JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

CAPTION GENERATOR OF IMAGE TO TEXT USING DEEP LEARNING

[1]PRATHIMA GAMINI, [2]PURURVA MANIKANTA BAVANA, [3]BHARGAVI SAI DHARMAVARAPU, [4]MADHULATHA AYIREDDY, [5]RAJESH GADHAM
[1]ASSISTANT PROFESSOR, [2]STUDENT, [3]STUDENT, [4]STUDENT, [5]STUDENT
Electronics and Communication Engineering Department, Sagi Ramakrishnam Raju Engineering College, Bhimavaram, West Godavari District, Andhra Pradesh, India

ABSTRACT: This paper presents an approach to image captioning using a combination of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. The ResNET-50 model serves as an encoder to extract meaningful features from images, while the LSTM-based decoder generates coherent and contextually relevant captions. The dataset utilized is the Flickr8k dataset, comprising 8,000 images, each associated with five human-generated captions. This dataset facilitated training the model to understand diverse contexts and generate descriptive captions capturing various aspects of the images. The pre-trained ResNET-50 encoder extracts high-level features from input images, which are then fed into the LSTM-based decoder responsible for generating sequential descriptions. The LSTM network will be trained to grasp temporal dependencies and relationship between the extracted features, thereby ensuring the production of accurate and contextually rich captions.

Keywords: - Deep Learning, CNN, ResNet50, VGG16, InceptionV3, Xception, MobileNet, BLEU Score, LSTM.

INTRODUCTION:

Image captioning is the process of describing the visual content of an image using natural language. This task requires both a visual understanding system and a language model capable of generating coherent sentences. While neuroscience research has shed light on the connection between human vision perception and language generation, the development of architectures in Artificial Intelligence to process images and generate language is a recent endeavor.

Fig. 1: Sample image for image captioning

The primary aim of such projects is to establish an effective pipeline that can process input images, extract meaningful visual representations and convert them into textual descriptions while ensuring linguistic fluency. Presently, image captioning predominantly leans on generative models rooted in deep learning. Within this conventional setup, the problem is typically structures as a transition from images to sequence, where the input consists of pixel data.

During the visual encoding steps, the input image is transformed into one or multiple feature vectors, these images are inputted into the language model to generate. This language model decodes the encoded features into a sequence of words or subwords, adhering to a predefined vocabulary.

This paper specifically focuses on utilizing the ResNet-50 model as an image encoder and Long Short-Term Memory (LSTM) networks as a decoder for image captioning. ResNet50, renowned for its exceptional performance in image classification, serves as an effective feature extractor, capturing intricate details and semantic information from input images. The LSTM network, on the other hand, functions as a sequential generator, learning to construct coherent and contextually relevant captions based on the extracted feature

ENCODER:

In the domain of neural networks, encoders play a pivotal role in transforming input data into a latent representation that captures meaningful features, facilitating downstream tasks such as classification, generation, and clustering. With the emergence of deep learning, encoder architectures have experienced significant progress, evolving from conventional convolutional and recurrent networks to more advanced variations such as transformers and autoencoders. These encoders have demonstrated remarkable efficiency across diverse fields encompassing computer vision, natural language processing, and audio processing.

CONVOLUTIONAL NEURAL NETWORK:

Convolutional Neural Networks (CNNs) have emerged as a cornerstone technology, revolutionizing tasks such as image classification, object detection and semantic segmentation. CNNs are a class of deep neural networks designed to automatically learn hierarchical representations of visual data through the application of convolutional filters and pooling operations. Their architecture is inspired by the organization of the visual cortex in animals. CNNs are used to effectively capture spatial dependencies and local patterns in images. Since their inception, CNNs have achieved remarkable success in various applications, including medical image analysis, autonomous driving and facial recognition.

VGG16 (Visual Geometry Group 16):

The VGG16 architecture, introduced by the Visual Geometry Group at Oxford, is known for its simplicity and effectiveness. It consists of 16 layers with small 3x3 convolutional filters, and it has been successfully used as an encoder in image captioning models.

ResNet-50 (Residual Network):

ResNet introduced the concept of residual learning, which helps in training very deep networks. ResNet architectures with varying depths (e.g., ResNet-50, ResNet-101) are commonly used as encoders. The skip connections within ResNet enable the network to grasp residual information, making it easier to train and improve the performance.

