Human Interaction and Emerging Technologies (IHIET-FS 2025), Vol. 196, 2025, 153–162
https://doi.org/10.54941/ahfe1005964

Vision Transformer-Based Image Captioning for the Visually Impaired

Nadeem Qazi[1], Indra Dewaji[2], and Nawaz Khan[3]

[1]University of East London, UK
[2]Universiti Teknologi Malaysia, Malaysia
[3]Walden University, USA

ABSTRACT

Digital accessibility remains a central concern in Human-Computer Interaction (HCI), particularly for visually impaired individuals who depend on assistive technologies to interpret visual content. While image captioning systems have shown notable progress in high-resource languages, languages such as Indonesian, despite having a large speaker base, continue to be underserved. This disparity stems from the lack of annotated datasets and models that account for linguistic and cultural nuances, thereby limiting equitable access to visual information for Indonesian-speaking users. To address this gap, we present a bilingual image captioning framework aimed at improving digital accessibility for visually impaired users in the Indonesian-speaking community. We propose an end-to-end system that integrates a neural machine translation component with three deep learning-based captioning architectures: CNN-RNN, Vision Transformer with GPT-2 (ViT-GPT2), and Generative Adversarial Networks (GANs). The Flickr30k dataset was translated into Indonesian using leading machine translation models, with Google Translate achieving the highest scores across BLEU, METEOR, and ROUGE metrics. These translated captions served as training data for evaluating the image captioning models. Experimental results demonstrate that the ViT-GPT2 model outperforms the others, achieving the highest BLEU (0.2599) and ROUGE (0.3004) scores, reflecting its effectiveness in generating accurate and contextually rich captions.

This work advances inclusive AI by developing culturally adaptive captioning models for underrepresented languages. Beyond its technical contributions, this research addresses key challenges in Human-Computer Interaction (HCI) by enabling inclusive, Bilingual assistive technologies. It supports the evolution of Next-Generation Work environments by equipping visually impaired individuals with tools to independently interpret visual information, an increasingly essential capability in AI-rich, visually oriented digital workspaces. In future work, the framework will be enhanced through multimodal pretraining and the integration of culturally enriched datasets, aiming to improve semantic accuracy and broaden its applicability to a wider range of linguistic communities.

Keywords: Image captioning, Multilingual AI, Vision transformer, Accessibility CNN-RNN indonesian language, GAN, Human computer interaction

INTRODUCTION

In today's digital world, visual content dominates social media, online news, and various media platforms. However, this visual-centric content presents significant barriers for individuals with visual impairments. By providing

automatic descriptions of visual content, we can empower visually impaired individuals to better understand and interact with digital media, significantly enhancing their quality of life.

Image captioning, the process of automatically generating textual descriptions for images, has emerged as a powerful solution to bridge this gap, enabling visually impaired individuals to access and interact with visual media (Sharma and Padha, 2023). Image captioning not only holds promise in accessibility applications but also in diverse fields such as healthcare (Zhao et al., 2017), where it can assist in interpreting medical imagery, and in education, where it can support language learning through visual aids.

Multilingual image captioning has seen notable progress, yet most systems remain English-centric and underperform in low-resource languages such as Indonesian. This limitation stems from the absence of culturally relevant datasets and tools capable of addressing linguistic and cultural subtleties. The Indonesian language is spoken by over 270 million people, including approximately 8 million individuals with visual impairments, one of the highest numbers in Southeast Asia. Despite this, it remains significantly underrepresented in captioning research. One of the primary barriers to advancing Indonesian captioning research is the limited availability of annotated datasets (Mulyawan et al., 2023). This scarcity makes it difficult for existing systems, which are primarily trained on Western-centric datasets, to capture the cultural nuances and contextual richness inherent to Indonesian visual media.

