Image Captioning Using an InceptionV3 Encoder and a Vanilla Transformer Decoder on the Flickr30k Dataset

Titis Dila Fajari[1], Ervin Yohannes[2]

[1,2] S1 Teknik Informatika, Universitas Negeri Surabaya
[1]titis.22105@mhs.unesa.ac.id
[2]ervinyohannes@unesa.ac.id

***Abstract*—Image captioning aims to automatically generate textual descriptions that accurately represent visual content. This study proposes a Sequence-to-Sequence image captioning framework that integrates an InceptionV3 encoder with a Vanilla Transformer decoder. The InceptionV3 network is employed to extract visual features from images, while the Vanilla Transformer generates captions by modeling contextual relationships between visual and textual representations. Experiments were conducted on the Flickr30k dataset using two data split scenarios (80:10:10 and 70:15:15) and two training configurations to evaluate the effectiveness of the proposed framework. Model performance was assessed using BLEU, METEOR, ROUGE-L, and CIDEr metrics. The best results were achieved using the 80:10:10 train-validation-test split and training configuration 1, obtaining BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores of 0.4252, 0.2946, 0.2176, and 0.1601, respectively, along with METEOR, ROUGE-L, and CIDEr scores of 0.4155, 0.4681, and 0.5059. These findings demonstrate that the proposed InceptionV3–Vanilla Transformer architecture is effective in generating accurate and contextually relevant image captions on the Flickr30k dataset.**

***Keywords*—Image Captioning, InceptionV3, Vanilla Transformer, Sequence-to-Sequence, Flickr30k.**

I. Introduction

The rapid growth of internet usage has significantly increased the volume of visual data shared through digital platforms. According to the Indonesian Internet Service Providers Association (APJII), internet users in Indonesia reached 229.4 million people, or approximately 80.66% of the population, in 2025 [1]. This growth creates a demand for automatic methods capable of understanding image content, where image captioning plays an important role by generating textual descriptions for applications such as image retrieval, content recommendation, and accessibility services[2], [3].

Most image captioning systems employ a Sequence-to-Sequence architecture that combines a Convolutional Neural Network (CNN) encoder with a Long Short-Term Memory (LSTM) decoder. Although effective, CNN–LSTM models often struggle to capture complex semantic relationships and may suffer from the information bottleneck problem, where image features are compressed into a fixed-length representation before decoding, leading to the loss of important visual information [4], [5].

Transformer-based architectures have emerged as a promising alternative because they can model long-range dependencies through efficient parallel processing and self-attention mechanisms [6], [7]. Previous studies have demonstrated the effectiveness of CNN–Transformer architectures for image captioning [8], [9]. Nevertheless, performance may vary depending on experimental settings such as data split strategies and training configurations.

To address this issue, this study investigates an image captioning framework that combines an InceptionV3 encoder with a Vanilla Transformer decoder on the Flickr30k dataset. Its performance is evaluated using BLEU, METEOR, ROUGE-L, and CIDEr metrics under different experimental settings.

II. Literature Review

A. Image Captioning

Image captioning is a multimodal task that combines computer vision and natural language processing (NLP) to automatically generate textual descriptions from images [10]. It is commonly formulated as a Sequence-to-Sequence problem, where visual features are translated into a sequence of words representing image content [11].

B. InceptionV3

InceptionV3 is a convolutional neural network architecture designed to improve computational efficiency while maintaining strong feature extraction capability [12], [13]. It employs Inception modules and convolution factorization techniques to reduce computational complexity without sacrificing performance [9].

C. Vanilla Transformer

The Transformer architecture consists of encoder and decoder components that utilize attention mechanisms to model dependencies within a sequence [11], [14]. In image captioning, the decoder generates captions by attending to both previously generated tokens and visual features extracted by the encoder. The attention mechanism is defined as:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

where Q, K, and V denote the query, key, and value matrices, respectively, and d_k represents the dimension of the key vector [15].

D. Filckr30k

Flickr30k is a publicly available benchmark dataset widely used in image captioning research and multimodal learning [16]. The dataset contains 31,783 images collected from Flickr, where each image is paired with five human-written captions describing its visual content [16].

