

Image captioning to aid blind and visually impaired outdoor navigation

Ruvita Faurina¹, Anisa Jelita¹, Arie Vatesia¹, Indra Agustian²

¹Department of Informatics, Faculty of Engineering, University of Bengkulu, Bengkulu, Indonesia

²Department of Electrical Engineering, Faculty of Engineering, University of Bengkulu, Bengkulu, Indonesia

Article Info

Article history:

Received Oct 14, 2022

Revised Oct 31, 2022

Accepted Dec 21, 2022

Keywords:

Attention mechanism

Convolutional neural network

Image captioning

Long short-term memory

Visually impaired

ABSTRACT

Artificial intelligence technology has dramatically improved the quality of services for human needs, one of which is technology to improve the quality of services for the blind and visually impaired, particularly technology that can help them understand visual sights to facilitate navigation in their daily lives. This study developed an image captioning model to aid the blind and visually impaired in outdoor navigation. The image captioning model employs the encoder-decoder method, with the convolutional neural network (CNN) feature extraction and attention layer as encoders and the long short-term memory (LSTM) as decoders. ResNet101 and ResNet152 are used in the encoder to extract image features. The results of the extraction and caption are forwarded to the attention layer and the LSTM network. The attention layer uses the Bahdanau attention mechanism. The accuracy of the model is calculated using the bilingual evaluation understudy score (BLEU), metric for evaluation of translation with explicit ordering (METEOR) and recall-oriented understudy for gisting evaluation-longest common subsequence (ROUGE-L). ResNet101 performed the best on BLEU-4, scoring 91.811% and 94.0337% in the METEOR evaluation. The captioning results show that the model is quite successful in displaying a simple caption that is suitable for each image.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Ruvita Faurina

Department of Informatics, Faculty of Engineering, University of Bengkulu

Jl. W.R Supratman, Kandang Limun, Bengkulu 38371, Indonesia

Email: ruvita.faurina@unib.ac.id

1. INTRODUCTION

Humans rely on their eyes to obtain visual information from their surroundings. Blind and visually impaired (BVI) people, defined as people with visual impairments who use assistive technology [1], face at least three significant limitations in their daily activities: limitations in the scope of experience, limitations in interacting with the environment, and limitations in mobility [2], [3]. Numerous factors impede BVI people's mobility and interaction with the environment, particularly in open public spaces such as roads, sidewalks, parks, and city recreation areas [4]. As discussed in [2], the guide stick, which is a tool for the blind when traveling, is still not effective enough in helping mobilize the blind, mainly outdoors. In an unfamiliar environment, common obstacles such as humans, animals, walls, potholes, stairs, or muddy road surfaces pose a high risk of injuring the blind person. The guide sticks commonly used by blind people do not provide enough specific information about what objects or obstacles are in front of them, making it difficult for the blind to quickly determine their reaction to things around them while moving or walking.

Advances in artificial intelligence technology to process various types of input data, such as audio, images, and even video, have triggered a lot of research and development in the field of assistive technology

(AT), including BVI. Research in [5] built a model that predicts emotions in human speech based on audio input. Meanwhile, research [6]–[8] builds a model for interpreting human sign language and gestures into text based on video input. Many technological and scientific advancements for BVI have been made by prior researchers, one of which is BVI with image captioning techniques [4]. Image captioning is a subset of artificial intelligence, which combines fields such as computer vision, natural language processing, and machine learning to produce descriptive sentences that are in sync with specific images, allowing people to understand the implied information in the image [9], [10].

One of the focuses of image captioning research as part of the visual assistance system [11] is providing image descriptions for blind people. The ability to automatically convert images to text can assist the visually impaired in navigation or generate information from images. Even so, providing this description will be difficult for the computer because, in addition to processing the image, the computer must be able to understand the image's content and provide a description in natural language [12]–[14]. Image captioning aims to create descriptive sentences that complement the image [15]. Convolutional neural network (CNN) has been widely used in captioning work [16], [17] since their popularity in dealing with computer vision problems such as classification [18]–[25] and object detection [26]–[30]. The ability of the long short-term memory (LSTM) network to learn order dependencies in sequence prediction problems in data series [31]–[36] makes it widely used for captioning tasks in generating sentence predictions. Research [37]–[40] utilizes a combination of CNN as feature extraction and LSTM to predict the output based on the order dependencies.

Research [11], [41], [42] reviewed several studies on image captioning tasks using various methods. The encoder-decoder framework, inspired by neural machine translation, is one of the methods discussed. The encoder network is in charge of encoding the image into a temporary representation, and the decoder forwards it to produce output in the form of sentences. The well-known neural image caption (NIC) model proposed by research [43] is an example of a representation of the encoder-decoder method in image captioning tasks. In another study, [43]–[47] used an encoder-decoder approach with CNN and an LSTM neural network on the decoder section, maximizing the benefits of LSTM, which can track essential data during input processing through “forget gates” to eliminate irrelevant details. The CNN on the encoder, like in the classification task, can be built from scratch or using a pre-trained model. Transfer learning is a model that was previously trained on a large dataset and then reused (with parameter adjustments) as a starting point for other tasks with a new dataset [18], [48]–[50]. The captioning task in [51], [52] was trained using three pre-trained models: visual geometry group (VGG) and ResNet. Meanwhile, research [52] employs the Inception-v3 encoder and gated recurrent unit (GRU) decoder to generate captions in Indonesian. On the other hand, several studies summarized in [53], [54] add an attention layer to the decoder network. This attention mechanism is effective in increasing accuracy because it allows the model to focus on essential parts when processing input into output sequences [55].

With many variations of CNN methods that can be used for image captioning, research [16] conducted a comparative study of 17 well-known transfer learning architectures combined with LSTM using an encoder-decoder approach. The model is trained on the flickr8k dataset using two methods: with and without an attention mechanism. In general, models with an attention layer outperform others in terms of accuracy and error reduction. The two CNN architectures that perform best in evaluating bilingual evaluation understudy (BLEU) metrics and metric for the evaluation of machine translation output (METEOR) are ResNet152 and ResNet 101. Flickr8k, Flickr30k, and microsoft common objects in context (MSCOCO) datasets are some well-known datasets commonly used in image captioning modeling training [42]. However, during its development, the existing modeling was also trained on other datasets for more specific captioning tasks. The study [56] uses a dataset of images related to local tourism in Yogyakarta gathered from the Google search engine. This research aims to create a unique image captioning model for Yogyakarta tourism that can be expanded into a chatbot system. Research [57] investigated captioning for car image datasets and lifelogging datasets [58].

Many research and developments to assist the BVI navigation have been carried out by previous researchers, both in algorithms and the design of tools or prototypes. Before the development of machine learning and massive deep learning, as it is today, previous studies were still based on heuristic algorithms, as in the following studies. In 2007, researchers [59] developed a portable device on blind sticks to capture the visually impaired's environmental information. By processing using field-programmable gate array (FPGA) and digital signal processor (DSP), the information captured by the stereoscopic camera is translated into a tactile feedback system consisting of 27 mini-actuators. In 2011, research [60] developed a smart vision prototype to help the visually impaired navigate devices consisting of a stereo camera, a portable computer in a pocket or shoulder, and small earphones. Research focuses on local navigation to detect path boundaries and obstacles in front of the user and beyond stick's reach so that the visually impaired can stay on the right track and be alert to detected objects. The method used is adapted hough space (AHS) for roads and static objects and optical flow for moving objects. In 2012, research [61] proposed a camera-based assistance system for the visually impaired to read text from nameplates and text on objects with complex backgrounds.

