ISSN 2723-6471

Generating Image Captions in Indonesian Using a Deep Learning Approach Based on Vision Transformer and IndoBERT Architectures

Ahmad Apandi[1,*], Achmad Benny Mutiara[2], Dharmayanti[3]

[1,2,3]*Doctoral Program in Information Technology, Gunadarma University, Depok 16424, Indonesia*

(Received: November 21, 2024; Revised: December 11, 2024; Accepted: January 21, 2025; Available online: March 15, 2025)

Abstract

The primary objective of this research is to develop an image captioning system in Indonesian by leveraging deep learning architectures, specifically Vision Transformer (ViT) and IndoBERT. This study addresses the challenge of generating accurate and contextually relevant captions for images, which is a crucial task in the fields of computer vision and natural language processing. The main contribution of this research lies in integrating ViT for visual feature extraction and IndoBERT for linguistic representation to enhance the quality of image captions in Indonesian. This approach aims to overcome limitations in existing models by improving semantic understanding and contextual relevance in generated captions. The methodology involves data preprocessing, model training, and evaluation using the Flickr8k dataset, which was translated into Indonesian. The research employs various data augmentation techniques to enhance model performance. The model is trained on a combined architecture where ViT extracts visual features and IndoBERT processes textual information. The experimental procedures include training the model on the Indonesian-translated Flickr8k dataset and evaluating its performance using BLEU and METEOR scores. The training loss and validation loss graphs provide insights into the model's learning process. The results indicate that the proposed model outperforms traditional CNN+LSTM and Transformer-based models in terms of BLEU and METEOR scores. A detailed analysis of these results highlights the advantages of using ViT and IndoBERT for this task. The findings of this research have significant implications for real-world applications, such as automatic image captioning for visually impaired users, content tagging for multimedia platforms, and improvements in machine translation. Future research can explore the integration of human evaluation metrics and the use of larger datasets to enhance generalizability.

Keywords: Computer Vision; Deep Learning; Image Captioning; indoBERT; Vision Transformer;

1. Introduction

This The rapid advancement of technology impacts various aspects of life, including Artificial Intelligence (AI), which is applied in fields such as healthcare, industry, transportation, and entertainment. One prominent subfield of AI is deep learning, which uses artificial neural networks to process data hierarchically, enabling complex solutions such as speech recognition, text analysis, and object detection in images [1].

Deep learning serves as the foundation for image captioning technology, which automatically generates descriptions for images. This technology combines Computer Vision to identify objects in images with Natural Language Processing (NLP) to convert visual information into descriptive text [1], [2], [3]. The applications of image captioning range from supporting visually impaired individuals, optimizing image search, and automating content production to enhancing intelligent transportation systems.

Most previous studies have focused on the textual aspects of image captioning by optimizing language modeling capabilities through various algorithms designed to generate relevant and coherent descriptions [4], [5]. For example, NLP based modeling plays a central role in producing text descriptions based on visual feature representations extracted by the model [5], [6]. This approach involves the use of algorithms such as Recurrent Neural Network (RNN) or Long Short-Term Memory (LSTM) to process sequential data, such as words in descriptive sentences [5], [7].

In the context of image captioning in Indonesian, the primary methods involve ViT for visual feature extraction and a modified Transformer architecture [5], [8], [9] integrated with IndoBERT to generate textual descriptions. Attention-

*Corresponding author: Ahmad Apandi (ahmadapandi@staff.gunadarma.ac.id)

DOI: https://doi.org/10.47738/jads.v6i2.672

based models further enhance this capability by focusing on relevant parts of the image, improving the quality of the generated captions. ViT is a deep learning model based on the Transformer architecture designed to process visual data, such as images, videos, or other pixel-based data. This model was first introduced in the paper titled "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" by the Google Research team in 2020 [5], [8]. It brings the Transformer concept, originally popularized in natural language processing into the realm of Computer Vision [10]. ViT addresses some limitations of traditional CNN while offering greater flexibility in understanding global relationships within visual data [9].

IndoBERT, on the other hand, is a NLP model based on the BERT (Bidirectional Encoder Representations from Transformers) architecture, specifically designed to handle Indonesian text [11], [12]. This model is an adaptation of BERT, originally developed by Google, but it is specifically trained on an Indonesian-language corpus. As a result, IndoBERT possesses a deep understanding of the structure, vocabulary, and patterns of the Indonesian language, making it highly effective for various NLP tasks in this language.

Further research in this field, particularly for the Indonesian language, is essential. Developing Indonesian-language image captioning methods can support the education, media, and e-commerce sectors, enhance accessibility, and enrich Indonesian-language resources in artificial intelligence. This research aims to contribute to image captioning technology using a deep learning approach tailored to local linguistic and cultural contexts.

2. Methodology

2.1. Research Stages

This research follows nine distinct stages: Data Collection, Data Preprocessing, Data Splitting, Image Feature Extraction, Model Architecture Development, Generating Image Captions, Model Evaluation, Model Implementation into a Software Prototype, and Testing with New Data. Figure 1 provides an overview of these stages.

