Extracting Keywords from Images Using Deep Learning for the Visually Challenged

Said Jaboob
University of Technology & Applied Sciences Salalah, Sultanate of Oman.

Munes Singh Chauhan
University of Technology & Applied Sciences Salalah, Sultanate of Oman.

Balaji Dhanasekaran
University of Technology & Applied Sciences Salalah, Sultanate of Oman.

Senthil Kumar Natarajan
University of Technology & Applied Sciences Salalah, Sultanate of Oman.

Abstract: Assistive technologies can in many ways facilitate the normal day-to-day lives of the disabled. As part of the ongoing research on assistive technologies at UTAS, Oman, that deals with augmenting and finding multimodal aspects of applications for the disabled, this paper aspires to investigate the role of deep learning in the field of image interpretation. Images are one of the most important mediums of conveying information among humans. Visually impaired persons especially with low cognitive abilities face insurmountable difficulties in understanding cues through images. This challenge is met by filtering words from image captions to facilitate understanding of the key notion conveyed by an image. This work utilizes the image captioning technique using deep learning frameworks such as convolution neural networks (CNN) and recurrent neural networks (RNN) to generate captions. These captions are fed to Rake, an NLP library that identifies keywords in the caption. The entire process is automated and uses transfer learning techniques for caption generation from images. This process is then further integrated with our main project, Finger Movement Multimodal Assistive System (FMAS) thereby incorporating text cues for interpreting images for the visually impaired.

Keywords: Finger Movement Multimodal Assistive System (FMAS), Visually impaired, Convolution neural networks (CNN)

Citation: Jaboob, S., Chauhan, M.S., Dhanasekaran, B., & Natarajan, S.K. (2022). Extracting Keywords from Images Using Deep Learning for the Visually Challenged. In A. Ben Attou, M. L. Ciddi, & M. Unal (Eds.), *Proceedings of ICSES 2022-- International Conference on Studies in Education and Social Sciences* (pp.554-561), Antalya, Türkiye. ISTES Organization.

Introduction

The disabled especially those with visual impairments, have a hard time understanding and interpreting images. This inability to comprehend images leads to the shrinking of a vast segment of visual cognitive abilities that otherwise would have been avoidable. Furthermore, some of the visually challenged also have prior cognitive impairments that further accentuate their handicap. This will also adversely affect their kinesis and response to particular situations in day-to-day life. Cognitive limitations among the visually impaired are seen equally among children as well as old people, especially those suffering from Parkinson's disease [1,2]. Various methodologies have been tried but explaining images has been for a long a logical challenge despite the availability of the latest technologies.

Image captioning is the process of generating a text description of a given image. This can be suitable modified for the visually incapacitated people with limited or no agility/ perceptibility, and thus can play a major role for them in dealing with the external world using our FMAS system. Image captioning lies at the confluence of computer vision, deep learning, and natural language processing. The FMAS is powered by GPU hardware which runs deep learning algorithms and possess inbuilt functionality to provide feedback to the FMAS via cloud. A recent work [3] uses a generative model [4] and RNN (recurrent neural network) [5] to generate captions. This allows for automatic caption generation with an accuracy of BLEU-1 score of 59 taken on the experiments on the Pascal dataset. Since the research work is ongoing we expect more robust and accurate models for image captioning in the near future. One of the limitations of such models is that they have to be trained on large image datasets which adds to the training time. Inherently, these models employ object detection and image segmentation to identify key contents in an image.

Keyword extraction is studied generally to summarize texts or emails to identify the specific category to which the text purports to. Keyword extraction mainly involves removing stop words ("is", "the", etc.) from a given text. Further, words are statistically chosen based on an information retrieval tool *tf-idf* that calculates the significance of a word based on its importance in a chosen corpus.

The focus of this work is on the keyword extraction which is an important component of the FMAS. The other FMAS system components include GSM and Bluetooth connectivity that sends images to the GPU backend hardware further processing. The final results are communicated to differently abled persons using the semi-automated devices like wheelchair or through output devices such as speaker or monitor.

