Detection and Caption Generation of Image Using Deep Learning

1.Jivan Mate **2.**Jaykumar Patil **3.**Nikita Waghmare **4.**Omkar Pawar **5.**Mrs.Y.N.Sakhare

[1,2,3,4]UG students, Department of Information Technology
[5]Asst. Prof., Department of Information Technology
Vidya Pratishthan's Kamalnayan Bajaj Institute of Engineering and Technology, Baramati

Abstract ***- In this research paper, we employ a combination of CNN and LSTM to tackle image caption generation, a task at the intersection of natural language processing and computer vision. We meticulously discuss key concepts in photograph captioning and its methodologies, leveraging resources like the Keras library, numpy, and Jupyter notebooks. Our research also delves into the flickr_dataset and CNN for photo classification, aiming to shed light on the intricate processes underlying image understanding and caption generation.***

Keywords*- CNN ,LSTM, Image captioning, Deep Learning.*

INTRODUCTION

In our daily encounters, we are inundated with photographs across various platforms, from social media to newspapers. Unlike machines, humans possess the innate ability to discern images without relying on captions. However, for machines, image captioning necessitates training data to automatically generate descriptive text.

The applications of image captioning are manifold. It can aid visually impaired individuals by providing real-time feedback through text-to-speech technology, enhance social media experiences by organizing photo captions and messages for speech, and aid in childhood education by helping children recognize objects while learning language..The utility of image captioning extends across diverse domains including biomedicine, commerce, internet search, and military applications. Social media platforms like Instagram and Facebook are poised to leverage automatic caption generation from images to enhance user experience.

This research paper primarily aims to delve into deep learning methodologies, particularly focusing on Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks for image classification. Through our exploration, we seek to gain insights into the potential of these techniques in advancing the field of image understanding and captioning.

IMAGE CAPTIONING TECHNIQUES

CNN - Convolutional Neural Networks (CNNs) are pivotal neural architectures adept at processing data with spatial structures, such as 2D grids, making them particularly suited for image-related tasks. Operating by scanning images from left to right and top to bottom, CNNs extract essential features crucial for image understanding. These features are then consolidated to classify and interpret images effectively.

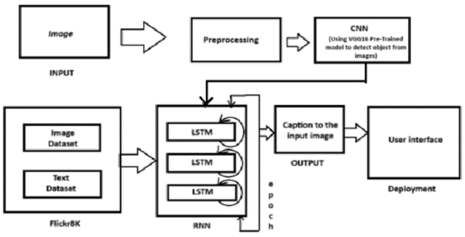


Fig.1 System Architecture

The structure of the curved system is like the neuronal network design inside human mind & is inspired by the way of the organization of the visual cortex Convolutional Neural Networks (ConvNets) demand relatively little preprocessing in contrast to alternative computational approaches. Despite the initial rudimentary design of channels, ConvNets possess the capability to learn these channels or features effectively with ample training data. The architecture of ConvNets mirrors the neural network structure found within the human brain, drawing parallels to the organization of the visual cortex. Singular neurons within ConvNets react to stimuli within confined regions of the visual field, termed receptive fields. The aggregation of these receptive fields encompasses the summation of visual regions, facilitating nuanced and comprehensive image analysis. This intrinsic structure aligns with the principles of hierarchical feature extraction observed in biological vision systems classification

CNN Architecture: A conventional fully connected neural network, where all neurons in one layer connect to all neurons in the next, proves inefficient for analyzing large images and videos. When dealing with a typical-sized image consisting of numerous pixels and three-tone colors (RGB), the sheer magnitude of parameters in such a network can lead to overfitting.

To mitigate this issue and enable effective feature recognition across significant portions of an image, Convolutional Neural Networks (CNNs) employ a 3D structure. In this arrangement, each set of neurons analyzes a small region or "feature" of the image. Instead of every neuron transmitting its activations to the next layer indiscriminately, groups of neurons specialize in detecting specific features such as a nose, a left ear, mouth, or a leg. The final output is a map of activations, illustrating the relevance of each feature to the class being classified. This hierarchical approach enables CNNs to efficiently process and interpret complex visual data, making them well- suited for tasks like image classification, object detection, and semantic segmentation.