Text on objects is recognized using optical character recognition (OCR) software, then converted into speech as the final output. The experimental results were evaluated on the ICDAR Robust Reading 2003 dataset. The results showed that the algorithm in this study outperformed the previous algorithm presented in the ICDAR dataset, with a precision score of 0.69, recall of 0.56, and time of 10.36. In 2014, research [62] introduced a cloud-based outdoor navigation system (COANS) for the visually impaired that uses an external GPS and can be accessed via an Android smartphone. The system's test scenarios include detecting traffic light status, zebra crossings, and benches near the user. In 2015, research [63] developed a voice-based assistive system that used color and infrared proximity sensors as obstacle detectors rather than images. The system's output is a device that can provide the user with alerts in the form of sound and vibration when in various fields such as grass, roads, paths, zebra crossings, and stairs.

With the advancement of computer technology and artificial intelligence, research into navigation aids for the visually impaired has begun to employ deep-learning image captioning algorithms. In 2018, researchers [64] created an image captioning prototype with a Raspberry Pi3 and an image captioning model based on the API provided by cloud-based microsoft cognitive services. When an obstacle is detected, the developed device produces audio, a vibration signal, and a ringtone. The study's findings are generally favorable, though there are some flaws. For example, the caption becomes meaningless when there is a shadow in the image. In 2019, researchers [65] created an image captioning model for the visually impaired by combining the Stanford CoreNLP model with visual feature extraction using the VGG16 architecture and the MSCOCO dataset. The resulting model consistently outperforms state-of-the-art approaches across multiple evaluation metrics. The same researcher from research [65], in the following study [66], developed the Android application "Eye of Horus" using an image captioning model based on VGG16 and LSTM, with the MSCOCO dataset as the source. The results show that the integrated platform has great potential to be used by the blind and deaf with advantages such as ease of portability, simple operation, and fast response.

In 2021, research [67] developed wearable devices to assist the navigation and communication of the visually impaired. The device can identify faces, recognize familiar objects, and read text on images. Meanwhile, Nasir *et al.* [68] developed an android-based object recognition model that can aid the blind in recognizing ordinary objects they encounter every day, such as money, clothing, and other items. In 2021, research [69] combined the deep learning image classification architecture VGG-16 with the RNN algorithm to develop an image captioning model for the visually impaired using the flickr8k dataset. The resulting model was evaluated using BLEU scores. BLEU-1 scored 0.545418, BLEU-2 0.290155, BLEU-3 0.193921, and BLEU-4 0.085051. Research [70] designed automatic image captioning for the visually impaired based on AoANet by utilizing the text detected in the image as an input feature and adding a pointer-generator mechanism. The dataset used is The Vizwiz Captions dataset. The model successfully outperformed the original AoANet, with a CIDEr score of 35% and a SPICE score of 16.2%. In 2022, research [71] developed an image captioning application with voice output on Android aimed at the visually impaired. This research uses LSTM and GRU algorithms with pre-trained model VGG16 and MS COCO datasets.

Aside from the image captioning method, several researchers have created tools for the visually impaired based on object detection and semantic segmentation. In 2019, researchers [72] created an implementation of single shot detection (SSD) and the MobileNet architecture for visually impaired object detection models. The MSCOCO dataset was used in the research. The model contains a Raspberry Pi with an audio output. In 2021, research [73] developed an object detection prototype with sound output for the visually impaired. The model was developed by combining a single-shot multibox detection framework with the MobileNet architecture to create a rapid real-time multi-object detection device that is compact, portable, and has a short response time.

The following are the main contributions of this study: we developed an image captioning model to aid BVI outdoor navigation. The image captioning model employs the encoder-decoder method, with the CNN feature extraction and the attention layer as encoders and the LSTM as decoders. Based on research [16], the CNN ResNet101 and ResNet152 feature extraction architectures are compared to determine the best architecture that can be used. The Bahdanau attention mechanism is used on the attention layer. The bottleneck issue in conventional encoder-decoder systems can be solved by the Bahdanau attention [74]. The model was created using a primary dataset of images of open public spaces in Bengkulu, Indonesia. Compared to previous related studies cited in this chapter, it is clear that this research has made a significant contribution.

2. METHOD

Modeling will typically perform two main functions. First, extract the image features. Second, using the previously extracted image features, create a suitable description. Figure 1 depicts the modeling workflow.

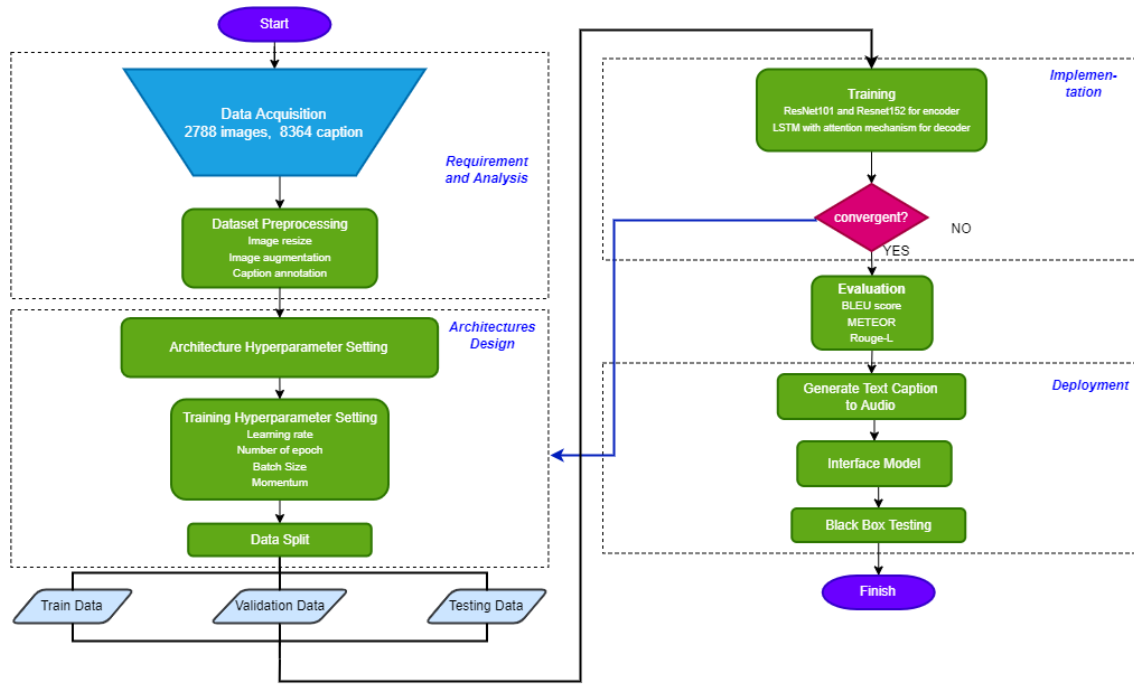


Figure 1. Modeling workflow

2.1. Data acquisition

The research focuses on image collection in open public spaces. The data collection method used in this study includes all stages of the data acquisition process required to build the model. As a result, we conducted interviews with blind people at the Dharma Bhakti Kesejahteraan Sosial Amal Mulia Foundation before gathering the data. Three of the ten questions asked concern respondents' activities in open public spaces:

- i) How often do you travel in open public spaces, and what are the common obstacles you face?
- ii) How can describing a visual scene help them create a mental map?
- iii) What types of devices are comfortable to use?

According to the interviews, three places in open public spaces are frequently passed by blind people in their activities: sidewalks, roads, and city parks. In addition to interviews, the field study (observation) method was used with one of the respondents, an active student at a private university who walks to campus every day. The findings of interviews and field studies are then used to compile datasets.

2.2. Dataset and caption annotation



The dataset consists the collection of public space photographs taken by the researcher in Bengkulu city, Indonesia. Images were taken with a mobile camera at a distance of 1.5-2 meters from the object, using eye angle and high angle techniques. Images in the primary dataset are classified into seven classes based on the similarity of the scene and the obstacles in the image: puddles, potholes on sidewalks, cars, motorcycles, intersections, ditches, and zebra crossings. The dataset comprises 2788 images, which are subsequently divided into training, validation, and testing sets with a ratio of 80:10:10. The distribution of training, validation, and testing datasets is shown in Table 1.

