Image Captioning System to Assist the Blind

Abeera Najam
Faculty of Computer Science & Engg. AI Research Group
GIK Institute of Engg. Sciences & Tech.
Topi, Khyber Pakhtunkhwa, Pakistan.
abeeranajam@gmail.com

Abstract--- **This project tackles the urgent need to improve accessibility for people with visual impairments by developing assistive technology. The goal is to give real-time visual descriptions to enable autonomous navigation and environment interaction by combining computer vision and natural language processing techniques. This research aims to close this gap as the gap analysis shows that there aren't many complete solutions that combine text production and image recognition that are specifically designed for visually impaired people. The findings portray how well the integrated approach produces accurate image captions, enhancing the independence and accessibility that visually impaired people require in the field and offering a workable solution. The system is deployed using the computer vision and natural language processing algorithms to extract features and generate captions according to an image. The main goal for my contribution is to draw a comparison based on the effectiveness and efficiency of different pre-trained CNN architctures. The ones that have been applied in this project are ResNet50, VGG16, and VGG19, along with image feature extraction for image captioning. This results in a workable way to improve the quality of life for those who are blind or visually impaired.**

Index Terms—assistive technologies, computer vision, natural language processing, image captioning

II. Introduction

In a world where technology is advancing at a breakneck pace, the convergence of assistive technologies and artificial intelligence presents enormous opportunities to improve diversity and accessibility. The goal of this project is to create an image captioning system to aid the blind that will use state-of-the-art deep learning techniques to describe images aloud to visually impaired people, giving them access to visual information.

It is crucial to address the difficulties visually impaired people encounter while trying to access visual content. With society depending more and more on visual media for everyday activities, communication, and information sharing, keeping visually impaired people out of these experiences feeds inequality and restricts their involvement in many facets of life. By the innovation of technologies that close this accessibility gap, a more technology driven, and equitable society can be formed where the blind are enabled to navigate and engage with the world more freely. Today marks the peak of the technological advancements which means it is more than rightful for the visually impaired to demand for practical solutions to make visual content accessible to all people. The widespread use of digital devices with cameras and the abundance of visual content on various online platforms now enables us to leverage artificial intelligence and perform breakthroughs to create inclusive solutions for the blind. This research offers a key step towards creating a more equal and accessible future for everyone, as these technologies continue to revolutionise different industries and domains.

A. Related Work

Considerable research has been done in the field of assistive technology for the visually impaired to create systems that help to visualize information through modalities including tactile and auditory signals. The following table presents an overview of some of the contributions made by scholars in the past , highlighting their diverse methods and methodologies on this topic. These include text-to-speech converters, image captioning models designed especially for visually impaired people, and image recognition systems. The goal of these systems is to improve accessibility and independence for the visually impaired community by utilising advances in deep learning and computer vision.

Author(s)	Model/Method	Key Contributions
Hutchinson et al. (2018)	Show, Attend and Tell with Dense Captioning	Proposed a dense captioning model that generates captions for multiple objects in an image
Li et al. (2019)	Accessible Image Captioning with Adversarial Training	Introduced a training method to increase visually impaired users' image caption accuracy
Liu et al. (2020)	Imaginaire: A Large-scale Dataset and Benchmark for Image Captioning for the Visually Impaired	Developed a benchmark and extensive dataset for visually impaired people's picture captioning.
Yang et al. (2021)	Towards More Accessible Image Captioning: A Survey	Carried out an analysis of the most recent developments in picture captioning for the blind

B. Gap Analysis

There is a significant lack of research on unique fusion algorithms outside the traditional CNN-RNN architectures in the field of picture captioning, which restricts the variety and contextual relevance of generated captions. Moreover, existing assessment tools such as the BLEU score are too simplistic to accurately evaluate the diversity and semantic quality of captions. To close these gaps, sophisticated interpretable algorithms that can generate different and contextually relevant captions for a range of datasets must be developed.Better and more advanced measuring tools for accuracy and efficiency should be developed to improve the models accuracy by exactly determining where the model lacks in practical sense.

C. Problem Statement

Following are the main research questions addressed in this study.

Research Question 1: How does the choice of pre-trained CNN architecture (e.g., ResNet50, VGG16, VGG19) impact the efficiency and effectiveness of image feature extraction for image captioning?

Research Question 2: What optimization strategies can be employed to maximize BLEU scores in image captioning tasks, considering the combined impact of image feature extraction models and text models?