E. Related Work

TABEL I
SUMMARY OF RELATED WORK

Ref	Method	Key Results
[17]	InceptionV3 + Transformer	BLEU-4:0.68, CIDEr:0.57, METEOR: 0.26
[18]	Multiple CNN encoders + Transformer	ResNet50 achieved the best performance. BLEU-4: 21.82, CIDEr: 62.63, METEOR: 21.11
[5]	InceptionV3 + LSTM	BLEU-4: 14.4, METEOR: 26.7

Existing studies have demonstrated the effectiveness of Transformer-based image captioning models with different CNN encoders. However, variations in experimental settings, including data split strategies and training configurations, make it difficult to assess the robustness of reported results across studies.

III. RESEARCH METHOD

This study develops an image captioning framework based on a Sequence-to-Sequence architecture that combines an InceptionV3 encoder and a Vanilla Transformer decoder. The methodology comprises literature review, dataset preparation, text and image preprocessing, dataset splitting, model development, training and validation, and performance evaluation. The Flickr30k dataset is used as the experimental dataset, while model performance is assessed using BLEU, METEOR, ROUGE-L, and CIDEr metrics. The complete research workflow is shown in Figure 1.

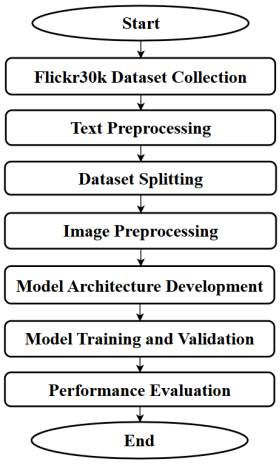

Figure 1. Research workflow

A. Flickr30k Dataset Collection

The Flickr30k dataset used in this study was obtained from Kaggle and consists of 31,783 images accompanied by textual descriptions. The dataset contains two main components: (1) a collection of images in JPEG format and (2) a caption file containing image identifiers and corresponding descriptions. Each image is paired with five human-written captions that describe its visual content from different perspectives.

B. Text Preprocessing

Text preprocessing converted image captions into numerical sequences for model training. The process included text cleaning, tokenization, vocabulary construction, caption encoding, and sequence padding. Captions were converted to lowercase, tokenized, and augmented with <SOS> and <EOS> tokens. A vocabulary was then constructed by mapping unique tokens to numerical indices, while out-of-vocabulary words were replaced with <UNK>. Finally, captions were encoded and padded with <PAD> tokens to ensure uniform sequence lengths.

C. Data Splitting

The Flickr30k dataset was partitioned into training, validation, and testing subsets using two split ratios, namely 80:10:10 and 70:15:15. The use of multiple data split scenarios enables an evaluation of model performance under different training data proportions. The distribution of images for each scenario is summarized in Table II.

TABEL II
DATASET SPLIT SCENARIOS

Split Ratio	Total Images	Training	Validation	Testing
80:10:10	31,783	25,426	3,178	3,179
70:15:15	31,783	22,248	4,767	4,768

D. Image Preprocessing

Image preprocessing was performed to adapt image inputs to the requirements of the InceptionV3 encoder and improve model generalization. The preprocessing pipeline consisted of image resizing, random cropping, data augmentation through random horizontal flipping, and normalization using ImageNet mean and standard deviation values. Images were resized to 320 × 320 pixels and cropped to 299 × 299 pixels before being fed into the InceptionV3 encoder.

E. Model Architecture Development

The proposed image captioning model employs a Sequence-to-Sequence architecture consisting of an InceptionV3 encoder and a Vanilla Transformer decoder.

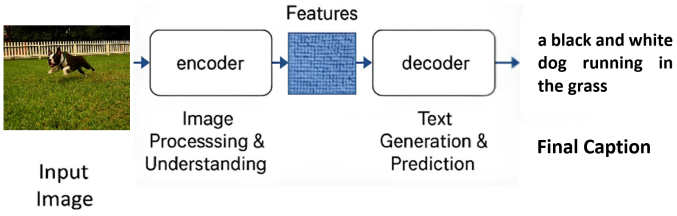

Figure 2. Architecture of the proposed InceptionV3–Vanilla Transformer model

1) Proposed Model Workflow

The workflow begins with the Flickr30k dataset, followed by text preprocessing, dataset splitting, and image preprocessing. The processed data are organized into datasets and dataloaders before being fed into the InceptionV3 encoder, where fine-tuning is applied to the Mixed_6 and Mixed_7 layers for visual feature extraction. The extracted visual features are projected into visual tokens and passed to the Vanilla Transformer decoder for caption generation. The encoder and decoder are

trained jointly, and the best-performing model is used to generate captions through Beam Search. Finally, the generated captions are evaluated using BLEU, METEOR, ROUGE-L, and CIDEr metrics.