Addressing this gap, the research presented in this paper targets the critical intersection of image captioning, multilingual AI, and accessibility. We aim to enhance digital inclusion for visually impaired individuals, particularly within the Indonesian context, by developing AI-assisted systems that account for diverse linguistic and cultural backgrounds. In doing so, this work contributes to advancing inclusive human-computer interaction on a global scale. Specifically, we investigate the following research question: How can we develop an effective image captioning system for the Indonesian language and culture, given the scarcity of datasets and the need for accurate object recognition?

To address this research question, we explore the application of transfer learning and generative deep learning techniques for image captioning in the Indonesian language. Our methodology involves experimenting with multiple architectures, including Vision Transformers, Generative Adversarial Networks (GANs), and hybrid CNN-RNN models. We evaluate how well these models generate captions that are both linguistically accurate and culturally appropriate, comparing approaches to understand how deep learning adapts to low-resource languages like Indonesian.

This work not only identifies the models most effective under data-scarce conditions but also contributes to the broader development of inclusive AI technologies designed to support diverse linguistic and cultural communities. The rest of the paper is organized in the following sections. The next section describes the related work, followed by the adopted methodology for the proposed system in Section 3. Section 4 presents the results and analysis of

the experiment, and finally, a conclusion is drawn in the last section of the paper.

LITERATURE REVIEW

Automatic image captioning is a multidisciplinary effort combining machine learning, natural language processing, and computer vision to create meaningful and context-rich captions for images. Early approaches relied on handcrafted visual features, such as Scale-Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG), combined with rule-based or template-driven text generation methods (Lowe, 2004). However, these techniques struggled to capture complex semantic relationships and lacked generalizability across diverse visual contexts (Young et al., 2014). The pivotal shift occurred in 2012 with the introduction of AlexNet (Krizhevsky et al., 2012), which demonstrated the superior capability of CNNs in large-scale image classification. This breakthrough revolutionized visual representation learning, enabling more robust feature extraction and paving the way for vision-language integration. Subsequent research explored the fusion of CNN-based visual encoders with Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, for sequential text generation (Vinyals et al., 2015).

Image captioning research began with English datasets, but now includes multilingual data, adapted models, and translation efforts for diverse languages. The goal is to account for grammar, context, and cultural relevance, thereby improving the global inclusivity and effectiveness of image captioning. Today, state-of-the-art systems leverage Transformer-based architectures (Devlin et al., 2019), and multimodal pre-training (Li et al., 2019), further advancing the field's capabilities in generating human-like descriptions. Recent advances in image captioning reflect a global push for inclusive AI, as researchers explore frameworks for languages beyond English. Multilingual datasets are advancing image captioning to languages like Chinese, Hindi, Japanese, and Punjabi, with machine learning models customized for grammar, vocabulary, and semantics rather than just translating English datasets (Sharma and Padha, 2023).

Numerous researchers are actively developing image captioning systems for diverse languages. Contributions by (Hou et al., 2021) and (Park et al., 2021) are noted for Medical Image Captioning (MIC), while (Cheng et al., 2021) emphasize the visual properties of remote sensing images in Remote Sensing Image Captioning (RSIC). These developments are crucial as they can make technology more accessible globally and aid in critical applications like healthcare, where accurate descriptions of medical images can significantly improve diagnoses and treatments.

To the best of our knowledge, while image captioning research has made significant strides in various languages, there remains limited focus on generating captions in Indonesian, particularly with Vision Transformers. Our work is significant in that it compares the performance of multiple models for this task, providing a broader evaluation of potential approaches for generating high-quality Indonesian captions.

METHODLOGY

This research proposes a unified framework for generating image captions in the Indonesian Language by integrating three distinct models with a machine translation system, shown in Figure 1. The proposed framework consists of several stages: data pre-processing; caption translation into Indonesian using the IndoCaption Generator, a Neural Machine Translation (NMT) model; model selection through the ImageCaption Model Selector, which incorporates three deep learning models; and finally, testing, training, and performance evaluation. First, fluent and contextually accurate captions from the Flickr30k dataset are translated into Indonesian using a neural translation model, which is then employed to evaluate and select the most effective among three image captioning models. This pipeline enables automated image caption generation across languages and enhances the accessibility of image understanding systems in multilingual environments. We now define each of these stages in detail in the following subsections.