The Flickr8k dataset [42], one of the most commonly used datasets for captioning tasks, is used for image naming formats and caption file storage, training, validation, and testing. Each image has three captions that are related to the description and explain the situation in the image. Table 2 shows the sample of images and captions in the dataset, and Figure 2 represents some of the most and least words in the dataset.

Table 1. Dataset distribution

Stage	Total
Training	2447
Validation	223
Testing	118

Table 2. Sample of images and captions

Image	file_name	caption
	trotoar-berlubang-42.jpg#0	<i>Hole in the middle of the sidewalk.</i>
	trotoar-berlubang-42.jpg#1	<i>Pavement with holes.</i>
	trotoar-berlubang-42.jpg#2	<i>Be careful; there is a hole in the middle of the sidewalk.</i>
	selokan-kiri-m-1.jpg#0	<i>Ditches and cars are on the left side of the road.</i>
	selokan-kiri-m-1.jpg#1	<i>Road with a car and a ditch on the left.</i>
	selokan-kiri-m-1.jpg#2	<i>Be careful; there are ditches and cars on the left side of the road.</i>

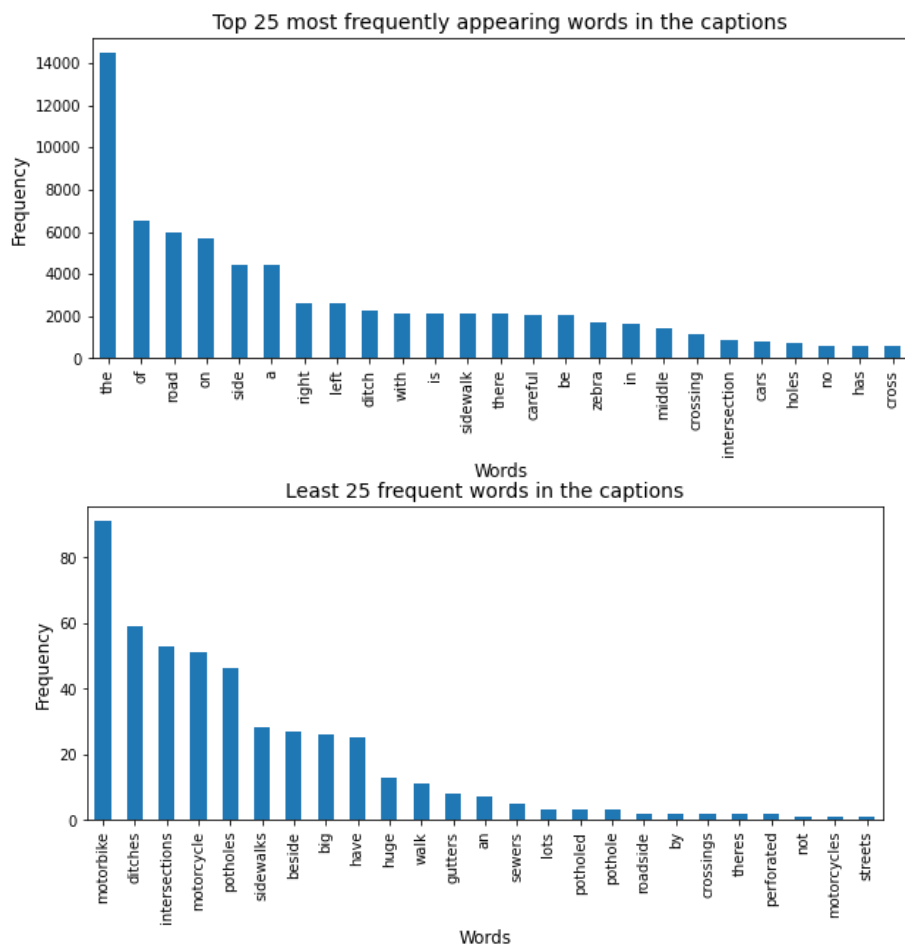


Figure 2. Top 25 most and least words in captions

2.3. Dataset preprocessing

The size of the image datasets collected varies. As a result, resizing the image to conform to the input standard specified in the transfer learning architecture rules is critical. The input size used is 224×224 with RGB image format. In addition to resizing, images are augmented to enrich the dataset. Random brightness, random blur, and random rotation are some of the augmentations used.

While preprocessing on captions is done to convert previous words sentences into a series of tokens based on a unique word index in the dictionary. Caption preprocessing consists of the following steps:

- Special token*: This special token consists of <eos> as token added at the beginning of a sentence, <eos> at the end of a sentence, as replacement token for unmatched words of the vocabulary in the dictionary, and <pad> as an additional token to equalize the sequence entire length.
- Lower case*: Captions will be changed to lowercase so the vocabulary will have a unique, non-repeating vocabulary.
- Tokenization*: It is the process of breaking text into smaller pieces (tokens). In the caption dataset, sentences will be broken down into word sequences. Each word will then be numbered for the computer to understand the data.
- Convert the token to a sequence of token*: The tokens will be converted back into sequences after the caption is changed to a token with unique numbering. In the sentence, this process generates a word order vector.
- Padding Sequence*: We must equalize the length of the sequences in order for the model to be trained. Padding sequences will add 0 to each word sequence vector whose length is less than the maximum length specified.

2.4. Modeling

The modeling stage will typically perform two significant functions. First, image features must be extracted. The second task is to create a suitable description for each image. At the training stage, there are two types of inputs. First, the input is an image that has been resized to match the standard size established as a rule in the transfer learning architecture. Because we will be using the ResNet architecture, we must ensure that the image has a size of 224,224,3 before proceeding. The second input is a caption preprocessed to become a token index sequence.

Figure 3 shows the architecture of image captioning used in this research. Feature extraction on the encoder uses CNN with ResNet101 architecture compared to ResNet152. Both pre-trained models were previously trained for the 1000 class classification task on the ImageNet dataset. The ResNet101 and ResNet152 architectures are then loaded using the PyTorch framework. The final two layers, the classification layer, will be removed because they will be used for a different task, in this case, image captioning. The feature extraction output from the encoder will then be forwarded to the attention layer. The attention layer is a connecting layer that aims to improve the performance of the encoder-decoder. The attention layer will learn to focus on certain parts of the image that are considered necessary by dividing the image into n parts, then calculating attention weights. In this layer, the input will then be mapped to the vector space of the hidden state and used to initialize the hidden and cell states. The result is a context vector which is then forwarded to the LSTM network. The decoder will then loop to generate word predictions, starting with <eos> as the initial initialization. The process will stop when the decoder generates a <eos> token.

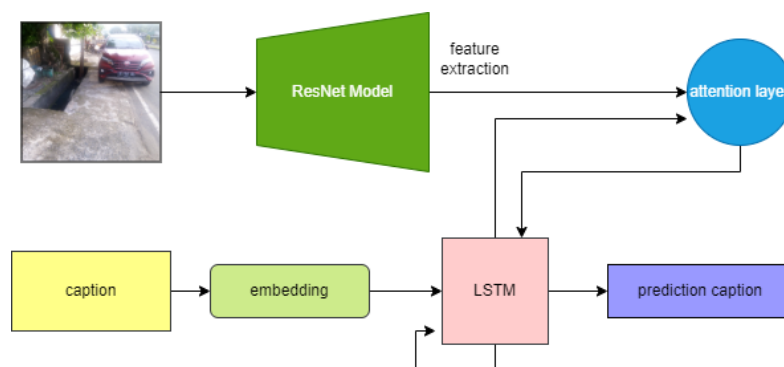


Figure 3. Image captioning architecture

2.5. Evaluation

The model's performance will be evaluated using the BLEU matrix. The BLEU score is a metric for evaluating machine-translated texts automatically. The BLEU value represents how similar the candidate text or machine-generated prediction is to the reference text [75], [76]. The BLEU score compares text from the engine to a high-quality reference set using a number between zero and one. A value of 0 indicates that the

engine's text results are of low quality, whereas a value closer to 1 indicates that the engine's text results are of high quality.

