

# Transformer+transformer architecture for image captioning in Indonesian language

Bryan Christofer Wijaya, Hendrik Santoso Sugiarto

Department of Information Technology, Faculty of Science and Technology, Calvin Institute of Technology, Jakarta, Indonesia

## Article Info

### Article history:

Received Sep 17, 2024

Revised Feb 14, 2025

Accepted Mar 15, 2025

### Keywords:

Computer vision

Deep learning

Image captioning

Natural language processing

Transformer architecture

## ABSTRACT

Image captioning in Indonesian language poses a significant challenge due to the complex interplay between visual and linguistic comprehension, as well as the scarcity of publicly available datasets. Despite considerable advancements in this field, research specifically targeting the Indonesian language remains scarce. In this paper, we propose a novel image captioning model employing a transformer-based architecture for both the encoder and decoder components. Our model is trained and evaluated on the pre-translated Flickr30k dataset in the Indonesian language. We conduct a comparative analysis of various transformer-transformer configurations and convolutional neural network (CNN)-recurrent neural network (RNN) architectures. Our findings highlight the superior performance of a vision transformer (ViT) as the visual encoder, combined with IndoBERT as the textual decoder. This architecture achieved a BLEU-4 score of 0.223 and a ROUGE-L score of 0.472.

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



## Corresponding Author:

Hendrik Santoso Sugiarto

Department of Information Technology, Faculty of Science and Technology, Calvin Institute of Technology  
Jakarta, Indonesia

Email: hendrik.sugiarto@calvin.ac.id

## 1. INTRODUCTION

Image captioning is a complex task that requires the seamless integration of visual and textual patterns to generate coherent and contextually accurate descriptions. Many existing models have achieved remarkable results on benchmark datasets in English [1]–[3], but little work has been done on image captioning in other languages, including Indonesia. As the fourth most populous country, Indonesia is home to a vast population with a rich culture. Capturing the essence of images in its language is a crucial task that enables numerous applications, including image retrieval, visual storytelling, and enhancing accessibility for visually impaired individuals. Developing an accurate and effective Indonesian image captioning model not only holds significant practical value but also has the potential to make a meaningful societal impact by bridging linguistic gaps.

Image captioning has been an active research area at the intersection of computer vision and natural language processing. Numerous approaches have been developed to address this challenge, with encoder-decoder models proving particularly effective in producing textual descriptions for images. A widely adopted encoder-decoder strategy involves using a convolutional neural network (CNN) such as VGG [4], ResNet [5], Inception [6] as image encoders, and a recurrent neural network (RNN) such as gated recurrent unit (GRU) [7], long short-term memory (LSTM) [8] as language decoders. These CNN+RNN architectures [9]–[11] have achieved impressive results on benchmark datasets such as the COCO [12] and the Flickr30k [13].

In the context of the Indonesian language, Nugraha *et al.* [14] proposed a CNN+RNN image captioning model for the Indonesian language using InceptionV3 [15] as the encoder and GRU [7] as the decoder. Their

model successfully generated grammatically correct and contextually appropriate captions, although some captions occasionally failed to align with the image content. Furthermore, Mahadi *et al.* [16] proposed a different encoder-decoder architecture using ResNet101 [5] as the encoder and LSTM [8] as the decoder. They also added adaptive attention mechanism [17] to the decoder to decide which region are attended to generate the next word for the caption. Patwari and Naik [18] also used a similar architecture, with InceptionV3 as the encoder and GRU as the decoder along with attention mechanism, which produces similar results as [16]. These models however, rely on RNN architectures, which tend to perform poorly with long sentences.

Recently, transformer-based models [19] have shown impressive performance on a wide range of language processing tasks, such as language modeling (LM) [20], machine translation [21], and question answering [22]. This success is attributed to their ability to model long-range dependencies and capture contextual information more effectively. Notable examples include generative pre-trained transformer (GPT) [23] and bidirectional encoder representations from transformers (BERT) [24]. Inspired by the success of the transformer architecture [19], Al-Faruq and Fudholi [25] developed an image captioning model on COCO dataset using CNN+transformer architecture, i.e., EfficientNet [26] as the encoder and transformer [19] as the decoder. The transformer architecture accelerates the learning process since it is no longer relied on RNN.