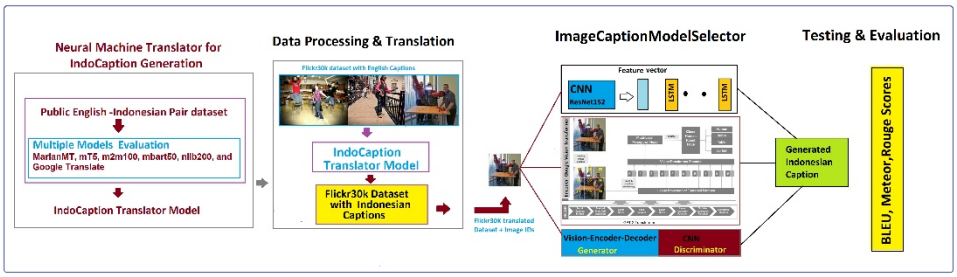

Figure 1: Proposed deep learning framework for bilingual image caption translation.

NEURAL MACHINE TRANSLATOR

The first component of the proposed framework is the Neural Machine Translator (NMT), referred to as the IndoCaption Generator in our system, which is designed to enable Bilingual caption generation. This component is responsible for translating English captions into Indonesian, thereby enhancing the framework's bilingual capabilities and facilitating effective cross-lingual content generation. For the English-to-Indonesian translation task, we leverage the SEACrowd/indo_general_mt_en_id benchmark dataset, which contains over 800,000 high-quality parallel sentences (Guntara et al., 2020).

This dataset, specifically tailored for English-Indonesian translation, was chosen for its relevance, high quality, and public availability, making it an ideal resource for model evaluation. To optimize translation performance, we employ state-of-the-art models, including Google Translate, MarianMT, mT5, M2M-100, MBART50, and NLLB-200, all fine-tuned to this dataset using advanced deep transfer learning techniques.

The Hugging Face transformers library was employed to load and process the dataset, ensuring seamless integration with the translation models. After

loading the data, comprehensive preprocessing was performed. This included resizing and normalizing images for captioning, and tokenizing and padding text for translation. During this preprocessing phase, the source and target languages were carefully aligned, preparing the data for optimal model input. These preprocessing steps ensured that the data was in the most suitable format for both the image captioning and translation models.

After training, the models were evaluated using standard translation metrics BLEU, METEOR, and ROUGE, to assess their performance in converting English captions into high-quality Indonesian. The model that achieved the best results was selected for generating Indonesian captions, which were then used for subsequent image captioning tasks. The evaluation results are presented in Table 1. As shown in the table, Google Translate emerged as the top performer across all three metrics, demonstrating strong precision, recall, and semantic accuracy. Consequently, it was selected for the machine translation layer of our proposed framework.

Table 1: Translational module evaluation

Model	BLEU Score	ROUGE Score	METEOR Score
Helsinki-NLP/opus-mt-en-id	0.27239	0.49503	0.48556
acul3/mt5-translate-en-id	0.19336	0.51563	0.51734
dl-translate-m2m100	0.27438743	0.49220	0.47123
dl-translate-mbart50	0.26232908	0.46267	0.46461
dl-translate-nllb200	0.30214486	0.51531	0.49632
google-translate	0.32593337	0.52807	0.51039

DATA PREPROCESSING FOR IMAGE CAPTIONING

Large-scale datasets are essential for developing and evaluating image captioning models, as they offer a diverse range of visual and linguistic content. Such datasets enhance model accuracy, predictive performance, generalization, and the ability to handle complex tasks (Al-Malla et al., 2022). MS COCO and Flickr30k are two prominent image captioning datasets. MS COCO contains over 330,000 images (200,000+ annotated), each with five human-written captions, covering a wide variety of everyday scenes. Flickr30k, while smaller with approximately 31,783 images, also provides five captions per image and focuses on real-world scenarios. Its rich, varied captions and manageable size make it ideal for training models in academic settings with limited resources, and it was consequently chosen for our experimental framework.

