VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v10i4.6657

DEVELOPMENT OF CNN-LSTM-BASED IMAGE CAPTIONING DATASET TO ENHANCE VISUAL ACCESSIBILITY FOR DISABILITIES

Muhammad Rifki[1*]; Ade Bastian[2]; Ardi Mardiana[3]

Informatics[1,2,3]
Universitas Majalengka, Majalengka, Indonesia[1,2,3]
www.unma.ac.id[1,2,3]
211410128@unma.ac.id[1*] , adebastian@unma.ac.id[2], aim@unma.ac.id[3]

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract*—Visual accessibility in public spaces remains limited for individuals with visual impairments in Indonesia, despite technological advancements such as image captioning. This study aims to develop a custom dataset and a baseline CNN-LSTM image captioning model capable of describing sidewalk accessibility conditions in Indonesian language. The methodology includes collecting 748 annotated images from various Indonesian cities, with captions manually crafted to reflect accessibility features. The model employs DenseNet201 as the CNN encoder and LSTM as the decoder, with 70% of the data used for training and 30% for validation. Evaluation was conducted using BLEU and CIDEr metrics. Results show a BLEU-4 score of 0.27 and a CIDEr score of 0.56, indicating moderate alignment between model-generated and reference captions. While the absence of an attention mechanism and the limited dataset size constrain overall performance, the model demonstrates the ability to identify key elements such as tactile paving, signage, and pedestrian barriers. This study contributes to assistive technology development in a low-resource language context, providing foundational work for future research. Enhancements through data expansion, incorporation of attention mechanisms, and transformer-based models are recommended to improve descriptive richness and accuracy.*

Keywords*: accessibility, CNN-LSTM, image captioning, sidewalk, visual impairment.*

Intisari—*Aksesibilitas visual bagi penyandang disabilitas penglihatan di ruang publik Indonesia masih sangat terbatas, meskipun perkembangan teknologi memungkinkan solusi inovatif seperti image captioning. Penelitian ini bertujuan mengembangkan dataset dan model baseline image captioning berbasis CNN-LSTM yang dapat mendeskripsikan kondisi aksesibilitas trotoar di berbagai kota di Indonesia dalam bahasa Indonesia. Metode yang digunakan meliputi pengumpulan 748 gambar dari ruang publik, anotasi manual caption kontekstual, serta pelatihan model deep learning dengan arsitektur DenseNet201 sebagai encoder dan LSTM sebagai decoder. Data dibagi dalam proporsi 70:30 untuk pelatihan dan validasi, serta diuji menggunakan metrik BLEU dan CIDEr. Hasil menunjukkan nilai BLEU-4 sebesar 0,27 dan CIDEr sebesar 0,56, mencerminkan kemiripan sedang antara caption hasil model dan referensi manual. Meskipun akurasi belum optimal akibat keterbatasan jumlah data dan tidak digunakannya mekanisme attention, model berhasil mengenali elemen-elemen penting seperti jalur pemandu, rambu, dan pembatas trotoar dalam konteks lokal. Studi ini menyediakan kontribusi awal dalam pengembangan teknologi deskripsi gambar berbasis bahasa Indonesia untuk mendukung kemandirian penyandang tunanetra di ruang publik. Penelitian lanjutan disarankan dengan penambahan data, integrasi attention mechanism, serta eksplorasi model berbasis transformer untuk peningkatan akurasi dan kompleksitas deskripsi.kunci).*

Kata Kunci*: aksesibilitas, CNN-LSTM, image captioning, trotoar, gangguan penglihatan*

INTRODUCTION

In Indonesia, as in numerous other nations, individuals with visual impairments face considerable challenges in accessing visual information within public spaces. According to data from World Health Organization (WHO), approximately 253 million individuals globally experience visual impairment, with 36 million classified as blind [1]. With blindness prevalence ranging from 1.5% to 3.0% and an estimated 8 million visually impaired people overall, Indonesia suffers a great load of visual disability [2], [3], [4].

Individuals with visual impairments encounter obstacles in employment, education, social engagement, and face discrimination stemming from inadequate accessibility technology. Despite the implementation of regulations concerning accessibility, the availability of facilities that accommodation disabilities in Indonesia public spaces remains quite limited [5]. This condition necessitates that individuals with visual impairments rely on assistance from other to comprehend their circumstances and navigate their environment effectively. The ability to access visual information independently in public spaces plays a significant role in enhancing individuals' quality of life.

Technology presents numerous opportunities to enhance accessibility. Recent studies have explored the integration of image captioning into navigation tools for people with visual impairments, producing promising results through the application of encoder-decoder architectures that combine CNN and LSTM, as well as their variations [6]. A variety of assistive technologies utilising artificial intelligence, particularly deep learning, have been developed to allow individuals with sensory disabilities to obtain visual information through different modalities [7], [8], [9].

An important advancement in assistive technology is image captioning, which autonomously produces descriptive sentences based on visual images. This method combines visual analysis and language understanding to efficiently deliver audio descriptions for individuals with visual impairments [10]. The primary obstacle in this domain is the amalgamation of two areas: utilizing computer vision to comprehend image content and employing natural language processing to produce effective descriptions.

Image captioning serves a variety of purposes. For example, it assists individuals with visual impairments by providing descriptions of their environment [11]. Given that image captions enhance accessibility and boost user engagement, this software finds application in both news media and social media platforms [12]. In the medical domain, techniques for image captioning have been utilized to generate reports or descriptions for medical images (such as radiology scans), assisting clinicians in their diagnostic processes [13]. The variety of applications highlights the necessity for the creation of precise and dependable image captioning systems. Simultaneously, they emphasize the necessity for captioning models capable of generalizing across various domains and languages.

Contemporary image captioning systems employ an encoder-decoder architecture. This system employs a deep convolutional neural network (CNN) to transform the input image into a concise feature representation (encoder) and utilises a decoding network to produce a phrase. Early models for deep learning captioning employed convolutional neural networks such as VGG or ResNet to extract visual features. A recurrent neural network, typically an LSTM, subsequently produced the caption one word at a time [14].

The encoder-decoder framework, as illustrated by Vinyals et al. and others, addressed image captioning in a manner akin to machine translation (converting an image into a sentence). At every time step, the CNN encoder produces a "image context vector" that the decoder utilises to forecast the subsequent word. This technology surpassed templating or retrieval-based captioning and established itself as the standard in the industry[15], [16].

Building on this foundation, subsequent experiments incorporated attention processes to enhance captioning. Rather than relying on a global image vector, visual attention enables the decoder to concentrate on particular sections of the image while generating words. Xu et al. discovered that adjusting the priority of image features during the decoding process enhances both the descriptiveness and accuracy of captions, such as emphasising the "dog" area when generating the word "dog." A variety of attention-based captioning models have recently been developed.

