Image Caption Generator

Sai Vamsi Gorle

cs.ualr.edu

Abstract

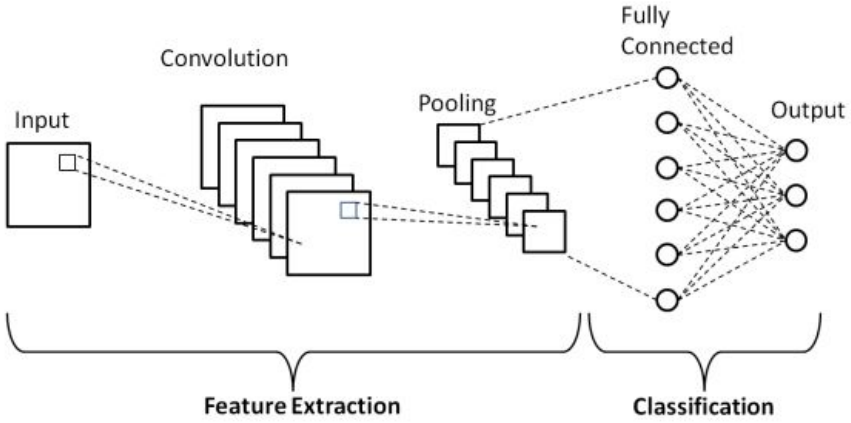
- This project introduces an Image Caption Generator that employs traditional image processing methods, including filtering and segmentation, to extract crucial features from images
- By utilizing Convolutional Neural Networks (CNN), the system identifies intricate patterns and textures within the images
- Integrating Natural Language Processing (NLP) techniques to convert extracted visual features into coherent textual descriptions, enhancing interpretability

Introduction

- The increasing integration of image processing techniques and natural language processing has led to the development of innovative systems that bridge the gap between visual content and textual understanding
- "Image Caption Generator" project aims to leverage traditional image processing methods alongside advanced deep learning algorithms to automatically generate descriptive captions for a diverse range of images
- By combining the strengths of Convolutional Neural Networks (CNNs) for feature extraction and Long Short-Term Memory Networks (LSTMs) for language modeling, the system not only identifies intricate visual patterns but also generates coherent textual interpretations, thereby facilitating a comprehensive understanding of the visual data
- This integration holds significant promise in various practical applications, including aiding the visually impaired, improving image search capabilities, and enhancing the accessibility of visual content

Convolutional Neural Network(CNN)

- CNN is a deep learning algorithm commonly used in image processing
- It applies filters to extract essential features such as edges, textures, and patterns from images, enabling the system to comprehend complex visual data effectively



Algorithm

- **Long Short-Term Memory Network (LSTM)**
  - LSTM is a type of recurrent neural network (RNN) that is widely used in natural language processing tasks
  - It facilitates the generation of coherent and contextually relevant textual descriptions based on the visual features extracted from the images, enhancing the system's ability to understand and interpret the data.

Dataset

- **Flickr 8k Dataset**
  - It consists of 8000 images and 5 captions for each image. The features are extracted from both the image and the text captions for input
  - Dataset: [Kaggle-Image Caption Generator]

Results

Input: generate_caption("1002674143_1b742ab4b8.jpg")

Output:

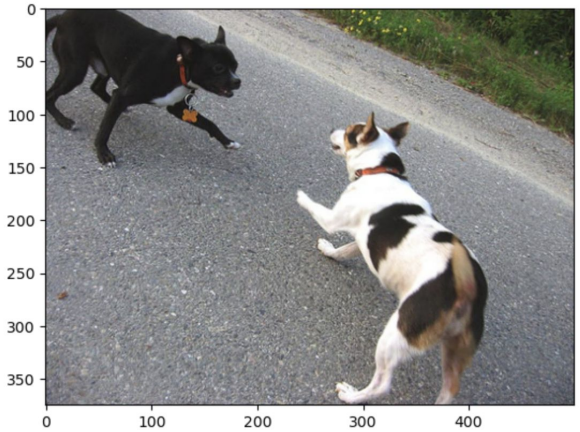
```
----------------------Actual----------------------
startseq little girl covered in paint sits in front of painted rainbow with her hands in bowl endseq
startseq little girl is sitting in front of large painted rainbow endseq
startseq small girl in the grass plays with fingerpaints in front of white canvas with rainbow on it endseq
startseq there is girl with pigtails sitting in front of rainbow painting endseq
startseq young girl with pigtails painting outside in the grass endseq
---------------------Predicted---------------------
startseq little girl is throwing fingerpaints in front of rainbow painting endseq
```

Results

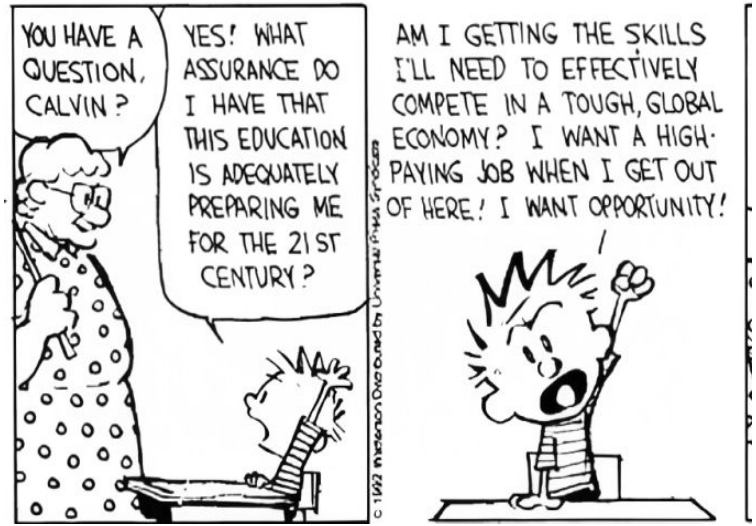
Input: generate_caption("1001773457_577c3a7d70.jpg")

Output:

```
----------------------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 each other in the snow endseq
```

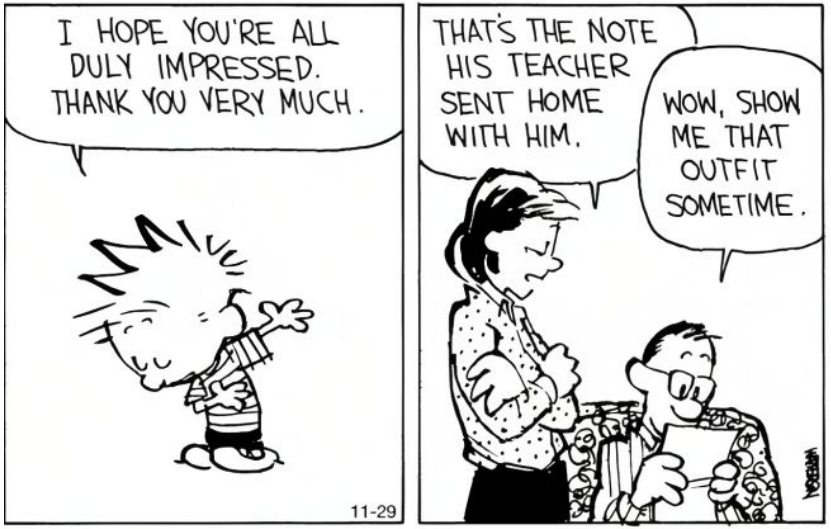


Questions?



https://dougslangandlit.blog/2022/03/07/you-have-a-question-calvin/

Thank You



shorturl.at/xBN24