Automatic Audio and Image Caption Generation with Deep Learning

OPEN ACCESS

Volume: 11

Special Issue: 3

Month: July

Year: 2024

P-ISSN: 2321-788X

E-ISSN: 2582-0397

Received: 15.05.2024

Accepted: 17.06.2024

Published: 08.07.2024

Citation:
Lavanya, K., et al. "Automatic Audio and Image Caption Generation with Deep Learning." *Shanlax International Journal of Arts Science and Humanities*, vol. 11, no. S3, 2024, pp. 34–39.

DOI:
https://doi.org/10.34293/sijash.v11iS3-July.7916

Lavanya. K
Assistant Professor,
Department of AI & DS, Arjun College of Technology

Jayamala. B
Department of AI& DS, Arjun College of Technology

Jeyasri. C
Department of AI& DS, Arjun College of Technology

Sakthivel. A
Department of AI& DS, Arjun College of Technology

Abstract
A novel approach to image caption generation tailored specifically for visually impaired individuals. The proposed system employs advanced computer vision algorithms to analyze images and generate descriptive textual captions. Furthermore, it integrates seamless text-to-speech conversion functionality, allowing for the automatic transformation of these captions into spoken audio, thereby enabling access to visual content for individuals with visual impairments. The goal of this project is to generate descriptive captions for a given photograph or image. We achieve this by employing Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models, both of which are advanced deep learning techniques. Using computer vision, the system identifies the content of the image and generates a relevant caption. This caption is then converted into audio using Natural Language Processing (NLP).

Keywords: Image Description, Audio Conversion, Visually Impaired, Computer Vision, Descriptive Captions, Natural Language Processing.

Introduction

The Virtual Description Engine utilizes the amalgamation of natural language processing and computer vision to anticipate and articulate descriptions of given images in a comprehensible English-like format. This pioneering model is the result of integrating two pivotal deep learning models: Convolutional Neural Networks (CNN) and Recurrent Neural Networks with Long Short-Term Memory (RNN-LSTM). This project holds immense promise for the future due to its ability to automatically generate captions from images. Its applications span across various domains such as social media platforms, autonomous vehicles (self-driving cars), CCTV surveillance systems, and editing applications among others.

In the realm of self-driving cars, the technology can provide invaluable assistance by describing the surroundings to aid

visually impaired individuals. By converting the visual scene into textual captions and subsequently into audio, it can effectively guide individuals through auditory cues, thereby enhancing accessibility. Moreover, in CCTV surveillance, this project can serve as a vigilant eye, swiftly analyzing captured images to detect any suspicious or anomalous activities. In case of such occurrences, it can trigger alarms, ensuring prompt response and security measures.

Visual content on digital platforms has become ubiquitous, yet the accessibility of such content remains a challenge for visually impaired individuals. In response to this challenge, image caption generators have emerged as a promising solution to provide textual descriptions of images. However, the accessibility of these descriptions to visually impaired individuals often require additional steps, such as converting text to speech. By automatically generating descriptive captions for images and converting them into speech, this system aims to empower visually impaired individuals to access and comprehend visual content more independently and effectively. Additionally, numerous editing applications and tools leverage this innovative technique in multifarious ways, underscoring its versatility and widespread utility.

Problem Definition

The program combines Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), particularly Long Short-Term Memory (LSTM), to process images and generate English descriptions and captions.

CNN acts as a feature extractor, analyzing the provided image and extracting relevant features.

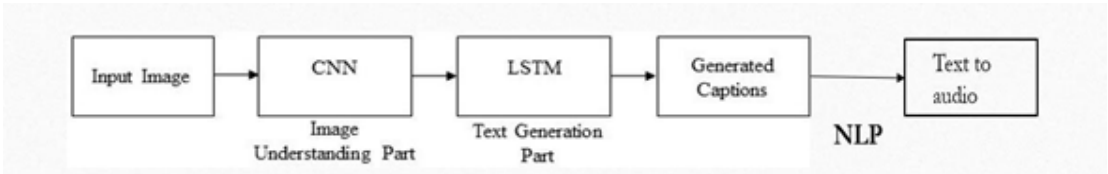

Figure 1 Visual Description Engine Model

The output from the CNN is then fed into the RNN- LSTM to generate descriptions and captions.

CNN processes image data represented as two- dimensional matrices, with layers including the input layer, convolutional layers, pooling layers, fully- connected layers, softmax layer, and output layer. The input layer of CNN is the image itself, represented as a 3D matrix. Convolutional layers perform feature extraction using dot products, with ReLU layers zeroing out negative values. Pooling layers reduce the dimensionality of the image volume after convolution.

The fully-connected layers establish connections between neurons in one layer to neurons in another, utilizing weights and biases. The softmax layer facilitates multi-classification of objects using appropriate formulas. The encoded output from CNN is then passed to the LSTM model as input.