While all previous studies have explored the benefit of some transformer architectures, their architecture is still partially relied on either CNNs or RNNs (including their architectural limitations). In contrast, our research pursues the development of an image captioning model that operates solely on transformer architectures, inspired by recent advances in vision transformer (ViT) framework. ViT is a pure transformer model (without any CNNs mechanism) that can be applied directly to sequences of image patches. ViT has demonstrated superior performance compared to state-of-the-art CNNs while significantly reducing computational resources required for training [27]. This advancement presents a unique opportunity to create architectures that can simultaneously process both visual and linguistic data using a unified transformer framework. In this context, we propose a novel architecture for Indonesian image captioning, utilizing a pure transformer-based architecture for both the encoder and decoder.

The main contributions of this paper include the development of an accurate and effective image captioning model for the Indonesian language based on transformer+transformer architecture. It also includes evaluating its performance (both quantitatively and qualitatively) on the pre-translated Flickr30 dataset, and comparing it with other state-of-the-art image captioning models. Our experiments demonstrate the effectiveness of our proposed architecture in generating accurate and fluent Indonesian captions for images while outperforming the other alternative baselines.

## 2. METHOD

The workflow of this paper is depicted in Figure 1(a), with each part described in subsections below.

### 2.1. Dataset

The Flickr30k dataset [13] is a popular benchmark dataset in the computer vision community that is widely used for evaluating image captioning models. The images were harvested from the Flickr website, which cover a wide range of events, scenes, and activities. This dataset consists of 31,783 images, each paired with 5 corresponding captions (obtained via crowdsourcing). In this paper, we used the pre-translated Flickr30k dataset in the Indonesian language. The dataset (publicly available) was translated using Google Translate and then corrected manually using the crowdsourcing method [14].

### 2.2. Image preprocessing

As ViT expects each image to be of the same size (resolution), the images are resized into  $224 \times 224$  pixels and the pixel values are normalized with average and standard deviation of 0.5. Resizing the images to a fixed size ensures that they can be fed into the model without any compatibility issues. Normalizing the pixel values ensures that the input values are within a certain range, which helps the model training convergence.

### 2.3. Caption preprocessing

We employed the BERT tokenizer to convert the captions into integer sequences that can be fed into the model. The BERT tokenizer relies on WordPiece tokenization [28], a subword tokenization method, that adeptly addresses the challenge of handling out-of-vocabulary words encountered during training, while keeping the sequence length reasonably compact. Before tokenizing the captions, we added special "[CLS]" and

”[SEP]” tokens to the beginning and end of each caption, to indicate the start and end of the caption generation process. Additionally, we added a padding token ”[PAD]” to ensure that all input sequences have the same length. Finally, augmented sentences (with these special tokens) were converted into integer sequences.

## 2.4. Encoder (visual extractor)

The encoder plays a crucial role in image captioning, as it is responsible for extracting visual features from the input image. We used the pretrained ViT to initialize our image encoder weight parameter. ViT is a transformer-based architecture that has shown to be effective in capturing visual features from images [27]. ViT adapts the transformer’s encoder architecture [19] to process 2-dimensional images. It first decomposes the image input  $x$  into a sequence of 2D patches, denoted as  $x_p$ . These patches are obtained by dividing the original image with resolution  $(H, W)$  and channel  $C$  into non-overlapping patches of resolution  $(P, P)$ . This results in  $N$  patches, where  $N = HW/P^2$ . ViT then flatten  $i$ -th patch into 1D sequence  $y_i \in \mathbb{R}^{P^2 \cdot C}$  and performs a linear projection (1) using trainable parameters  $E$  into its base (0-th) layer embedding  $z_i^{(0)} \in \mathbb{R}^D$ .  $D$  is the size of transformer latent vector in all of its layers. Similar to BERT tokenization process, ViT also prepend [CLS] embedding ( $z_{class}^{(0)}$ ) that will serve as the whole image representation at the output of the encoder. In addition, a positional embedding  $E_{pos} \in \mathbb{R}^{(N+1) \times D}$  is added to the tokens to retain positional information.