The Attention on Attention network developed by Das et al. enhances the traditional attention module by assessing the relevance of attention outcomes to the query, thereby refining focus and enhancing caption accuracy [17]. In a different strategy, Pan et al. presented an X-linear attention network that employs bilinear pooling to capture higher-order feature interactions, thereby improving the model's capacity to focus on important image details [18]. The implementation of these attention-based innovations has greatly enhanced the descriptive richness of generated captions, establishing them as a standard element in cutting-edge models.

Recently, the field has adopted transformer-based architectures, resulting in exceptional performance on image captioning benchmarks [19], [20]. Self-attention techniques and transformer models effectively capture sequence dependencies without the need for recurrent units. Self-attention layers are employed in the decoder of transformer models, while a CNN or transformer encoder is utilised for image captioning. The Meshed-Memory Transformer developed by Cornia et al. employs a multi-layer transformer decoder along with learnable memory tokens to effectively capture visual context across different tiers.

The advancements achieved through cross-modal pre-training are significant. OSCAR, created by Li et al., pre-trains a transformer on an extensive dataset of image–text pairs that include object tags, resulting in a model that can be fine-tuned for generating high-quality captions. The subsequent research conducted by Zhang et al. introduced the VinYL model, which integrates a more advanced object detector to enhance visual feature richness; this model achieved new state-of-the-art results across multiple captioning benchmarks [21]. Due to recent advancements, the leading captioning models of today are capable of describing images with remarkable precision, attaining elevated scores on evaluation metrics such as BLEU and CIDEr.

Image captioning provides visually impaired individuals with audio descriptions of text, faces, and surrounding objects, as demonstrated by Microsoft's Seeing AI. A global event such as the VizWiz Grand Challenge encourages the development of image captioning systems aimed at enhancing accessibility for individuals with visual impairments [22]. Since 2015, advancements in deep learning architectures utilising CNNs for image analysis and LSTM networks for language synthesis have led to remarkable developments in automatic picture captioning [23]. Due to its capacity to extract visual cues and generate meaningful descriptive phrases, CNN-LSTM is extensively utilised.

The convolutional neural network identifies essential visual elements, whereas the long short-term memory network generates cohesive sentences based on those elements [24]. The effectiveness of this approach in Indonesia is clear from studies showing significant accuracy in producing image descriptions in the Indonesian language, with a BLEU-4 score of approximately 0.60 on the MS COCO dataset.

In light of these developments, numerous obstacles and areas for further investigation persist. Models that mimic unseen features reduce object hallucination [25]. Too much topical focus can disregard model context. Errors aggravate complex circumstances. Overuse of training datasets is another issue. Train and assess top models using tens of thousands of MS COCO pictures. Many practical or specialized applications lack large labeled datasets. Lack of data or training hinders captioning algorithms. Culturally diverse scientific images or situations may challenge models. New language and location databases and models address these issues.

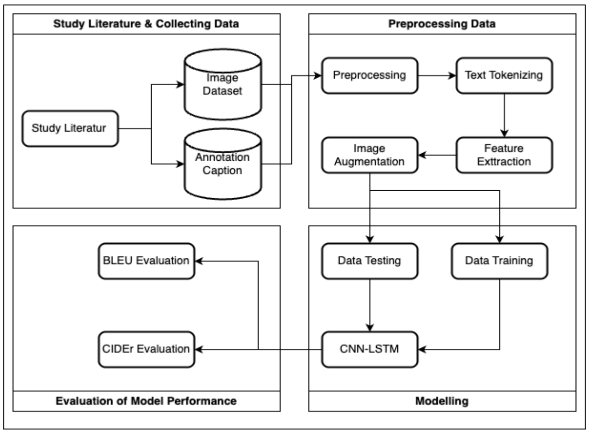
Rock properties were retrieved from geological rock core photos using deep learning [26]. Caption photographs in multiple languages [27]. An Indonesian game with object-attention descriptions and heritage photos. These initiatives demonstrate the growing need for captioning in non-English and specialized domains with minimal data and unstandard graphics. Every scenario typically necessitates a new dataset along with modifications to the model architecture or training strategies to effectively tackle the unique challenges presented [28].

Therefore, this study aims to develop a custom image captioning dataset that reflects sidewalk accessibility in Indonesian public spaces and to evaluate the performance of a baseline CNN-LSTM model trained on that dataset. The primary objective is to determine the model's ability to generate meaningful Indonesian-language captions related to accessibility features, serving as an early step toward localized assistive technologies for the visually impaired.

MATERIALS AND METHODS

The development of the proposed image captioning model involves several key stages, including the study literature, collection and preprocessing of image datasets with captions in Bahasa Indonesia, text tokenization, visual feature extraction utilizing DenseNet201, and image augmentation to enhance the model's generalization capabilities. The dataset was subsequently partitioned into training and validation subsets, allocated at proportions of 70% and 30% respectively. The evaluation of model performance is carried out through the application of BLEU and CIDEr metrics, which serve to assess the quality of the generated captions.

VOL. 10. NO. 4 MAY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i4.6657



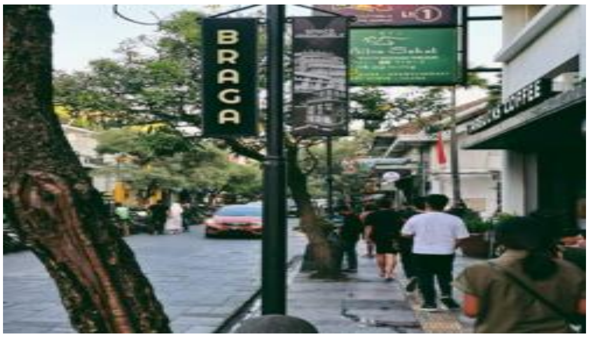
Source : (Research Result, 2025)

Figure 1. Research Workflow

Collecting the Datasets

This study utilizes an image dataset comprising photographs of sidewalk conditions and pedestrian facilities collected from multiple cities across Indonesia. A total of 748 images were gathered from over 20 distinct locations or cities, including Jakarta, Bandung, Banjarmasin, Palembang, Yogyakarta, Bali, along with several cities in Kalimantan, Nusa Tenggara, and Maluku. Every image illustrates the state of the street or station surroundings, focusing specifically on the availability of guiding blocks, sidewalk obstacles, crossing facilities, and various other elements related to accessibility. The images were collected directly from the field and through publicly available documentation, subsequently chosen based on their relevance to the context of accessibility for individuals with visual impairments.

The annotation process involved manually crafting a descriptive sentence in Indonesian for each image. Every image comes with a caption that elucidates the most important visual aspect. Below is an example of an image accompanied by a caption that outlines the primary visual conditions.