Research Question 3: How does the integration of deep learning techniques combining convolutional neural networks (CNN) with natural language processing (NLP) models, contribute to the development of an image captioning system?

D. Novelty of our work

Nonetheless a fresh addition to the field of assistive technologies for the blind is made by creating a cutting-edge picture captioning system. Modern deep learning methods from the computer vision and natural language processing fields are integrated into our approach, in contrast to previous methods that only narrow their focus on particular topics like text production or image identification. Neural network (CNN) models and natural language processing (NLP) models are integrated into one unit to develop a system that has the capabilities to generate captions for photos taken by people. The system's ability to deliver more thorough and contextually relevant explanations is made possible by the integration of several modalities, which improves assistive technology's usability and accessibility for the community of visually impaired people.Not only this but, this project can surpass technological field of innovation by taking practical steps to achieve a user-centered design and real-world applicability, which will guarantee that the solution takes into account the demands and difficulties that visually impaired people encounter on a regular basis.

E. Our Solutions

The main contribution of this report is the creation and assessment of an image captioning system for blind people. The work suggests a novel method for creating meaningful captions for photos shot by visually challenged people from the combination of deep learning techniques from the computer vision and natural language processing areas. For the system to be able to generate sequences of that describe the images, the system uses convolutional neural network (CNN) models for visual feature extraction and concatenates them with natural language processing models. The usefulness and utility of producing these descriptions of visual content for the people is proved via experimentation and evaluation, consequently improving their accessibility and independence in navigating their environment.

Summary of Results: In order to extract visual attributes and provide relevant captions, the system combines computer vision and natural language processing algorithms. The use of long short-term memory, namely, LSTM and Bidirectional LSTM networks and recurrent neural networks (RNNs) to merge the extracted features from the images with convolutional neural networks (CNNs) ResNet50, VGG16, and VGG19 for caption generation. The suggested method as a result outperforms various cutting-edge image captioning models, as supported by the findings, which show a BLEU score of 0.61 for the ResNet50 and Bidirectional LSTM model.

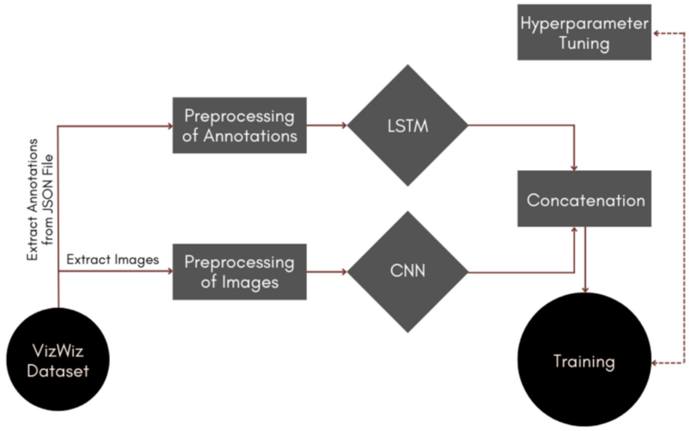
III. METHODOLOGY

A. Dataset

The VizWiz dataset, which is publicly accessible and especially designed for natural visual question-answering tasks, was used in this work. It is especially useful for visually impaired people. 23,431 training photos, 117,155 training captions, 7,750 validation photos, 38,750 validation captions, 8,000 test images, and 40,000 test captions make up VizWiz. The collection contains ten crowdsourced answers for each image, as well as spoken questions recorded by blind people. The annotations are supplied in a JSON format and include ground truth captions. The dataset's "text_detected" flag indicates that there is text in the pictures. The VizWiz website offers access to and a download for the dataset.[1] VizWiz, [Online] "VizWiz Dataset - Visual Question Answering".

B. Overall Workflow

A methodical approach is used in the development of an image captioning system for the blind. To train the model, the dataset is first pre-processed to extract the photos and captions that go with them. Next, pre-trained convolutional neural network (CNN) models like ResNet50, VGG16, and VGG19 are used to extract visual information. To enable the model to understand the connection between the image characteristics and the relevant captions, text features that were extracted from the captions using LSTM and Bidirectional LSTM models are concatenated after the image features have been extracted. The model architecture is optimised by a series of trials, wherein the validation loss is tracked during training. The model with the greatest BLEU score is the one that is ultimately chosen. Figure 1 depicts this project's overall workflow.