$$z_i^{(0)} = y_i E, \quad E \in \mathbb{R}^{(P^2 \cdot C) \times D} \quad (1)$$

$$z^{(0)} = [z_{class}^{(0)}; z_1^{(0)}; z_2^{(0)}; \dots; z_N^{(0)}] + E_{pos} \quad (2)$$

The tokens are then passed to an encoder consisting of a sequence of  $L$  transformer layers. Each layer  $l$  consists of multihead self-attention (MSA), pre-layer normalization (LN), and multi layer perceptron (MLP) blocks as in (3) and (4). The model’s output is the class token embedding after layer normalization.

$$\tilde{z}^{(l)} = MSA(LN(z^{(l-1)})) + z^{(l-1)} \quad (3)$$

$$z^{(l)} = MLP(LN(\tilde{z}^{(l)})) + \tilde{z}^{(l)} \quad (4)$$

Where  $\tilde{z}^{(l)}$  represents an intermediate pre-processed embedding for calculating the  $(l)$ -th layer embedding.

We initialized the encoder weight parameters using a pretrained ViT weight parameters, which had undergone pretraining on the ImageNet-21k dataset and further fine-tuned on ImageNet 2012 (ILSVRC2012) dataset. Afterward, we implement additional model fine-tuning on our dataset. This fine-tuning step is particularly crucial due to the presence of randomly initialized cross-attention layers, as elaborated in the subsection 2.5. Moreover, this process equips the model with an internal representation of images, enabling it to effectively extract essential features for image captioning tasks.

## 2.5. Decoder (text generator)

The decoder is responsible for generating a caption for the input image, based on the visual features extracted by the encoder. In our model, the decoder weights are based on pretrained transformer-based encoder (BERT) [24]. Specifically, this paper uses a specific pretrained BERT known as IndoBERT [29], which has been trained on an extensive corpus of approximately 4 billion Indonesian words, with around 250 million sentences. BERT is a well-known model for natural language understanding and generation tasks, and it is adapted here to generate captions for the Indonesian language. Both ViT and BERT draw inspiration from the encoder architecture of the original transformer model. As a result, they share a similar underlying architecture. However, BERT still maintains the use of post-normalization, aligning with the original transformer’s design. Specifically, original BERT embedding ( $\zeta^{(l)}$ ) employs post-normalization as shown in (5) and (6).

$$\tilde{\zeta}^{(l)} = LN(MSA(\zeta^{(l-1)})) + \zeta^{(l-1)} \quad (5)$$

$$\zeta^{(l)} = LN(MLP(\tilde{\zeta}^{(l)})) + \tilde{\zeta}^{(l)} \quad (6)$$

Where  $\tilde{\zeta}^{(l)}$  represents an intermediate pre-processed embedding for calculating the  $(l)$ -th layer embedding.

However, BERT is originally designed as an encoder for various natural language processing tasks. To adapt BERT as a decoder in our encoder-decoder model for image captioning, some modifications are necessary. We employ a warm-starting approach to reconfigure the BERT model as the decoder, as implemented by Rothe *et al.* [30]. There are mainly 3 modifications:

- Unlike BERT's original architecture, where it only relies on self-attention layers, the decoder needs to be conditioned on the contextualized encoded sequence of the image features. To achieve this, cross-attention layers need to be added between the self-attention layer and the two feed-forward layers in each BERT block. As BERT does not inherently incorporate cross-attention layers, we added randomly initialized cross-attention layers that necessitate subsequent fine-tuning.
- To enable auto-regressive caption generation, the bi-directional self-attention in BERT is replaced by uni-directional self-attention layers, ensuring that the model focuses only on the previous tokens during caption generation. While the key, query, and value projection weights of the decoder's uni-directional self-attention layers are initialized with those of BERT's bi-directional self-attention layers.
- In order to define the conditional probability distribution of the output sequence, we need to add a LM head layer on top of the last decoder block. The LM head layer is responsible for generating a sequence of logit vectors. The weight parameters of the LM head layer correspond to the weight parameters of BERT's word embeddings, and hence, are not randomly initialized. The final encoder-decoder architecture is depicted in Figure 1(b).