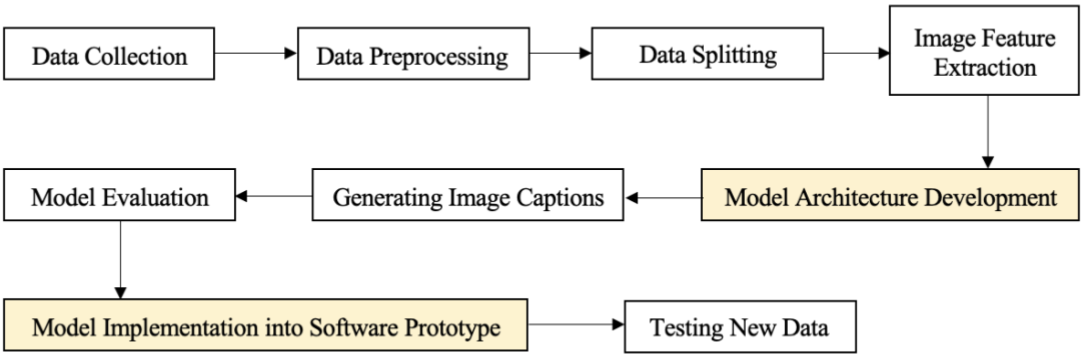


Figure 1. Model architecture for Indonesian image captioning

2.2. Data Collection

In this research, the data used is open-source data from Flickr8k, which was then translated into Indonesian. The translation of the Flickr8k dataset into Indonesian was conducted using a combination of automated and human translation methods. Initially, automated translation tools were utilized to provide a quick and efficient translation of the English captions into Indonesian. However, relying solely on automated translation posed challenges in terms of accuracy and contextual appropriateness. To address this, a manual review process was implemented, where human translators carefully examined and refined the translations to ensure fluency, coherence, and relevance to the images. During the translation process, several challenges and issues were encountered. One of the main challenges was contextual ambiguity. Some words or phrases in the original English captions had multiple possible meanings in Indonesian, which required human translators to interpret the context correctly. Without proper context, certain translations could lead to misinterpretations, affecting the overall quality of the dataset. The Flickr8k dataset is one of the datasets commonly used in image captioning research. This dataset contains 8,000 images, each accompanied by five descriptions that illustrate the content of the images. However, to use this dataset in Indonesian, a translation or adaptation process is required to convert the English descriptions into Indonesian"

2.3. Data Preprocessing

The preprocessing pipeline is an essential step in preparing the dataset for training models in image captioning tasks. It begins with loading the data and mapping textual captions to their corresponding images. This ensures that each caption is correctly paired with the appropriate image, forming the basis for effective learning. The dataset is then split into two subsets: a training set used to teach the model and a validation set to evaluate its performance and prevent overfitting [13], [14].

For the text data, preprocessing involves custom standardization techniques. These include converting all text to lowercase to maintain uniformity and removing specific unwanted characters that may not contribute meaningfully to the model's learning process. After standardization, the text is passed through a Text Vectorization layer, which converts the cleaned text into sequences of integers [10], [11]. This step is crucial, as machine learning models operate on numerical data rather than raw text. The vectorized sequences represent the vocabulary of the captions in a format that the model can process effectively.

On the image side, preprocessing focuses on data augmentation to enhance the diversity of the training dataset. Using a Sequential model, various augmentation techniques are applied to the images, such as brightness adjustment, contrast adjustment, and random cropping. These transformations mimic different real-world scenarios and variations, enabling the model to become more robust and generalize better to unseen data. Data augmentation also helps address potential overfitting by artificially expanding the size of the training dataset without requiring additional images.

By combining these steps, the preprocessing pipeline ensures that both the textual and visual data are in a format suitable for model training. The pipeline enhances the dataset's quality and variability, creating a strong foundation for the development of accurate and reliable image captioning models [15], [16].

2.4. Data Splitting

Dataset splitting is a step in preparing data for machine learning and ensures the model's performance is both robust and generalizable. It involves dividing the available dataset into distinct subsets, typically used for training, validation, and testing purposes. The primary objective of this process is to provide the model with different sets of data for training and evaluation. This approach ensures that the model does not overfit the training data and can perform well on unseen or new data [17].

The training dataset, which constitutes the largest portion of the split, is used to teach the model. In this study, 70% of the dataset, amounting to 5,600 images, is allocated for training. This dataset allows the model to learn patterns, features, and relationships within the data to minimize errors and improve predictions.

The validation dataset serves a different purpose. It is used during the model training process to fine-tune hyperparameters and evaluate the model's intermediate performance. By allocating 15% of the dataset (1,200 images) to validation, researchers can monitor how well the model generalizes to data that it has not directly trained on. This helps identify potential issues like overfitting, where the model may perform well on the training data but poorly on unseen data

Finally, the testing dataset, also 15% of the total dataset (1,200 images), is reserved for the final evaluation of the model. This dataset is entirely separate from the training and validation data and is used to provide an unbiased assessment of the model's performance. Testing ensures that the model's predictive capabilities are evaluated under conditions similar to real-world usage, offering insights into its reliability and accuracy on new data.

In this study, splitting the dataset into these proportions—70% for training, 15% for validation, and 15% for testing—provides a balanced and systematic approach to model development [15]. It ensures that the model is not only trained effectively but is also rigorously evaluated for its ability to generalize to new, unseen data, making it more reliable for practical applications.

2.5. Image Feature Extraction

Image feature extraction is a critical step in computer vision tasks where relevant visual information from images is transformed into a numerical format suitable for machine learning models. Traditional methods rely on Convolutional Neural Network (CNN) to extract features hierarchically through convolutional layers. However, ViT has emerged as a powerful alternative, leveraging the transformer architecture originally designed for natural language processing tasks to process visual data. ViT achieve this by treating images as a sequence of patches, enabling them to capture long-range dependencies and contextual relationships in an image [8], [9], [14].