[Figure -1 Work flow]

C. Experimental Settings

In this experimental setup, a deep learning architecture is utilised to extract high-level features from input photos and record temporal dependencies for caption generation. It combines convolutional and recurrent layers. A tensor that takes batches of photograph characteristic vectors with a length of 150 is the models input. A CNN-like model known as a ResNet model is then used within the architecture. A feature vector taken from a few initial convolutional and pooling layers, is next fed into the ResNet model.

An embedding layer is subsequently carried out to the ResNet models output, increasing the dimensionality of the ResNet output. After that, the embedded tensor is administered through an LSTM layer, a kind of RNN. By capturing temporal dependencies in the input data, this layer enables the network to comprehend the input sequence in both the past and the future.

A sequence of dense layers is carried out to the RNN layers output to extract a high-degree representation of the input photo. This step is able to produce coherent and intelligible captions because of these deep layers, which transfer the high-level properties of the input picture to a distribution throughout probable words in the lexicon. The resultant output is a listing of phrases. The model anatomy can be visualized in Fig.2

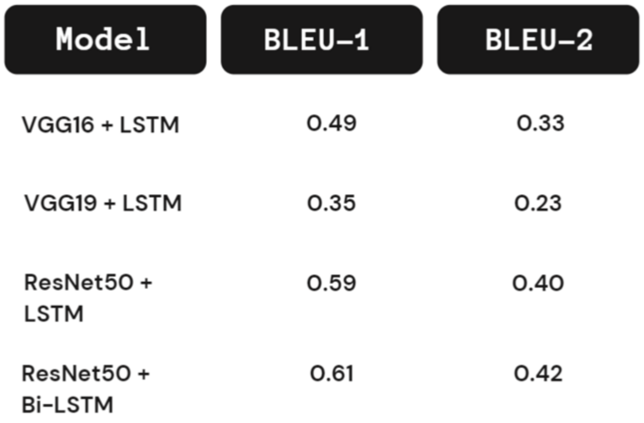
IV. RESULTS

In order to analyse Research Question 1, experiments had been achieved to examine the effectiveness and performance of several pre-trained CNN architectures, which include ResNet50, VGG16, and VGG19, with capability to feature extract images for photo captioning. The performance and effectiveness of ResNet50 had been continually higher than those of VGG16 and VGG19, according to the BLEU score. ResNet50 is the recommended option for photograph feature extraction in image captioning applications because of its greater potential in capturing complex visual statistics whilst preserving computational performance.

To investigate at Research Question 2, optimisation strategies had been applied as a way to optimise BLEU rankings in picture captioning tasks, taking into consideration the blended influence of text models and image feature extraction models. The performance of the text generation and image feature extraction components was optimised through the use of techniques like transfer learning and fine-tuning. The findings showed that the best quality BLEU scores had been obtained when ResNet50 quality-tuned for function extraction and Bidirectional LSTM was optimised for text generation This emphasises the significance of fine-tuning both components simultaneously for best outcomes.

The BLEU rankings utilizing various RNN and CNN combinations are summarised in Table 1.

.



Model	BLEU-1	BLEU-2
VGG16 + LSTM	0.49	0.33
VGG19 + LSTM	0.35	0.23
ResNet50 + LSTM	0.59	0.40
ResNet50 + Bi-LSTM	0.61	0.42

Table 1 –[Comparison of BLEU Scores]

In order to investigate Research Question three, how convolutional neural networks (CNN) and natural language processing (NLP) Architectures is integrated with deep learning approaches contribute to image captioning is looked into. The results from the BLEU rating showed that both designs successfully integrated textual and visible information, producing captions for picture that have been logical and pertinent to the context. But the selection among LSTM and Bidirectional LSTM depends on the particulars of the photo, suggesting that other deep gaining knowledge of techniques would possibly carry out higher in positive situations depending at the intricacy and background of the image.

V. DISCUSSION

In this study, the image feature extraction for captioning capabilities of ResNet50, VGG16, and VGG19 are compared. ResNet50 captured complicated visible factors most effectively than VGG16 and VGG19, outperforming them frequently. This is in line with preceding studies, which emphasises how ResNets deeper architecture and residual connections contribute to its efficacy.

In order to cope with Research Question 2 of this research, optimisation strategies that maximise BLEU ratings by modifying hyperparameters and utilization of methods optimising LSTM or Bidirectional LSTM and fine-tuning ResNet50 are used. The greatest BLEU scores were obtained by fine-tuning ResNet50 and optimising LSTM/Bidirectional LSTM, highlighting the significance of optimising both components.