The Proposed Method

The proposed solution is based on the utilization of image captioning based on transfer learning. The CNN model is trained on the ImageNet dataset and thus is capable of identifying specific objects in various types of images seen in daily life. Our image captioning model is a type of encoder-decoder model based on visual-space

methodology instead of using modal representations as in some other cases. The architecture of the model is a CNN-RNN combination [6] as shown in Figure 1.

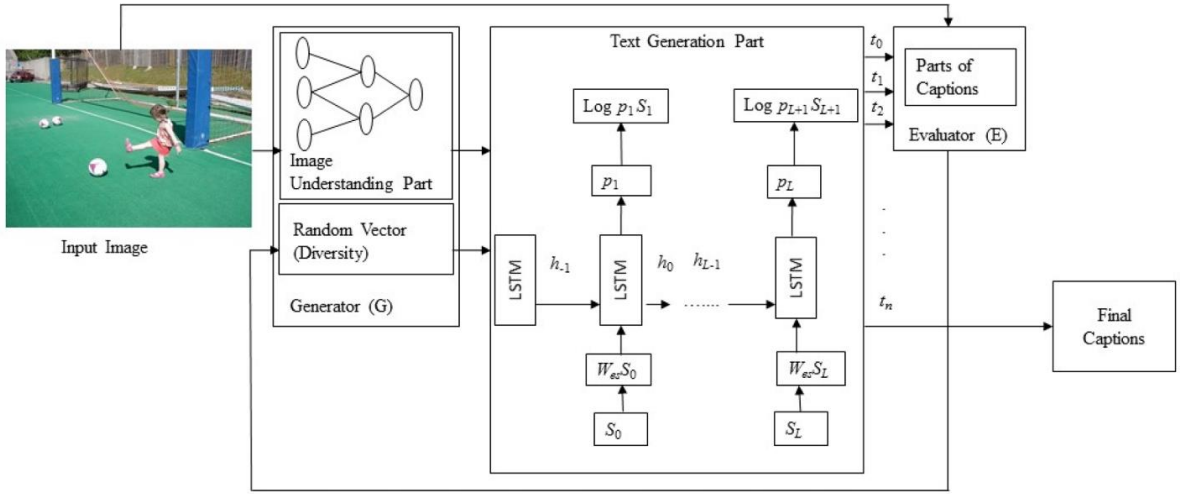

Figure 1. A CNN-RNN Paired Model for Image Captioning [6]

The second stage of the proposed network uses LSTM (Long Short-Term Memory) RNN model for generating captions using Glove word embeddings. This allows for the generation of captions. The method for generating caption is not very robust but in most cases, it helps in identifying key objects in an image which is sufficient in our case. We further input our generated captions to an existing NLP library, Rake, which further zeroes in on the prominent keywords. Our overall model is depicted in Figure 2.

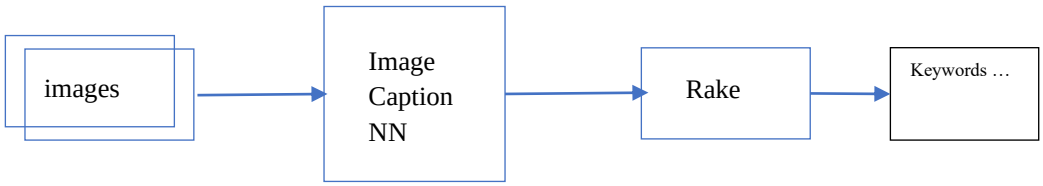

Figure 2. Keyword Generator NN from Images

Our image captioning CNN model is based on transfer learning. We train our model on a pre-trained InceptionV3 CNN model [7] trained on the ImageNet dataset. We slice the top layer of the model in order to customize it and fit our RNN (LSTM) model. The main idea is to capture the latent representation [8] of image features and pass it on to the next RNN network.

Our captions are sourced from a vocabulary provided by GloVe encodings [9]. This allows for access to a pretrained unsupervised learning algorithm that represents words in terms of vectors. These vectors are calculated

based on the nearest neighbours principle thus words with similar meanings are closer in terms of Euclidean distance (using vectors) than those which are not. We limit our embedding dimensions to 200 for the sake of fast convergence. The RNN network that deals with word embeddings is described in Figure 3.

