Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 19 No. 9 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijim.v19i09.53441

PAPER

Attention-Driven Image Captioning for Mobile Accessibility of the Visually Impaired

Dessy Santi[1,2], Amil Ahmad Ilham[3](✉), Syafaruddin[1], Ingrid Nurtanio[3]

[1]Department of Electrical Engineering, Universitas Hasanuddin, Gowa, Sulawesi Selatan, Indonesia

[2]Department of Information Technology, Universitas Tadulako, Palu, Sulawesi Tengah, Indonesia

[3]Department of Informatics, Universitas Hasanuddin, Gowa, Sulawesi Selatan, Indonesia

amil@unhas.ac.id

ABSTRACT

In a world increasingly reliant on visual information, individuals with visual impairments face significant challenges in understanding their environment. This paper introduces an attention-based image captioning model to improve accessibility for visually impaired users. The model integrates ResNet-152 for visual feature extraction, long short-term memory (LSTM) for text processing, and an attention mechanism to generate contextual image descriptions. Captured images are processed via a mobile device, then the description text is translated into Bahasa and converted to speech in real-time using text-to-speech technology. The system shows an average inference time of 2.99 seconds per image, enabling real-time use. The model is tested on the Flickr dataset and new datasets covering a variety of environments and object interactions. Experimental results show superior performance on the Flickr dataset (bilingual evaluation understudy (BLEU)-1: 0.59, metric for evaluation of translation with explicit ordering (METEOR): 0.25). Performance on real-world datasets is slightly lower, indicating challenges in generalizing to scenarios with occluded objects and inconsistent text. Future research will focus on scaling up real-world datasets, adversarial training, and integrating the system into devices such as smart glasses or canes for wider accessibility.

KEYWORDS

attention, image captioning, mobile accessibility, ResNet, visually impaired

1 INTRODUCTION

The development of numerous ways to increase accessibility for those with visual impairments has been made possible by technological advancements. The difficulty they have most often stems from the restricted availability of visual information in their daily surroundings. One possible way to deal with this problem is to use image captioning technology, which attempts to automatically describe images [1]. Image captioning allows people with visual disabilities to receive relevant text information about what is in an image, giving them a better understanding of the visual environment around them [2].

Santi, D., Ilham, A.A., Syafaruddin, Nurtanio, I. (2025). Attention-Driven Image Captioning for Mobile Accessibility of the Visually Impaired. *International Journal of Interactive Mobile Technologies (iJIM)*, 19(9), pp. 4–18. https://doi.org/10.3991/ijim.v19i09.53441

Article submitted 2024-11-22. Revision uploaded 2025-02-25. Final acceptance 2025-02-25.

However, traditional image captioning models face several limitations. Most early models are only able to provide simple descriptions without taking into account deeper contextual details, which are critical for people with visual disabilities [3–5]. The use of this technology may be limited by descriptions that are overly general or taken out of context. Users require precise and in-depth details regarding the objects, how they interact with one another, and the environments in which they are found. As a result, the descriptions generated by image captioning algorithms need to be improved.

Using attention mechanisms is one strategy that has shown to be successful in improving the relevance and accuracy of descriptions [6]. This mechanism works by focusing attention on more relevant parts of the image, allowing the model to select the most important objects or details in a given context. Attention mechanisms, both in the form of soft attention and hard attention, have been shown to provide significant improvements in image captioning performance, especially in producing more in-depth and contextual descriptions [7–11].

In the context of developing mobile technology for people with visual disabilities, the use of attention mechanisms can open up new opportunities for creating more responsive and informative applications [12–14]. Mobile devices, which are increasingly used by people with visual disabilities, require systems that are not only energy efficient but also capable of providing fast and accurate descriptions in a variety of situations. For example, users may rely on mobile applications to identify objects in the surrounding environment, read text from images, or understand interactions between objects in public spaces. Attention mechanisms can help generate more specific descriptions according to the needs of users in certain situations.

For the purpose of enhancing technology's usefulness for visually challenged people, this study puts forth a suggestion. The plan is to develop an Android app that allows users to take images of their surroundings. Next, the app will use a unique model that generates captions for the photographs to represent what it sees. The project's main goal is to improve the captioning model's performance on mobile devices, particularly for the benefit of visually challenged users. Users may hear descriptions of what's around them right away because the software will read the captions aloud in real-time. They will be able to navigate their environment more easily and gain a greater understanding of their surroundings as a result. The captions will be available in Indonesian, making the content more accessible to viewers who know the language.

The study is divided up into various sections. Section 2 examines current study on picture description and improving accessibility of images on mobile devices for visually impaired individuals. The Section 3 describes the study methodology and provides an explanation of a novel model that generates visual descriptions through a unique focus strategy. The testing setup, findings, and their significance are discussed in detail in Section 4. Section 5 provides a comprehensive summary, concisely highlights the key findings, and offers valuable suggestions for future research directions.