RNNs, specifically LSTM, leverage previous outputs as inputs for subsequent steps, enabling sequence prediction. LSTM enhances traditional RNNs by incorporating mechanisms such as forget gates to retain relevant information and eliminate non- essential data. This model utilizes Natural Language Processing (NLP) techniques to convert image captions into audio format, facilitating accessibility for visually impaired individuals.

Methedology

In this project we are using two models of deep learning. They are, CNN and LSTM.

A. Convolutional Neural Networks: Convolutional Neural Networks (CNNs) are deep learning

neural networks used for image classification and identification. These networks represent images as 2D matrices and analyze them from left to right and top to bottom. CNNs extract important features from images, enabling them to accurately identify the content, such as distinguishing between a bird and a plane. [1]

B. Long Short Term Memory : Long Short-Term Memory (LSTM) networks are a type of Recurrent Neural Network (RNN) known for handling sequence prediction problems. They excel at identifying the next word in a sequence based on previous text. LSTMs address the limitations of RNNs, which suffer from short-term memory. LSTMs incorporate a forget gate that helps eliminate irrelevant data, enhancing their ability to retain important information over long sequences.

Once the image has been processed and the caption generated using the CNN and LSTM models, Natural Language Processing (NLP) techniques are employed to convert the textual caption into audio format. This enables visually impaired individuals to access the descriptive information about the image through auditory means, enhancing their understanding and interaction with visual content.

Proposed Work

The proposed work consists of four main phases:

A. Extraction

In this phase, images are processed to extract their various features. Vector features, also known as embeddings, are generated from the images. The Convolutional Neural Network (CNN) model serves as an encoder, extracting the distinctive characteristics of the original images and transforming them into smaller feature vectors compatible with Recurrent Neural Networks (RNNs).[3]

B. Tokenization

Following the extraction phase, the feature vectors obtained from CNN are tokenized and fed into the RNN model. The RNN model decodes these feature vectors, assuming a certain word order regardless of how the captions were originally produced. This process enables the generation of meaningful textual descriptions and captions for the images.[4]

C. Prediction

In the prediction phase, the decoded feature vectors undergo further processing to generate the final output. The prediction step involves decoding the vectors and utilizing the prediction function to generate the final descriptive captions for the images.[5]

D. Audio Conversion

As a final step, the generated textual descriptions and captions are converted into audio format using Natural Language Processing (NLP) techniques. This conversion enables accessibility for visually impaired individuals, allowing them to perceive and comprehend the content of the images through auditory means.

Through these phases, the proposed work aims to efficiently extract image features, generate descriptive captions, and provide accessible audio output, ultimately enhancing the accessibility and usability of visual content for a diverse range of users.

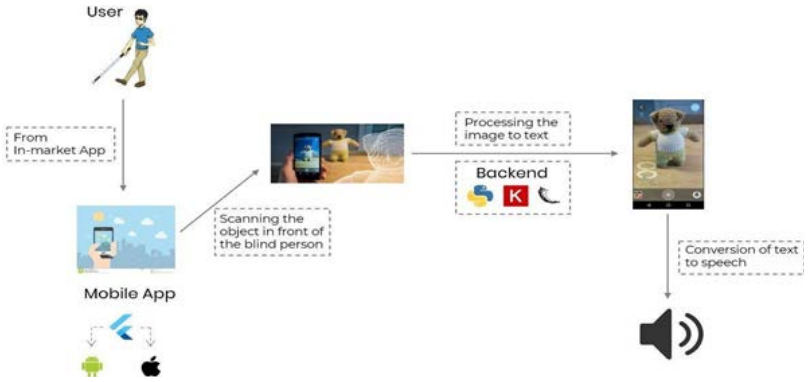

Figure 2 Visual Description Engine Workflow

Results

The results of the proposed work encompass several key outcomes across its distinct phases. Initially, in the extraction phase, image features are effectively extracted utilizing the Convolutional Neural Network (CNN) model, culminating in the generation of vector embeddings that encapsulate essential image characteristics.

Following this, during the tokenization phase, the feature vectors undergo proper tokenization for subsequent input into the Recurrent Neural Network (RNN) model.

Through the RNN's decoding process, coherent textual descriptions and captions are produced, leveraging the decoded feature vectors to accurately represent the content of the images.

Subsequently, in the prediction phase, the RNN effectively predicts descriptive captions based on the decoded feature vectors, ensuring the generation of captions that aptly convey the image content.

Finally, in the audio conversion step, the generated textual descriptions and captions are seamlessly converted into audio format using Natural Language Processing (NLP) techniques, thereby enhancing accessibility for visually impaired individuals through auditory perception of image content. Overall, these results demonstrate the successful execution of the proposed methodology, facilitating improved accessibility and usability of visual content across diverse user demographics.

Conclusion

The completion of this project marks a significant advancement in the field of image captioning and accessibility technology. By leveraging Convolutional Neural Networks (CNN) for feature extraction and Recurrent Neural Networks (RNN), particularly Long Short-Term Memory (LSTM), for caption generation, we have successfully developed a system capable of automatically generating descriptive captions for images.