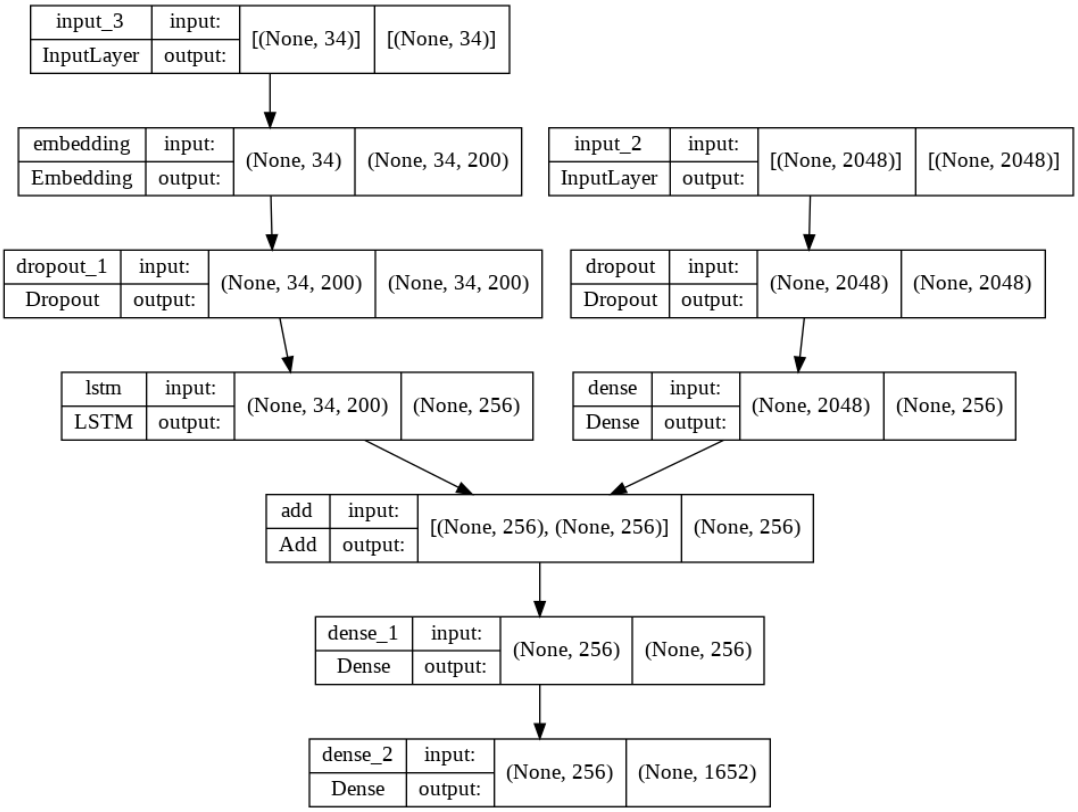

Figure 3. The Caption Generator RNN Model

Some of the outputs of the generated captions are shown in Figure 4.

