e-ISSN: 3089-5081 | p-ISSN: 1979-8911
Vol. 14 No. 2 December 2025: 88-94
DOI: 10.15575/istek.v14i2.2599

Transforming Story Ideas from Images to Text Using Convolutional Neural Networks (CNN) and Generative Pre-trained Transformer (GPT-2)

Moh Hasbi Rizqulloh[1], Eva Nurlatifah[2], Wildan Budiawan Zulfikar[3]
[1,2,3]Department of Informatics, UIN Sunan Gunung Djati Bandung, Indonesia

Article Info

Article history:

Keywords:

CNN
GPT-2
Image-to-Text
Story Generation

ABSTRACT (10 PT)

The gap between rich visual inspiration and the challenge of creative articulation (writer's block) remains a major obstacle in the writing process. This study aims to bridge this gap by designing a two-stage artificial intelligence system based on deep learning to provide automated narrative stimuli. The proposed method implements a custom Convolutional Neural Network (CNN) architecture to detect seven classes of natural objects from 4,362 images. The detected objects are then used as prompts for a fine-tuned Generative Pre-trained Transformer (GPT-2) model to generate poetic narratives. Experimental results indicate that the CNN module achieved a peak classification accuracy of 61.96%. Confusion matrix analysis reveals that this limitation is not caused by overfitting, but rather by high inter-class visual ambiguity. Although the GPT-2 module is capable of generating narratives with a BERTScore F1 of up to 0.6455, the primary finding of this study is that the overall narrative quality is highly dependent on the accuracy of the CNN output, which acts as a critical bottleneck in the system.

Corresponding Author:

Eva Nurlatifah
Informatics Department, Faculty of Science & Technology, UIN Sunan Gunung Djati Bandung
Jl. A. H. Nasution No. 105, Cibiru, Bandung, Indonesia. 40614
Email: evanurlatifah@uinsgd.ac.id

1. INTRODUCTION

In the world of writing, visual inspiration derived from images can trigger the imagination to create interesting works [1], [2], [3]. Although the development of artificial intelligence technology has brought significant changes to the creative process, the phenomenon of writer's block or deadlock of ideas is still a common problem that hinders writer's productivity [4], [5], [6], [7]. A study by SJ Ahmed confirmed through a survey of 146 writers that writer's block is a valid and measurable problem that is often triggered by uncertainty in starting or developing a story line [8]. The gap between the rich visual perception of story potential and the difficulty in pouring it into narrative text is a fundamental challenge, so a new approach is needed that can function as a stimulus to overcome initial obstacles in the writing process [9].

Various previous studies have explored the realm of image-to-text transformation and story generation. For example, one study developed a model to generate story endings based on image guidance [10]. Another study focused on the generation of controlled, image-grounded, stylized stories [11]. A similar approach has also been taken by applying deep learning directly to image -based storytelling [12]. The flexibility of this technology is also seen in its application in other domains, where the integration of ViT, Faster R-CNN, and GPT-2 was used to analyze medical brain images and automatically generate textual reports [13]. Meanwhile, another study focused on the aspect of narrative structure by developing a model capable of generating explainable plots to support the story generation process [14].

While these studies have shown significant progress, most tend to be limited to narrow applications or have not optimized specific architecture combinations for creative inspiration. This research fills this gap by proposing a pipeline that specifically integrates Convolutional Neural Networks (CNNs) for visual object detection [15] with Generative Pre-trained Transformer 2 (GPT-2) fine -tuned to produce poetic-style narratives, rather than literal descriptions [16]. The main contribution of this research lies in designing a conceptual “bridge” that transforms the structured output of an object detector into context-rich prompts to trigger the imagination of a language model [17]. Therefore, this study aims to (1) develop a program that implements CNN and GPT-2 techniques to generate narrative text from images, and (2) test the level of accuracy of the generated narrative text in representing the contents of the image.