2 LITERATURE REVIEW

2.1 The evolution of accessibility technologies for individuals with visual impairments

In recent decades, technology has played an increasingly important role in enhancing accessibility for individuals with visual impairments. Image recognition

and natural language processing (NLP) technologies have advanced rapidly, enabling the creation of systems that assist visually impaired individuals in accessing visual information. One approach used is image captioning, a technology that automatically generates textual descriptions from an image, allowing visually impaired users to "see" the world through words [15]. However, early image captioning technology still faces many limitations, particularly in providing detailed and relevant descriptions for individuals with visual impairments. These models often generate descriptions that are too generic or not specific enough to meet accessibility needs [16]. This highlights the urgent need to develop more complex and context-aware models.

2.2 Challenges in traditional image captioning

Before the introduction of attention mechanisms, traditional image captioning models often struggled to provide contextually appropriate descriptions of images. Most early models were based on convolutional neural network (CNN) architectures for visual feature extraction and recurrent neural network (RNN) or long short-term memory (LSTM) for generating descriptive text. The combination of CNN and LSTM is widely used, with CNN extracting spatial features from images and LSTM processing sequential data, making it effective for tasks involving both visual and temporal information, as in research [17]. Although this method performed reasonably well in understanding the overall content of an image, the models often missed important details and failed to capture interactions between objects within the image. For instance, when two objects appear in the same image, traditional models might struggle to grasp the contextual relationship between those objects. A study by Xu et al. [18] explored how attention mechanisms could address this issue by allowing models to focus on specific parts of an image while generating captions. This approach enables the model to attend to relevant regions and capture more detailed and contextually accurate descriptions, especially in cases where multiple objects and their interactions need to be understood. This issue becomes more problematic in mobile applications designed for individuals with visual impairments, where detail and context are crucial.

2.3 Attention mechanism in image captioning

To overcome the limitations of traditional image captioning, attention mechanisms are introduced into image captioning models. These mechanisms allow the model to focus on more relevant parts of the image, improving the model's ability to understand relationships between objects and provide more contextual descriptions. In this context, soft attention and hard attention are two commonly used approaches. Soft attention allows the model to distribute its focus across the entire image with varying levels of emphasis, enabling the model to select the most important parts of the image for each word in the description [19]. Conversely, hard attention is more focused on specific parts of the image, but this approach requires more complex sampling methods to select the most important regions [20]. Research by Anderson et al. [21] demonstrates that attention mechanisms can significantly enhance captioning performance, particularly in handling images with multiple objects or complex compositions.

2.4 Implementation of attention mechanisms for the visually impaired

The use of mobile technology in daily life by individuals with visual impairments is increasingly on the rise. Text-to-speech technology empowers visually impaired individuals by converting text or visual descriptions into audible speech, enabling easier access to information and improved daily navigation [22]. In the study [23], there are several dimensions that have a significant impact on the satisfaction of moderate and severe blind users who use mobile applications. These dimensions are efficiency, effectiveness, satisfaction, errors, accessibility, and understanding. Attention mechanisms help focus on objects relevant to the user, such as road signs, people's faces, or text in their surroundings. This enables users to gain a more contextual and immersive understanding of their environment.

Several studies have shown the advantages of using attention mechanisms in image captioning. Research by Zaman et al. [24] indicated that the model not only assists individuals with visual impairments but also those who are deaf-blind by converting images into text that can be read in Braille. This text is then processed by a microcontroller-based system using a push-pull solenoid arrangement (2×3) to produce tactile output for the visually impaired. They also found that with further optimization, mobile applications could generate relevant captions with quick response times, which is crucial in the context of real-time use by individuals with visual impairments.

In mobile applications, the implementation of attention mechanisms becomes more complex, particularly due to the processing power and battery limitations of mobile devices. Nevertheless, research indicates that attention mechanisms still hold significant potential for enhancing the quality of image captioning on mobile platforms. Attention mechanisms can be optimized for mobile devices by reducing the complexity of neural networks. Utilizing mobile devices facilitates accessibility for visually impaired individuals.

2.5 Challenges and future opportunities

Although attention mechanisms have provided numerous benefits in image captioning, several challenges remain in their implementation on mobile devices. Higher processing power requirements and battery consumption are two primary challenges. However, significant opportunities still exist to advance this technology. Future research could focus on developing more efficient model compression techniques and applying transfer learning to enhance performance without requiring excessive computational resources.

Additionally, research could explore the use of reinforcement learning to train more efficient attention models that can adapt their focus to the most relevant parts of an image based on user objectives. This could provide more tailored and responsive solutions, particularly in the context of mobile accessibility for individuals with visual impairments.