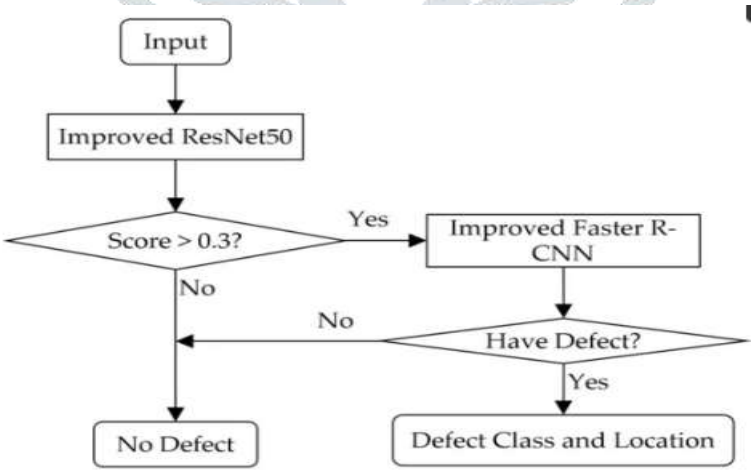

Fig.2: Flowchart for ResNet50

InceptionV3:

The Inception architecture, also known as Google Net, utilizes inception modules with multiple filter sizes in parallel. In image captioning models, Inception-based CNNs have been utilized as encoders, providing good performance.

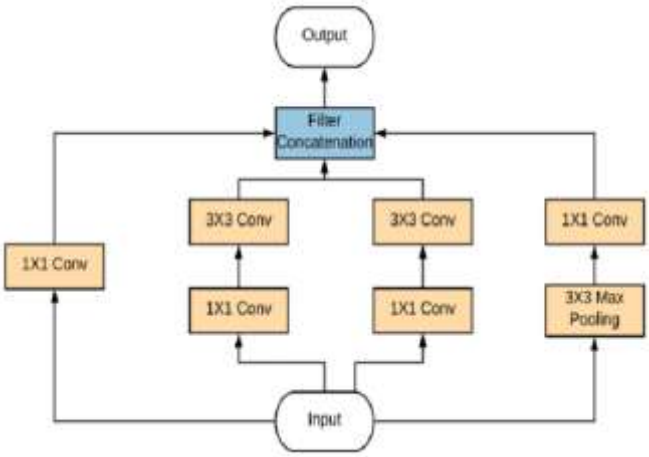

Fig.3: Architecture of InceptionV

MobileNet:

MobileNet is designed for efficiency and is often used in scenarios with limited computational resources, such as mobile devices. It employs depth-wise separable convolutions to decrease parameter count while preserving performance quality. MobileNet architectures can serve as lightweight encoders for image captioning.

Dense Net (Densely Connected Convolutional Networks):

Dense Net establishes connections between each layer in a feed-forward manner, facilitating dense connectivity. This approach fosters feature reuse and ensures smoother gradient flow, thereby enhancing information propagation within the network. Image captioning models leverage Dense Net-based encoders to capture intricate feature representations effectively.