Source : (https://travel.detik.com/galeri-foto/d-5414126/jalan-braga-bandung-tempo-dulu. [Accessed: March. 05,2025])

Figure 2. bdg1

The caption provided above is : "*Jalan Braga dengan trotoar berpola batu kotak-kotak, terdapat papan penunjuk bertuliskan 'Braga'. Pejalan kaki ramai, dan terdapat rambu di sekitar troto*ar". The captions were crafted to accurately represent the key elements within the images, alongside the challenges encountered by pedestrians with disabilities. The dataset comprises a CSV file that contains two columns: one for the image file name and another for the corresponding text description. Table 1 presents an overview of sample image and caption pairs:

Table 1. Datasets Collection

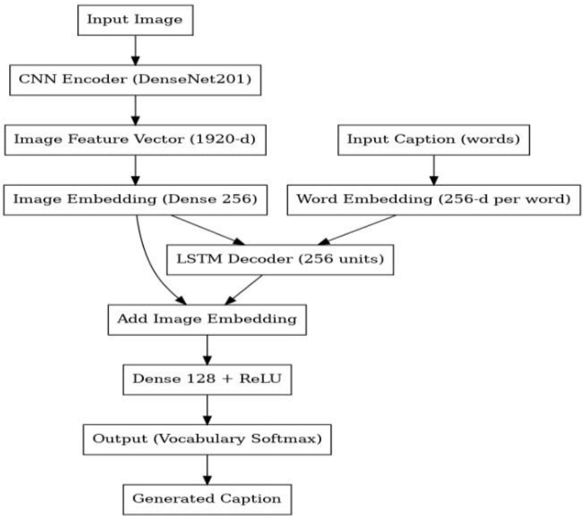
Image	Caption (Indonesian)	Caption (English)
bdg1.jpg	Jalan Braga dengan trotoar berpola batu kotak-kotak, terdapat papan penunjuk bertuliskan 'Braga'. Pejalan kaki ramai, dan terdapat rambu di sekitar trotoar.	Braga Street with a checkered stone-patterned sidewalk, a signboard reading 'Braga', many pedestrians, and a traffic sign nearby.
bdg2.jpg	Jalur pemandu kuning di tengah trotoar, bangku kayu di sisi kiri, dan bola beton pembatas di sisi kanan sebelum jalan raya	Yellow guiding block in the center of the sidewalk, wooden bench on the left, and concrete bollards on the right before the roadway.
bdg3.jpg	Jalur pemandu terganggu oleh mobil dan motor parkir di trotoar, menyulitkan akses pejalan kaki tunanetra.	The guiding block is obstructed by parked cars and motorcycles on the sidewalk, hindering access for visually impaired pedestrians.
jkt1.jpg	Trotoar dengan jalur pemandu berwarna kuning di tengah, diapit taman kecil dan jalan raya di sebelah kiri.	Sidewalk with a yellow guiding path in the center, flanked by a small garden and a roadway on the left.
jkt2.jpg	Jembatan penyeberangan beratap dengan pegangan tangan di kedua sisi. Tanda arah ke halte bus terlihat di bagian atas.	Covered pedestrian bridge with handrails on both sides. A sign pointing to the bus stop is visible at the top.
Jkt3.jpg	Pejalan kaki berjalan di trotoar sempit dengan dinding di sebelah kiri dan jalan raya tanpa pembatas di kanan.	Pedestrians walking on a narrow sidewalk with a wall on the left and an unguarded roadway on the right.
ntb1.jpg	Jalur pemandu taktile berada di tengah trotoar yang luas, dengan papan informasi berdiri di sebelah kanan jalur.	Tactile guiding block is in the center of the sidewalk, with an information board standing on the right side.
ntb2.jpg	Jalur taktile membentang lurus di tengah trotoar, terdapat bangku di sisi kiri dan pohon di kanan.	The tactile block stretches straight through the center of the sidewalk, with a bench on the left and a tree on the right.
ntb3.jpg	Jalur pemandu di sisi kanan trotoar, berada	The guiding block is on the right side of

Image	Caption (Indonesian)	Caption (English)
	sejajar dengan gedung tinggi dan taman kota.	the sidewalk, aligned with tall buildings and a city park.
yk1.jpg	Trotoar luas dengan jalur pemandu untuk tunanetra di sisi kanan. Kursi kayu tersedia di sepanjang jalur pejalan kaki.	Wide sidewalk with a guiding block for the blind on the right side. Wooden benches are available along the pedestrian path.
yk2.jpg	Jalur pemandu berada di tengah trotoar, tapi permukaan licin akibat hujan. Banyak pejalan kaki, dan kendaraan parkir di sebelah kanan.	The guiding block is in the center of the sidewalk, but the surface is slippery due to rain. Many pedestrians and parked vehicles on the right.
yk3.jpg	Persimpangan besar dengan trotoar luas dan jalur penyeberangan. Lampu lalu lintas tersedia untuk penyeberang jalan.	Large intersection with wide sidewalks and a pedestrian crossing. Traffic lights are available for pedestrians.

Source : (Research Result, 2025)

Architecture of an Image Captioning System

The architecture proposed for the image captioning model, which incorporates object detection, is illustrated in Figure 2. This study presents the development of an image captioning system utilizing a deep learning methodology grounded in a CNN-LSTM encoder-decoder architecture. This method has been widely adopted in numerous prior investigations because of its efficiency in integrating visual and linguistic elements concurrently. This architecture is composed of two primary elements: a Convolutional Neural Network (CNN) serving as the image encoder and a Long Short-Term Memory (LSTM) functioning as the text decoder.


Input Image
CNN Encoder (DenseNet201)
Image Feature Vector (1920-d)
Input Caption (words)
Image Embedding (Dense 256)
Word Embedding (256-d per word)
LSTM Decoder (256 units)
Add Image Embedding
Dense 128 + ReLU
Output (Vocabulary Softmax)
Generated Caption

Source : (Research Result, 2025)

Figure 3. Proposed Model for Image Captioining

This study employs DenseNet201 as the CNN encoder component within its architecture. The choice of DenseNet201 stems from its capacity to generate a more nuanced and effective feature representation when compared to other CNN architectures like VGG16 and ResNet50, which have been frequently utilized in earlier research [10], [24].

DenseNet201 produces high-dimensional image feature vectors (1920 dimensions), which are subsequently reduced to 256-dimensional image embedding vectors via a fully connected layer. This compression seeks to simplify information complexity while preserving key characteristics in the visual depiction of the image. The decoder section involves a preprocessing stage for the caption text, which encompasses tokenization and word embedding. Every word is transformed into a 256-dimensional embedding vector, resulting in a continuous representation that facilitates the prediction of the subsequent word.

The image embedding and word embedding are subsequently integrated within a 256-unit LSTM to effectively capture the temporal patterns between words and their relationship with the visual features of the image. The distinction of this study from the methodology presented in the prior RESTI journal by Akbar et al. (2023), which employed a VGG16-based object attention technique and bounding box objects for captioning traditional game images, lies in the integration of visual features [28]. Akbar et al. conducted a study that employs an attention mechanism which explicitly considers the location of detected objects, resulting in more precise captions, particularly for specific and localized objects within the image.

This study adopts a more targeted approach by updating a specific dataset, specifically public space environments in Indonesia. It does so without incorporating additional attention mechanisms, instead leveraging DenseNet201's capacity to capture intricate global visual representations. The integration of visual features is achieved through an element-wise summation process involving the LSTM output and the embedded image, thereby maintaining a robust connection of visual features in the generation of captions.