2.6 Evaluation metric

Bilingual evaluation understudy (BLEU). BLEU is the evaluation metric used [25]. In Equation (1) we can see the calculation of *n*-gram precision as the

ratio of the total number of *n*-gram frequencies in the candidate to the number of minimum frequency values of the matching *n*-grams. BLEU measures the similarity between *n*-grams of the predicted caption and the reference caption. The basic formula for BLEU can be seen in equation 1 as follows:

$$BLEU = BP.exp\left(\sum_{n=1}^{N} w_{n\log Pn}\right) \tag{1}$$

Where, *BP* is the brevity penalty, p_n is the precision for *n*-grams of length *n*, w_n is the weight for a given *n*-gram $\left(w_n = \frac{1}{N}\right)$, and N is the maximum length of *n*-grams to be evaluated (e.g., N = 4 for 4-gram BLEU). Equation 2 represents the penalty that BLEU applies via BP in the event that the forecast length is less than the reference:

$$BP = \begin{cases} 1, & if\ r > c \\ exp\left(1 - \frac{r}{c}\right), & if\ r \le c \end{cases} \tag{2}$$

Where, *c* is the total length of the prediction caption and *r* is the total length of the reference caption.

Metric for evaluation of translation with explicit ordering (METEOR). Combines precision and recall of the alignment between predicted and reference words. Equation 3 shows the formula for METEOR [26].

$$\text{METEOR} = F_{mean}.(1 - Penalty) \tag{3}$$

Where, $F_{mean} = \frac{10.Precision.recall}{9.Precision + Recall}$ is the F-mean score that combines precision and recall. Penalty is a penalty applied to broken or unordered word segments, as seen in equation 3.

$$Penalty = 0.5 \cdot \left(\frac{chunks}{matches}\right)^3 \tag{4}$$

Where, chunks are the number of unordered word fragments, and matches is the number of words that match between the predicted and reference captions.

3 MATERIAL AND METHOD

This study presents a comprehensive implementation of a CNN-based image captioning model combining ResNet152, LSTM, and Bahdanau Attention. The model processes images and text data using a systematic pipeline, including feature extraction, sequence generation, and attention mechanisms. Dataset preprocessing involved resizing images to 224 × 224 pixels and tokenizing captions using FastText embeddings. The trained model was deployed as a cloud-based API and integrated with text-to-speech technology to enhance accessibility for visually impaired users. The implementation achieves significant improvements in generating contextually relevant captions, as validated by BLEU and METEOR metrics. An overview of image captioning architecture using the Flickr30k dataset with CNN and LSTM with attention can be seen in Figure 1.

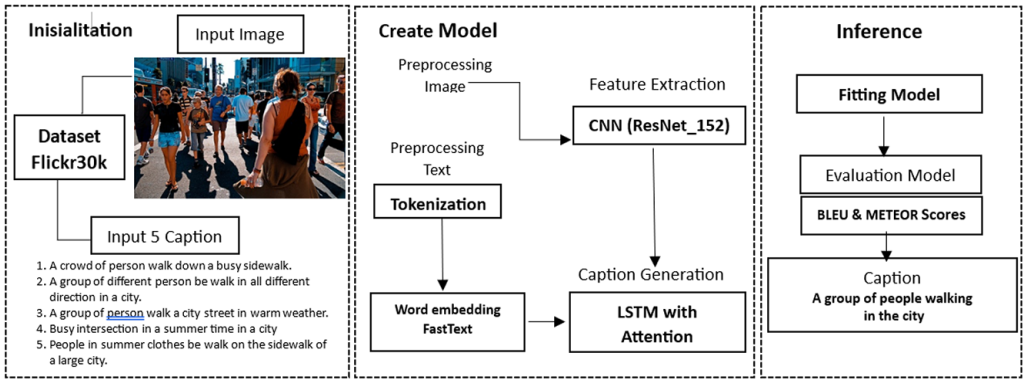

Fig. 1. Architecture overview: of Image captioning using Flickr30k Dataset with CNN and LSTM with attention

3.1 Dataset and preprocessing

For the purpose of enabling the model to produce captions based on visual input, this study set out to train it to correlate photos with related textual descriptions. Thirty thousand photos from Flickr [27] were used in the dataset, each with five different captions. Three subsets of the dataset were created: 60% were used to train the model, 20% were used for validation, and the remaining 20% were used to assess the model's performance. The photos were preprocessed to fulfill the model's input requirements and shrunk to 224×224 pixels in order to standardize them for processing. In terms of the textual data, all text in the captions was converted to lowercase, and superfluous characters were eliminated. The various phrases in the dataset were then used to create a lexicon, and each word was given a numerical token for quick access. Sentences longer than the maximum length were clipped, and captions were padded to match the longest sentence. Special tokens <start> and <end> were appended to the beginning and end of each caption to denote the sequence boundaries.