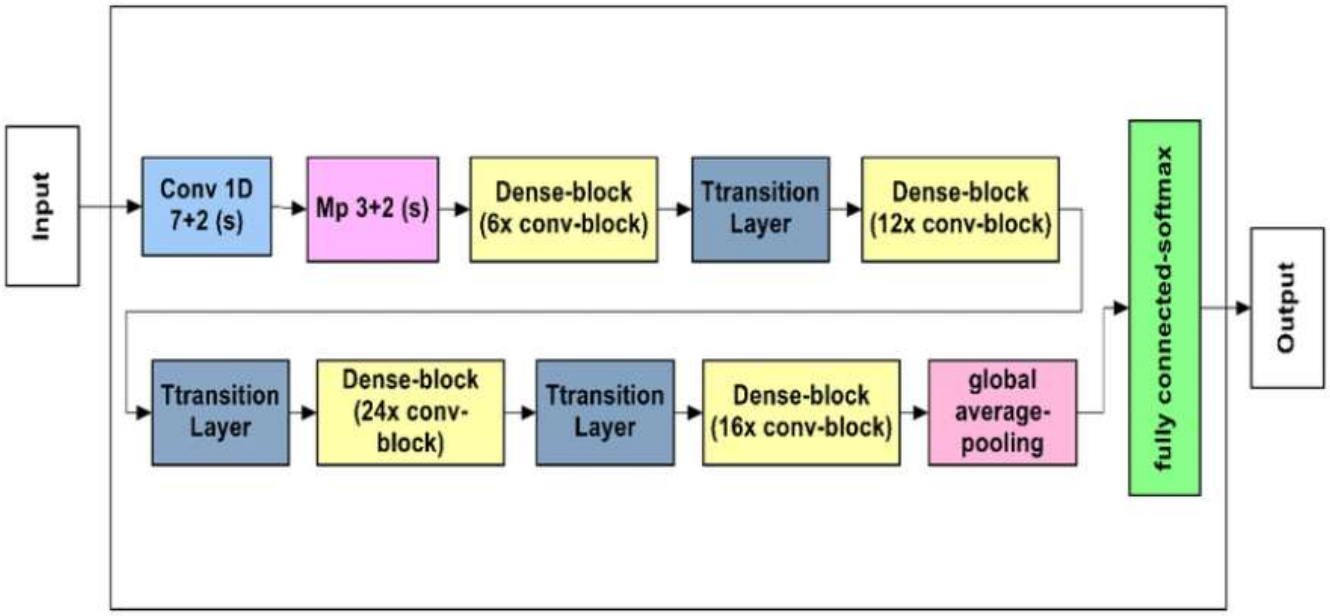

Fig.4: Architecture of DenseNet

EfficientNet:

EfficientNet represents a significant advancement in convolutional neural network (CNN) architectures, particularly within the realm of computer vision. EfficientNet strives to achieve superior performance with significantly fewer parameters compared to traditional CNNs. This is accomplished through a novel compound scaling method that scales the network width, depth, and resolution simultaneously. By carefully balancing these scaling factors, EfficientNet attains state-of-the-art performance across diverse image classification tasks while maintaining computational efficiency. The architecture's ability to achieve remarkable accuracy with fewer parameters makes it well-suited for deployment on devices with limited resources and for real-time applications.

.

XCEPTION:

Xception architecture, leveraging depthwise separable convolution, is adept at serving as an encoder in diverse computer vision tasks like image classification, object detection, and image captioning. Its innovative design allows for efficient feature extraction, making it suitable for resource-constrained environments such as mobile devices. In image classification, Xception excels at capturing hierarchical visual representations, leading to accurate classification. Similarly in object detection, its lightweight nature enables fast and precise detection of objects in images. Moreover, in image captioning, Xception efficiently extracts rich visual features for generating descriptive captions that accurately depict image content. Overall, Xception's versatility and efficiency make it as asset across various computer vision applications.

DECODER:

The decoder serves as a pivotal component in sequence-to-sequence tasks, playing a vital role in generating output sequences based on encoded contextual information. In various applications such as machine translation, summarizing text and captioning images, the decoder network is responsible for producing coherent and contextually relevant sequences. By leveraging the encoded representations provided by an encoder network, the decoder generates output tokens step-by-step, considering both the current input token and the previously generated tokens. This sequential generation process enables the decoder to produce meaningful outputs that reflect the underlying input context. In this literature survey, delve into the intricacies of decoder architectures, exploring their design principles, training methodologies, and applications across diverse domains. By analyzing the latest research and advancements in decoder networks, this survey aims to provide insights into their effectiveness, challenges, and potential avenues for further exploration and optimization in sequence-to-sequence tasks.

LONG SHORT-TERM MEMORY:

In image captioning, the LSTM (Long Short-Term Memory) decoder is responsible for generating a series of words generated from the extracted features of the image. The LSTM decoder is trained to anticipate the subsequent word in the sequence. In practice, use the previously trained CNN for feature extraction and feed these features into the LSTM decoder during the training.

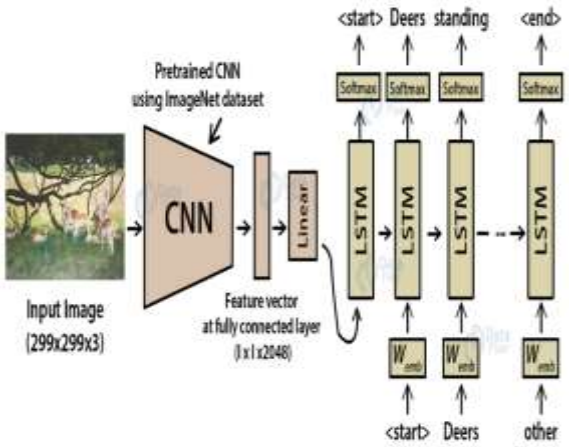

Fig.5: CNN-LSTM MODEL

At inference time, employ the trained LSTM decoder to generate captions for new images by providing the image features and continuously predict the subsequent word until reaching an end token or reaching the maximum sequence length. This decoder will receive the transformed features from the encoder. Then processes this input sequence (initially starting with a special token like "start") and predicts the subsequent word at each time step. At each time step, the LSTM hidden state is modified to capture context and produce sequential output.