Subsequently, this integrated representation is forwarded to an extra Dense layer of size 128 utilizing the ReLU activation function, designed to enhance the feature representation for producing more detailed and precise captions. In the concluding phase, the representation is forwarded to the softmax layer, which produces the probability distribution of the existing vocabulary.

This iterative process persists until a specific token is generated to signify the conclusion of the caption. In the concluding phase, the representation is forwarded to the softmax layer, which produces the probability distribution of the existing vocabulary. The iteration process persists until a specific token is produced to signify the conclusion of the caption. Therefore, this investigative method not only enhances the effective application of the DenseNet201-based CNN-LSTM architecture, but also offers a distinctive contribution through an updated dataset that reflects the actual conditions of public spaces in Indonesia.

This sets it apart from earlier investigations that typically rely on broad datasets like MS COCO or images featuring highly specific and localized objects, as demonstrated by Akbar et al. This study thoroughly assesses the efficacy of the CNN-LSTM model within the confines of a particular and restricted dataset, serving as a foundational reference for future inquiries, including the potential incorporation of attention mechanisms or transformer-based components.

Training Model

Prior to the training process, all images underwent resizing to dimensions of 224×224 pixels and were normalized by dividing the pixel values by 255, ensuring compatibility with the input requirements of DenseNet201. In the process of preparing caption text, tokenization involves the inclusion of special tokens and at the beginning and end of each caption sentence.

The incorporation of these tokens is intended to signify the start and conclusion of the sequence throughout the training and inference phases. Subsequently, every word is transformed into an integer index through the Keras Tokenizer, and the resulting word index sequence is padded to a specified maximum length, which corresponds to the longest caption, approximately 30 words including tokens.

The tokenization process has produced approximately 3.3 thousand distinct words, with an adjustment of +1 for tokens. This procedure guarantees that all text inputs maintain consistent dimensions upon entering the LSTM network. Image augmentation is utilized to enhance the diversity of the training data. Every image in the training batch undergoes a series of random transformations, including rotation (up to 20°), width/height shifts (up to 20%), shearing, zooming (±20%), and horizontal flipping. The implementation of augmentation is anticipated to enhance the model's robustness in object recognition, even when faced with variations in viewing angles or lighting conditions. The dataset is subsequently partitioned into two subsets: 70% of the images, approximately 523 images, are allocated for training, while the remaining 30%, around 225 images, are designated for validation. The division is executed according to a compilation of distinct image file names, ensuring that all captions associated with a single image are exclusively contained within one of the subsets. This validation dataset serves as a simulated testing ground to assess the model's performance on images it has not encountered before.

The model utilized Adam's optimizer for training, starting with a learning rate of 0.001. The employed loss function is categorical cross-entropy, given that the output pertains to the classification of the next word from a set of words within the vocabulary. The training process was executed using mini-batch mode, employing a batch size of 32 samples. The maximum number of epochs was established at 50, while an Early Stopping mechanism with a patience of 5 was implemented to halt training if the validation accuracy failed to improve after 5 consecutive epochs.

Furthermore, ReduceLROnPlateau was implemented, which lowers the learning rate by a factor of 0.2 if there is no enhancement in validation accuracy over a span of 3 epochs. Throughout the training process, at every iteration, the model is provided with two types of inputs: (1) image features represented as 1920-dimensional vectors derived from CNN results, and (2) word sequences formatted as padded sequences. The objective is to predict the subsequent word for each caption prefix, employing the teacher forcing method.

The evaluation of the image captioning model's performance is conducted through the use of BLEU and CIDEr automatic caption evaluation metrics. The BLEU metric, or Bilingual Evaluation Understudy, assesses the n-gram similarity between the generated caption and the reference caption, which serves as the ground truth. This analysis involved calculating BLEU-4 scores, utilizing 1 to 4-grams, for the average across all test images. In the meantime, the CIDEr metric assesses the cosine tf-idf similarity between the n-grams of the generated caption and the reference caption, tailored specifically for evaluating image descriptions. Below is a summary of the primary hyperparameters of the model along with the evaluation scores achieved:

Table 2. Key Hyperparameters of the CNN-LSTM Image Captioning Model

Hyperparameter	Specification
CNN Encoder (Backbone)	DenseNet201 (Pretrained)
Image Feature Dimension	1920 (Global Average Pool)

Hyperparameter	Specification
Image Embedding (Dense)	256 Neurons, ReLu
Word Embedding	256 Dimension
Units LSTM (Decoder)	256 units
Dropout	0.5 (After LSTM & Dense)
Optimizer	Adam (learning_rate=001)
Loss Function	Categorical Crossentropy
Batch Size	32
Epoch Maximum	50 (Early Stop ~22)
Data Augmentation	Rotasi, Zoom, Shear, Flip
Proportion of training: validation data	70:30 (% Images)

Source : (Research Result, 2025)

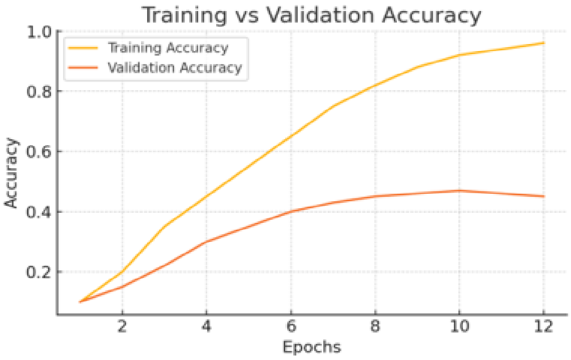
RESULTS AND DISCUSSION

This section presents the findings and evaluation of the CNN-LSTM model's performance on the newly proposed Indonesian sidewalk accessibility dataset. An analysis was conducted focusing on the accuracy metrics, particularly the BLEU and CIDEr scores derived from the trained model. Furthermore, the analysis of learning curves related to accuracy and loss was conducted to assess the stability of the training process and to pinpoint any instances of potential overfitting.

Quantitative Performance

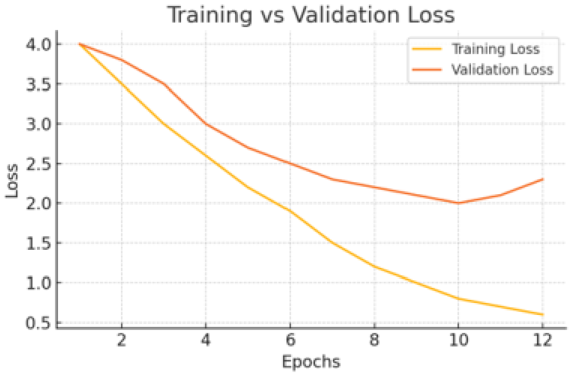
The CNN-LSTM model was trained for a maximum of 50 epochs, incorporating early stopping to mitigate the risk of overfitting. The model demonstrated rapid convergence attributed to the limited size of the dataset, achieving optimal performance after 22 epochs. Figures 3 and 4 depict the training process. Figure 3 presents the image for our analysis. The following illustrates the comparison of training and validation accuracy throughout the epochs.