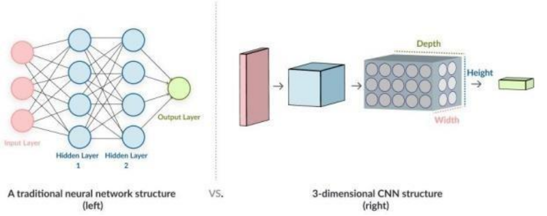
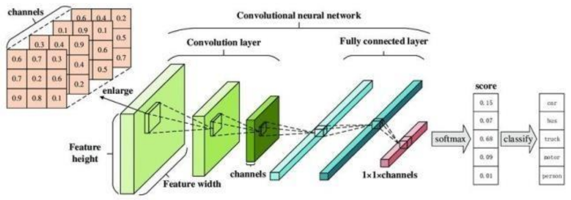

Input Layer
Hidden Layer 1
Hidden Layer 2
Output Layer
A traditional neural network structure (left)
VS.
Depth
Height
Width
3-dimensional CNN structure (right)

Fig.2 Working of CNN

How does the CNN works ?

CNNs work by leveraging a specialized architecture designed to effectively process and interpret visual data. Unlike fully connected neural networks, where each neuron in one layer is connected to every neuron in the subsequent layer, CNNs employ a different approach.



In CNNs, neurons in a layer are connected to a specific region or receptive field in the preceding layer, rather than being connected to all neurons in a similar manner. This means that each neuron in a CNN focuses on analyzing a local region of the input data, allowing for the extraction of meaningful features from the input images.

By operating in this manner, CNNs can efficiently capture spatial hierarchies and local patterns in the data. This hierarchical feature extraction enables CNNs to effectively recognize complex patterns and objects within images, making them particularly well-suited for tasks such as image classification, object detection, and segmentation.

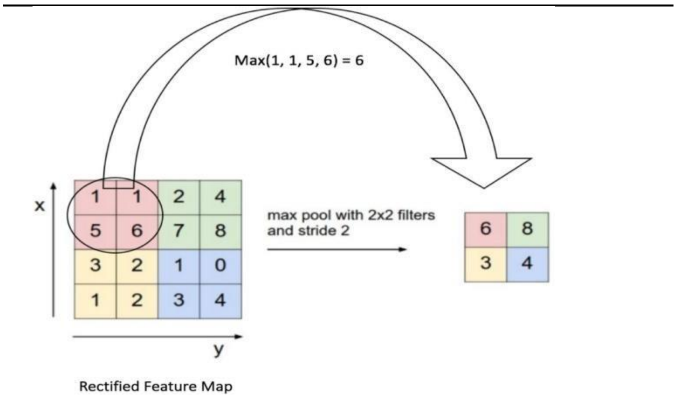


Fig.3 Feature map of CNN picture

CNNs are favored for their unique capability to directly process image data. They classify pixels based on similarities and differences, generating feature maps that group similar pixels together. These feature maps are instrumental in identifying the essence of objects within the input image. By efficiently highlighting relevant features while suppressing noise, CNNs excel in tasks like object recognition and image classification.

More about CNN :

There are 3 kinds of layers in given model

1. Convolutional
2. Pooling
3. Fully connected

In the initial layer of a CNN, the input image is processed to create a feature map. This feature map serves as the input for subsequent layers, such as the Pooling layer. In the Pooling layer, the feature map is subdivided into smaller parts to better analyze the image context, resulting in a denser representation of the features.

The Convolutional and Pooling layers are iteratively applied multiple times, depending on the complexity of the image, to extract denser information. The resulting denser feature map is then utilized by the final layer, the Fully Connected layer, for further analysis or classification. This iterative process of convolution and pooling enables the CNN to progressively extract and refine features from the input image, leading to effective image understanding and classification.