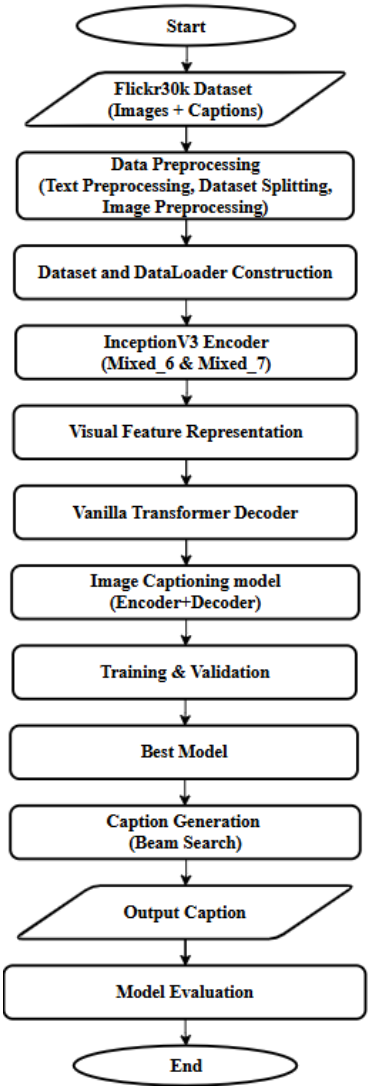

Figure 3. Workflow of the proposed InceptionV3–Vanilla Transformer model

TABEL III
MODEL CONFIGURATION

Parameter	Value
Fine-Tuning Layers	Mixed_6, Mixed_7
Decoder Layers	3
Attention Heads	8
Embedding Dimension	512
Feed-Forward Dimension	2048
Dropout Rate	0.2

F. *Model Training and Validation*

The proposed InceptionV3–Vanilla Transformer model was trained using two configurations to evaluate the impact of different learning rates and batch sizes on model performance. The training settings are summarized in Table III.

TABEL IV
TRAINING CONFIGURATIONS

Parameter	Configuration 1	Configuration 2
Learning Rate	4×10^{-5}	1×10^{-4}
Batch Size	32	64
Optimizer	AdamW	AdamW
Loss Function	Label Smoothing Loss	Label Smoothing Loss
Maximum Epochs	50	50
Gradient Clipping	1.0	1.0
Scheduler	ReduceLROnPlateau	ReduceLROnPlateau
Early Stopping Patience	10	10

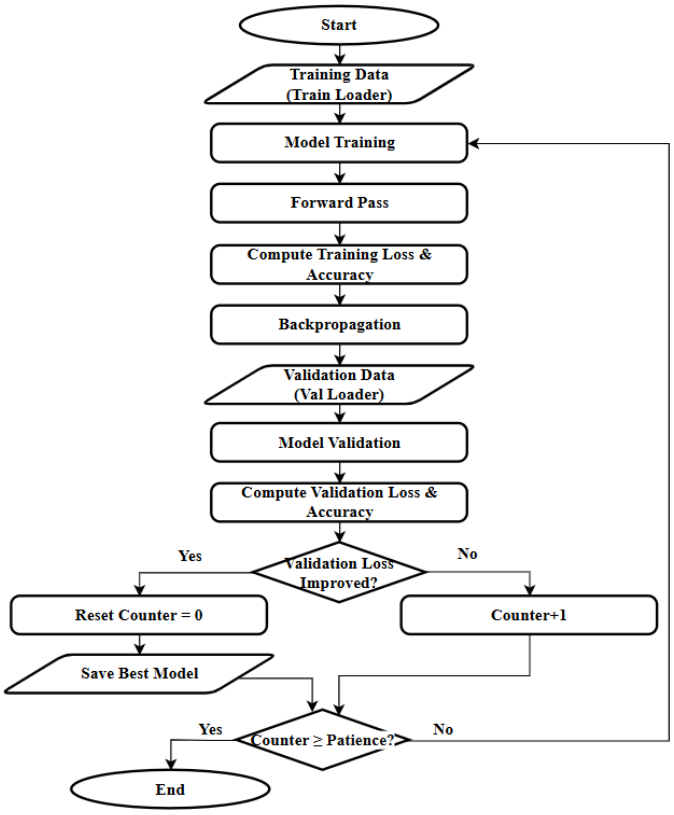

Figure 4. Training and validation workflow

Based on Figure 4, the training process begins with data from the training set, followed by forward propagation, loss computation, and backpropagation to update model parameters. After each epoch, the model is evaluated on the validation set by computing validation loss and accuracy. Early stopping is applied based on validation loss, where the best-performing model is saved whenever an improvement is observed. Training is terminated when the validation loss fails to improve for a predefined number of consecutive epochs.