METEOR is used to evaluate the translated text by clearly matching word-to-word between the reference text and the predicted text using a combination of unigram-precision, unigram recall, and fragmentation measures to measure how well the word order is between the predictive and reference texts [77], [78]. The match score will be calculated against each reference independently to get the best score. It also applies when the model has more than one reference text for each predictive result. Meanwhile, in the ROUGE-L evaluation [79], [80], the precision and recall scores did not depend on the n-grams match sequentially but used longest common subsequences (LCS) to measure the similarity between the reference text and the predicted text, where LCS was the most extended series of words that were both found in reference and predictive texts.

3. RESULTS AND DISCUSSION

The research uses the ResNet architecture as an encoder to extract image features and LSTM with an attention layer on the decoder side to produce word sequences as caption predictions. We compare two ResNet architectures on the encoder side: ResNet101 and ResNet152. During the training process, we applied a learning rate of 0.0001 for both models. Figure 4 depicts the graphs of loss and accuracy during the model training process. ResNet101 training and validation loss, ResNet101 training and validation accuracy, ResNet152 training and validation loss, and ResNet152 training and validation loss are all depicted in Figures 4(a) to 4(d), respectively. As epochs increase, graph values are automatically updated. We also implement an early stop 70 to ensure the model training process runs optimally to achieve the minimum loss value and the highest possible accuracy score. The training process will be automatically stopped if, during the addition of seventy epochs, the accuracy and validation score of the model no longer increase. Then the training process will be automatically completed.

ResNet101's best accuracy is 66.303% in training and 65.664% in validation. Hence, the training loss of the ResNet101 model was reduced to 0.772 for training and 0.894 for validation in the 70th epoch. The model becomes increasingly overfit in the next epoch increment, where the training loss score no longer decreases significantly while the validation loss score increases periodically. When early stopping reached its maximum limit at the 140th epoch, the training process ended.

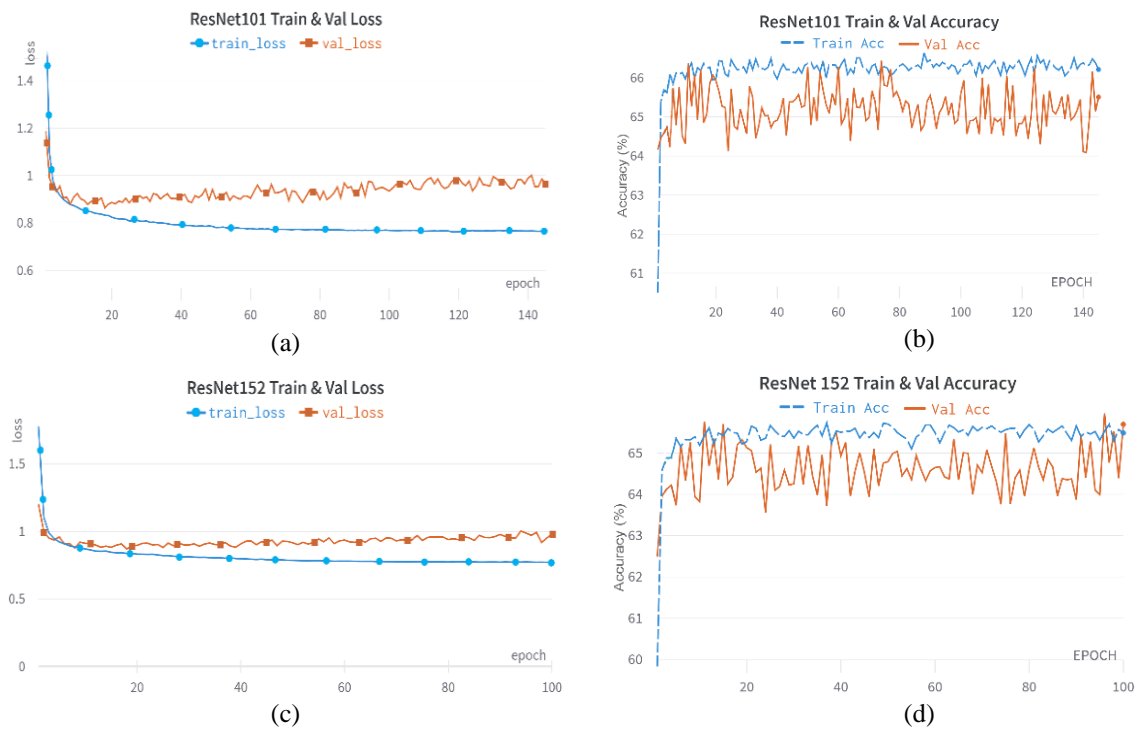


Figure 4. Loss and accuracy rate on training and validation dataset during the training phase, (a) training and validation loss of ResNet101, (b) training and validation accuracy of ResNet101, (c) training and validation loss of ResNet 152, and (d) training and validation loss of ResNet152

The training loss in the ResNet152 model has been reduced to 0.748, and the validation loss has been reduced to 0.902. Meanwhile, training accuracy has reached the highest value of 65.624%, while validation accuracy has reached 64.986%. The ResNet152 training process, like the first, ends at the 100th epoch when the training loss no longer decreases. The validation loss score, on the other hand, continues to rise, indicating that the model has become overfit, as illustrated in Figure 5. Figures 5(a) and 5(b) show the BLEU score for ResNet101 and ResNet 152, Figures 5(c) and 5(d) show the Rouge-L score for ResNet101 and ResNet 152, Figures 5(e) and 5(f) show the METEOR score for ResNet101 and ResNet 152. After reaching the maximum early stop, the training process was terminated, and no further training was conducted. For training, there is no set epoch limit. However, if the validation error value rises, the training process can be halted because, in some cases, a higher number of epochs may cause the model to overfit, as described in the paper [28].

Quantitative evaluation was carried out using BLEU, METEOR, and ROUGE-L metrics. These three are metrics commonly used in natural language generation (NLG) evaluation to compare the predicted text with a collection of reference texts from the dataset. The evaluation results are shown in Table 3 and Figure 5. The ResNet101 dominates the best evaluation scores on the BLEU-2, BLEU-3, BLEU-4, and Rouge-L metrics. However, in evaluating BLEU-1 and Meteor, ResNet152 got better results than ResNet101. Further, Figure 5 displays the evaluation graphs for the two trained models.