Data Pre-processing: The Flickr30k dataset was first translated into Indonesian using the Neural Machine Translation component of the framework. Following the necessary preprocessing steps required by each model, the dataset was subsequently fed into the image captioning models.

For image preprocessing, all images of the translated dataset were resized to a fixed resolution suitable for the input requirements of the Vision Image Transformer (ViT) model. Standard normalization techniques were applied

to match the distribution based on the encoder's pre-training, ensuring effective use of pre-learned visual representations. For textual data, captions were tokenized using a subword-level tokenizer aligned with the GPT-2 vocabulary. Special tokens were added to indicate the beginning and end of sequences. Padding and truncation strategies were employed to manage variable-length captions, ensuring consistency in batch processing. The translated dataset was then divided into three standard subsets: a training set for model optimization, a validation set for Hyperparameter tuning and overfitting prevention, and a test set for final evaluation.

IMAGE CAPTION MODEL

The framework integrates three image captioning architectures, including CNN-RNN, ViT-GPT2, and Generative Adversarial Network (GAN), chosen for their strengths in feature extraction, language modelling, and generative training. This enables a comparative analysis across traditional, attention-driven, and generative paradigms. The detailed description of each of these architectures is provided in the following subsections.

CNN-RNN Captioning Model: The CNN-RNN architecture adopts a classical encoder-decoder structure. The encoder, a fine-tuned ResNet-152, extracts high-level visual features from images, which are then processed by an LSTM-based decoder to generate captions. Words are represented using GloVe embeddings, capturing semantic relationships.

The model was trained using the Adam optimizer (learning rate 1e-4, batch size 32) with cross-entropy loss to guide the caption generation. A dropout rate of 0.5 was applied for regularization, and early stopping was used to prevent overfitting. Captions were generated during inference using greedy search, ensuring coherent and contextually relevant output.

VIT-GPT2 Transformer Model: This approach integrates a pre-trained Vision Transformer (ViT), i.e., (google/vit-base-patch16-224-in21k) as the encoder and a Generative Pretrained Transformer 2 (GPT-2) as the decoder for bilingual image captioning. The framework enables direct mapping from visual features to natural language descriptions.

Vision Transformer (ViT) Encoder The ViT encoder segments each input image into fixed-size patches, embeds them with positional information, and processes them through multi-head self-attention layers to capture local and global visual patterns. We employed a Google Vision Transformer (ViT) model (google/vit-base-patch16-224-in21k) pre-trained ViT model, fine-tuned on the captioning task, to extract dense, high-level semantic representations suitable for conditioning the language model.

GPT-2 Transformer Decoder: For text generation, we employed pre-trained GPT-2 on a large corpus of English, leveraging its strong autoregressive capabilities. To integrate image features into the language model, we introduced a learnable projection layer that maps the ViT-generated image embeddings into the same space as GPT-2's input token embeddings. These projected visual vectors are prepended to the input token sequence, effectively serving as a contextual prefix that conditions caption generation on the image content. This simple yet effective integration strategy

enables multimodal conditioning without introducing additional complexity such as cross-attention mechanisms.

The entire model was trained end-to-end to minimize cross-entropy loss between predicted and reference captions, employing teacher forcing. We used the AdamW optimizer with scheduled learning rate decay. Image embeddings are fixed during decoding to maintain consistent visual context. Hyperparameters such as learning rate, batch size, and number of epochs were selected empirically based on validation performance.

Generative Adversarial Network: In the third approach of caption generation over the Flickr30k dataset, we employed the Generative Adversarial Network (GAN) approach. It involves two main components: the Generator and the Discriminator. The Generator is responsible for producing captions based on input images, while the Discriminator differentiates between real (human-written) captions and those generated by the Generator.