3.2 Image feature extraction

Following image preprocessing, the next step involves utilizing ResNet-152. The selection of ResNet-152, in combination with LSTM and Bahdanau Attention, is driven by its capability to extract fine-grained visual features and capture sequential dependencies in natural language descriptions. Moreover, ResNet-152 effectively mitigates the issue of information loss in deeper networks, leveraging its increased depth to learn complex and abstract features from images [28]. ResNet architecture is composed of multiple residual blocks, as illustrated in Figure 2. Its distinctive feature lies in the use of shortcut connections, also known as skip connections, which help preserve gradient flow and improve training efficiency in deep networks.

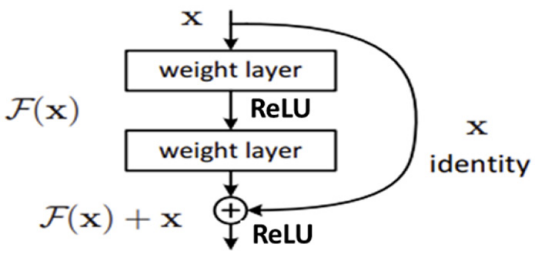

Fig. 2. Residual learning and building block

The image is passed through a ResNet152 network pre-trained on ImageNet. The last convolutional layer generates a 2048-dimensional feature vector for each spatial location, resulting in a $7 \times 7 \times 2048$ feature map. These features serve as input to the attention mechanism.

3.3 Caption generation: LSTM with attention mechanism

After the image features are extracted by CNN with ResNet-152 architecture, the features are used by LSTM decoder to generate a series of words (captions) one by one in natural language. At each time step, the decoder generates one word based on the context of the image and the previously generated words.

The attention mechanism is used to select the parts of the image (feature vectors) that are most relevant to generating a particular word. At each time step, the LSTM makes a decision to generate the next word based on two things:

1. The hidden state of the LSTM at the previous time step.
2. The context vector generated by the attention mechanism, which is a weighted sum of the image feature vectors based on relevance (attention scores).

Attention score computation process. At each time step, the LSTM hidden state (s_t) is computed. The relevance score between this hidden state and each image feature vector (h_i) is computed using a feedforward neural network function, as in the Bahdanau attention mechanism shown in equation 5 below:

$$e_{ti} = v_a^T \tanh\left(w_a \left[S_t; h_i\right]\right) \tag{5}$$

Where, s_t is the hidden state of the decoder at time step t, h_i is the feature vector of the encoder at time step i, and w_a is the weight matrix used for linear transformation of the hidden state decoder and the encoder feature vector, v_a is the weight vector used to generate relevance score, and e_{ti} are attention scores between hidden state decoder s_t.

The obtained scores are then converted into attention weights using softmax, which indicate how important each part of the image is at that time.

Context vector. The context vector is obtained by calculating the weighted sum of the image feature vectors after calculating the attention weights (α_{ti}). The important details of the currently most relevant image area are captured by this context vector. The equation for the context vector can be seen in equation 6.

$$c_t = \sum_{i-1}^{n} \alpha_{ti} h_i \tag{6}$$

Where, h_i is the feature vector from the encoder that represents the image parts. α_{ti} is the attention weight that shows how important the feature vector h_i is at time step t.

Caption generation. The context vector generated by the attention mechanism is then combined with the hidden state of the LSTM to generate the next word. This process is repeated for each generated word, with the part of the image that is paid attention to changing at each time step, depending on the context. This mechanism repeats until the entire caption is generated. At each time step, the model focuses attention on a different area of the image, resulting in a more detailed and contextual description. The entire image is represented as a single global feature vector in an attention-deficient LSTM model. Because the model is unable to concentrate on

certain, pertinent areas of the image, the descriptions that emerge are typically less precise or thorough. On the other hand, an attention-based LSTM allows the model to focus on distinct regions of the image for every phrase.

