informative and diverse image captions.

THANK YOU