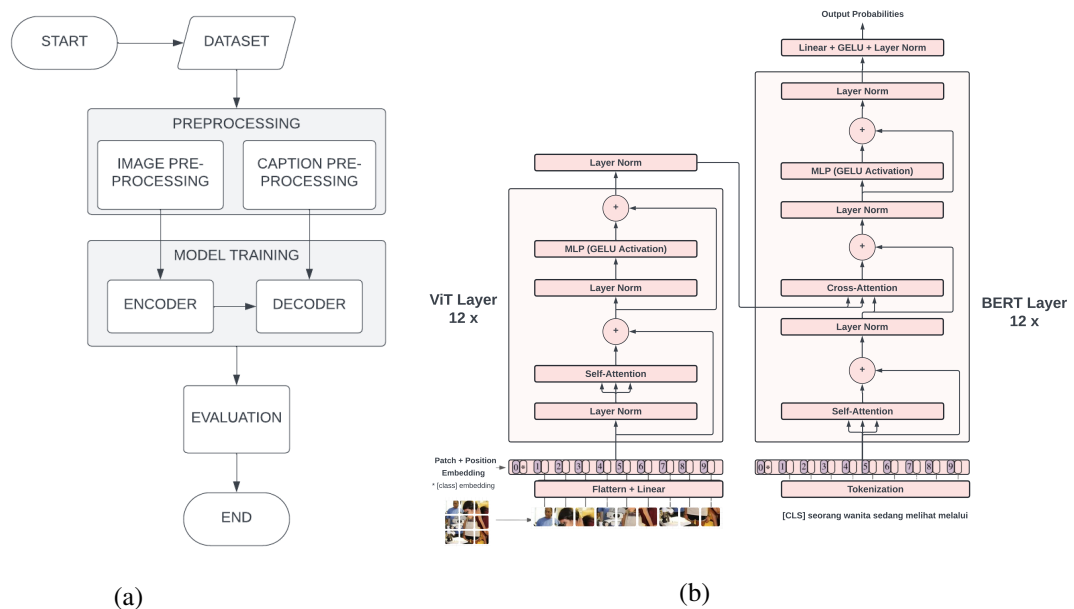


Figure 1. The overall framework of the proposed model: (a) system flowchart and (b) ViT + BERT architecture

## 2.6. Model evaluation

We used the Flickr30k dataset with the Indonesian captions to train and evaluate our model. We split the dataset into training set (80%) and test set (20%). To evaluate the performance of our proposed model, we used several standard metrics in image captioning, including BLEU and ROUGE scores. These metrics measure the similarity between the generated captions and the ground truth captions. BLEU, which stands for bilingual evaluation understudy, is a widely used metric for evaluating the quality of machine-generated texts, such as machine translation or image captions [31]. It is designed to measure the similarity between the generated text and one or more reference texts (ground truth). BLEU calculates a precision score, i.e., how much of the generated text is present in the reference text, based on  $n$ -grams, which are contiguous sequences of  $n$  words, e.g., BLEU-3 measures precision of trigrams. The BLEU score ranges from 0 to 1, with 1 indicating a perfect match with the reference text. ROUGE, which stands for recall-oriented understudy for gisting evaluation, is another set of metrics used for the automatic evaluation of text generated by natural language processing systems [32]. In contrast to BLEU, ROUGE places a specific emphasis on recall, which measure how much of the reference text is effectively captured by the generated text. In this paper, we use ROUGE-N and ROUGE-L. ROUGE-N measures the overlap of  $n$ -grams between the generated text and the reference text, while ROUGE-

L calculates the longest common subsequence between the generated text and the reference text, rewarding longer and more meaningful matches. Similar to BLEU, a higher ROUGE score indicates a better quality in the generated text, with 1 being a perfect match with the reference text.