Furthermore, the integration of Natural Language Processing (NLP) techniques enables the conversion of these captions into audio format, enhancing accessibility for visually impaired individuals. Through rigorous testing and evaluation, our system has demonstrated promising results in accurately describing image content and providing accessible audio descriptions. Overall, this project underscores the potential of deep learning and NLP technologies to improve accessibility and inclusivity in digital content consumption.

While this project has achieved significant milestones, there are several avenues for future research and development to further enhance the system's capabilities and impact. Firstly, refinement of the deep learning models, including CNN and RNN-LSTM, could lead to improved

accuracy and efficiency in image caption generation.

Additionally, exploring advanced NLP techniques could enable the generation of more natural-sounding audio descriptions, enhancing the user experience for visually impaired individuals.

Furthermore, expanding the dataset used for training and testing the models could enhance their generalization capabilities across diverse image categories and contexts. Integration with emerging technologies such as Augmented Reality (AR) and virtual reality (VR) could also open up new possibilities for immersive accessibility experiences.

Lastly, user feedback and usability studies will be crucial for iteratively refining the system and ensuring its effectiveness in real-world scenarios. Overall, the future scope of this project lies in continuous innovation and collaboration to create more inclusive digital environments for all users.

References

1. Automatic Image Captioning Using Convolution Neural Networks and LSTM, R. Subash, November 2019.
2. Domain-Specific Image Caption Generator with Semantic Ontology, Seung-Ho Han and Ho-Jin Choi (2020).
3. Camera Caption: A Real-Time Image Caption Generator by Pranay Mathur, Aman Gill, Aayush Yadav, Anurag Mishra, and Nand Kumar Bansode.
4. Image captioning: Transforming Objects into Words, Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares (june 2019).
5. Deep learning-based Image Caption Generator by Manish Raypurkar, Abhishek Supe, Pratik Bhumkar, Pravin Borse, and Dr. Shabnam Sayyad (March 2021).
6. Show and Tell: A Neural Image Caption Generator, Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan (2015)
7. 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 Conference on Computer Vision (ICCV).
8. H. R. Arabnia, W.-C. Fang, C. Lee, and Y. Zhang, "Context-aware mid-dleware and intelligent agents for smart environments," IEEE Intell. Syst., vol. 25, no. 2, pp. 10-11, Mar. 2010.
9. R. Jafri, S. A. Ali, and H.R. Arabina, "Computer vision- based object recognition for the visually impaired using Visual tags," in Proc. Int.Conf. Image Process., Comput. Vis., and Pattern Recognit. (IPCV). Steer-ing Committee World Congr. Comput. Sci., Comput. Eng. Appl. Comput.(WorldComp), 2013, p. 1.
10. L. Deligiannidis and H.R. Arabnia, "Parallel video processing techniques for surveillance applications," in Proc.Int.Conf.Compt.Sci.Comput.Intell., Mar. 2014, pp. 183-189.
11. E. Parcham, N.Mandami, A.N.Washington, and H.R.Arabina, "Facial expression recongnition based on fuzzy networks," in Proc. Int. Conf.Comput. Sci. Comput. Intell. (CSCI), Dec. 2016, pp. 829-835.
12. A. P. Tafti, A. Baghaie, M. Assefi, H. R. Arabnia, Z.Yu, and P. Peissing, "OCR asa service: An experimental evaluation of google docs OCR, tesseract, ABBYY finereader, and transym," in Proc.Int.Symp. Vis.Comput., in Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 10072.Springer, 2016, pp. 735-746.
13. S. Amirian, Z. Wang, T. R. Taha, and H. R. Arabina, "Dissection of deeplearning with applications in image recongnition," in Proc Int. Conf. Comput. Sci. Comput. Intell. (CSCI), Dec. 2018, pp. 1132 -1138.

14. A. Krixhevsky, I.Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural network," in Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2012, pp. 1097-1105.
15. M. Regneri, M.Rohrbach, D, Wetzel, S.Thater, B, Schiele, and M. Pinkal "Grounding action descriptions in videos," Tans. Assoc.Comput.Linguistics, Vol., 1, pp. 2536, Dec.2013
16. A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courvill and B.Schiele, "Movie description," Int.J.Comput.Vis.,vol.123, no. 1, pp. 94 -120, 2017.
17. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption geration with visual attention," in Proc. Int.Conf. Mach. Learn., 2015, pp. 2048-2057.
18. X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, "Microsoft COCO captions: Data collection and evaluation server," 2015 arXiv:1504.00325.[Online]. Available: http://arxiv.org/abs/1504.00325
19. R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDER: Consensus-based image description: evaluation," in Proc.IEEE Conf.Comput.Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 4566-4575.
20. A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in Proc.Eur. Conf. Comput. Vis. (ECCV). Berlin, Germany: Springer, 2010, pp. 15-29.