3.4 Implementation of the caption generation model with a mobile application

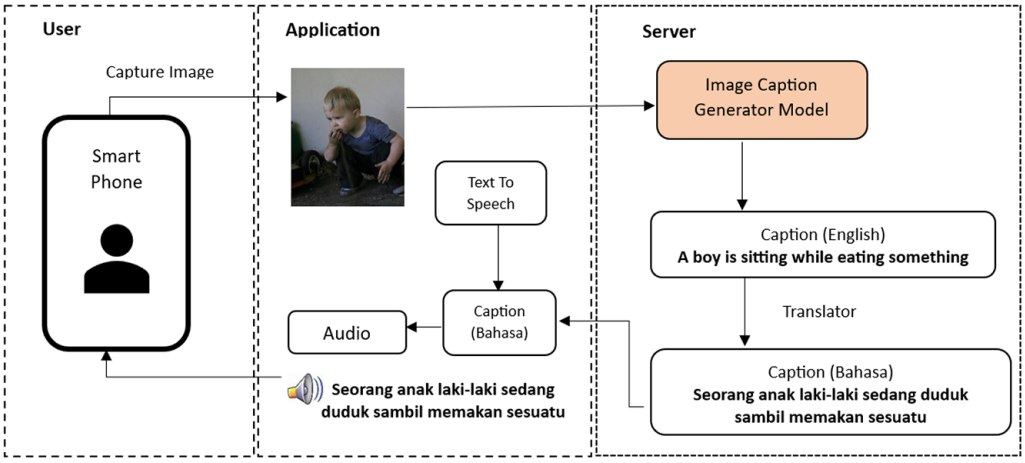

Fig. 3. Implementation of the caption generation model with a mobile application

The application architecture intended to automatically create captions for photos and provide them to users in text and audio forms is depicted in Figure 3. The goal of this implementation is to aid users with comprehending the content of taken photographs, particularly those who are visually impaired. The following is a description of the application workflow:

User interface: By utilizing the application to capture images, the user initiates the procedure.

Application layer:

- **Received image:** The application receives the captured image and processes it before preparing it for further use.
- **Caption text (English):** The application receives the English descriptive text about the image after it has been processed by the server. The application can display it or proceed to the next step.

Server layer:

- **Integrate API:** Python code that uses the Flask framework to build an API that takes image requests and generates a caption:

```
python
from flask import Flask, request, jsonify
app = Flask(__name__)
@app.route('/caption', methods=['POST'])
def caption_image():
    image = request.files['image']
    # Preprocessing image and running model inference
    caption = generate_caption(image)
    return jsonify({'caption': caption})
```

Overall, this code builds an API endpoint to accept an image and return the caption of that image in JSON format after processing it with a model.

- **Image processing:** The application sends the image to the server, where the English description is produced by the image caption model. The ResNet-152 image caption model and LSTM with attention mechanism
- **Translate to Bahasa:** After the English descriptive text is generated, it is routed to a different server module to be translated to Bahasa. This stage shows that bilingual output is supported by the system design.

Application layer:

- **Android client:** The following code is configured to send a request to the server, get a caption for an image, and handle the response or failure.

```
java
Retrofit retrofit = new Retrofit.Builder()
        .baseUrl("https://yourserver.com/")
        .addConverterFactory(GsonConverterFactory.create())
        .build();
ApiService apiService = retrofit.create(ApiService.class);
Call<CaptionResponse> call = apiService.getCaption(image);
call.enqueue(new Callback<CaptionResponse>() {
    @Override
    public void onResponse(Call<CaptionResponse> call, Response<CaptionResponse> response) {
        String caption = response.body().getCaption();
        // Show captions in the app
    }
    @Override
    public void onFailure(Call<CaptionResponse> call, Throwable t) {
        // Handle errors
    }
});
```

- **Translated descriptive text:** The application receives the translated descriptive text (in Bahasa), which can be displayed or further processed.
- **Text to speech:** The final stage involves utilizing a text-to-speech engine to translate descriptive text into spoken words so that people with visual impairments can hear the image description in Bahasa.

4 RESULTS AND DISCUSSION

The results of the study and experiment on the image captioning model with attention are discussed in this section.

4.1 Quantitative analysis model

To evaluate the performance of our image captioning model, we use the BLEU-1 to BLEU-4 evacuation metrics and METEOR. The generated captions have a BLEU-1 score of 0.59 and a METEOR score of 0.25. Based on the results, our approach is proven to improve BLEU-1 because the model is able to generate more precise words in the correct order, according to the reference. In addition, METEOR, which combines precision and recall, also improves because the model is able to handle

synonyms, stemming, and word matches better. In other words, this approach produces captions that are not only more accurate in terms of words but also better in reflecting the semantic meaning of the image, which contributes to the improvement of BLEU-1 and METEOR scores. As shown in Table 1 when compared to other models, our approach shows superior performance in both BLEU-1 and METEOR metrics.

Table 1. Evaluation metrics: BLEU and METEOR scores

MODEL	BLEU				METEOR
	1	2	3	4	
Attention-Enhanced ResNet-LSTM (Proposed Model)	**0.59**	**0.52**	**0.50**	**0.49**	**0.25**
Inceptionresnetv2-RNN and word embedding Glove with [29]	0.56	0.50	0.48	0.45	–
LSTM with Word Embedding Glove [30]	0.53	0.25	0.12	0.47	0.22

Note: Bold values indicate the best performance scores across the models for each metric.

4.2 Users qualitative evaluation model: User feedback trial

An experiment was carried out to assess the efficacy of a mobile application intended to help people with vision impairments. Wearing a blindfold, a sighted participant used an Android phone loaded with the program to find their way around a university campus. The goal was to evaluate how well the app described the surroundings for a user who is sight impaired. Though it periodically duplicated information, the software was able to identify objects such as stairs, a young man, a computer, and a whiteboard with success. The user thought the software was useful overall and proposed several changes, such as adding the ability to read text from signs, a capability that would necessitate the use of optical character recognition (OCR) technology. The experiment for the campus environment can be seen in Figure 4.