Unlike CNN, which processes images using localized convolutional filters, Vision Transformer begin by dividing an input image into smaller fixed-size patches. For instance, a 224x224 image can be split into 16x16 patches, resulting in a grid of patches. Each patch is then flattened into a 1D vector and embedded using a learnable linear projection to create patch embeddings. These embeddings are supplemented with positional encodings to retain spatial information about the image's structure [16].

Once the patches are encoded, they are fed into a standard transformer encoder, which consists of layers of multi-head self-attention and feed-forward neural networks. The self-attention mechanism allows the model to capture global relationships between patches, enabling it to understand how different parts of the image relate to one another. This capability makes ViT particularly effective in extracting high-quality features from complex images, as they are not limited to local context like CNN.

ViT has several advantages for feature extraction. First, their ability to model global relationships within an image makes them well-suited for capturing detailed and meaningful features. This global understanding is especially beneficial for tasks that require identifying objects in challenging conditions, such as overlapping objects or objects with intricate details. Second, because ViT does not rely on convolutional operations, they can generalize better across different domains when pretrained on large datasets like ImageNet. Additionally, Vision Transformers are highly scalable [14]. By increasing the number of layers or heads in the transformer encoder, the model's capacity to learn complex representations can be enhanced. Pretrained ViT models can also serve as feature extractors for downstream tasks. For example, the output from the final transformer layer, representing the global feature embedding of the image, can be used as input for models performing classification, object detection, or image captioning [1], [3], [18].

2.6. Proposed Model Architecture Development

In this research, the model architecture was built using IndoBERT as the embedding used as NLP, the Vision Transformer model was used for image feature extraction. This model is designed to produce rich and efficient feature representations from input images. Before an image is provided as input to a Vision Transformer model, a preprocessing step is required to ensure data consistency. Figure 2 is the architectural design that was built.

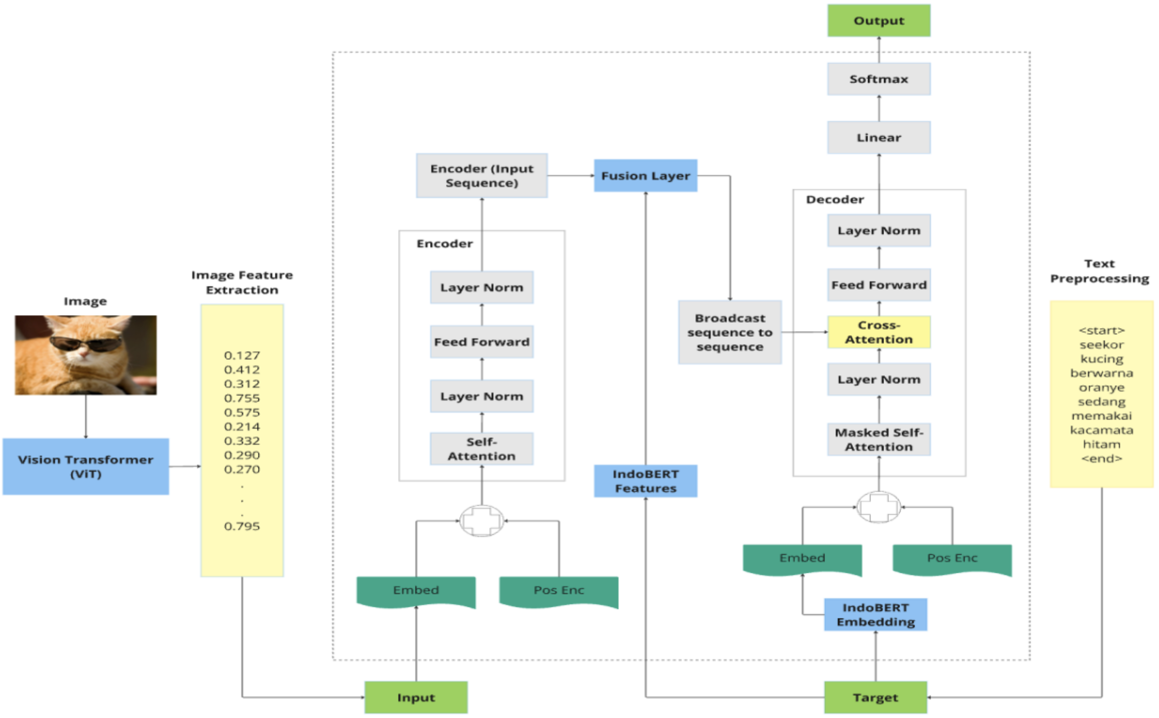


Figure 2. Proposed model architecture for Indonesian image captioning

The diagram outlines a multimodal pipeline for image captioning that integrates a ViT for image feature extraction and IndoBERT for text processing, creating a seamless framework to generate meaningful image captions. The process begins with the image feature extraction stage, where the input image (e.g., a cat wearing sunglasses) is processed by a Vision Transformer. ViT divides the image into smaller patches, encodes them into feature embeddings, and outputs a feature vector (e.g., [0.127, 0.412, 0.312, ...]) that captures the semantic and spatial attributes of the image. These feature vectors are critical as they serve as the visual representation of the image for the subsequent stages.

Simultaneously, the text preprocessing stage handles the target captions. The textual data is tokenized into meaningful units, such as <start> *seekor kucing berwarna oranye sedang memakai kacamata hitam* (an orange cat is wearing sunglasses) <end>. These tokens are then embedded using IndoBERT, a pretrained language model tailored for Indonesian, which generates embeddings that represent the contextual meaning of the words. Positional encodings are added to these embeddings to retain the sequential structure of the caption, ensuring the model understands the order of the words.