The training accuracy demonstrates a rapid increase, exceeding 90%, whereas the validation accuracy peaks at around 47% before showing a downward trajectory. This inconsistency indicates that the model begins to overly conform to the training data after reaching a certain limit. The implementation of data augmentation methods like rotation, zoom, shear, and flip provided a certain level of regularization; nonetheless, the restricted size of the dataset ultimately limited the validation accuracy that could be achieved. The early stopping mechanism halted training when there was no further improvement in validation performance, which occurred after just a dozen epochs, highlighting the rapid exhaustion of the information present in the dataset by the model.



Source: (Research Result, 2025)
Figure 4. Training vs Validation Accuracy Across Epochs

The image for our analysis is illustrated in Figure 4. The training loss, represented by the orange line, consistently declines to below 1.0, while the validation loss, indicated by the red line, hits a minimum near epoch 10 before experiencing a slight increase. This trend suggests that after the 10th epoch, the model began to overfit: it kept reducing loss on the training captions but lost generalizability, as shown by the rising loss on the validation set. The model exhibiting the lowest validation loss has been preserved for the final assessment.



Source : (Research Result, 2025)
Figure 5. Training vs Validation Loss Across Epochs

Following the training phase, the quality of the model's captions was assessed through established metrics, specifically BLEU and CIDEr, utilizing the held-out test set for evaluation. We calculated BLEU scores ranging from 1-grams to 4-grams (BLEU-1 to BLEU-4) and the CIDEr score for each generated caption in comparison to its reference. The final evaluation metrics achieved are presented in Table 3. The CNN encoder employed a pretrained DenseNet-201 backbone to derive 1920-dimensional features from each image, subsequently projecting these features into a 256-length vector to create the image embedding. The

LSTM decoder was set up with 256 units and a word embedding dimension of 256. The Adam optimizer was utilized with a learning rate of 0.001, alongside a categorical cross-entropy loss, to train the network using a conventional teacher-forcing approach. A batch size of 32 was employed, along with a 70:30 split for training and validation purposes. The settings outlined in Table 2 demonstrated stable training dynamics.

Table 3. Model Evaluation Metrics on Test Captions

Evaluation Metric	Score
BLEU-1 (unigram)	0.53
BLEU-2 (bigram)	0.42
BLEU-3 (trigram)	0.34
BLEU-4 (4-gram)	0.27

Source : (Research Result, 2025)

The findings presented in Table 3 demonstrate that the model successfully produces captions that exhibit a moderate degree of similarity to the actual descriptions provided. The BLEU-4 score, which takes into account sequences of up to four words, stands at approximately 0.27 (27%). In contrast, the cumulative scores for BLEU-1 to BLEU-3 vary from around 0.50 to 0.34. The CIDEr score is approximately 0.56. Both metrics are capped at 1.0, indicating that while our scores are not ideal, they do suggest that the model has effectively learned to generate captions that significantly align with the reference descriptions.

In absolute terms, a BLEU-4 score of approximately 25–30% and a CIDEr score around 0.5 align with the anticipated outcomes for an initial model developed on a new, constrained dataset. In contrast, advanced image captioning models that have been trained on extensive benchmarks such as MS COCO can achieve BLEU-4 scores of approximately 0.60, focusing on the disparity between our existing outcomes and the optimal performance that could be attained with ample training data and sophisticated architectures. The lower performance of our model can be linked to the restricted size and specific characteristics of the dataset, along with the lack of advanced attention mechanisms.

These results provide a valuable foundation: they indicate that the model successfully captures key elements of the images (as evidenced by non-zero BLEU/CIDEr scores), while also highlighting significant opportunities for enhancement in the future. The BLEU metric assesses n-gram overlap between the generated caption and the reference caption, while the CIDEr metric calculates a TF-IDF weighted cosine similarity of n-grams, specifically designed for evaluating image descriptions. All metrics are adjusted so that a score of 1.0 signifies an ideal alignment with the reference.

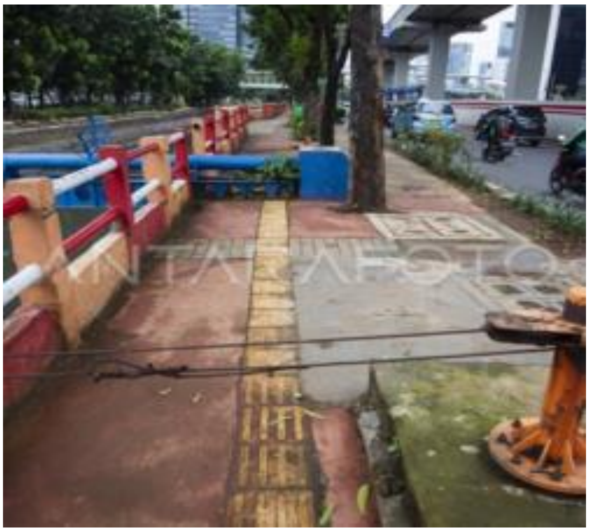
The BLEU-1 score of 0.53 indicates that, on average, slightly more than half of the unigrams (individual words) in the captions generated by the model correspond with the ground truth. The scores consistently decline for higher-order n-grams, falling to 0.27 for BLEU-4, which aligns with the expectation that longer exact phrase matches are more challenging to obtain. The CIDEr score of approximately 0.56, which takes into account the significance of each word through TF-IDF weighting, indicates that the generated captions adequately reflect the key elements of the scene.

However, there are notable variations in phrasing and detail when compared to captions created by humans. The quantitative results indicate that the model is effectively acquiring the “language” of Indonesian sidewalk scenes, as it frequently identifies and describes key elements such as trotoar (sidewalk), jalur pemandu (guiding block), bangku (bench), halte (bus stop), and others in the images. Nonetheless, the moderate scores of BLEU-4 and CIDEr indicate certain omissions or discrepancies in some of the generated descriptions, which we will investigate further through qualitative analysis in the following section.

Qualitative Result

In order to gain deeper insights into the model's performance, we analyzed instances of the captions it produces for images within the test set. Figures 5 and 6 present two sample images from the dataset, accompanied by their corresponding captions. For every instance, we present the accurate caption (the reference authored by a human) alongside the caption generated by our CNN-LSTM model. These examples illuminate the strengths and limitations of the model's descriptive capability in real-world scenarios.

Figure 5 depicts a tactile guide block positioned along the sidewalk, which is notably obstructed by a low-hanging wire, as illustrated in the photo across the street. The reference caption clearly indicates the obstruction caused by cable and pipe, underscoring a significant accessibility concern. The generated caption from the model accurately recognizes the yellow guiding block on the sidewalk and mentions a “barrier at the side,” which probably pertains to the railing or curb at the edge. Nevertheless, the model does not address the wire intersecting the path. This oversight is expected – the wire is a relatively slender object, and the model, without an attention mechanism, might not prioritize it. The model primarily focuses on the more prominent characteristics, such as the guiding block and the overall configuration of the sidewalk. This example demonstrates that although the model understands the overall context (a sidewalk featuring a guiding block and some boundaries), it may overlook essential finer details that influence accessibility. In practical assistive applications, overlooked details (such as an obstacle on the path) can be crucial, highlighting the necessity for enhancements to the model.