The final layer of the CNN performs classification by categorizing pixels based on their similarities and differences. This process is carried out to an exceptional degree to distill the essence of the picture, aiding in the identification of objects, persons, and other elements within the image. Through classification, the CNN assigns labels or probabilities to different classes, enabling accurate recognition and interpretation of the visual content.

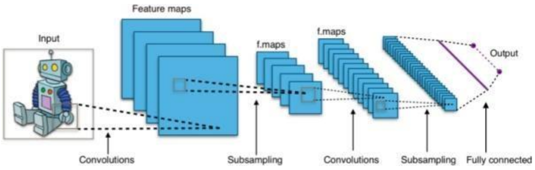


Fig.4 Layers of scanned picture

These layers within a CNN play a crucial role in precisely locating and extracting features from images. They enable the extraction of essential features from images with variable-length inputs, transforming them into fixed-size outputs. This process ensures that relevant information is captured effectively, facilitating accurate analysis and interpretation of visual data.CNN techniques are very much in usage.

- **Computer Vision in Medicine:** CNNs are extensively utilized for image analysis in medical sciences, aiding in tasks such as diagnosing diseases
- **Mobile Applications**: In mobile phones, CNNs serve a multitude of purposes, including facial recognition for determining age, gender, or identity verification.
- **Pharmaceuticals Discovery:** CNNs play a significant role in pharmaceutical research for drug discovery and development.

LSTM -

LSTM, short for Long Short-Term Memory, was pioneered by German researchers Sepp Hochreiter and Jurgen Schmidhuber in 1997. Within the realm of recurrent neural networks (RNNs) in deep learning, LSTM holds a significant position. What sets LSTM apart is its ability not just to retain input data but also to make predictions about future data points based on its own internal state. Unlike traditional RNNs, LSTM networks can store information for extended periods and use it to forecast future values accurately. This capability makes LSTM particularly valuable for tasks requiring long- range dependencies and precise temporal modeling, explaining its widespread usage in various fields of machine learning and artificial intelligence.

The Problem with RNNs -

The problem with RNNs (Recurrent Neural Networks) lies in their struggle to effectively handle long-term dependencies in sequential data. While RNNs are powerful algorithms used for complex tasks such as item classification and speech recognition, they face challenges when confronted with sequences of data where each step depends heavily on information from previous steps.

Specifically, RNNs are favored for tasks involving extensive datasets and high-dimensional features. They are capable of addressing real-life problems such as inventory forecasting and improving speech recognition accuracy. However, the practical application of RNNs is hindered by a phenomenon known as the Vanishing Gradient problem.

This issue occurs during the training of RNNs, where gradients become extremely small as they are backpropagated through time. As a result, the network struggles to learn long- term dependencies effectively, limiting its ability to accurately model sequential data over extended periods. This limitation has led researchers to explore alternative architectures like LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit) to mitigate the Vanishing Gradient problem and enhance the performance of recurrent neural networks in real-world applications.

Vanishing Gradient Problem –

The vanishing gradient problem presents a significant challenge for the effective operation of RNNs (Recurrent Neural Networks). Typically, RNNs are engineered to retain information for a short period, limiting their ability to remember long sequences of data. This problem becomes more pronounced compared to traditional RNNs when solving tasks that require many time steps. As the number of time steps increases, RNNs struggle to store and propagate data values through backpropagation, leading to the loss of information. This limitation stems from the inherent difficulty of storing and processing extensive amounts of data over extended periods, ultimately resulting in the vanishing gradient problem.

Solutions To address the vanishing gradient problem in RNNs:

Initialization Methods : Choose appropriate initialization methods for network parameters to ensure gradients neither vanish nor explode during training.

Gradient Clipping : Scale gradients during backpropagation to prevent them from becoming too large or too small, stabilizing training.