G. *Performance Evaluation*

Model performance was evaluated on the Flickr30k test set using Beam Search decoding. The generated captions were compared with reference captions using four evaluation metrics: BLEU, ROUGE-L, METEOR, and CIDEr. These metrics

assess caption quality from lexical, structural, and semantic perspectives.

1) *Bilingual Evaluation Understudy (BLEU)*

BLEU measures the n-gram overlap between generated and reference captions and is widely used in image captioning evaluation [10]. Let p_n denote the n-gram precision, w_n the corresponding weight, and $BP$the brevity penalty. The BLEU score is computed as:

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (2)$$

2) *Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L)*

ROUGE-L evaluates the similarity between generated and reference captions based on the Longest Common Subsequence (LCS), capturing sentence structure and word order [17]. Let R_{LCS} and P_{LCS} denote the LCS-based recall and precision, respectively. The ROUGE-L score is defined as:

$$F_{LCS} = \frac{(1+\beta^2) \times R_{LCS} \times P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \quad (3)$$

3) *Metric for Evaluation of Translation with Explicit Ordering (METEOR)*

METEOR evaluates caption quality by considering precision, recall, and word ordering, providing better correlation with human judgment than exact word matching alone [16]. The metric is calculated as:

$$METEOR = F_{mean} \times (1 - Penalty) \quad (4)$$

4) *Consensus-based Image Description Evaluation (CIDEr)*

CIDEr measures the consensus between generated captions and multiple human-written references using TF-IDF weighted n-grams[17], where g^ndenotes the TF-IDF representation of n-grams and $w_n$represents the weight assigned to each n-gram level. The score is computed as:

$$CIDEr(c_i, S_i) = \frac{1}{N} \sum_{n=1}^{N} w_n \frac{g^n(c_i) \cdot g^n(S_i)}{\| g^n(c_i) \| \| g^n(S_i) \|} \quad (5)$$

IV. Result and Discussion

A. *Training Results*

The InceptionV3–Vanilla Transformer model was trained using two data split scenarios (80:10:10 and 70:15:15) and two training configurations. Training performance was analyzed using loss and accuracy curves to evaluate the convergence behavior and stability of the model under different experimental settings. The results for each scenario are presented as follows.

1) *80:10:10 Split – Configuration 1*

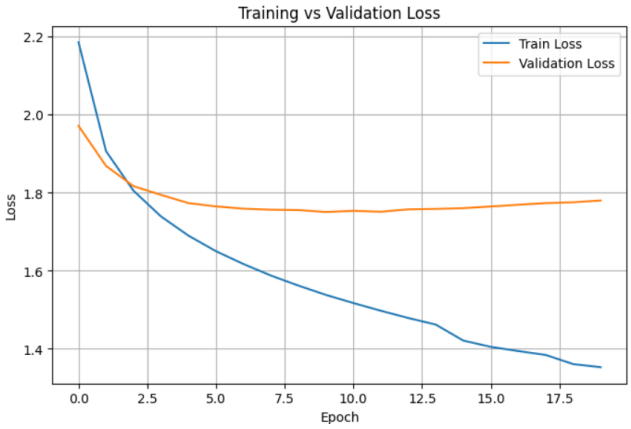

Figure 5. Training and validation loss curves for the 80:10:10 split (Configuration 1)