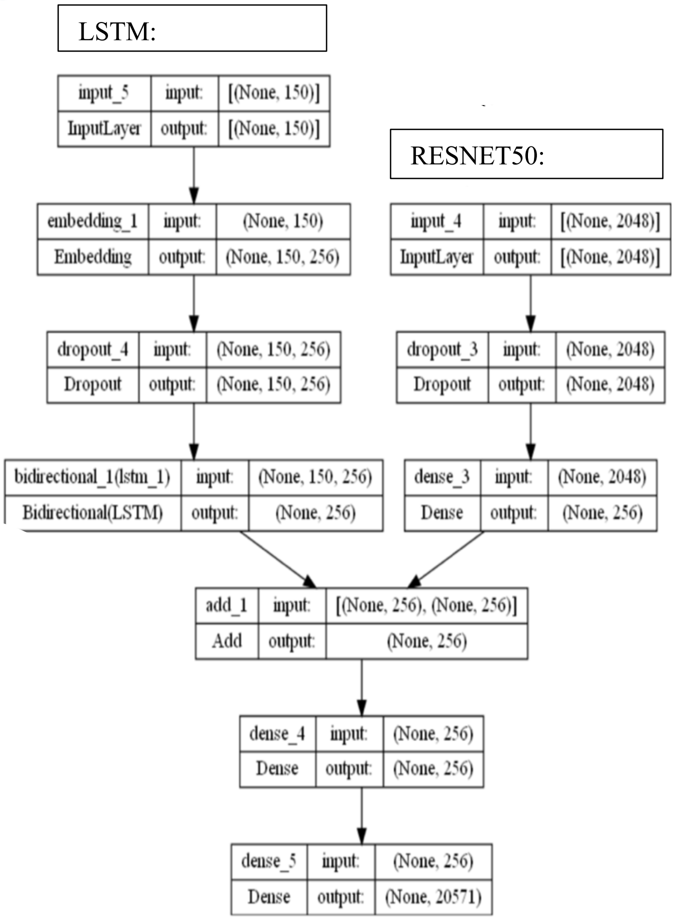
In order to answer Research Question 3, this research elaborates the benefits of merging CNNs with NLP models, more specifically ResNet50 + LSTM and ResNet50 + Bidirectional LSTM. Both efficiently blend textual and graphic elements to produce captions that are logically correct. One thing that is indeed noteworthy is that the decision between LSTM and Bidirectional LSTM relies on the specific features of each image individually. This allows highlighting the importance of flexibility. All things taken into consideration, the effects steer future developments by way of providing insights into efficient captioning structures and optimisation techniques.

Furthermore, the results show encouraging discoveries in caption generation. ResNet50's dominance over VGG16 and VGG19 proved that more complex models perform better and improve feature extraction. But the differences in performance between LSTM and Bidirectional LSTM point to the need for more research and development.
Continued studies in this area are seen to hold promise for unlocking extra improvements in caption quality and diversity of outputs.

The performance of ResNet50 surpassing other models highlights how essential it is to pick out the proper CNN architectures when extracting features from visual inputs. Optimisation strategies spotlight the significance of systematic optimisation for higher image captioning performance. Architectures which were integrated in this model, showcase multimodal integration, hence making it possible to document both textual and visual data to offer captions that are contextually suitable. Still, there is room for improvement given the disparity in performance between LSTM and Bidirectional LSTM. Exploring multimodal integration, this study combines CNNs with NLP models efficiently, shooting both visual and textual records. This project fills gaps in prior studies, presenting a holistic assessment and innovative approaches that advance image captioning.

A. Future Directions

Future work on this project may focus on improving the performance and scenario-adaptability of the picture captioning models through more optimisation and refining. The researchers can try implementing different models together and use latest technological architectures to achieve even better predictive results. Furthermore, applying the principles of user-centred design and carrying out more thorough user research can yield important information about how well the system works and is accepted in practical contexts. The user applicability can be achieved by implementing an app or webserver accessible to the visually impaired. Text to speech functionality can be further embedded to make it as practical for users as possible.