Use LSTM and GRU: Employ specialized RNN architectures like LSTM and GRU, which incorporate gating mechanisms to regulate gradient flow and better retain information over longer sequences.

Uses of LSTM Networks :-

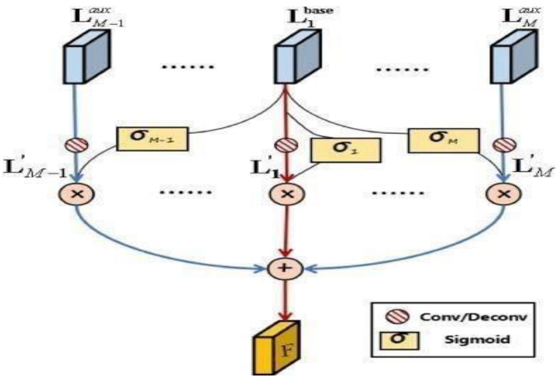


Fig.5 Gates in LSTM

In contrast to traditional RNNs, LSTM architectures incorporate several gates to manage and process data effectively. These gates enable LSTM networks to shape and manipulate data in various ways, including storing and reviewing it over time. Each gate within an LSTM network is independently capable of making judgments about the data, determining whether to allow information to pass through or be stored for later use.

The ability of LSTM gates to retain and manipulate data over extended periods offers distinct advantages over traditional RNNs. By selectively opening and closing gates, LSTMs can control the flow of information and better capture long-term dependencies in sequential data, making them particularly well-suited for tasks requiring memory and context retention.

Architecture Of LSTM :-

The architecture of LSTM is indeed straightforward, consisting of three major gates designed to address challenges that traditional RNNs face:

1. Forget Gate: The forget gate filters and selectively discards irrelevant information that is not needed for solving a specific task. It plays a crucial role in optimizing the overall performance of the LSTM by controlling the retention of past information.

2. Input Gate: Serving as the entry point of the LSTM, the input gate receives input from the user and distributes this input data to the other gates within the network. It initiates the processing of new information.

3. Output Gate: The output gate is responsible for presenting the final output in a coherent and meaningful manner, showcasing the desired result derived from the processed information within the LSTM.

LSTMs are extensively utilized in various deep learning tasks, particularly for data forecasting based on historical data. Two prominent examples include text prediction and stock market prediction.

- **Text Prediction**: LSTMs excel in predicting the next words in a sequence due to their long-term memory capabilities. By understanding the context, style, and usage of words, LSTMs can autonomously predict subsequent words in sentences. Chatbots in eCommerce websites and mobile applications are prime examples of text prediction applications leveraging LSTMs.

- **Stock Market Prediction** : LSTMs analyze historical market trends to predict future variations and trends in the stock market. Despite the complexity of market fluctuations, LSTMs can be trained to provide accurate forecasts. However, this requires extensive training with large datasets spanning significant periods, sometimes even days

- **Healthcare:** In healthcare, LSTMs are employed for tasks like patient monitoring, disease prediction, and medical diagnosis. They can analyze sequential patient data, such as vital signs and medical records, to detect anomalies and predict health outcomes.

- **Financial Forecasting:** LSTMs are utilized in financial markets for predicting stock prices, currency exchange rates, and market trends. They can analyze historical financial data to identify patterns and make informed predictions about future market movements.

- **Gesture Recognition:** LSTMs can be used in gesture recognition systems to interpret and classify hand movements captured by sensors or cameras. They can recognize and interpret complex gestures for applications like sign language recognition and gesture-based interfaces.

- **Autonomous Vehicles:** LSTMs are employed in autonomous vehicles for tasks such as object detection, pedestrian tracking, and trajectory prediction. They can analyze sequential data from sensors like cameras and LiDAR to make real-time decisions in complex driving scenario.

Image Caption Generation Model:-

By integrating CNN and LSTM, the model leverages the power of both architectures: CNN for image feature extraction and LSTM for sequence modeling and caption generation. This hybrid approach allows the model to produce accurate and contextually relevant captions for a wide range of input images.