During the experiment, both the Generator and Discriminator programs were developed. The Generator class utilizes a pre-trained Vision-Encoder-Decoder model, specifically the Vision Transformer (ViT) paired with GPT-2, to create captions from the visual input. This combination leverages ViT's ability to extract detailed visual features and GPT-2's strength in generating coherent text and producing high-quality captions.

The Discriminator class, on the other hand, employs a straightforward Convolutional Neural Network (CNN). This CNN processes both the images and the accompanying captions, assessing whether the captions are real or generated. The discriminator's effectiveness is crucial as it provides feedback to the Generator, guiding it to improve its captioning accuracy over time through the adversarial training process.

RESULT & DISCUSSION

The final component of the proposed framework evaluated the three models on the testing dataset from the translated Flickr30k dataset using standard metrics, including BLEU, METEOR, and ROUGE. The results of this evaluation are summarized in Figure 2a, while Figure 2b shows two samples of the image caption translation through our proposed framework using Vision Transformer as an image caption model. We chose BLEU because it measures n-gram overlap, emphasizing syntactic precision. It is a commonly used metric for evaluating machine translation and image captioning models. It compares a generated sentence with target sentences, giving a score of 1.0 for exact matches and 0.0 for no similarity.

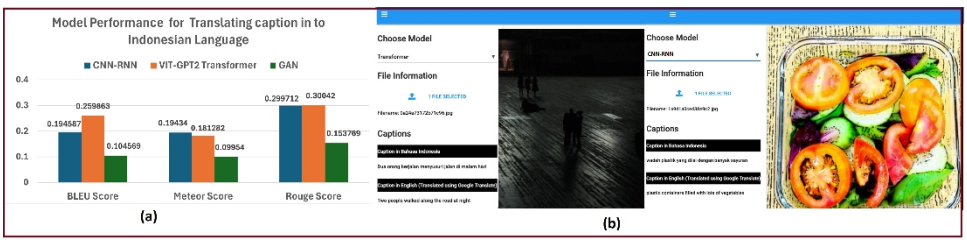

Figure 2: Proposed deep learning framework for bilingual image caption translation.

The closer the generated sentence is to the target, the higher the BLEU score. METEOR accounts for synonyms and paraphrasing, providing a more semantically aware assessment. ROUGE evaluates content recall based on the longest common subsequence. Together, they offer a comprehensive view of both lexical and semantic fidelity.

As can be seen in Figure 2a, the evaluation metrics, i.e., BLEU, METEOR, and ROUGE, for the CNN-RNN models are relatively well-balanced. The close alignment between BLEU and METEOR reflects a good balance between lexical accuracy and semantic similarity. The model's higher METEOR score further highlights its effectiveness in capturing semantically meaningful information, and its ROUGE score aligns with its ability to cover relevant content.

In contrast, the ViT-GPT2-Transformer exhibits a notably higher BLEU score relative to its METEOR score, indicating a preference for exact n-gram matches over semantically varied paraphrases. The model also achieved the highest ROUGE score (0.300), suggesting superior coverage of relevant content compared to the other models. The CNN-RNN model followed closely with a ROUGE score of 0.299, while the GAN model performed considerably worse with a score of 0.154. This suggests the model favours surface-level similarity, though its high ROUGE score indicates strong content relevance overall.

The GAN model, by comparison, yields consistently low scores across all metrics, suggesting limited ability to generate coherent, accurate, and semantically relevant captions. Its poor performance may stem from common challenges in GAN-based text generation, such as mode collapse and training instability, which may require architectural or methodological improvements.

Overall, the ViT-GPT2 Transformer outperformed the other models, particularly in the BLEU score of 0.259 as compared to the corresponding BLEU scores of 0.194 and 0.10 achieved through CNN-RNN and GAN, respectively. The ROUGE scores of ViT-GPT2 Transformer model are also higher than the other two models, indicating high precision and relevance in the generated captions.

