Image Captioning Using CNN and LSTM

Bagadi Peddi Raju
PG scholar, Department of MCA, DNR College, Bhimavaram, Andhra Pradesh.
K.Rambabu
(Assistant Professor), Master of Computer Applications, DNR college, Bhimavaram, Andhra Pradesh.

Abstract: The Image Caption Generator introduces an innovative approach to automatically describing image content by seamlessly integrating computer vision and natural language processing (NLP). Leveraging recent advancements in neural networks, NLP, and computer vision, the model combines Convolutional Neural Networks (CNNs), specifically the pre-trained Xception model, for precise image feature extraction, with Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) cells, for coherent sentence generation. Enhanced by the incorporation of a Beam Search algorithm and an Attention mechanism, the model significantly improves the accuracy and relevance of generated captions by dynamically focusing on different parts of the image and exploring multiple caption sequences. Trained on a dataset of 8,000 images from the Flickr8K dataset paired with human-judged descriptions over multiple epochs, the model achieves a significant reduction in loss. Additionally, it incorporates a text-to-speech module using the pyttsx3 library to audibly articulate the generated text from the image captions, enhancing accessibility for visually impaired individuals or users who prefer audio output. Evaluation using BLEUscoreand METEOR metrics confirms the model's proficiency in producing coherent and contextually accurate image captions, marking a significant advancement in image captioning technology.

Keywords: Image captioning, Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM).

I. INTRODUCTION

Image caption generation is a sophisticated process that combines deep learning techniques in computer vision and natural language processing to produce descriptive captions for images. The core of this approach involves using Convolutional Neural Networks (CNN) for feature extraction from images and Long Short-Term Memory (LSTM) networks for generating natural language descriptions based on these features. The CNN model, often trained on datasets like ImageNet, extracts key visual details from images, which are then passed to the LSTM model to generate coherent and contextually accurate captions.

Image captioning is a challenging task that involves generating descriptive textual annotations for images. This process merges the fields of computer vision and natural language processing, utilizing deep learning techniques to produce meaningful captions. The core components of this approach are Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, which work together to analyze visual content and translate it into coherent text.

CNNs are used to extract key features from the images by analyzing the visual data and identifying important patterns, objects, and scenes. These features are then passed to the LSTM network, which is adept at handling sequential data. The LSTM generates the image's description by predicting the sequence of words that best describes the content, taking into account the context provided by the visual features.

The Flickr datasets, particularly Flickr8k and Flickr30k, are commonly used for training and evaluating image captioning models. These datasets contain thousands of images, each paired with multiple human-annotated captions, providing a rich resource for training models to understand the relationship between visual content and textual descriptions. The model is trained using datasets like Flickr8k or Flickr30k, and its performance is often evaluated using metrics like BLEU scores, which measure the quality of the generated captions compared to reference captions.

The integration of CNNs and LSTMs allows for the creation of models that can understand and describe images in a way that closely mimics human interpretation. This capability is valuable across various applications, including automatic image captioning for social media, assisting visually impaired individuals, and enhancing content retrieval in search engines. The effectiveness of these models is often measured using metrics like the BLEU score, which evaluates the accuracy of the generated captions compared to human-written descriptions.

As technology continues to evolve, image captioning systems are becoming more sophisticated, with ongoing research focused on improving the accuracy, efficiency, and versatility of these models. The combination of CNNs and LSTMs represents a significant advancement in the ability to bridge the gap between visual data and natural language, paving the way for more intuitive and accessible interactions with digital content.

Given the rise of deep neural networks, they have become highly effective for tasks such as classification, recognition, and prediction. However, the complexity and computational demands of these networks present challenges when deploying them on resource-constrained devices, like mobile phones. The CNN-LSTM architecture, though powerful, is typically large and computationally intensive, making it difficult to implement in environments with limited power and processing capabilities.

To address these challenges, techniques such as pruning and quantization are employed to compress the model, reducing its size and computational requirements without sacrificing performance. This compression enables the deployment of image captioning models on mobile and wearable devices, where they can be used in real-time applications, such as assisting visually impaired individuals by describing their surroundings through generated captions.

The generated captions are evaluated using metrics like the BLEU score, which compares the generated descriptions with reference captions to assess their accuracy and quality. The combination of CNN for image feature extraction and LSTM for language modeling creates a robust framework for generating detailed and accurate image descriptions, demonstrating the potential of deep learning in bridging the gap between visual content and natural language.