1. **CNN (Convolutional Neural Network):** CNN is utilized to extract essential features from the input image. In this model, a pre-trained CNN model called Xception is employed. Xception is known for its ability to capture intricate features from images effectively.

2. **LSTM (Long Short-Term Memory):** LSTM is responsible for storing and processing the features extracted by the CNN model.

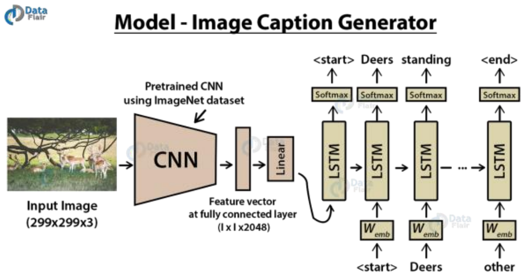
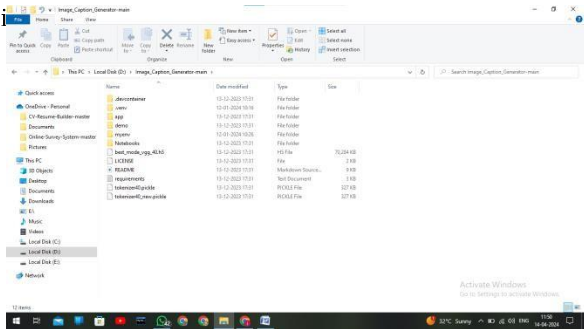

Data Flair
Model - Image Caption Generator
Pretrained CNN using ImageNet dataset
Input Image (299x299x3)
CNN
Linear
Feature vector at fully connected layer (l x l x2048)
<start> Deers standing <end>
Softmax
Softmax
Softmax
Softmax
LSTM
LSTM
LSTM
LSTM
W emb
W emb
W emb
<start> Deers other

Fig.7 CNN-LSTM model

6. **Myenv:** Environment to run the model. Virtual environments allow you to isolate the dependencies of di


Image_Caption_Generator-main
File Home Share View
Pin to Quick access Copy Paste Cut Copy path Paste shortcut
Move to Copy to Delete Rename
New folder New item Easy access
Properties Open Edit History
Select all Select none Invert selection
Clipboard Organize New Open Select
This PC > Local Disk (D:) > Image_Caption_Generator-main
Quick access
OneDrive - Personal
CV-Resume-Builder-master
Documents
Online-Survey-System-master
Pictures
This PC
3D Objects
Desktop
Documents
Downloads
E:\
Music
Videos
Local Disk (C:)
Local Disk (D:)
Local Disk (E:)
Network
12 items
Activate Windows

Name	Date modified	Type	Size
.devcontainer	13-12-2023 17:31	File folder	
.venv	12-01-2024 10:18	File folder	
app	13-12-2023 17:31	File folder	
demo	13-12-2023 17:31	File folder	
myenv	12-01-2024 10:26	File folder	
Notebooks	13-12-2023 17:31	File folder	
best_mode_vgg_40.h5	13-12-2023 17:31	H5 File	70,284 KB
LICENSE	13-12-2023 17:31	File	2 KB
README	13-12-2023 17:31	Markdown Source...	9 KB
requirements	13-12-2023 17:31	Text Document	3 KB
tokenizer40.pickle	13-12-2023 17:31	PICKLE File	327 KB
tokenizer40_new.pickle	13-12-2023 17:31	PICKLE File	327 KB

.