### 3. RESULTS AND DISCUSSION

#### 3.1. Model comparison

We compared the performance of our proposed ViT+IndoBERT model with other architectures, including replicated versions of previous related works (InceptionV3+GRU [14], EfficientNet+transformer [25]). Since their original datasets are different, we reimplemented their models and evaluated them on our pre-translated Flickr30k dataset using the same train-test split. This allowed us to perform a fair comparison across all models. Since no pretrained RNN models in Indonesian are available, all RNN decoders were trained from scratch using the training data. In contrast, the transformer decoders were fine-tuned from their pretrained Indonesian models using the training data. Meanwhile, all encoders parameters were extracted from their respective pretrained versions. The results are summarized in Table 1.


As shown in Table 1, our proposed model (ViT+IndoBERT) achieved the highest BLEU and ROUGE scores, indicating that it outperformed other architectures in generating accurate and relevant Indonesian captions for images. In general, the architectural combination of transformer encoder and transformer decoder yields the best performance in comparison with other alternative encoder and decoder combinations. Specifically, transformer+transformer delivers the best performance, followed by transformer+RNN in second place and CNN+transformer in third, while CNN+RNN performs the worst. Additionally, combinations within the same model family (e.g., ViT+LSTM and ViT+GRU) exhibit comparable performance.

Table 1. Model comparisons

Enc-Dec	Architecture	BLEU 1	BLEU 2	BLEU 3	BLEU 4	ROUGE 1	ROUGE 2	ROUGE L
CNN + RNN	InceptionV3+GRU	0.282	0.127	0.065	0.035	0.222	0.051	0.210
	ResNet+LSTM	0.467	0.292	0.176	0.104	0.377	0.168	0.355
	ResNet+GRU	0.459	0.287	0.173	0.102	0.379	0.169	0.357
CNN + Tx	EfficientNet+Transformer	0.484	0.304	0.185	0.111	0.380	0.170	0.358
Tx + RNN	ViT+LSTM	0.516	0.342	0.217	0.133	0.417	0.205	0.394
	ViT+GRU	0.516	0.342	0.216	0.133	0.417	0.205	0.394
Tx + Tx	ViT+IndoBERT	0.585	0.414	0.287	0.199	0.476	0.261	0.453

Table 2 further illustrates this point by providing examples of captions generated by different architectures for the same image. In this example, the ResNet+LSTM model failed to identify the microscope, the central object in the image, while the ViT+LSTM model struggled to differentiate which woman was looking into it. The examples show that the proposed model captures contextual details more effectively than other architectures. This result demonstrates the effectiveness of using transformer-based architectures for both image and language representations in generating better captions.

Table 2. Example of Indonesian image captioning results from different architecture

Image	Architecture	Caption generated	Google translation
	ResNet+LSTM	<i>seorang pria dengan kemeja biru dan celana pendek hitam berdiri di atas bukit</i>	a man in a blue shirt and black shorts standing on a hill
	ViT+LSTM	<i>seorang wanita dengan kemeja putih sedang melihat melalui mikroskop</i>	a woman in a white shirt is looking through a microscope
	ViT+IndoBERT	<i>seorang wanita melihat melalui mikroskop sementara seorang pria melihat</i>	a woman looks through a microscope while a man looks

#### 3.2. Encoder fine-tuning

In our experiments, we observed that fine-tuning the encoder part of our model further improved the results. As Table 3 demonstrates, the fine-tuned model achieved higher BLEU and ROUGE scores compared to the model using extracted image features, which shows the importance of encoder fine-tuning for language-specific tasks. It aligns well with our objective of accurately captioning images in the Indonesian language.

Table 3. Encoder fine-tuning

Fine-tune encoder	BLEU 1	BLEU 2	BLEU 3	BLEU 4	ROUGE 1	ROUGE 2	ROUGE L
No	0.585	0.414	0.287	0.199	0.476	0.261	0.453
Yes	0.595	0.433	0.310	0.222	0.496	0.282	0.469

After fine-tuning the encoder part, we also conducted a comparison of our proposed encoder with various other state-of-the-art pretrained vision transformer models, including data-efficient image transformer (DEIT) [33], BERT pre-training of image transformers (BEIT) [34], self-distillation with no labels (DINO) [35], and DINOv2 [36]. This comparison aims to assess the performance of ViT encoder against other alternative pretrained transformer encoders for image captioning (Table 4). It's worth noting that all transformer-based encoder models outperformed the traditional models (CNN+RNN) in terms of caption quality. This highlights the superior performance of transformer-based encoders in the context of image captioning. While ViT emerged as the best-performing model in this specific experiment, it is essential to recognize the overall effectiveness of transformer-based encoders for this task.