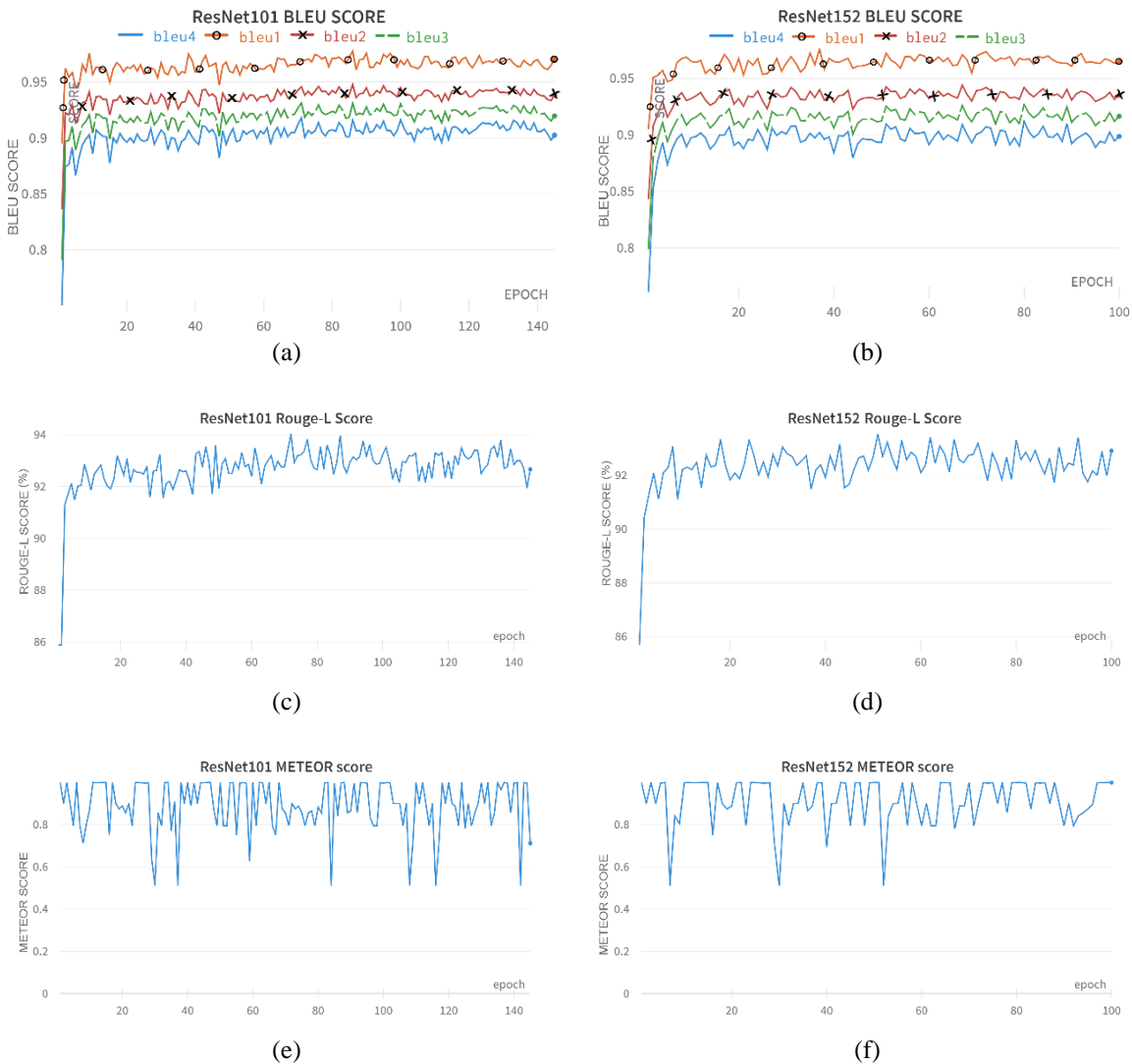


Figure 5. Graph of BLEU, ROUGE-L, METEOR evaluation for (a) ResNet101, (b) ResNet 152, (c) ResNet101, (d) ResNet 152, (e) ResNet101 and (f) ResNet 152

Table 3. Models score

Evaluation score (%)	Architecture	
	ResNet101	ResNet152
BLEU-1	96.946%	97.236%
BLEU-2	94.589%	94.266%
BLEU-3	93.144%	92.543%
BLEU-4	91.811%	90.965%
METEOR	79.380%	79.950%
ROUGE-L	94.0337%	93.5453%

The best model is taken from the results of the ResNet101 encoder training in the 70th epoch when the model gets the minimum validation loss value and the highest BLEU-4 evaluation value. Furthermore, we also tested the best modeling on the test data set, a new data set that the model has never seen before. Some of the testing results can be seen in Figures 6 to 9. In general, the testing showed good prediction results for each image, although there were still some prediction errors and poor wording in some prediction results. In Figure 6, the three reference captions read ('#0 The car parked in the middle of the sidewalk.', '#1 Watch out for cars in the middle of the sidewalk.', '#2 Be careful of parking cars in the middle of the sidewalk.'). It can be seen that the model can recognize cars and pavement objects well and gives predictive text results that are similar to the caption in the reference, which reads 'the car parked in the middle of the pavement'. Figure 9 shows the three reference captions (#0 hole). In the middle of the sidewalk, #1 pavement with holes, #2 Be careful there is a hole in the middle of the sidewalk). The model appears to show good prediction results by recognizing the 'hole' object in the middle of the sidewalk and displaying the sentence 'hole in the middle of the sidewalk' as a prediction. This result is similar to one of the captions in the reference.

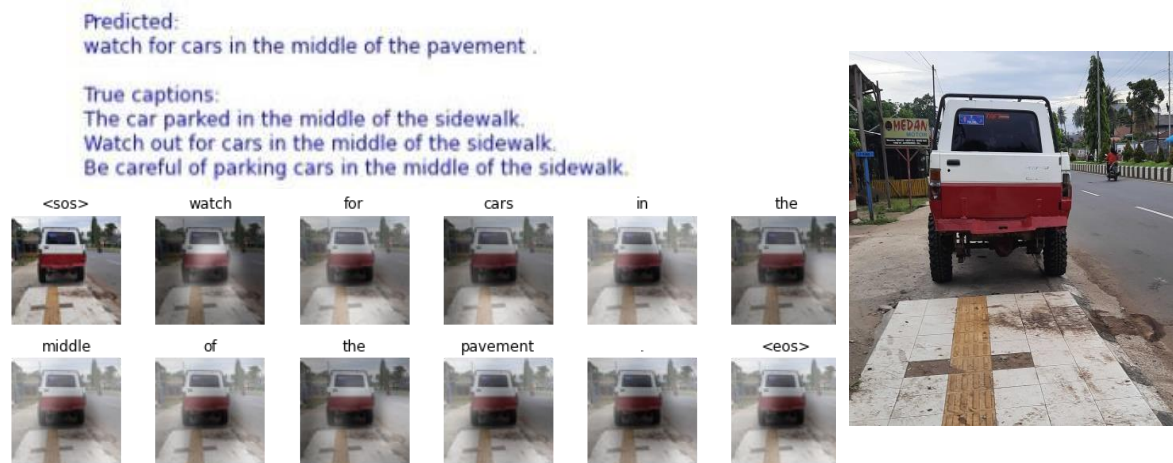


Figure 6. Caption prediction 1

The unsatisfactory result is shown in Figure 7 with three reference captions (#0 Motorcycle parked in the middle of the sidewalk, #1 Be careful of parking motorbikes in the middle of the sidewalk, #3 Be careful the motorbikes parked in the middle of the sidewalk). The model seems to misrecognize the motor object as 'cars' to display the prediction result for the caption 'watch for cars in the middle of the.' Meanwhile, in Figure 8, the model displays the result of the caption prediction, which is quite different from the reference caption. The reference caption writes an image caption in the form (#0 Ditch on the left side of the road, #1 The road with the gutter on the left side, #2 Be careful there is a ditch on the left side of the road). However, the model seems to recognize the ditch and sidewalk objects in the image so that it displays a caption prediction in the form of 'sidewalk with a gutter on the left side.' Although slightly different, this prediction seems reasonable because if we look closely at Figure 8, it shows a visual scene in the form of a sidewalk with a gutter on its left side instead of a sidewalk and a road. Interestingly enough, the model can recognize the gutter's position on the sidewalk's left side.

Predicted:
watch for cars in the middle of the .

True captions:
Motorcycle parked in the middle of the sidewalk.
Be careful of parking motorbikes in the middle of the sidewalk.
Be careful there are motorbikes parked in the middle of the sidewalk.



Figure 7. Caption prediction 2

Predicted:
sidewalk with gutter on left side .

True captions:
Ditch on the left side of the road.
The road with the gutter on the left.
Be careful there is a ditch on the left side of the road.