In the fusion layer, the extracted image features from the ViT are aligned and combined with the textual sequence. This fusion involves broadcasting the image feature vector to match the sequence length of the text input, allowing the encoder to process both image and text data effectively. The combined sequence is then passed through the encoder, which comprises layers of self-attention, feed-forward networks, and layer normalization. The encoder learns to capture the relationships within the input sequence, refining the representations for both image and text features.

Next, the decoder processes the caption generation. It uses masked self-attention to analyze the partial captions generated so far while ensuring the model doesn't see future tokens during training [1], [3], [19]. The decoder also employs cross-attention, where it focuses on the image features passed from the encoder, enabling it to relate specific regions of the image to the words being generated. Feed-forward layers and layer normalization further refine the decoder's outputs at each step.

Finally, the decoder's output is passed through a linear layer and a softmax activation function, which converts the outputs into probabilities over the vocabulary. The model selects the most probable word at each step, generating a coherent caption, such as "*seekor kucing berwarna oranye sedang memakai kacamata hitam*" (an orange cat is wearing sunglasses). This integration of visual and textual features, facilitated by transformers, ensures the model generates accurate and contextually relevant captions for the input image.

2.7. Generating Image Captions

After the model is created, it can be used to produce image captions by leveraging the features learned during training. The process of generating captions begins with feeding a new image into the model's ViT component. The ViT breaks the input image into smaller patches and processes these through multiple self-attention layers, resulting in a feature vector that encodes the semantic and spatial characteristics of the image. This feature vector serves as the visual representation of the input image, capturing details such as objects, their relationships, and the overall scene context [8], [9].

Once the image features are extracted, they are passed to the model's encoder, where they are combined with positional encodings to account for the spatial structure of the image. At the same time, the decoder is initialized with the <start> token, signaling the beginning of the caption. The decoder generates captions autoregressively, meaning it predicts one word at a time based on the words generated so far and the visual features provided by the encoder [20], [21], [22].

In each step of the decoding process, the decoder employs masked self-attention to process the partial caption generated so far. This ensures that the model only considers tokens up to the current position, preventing it from "peeking" at future words during generation. The decoder also uses cross-attention to focus on relevant parts of the image feature vector, allowing it to associate specific visual regions with the corresponding parts of the caption. For example, when generating the word "sunglasses," the decoder might attend to the region in the image containing the sunglasses.

As the decoder generates each word, it passes the output through a linear layer followed by a softmax function [23]. This converts the output into a probability distribution over the vocabulary. The model then selects the word with the highest probability as the next word in the caption. This process repeats until the <end> token is generated, signaling the completion of the caption.

The generated caption is designed to describe the content of the image accurately and coherently. For example, for an image of a cat wearing sunglasses, the caption might be "*seekor kucing berwarna oranye sedang memakai kacamata hitam*" (an orange cat is wearing sunglasses). The quality of the caption depends on how well the model was trained and how effectively it learned the relationships between visual and textual features during training.

What makes this process particularly effective is the interplay between the Vision Transformer and the text generation component. The Vision Transformer ensures that the visual features are rich and descriptive, capturing fine-grained details of the image. On the other hand, the decoder, supported by a pretrained language model such as IndoBERT, ensures that the generated text is fluent, contextually relevant, and grammatically correct [11], [12].

Moreover, the model's ability to focus on different parts of the image while generating specific words is a direct result of the cross-attention mechanism. This mechanism enables the decoder to dynamically shift its attention based on the current context of the caption, allowing it to describe complex scenes with multiple objects and interactions accurately.

Finally, this process is not limited to a single application. Once the model is built, it can be fine-tuned or extended for various tasks [6]. For instance, it can be adapted for multilingual captioning by incorporating language-specific embeddings or enhanced with additional modalities, such as audio or video inputs, for broader applications like video captioning. The ability to produce high-quality captions from images demonstrates the effectiveness of multimodal learning and opens the door to numerous applications, such as assisting visually impaired individuals, improving content discovery on the web, and enabling human-computer interactions in natural language.

2.8. Model Evaluation

Evaluating image captioning models involves assessing how well the generated captions match human-written captions that describe the same image. Two commonly used metrics for this purpose are the BLEU (Bilingual Evaluation Understudy) score and the METEOR (Metric for Evaluation of Translation with Explicit ORdering) score. These metrics are widely adopted in the field of natural language processing and are tailored to evaluate the quality of machine-generated text by comparing it to one or more reference texts [24], [25]. For instance, if the reference caption is "A young boy is playing with a ball," and the generated caption is "A kid is tossing a ball," BLEU might score it lower due to differences in word choice, but METEOR would score it higher because “kid” and “boy” are synonyms, and “tossing” is a reasonable variation of “playing.” By using both BLEU and METEOR, we ensure that our model is evaluated not just on exact word matches but also on semantic correctness and fluency, providing a more comprehensive assessment of the captioning quality.

BLEU is calculated based on the precision of n-grams (consecutive sequences of words) in the generated caption compared to reference captions. The BLEU score is computed using the following equation (1):

$$\mathrm{BLEU} = \mathrm{BP} \times \exp\left(\sum_{n=1}^{N} w_n \log p_n \right) \quad (1)$$

BP is the brevity penalty to handle length mismatches, w_n is the weight assigned to each n-gram precision, p_n is the precision of n-grams up to N.

METEOR takes into account synonym matching and word order. It is computed using the following equation (2):

$$METEOR = F_{mean} \times (1 - Penalty) \quad (2)$$

F_{mean} is the weighted harmonic mean of precision and recall, *penalty* is penalizing incorrect word order.

METEOR improves upon BLEU by considering semantic similarities and partial matches between words, making it more aligned with human judgment.