Table 4. Comparison with other pretrained transformer-based encoder architecture

Encoder	BLEU 1	BLEU 2	BLEU 3	BLEU 4	ROUGE 1	ROUGE 2	ROUGE L
ViT	0.595	0.433	0.310	0.222	0.496	0.282	0.469
DEIT(distilled)	0.568	0.407	0.287	0.203	0.476	0.265	0.449
DINO	0.534	0.374	0.259	0.180	0.452	0.243	0.425
BEIT	0.419	0.219	0.087	0.030	0.369	0.119	0.358
DINOv2	0.562	0.402	0.283	0.200	0.471	0.260	0.443

### 3.3. Decoding with beam search

We also experimented with different decoding methods to enhance the quality of the generated captions. In particular, we applied beam search decoding to improve the fluency and coherence of the captions. Table 5 summarizes the results of using various beam search widths, showing that a width of 2 provided the best BLEU and ROUGE scores. By implementing beam search, we observed improvements of generated captions, with more contextually relevant and coherent sentences due to exploration of multiple candidate words.

Table 5. Comparison of different beam search widths

Beam search width	BLEU 1	BLEU 2	BLEU 3	BLEU 4	ROUGE 1	ROUGE 2	ROUGE L
1 (Greedy search)	0.595	0.433	0.310	0.222	0.476	0.261	0.453
2	0.602	0.437	0.311	0.223	0.498	0.283	0.472
3	0.595	0.433	0.310	0.222	0.495	0.282	0.469
4	0.592	0.430	0.307	0.219	0.494	0.281	0.468

### 3.4. Qualitative inspection

To provide a qualitative assessment of our model's performance, we present some image captions samples generated by our ViT+IndoBERT model. Figures 2(a) to 2(c) displays few images from the test dataset along with their corresponding generated captions. As shown in the figure, our model generated captions that capture the essence of the images effectively. This qualitative evaluation reinforces the quantitative results and demonstrates the practical applicability of our model for image captioning tasks in the Indonesian language.

### 3.5. Limitations

This study's findings, while promising, are subject to several limitations that may affect the generalizability and performance of the proposed model. The dataset used was derived from the English-language Flickr30k dataset, with captions translated using Google Translate and subsequently corrected through crowd-sourcing. Although this approach enabled us to create an Indonesian-captioned dataset, it introduces potential translation inconsistencies that may affect the accuracy of the captions. Limited available Indonesian pretrained transformer decoder models (IndoBERT, IndoGPT) also constraint us in experimenting other alternative transformer architectures. While IndoBERT was effective and chosen for this study, IndoGPT (not included in this paper) performed poorly with less coherent captions. These limitations highlight the need for a native Indonesian image-captioning dataset and the development of additional pretrained models for Indonesian.