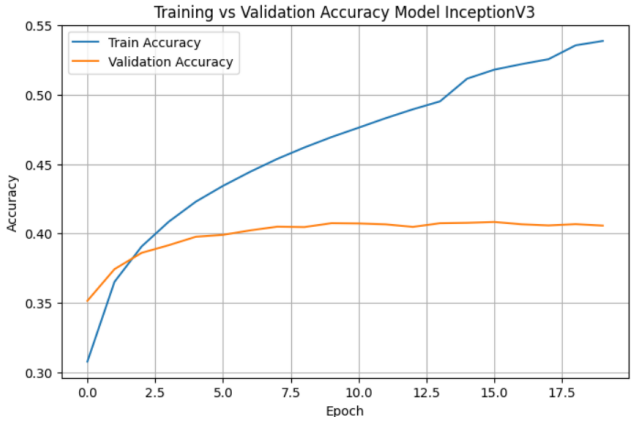

Figure 6. Training and validation accuracy curves for the 80:10:10 split (Configuration 1)

The best model was obtained at epoch 10, while training stopped at epoch 20 due to early stopping. As shown in Figure 5 and Figure 6, training loss decreased consistently and accuracy improved throughout training, whereas validation performance peaked at epoch 10 with a validation accuracy of 0.4076. These results indicate stable model convergence under this configuration.

2) *80:10:10 Split – Configuration 2*

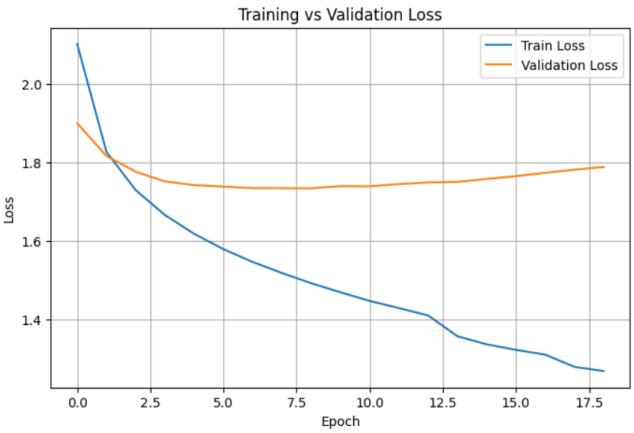

Figure 7. Training and validation loss curves for the 80:10:10 split (Configuration 2)

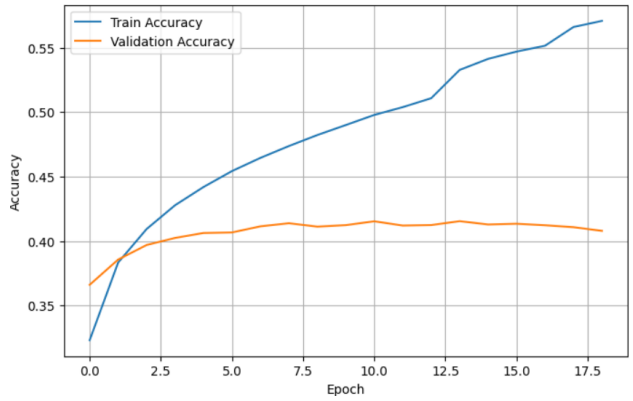

Figure 8. Training and validation accuracy curves for the 80:10:10 split (Configuration 2)

The best model was obtained at epoch 9, while training stopped at epoch 19 due to early stopping. As shown in Figure 7 and Figure 8, training loss decreased consistently and accuracy improved throughout training, whereas validation performance peaked at epoch 9 with a validation accuracy of 0.4112. These results indicate stable model convergence under this configuration.

3) *70:15:15 Split – Configuration 1*

Figure 9. Training and validation loss curves for the 70:15:15 split (Configuration 1)

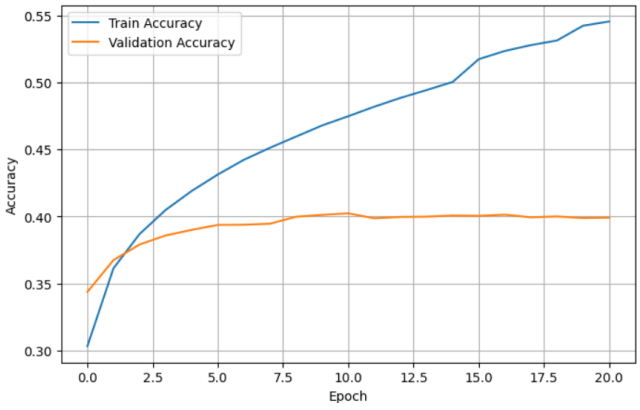

Figure 10. Training and validation accuracy curves for the 70:15:15 split (Configuration 1)

The best model was obtained at epoch 11, while training stopped at epoch 21 due to early stopping. As shown in Figure 9 and Figure 10, training loss decreased consistently and accuracy improved throughout training, whereas validation performance peaked at epoch 11 with a validation accuracy of 0.4020. These results indicate stable model convergence under this configuration.