Both BLEU and METEOR are complementary metrics in evaluating image captioning models. BLEU is useful for measuring surface-level matches and is effective in scenarios where exact word matches are critical, such as technical or highly specific domains. METEOR, with its focus on semantic and grammatical alignment, is better suited for evaluating captions that prioritize meaning over exact word matches.

The combined use of BLEU and METEOR provides a more comprehensive evaluation of image captioning models. BLEU offers insights into the lexical precision of the captions, while METEOR evaluates their semantic richness and fluency. High scores on both metrics indicate that the model generates captions that are both accurate in terms of word choice and meaningful in terms of content [26].

However, both metrics have their limitations. Neither can fully capture the subjective nature of captioning, where multiple valid captions may exist for the same image. For this reason, these metrics are often supplemented with human evaluation to ensure that the captions are not only technically accurate but also natural and engaging. Together, BLEU and METEOR provide a robust framework for assessing the performance of image captioning models [27].

2.9. Model Implementation into Prototype

Testing is conducted by processing new image data that the model has neither been trained on nor encountered before. In this study, the testing process mirrors the training steps, starting with data preprocessing. The new image data is then inputted into the previously saved optimal model for caption generation. The model will produce image captions in the Indonesian language.

Within the developed prototype, an input form is displayed for the user. The user can browse and upload a new image they wish to caption. After selecting the image, they can click the “Generate Caption” button. The model will then process the input and generate a caption for the uploaded image.

3. Results and Discussion

3.1.Model Evaluation Results

The training loss and validation loss results for the proposed model are shown in figure 3. The graph illustrates the progression of Training Loss and Validation Loss for the proposed model during the training process. These two metrics are used to evaluate the model’s performance in learning from the training data and predicting new data. The Training Loss (solid blue line) shows a significant decrease at the start of training, especially during the first two epochs. This indicates that the model quickly learns to minimize errors on the training data. After approximately the 5th epoch, the Training Loss stabilizes, suggesting that the model has reached an optimal state for the training data.

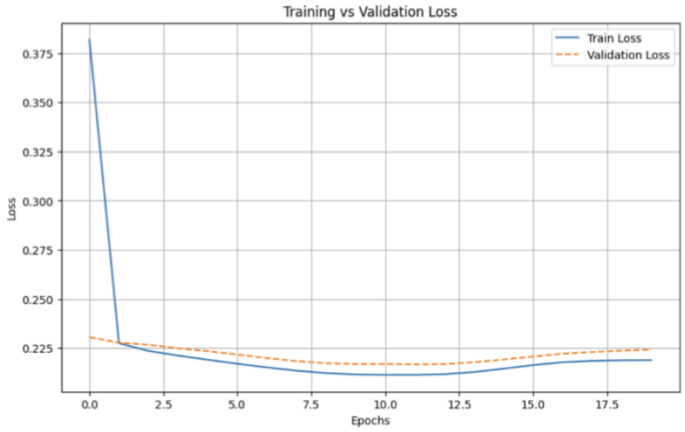


Figure 3. Graph of training lost and validation lost of the proposed model

The Validation Loss (dashed orange line) follows a similar pattern to the Training Loss at the start of training. However, the Validation Loss remains slightly higher than the Training Loss throughout the training process [21], [23], which is a common indication that the model performs better on the training data than on the validation data. In the final epoch, there is a slight increase in Validation Loss, which may indicate the onset of overfitting, where the model becomes too tailored to the training data, leading to reduced performance on new data.

This graph shows that the proposed model can learn to generate captions consistent with the training data. However, the slight discrepancy between Training Loss and Validation Loss highlights a potential area for improving the model's generalization to better generate descriptions for unseen images. Overall, the model demonstrates a reasonably good training process with stability at the end, though more attention should be given to validation performance to ensure the model can work optimally on new data.

Figure 4 shows a graph illustrating the development of BLEU and METEOR scores as the number of epochs increases during the model training process. These two metrics are used to evaluate the quality of the model's output. The BLEU score (solid blue line) in the graph shows a significant increase during the initial epochs, reflecting that the model learns quickly and produces outputs increasingly similar to the references. However, after around 5–7 epochs, the improvement slows, indicating that the model is nearing its optimal performance in understanding the training data patterns.

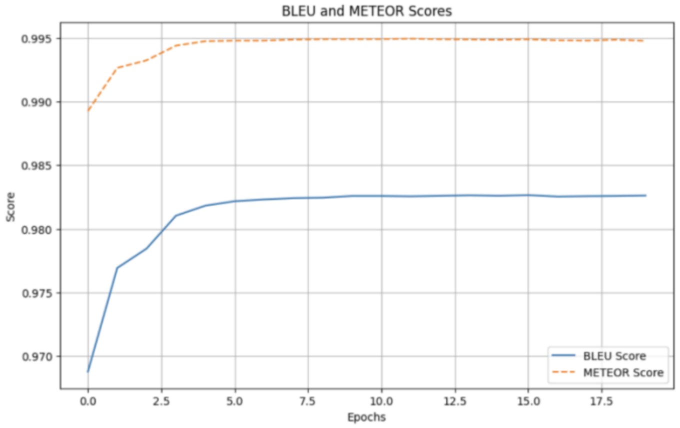


Figure 4. Graph of BLEU scores and METEOR scores of the proposed model

On the other hand, the METEOR score (dashed orange line) demonstrates more stable and consistent performance. This score reaches near-maximum values in the early epochs (around epochs 2 to 5) and then remains steady until the end. METEOR, often considered more sensitive to word order and synonyms, indicates that the model captures the text semantics quite well from the beginning.