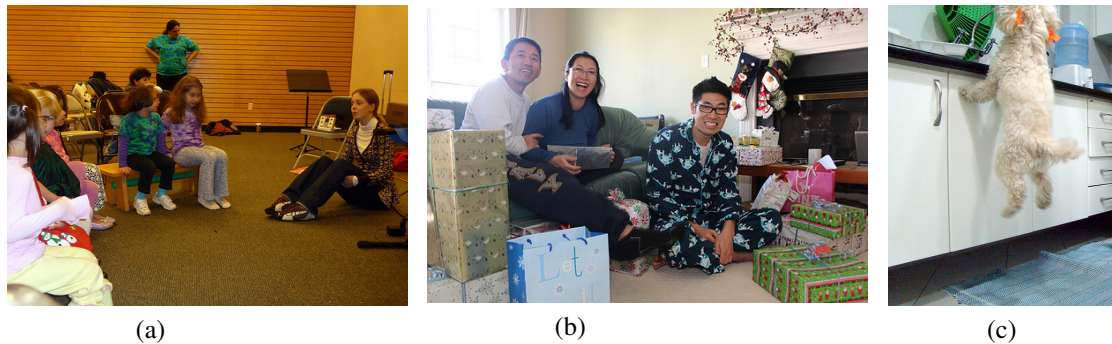


Figure 2. Generated image captions samples (a) *sekelompok anak - anak duduk di kursi di ruang kelas* (group of children sitting on chairs in classroom), (b) *dua pria dan seorang wanita duduk di lantai di sebuah ruangan dengan banyak hadiah* (two men and a woman sit on the floor in a room with many gifts), and (c) *seekor anjing putih berbulu melompat di udara di dapur* (a fluffy white dog jumps in the air at the kitchen)

#### 4. CONCLUSION

In this paper, we implemented a transformer+transformer architecture for image captioning in the Indonesian language, utilizing the ViT for image encoder and modified IndoBERT for text decoder. Our study demonstrates that transformer+transformer architectures (our proposed model) outperformed traditional CNN+RNN architectures, as well as hybrid transformer+RNN and CNN+transformer approaches. Additionally, we demonstrated that fine-tuning the encoder and incorporating beam search for decoding further enhanced the model's performance. Qualitative examples showcased the model's ability to generate coherent and contextually relevant captions. These results highlight the strong potential of transformer-based architectures for image captioning, particularly when adapted to specific languages and fine-tuned for task-specific nuances. However, the study faced limitations, including reliance on a translated dataset and limited pretrained models for Indonesian, which may influence the consistency and generalizability of our findings. Future research could benefit from native Indonesian datasets and an expanded range of pretrained models to improve adaptability across diverse language contexts. Overall, our work advances the field of image captioning by demonstrating its applicability to a broader range of languages and cultural contexts, specifically in the Indonesian language.

#### FUNDING INFORMATION

This research is supported by Calvin Institute of Technology (LPPM-R-DR-2024-01-01-005).

#### AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Bryan Christofer Wijaya	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓			
Hendrik Santoso Sugiarto	✓	✓					✓		✓	✓		✓	✓	✓

C : **C**onceptualization

M : **M**ethodology

So : **S**oftware

Va : **V**alidation

Fo : **F**ormal Analysis

I : **I**nterpretation

R : **R**esources

D : **D**ata Curation

O : **O**rganizing - **O**riginal Draft

E : **E**valuation - **R**eview & **E**dit

Vi : **V**isualization

Su : **S**upervision

P : **P**roject Administration

Fu : **F**unding Acquisition



## CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

## DATA AVAILABILITY

The authors confirm that the data supporting the findings of this study are available within the article.