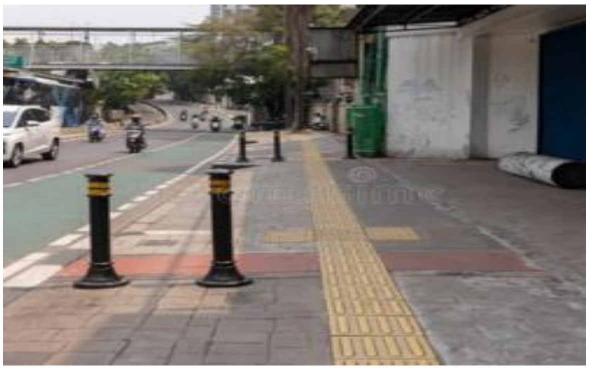


Source : (https://news.detik.com/foto-news/d-6618508/nestapa-jalur-tunanetra-di-jakarta-nabrak-tiang-diserobot-pedagang. [Accessed: March. 05,2025])

Figure 6. jkt1

Ground Truth: “Jalur pemandu kuning di trotoar Kuningan, Jakarta, terhalang oleh instalasi kabel dan pipa”. (The yellow guiding block on the sidewalk in Kuningan, Jakarta is obstructed by a cable and pipe installation). Model-generated: “Trotoar dengan jalur pemandu kuning di tengah, terdapat pembatas di samping trotoar.” (Sidewalk with a yellow guiding path in the middle, with a barrier at the side of the sidewalk.).

Figure 6. Another example of model captioning on a Jakarta street scene. Ground truth: “Trotoar di Jakarta dengan jalur pemandu berwarna kuning di tengah dan bollard hitam sebagai pembatas jalan.” (A sidewalk in Jakarta with a yellow guiding block in the middle and black bollards as road separators.) Model-generated: “Trotoar dengan jalur pemandu kuning di tengah dan beberapa bollard di sisi jalan.” (Sidewalk with a yellow guiding path in the middle and several bollards at the side of the road.)

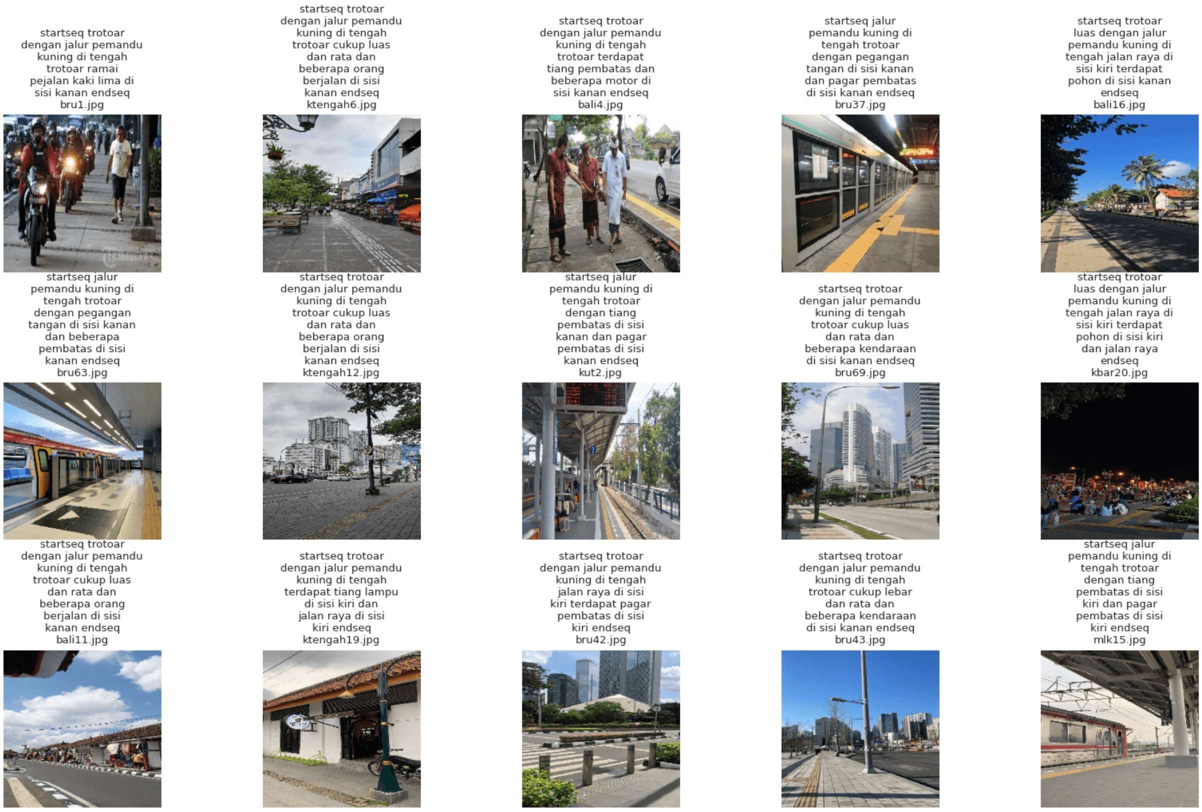


Source : (https://sebaraya.com/mengenal-jalur-kuning-pemandu-garis-berwarna-kuning-di-trotoar/. [Accessed: March. 05,2025])

Figure 7. jkt2

The model's functionality is demonstrated in Figure 6 above. A paved sidewalk with a yellow tactile guiding strip down its center and many black bollards (short posts) along the curb separates it from the road. Reference captions note these elements. The model-generated caption captures the scene’s major elements, including the jalur pemandu kuning (yellow guiding route) and side bollards. Using “beberapa bollard” instead of “bollard hitam” is slightly different, but the meaning is the same. After learning about bollards and guide blocks from the training data, the model can use them correctly in images. This graphic shows that the model can accurately recognize and characterize accessibility characteristics like tactile paving and barriers in Indonesian. When developed, such captions could help visually challenged individuals understand their environment. Even in this case, the model lists visible items in short sentences. The caption may be richer and more expressive by mentioning a bus lane or sidewalk condition. This shows that training on a small dataset with largely single-sentence captions narrows the model’s description style.