Fig. 4. Real time and real-world experiment: Human subject utilizing mobile accessibility

The attention mechanism will dynamically change emphasis to the areas of the image that are relevant for each word in the sentence "a man sitting in front of a computer." When creating "a man," the area of the image displaying the man's figure will catch the model's eye. Focus turns to the man's body as he forms the word "sit," indicating a sitting posture. Next, while producing "in front of," the man's spatial relationship with other objects such as the computer will be taken into consideration. Lastly, the model will concentrate on the region that displays the computer when producing "a computer," guaranteeing an accurate description that corresponds with the image's content. We also calculated the duration of

some software components to ensure that the program operates quickly in practical scenarios. The processing time for critical components and the total performance were monitored for ten different scenarios, and their effectiveness was assessed. The experimental results showed that the average duration time was 2.99 seconds for each image, as can be seen in Table 2.

Table 2. Duration of model inference and text-to-speech execution

#Image	Caption in English (s)	Bahasa Audio Captioning (s)	Total
1	1.99	1.25	3.24
2	1.98	1.22	3.2
3	1.88	1.25	3.13
4	1.98	1.22	3.2
5	1.79	1.21	3
6	1.43	1.1	2.53
7	1.78	1.1	2.88
8	1.98	1.02	3
9	1.54	1.01	2.55
10	1.97	1.2	3.17
Mean	1.832	1.158	2.99

4.3 Comparison of testing between the Flickr dataset and a new real-world dataset

Additionally, we compared the Flickr dataset with a test dataset of 1,000 photos from a real-world dataset that we generated and gathered from a variety of sources, such as public repositories, social media, and mobile phone cameras, and annotated. This dataset records a variety of difficult situations, including low light levels and partially obscured objects. In contrast to a real-world test dataset that contains a variety of unstructured scenarios, the comparison attempts to assess the model's generalization ability and robustness when evaluated on a carefully selected benchmark dataset (Flickr30k). Table 3 displays the comparing results.

Table 3. Test results of Flickr dataset and new real-world dataset

Dataset	BLEU-1	BLEU-4	METEOR
Flickr30k	0.59	0.29	0.25
Real-World Data	0.48	0.22	0.22

Table 3 shows that the model achieves higher BLEU-4 and METEOR scores on Flickr30k due to its structured and curated text. However, its performance on the real-world test dataset is lower, indicating challenges in generalizing to scenarios with occluded objects, ambiguous visual context, and inconsistent text. The lower performance on the real-world test dataset indicates that while the model outperforms the curated dataset, further improvements are needed to handle a variety of ambiguous real-world scenarios. Improvements such as incorporating domain adaptation or

fine-tuning on real-world data can improve its robustness. Future work will focus on augmenting the training dataset with samples from real-world scenarios and incorporating adversarial training to improve robustness to ambiguous inputs.

5 CONCLUSION

This study aims to develop and evaluate a system designed to assist individuals with visual impairments by providing easy-to-understand image descriptions. The system captures images, interprets their content, and converts the information into Bahasa and into spoken form, offering a practical solution to improving accessibility. Testing shows that the system effectively identifies objects and produces accurate descriptions, albeit with occasional repetition. The system operates efficiently, with a response time of just over two seconds, making it suitable for everyday use. However, its lower performance on a real-world test dataset highlights the need for improvements, such as domain adaptation, fine-tuning with real-world data, and adversarial training, to improve robustness. Future work will focus on expanding the training dataset, enabling text reading from images, and integrating the system into smart glasses or canes for real-time feedback and greater accessibility.

6 ACKNOWLEDGEMENT

With deep gratitude, we would like to express our sincere appreciation to Hasanuddin University for the support provided throughout this study. We extend our heartfelt thanks to our colleagues for their valuable discussions, constructive feedback, and continuous support, which have significantly contributed to the improvement of this study. Their insights and encouragement have been instrumental in refining our approach and enhancing the quality of this study. Additionally, we would like to acknowledge the Laboratory of the Department of Electrical Engineering, Hasanuddin University, for providing the necessary facilities and computational resources that enabled the implementation and evaluation of our proposed model. Finally, we express our deepest gratitude to our family and friends for their unwavering support and encouragement throughout this study journey. Their motivation and belief in our work have been a great source of strength.