This research also tested several deep learning models for the image captioning task, such as CNN+LSTM and Transformer. The BLEU and METEOR results for all these models are presented in table 1.

Table 1. Test Results of Several Deep Learning Models

Model	Dataset	Data Ratio	BLEU	METEOR
Proposed Model	Flickr8K	70:15:15	0.9826	0.9948
CNN+LSTM	Flickr8K	70:15:15	0.8703	0.8898
Transformer	Flickr8K	70:15:15	0.9011	0.9814

Based on the testing results shown in table 1, the performance of several deep learning models for the image captioning task is presented. The models include the proposed model, CNN+LSTM, and Transformer. The proposed model demonstrates the best performance with a BLEU score of 0.9826 and a METEOR score of 0.9948. This indicates that the model can generate high-quality image captions compared to the reference, both in word selection and sentence structure. This model leverages the combination of ViT and IndoBERT, which effectively supports visual and textual processing.

The CNN+LSTM and Transformer models show lower performance compared to the two ViT-based models. CNN+LSTM records a BLEU score of 0.8703 and a METEOR score of 0.8898, while the Transformer records a BLEU score of 0.9011 and a METEOR score of 0.9814. This performance indicates that ViT-based approaches, supported by language models like BERT, have a superior ability to understand visual features and link them with textual representations compared to traditional approaches such as CNN+LSTM.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include ACKNOWLEDGMENTS and REFERENCES, and for these, the correct style to use is “Heading 5.” Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract,” will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1,” “Heading 2,” “Heading 3,” and “Heading 4” are prescribed.

3.2.Results of Model Implementation into Prototype

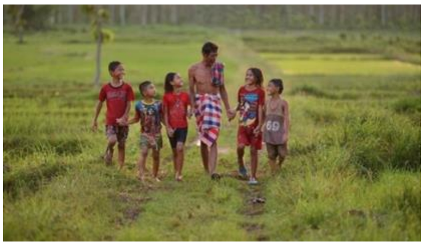
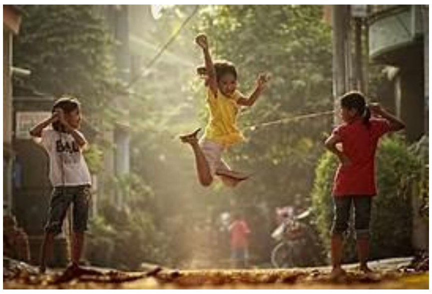
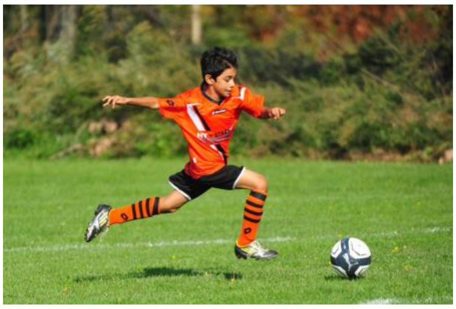
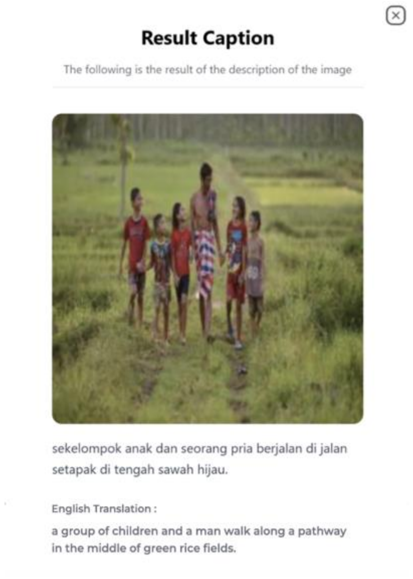
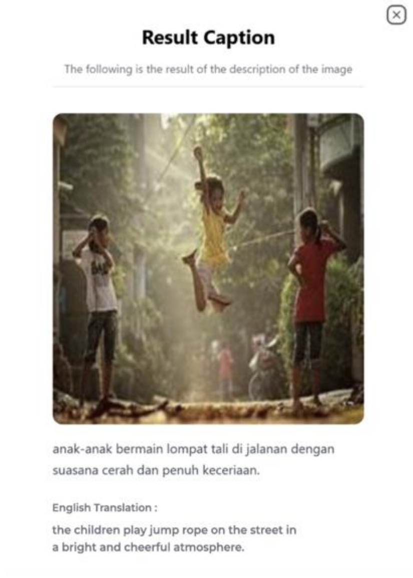
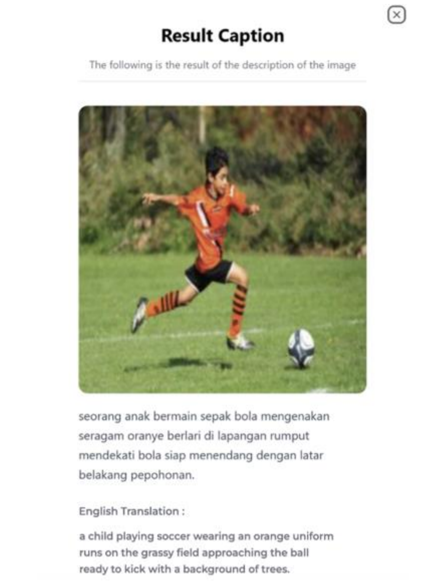
The proposed model, which has undergone the training process and achieved optimal accuracy, is then saved. The saved model can subsequently be implemented into a software prototype to generate automatic captions for images that have not been previously trained on or recognized. New data, which the model has not been trained on, is used as input for testing the model. Figure 5 illustrates a prototype for uploading new images to generate captions automatically.