2. METHOD

This research uses a quantitative approach by adopting a modification of the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, a standard framework in machine learning projects [18]. The system architecture consists of two main algorithms, namely Convolutional Neural Network (CNN), which is a deep learning architecture for processing image data [19], and Generative Pre-trained Transformer 2 (GPT-2), a language model based on the Transformer architecture for text generation [20]. The methodology consists of six main stages.

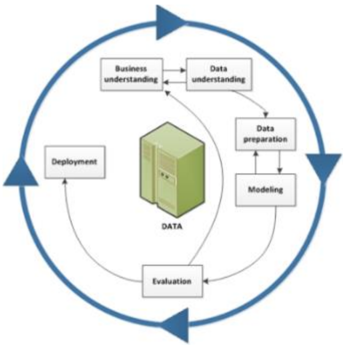

Figure 1. CRISP-DM Diagram [21]

2.1. Business Understanding

This initial phase focused on identifying the fundamental problem to be solved: writer's block, a phenomenon that prevents writers from transforming visual inspiration into narrative. Literature studies, including research by SJ Ahmed, confirmed that idea block is a real challenge often triggered by the difficulty of starting from a blank page (blank page syndrome). In response, the solution designed is a two-stage artificial intelligence system that aims to provide poetic narrative stimuli, rather than literal descriptions. The proposed system architecture consists of a Vision Module using a Convolutional Neural Network (CNN) to detect objects as keywords, and a Language Module using a fine-tuned Generative Pre-trained Transformer (GPT-2) to accept those keywords as prompts and generate narrative text.

2.2. Data Understanding

At this stage, two types of datasets were collected and explored. The first is an image dataset, consisting of 4,362 images of natural scenery collected from the Unsplash platform. This dataset focuses on seven classes of natural objects: mountains, rivers, trees, lakes, skies, seas, and rocks. The second is a text dataset, a custom corpus of 1,660 input-output data pairs built for the GPT-2 fine-tuning process. This dataset was created by combining drafts from generative AI with manual curation and rewriting to ensure the narrative quality is poetic and in line with the desired style.

2.3. Data Preparation

Image data was prepared through manual annotation using Roboflow to assign class labels and bounding boxes, resulting in a total of 18,726 annotations. The images were then processed by resizing them to 224x224 pixels, normalizing their values, and padding them to standardize the number of objects per image to seven. For text data, the preparation process included formatting into a prompt-completion structure and tokenization using the standard GPT-2 tokenizer. The length of each text was standardized to 512 tokens through truncation and padding processes. Finally, both datasets (image and text) were divided into 80% training data and 20% validation data for model training and evaluation purposes.

2.4. Modeling

Two main models were built and trained separately. The CNN model was designed as a custom architecture from scratch, consisting of three convolutional blocks for feature extraction and a head for bounding box prediction and class classification. The model was compiled using the Adam optimizer and trained with callbacks such as EarlyStopping and ReduceLROnPlateau to prevent overfitting . The text generation model uses the pre-trained cahya/gpt2-small-indonesian-522M architecture. The model was then fine -tuned on a custom narrative dataset for 10 epochs using the AdamW optimizer. This process aimed to refine the model's capabilities to generate text with a consistent poetic style.

2.5. Evaluation

Following the modeling process, a performance evaluation will be conducted to measure the effectiveness of each model. The performance of the CNN model will be quantitatively evaluated on classification and detection tasks. Metrics used include Accuracy, Precision, Recall, F1-Score, and in-depth analysis using a Confusion Matrix. For the GPT-2 model, the quality of the generated narrative text will be evaluated using a dual approach. An automated evaluation will be conducted using the BERTScore metric to measure semantic similarity. Additionally, a qualitative manual evaluation will be conducted to assess aspects such as narrative coherence, relevance, and creativity.