## REFERENCES





- [1] C. Li et al., "mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 7241–7259, doi: 10.18653/v1/2022.emnlp-main.488.
- [2] P. Wang et al., "OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," *arXiv-Computer Science*, pp. 1–26, 2022.
- [3] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and VQA," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 13041–13049, 2020, doi: 10.1609/aaai.v34i07.7005.
- [4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015, pp. 1–14.
- [5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
- [6] C. Szegedy et al., "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.
- [7] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734, doi: 10.3115/v1/D14-1179.
- [8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
- [9] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3128–3137, doi: 10.1109/CVPR.2015.7298932.
- [10] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3156–3164, doi: 10.1109/CVPR.2015.7298935.
- [11] J. Mao, W. Xu, Y. Yang, J. Wang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," in *3rd International Conference on Learning Representations*, 2015, pp. 1–15.
- [12] T.-Y. Lin et al., "Microsoft COCO: common objects in context," *arXiv-Computer Science*, pp. 1–15, 2015.
- [13] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014, doi: 10.1162/tacl.a.00166.
- [14] A. A. Nugraha, A. Arifianto, and Suyanto, "Generating image description on Indonesian language using convolutional neural network and gated recurrent unit," in *2019 7th International Conference on Information and Communication Technology (ICoICT)*, 2019, pp. 1–6, doi: 10.1109/ICoICT.2019.8835370.
- [15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.
- [16] M. R. S. Mahadi, A. Arifianto, and K. N. Ramadhani, "Adaptive attention generation for Indonesian image captioning," in *2020 8th International Conference on Information and Communication Technology (ICoICT)*, 2020, pp. 1–6, doi: 10.1109/ICoICT49345.2020.9166244.
- [17] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3242–3250, doi: 10.1109/CVPR.2017.345.
- [18] N. Patwari and D. Naik, "En-De-Cap: An encoder decoder model for image captioning," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 2021, pp. 1192–1196, doi: 10.1109/ICCMC51019.2021.9418414.
- [19] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5999–6009.
- [20] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed length context," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2978–2988, doi: 10.18653/v1/P19-1285.
- [21] Q. Wang et al., "Learning deep transformer models for machine translation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1810–1822, doi: 10.18653/v1/P19-1176.
- [22] T. Shao, Y. Guo, H. Chen, and Z. Hao, "Transformer-based neural network for answer selection in question answering," *IEEE Access*, vol. 7, pp. 26146–26156, 2019, doi: 10.1109/ACCESS.2019.2900753.
- [23] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *Open AI*, pp. 1–12, 2018.
- [24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
- [25] U. A. A. Al-Faruq and D. H. Fudholi, "EfficientNet-Transformer for image captioning in Bahasa," in *AIP Conference Proceedings*, 2023, doi: 10.1063/5.0118155.
- [26] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *36th International Conference on Machine Learning, ICLR 2019*, 2019, pp. 1–11.
- [27] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR 2021 - 9th International Conference on Learning Representations*, 2021, pp. 1–22.







- [28] Y. Wu et al., “Google’s neural machine translation system: Bridging the gap between human and machine translation,” *arXiv-Computer Science*, vol. 2, pp. 1–23, 2016.
- [29] B. Wilie et al., “IndoNLU: Benchmark and resources for evaluating indonesian natural language understanding,” in *1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020, pp. 843–857.
- [30] S. Rothe, S. Narayan, and A. Severyn, “Leveraging pre-trained checkpoints for sequence generation tasks,” *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 264–280, 2020, doi: 10.1162/tac1.a.00313.
- [31] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 2001, pp. 311–318, doi: 10.3115/1073083.1073135.
- [32] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Association for Computational Linguistics*, Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81.
- [33] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers and distillation through attention,” *arXiv-Computer Science*, vol. 1, pp. 1–22, 2020.
- [34] H. Bao, L. Dong, S. Piao, and F. Wei, “BEiT: BERT pre-training of image transformers,” in *ICLR 2022 - 10th International Conference on Learning Representations*, 2022, pp. 1–18.
- [35] M. Caron et al., “Emerging properties in self-supervised vision transformers,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9630–9640, doi: 10.1109/ICCV48922.2021.00951.
- [36] M. Oquab et al., “DINOv2: Learning Robust Visual Features without Supervision,” *arXiv-Computer Science*, vol. 1, pp. 1–32, 2023.

## BIOGRAPHIES OF AUTHORS



**Bryan Christofer Wijaya**     is currently a Junior Data Scientist at Fata Organa Solusi, Indonesia. He earned his bachelor’s degree in IT and big data analytics from Calvin Institute of Technology, Indonesia. He has also served as a deep learning data analysis assistant at the Paul Scherrer Institut in Switzerland, where he contributed in improving deep learning models for image classification and clustering. His professional interests include web programming, data science, and artificial intelligence. He can be contacted at email: bryanchristoferwijaya@gmail.com.



**Hendrik Santoso Sugiarto**     obtained his Ph.D. in physics of complex systems from Nanyang Technological University. His research interests ranged from phase transition (in nature and society), nonlinear dynamics, complex network, probabilistic graphical modeling, deep learning, and quantum machine learning; in which he has published several research articles on those topics. He has various professional experiences: as a research scientist at a smart nation research center in Singapore, as a data scientist at the leading startup company in Indonesia, and as a visiting scholar at the largest research institute in Switzerland. Currently, he is the head of Department of IT and Big Data Analytics in Calvin Institute of Technology, Indonesia. He can be contacted at email: hendrik.sugiarto@calvin.ac.id.