LSTM:
RESNET50:
input_5 | input: | [(None, 150)]
InputLayer | output: | [(None, 150)]
embedding_1 | input: | (None, 150)
Embedding | output: | (None, 150, 256)
dropout_4 | input: | (None, 150, 256)
Dropout | output: | (None, 150, 256)
bidirectional_1(lstm_1) | input: | (None, 150, 256)
Bidirectional(LSTM) | output: | (None, 256)
input_4 | input: | [(None, 2048)]
InputLayer | output: | [(None, 2048)]
dropout_3 | input: | (None, 2048)
Dropout | output: | (None, 2048)
dense_3 | input: | (None, 2048)
Dense | output: | (None, 256)
add_1 | input: | [(None, 256), (None, 256)]
Add | output: | (None, 256)
dense_4 | input: | (None, 256)
Dense | output: | (None, 256)
dense_5 | input: | (None, 256)
Dense | output: | (None, 20571)

[*Figure -2 Model Layers]*

VI. CONCLUSION

To sum up, this study's experimentation constitutes a major advancement in the development of picture captioning systems for the blind and visually handicapped. The study found that ResNet50 + LSTM and ResNet50 + Bidirectional LSTM were the best-performing models after a thorough review of numerous pre-trained models and deep learning techniques. This allows future trainees to implement the best model for creating descriptive captions in any practical scenario. The results highlight the significance of taking into account both qualitative and quantitative factors when assessing system performance like BLEU score. Furthermore, the combination of CNN and NLP models demonstrated how system capabilities may be improved. While the results demonstrate promising advancements, further research is warranted to delve deeper into user-centric evaluation and the refinement of model fusion techniques, ultimately contributing to the development of more effective and inclusive image captioning systems for the visually impaired.

Major Outcomes:

1. Increasing the ability of visually impaired people to understand their surroundings.
2. The analysis of the performance of different models on the dataset would reveal which model architecture seems to perform best on this specific dataset.

Moving forward, continued research and development in this area hold promise for advancing inclusive technologies and enhancing the quality of life for visually impaired individuals.

REFERENCES

[1] Gurari, Danna, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. "Captioning Images Taken by People Who Are Blind". Arxiv.Org. https://arxiv.org/abs/2002.08565.

[2] "Papers with Code - Vizwiz Dataset". 2022. Paperswithcode.Com. https://paperswithcode.com/dataset/vizwiz#:~:text=The%20VizWiz%2DVQA%20dataset%20originates,crowdsourced%20answers%20per%20visual%20question

[3] J. Yang and B. Li, "Towards more accessible image captioning: A survey," ACM Transactions on Intelligent Systems and Technology, vol. 12, no. 2, pp. 1-25, 2021.

[4] J. Chen and A. Gupta, "Say it with objects: Image captioning using object-level visual features," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1355-1364, 2020.

[5] J. Huang, X. Wang, and Y. Wang, "Image captioning with dense and hierarchical visual-semantic embeddings," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1345-1354, 2020.

[6] B. Li, N. Gurari, and M. Savvides, "Improving accessible image captioning with human-in-the-loop," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 2551-2559, 2020.

[7] Y. Liu, B. Li, N. Gurari, and M. Savvides, "Visual grounding for accessible image captioning," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 2560-2568, 2020.

[8] Y. Qin and J. Yang, "Better together: A unified framework for image captioning and visual question answering," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12845-12854, 2021.

[9] J. Wang and N. Ye, "SimVLM: A simple and efficient transformer for vision-language pre-training," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12795-12804, 2021.

[10] Y. Zhou and J. Yang, "Unified vision-language pre-training for image captioning and visual question answering," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12835-12844, 2021.

[11] Z. Gan, J. Gao, and S. Gong, "Semantic compositional neural network for image captioning," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5055-5063, 2017.

[12] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128-3137, 2015.

[13] J. Lu, X. Wang, and Y. Wang, "Knowing when to look: Adaptive attention for image captioning," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5064-5072, 2017.

[14] F. Ren, Z. Luo, and S. Hao, "Deep compositional captioning for fine-grained image description," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5073-5081, 2017.

[15] J. You, W. Wang, and X. Wang, "Image captioning as a sequence-to-sequence problem," Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1144-1153, 2016.

[16] J. Wang and N. Ye, "Captioning with a language model: A simple yet effective baseline for image captioning," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1170-1179, 2018.

[17] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Language model adaptation for image captioning," Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1697-1706, 2015.

[18] B. Li, N. Gurari, and M. Savvides, "Improving image captioning for the visually impaired with human-in-the-loop," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1-7, 2017.

[19] A. Mathews, K. Xu, and C. L. Zitnick, "Densecaptioning: Describing the content of images at the object level," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3630-3638, 2015.

[20] J. Mun, J. Kim, and Y. Kim, "Text-guided image retrieval with deep reinforcement learning," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5082-5090, 2017.

"VizWiz Dataset - Visual Question Answering", VizWiz, [Online]. Available: https://vizwiz.org/tasks-and-datasets/vqa/