2.6. Deployment

As a final step, the trained and evaluated models were integrated into a functional web-based application prototype. The application was built using the Gradio library, chosen for its ability to quickly build interactive user interfaces (UIs) for machine learning models. The application's functionality allows users to upload an image, which is then processed by an end-to-end inference pipeline. The system displays two types of output simultaneously: the original image with a bounding box overlay from CNN detection, and a text box containing a short, relevant narrative generated by GPT-2. This implementation successfully demonstrates the system's functionality in directly transforming ideas from visual to text.

3. RESULT AND DISCUSSION

This section discusses the results obtained from each stage of the methodology used in the study. Each step, from data understanding to final analysis, is explained in detail to provide a comprehensive overview of the performance of the developed system. Data understanding began with the analysis of 4,362 natural landscape images and 1,660 narrative text data pairs. A crucial image data preparation process was manual annotation using the Roboflow platform, where each image was assigned a class label and a precise bounding box. This process resulted in a total of 18,726 labeled objects distributed among seven main classes. The distribution of annotations for each class was kept balanced to avoid bias during training, as presented in Table 1.

Table 1. Distribution of Object Annotations in the Dataset

Class	Number of Annotations
Tree	2,791
Mountain	2,741
Sky	2,787
River	2,539
Rock	2,565
Lake	2,536
Sea	2,767

After annotation, all images were resized to a uniform resolution of 224x224 pixels and normalized. The final preparation step was to divide the dataset proportionally into 80% training data (3,488 samples) and 20% validation data (872 samples) to ensure objective model evaluation.

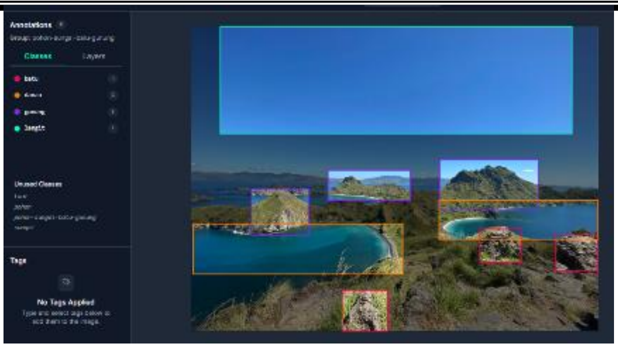

Figure 2. Example of Image Annotation Results on the Roboflow Platform

3.1. Model Architecture and Training

The Object Detection Network (CNN) model has three main convolutional blocks for feature extraction and a head that produces two outputs: bounding box prediction and object class classification. The model is compiled using the Adam optimizer with Mean Squared Error (MSE) loss functions for bounding boxes and Categorical Crossentropy for classification. Training is tightly controlled using callbacks such as EarlyStopping to prevent overfitting.

Table 2. Summary of CNN Model Architecture

Layer (type)	Output Shape	Param#
Input Layer	(None, 224, 224, 3)	0
3x Convolution Block	(None, 28, 28, 256)	-
GlobalAveragePooling2D	(None, 256)	0
Output Head (bbox & class)	(None, 5, 4) & (None, 5, 8)	-
Total Parameters		2,565

For text generation, a pre-trained cahya/gpt2-small-indonesian-522M model adapted through fine-tuning on a custom narrative dataset was used. Training was run for 10 epochs using the AdamW optimizer to specialize the model for generating poetic-style narratives.

3. 2. Model Performance Evaluation Results

The evaluation showed contrasting performance between the two models and their tasks. The training curve for the CNN model (Figure 3) shows a healthy learning process without significant overfitting, with the loss on both the training and validation data decreasing and reaching convergence.