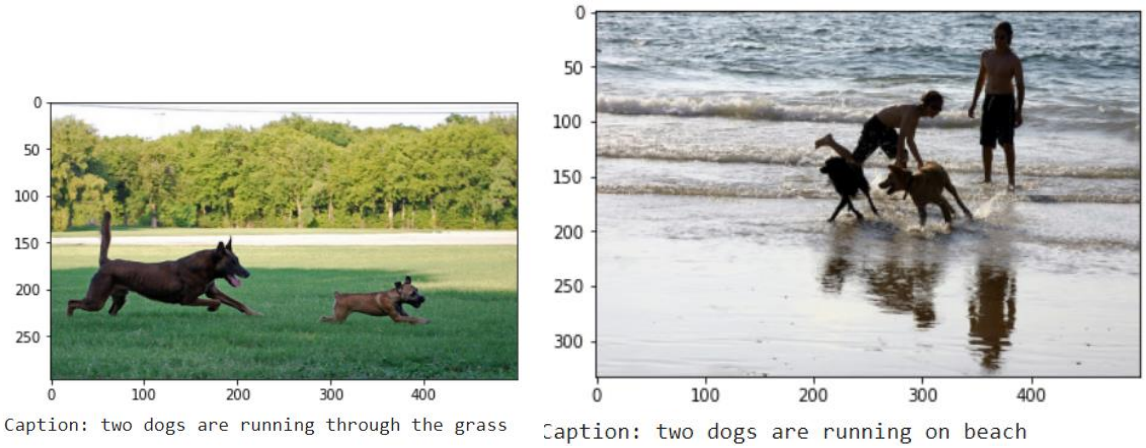

Figure 4: Generated Captions

The captions generated are then fed to the NLP Rake library which outputs the critical keywords. As mentioned before, the visually challenged with some sort of cognitive limitations can fathom quite easily an image about its context and meaning. Figure 5 provides the keywords generated from a given image.

Figure 5. Caption with Accompanying Keywords Generated by Rake

Performance

We use BLEU (Bilingual Evaluation Understudy) metric [10] for comparing the generated captions with given training labels. As each training image is provided with 5 descriptions, we analyze our BLEU score with our generated caption. BLEU score varies between 0 and 1, where 1 is considered the optimal match and 0 is a complete mismatch. The score uses various sizes of n-grams and then summarizes the score using a geometric mean.

The general formula of the BEU score calculations is given as under:

$$exp\left(1 - \frac{\text{lengh of reference}}{\text{length of hypothesis}}\right)$$

Figure 6 shows the BLEU score for each generated caption for some samples. It is clearly seen that our model

provides on average 70% accuracy in most cases when sampled over the training and test data. In actual use, the efficiency may fall to 20% on average.