Upload your image that you want to caption in Indonesian

Figure 5. Home page of the application prototype

On the home page, users can enter an image for which a caption will be created automatically by clicking the Upload Image button. Table 2 presents the captioning results produced by the developed model.

Table 2. Caption Results from New Images

Experiment 1	Experiment 2	Experiment 3
Input		
		
Output		
 Result Caption The following is the result of the description of the image sekelompok anak dan seorang pria berjalan di jalan setapak di tengah sawah hijau. English Translation : a group of children and a man walk along a pathway in the middle of green rice fields.	 Result Caption The following is the result of the description of the image anak-anak bermain lompat tali di jalanan dengan suasana cerah dan penuh keceriaan. English Translation : the children play jump rope on the street in a bright and cheerful atmosphere.	 Result Caption The following is the result of the description of the image seorang anak bermain sepak bola mengenakan seragam oranye berlari di lapangan rumput mendekati bola siap menendang dengan latar belakang pepohonan. English Translation : a child playing soccer wearing an orange uniform runs on the grassy field approaching the ball ready to kick with a background of trees.

4. Conclusion

Based on the analysis of the testing results, several conclusions can be drawn: This research successfully conducted the process of generating captions for images in Indonesian. The generated captions can serve as references for developing automated captioning systems. A model for generating captions in Indonesian has been successfully developed in this research. The designed model is a combination of a Vision Transformer to extract visual features and an adaptation of IndoBERT to generate text captions in Indonesian. The results show that the developed model achieved a BLEU score of 0.9826 and a METEOR score of 0.9948. Compared to other models, such as CNN+LSTM and Transformer, the proposed model demonstrates higher evaluation accuracy. The developed model was then implemented as a software prototype. This prototype allows users to input new images that the model has not previously trained on or recognized, automatically generating captions for those images.

Recommendations for future work related to automated image captioning include expanding and diversifying the dataset used to improve the quality of model training. This would further enhance the accuracy achieved, future research could explore adding hidden layers and varying the number of neurons to evaluate the impact on the model's accuracy. Additionally future studies could incorporate human evaluation metrics, such as expert ratings and user studies, to assess the quality and relevance of the generated captions.

5. Declarations

5.1. Author Contributions

Conceptualization: A.A., A.B.M., and D.; Methodology: A.B.M.; Software: A.A.; Validation: A.A., A.B.M., and D.; Formal Analysis: A.A., A.B.M., and D.; Investigation: A.A.; Resources: A.B.M.; Data Curation: A.B.M.; Writing Original Draft Preparation: A.A., A.B.M., and D.; Writing Review and Editing: A.B.M., A.A., and D.; Visualization: A.A. All authors have read and agreed to the published version of the manuscript.

5.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

5.3. Funding

The authors would like to thank the Doctoral Program of Information Technology, Gunadarma University, which has supported this research.

5.4. Institutional Review Board Statement

Not applicable.

5.5. Informed Consent Statement

Not applicable.

5.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Bi, Z. Zhu and Q. Meng, "Transformer in Computer Vision," *2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*, Fuzhou, China, vol. 2021, no, 1, pp. 178-188, 2021, doi: 10.1109/CEI52496.2021.9574462.

[2] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural Language Processing: State of The Art, Current Trends and Challenges," *Multimedia Tools and Applications*, vol. 81, no. 3, pp. 9657–9684, 2022.

[3] Z. U. Kamangar, G. M. Shaikh, S. Hassan, N. Mughal and U. A. Kamangar, "Image Caption Generation Related to Object Detection and Colour Recognition Using Transformer-Decoder," *2023 4th International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, Sukkur, Pakistan, vol. 2023, no. 1, pp. 1-5, 2023, doi: 10.1109/iCoMET57998.2023.10099161.

[4] J. Sudhakar, V. V. Iyer and S. T. Sharmila, "Image Caption Generation using Deep Neural Networks," *2022 International Conference for Advancement in Technology (ICONAT)*, Goa, India, vol. 2023, no. 1, pp. 1-3, 2023 doi: 10.1109/ICONAT53423.2022.9726074

[5] S. K. Satti, G. N. V. Rajareddy, P. Maddula and N. V. Vishnumurthy Ravipati, "Image Caption Generation using ResNET-50 and LSTM," *2023 IEEE Silchar Subsection Conference (SILCON)*, Silchar, India, vol. 2023, no. 1, pp. 1-6, 2023, doi: 10.1109/SILCON59133.2023.10404600.

[6] J. Y. Koh, D. Fried, and R. Salakhutdinov, "Generating Images with Multimodal Language Models," *in Proceedings of the 37th Conference on Neural Information Processing Systems*, vol. 2023 no. 1, pp. 1-12, 2023.

[7] D. Santi, A. A. Ilham, Syafaruddin and I. Nurtanio, "Image Caption Generation Through the Integration of CNN-Based Residual Network Architectures and LSTM," *2024 7th International Conference on Informatics and Computational Sciences (ICICoS)*, Semarang, Indonesia, vol. 2024, no. 1, pp. 227-232, 2024, doi: 10.1109/ICICoS62600.2024.10636926.

[8] Shourya Tyagi, Olukayode Ayodele Oki, Vineet Verma, Swati Gupta, Meenu Vijarania, Joseph Bamidele Awotunde, Abdulraufh Olanrewaju Babatunde, "Novel Advance Image Caption Generation Utilizing Vision Transformer and Generative Adversarial Networks", *Computers*, vol.13, no.12, pp.305, 2024.