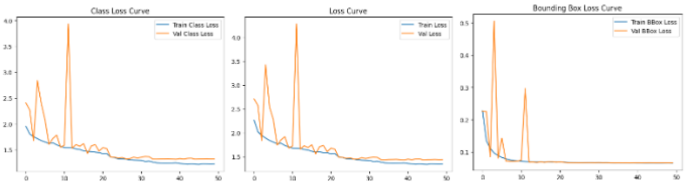

Figure 3. Loss Metric Performance Curve During CNN Model Training

Quantitatively, the CNN model achieved a peak classification accuracy of 61.96%. However, it demonstrated very high reliability in object localization, as evidenced by its very low Bounding Box Loss (MSE) value of 0.0582. Confusion Matrix analysis (Figure 4) revealed that this limited classification performance was not caused by model issues, but rather by high visual ambiguity between classes in the dataset, such as significant confusion between the class 'lake' and 'mountain' (180 misclassification cases) and 'rock' and 'river' (119 misclassification cases).

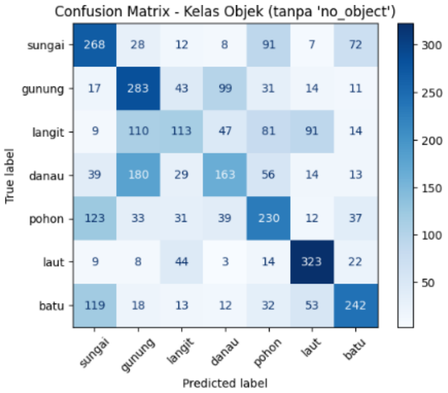

Figure 4. Confusion Matrix for CNN Model Classification Performance

On the other hand, the fine -tuned GPT-2 module was able to generate semantically relevant narratives, with the highest BERTScore F1 score reaching 0.6455. This result indicates that the model successfully captured the meaningful elements of the given prompt and structured them into a coherent sentence.

The following are the results of a comparison of the results of GPT-2 text generation with human-generated text taken from Umar Kayam's short story entitled "A Thousand Fireflies in Manhattan" which can be seen in Table 3.

Table 3. Comparison of GPT-2 Text Generation Results with References

No	GPT-2 Generated Description	References (Literary Citations)	BERTS core (F1)
1	"A knotty pair of tree leaves grips the rock surface on the riverbank..."	"He looked down and a jungle of skyscrapers was sleeping beneath him..."	
2	"Nurse log is a promise of the infinite beauty of the forest..."	"Marno started lighting a cigarette and then went to stand by the window. The sky was clear that night..."	

The key finding of this study is that the overall system performance is heavily influenced by the CNN module in the early stages. While the GPT-2 module can generate high-quality narratives, the relevance and richness of those stories directly depend on the accuracy of the object lists received from the CNN. With a classification accuracy of 61.96%, the generated prompts are often not entirely accurate, making the CNN module a critical bottleneck. This approach, which integrates vision and language models, aligns with previous research that has also successfully transformed image descriptions into visual storytelling [17], but highlights the importance of accuracy in the visual perception module.

This accuracy limitation does not stem from overfitting, but rather from a fundamental challenge in the natural scene dataset itself, namely the high visual ambiguity between classes. This challenge is a common problem in object detection in uncontrolled environments, which has driven the development of more sophisticated architectures such as the YOLO detection model [22]. Therefore, for future development, it is highly recommended to use transfer learning techniques with a more robust backbone. This approach is supported by trends in the literature, where models such as ViT and Faster R-CNN have proven effective for complex visual analysis tasks [13]. Ultimately, the system implemented in the Gradio interface (Figure 5) successfully proves its end-to-end functionality as a creativity-inducing tool.