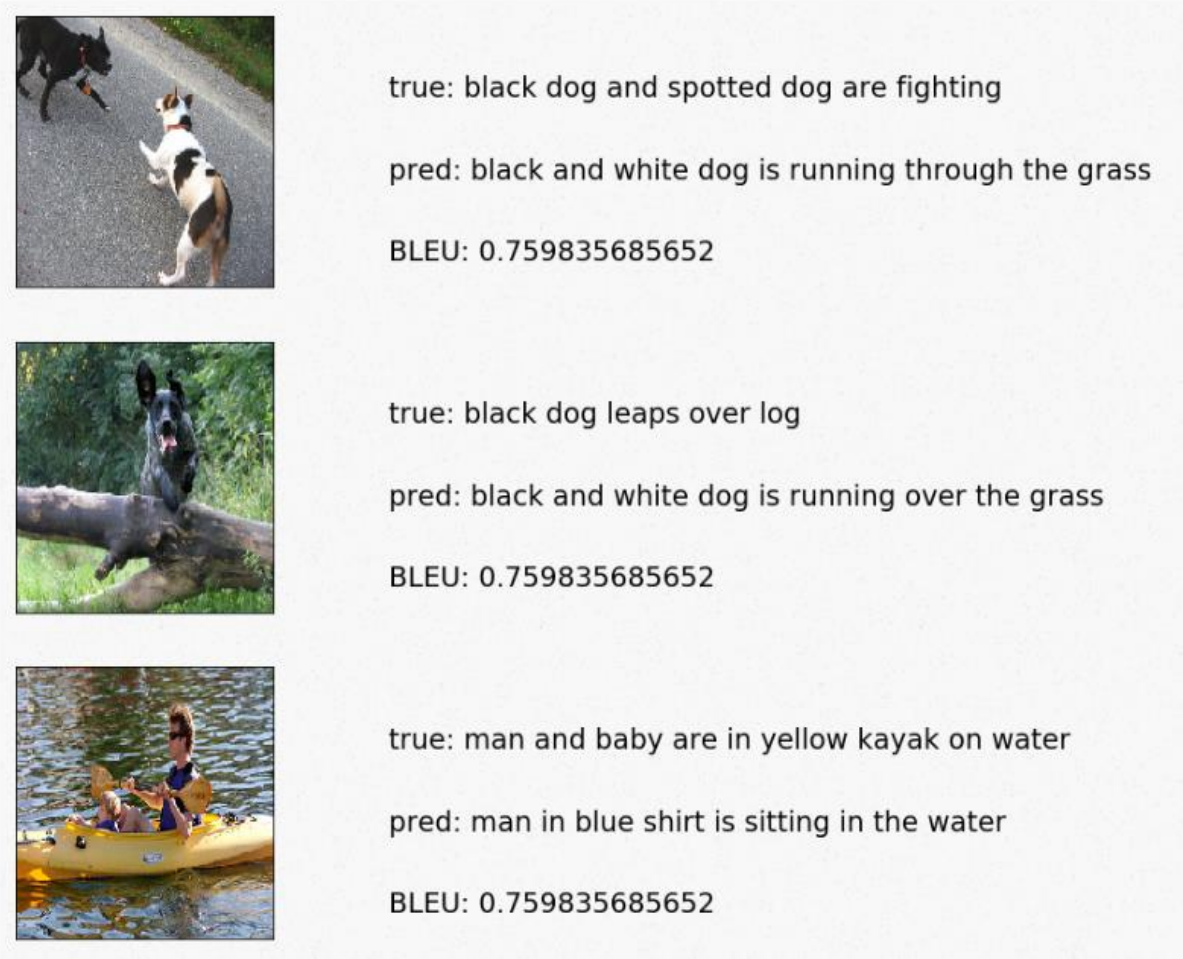

Figure 6. Examples with Captions and Corresponding BLEU Scores

Conclusion

We show empirically how an image's content can be isolated and displayed as a keyword, especially for visually challenged persons. The process uses various technologies, computer vision, transfer learning, convolution networks, recurrent neural networks, and unsupervised NLP-based algorithms for keyword generation. The system still does not provide accurate contents in an image and hence needs to be further fine-tuned. The main aim of providing keywords is still met and as such can be of immediate relevance to the visually impaired community at large.

The efficiency of our model depends on various factors, such as the volume of training data, the CNN model used, and also the quality of the images used for prediction. The proposed key extraction work will be fused with the existing FMAS project that uses cloud and the GSM system connected with the Raspberry Pi. The system

aims to make the communication effortless and can be individually personalized using suitable data using our deep learning component. The system efficacy can be further enhanced with the help of an eye tracker device. The system is then integrated with the Google Mini device thus allowing multifaceted Google functionalities in different aspects of usages for the disabled.

Acknowledgment

This work is wholly supported by the project titled "Finger-Movement Multimodal Assistive System (FMAS) for Smart People, including People with Special Needs" funded by the Ministry of Higher Education, Research and Innovation, Sultanate of Oman bearing Project ID BFP/RGP/ICT/19/152 under the Research Grant (RG) program.

References

E. McKillop et al., "Problems experienced by children with cognitive visual dysfunction due to cerebral visual impairment – and the approaches which parents have adopted to deal with these problems," Br. J. Vis. Impair., vol. 24, no. 3, pp. 121–127, 2006.

A. Antal et al., "Electrophysiological correlates of visual categorization: evidence for cognitive dysfunctions in early Parkinson's disease," Brain Res. Cogn. Brain Res., vol. 13, no. 2, pp. 153–158, 2002.

O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

I. Goodfellow et al., "Generative adversarial networks," Commun. ACM, vol. 63, no. 11, pp. 139–144, 2020.

W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," arXiv [cs.NE], 2014.

M. D. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," ACM Comput. Surv., vol. 51, no. 6, pp. 1–36, 2019.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

I. Gat, G. Lorberbom, I. Schwartz, and T. Hazan, "Latent space explanation by intervention," Proc. Conf. AAAI Artif. Intell., vol. 36, no. 1, pp. 679–687, 2022.

J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02, 2001.

Appendix

The implementation code can be accessed at the following Google Colab link.

https://tinyurl.com/2p8848ab