To provide a deeper insight into the qualitative performance of the proposed CNN-LSTM model, Figure 7 showcases several examples from the dataset along with their corresponding captions. The images illustrate typical situations found in public spaces in Indonesia, highlighting features related to accessibility such as tactile guide paths, pedestrian barriers, and pavement conditions. The examples demonstrate the model's ability to effectively identify and express key visual elements relevant to pedestrians with visual impairments, despite certain limitations in detail specificity and object recognition.


startseq trotoar dengan jalur pemandu kuning di tengah trotoar ramai pejalan kaki lima di sisi kanan endseq bru1.jpg
startseq trotoar dengan jalur pemandu kuning di tengah trotoar cukup luas dan rata dan beberapa orang berjalan di sisi kanan endseq ktengah6.jpg
startseq trotoar dengan jalur pemandu kuning di tengah trotoar terdapat tiang pembatas dan beberapa motor di sisi kanan endseq bali4.jpg
startseq jalur pemandu kuning di tengah trotoar dengan pegangan tangan di sisi kanan dan pagar pembatas di sisi kanan endseq bru37.jpg
startseq trotoar luas dengan jalur pemandu kuning di tengah jalan raya di sisi kiri terdapat pohon di sisi kanan endseq bali16.jpg
startseq jalur pemandu kuning di tengah trotoar dengan pegangan tangan di sisi kanan dan beberapa pembatas di sisi kanan endseq bru63.jpg
startseq trotoar dengan jalur pemandu kuning di tengah trotoar cukup luas dan rata dan beberapa orang berjalan di sisi kanan endseq ktengah12.jpg
startseq jalur pemandu kuning di tengah trotoar dengan tiang pembatas di sisi kanan dan pagar pembatas di sisi kanan endseq kut2.jpg
startseq trotoar dengan jalur pemandu kuning di tengah trotoar cukup luas dan rata dan beberapa kendaraan di sisi kanan endseq bru69.jpg
startseq trotoar luas dengan jalur pemandu kuning di tengah jalan raya di sisi kiri terdapat pohon di sisi kiri dan jalan raya endseq kbar20.jpg
startseq trotoar dengan jalur pemandu kuning di tengah trotoar cukup luas dan rata dan beberapa orang berjalan di sisi kanan endseq bali11.jpg
startseq trotoar dengan jalur pemandu kuning di tengah terdapat tiang lampu di sisi kiri dan jalan raya di sisi kiri endseq ktengah19.jpg
startseq trotoar dengan jalur pemandu kuning di tengah jalan raya di sisi kiri terdapat pagar pembatas di sisi kiri endseq bru42.jpg
startseq trotoar dengan jalur pemandu kuning di tengah trotoar cukup lebar dan rata dan beberapa kendaraan di sisi kanan endseq bru43.jpg
startseq jalur pemandu kuning di tengah trotoar dengan tiang pembatas di sisi kiri dan pagar pembatas di sisi kiri endseq mlk15.jpg

Source : (Research Result, 2025)

Figure 8. Illustration of qualitative findings from the CNN-LSTM image captioning model applied to the Indonesian sidewalk accessibility dataset

Overall, the qualitative results suggest that the model accurately detects visual structures and accessibility aspects. The model commonly captions images with a clear guide block, pedestrian crossing, or sidewalk barrier like a bench or parked vehicle in numerous test scenarios. This is promising since it emphasizes visual elements crucial to visually impaired users. A CNN (DenseNet201) pretrained on ImageNet delivers robust feature extraction, which may help the model recognize these objects with little training instances. Captions in Bahasa Indonesia have been generated, showing that a deep learning model can describe images in a low-resource language. Most picture captioning systems are designed for English or other high-resource languages, therefore this fills a crucial need. We improve local assisted vision technology accessibility by creating an Indonesian-captioned dataset and model.

Examples show various drawbacks despite these benefits. The resulting captions typically lack detail. The model mentions 1–3 key objects or properties and little else. Little elements like Figure 5's blockage, weather/lighting conditions, signs language, etc. are often missing. Due to the training captions' brief single sentences, the model was never exposed to longer descriptive statements. Second, the model sometimes misidentified objects if they were visually similar or had learned a bias. We saw the model mention a guide block in a couple outputs even though the sidewalk was smooth with no tactile pavement. Such inaccuracies suggest bias—the model may hallucinate guiding blocks in unexpected contexts because it has seen many sidewalks with them. Object hallucination, a common difficulty in picture captioning research, refers to a model stating objects that are not present. Our model's simplistic architecture (no attention mechanism) and insufficient training set variation may render it susceptible to this issue. Finally, caption language fluency and complexity can be improved. The generated Indonesian sentences are usually valid, however they are repetitious (frequently opening with "Trotoar dengan jalur pemandu...") and lack diversity. Due to the tiny dataset (~750 phrases), the model has limited style and vocabulary.

Comparison with Previous Works

This study expands upon earlier investigations in image captioning, with a specific focus on enhancing accessibility. A multitude of studies leverage extensive datasets such as MS COCO and utilize attention or transformer mechanisms to enhance accuracy. Nonetheless, the exploration of Indonesian-language captioning in public pedestrian settings is still quite scarce.

A significant study conducted by Akbar et al. employed CNN-LSTM with attention on images of Indonesian folk games, resulting in a 36% increase in BLEU-4 and a 43% improvement in CIDEr. This study distinguishes itself by concentrating on urban sidewalks and does not incorporate attention mechanisms. A novel dataset was developed specifically for assessing street accessibility in Indonesia, achieving baseline performance that aligns with pilot studies, and attaining a BLEU-4 score of 0.27.

Unlike prior models using lighter CNNs like MobileNet, we employed DenseNet201 for richer feature extraction. Though it increased computational demands, it successfully captured objects such as bollards. This confirms the potential of deeper models for limited-data contexts, but attention mechanisms are recommended for future improvement.

Table 4. Comparison of Related Research Results

Researcher & year	BLEU-4	CIDEr	Distinctive Feature/Excellence
Akbar et al. (2023)	0.30	0.60	Object detection with attention, local cultural domain
Nursikuwagus et al. (2022)	-	-	Geological domain, technical descriptive captions
Rifki et al. (2025)	0.27	0.56	New local dataset, baseline without attention

Source: (Research Result, 2025)

CONCLUSION

A novel dataset and baseline model for describing Indonesian sidewalk scenes has been introduced in the study titled "CNN-LSTM-Based Image Captioning Dataset to Enhance Visual Accessibility for Disabilities." The feasibility of employing deep learning to convert visual information into accessible language for individuals with visual impairments has been demonstrated, even within a low-resource environment. The results underscore the possibilities and obstacles: our model effectively communicates the essential information of a scene, yet enhancing detail, accuracy, and generalization will necessitate additional data and more sophisticated methods. This study represents a crucial initial phase in the development of automated image description tools tailored for the Indonesian context, ultimately seeking to enhance the ability of individuals with visual impairments to navigate public spaces with greater safety and independence. The dataset and insights from this study are expected to stimulate additional exploration and innovation, potentially resulting in improved models utilizing attention mechanisms and transformers, which will contribute to the advancement of assistive technologies in the near future. It is recommended that future studies broaden the dataset to include diverse regions, integrate multilingual captioning to enhance accessibility, and assess the model's effectiveness through user-focused usability testing with visually impaired participants.

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to Universitas Majalengka for the support and facilities provided during the preparation and completion of this study.

REFERENCE

[1] B. Arystanbekov, A. Kuzdeuov, S. Nurgaliyev, and H. A. Varol, "Image Captioning for the Visually Impaired and Blind: A Recipe for Low-Resource Languages," Feb. 23, 2023. doi: 10.36227/techrxiv.22133894.v1.