Comparing our work with the other researchers' work (Mahadi et al., 2020) achieved BLEU-4 scores of 0.247 employing VGG16 and 0.274 through ResNet101 as encoders and LSTM as decoder over both Flickr30k and MSCOCO. In contrast, our BLEU-4 scores validated on translated Flickr30K were 0.194587 and 0.259863 with CNN-RNN and VIT-GPT2 Transformer, respectively. Additionally, our scores significantly outperform the 0.024 BLEU-4 reported by (Nugraha et al., 2019), demonstrating the effectiveness of our approach.

CONCLUSION

In conclusion, this work presents a significant advancement in Indonesian image captioning by leveraging a translated version of the Flickr30k dataset, providing a broader and more diverse evaluation benchmark. Our unified framework integrates three image captioning models and a machine translation system, generating bilingual captions in both English

and Indonesian. The ViT-GPT2 Trans- former model outperformed others in both BLEU and ROUGE scores, ensuring high-quality captions. This work is especially valuable for visually impaired individuals, providing an accessible tool for understanding visual content in Indonesian.

Future work will explore more sophisticated evaluation methods and multimodal pre-training strategies to enhance semantic accuracy. By addressing the lack of linguistically and culturally inclusive AI tools, this framework tackles a key challenge in Human-Computer Interaction-ensuring accessible and context-aware interactions for users in under-represented language communities. It can be integrated into assistive technologies, educational platforms, and digital interfaces to foster more inclusive, adaptive, and human-centred AI experiences.

REFERENCES

Al-Malla, M. A., A. Jafar, and N. Ghneim (2022). Image captioning model using attention and object features to mimic human image understanding. Journal of Big Data 9 (1), 20.

Cheng, Q., Y. Zhou, P. Fu, Y. Xu, and L. Zhang (2021). A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14, 4284–4297.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, Volume 1 (long and short papers), pp. 4171–4186.

Guntara, T. W., A. F. Aji, and R. E. Prasojo (2020, May). Benchmarking multi-domain English-Indonesian machine translation. In R. Rapp, P. Zweigenbaum, and S. Sharoff (Eds.), Proceedings of the 13th Workshop on Building and Using Comparable Corpora, Marseille, France, pp. 35–43. European Language Resources Association.

Hou, D., Z. Zhao, Y. Liu, F. Chang, and S. Hu (2021). Automatic report generation for chest x-ray images via adversarial reinforcement learning. IEEE Access 9, 21236–21250.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25.

Li, L. H., M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang (2019). Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110.

Mahadi, M. R. S., A. Arifianto, and K. N. Ramadhani (2020). Adaptive attention generation for Indonesian image captioning. In 2020 8th International Conference on Information and Communication Technology (ICoICT), pp. 1–6.

Mulyawan, R., A. Sunyoto, and A. H. M. Muhammad (2023). Pre-trained cnn architecture analysis for transformer-based Indonesian image caption generation model. JOIV: International Journal on Informatics Visualization 7(2), 487–493.

Nugraha, A. A., A. Arifianto, and Suyanto (2019). Generating image description in Indonesian language using a convolutional neural network and gated recurrent unit. In 2019 7th International Conference on Information and Communication Technology (ICoICT), pp. 1–6.

Park, H., K. Kim, S. Park, and J. Choi (2021). Medical image captioning model to convey more details: Methodological comparison of feature difference generation. IEEE Access 9, 150560–150568.

Sharma, H. and D. Padha (2023, April). A comprehensive survey on image captioning: From handcrafted to deep learning-based techniques, a taxonomy, and open research issues. Artif. Intell. Rev. 56 (11), 13619–13661.

Vinyals, O., A. Toshev, S. Bengio, and D. Erhan (2015, June). Show and tell: A neural image caption generator. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, CA, USA, pp. 3156–3164. IEEE Computer Society.

Young, P., A. Lai, M. Hodosh, and J. Hockenmaier (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics 2, 67–78.

Zhao, Y., S. Wu, L. Reynolds, and S. Azenkot (2017). The effect of computer-generated descriptions on photo-sharing experiences of people with visual impairments. Proceedings of the ACM on Human-Computer Interaction 1, 1–22.