4) *70:15:15 Split – Configuration 2*

Figure 11. Training and validation loss curves for the 70:15:15 split (Configuration 2)

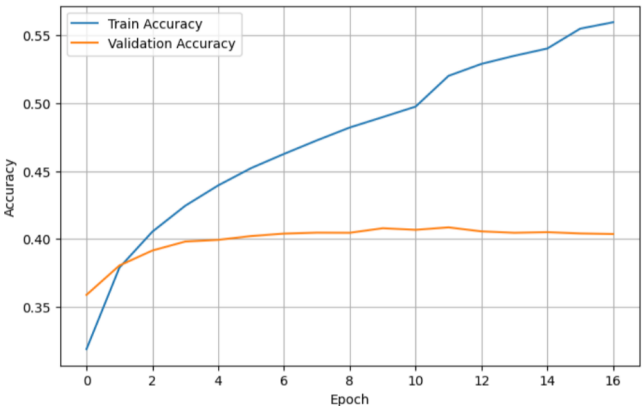

Figure 12. Training and validation accuracy curves for the 70:15:15 split (Configuration 2)

The best model was obtained at epoch 7, while training stopped at epoch 17 due to early stopping. As shown in Figure 11 and Figure 12, training loss decreased consistently throughout training, whereas validation loss reached its minimum value of 1.7536 at epoch 7 before gradually increasing. Validation performance also peaked at epoch 7, achieving the highest validation accuracy of 0.4040. These results indicate stable model convergence under this configuration.

B. *Model Evaluation Results*

To evaluate the effectiveness of the proposed decoder, the InceptionV3–Vanilla Transformer model was compared with the baseline InceptionV3–LSTM model under two data split scenarios and two training configurations. The evaluation results are presented in Table V and Figure 13.

TABEL V
PERFORMANCE COMPARISON OF INCEPTIONV3–LSTM AND INCEPTIONV3–VANILLA TRANSFORMER

Model	BLEU-4	METEOR	ROUGE-L	CIDEr
InceptionV3-LSTM (80:10:10, C1)	0.1477	0.3953	0.4523	0.4252
InceptionV3-Transformer (80:10:10, C1)	0.1601	0.4155	0.4681	0.5059
InceptionV3-LSTM (80:10:10, C2)	0.1462	0.3925	0.4525	0.4135
InceptionV3-Transformer (80:10:10, C2)	0.1548	0.4178	0.4654	0.4912
InceptionV3-LSTM (70:15:15, C1)	0.1446	0.3905	0.4486	0.4074
InceptionV3-Transformer (70:15:15, C1)	0.1541	0.4069	0.4615	0.4690
InceptionV3-LSTM (70:15:15, C2)	0.1441	0.3880	0.4473	0.4035
InceptionV3-Transformer (70:15:15, C2)	0.1582	0.4128	0.4664	0.4859

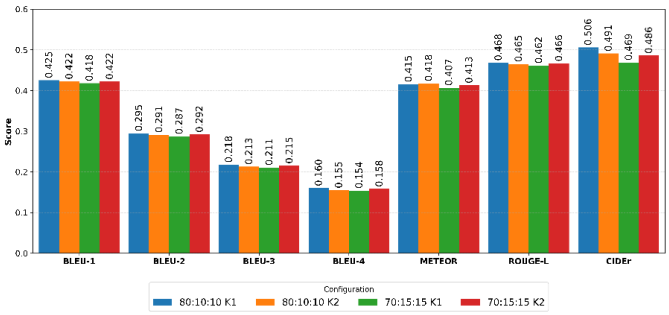

Figure 13. Comparison of image captioning performance across different data split scenarios and training configurations

The evaluation results demonstrate that the InceptionV3–Vanilla Transformer consistently outperformed the baseline InceptionV3–LSTM across all experimental settings, indicating the effectiveness of the Transformer decoder in modeling visual–textual relationships. According to Table V and Figure 13, the best performance was achieved using the 80:10:10 split with Configuration 1. Overall, the 80:10:10 split produced better results than the 70:15:15 split, suggesting that a larger training set contributes to improved caption generation performance.