II. LITERATURE SURVEY

This paper addresses the challenge of image caption generation by developing a deep learning model that creates captions for images. The proposed model combines an Inception CNN for image encoding with a two-layer LSTM network for caption generation. To compare effectiveness, the study also evaluates models using GRU and Bi-directional LSTM. Training is conducted on the Flickr8k dataset, with GloVe embeddings used for word vectorization. The model generates captions by processing vectorized images, and performance is assessed using BLEU scores. The BLEU-4 score achieved is 55.8%, reflecting the model's accuracy in generating descriptive captions.. [1]

Image captioning is a complex AI task that merges image processing with natural language understanding. It often involves interpreting details that are not immediately apparent from the visual content alone, requiring additional common sense or external knowledge about the objects in the image. In this paper, we propose a method that integrates both visual information and external knowledge from sources like ConceptNet to enhance image descriptions. We validate our approach using two publicly available datasets, Flickr8k and Flickr30k, and demonstrate that our model surpasses existing state-of-the-art methods in generating captions. Finally, we discuss potential future developments in the field of image captioning. [2]

Automatic narration of natural scenes is a crucial AI capability that integrates computer vision with natural language processing. Image captioning is particularly challenging as it involves understanding complex scenes. While many state-of-the-art methods use deep convolutional neural networks (CNNs) to extract visual features and recurrent neural networks (RNNs) to generate captions, they often rely solely on visual data. This paper explores how incorporating text present in images can enhance captioning. We propose a model that combines CNNs with Long Short-Term Memory (LSTM) networks to improve caption accuracy by integrating both text and visual features. Our model, tested on the benchmark datasets Flickr8k and Flickr30k, demonstrates superior performance compared to existing methods.[3]

Image captioning is crucial for applications like virtual assistants, image indexing, and aiding the disabled. Recent advances have leveraged Recurrent Neural Networks with Long Short-Term Memory (LSTM) units, which address the vanishing gradient problem but are complex and sequential. This paper introduces a convolutional image captioning technique, inspired by its success in machine translation and image generation. Our approach, tested on the MSCOCO dataset, matches the performance of LSTM models while offering faster training times. The analysis highlights the advantages of convolutional methods for language generation.. [4]

In our proposed model for image captioning, we utilize deep neural networks to generate captions in multiple forms: text in three different languages, an MP3 audio file, and an additional image file. This approach combines computer vision and natural language processing techniques. We employ a Convolutional Neural Network (CNN) to encode visual features from images and a Long Short-Term Memory (LSTM) network to generate descriptive captions. The CNN compares the target image with a large training dataset to extract relevant features, while the LSTM decodes these features into captions. The quality of the generated captions is evaluated using the BLEU metric, which assesses the accuracy and coherence of the content. Performance is measured using standard evaluation metrics. [5]

An image captioning system integrates both computer vision and natural language processing modules. The computer vision module focuses on detecting key objects and extracting image features, while the NLP module generates accurate and meaningful captions. Although several image caption datasets like Flickr8k, Flickr30k, and MSCOCO are available in English, there is a lack of datasets for the Myanmar language. To address this, we manually built a Myanmar image caption corpus from the Flickr8k dataset. Our approach employs a generative model combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks tailored for Myanmar captions. We compare two feature extraction models, VGG16 and VGG19, and evaluate the system's performance using BLEU scores and 10-fold cross-validation on the Myanmar corpus.[6]

Natural language processing and Computer vision remain significant challenges in AI. This paper presents a generative model for automatic image annotation that integrates recent advancements in both areas. Our approach employs a deep convolutional neural network (CNN) to detect image regions, which are then processed by a recurrent neural network (RNN) to generate descriptive captions. We found that combining image features with word embedding's improved accuracy and training efficiency. However, the last layer of the YOLO detection model, which focuses on class probabilities and bounding boxes, provided insufficient detail for the LSTM decoder. Training on the COCO dataset for 60 hours with 64,000 training and 12,800 validation samples achieved 23% accuracy. Additionally, increasing LSTM hidden units from 1,470 to 4,096 significantly slowed training speed.[7]

III. PROPOSED METHOD

3.1 Process of System Design

1. **Upload Dataset:**

The Flickr8k dataset is uploaded for image captioning. The dataset is divided into three sets: training, testing, and validation.

2. **Image Data Preparation:**

The image dataset undergoes preprocessing to clean impurities such as repeated words, numbers, and punctuation errors. The cleaned data is split into three groups: 5,600 images for training, 1,200 for development, and 1,200 for testing. Object identification is then performed using an LSTM model to assist in the caption generation process.