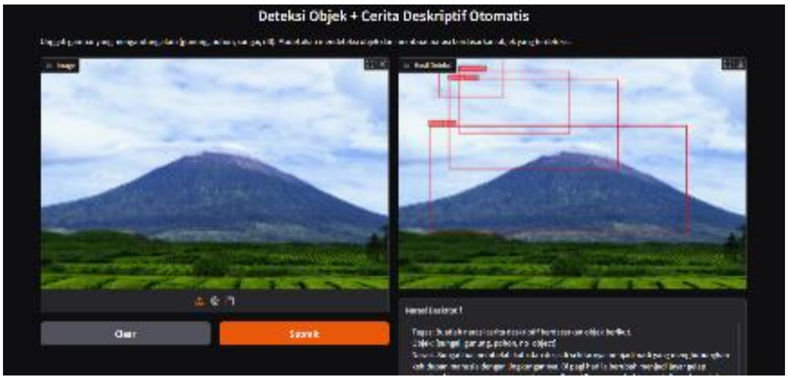

Figure 5. Prototype Application Interface Display

4. CONCLUSION

This research successfully implemented a functional end-to-end system for transforming images into narrative text by integrating a Convolutional Neural Network (CNN) and a Generative Pre-trained Transformer (GPT-2). Key findings indicate that although the system was successfully implemented, its effectiveness was largely determined by the performance of the CNN module in the initial stage. With a peak classification accuracy of 61.96%, the CNN module proved to be a critical bottleneck for the entire system. This limitation was not caused by overfitting, but rather by the fundamental challenge of high visual ambiguity between natural object classes in the dataset used. The main contribution of this research is the design of a conceptual "bridge" that transforms the structured output of the object detector into poetic and interpretive narrative stimuli, rather than mere literal descriptions. Consequently, this system offers an applicable solution as a creativity aid that can be a catalyst for ideas for educators, content creators, or writers facing the challenge of writer's block.

This research demonstrates that the integration of vision and language technologies can be extended from descriptive tasks to aesthetically inspiring tools. Based on the findings and limitations, further research can be directed at several areas. First, implementing transfer learning techniques using pre-trained backbones (such as EfficientNet or ResNet) is highly recommended to significantly improve object detection accuracy. Second, to address visual ambiguity, targeted data collection specifically targeting classes that frequently cause confusion is necessary. Finally, experiments with larger or more sophisticated language models can be explored to increase the complexity and creativity of the generated narratives.

REFERENCES

[1] G. Sakkir, “the Effectiveness of Pictures in Enhance Writing Skill of Senior High School Students,” Interf. J. Lang. Lit. Linguist., vol. 1, no. 1, 2020, doi: 10.26858/interference.v1i1.12803.

[2] Listyani, “The use of a visual image to promote narrative writing ability and creativity,” Eurasian J. Educ. Res., vol. 2019, no. 80, pp. 193–224, 2019, doi: 10.14689/ejer.2019.80.10.

[3] H. M. Romadlona and Z. A. Khofshoh, “The effectiveness of using picture series media on student’s writing narrative text,” Karangan, J. Kependidikan, Pembelajaran, dan Pengemb., vol. 5, no. 1, pp. 30–35, 2023.

[4] Joseph Patrick Pascale, “George R. R. Martin: An Epic Case of Writer’s Block,” https://medium.com/. Accessed: Oct. 30, 2024. [Online]. Available: https://medium.com/@josephpatrickpascale/george-r-r-martin-an-epic-case-of-writers-block-5e6e0535ccff

[5] N. S. Pasaribu, N. Annisa, and S. H. Harahap, “Pengaruh Teknologi Terhadap Gaya Menulis dan Komunikasi,” IJEDR Indones. J. Educ. Dev. Res., vol. 2, no. 1, pp. 315–319, 2024, doi: 10.57235/ijedr.v2i1.1764.

[6] N. Amado, “Psychoanalytic views of ‘writer’s block’: Artistic creation and its discontents,” Int. Forum Psychoanal., vol. 31, no. 2, pp. 100–107, 2022, doi: 10.1080/0803706X.2021.1887518.

[7] J. Rowling, “The Times publishes a new interview with J.K. Rowling about her writing process,” https://www.therowlinglibrary.com/. Accessed: Oct. 30, 2024. [Online]. Available: https://www.therowlinglibrary.com/2024/05/05/the-times-publishes-a-new-interview-with-j-k-rowling-about-her-writing-process/#:~:text=I’ve only ever once,t see my way forward.