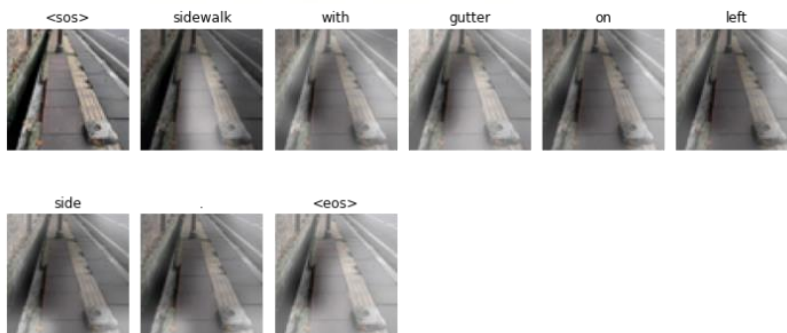


Figure 8. Caption prediction 3

Predicted:
hole in the middle of the sidewalk .

True captions:
Hole in the middle of the sidewalk.
Pavement with holes.
Be careful there is a hole in the middle of the sidewalk.

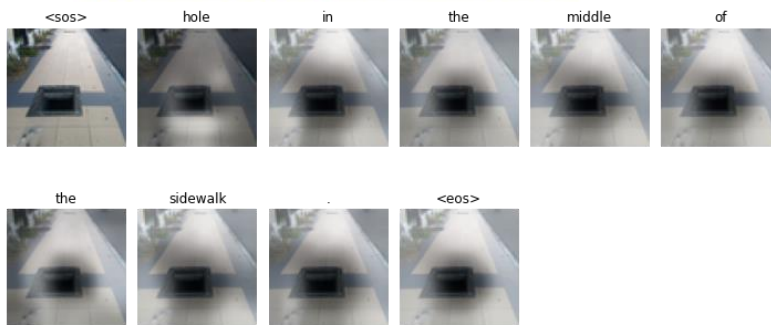


Figure 9. Caption prediction 4

4. CONCLUSION

An image captioning model was developed in this study to describe visual sights in an outdoor environment, thereby assisting BVI people's navigation and creating a mental map for obstacles or obstacles that may be encountered in the outdoor environment. The image captioning model employs an encoder-decoder approach with an attention layer based on CNN and LSTM. ResNet 101 and ResNet 152 are the CNN architectures used in this study. The dataset contains 2788 images divided into seven categories: puddles, sidewalk potholes, cars, motorcycles, intersections, ditches, and zebra crosses. Based on the evaluation metrics of BLEU, METEOR, and ROUGE-L, it can be concluded that the image captioning model generated from the training process using a set of datasets in open public spaces has pretty good performance. The ResNet 101 encoder has a slightly better performance than the ResNet 152. The image captioning model with the ResNet101 encoder has the best score on evaluation with BLEU and ROUGE-L metrics. In BLEU-4, which calculates precision weights up to 4 grams, the ResNet101 encoder gets a score of 91.811%, which is the highest BLEU-4 score in the training process, while the ROUGE-L evaluation metric shows a score of 94.0337%. Further development for future research, datasets can be added that discuss more diverse scenes in the public space environment and create richer captions by recognizing more objects and small details in images and making more diverse variations of descriptive sentences.




REFERENCES

- [1] A. Bhowmick and S. M. Hazarika, "An insight into assistive technology for the visually impaired and blind people: state-of-the-art and future trends," *Journal on Multimodal User Interfaces*, vol. 11, no. 2, pp. 149–172, Jun. 2017, doi: 10.1007/s12193-016-0235-6.
- [2] N. S. Ramaiah, R. Mishra, A. Sharma, and T. Iwoni, "A research paper on third eye for blind," in *Integrated Emerging Methods of Artificial Intelligence & Cloud Computing*, 2022, pp. 390–399.
- [3] R. G. Golledge, "Geography and the disabled: a survey with special reference to vision impaired and blind populations," *Transactions of the Institute of British Geographers*, vol. 18, no. 1, p. 63, 1993, doi: 10.2307/623069.
- [4] B. Kuriakose, R. Shrestha, and F. E. Sandnes, "Tools and technologies for blind and visually impaired navigation support: a review," *IETE Technical Review*, vol. 39, no. 1, pp. 3–18, Jan. 2022, doi: 10.1080/02564602.2020.1819893.
- [5] F. Reggiswarashari and S. W. Sihwi, "Speech emotion recognition using 2D-convolutional neural network," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 6, p. 6594, Dec. 2022, doi: 10.11591/ijece.v12i6.pp6594-6601.
- [6] E. Rakun, I. G. B. H. Widhinugraha, and N. F. Putra Setyono, "Word recognition and automated epenthesis removal for Indonesian sign system sentence gestures," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 3, p. 1402, Jun. 2022, doi: 10.11591/ijeecs.v26.i3.pp1402-1414.
- [7] M. H. Ismail, S. A. Dawwd, and F. H. Ali, "Dynamic hand gesture recognition of Arabic sign language by using deep convolutional neural networks," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 2, p. 952, Feb. 2022, doi: 10.11591/ijeecs.v25.i2.pp952-962.
- [8] N. H. Ali, M. E. Abdulmunem, and A. E. Ali, "Constructed model for micro-content recognition in lip reading based deep learning," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2557–2565, Oct. 2021, doi: 10.11591/eei.v10i5.2927.
- [9] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: transforming objects into words," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 11137–11147.
- [10] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, 2009.
- [11] M. D. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Computing Surveys*, vol. 51, no. 6, pp. 1–36, Nov. 2019, doi: 10.1145/3295748.
- [12] D. S. McNamara, L. K. Allen, S. A. Crossley, M. Dascalu, and C. A. Perret, "Natural language processing and learning analytics," in *Handbook of Learning Analytics*, Society for Learning Analytics Research (SoLAR), 2017, pp. 93–104.
- [13] R. Mitkov, Ed., *The Oxford handbook of computational linguistics 2nd edition*. Oxford University Press, 2014.
- [14] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: an introduction," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 544–551, Sep. 2011, doi: 10.1136/amiajnl-2011-000464.
- [15] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 4904–4912, doi: 10.1109/ICCV.2017.524.
- [16] S. Katiyar and S. Kumar, "Comparative evaluation of CNN architectures for image caption generation," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, 2020, doi: 10.14569/IJACSA.2020.0111291.
- [17] G. Ding, M. Chen, S. Zhao, H. Chen, J. Han, and Q. Liu, "Neural image caption generation with weighted training and reference," *Cognitive Computation*, vol. 11, no. 6, pp. 763–777, Dec. 2019, doi: 10.1007/s12559-018-9581-x.
- [18] M. Hussain, J. J. Bird, and D. R. Faria, "A study on cnn transfer learning for image classification," in *UK Workshop on Computational Intelligence*, 2019, pp. 191–202.
- [19] A. A. Elngar *et al.*, "Image classification based on CNN: a survey," *Journal of Cybersecurity and Information Management*, vol. 6, no. 1, p. PP. 18–50, 2021, doi: 10.54216/JCIM.060102.
- [20] M. Alagarsamy, J. M. J. Vedam, N. Shanmugam, P. M. Eswaran, G. Sankaraiyer, and K. Suriyan, "Performing the classification of pulsation cardiac beats automatically by using CNN with various dimensions of kernels," *International Journal of Reconfigurable and Embedded Systems (IJRES)*, vol. 11, no. 3, p. 249, Nov. 2022, doi: 10.11591/ijres.v11.i3.pp249-257.
- [21] S. Bekhet, A. M. Alghamdi, and I. F. Taj-Eddin, "Gender recognition from unconstrained selfie images: a convolutional neural network approach," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 2, p. 2066, Apr. 2022, doi: 10.11591/ijece.v12i2.pp2066-2078.
- [22] A. Issam, A. K. Mounir, E. M. Saida, and E. M. Fatma, "Financial sentiment analysis of tweets based on deep learning approach," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 3, p. 1759, Mar. 2022,