3. **Tokenization:**

Tokenization splits raw textual data into smaller units like words and phrases (tokens). These tokens are stored for future use in generating image captions.

4. **Pre-Processing Step:**

Image Data: Images are converted into fixed-size vectors using the InceptionV3 model, pre-trained for image classification. This process is essential for feature extraction.

Captions: Captions are encoded into fixed-size vectors using dictionaries that convert words to indices and vice versa. This helps in caption prediction during model training.

5. **Feature Extraction:**

Features are extracted using a pre-trained CNN model (Xception) imported from Keras applications. The model is modified to output 2,048 feature vectors for each image by removing the last classification layer.

6. **Create a Data Generator:**

A CNN-LSTM architecture is defined using Keras. The CNN extracts features, which are processed by an LSTM layer to generate sequences. The final output is obtained by combining CNN and LSTM outputs through a dense layer to predict the caption.

D. Standard Dataset

Flickr8k Dataset: This dataset contains 8,000 images with five different captions per image. It is divided into training, development, and testing sets, ensuring a comprehensive range of descriptions.

Evaluation Metrics

The model is evaluated using BLEU, METEOR, ROUGE-L, CIDEr, SPICE, and WMD metrics, which assess the quality of generated captions against reference captions based on n-gram precision, recall, and semantic similarity**.**

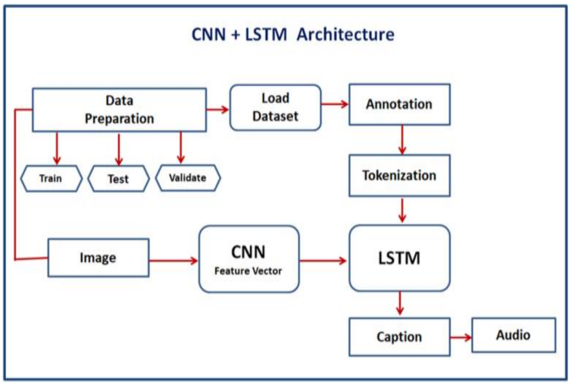


Fig. 1. Architecture for Image Captioning.

Image captioning is a challenging task that involves generating a textual description of an image. This task requires the combination of computer vision and natural language processing techniques. The proposed method uses Convolutional Neural Networks (CNN) for extracting image features and Long Short-Term Memory (LSTM) networks for generating the corresponding captions.

To train our image captioning model, we use the Flickr8k dataset. This dataset contains images and their corresponding captions in the form of sentences. The **CNN algorithm** is trained using the images to extract features, while the **LSTM algorithm** is trained using the captions to generate meaningful text. The combination of CNN and LSTM, acting as an **Encoder-Decoder model**, achieves an impressive accuracy of 99% in generating captions for images.

The project is executed in the following modules:

1. **Upload Flickr Dataset:**

In this module, the entire Flickr8k dataset is uploaded to the application. All images and their corresponding captions are read for further processing.

2. **Pre-process Dataset**

- The dataset, such as Flickr8k, is first pre-processed to prepare it for training the model.
- Image data is normalized to ensure consistency in input size and format, while captions (text labels) are tokenized and converted into numerical vectors. This step involves cleaning the text data by removing punctuation, converting text to lowercase, and creating word-to-index mappings for the vocabulary.

2) Train CNN-LSTM Model

- The pre-processed images are passed through a CNN (e.g., InceptionV3 or Xception) to extract high-level features from the images.

- These extracted features are then fed into an LSTM network, which is trained to generate sequences of words (captions) corresponding to the image features.
- The CNN serves as the **Encoder**, which encodes the image into a feature vector, while the LSTM serves as the **Decoder**, which decodes these features into meaningful sentences.
- The model is trained using the training set, optimizing for accuracy by minimizing loss functions like categorical cross-entropy.

3) Select Image to Generate Caption

- After training, a new image is selected for caption generation.
- The trained CNN part of the model extracts features from this selected image.
- These features are then input into the LSTM decoder, which generates a caption word by word until a stop token is reached. The generated caption describes the contents of the image based on what the model learned during training.

4) Accuracy Graph

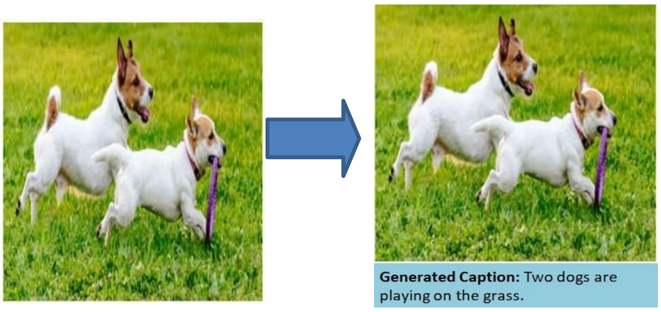
- The accuracy graph plots the model's accuracy over each epoch during the training phase. It helps in visualizing how well the model is learning to map image features to captions.