[8] S. J. Ahmed, “An analysis of writer’s block: causes, characteristics, and solutions,” University of North Florida, 2019. [Online]. Available: https://digitalcommons.unf.edu/etd

[9] J. E. R. Marantika, “The Contribution Of Visual Literacy And Creative Thinking On Writing Skills,” J. Int. Semin. Lang. …, vol. 1, no. 1, pp. 2017–2020, 2019.

[10] Y. Wang, W. Hu, and R. Hong, “Iterative Adversarial Attack on Image-Guided Story Ending Generation,” IEEE Trans. Multimed., vol. 26, pp. 6117–6130, 2024, doi: 10.1109/TMM.2023.3345167.

[11] H. Lovenia, B. Wilie, R. Barraud, S. Cahyawijaya, W. Chung, and P. Fung, “Every picture tells a story: Image-grounded controllable stylistic story generation,” Proc. - Int. Conf. Comput. Linguist. COLING, vol. 29, no. 3, pp. 40–52, 2022.

[12] Y. Zhu and W. Q. Yan, “Image-Based Storytelling Using Deep Learning,” ACM Int. Conf. Proceeding Ser., pp. 179–186, 2022, doi: 10.1145/3561613.3561641.

[13] V. C. Sai Santhosh, T. Nikhil Eshwar, R. Ponraj, and K. Kiran, “Comprehensive Strategy for Analyzing Dementia Brain Images and Generating Textual Reports through ViT, Faster R-CNN and GPT-2 Integration,” 2023 1st Int. Conf. Adv. Electr. Electron. Comput. Intell. ICAEECI 2023, pp. 1–10, 2023, doi: 10.1109/ICAEECI58247.2023.10370864.

[14] G. Chen, Y. Liu, H. Luan, M. Zhang, Q. Liu, and M. Sun, “Learning to Generate Explainable Plots for Neural Story Generation,” IEEE/ACM Trans. Audio Speech Lang. Process., vol. 29, pp. 585–593, 2021, doi: 10.1109/TASLP.2020.3039606.

[15] J. A. Cahyono and J. N. Jusuf, “Automated Image Captioning with CNNs and Transformers,” pp. 1–13, 2024, [Online]. Available: http://arxiv.org/abs/2412.10511

[16] J. Li, D. M. Vo, A. Sugimoto, and H. Nakayama, “EVC AP : Retrieval-Augmented Image Captioning with External Visual – Name Memory for Open-World Comprehension,” pp. 13733–13742.

[17] R. Patankar, H. Sethi, A. Sadhukha, N. Banjade, and A. Mathur, “Image Captioning with Audio Reinforcement using RNN and CNN,” Int. Conf. Sustain. Comput. Smart Syst. ICSCSS 2023 - Proc., vol. 2, no. Icscss, pp. 591–596, 2023, doi: 10.1109/ICSCSS57650.2023.10169692.

[18] M. Bautista, S. Alfaro, and L. Wong, “Framework for the Adaptive Learning of Higher Education Students in Virtual Classes in Peru Using CRISP-DM and Machine Learning,” J. Comput. Sci., vol. 20, no. 5, pp. 522–534, 2024, doi: 10.3844/jcsp.2024.522.534.

[19] M. Bhalekar and M. Bedekar, “D-CNN: A New model for Generating Image Captions with Text Extraction Using Deep Learning for Visually Challenged Individuals,” Eng. Technol. Appl. Sci. Res., vol. 12, no. 2, pp. 8366–8373, 2022, doi: 10.48084/etasr.4772.

[20] A. Rahali and M. A. Akhloufi, “End-to-End Transformer-Based Models in Textual-Based NLP,” AI, vol. 4, no. 1, pp. 54–110, 2023, doi: 10.3390/ai4010004.

[21] IBM, “IBM SPSS Modeler CRISP-DM Guide,” https://www.ibm.com/docs/it/SS3RA7_18.3.0/pdf/ModelerCRISPDM.pdf, 2021.

[22] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, “YOLO-World : Real-Time Open-Vocabulary Object Detection,” pp. 16901–16911.