7 REFERENCES

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *2015 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3156–3164. https://doi.org/10.1109/CVPR.2015.7298935

[2] P. Dognin *et al.*, "Image captioning as an assistive technology: Lessons learned from VizWiz 2020 challenge," *J. Artif. Intell. Res. (JAIR)*, vol. 73, pp. 437–459, 2022. https://doi.org/10.1613/jair.1.13113

[3] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in *2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7219–7228. https://doi.org/10.1109/CVPR.2018.00754

[4] P. Voditel *et al.*, "Image captioning – A deep learning approach using CNN and LSTM network," in *Proceedings – 2023 3rd International Conference on Pervasive Computing and Social Networking*, ICPCSN 2023, 2023, pp. 343–348. Available at: https://doi.org/10.1109/ICPCSN58827.2023.00062

[5] K. Anitha Kumari, C. Mouneeshwari, R. B. Udhaya, and R. Jasmitha, "Automated image captioning for Flickr8K dataset," in *Proc. Int. Conf. Artif. Intell., Smart Grid and Smart City Appl., AISGSC 2019*, L. Kumar, L. Jayashree, and R. Manimegalai, Eds., Springer, Cham, 2020, pp. 679–687. https://doi.org/10.1007/978-3-030-24051-6_62

[6] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2015. https://doi.org/10.48550/arXiv.1409.0473

[7] J. Yuan, L. Zhang, S. Guo, Y. Xiao, and Z. Li, "Image captioning with a joint attention mechanism by visual concept samples," *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)*, vol. 16, no. 3, pp. 1–22, 2020. https://doi.org/10.1145/3394955

[8] V. Agrawal, S. Dhekane, N. Tuniya, and V. Vyas, "Image caption generator using attention mechanism," in *2021 12th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, 2021, pp. 1–6. https://doi.org/10.1109/ICCCNT51525.2021.9579967

[9] S. Ayoub, Y. Gulzar, F. A. Reegu, and S. Turaev, "Generating image captions using bahdanau attention mechanism and transfer learning," *Symmetry*, vol. 14, no. 12, p. 2681, 2022. https://doi.org/10.3390/sym14122681

[10] L. Huang, W. Wang, J. Chen, and X. Y. Wei, "Attention on attention for image captioning," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4633–4642. https://doi.org/10.1109/ICCV.2019.00473

[11] P. Kinghorn, L. Zhang, and L. Shao, "A region-based image caption generator with refined descriptions," *Neurocomputing*, vol. 272, pp. 416–424, 2018. https://doi.org/10.1016/j.neucom.2017.07.014

[12] C. Rane, A. Lashkare, A. Karande, and Y. S. Rao, "Image captioning based smart navigation system for visually impaired," in *2021 Int. Conf. Commun. Inf. Comput. Technol. (ICCICT)*, 2021, pp. 1–5. https://doi.org/10.1109/ICCICT50803.2021.9510102

[13] H. Ahsan, N. Bhalla, D. Bhatt, and K. Shah, "Multi-modal image captioning for the visually impaired," *arXiv preprint arXiv:2105.08106*, 2021. https://doi.org/10.48550/arXiv.2105.08106

[14] D. Gurari, Y. Zhao, M. Zhang, and N. Bhattacharya, "Captioning images taken by people who are blind," in *Computer Vision – ECCV 2020,* in Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, J. M. Frahm, Eds., vol. 12362, 2020, pp. 417–434. https://doi.org/10.1007/978-3-030-58520-4_25

[15] Ö. Çaylı, B. Makav, V. Kılıç, and A. Onan, "Mobile application based automatic caption generation for visually impaired," in *Intelligent and Fuzzy Techniques: Smart and Innovative Solutions, INFUS 2020, Advances in Intelligent Systems and Computing*, C. Kahraman, S. Cevik Onar, B. Oztaysi, I. Sari, S. Cebi, and A. Tolga, Eds., vol. 1197, 2021, pp. 1532–1539. https://doi.org/10.1007/978-3-030-51156-2_178

[16] N. Li and Z. Chen, "Image cationing with visual-semantic LSTM," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, 2018, pp. 793–799. [Online]. Available: https://www.ijcai.org/Proceedings/2018/0110.pdf

[17] S. Alshattnawi and H. R. Alshboul, "Combined deep learning approaches for intrusion detection systems," *Int. J. Interact. Mob. Technol. (iJIM)*, vol. 18, no. 19, pp. 144–155, 2024. https://doi.org/10.3991/ijim.v18i19.49907

[18] H. Liu and T. Brailsford, "Reproducing 'show, attend and tell: Neural image caption generation with visual attention,'" *J. Phys. Conf. Ser.*, vol. 2589, pp. 1–7, 2023. https://doi.org/10.1088/1742-6596/2589/1/012012

[19] Y. Chu, X. Yue, L. Yu, M. Sergei, and Z. Wang, "Automatic image captioning based on ResNet50 and LSTM with Soft Attention," *Wireless Communications and Mobile Computing*, vol. 2020, no. 1, pp. 1–7, 2020. https://doi.org/10.1155/2020/8909458

[20] S. Yan, J. S. Smith, W. Lu, and B. Zhang, "Hierarchical multi-scale attention networks for action recognition," *Signal Process. Image Commun.*, vol. 61, pp. 73–84, 2018. https://doi.org/10.1016/j.image.2017.11.005