C. *Caption Generation Results*

This section presents examples of captions generated by the proposed model under different data split scenarios and training configurations. Both Flickr30k test images and unseen images are used to evaluate the quality and generalization capability of the generated captions.

1) *80:10:10 Split – Configuration 1*

Figure 14. Generated caption on a Flickr30k test image (80:10:10 Split – Configuration 1)

Figure 15. Generated caption on an unseen image (80:10:10 Split – Configuration 1)

Based on Figure 14 and Figure 15, the generated captions successfully identified the main objects and activities in both images; however, for Figure 14, the model incorrectly guessed the color, even though it produced a description that was semantically consistent with the visual content.

2) *80:10:10 Split – Configuration 2*

Figure 16. Generated caption on a Flickr30k test image (80:10:10 Split – Configuration 2)

Figure 17. Generated caption on an unseen image (80:10:10 Split – Configuration 2)

Based on Figure 16 and Figure 17, the generated captions successfully describe the main objects and activities. However, some visual attributes are inaccurately identified, particularly in Figure 17, where the model incorrectly adds a book-reading activity that is not present in the image.

3) *70:15:15 Split – Configuration 1*

Figure 18. Generated caption on a Flickr30k test image (70:15:15 Split – Configuration 1)

Figure 19. Generated caption on an unseen image (70:15:15 Split – Configuration 1)

Based on Figure 18 and Figure 19, the generated captions successfully describe the main objects and activities. However, some visual attributes are inaccurately identified, particularly in Figure 19, where the model incorrectly includes reading a newspaper activity that is not present in the image.

4) *70:15:15 Split – Configuration 2*

Figure 20. Generated caption on a Flickr30k test image (70:15:15 Split – Configuration 2)

Figure 21. Generated caption on an unseen image (70:15:15 Split – Configuration 2)

Based on Figures 20 and 21, the generated captions generally provide an appropriate description of the primary objects and activities depicted in the images. Nevertheless, inaccuracies remain in certain visual details, particularly in Figure 21, where the model incorrectly introduces reading a book activity that is not present in the scene.

Overall, the 80:10:10 split with Configuration 1 produced the most accurate and contextually relevant captions, consistent with the quantitative evaluation results presented in Table V. Performance on the Flickr30k test set was assessed using BLEU, METEOR, ROUGE-L, and CIDEr metrics, while generalization ability was further evaluated on 10 unseen images outside the dataset. The model correctly generated relevant captions for 9 of the 10 images, demonstrating its capability to describe previously unseen visual content. These findings indicate that the proposed model can effectively recognize the main objects and activities while producing semantically meaningful captions across different image sources.

V. CONCLUSION

This study proposed a Sequence-to-Sequence image captioning framework that integrates an InceptionV3 encoder with a Vanilla Transformer decoder for automatic image caption generation on the Flickr30k dataset. Experimental results demonstrate that the proposed architecture is capable of generating captions that accurately describe the main objects and activities depicted in images.

The model was evaluated under two data split scenarios (80:10:10 and 70:15:15) and two training configurations using BLEU, METEOR, ROUGE-L, and CIDEr metrics. The best performance was achieved with the 80:10:10 split and Configuration 1, obtaining BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-L, and CIDEr scores of 0.4252, 0.2946, 0.2176, 0.1601, 0.4155, 0.4681, and 0.5059, respectively. Furthermore, qualitative results indicate that the proposed model can generate contextually relevant captions for both Flickr30k test images and unseen images. These findings suggest that the combination of InceptionV3 and a Vanilla Transformer is effective for image captioning tasks on general-purpose image datasets.

ACKNOWLEDGMENT

The author gratefully acknowledges the support and guidance received throughout this research. Special thanks are extended to the author's parents and family for their continuous encouragement and prayers. The author also thanks the supervisor and examiners for their valuable suggestions and feedback, as well as friends and colleagues who contributed support during the completion of this study.