- doi: 10.11591/ijeecs.v25.i3.pp1759-1770.
- [23] Y. Hu and G. Mogos, "Music genres classification by deep learning," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 2, p. 1186, Feb. 2022, doi: 10.11591/ijeecs.v25.i2.pp1186-1198.
- [24] H. Abdulkarim and M. Z. Al-Faiz, "Online multiclass EEG feature extraction and recognition using modified convolutional neural network method," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 5, p. 4016, Oct. 2021, doi: 10.11591/ijece.v11i5.pp4016-4026.
- [25] A. W. Sugiyarto, A. M. Abadi, and S. Sumarna, "Classification of heart disease based on PCG signal using CNN," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 19, no. 5, p. 1697, Oct. 2021, doi: 10.12928/telkomnika.v19i5.20486.
- [26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.
- [27] Z. Cai and N. Vasconcelos, "Cascade R-CNN: delving into high quality object detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 6154–6162, doi: 10.1109/CVPR.2018.00644.
- [28] P. Vasavi, A. Punitha, and T. V. Narayana Rao, "Crop leaf disease detection and classification using machine learning and deep learning algorithms by visual symptoms: a review," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 2, p. 2079, Apr. 2022, doi: 10.11591/ijece.v12i2.pp2079-2086.
- [29] M. Shivanandappa and M. M. Patil, "Extraction of image resampling using correlation aware convolution neural networks for image tampering detection," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 3, p. 3033, Jun. 2022, doi: 10.11591/ijece.v12i3.pp3033-3043.
- [30] M. Mohebbanaaz, Y. P. Sai, and L. V. R. Kumari, "Detection of cardiac arrhythmia using deep CNN and optimized SVM," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 24, no. 1, p. 217, Oct. 2021, doi: 10.11591/ijeecs.v24.i1.pp217-225.
- [31] P. Wijonarko and A. Zahra, "Spoken language identification on 4 Indonesian local languages using deep learning," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 6, pp. 3288–3293, Dec. 2022, doi: 10.11591/eei.v11i6.4166.
- [32] M. V. V. Prasad Kantipud and S. Kumar, "A computationally efficient learning model to classify audio signal attributes," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 5, p. 4926, Oct. 2022, doi: 10.11591/ijece.v12i5.pp4926-4934.
- [33] T. Mathu and K. Raimond, "A novel deep learning architecture for drug named entity recognition," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 19, no. 6, p. 1884, Dec. 2021, doi: 10.12928/telkomnika.v19i6.21667.
- [34] A. Darmawahyuni, S. Nurmaini, M. N. Rachmatullah, F. Firdaus, and B. Tutuko, "Unidirectional-bidirectional recurrent networks for cardiac disorders classification," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 19, no. 3, p. 902, Jun. 2021, doi: 10.12928/telkomnika.v19i3.18876.
- [35] I. W. A. Suranata, I. N. K. Wardana, N. Jawas, and I. K. A. A. Aryanto, "Feature engineering and long short-term memory for energy use of appliances prediction," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 19, no. 3, p. 920, Jun. 2021, doi: 10.12928/telkomnika.v19i3.17882.
- [36] C. G. Pachon-Suescun, J. O. Pinzon-Arenas, and R. Jimenez-Moreno, "Abnormal gait detection by means of LSTM," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 2, p. 1495, Apr. 2020, doi: 10.11591/ijece.v10i2.pp1495-1506.
- [37] G. L. A. Kumari, P. Padmaja, and J. G. Suma, "A novel method for prediction of diabetes mellitus using deep convolutional neural network and long short-term memory," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 1, p. 404, Apr. 2022, doi: 10.11591/ijeecs.v26.i1.pp404-413.
- [38] M. A. H. Muhammad Fadzli, M. F. Abu Hassan, and N. Ibrahim, "Explicit kissing scene detection in cartoon using convolutional long short-term memory," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 1, pp. 213–220, Feb. 2022, doi: 10.11591/eei.v11i1.3542.
- [39] D. Munandar, A. F. Rozie, and A. Arisal, "A multi domains short message sentiment classification using hybrid neural network architecture," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 4, pp. 2181–2191, Aug. 2021, doi: 10.11591/eei.v10i4.2790.
- [40] S. Bhanja and A. Das, "A hybrid deep learning model for air quality time series prediction," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 3, p. 1611, Jun. 2021, doi: 10.11591/ijeecs.v22.i3.pp1611-1618.
- [41] S. Bai and S. An, "A survey on automatic image caption generation," *Neurocomputing*, vol. 311, pp. 291–304, Oct. 2018, doi: 10.1016/j.neucom.2018.05.080.
- [42] S. Takkar, A. Jain, and P. Adlakha, "Comparative study of different image captioning models," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, Apr. 2021, pp. 1366–1371, doi: 10.1109/ICCMC51019.2021.9418451.
- [43] 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)*, Jun. 2015, pp. 3156–3164, doi: 10.1109/CVPR.2015.7298935.
- [44] Hartatik, H. Al Fatta, and U. Fajar, "Captioning image using convolutional neural network (CNN) and long-short term memory (LSTM)," in *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Dec. 2019, pp. 263–268, doi: 10.1109/ISRITI48646.2019.9034562.
- [45] H. Sharma and A. S. Jalal, "Incorporating external knowledge for image captioning using CNN and LSTM," *Modern Physics Letters B*, vol. 34, no. 28, p. 2050315, Oct. 2020, doi: 10.1142/S0217984920503157.
- [46] J. Bineeshia, "Image caption generation using CNN-LSTM based approach," 2021, doi: 10.4108/eai.7-12-2021.2314958.
- [47] K. R. Suresh, A. Jarapala, and P. V Sudeep, "Image captioning encoder–decoder models using CNN-RNN architectures: A comparative study," *Circuits, Systems, and Signal Processing*, vol. 41, no. 10, pp. 5719–5742, Oct. 2022, doi: 10.1007/s00034-022-02050-2.
- [48] M. T. Younis, Y. T. Younus, J. N. Hasoon, A. H. Fadhil, and S. A. Mostafa, "An accurate Alzheimer's disease detection using a developed convolutional neural network model," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 4, pp. 2005–2012, Aug. 2022, doi: 10.11591/eei.v11i4.3659.
- [49] F. M. J. M. Shamrat *et al.*, "Analysing most efficient deep learning model to detect COVID-19 from computer tomography images," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 1, p. 462, Apr. 2022, doi: 10.11591/ijeecs.v26.i1.pp462-471.
- [50] V. Atliha and D. Sesok, "Comparison of VGG and resnet used as encoders for image captioning," in *2020 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*, Apr. 2020, pp. 1–4, doi: 10.1109/eStream50540.2020.9108880.
- [51] M. Bhalekar, S. Sureka, S. Joshi, and M. Bedekar, "Generation of image captions using VGG and ResNet CNN models cascaded with RNN approach," 2020, pp. 27–42.