[9] S. Mishra et al., "Image Caption Generation using Vision Transformer and GPT Architecture," *2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, Gharuan, India, vol. 2024, no. 1, pp. 1-6, 2024, doi: 10.1109/InCACCT61598.2024.10551257.

[10] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *in Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, Minneapolis, MN, USA, vol. 2019, no. 1, pp. 2–7, 2019

[11] B. Juarto and Yulianto, "Indonesian News Classification Using IndoBERT," *International Journal of Intelligent Systems and Applications in Engineering (IJISAE)*, vol. 11, no. 2, pp. 454–460, 2023.

[12] P. Sayarizki, Hasmawati, H. Nurrahmi, "Implementation of IndoBERT for Sentiment Analysis of Indonesian Presidential Candidates," *Journal on Computing*, vol. 9, no. 2, pp. 61–72, 2024, doi:10.34818/indojc.2024.9.2.934.

[13] P. Purwono, A. Ma'arif, W. Rahmaniar, H. Fathurrahman, A. Frisky, and Q. Haq, "Understanding of Convolutional Neural Network (CNN): A review," *International Journal of Robotics and Control Systems*, vol. 2, no. 4, pp. 739–748, 2023.

[14] K. N. Lam, H. T. Nguyen, V. P. Mai and J. Kalita, "Deep Vision Transformer and T5-Based for Image Captioning," *2023 RIVF International Conference on Computing and Communication Technologies (RIVF)*, Hanoi, Vietnam, vol. 2023, no. 1, pp. 306-311, 2023, doi: 10.1109/RIVF60135.2023.10471815.

[15] A. A. Nugraha, A. Arifinato, and Suyanto, "Generating Image Description in Indonesian Language using Convolutional Neural Network and Gated Recurrent Unit," *in Proceedings of the 7th International Conference on Information and Communication Technology (ICoICT)*, vol. 2019, no. 1, pp. 1-6, 2019. doi: 10.1109/ICoICT.2019.8835370.

[16] S. Yıldız, A. Memiş and S. Varlı, "Turkish Image Captioning with Vision Transformer Based Encoders and Text Decoders," *2024 32nd Signal Processing and Communications Applications Conference (SIU)*, Mersin, Turkiye, vol. 2024, no. 1, pp. 1-4, 2024, doi: 10.1109/SIU61531.2024.10600738.

[17] L. V. Prasad, B. S. Mounika, P. Vijaybabu, A. Teethesbabu, and C. Srikanth, "Image Caption Generator Using CNN and LSTM," *South Asian Journal of Engineering and Technology Research*, vol. 12 no. 3, 2022.

[18] R. Mulyawan, A. Sunyoto and A. H. Muhammad, "Automatic Indonesian Image Captioning using CNN and Transformer-Based Model Approach," *2022 5th International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, Indonesia, 2022, vol. 2022, no. 1, pp. 355-360, 2022, doi: 10.1109/ICOIACT55506.2022.9971855.

[19] C. Bhatt, S. Rai, R. Chauhan, D. Dua, M. Kumar and S. Sharma, "Deep Fusion: A CNN-LSTM Image Caption Generator for Enhanced Visual Understanding," *2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT)*, Dehradun, India, 2023, vol. 2023, no. 1, pp. 1-4, 2023, doi: 10.1109/CISCT57197.2023.10351389.

[20] F. Attar, F. Khan, A. Ansari, M. Saklen, A. Shaikh and D. Khan, "Image Caption Generator using Deep Learning," *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*, vol. 4, no.7, pp. 540-545 Apr. 2024.

[21] R. Kumar and G. Goel, "Image Caption using CNN in Computer Vision," *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)*, Greater Noida, India, 2023, vol. 2024, no. 1, pp. 874-878, 2024, doi: 10.1109/AISC56616.2023.10085162.

[22] G. Sairam, M. Mandha, P. Prashanth, and P. Swetha, "Image Captioning using CNN and LSTM," *in Proceedings of the 4th Smart Cities Symposium (SCS 2021)*, IET Conference Proceedings, vol. 2021, no. 11, pp. 1–6, 2021.

[23] Y. Dongare, B. M. Hardas, R. Srinivasan, V. Meshram, M. G. Aush, and A. Kulkarni, "Deep Neural Networks for Automated Image Captioning to Improve Accessibility for Visually Impaired Users", *Int J Intell Syst Appl Eng*, vol. 12, no. 2, pp. 267–281, Oct. 2023.

[24] H. Saadany dan C. Orasan, "BLEU, METEOR, BERTScore: Evaluation of Metrics Performance in Assessing Critical Translation Errors in Sentiment-Oriented Text," *Proceedings of the Translation and Interpreting Technology Online Conference*, vol. 2021, no. 1, pp. 45-54, 2021.

[25] M. Sailaja, K. Harika, B. Sridhar, and R. Singh, "Image Caption Generator using Deep Learning," *in Proceedings of the International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*, vol. 2022, no. 1, pp. 1-5, 2022.

[26] R. Subash, R. Jebakumar, Y. Kamdar, and N. Bhatt, "Automatic Image Captioning Using Convolution Neural Networks and LSTM," *in Proceedings of the International Conference on Physics and Photonics Processes in Nano Sciences*, vol. 2019, no. 1, pp. 1-10, 2019.

[27] S. E. Fatima, K. Gupta, D. Goyal and S. K. Mishra, "Image Caption Generation Using Deep Learning Algorithm," *KUEY*, vol. 30, no. 5, pp. 8118–8128, 2024.