[21] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086. https://doi.org/10.1109/CVPR.2018.00636

[22] S. Nayak and C. B. Chandrakala, "Assistive mobile application for visually impaired people," *Int. J. Interact. Mob. Technol. (iJIM)*, vol. 14, no. 16, pp. 52–69, 2020. https://doi.org/10.3991/ijim.v14i16.15295

[23] A. Hussain and A. M. Omar, "Usability evaluation model for mobile visually impaired applications," *Int. J. Interact. Mob. Technol. (iJIM)*, vol. 14, no. 5, pp. 95–107, 2020. https://doi.org/10.3991/ijim.v14i05.13349

[24] S. Zaman, M. A. Abrar, M. M. Hassan, and A. N. M. Nafiul Islam, "A recurrent neural network approach to image captioning in braille for blind-deaf people," in *2019 IEEE Int. Conf. Signal Processing Information, Commun. Syst. (SPICSCON)*, 2019, pp. 49–53. https://doi.org/10.1109/SPICSCON48833.2019.9065144

[25] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. of the 40th Annu. Meet. Assoc. Comput. Linguist. (ACL)*, 2002, pp. 311–318. https://doi.org/10.3115/1073083.1073135

[26] A. Lavie and A. Agarwal, "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation*, 2007, pp. 228–231. https://doi.org/10.3115/1626355.1626389

[27] Hsankesara, Flickr Image dataset|Kaggle, 2018. Available at: https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset

[28] D. Santi, A. A. Ilham, Syafaruddin, and I. Nurtanio, "Image caption generation through the integration of CNN-based residual network architectures and LSTM," in *2024 7th International Conference on Informatics and Computational Sciences (ICICoS)*, 2024, pp. 227–232. https://doi.org/10.1109/ICICoS62600.2024.10636926

[29] Y. Bhatia, A. Bajpayee, D. Raghuvanshi, and H. Mittal, "Image captioning using Google's inception-resnet-v2 and recurrent neural network," in *2019 12th Int. Conf. Contemp. Comput. (IC3)*, 2019, pp. 1–6. https://doi.org/10.1109/IC3.2019.8844921

[30] A. Singh, T. D. Singh, and S. Bandyopadhyay, "An encoder-decoder based framework for hindi image caption generation," *Multimed. Tools Appl.*, vol. 80, nos. 28–29, pp. 35721–35740, 2021. https://doi.org/10.1007/s11042-021-11106-5

8 AUTHORS

Dessy Santi is a Lecturer in the Department of Information Technology, Faculty of Engineering, Universitas Tadulako, Indonesia. She is currently pursuing a doctoral degree in the Department of Electrical Engineering with a concentration in Informatics Engineering at Universitas Hasanuddin, Indonesia. Her research interests include artificial intelligence, computer vision, natural language processing, and data mining (E-mail: dessy@untad.ac.id).

Amil Ahmad Ilham is a Professor at the Department of Informatics, Faculty of Engineering, Universitas Hasanuddin, Indonesia. He is currently the Vice Dean for Academic and Student Affairs at the Faculty of Engineering, Universitas Hasanuddin. His research interests include various fields in information systems, big data, cloud computing, and computer systems. Additionally, he also has interests in image

processing, data mining, machine learning, deep learning, and artificial intelligence (E-mail: amil@unhas.ac.id).

Syafaruddin received his B.Eng degree in Electrical Engineering from Universitas Hasanuddin, Indonesia, in 1996, a Master of Engineering (M.Eng) degree in Electrical Engineering from the University of Queensland, Australia, in 2004, and a Doctor of Engineering (D.Eng) degree from Kumamoto University, Japan, in 2009. He has been working as a project assistant professor in the Frontier Technology for Electrical Energy Department of Computer Science and Electrical Engineering at Kumamoto University, Japan, from December 2009 to March 2011. He has been a Professor in Energy Conversion at the Department of Electrical Engineering, Universitas Hasanuddin, since 2017. His research interests are distributed generation planning, real-time simulation, and renewable energy, especially photovoltaic system, including MPP control, inverter configuration, partially shaded cases, and fault tolerance for household PV applications, and intelligent control algorithms and applications to power systems (E-mail: syafaruddin@unhas.ac.id).

Ingrid Nurtanio received the Bachelor's degree in Electrical Engineering from Universitas Hasanuddin, Indonesia, in 1986. She received her master of technology from Universitas Hasanuddin, Indonesia, in 2002. She received her doctoral degree from Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia, in 2013. Her research interests are digital image processing, computer vision, and intelligent systems. Currently, she is on the staff of the Department of Informatics, Faculty of Engineering, Universitas Hasanuddin. She is a member of IAENG and IEEE (E-mail: ingrid@unhas.ac.id).