- [52] A. A. Nugraha, A. Arifianto, and Suyanto, "Generating image description on Indonesian language using convolutional neural network and gated recurrent unit," in *2019 7th International Conference on Information and Communication Technology (ICoICT)*, Jul. 2019, pp. 1–6, doi: 10.1109/ICoICT.2019.8835370.
- [53] P. Phyu Khaing and M. The` Yu, "Attention-based deep learning model for image captioning: a comparative study," *International Journal of Image, Graphics and Signal Processing*, vol. 11, no. 6, pp. 1–8, Jun. 2019, doi: 10.5815/ijgsp.2019.06.01.
- [54] A. A. Jadhav, J. Kundale, A. Y. Joshi, and A. V. Kale, "Neural dense captioning with visual attention," in *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*, Jun. 2021, pp. 622–637, doi: 10.1109/CSNT51715.2021.9509657.
- [55] M. A. Al-Malla, A. Jafar, and N. Ghneim, "Image captioning model using attention and object features to mimic human image understanding," *Journal of Big Data*, vol. 9, no. 1, p. 20, Dec. 2022, doi: 10.1186/s40537-022-00571-w.
- [56] D. H. Fudholi *et al.*, "Image captioning with attention for smart local tourism using efficientnet," *IOP Conference Series: Materials Science and Engineering*, vol. 1077, no. 1, p. 12038, Feb. 2021, doi: 10.1088/1757-899X/1077/1/012038.
- [57] L. Chen, Y. He, and L. Fan, "Let the robot tell: Describe car image with natural language via LSTM," *Pattern Recognition Letters*, vol. 98, pp. 75–82, Oct. 2017, doi: 10.1016/j.patrec.2017.09.007.
- [58] C. Fan, Z. Zhang, and D. J. Crandall, "Deepdiary: Lifelogging image captioning and summarization," *Journal of Visual Communication and Image Representation*, vol. 55, pp. 40–55, Aug. 2018, doi: 10.1016/j.jvcir.2018.05.008.
- [59] G. Costa, A. Gusberti, J. P. Graffigna, M. Guzzo, and O. Nasisi, "Mobility and orientation aid for blind persons using artificial vision," *Journal of Physics: Conference Series*, vol. 90, p. 12090, Nov. 2007, doi: 10.1088/1742-6596/90/1/012090.
- [60] J. José, M. Farrajota, J. M. F. Rodrigues, and J. M. H. du Buf, "The SmartVision local navigation aid for blind and visually impaired persons," *International Journal of Digital Content Technology and its Applications*, vol. 5, no. 5, pp. 362–375, May 2011, doi: 10.4156/jdcta.vol5.issue5.40.
- [61] C. Yi and Y. Tian, "Assistive text reading from complex background for blind persons," in *Camera-Based Document Analysis and Recognition: 4th International Workshop*, Beijing, China, 2012, pp. 15–28.
- [62] A. N. Lapyko, L.-P. Tung, and B.-S. P. Lin, "A cloud-based outdoor assistive navigation system for the blind and visually impaired," in *2014 7th IFIP Wireless and Mobile Networking Conference (WMNC)*, May 2014, pp. 1–8, doi: 10.1109/WMNC.2014.6878884.
- [63] C. K. Lakde and P. S. Prasad, "Navigation system for visually impaired people," in *2015 International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC)*, Apr. 2015, pp. 93–98, doi: 10.1109/ICCPEIC.2015.7259447.
- [64] F. Ahmed, M. S. Mahmud, R. Al-Fahad, S. Alam, and M. Yeasin, "Image captioning for ambient awareness on a sidewalk," in *2018 1st International Conference on Data Intelligence and Security (ICDIS)*, Apr. 2018, pp. 85–91, doi: 10.1109/ICDIS.2018.00020.
- [65] B. Makav and V. Kilic, "A new image captioning approach for visually impaired people," in *2019 11th International Conference on Electrical and Electronics Engineering (ELECO)*, Nov. 2019, pp. 945–949, doi: 10.23919/ELECO47770.2019.8990630.
- [66] B. Makav and V. Kilic, "Smartphone-based image captioning for visually and hearing impaired," in *2019 11th International Conference on Electrical and Electronics Engineering (ELECO)*, Nov. 2019, pp. 950–953, doi: 10.23919/ELECO47770.2019.8990395.
- [67] S. Saha, F. H. Shakal, and M. Mahmood, "Visual, navigation and communication aid for visually impaired person," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 2, p. 1276, Apr. 2021, doi: 10.11591/ijece.v11i2.pp1276-1283.
- [68] H. Mohd Nasir, N. M. A. Brahin, M. M. Mohamed Aminuddin, M. S. Mispan, and M. F. Zulkifli, "Android based application for visually impaired using deep learning approach," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 4, p. 879, Dec. 2021, doi: 10.11591/ijai.v10.i4.pp879-888.
- [69] M. Grover, R. Rathi, C. Manchanda, K. Garg, and R. Beniwal, "AI optics: Object recognition and caption generation for blinds using deep learning methodologies," in *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, Feb. 2021, pp. 354–359, doi: 10.1109/ICCCIS51004.2021.9397143.
- [70] H. Ahsan, D. Bhatt, K. Shah, and N. Bhalla, "Multi-modal image captioning for the visually impaired," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, 2021, pp. 53–60, doi: 10.18653/v1/2021.naacl-srw.8.
- [71] S. P. Manay, S. A. Yaligar, Y. Thathva Sri Sai Reddy, and N. J. Saunshimath, "Image captioning for the visually impaired," 2022, pp. 511–522.
- [72] A. Arora, A. Grover, R. Chugh, and S. S. Reka, "Real time multi object detection for blind using single shot multibox detector," *Wireless Personal Communications*, vol. 107, no. 1, pp. 651–661, Jul. 2019, doi: 10.1007/s11277-019-06294-1.
- [73] A. Stangl, N. Verma, K. R. Fleischmann, M. R. Morris, and D. Gurari, "Going beyond one-size-fits-all image descriptions to satisfy the information wants of people who are blind or have low vision," in *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, Oct. 2021, pp. 1–15, doi: 10.1145/3441852.3471233.
- [74] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 4945–4949, doi: 10.1109/ICASSP.2016.7472618.
- [75] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 2001, p. 311, doi: 10.3115/1073083.1073135.
- [76] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, 2018, pp. 186–191, doi: 10.18653/v1/W18-6319.
- [77] A. Lavie and M. J. Denkowski, "The meteor metric for automatic evaluation of machine translation," *Machine Translation*, vol. 23, no. 2–3, pp. 105–115, Sep. 2009, doi: 10.1007/s10590-009-9059-4.
- [78] A. Lavie and A. Agarwal, "Meteor," in *Proceedings of the Second Workshop on Statistical Machine Translation - StatMT '07*, 2007, pp. 228–231, doi: 10.3115/1626355.1626389.
- [79] M. Barbella and G. Tortora, "Rouge metric evaluation for text summarization techniques," *SSRN Electronic Journal*, 2022, doi: 10.2139/ssrn.4120317.
- [80] F. Kiyoumars, "Evaluation of automatic text summarizations based on human summaries," *Procedia - Social and Behavioral Sciences*, vol. 192, pp. 83–91, Jun. 2015, doi: 10.1016/j.sbspro.2015.06.013.




BIOGRAPHIES OF AUTHORS

Ruvita Faurina    received the S.T. degree in informatics from University of Bengkulu, Indonesia and the M.Eng. degree in electrical engineering and information technology from University of Gadjah Mada, Indonesia, respectively. Currently, she is an Assistant Professor at the Department of Informatics, University of Bengkulu. Her research interests include artificial intelligence, natural language processing, image captioning and attention mechanism. She can be contacted at email: ruvita.faurina@unib.ac.id.






Anisa Jelita    currently is an undergraduate student in department of Informatics, University of Bengkulu. Her research includes artificial intelligence and image captioning. She can be contacted at email: anisaarpan99@gmail.com.



Arie Vatesia    received the S.T. degree in informatics from University of Bengkulu, Indonesia, M.T.I degree in Information Technology from The University of Indonesia, and Ph.D degree in Computer Science from University Of Birmingham, United Kingdom. Currently, she is an Associate Professor at the Department of Informatics, University of Bengkulu. Her research interests include artificial intelligence, geoinformatics, spatial analysis, GIS and data mining. She can be contacted at email: arie.vatesia@unib.ac.id.



Indra Agustian    received the S.T. and M.Eng. degrees in electrical engineering from Electrical Engineering, University of Gadjah Mada, Indonesia. Currently, he is an Assistant Professor at the Department of Electrical Engineering, University of Bengkulu, Indonesia. His research interests include control engineering, robotics and artificial intelligence. He can be contacted at email: indraagustian@unib.ac.id.