Fig.8 Project file structure

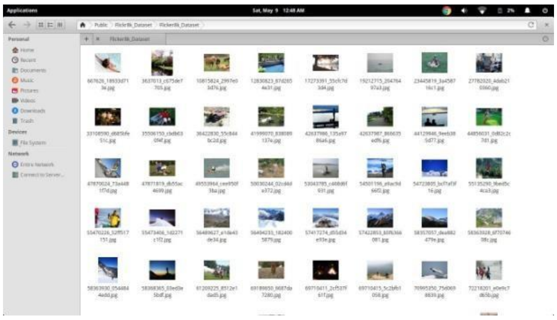


Fig.9 Flicker 8k Dataset

Project File Architecture :-

It seems like you have a comprehensive dataset and supporting files for your research on image caption generation. Here's a brief overview of each file:

1. **Flickr8k_Datasets:** Contains 8091 images for training the model.

2. **Description.txt** This file will store the picture names and their related captions after preprocessing.

3. **Tokenizers.pkl:** Contains tokens.

4. **.app.py**: Python script for generating captions for input pictures.

5. **Image_captioning_by_VGG16.ipynb :** Jupyter notebook used for training the model and generating captions for input pictures.

CONCLUSION

In conclusion, the CNN-LSTM model was developed with the goal of generating captions for input pictures, offering a versatile solution applicable to various tasks. Through our study, we delved into the concepts of CNN, RNN, and LSTM models, understanding their individual contributions to the overall architecture. By combining these models, we successfully validated the capability of the CNN-LSTM model to generate descriptive captions for input images. This model holds promise for a wide range of applications, demonstrating its potential to enhance tasks requiring image understanding and interpretation.

REFERENCES

[1] Micah Hodosh, Peter Young, Julia Hockenmaier(2021) "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics".

[2] Aishwarya Maroju, Sneha Sri Doma, Lahari Chandarlapati, 2021, "Image Caption Generating Deep Learning Mode",INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH TECHNOLOGY (IJERT) Volume 10, Issue 09 (September 2021).

[3] Imtinan Azha, Imad Afyouni, Ashraf Elnagar 2021 55th Annual Conference on Information Sciences and Systems (CISS) "Facilitated Deep Learning Models for Image Captioning"
.

[4] Facilitated Deep Learning Models for Image Captioning", 021 55th Annual Conference on Information Sciences and Systems (CISS) | 978- 1-6654-1268-1/21/$31.00 ©2021 IEEE

[5] Soheyla amirian,KhaledRasheed,Thiab R taha and Hamid R. arabnia 2020 "Automatic Image and Video Caption Generation With Deep Learning:A Concise Review and Algorithmic Overlap".

[6] Geetha,T.Kirthigadevi, T.Karthik, G GODWIN,M.Safa, "Image Captioning Using Deep Convolutional Neural Networks(CNNs)" Published under license by IOP Publishing Ltd in Journal of Physics: Conference Series , Volume 1712, International Conference On Computational Physics in Emerging Technologies(ICCPET) 2020.

[7] Syed Haseeb, Srushti G M, Bhamidi Haripriya, Mrs. Madhura Prakash, 2019, "Image Captioning using Deep Learning", INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH TECHNOLOGY (IJERT) Volume 08, Issue 05 (May 2019).

[8] Varsha Kesavan,VaidehiMuley,Megha Kolhekar 2019 "Deep Learning based Automatic Image Caption Generation".

[9] Edy Mulyanto,Esther Irawati Setiawan,EkoMulyantoYuniarno,Mauridhi Hery Purnomo 2019 "Automatic Indonesian Image Caption Generation using CNN-LSTM Model and FEEH-ID Dataset".

[10] **"**Deep Learning based Automatic Image Caption Generation", 2019 Global Conference forAdvancement in Technology (GCAT) Bangalore, India. Oct 18-20, 2019

[11] "Image Caption Generation using Deep Learning Technique", 2019 Fourth International Conferenceon Computing Communication Control and Automation (ICCUBEA)

[12] Edy Mulyanto, Esther Irawati Setiawan, Eko MulyantoYuniarno, Mauridhi Hery Purnomo 2019 "Automatic Indonesian Image Caption Generation using CNN-LSTM Model and FEEH-ID Dataset".