TRAINING PHASE:

During the training phase, the Flickr8k dataset will undergo division into training, validation and test sets. Out of 8000 images, common splits include usage of 6,000 images for training, in which each image will be associated with its corresponding set of five captions. 1,000 images will be used for validation and the remaining 1,000 images will be used for testing.

Preprocess the images using the ResNET-50 preprocessing steps, ensuring that they will be compatible with the ResNET-50 model. Tokenize the captions into words and create a vocabulary, then convert the captions into numerical sequences using this created vocabulary and finally truncate the sequences to a fixed length for consistency.

Connect the ResNET-50 output to the LSTM network and design the output layer to anticipate the subsequent word in the sequence. Train the model using the training set. The input will consist of image features extracted by the ResNET-50 model, and the target output will be the caption.

Generate captions for a set of test images and compare them with the ground truth captions. Finally analyze the generated captions, considering factors such as fluency, diversity, and relevance to the images.

EVALUATION:

Evaluating an image captioning project involves assessing the generated captions quality in relation to the ground truth captions. Various metrics are employed to measure the model's performance. Considering the BLEU (Bilingual Evaluation Understudy), which quantifies the similarity between the generated and reference captions based on n-grams (unigrams, bigrams, trigrams, etc.). It is simple and widely used, promoting the generation of precise and varied captions.

In image captioning, BLEU is commonly employed to gauge the alignment between generated captions and reference captions based on n-gram overlap. However, while BLEU offers a quantitative measure, it may not fully capture the overall quality or semantic relevance of captions. Hence, it is often complemented with other metrics for a more comprehensive evaluation.

WEBSITE DESIGN:

Website Design:

- **Layout:** Create a simple webpage with a clear structure.
- **Header:** Incorporate a title to denote the literature survey project.
- **Content Area:** Allocate space for displaying literature entries.
- **Footer:** Add a section at the bottom with essential project details or contact information.

REFERENCE LINKS:

[1]. Shuang Liu, Liang Bai, Yanli Hu, Haoran Wang. "Image Captioning Based on Deep Neural Networks". November 2018. MATEC Web of Conference 232:01052.

[2]. Dr. P. Srinivasa Rao, Thipireddy Pavankumar, Raghu Mukkera, Gopu Hruthik Kiran, Velisala Hariprasad. "Image Caption Generation using Deep Learning Technique", Published By International Research Journal of Modernization in Engineering Technology and Science e-ISSN: 2582-5208. Volume:04/Issue:06/June-2022.

[3]. Aishwarya Maroju, Sneha Sri Doma, Lahari Chandarlapati. "Image Caption Generating Deep Learning Model". Published By International Journal of Engineering Research & Technology(IJERT) ISSN: 2278-0181. Vol. 10 Issue 09, September-2021.

[4]. Muhammad Abdelhadie AI-Malla, Assef Jafar, Nada Ghneim. "Image Captioning model using attention and object features to mimic human image understanding". Journal of Big Data 9, Article number: 20(2022).

[5]. Akash Verma, Arun Kumar Yadav, Mohit Kumar, Divakar Yadav. "Automatic Image Caption Generation Using Deep Learning". January 2022. DOL:10.21203/rs.3.rs-1282936/v1.

[6]. Hossain, M.Z.; Sohel, F.; Shiratuddin, M.F.; Laga, H. "A comprehensive survey of deep learning for image captioning". Article number: 9416431. Journal: IEEE Access, Volume:9, Published-26 Apr 2021.

[7]. R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in:IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 580-587

[8]. Harshitha Katpally and Ajay Bansal. "Ensemble learning on deep neural networks for image caption generation". Proceding – 14th IEEE International Conference on Semantic Computing, ICSC 2020.

[9]. Venugopalan, S.; Anne Hendricks, L.; Rohrbach, M.; Mooney, R.; Darrell, T.; Saenko, K. "Captioning images with diverse objects". In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA.

[10]. Huang, Q.; Smolensky, P.; He, X.; Deng, L.; Wu, D. "Tensor product generation networks for deep NLP modeling". In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), New Orleans, LA, USA.

[11]. Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin and Hamid Laga. "A comprehensive survey of Deep Learning for Image

Captioning". ACM comput. Surv. Article, 0(0):36, 2018.

[12]. Dongming Zhou, Jing Yang, Riqiang Bao. "Collaborative strategy network for spatial attention image captioning". Applied Intelligence, Volume 52, Issue 8, June 2022, pp 9017-9032.