REFERENSI

[1] APJII, "SURVEI PENETRASI INTERNET DAN PERILAKU PENGGUNAAN INTERNET," 2025. Accessed: Sep. 02, 2025. [Online]. Available: https://survei.apjii.or.id/

[2] R. Mokady, A. Hertz, and A. H. Bermano, "ClipCap: CLIP Prefix for Image Captioning," Nov. 2021, [Online]. Available: http://arxiv.org/abs/2111.09734

[3] H. Kristianto *et al.*, "Optimalisasi Deskripsi Produk E-Commerce: Manfaatkan AI ChatGPT Untuk Deskripsi Menarik & Kontekstual," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 11, no. 3, 2024, [Online]. Available: http://jurnal.mdp.ac.id

[4] J.-F. Yeh, K.-M. Lin, and C.-C. Chen, "Image Captioning Using Topic Faster R-CNN-LSTM Networks," *Information*, vol. 16, no. 9, p. 726, Aug. 2025, doi: 10.3390/info16090726.

[5] H. K. Gupta, "When Better Eyes Lead to Blindness: A Diagnostic Study of the Information Bottleneck in CNN-LSTM Image Captioning Models," Aug. 2025, doi: 10.5120/ijca2025925560.

[6] P. Zeng, H. Zhang, J. Song, and L. Gao, "S 2 Transformer for Image Captioning," *IJCAI*, 2022, [Online]. Available: https://github.com/zchoi/S2-

[7] M. A. R. Ridoy, M. M. Hasan, and S. Bhowmick, "Compressed Image Captioning using CNN-based Encoder-Decoder Framework," Apr. 2024, [Online]. Available: http://arxiv.org/abs/2404.18062

[8] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From Show to Tell: A Survey on Deep Learning-based Image Captioning," Nov. 2021, [Online]. Available: http://arxiv.org/abs/2107.06912

[9] P. Dandwate, C. Shahane, V. Jagtap, and S. C. Karande, "Comparative study of Transformer and LSTM Network with attention mechanism on Image Captioning," Mar. 2023, [Online]. Available: http://arxiv.org/abs/2303.02648

[10] J. Jean, V. Sengka, O. A. Lantang, and M. Dwisnanto Putro, "Aplikasi Image Captioning Menggunakan Convolutional Neural Network dan Long Short Term Memory," *Jurnal Teknik Informatika Unika ST. Thomas (JTIUST)*, vol. 10, no. 02, pp. 2657–1501, 2025.

[11] M. Faishal Dzaky and J. Pardede, "CNN dan Transformer pada Image Captioning," 2023.

[12] A. W. Kosman, Y. Wahyuningsih, and F. Mahendrasusila, "Pengujian Metode Inception V3 dalam Mengidentifikasi Penyakit Kanker Kulit," *Jurnal Teknologi Informatika dan Komputer*, vol. 10, no. 1, pp. 136–146, Mar. 2024, doi: 10.37012/jtik.v10i1.1940.

[13] M. Maryamah, N. A. Alya, M. H. Sudibyo, E. Liviani, and R. I. Thirafi, "Image Classification on Fashion Dataset Using Inception V3," *Journal of Advanced Technology and Multidiscipline (JATM)*, vol. 02, no. 01, pp. 10–15, 2023.

[14] M. Cornia, L. Baraldi, and R. Cucchiara, "Explaining transformer-based image captioning models: An empirical analysis," *AI Communications*, vol. 35, no. 2, pp. 111–129, 2022, doi: 10.3233/AIC-210172.

[15] S. Zhang, R. Fan, Y. Liu, S. Chen, Q. Liu, and W. Zeng, "Applications of transformer-based language models in bioinformatics: A survey," 2023, *Oxford University Press*. doi: 10.1093/bioadv/vbad001.

[16] H. D. Abdulgalil and O. A. Basir, "Next-generation image captioning: A survey of methodologies and emerging challenges from transformers to Multimodal Large Language Models," *Natural Language Processing Journal*, vol. 12, p. 100159, Sep. 2025, doi: 10.1016/j.nlp.2025.100159.

[17] J. Khatib Sulaiman, N. Pramesti Aprilia, T. Herlina Rochadiani, and U. Pradita, "Image Captioning untuk Gambar Rambu Lalu Lintas Indonesia Menggunakan Pretrained CNN dan Transformer," *Indonesian Journal of Computer Science*, vol. 13, no. 3, pp. 4627–4640, 2024.

[18] R. Mulyawan, A. Sunyoto, and A. H. Muhammad, "Pre-Trained CNN Architecture Analysis for Transformer-Based Indonesian Image Caption Generation Model," *INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION (JOIV)*, 2023, [Online]. Available: www.joiv.org/index.php/joiv