[2] L. Rif'Ati, A. Halim, Y. D. Lestari, N. F. Moeloek, and H. Limburg, "Blindness and Visual Impairment Situation in Indonesia Based on Rapid Assessment of Avoidable Blindness Surveys in 15 Provinces," Ophthalmic Epidemiol, vol. 28, no. 5, pp. 408–419, 2021, doi: 10.1080/09286586.2020.1853178.

[3] R. R. A. Bourne et al., "Trends in prevalence of blindness and distance and near vision impairment over 30 years: An analysis for the Global Burden of Disease Study," Lancet Glob Health, vol. 9, no. 2, pp. e130–e143, Feb. 2021, doi: 10.1016/S2214-109X(20)30425-3.

[4] R. Kesuma and A. Prasetyo, "Literature Review on the Prevalence of Vision Impairment and Age-Related Eye Diseases," 2024, pp. 1079–1085. doi: 10.2991/978-2-38476-273-6_111.

[5] D. Daniel, A. Nastiti, H. Y. Surbakti, and N. M. U. Dwipayanti, "Access to inclusive sanitation and participation in sanitation programs for people with disabilities in Indonesia," Sci Rep, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-30586-z.

[6] Y. Azhar, M. R. Anugerah, M. A. R. Fahlopy, and A. Yusriansyah, “Image Captioning using Hybrid of VGG16 and Bidirectional LSTM Model,” Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, Nov. 2022, doi: 10.22219/kinetik.v7i4.1568.

[7] P. Patel, S. Pampaniya, A. Ghosh, R. Raj, D. Karuppaih, and S. Kandasamy, “Enhancing Accessibility Through Machine Learning: A Review on Visual and Hearing Impairment Technologies,” 2025, Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/ACCESS.2025.3539081.

[8] B. Nurhakim, A. Rifai, D. A. Kurnia, D. Sudrajat, and U. Supriatna, “Smart Attendance Tracking System Employing Deep Learning For Face Anti-Spoofing Protection,” JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer), vol. 10, no. 3, pp. 496–505, Feb. 2025, doi: 10.33480/jitk.v10i3.5992.

[9] Moh. H. Fariz and E. B. Setiawan, “The Impact Of Word Embedding On Cyberbulliying Detection Using Hybrid Deep Learning CNN-BILSTM,” JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer), vol. 10, no. 3, pp. 661–671, Feb. 2025, doi: 10.33480/jitk.v10i3.6270.

[10] J. Ganesan, A. T. Azar, S. Alsenan, N. A. Kamal, B. Qureshi, and A. E. Hassanien, “Deep Learning Reader for Visually Impaired,” Electronics (Switzerland), vol. 11, no. 20, Oct. 2022, doi: 10.3390/electronics11203335.

[11] D. Gurari, Y. Zhao, M. Zhang, and N. Bhattacharya, “Captioning Images Taken by People Who Are Blind.” [Online]. Available: https://vizwiz.org.

[12] F. Liu, Y. Wang, T. Wang, and V. Ordonez, “Visual News: Benchmark and Challenges in News Image Captioning,” Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.03743

[13] Y. Yang, J. Yu, J. Zhang, W. Han, H. Jiang, and Q. Huang, “Joint Embedding of Deep Visual and Semantic Features for Medical Image Report Generation,” IEEE Trans Multimedia, vol. 25, pp. 167–178, 2023, doi: 10.1109/TMM.2021.3122542.

[14] Z. Zohourianshahzadi and J. K. Kalita, “Neural Attention for Image Captioning: Review of Outstanding Methods,” Nov. 2021, doi: 10.1007/s10462-021-10092-2.

[15] H. Sharma and D. Padha, “A comprehensive survey on image captioning: from handcrafted to deep learning-based techniques, a taxonomy and open research issues,” Artif Intell Rev, vol. 56, pp. 1–43, Apr. 2023, doi: 10.1007/s10462-023-10488-2.

[16] L. Xu, Q. Tang, J. Lv, B. Zheng, X. Zeng, and W. Li, “Deep image captioning: A review of methods, trends and future challenges,” Neurocomputing, vol. 546, p. 126287, 2023, doi: https://doi.org/10.1016/j.neucom.2023.126287.

[17] S. Das and R. Sharma, “A TextGCN-Based Decoding Approach for Improving Remote Sensing Image Captioning,” Sep. 2024, doi: 10.1109/LGRS.2024.3523134.

[18] D. Kumar, V. Srivastava, D. E. Popescu, and J. D. Hemanth, “Dual-Modal Transformer with Enhanced Inter-and Intra-Modality Interactions for Image Captioning,” Applied Sciences (Switzerland), vol. 12, no. 13, Jul. 2022, doi: 10.3390/app12136733.

[19] O. Ondeng, H. Ouma, and P. Akuon, “A Review of Transformer-Based Approaches for Image Captioning,” Oct. 01, 2023, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/app131911103.

[20] Q. Nguyen Van, M. Suganuma, and T. Okatani, “GRIT: Faster and Better Image Captioning Transformer Using Dual Visual Features,” 2022, pp. 167–184. doi: 10.1007/978-3-031-20059-5_10.

[21] P. Zhang et al., “VinVL: Revisiting Visual Representations in Vision-Language Models,” Jan. 2021, [Online]. Available: http://arxiv.org/abs/2101.00529

[22] L. Louis and O. Adedamola Aluko, “Enhancing accessibility : a pilot study for context-aware image-Enhancing accessibility : a pilot study for context-aware image-caption to American Sign Language (ASL) translation caption to American Sign Language (ASL) translation,” 2024. [Online]. Available: https://louis.uah.edu/uah-theses/720

[23] H. Hibatullah, A. Thobirin, S. Surono, and A. Dahlan University, “Deep Belief Network (DBN) Implementation For Multimodal Cclassification Of Sentiment Analysis,” vol. 10, no. 3, 2025, doi: 10.33480/jitk.v10i2.6257.

[24] D. Vala, K. Sharma, J. Rathod, and M. Holia, “Image Captioning Using Deep Learning-An Overview,” 2024. [Online]. Available: www.tnsroindia.org.in

[25] T. Ghandi, H. Pourreza, and H. Mahyar, "Deep Learning Approaches on Image Captioning: A Review," Jan. 2022, doi: 10.1145/3617592.

[26] A. Nursikuwagus, R. Munir, and M. L. Khodra, "Hybrid of Deep Learning and Word Embedding in Generating Captions: Image-Captioning Solution for Geological Rock Images," J Imaging, vol. 8, no. 11, Nov. 2022, doi: 10.3390/jimaging8110294.

[27] S. Cho and H. Oh, "Generalized Image Captioning for Multilingual Support," Applied Sciences (Switzerland), vol. 13, no. 4, Feb. 2023, doi: 10.3390/app13042446.

[28] S. Akbar et al., "Folk Games Image Captioning using Object Attention," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 7, no. 4, pp. 758–766, Aug. 2023, doi: 10.29207/resti.v7i4.4708.