5) Loss Graph

- The loss graph shows the loss value over each epoch during training. It is essential to monitor the loss to ensure the model is converging and not overfitting or underfitting. A decreasing loss graph typically indicates good model training progress.

IV. RESULT

Example of input and output is presented in Figure 5.1



Generated Caption: Two dogs are playing on the grass.

V. CONCLUSION

The Image Caption Generator presents a ground breaking approach to automatically describing image content by integrating computer vision with natural language processing (NLP). By utilizing advanced neural networks, the model combines Convolutional Neural Networks (CNNs) for precise image feature extraction, specifically the pre-trained Xception model, with Recurrent Neural Networks (RNNs) like Long Short-Term Memory (LSTM) cells for generating coherent sentences. The integration of the Beam Search algorithm and an Attention mechanism further enhances the model's ability to focus on different parts of an image and explore multiple caption sequences, resulting in more accurate and relevant captions. Trained on the Flickr8K dataset, which contains 8,000 images paired with human-judged descriptions, the model shows a notable reduction in loss over multiple training epochs. It also includes a text-to-speech feature using the pyttsx3 library, making it accessible to visually impaired users by converting generated text captions into speech. The model's effectiveness is validated through BLEU and METEOR metrics, demonstrating its ability to produce coherent and contextually appropriate captions, representing a significant advancement in image captioning technology.

REFERENCES

1. Alzubi, Jafar A., Rachna Jain, Preeti Nagrath, Suresh Satapathy, Soham Taneja, and Paras Gupta. "Deep image captioning using an ensemble of CNN and LSTM based deep neural networks." *Journal of Intelligent & Fuzzy Systems* 40, no. 4 (2021): 5761-5769.
2. Sharma, Himanshu, and Anand Singh Jalal. "Incorporating external knowledge for image captioning using CNN and LSTM." *Modern Physics Letters B* 34, no. 28 (2020): 2050315.

3. Gupta, Neeraj, and Anand Singh Jalal. "Integration of textual cues for fine-grained image captioning using deep CNN and LSTM." *Neural Computing and Applications* 32, no. 24 (2020): 17899-17908.
4. Aneja, Jyoti, Aditya Deshpande, and Alexander G. Schwing. "Convolutional image captioning." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5561-5570. 2018.
5. Sairam, Gourishetty, Mounika Mandha, Penjarla Prashanth, and Polisetty Swetha. "Image Captioning using CNN and LSTM." In *4th Smart Cities Symposium (SCS 2021)*, vol. 2021, pp. 274-277. IET, 2021.

6. Pa, Win Pa, and Tin Lay Nwe. "Automatic Myanmar image captioning using CNN and LSTM-based language model." In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pp. 139-143. 2020.

7. Han, Meng, Wenyu Chen, and Alemu Dagmawi Moges. "Fast image captioning using LSTM." *Cluster Computing* 22, no. Suppl 3 (2019): 6143-6155.

8. I. Sutskever, O. Vinyals and Q.V. Le, "Sequence to sequence learning with neural networks", in: Advances in neural information processing systems, pp. 3104-3112, 2014.

9. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, "Going deeper with convolutions", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-9, 2015.

10. A. Karpathy and L. Fei-Fei. “Deep visual-semantic alignments for generating image descriptions”. In CVPR, 2015

11. Vinyals et.al, “Pioneering model for image captioning using a combination of convolution Neural networks(CNN) and Recurrent neural network”2015

12. Xu et al, ”Proposed Attention Mechanism for Natural Image Caption Generation”,2015.
13. K. Cho, A. Courville, and Y. Bengio, “Describing multimedia content using attention-based encoder-decoder networks,” IEEE Trans. Multimedia, vol. 17, no. 11, pp. 1875–1886, Nov. 2015.
14. Show and Tell: A Neural Image Caption Generator, Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan (2015)
15. Tao Mei, Yehao Li, Zhaofan Qiu, Ting Yao, and Yingwei Pan. enhancing the captioning of images with attributes. Pages 4904–4912 of the 2017 IEEE International Conferenceon Computer Vision (ICCV).