**Computing and Information Processing Letters**
Vol. 4, No. 1, Desember 2024, pp. 01–08
http://jurnal.upnyk.ac.id/index.php/cip/index

# Image Captioning With Attention Mechanism Using Convolutional Neural Network and Gated Recurrent Unit on the Picture of Tourism in the Special Region of Yogyakarta

Basrizal Reza Astana[a,1], Dessyanto Boedi Prasetyo[b,2], Wilis Kaswidjanti[b,3]
[a,b,c] Informatika, Universitas Pembangunan Nasional Veteran Yogyakarta, Indonesia
[1]123170089@student.upnyk.ac.id; [2]dess@upnyk.ac.id*; [3]wilisk@upnyk.ac.id*
*Corresponding Author

**ARTICLE INFO**

**Article history**
Received: 15/12/2024
Revised: 23/12/2024
Accepted: 24/12/2024

**Keywords**
image captioning
attention mechanism
CNN
GRU

**ABSTRACT**

This study's purpose is to apply Convolutional Neural Network (CNN) using ResNet-50 feature extraction and GRU also using Attention Mechanism to generate image captions automatically with dataset of tourism pictures in Special Region of Yogyakarta using Indonesian language and to know the accuracy. Design/methodology/approach: Using CNN method with ResNet-50 and GRU and attention mechanism. Findings/result: The results of the test carried out using the BLEU score got the BLEU-1,2,3,4 score 42.64, 32.56, 18.55, 16.27. Originality/value/state of the art: This study uses datasets taken from 10 tourist attractions in Yogyakarta, and also in this study using the CNN method with feature extraction ResNet-50, GRU and using the Attention Mechanism.


## 1 Introduction

Yogyakarta is known as an educational center, a cultural city, and also a center of culture. However, it is undeniable that Yogyakarta is a tourist destination known for its wealth of natural charm. Yogyakarta is a tourist destination for both domestic and foreign tourists.

Many tourists capture their travel moments in photos, which are then uploaded to their personal social media accounts. However, a single image can contain a lot of information, and each individual can produce different information or descriptions for each image. Humans have limitations in parsing large amounts of information from an image without a description or caption [1].

Originating from this problem, and along with technological advancements today, the development of automatic image description by combining two artificial intelligence methods, Computer Vision and Natural Language Processing, has begun to flourish [2].

In previous research on image captioning using CNN, researchers compared several feature extraction methods, including ResNet-50, with others. It was found that ResNet-50 yielded the best results and had the highest BLEU score in generating image features [3, 4].

This research will use a combined method of CNN and GRU with an attention mechanism to perform image captioning and obtain automatic Indonesian-language image descriptions. The CNN uses feature extraction from ResNet50, chosen because previous research found that ResNet50 feature extraction achieved good accuracy compared to other feature extraction methods in image


**Figure 1:** Example of image captioning

processing. GRU is chosen because it is a development from LSTM where GRU has fewer weights and parameters to train compared to LSTM, making model training computation faster than LSTM, and GRU is a simplified model of LSTM [5]. LSTM has three gates: forget gate, input gate, and output gate, while GRU simplifies this to two gates where the forget gate and input gate are combined into an update gate, and the second gate of GRU is the reset gate. It is also chosen because it can handle the vanishing gradient problem [6]. The attention mechanism is chosen because it can unify important information from a sentence, focusing the GRU model's output on the target word so that the generated words are better and more appropriate. The attention mechanism is also added to handle long input sequence problems, thereby increasing accuracy [7].

# 2 Research Method

This research method falls under quantitative and implementative design research. The result of this research is a system that can be used by users to find out the description of a tourism image in the Special Region of Yogyakarta, which will be displayed in a web-based form. The dataset of images and image descriptions is then processed using the Convolutional Neural Network (CNN) algorithm with ResNet-50 feature extraction, Gated Recurrent Unit (GRU), and Attention Mechanism.

## 2.1 Data Collection

Data collection is the process of gathering the information and data needed to be processed later by the image captioning generation system. The data to be used consists of tourism place images in the form of secondary image data, where this data collection is done by taking images online from social media or the internet. The tourism locations taken are those in the Special Region of Yogyakarta, totaling 10 locations. Each location has 50 images taken, so the total number of images is 500.

## 2.2 Data Labeling

Data labeling is done manually by giving a label or description to each image. Each image will be given three different descriptions, so the total number of image descriptions used in this research is 1,500.

## 2.3 Image Captioning

Image captioning is the process of creating a textual description to explain an image. To generate an automatic description for an image, a combination of Computer Vision and Natural Language Processing processes is required [8]. Computer Vision functions to model a machine so it can detect and recognize objects in an image, and NLP enables computers to read text, interpret text, and determine important parts of text to produce a well-structured sentence [4].

## 2.4 Encoder–Decoder

The encoder-decoder architecture is used in this image captioning with attention mechanism, where the encoder contains a pre-trained Convolutional Neural Network to encode or extract the image into a feature vector. Then, the decoder is used to create a descriptive sentence, and the words of the

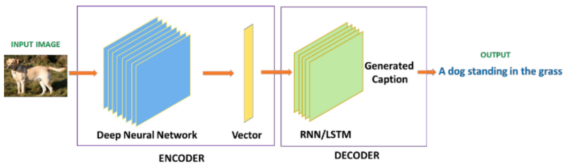


**Figure 2:** Encoder-Decoder Architecture

sentence use the Gated Recurrent Unit model as the decoder. The encoder-decoder architecture can be seen in Figure **??** [8].

## 2.5 Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is an artificial neural network developed from Multilayer Perceptron (MLP), which is a type of feed-forward network. MLP itself is a further development of Artificial Neural Network (ANN). CNN has been successfully used in video and for image classification or object detection [9]. The key stages of CNN used in this research are described as follows.

### 2.5.1 Input Layer

The input layer is the first layer in the CNN algorithm because this layer contains image data. The input layer is represented by 3 dimensions containing length, width, and height. This is usually written in the form width$\times$height$\times$depth in pixels. For image data input, it is displayed in matrix form. An example of an input image is $128 \times 128 \times 3$, where $128 \times 128$ is the width and height, and 3 is the depth representing the RGB channel [10].

### 2.5.2 Convolutional Layer

The convolutional layer consists of neurons forming a filter with a specific height, width, and thickness. This convolutional layer utilizes filters. The main goal of this layer is to identify the main features of the image inputted in the first layer [11]. The convolution equation is:

$$H_{m,n} = \sum_{a=-\infty}^{\infty} \sum_{b=-\infty}^{\infty} I_{m-a,n-b} \cdot K_{a,b} \tag{2.1}$$

where $H$ is the feature map, $I$ is the input image, and $K$ is the filter.

### 2.5.3 Rectified Linear Unit (ReLu)

CNN uses an activation function between layers, namely the convolutional layer and the pooling layer. The activation used is the ReLu or Rectified Linear Unit activation function [12]:

$$f(x) = \max(0, x) \tag{2.2}$$

### 2.5.4 Batch Normalization

Batch normalization is a stage performed to improve the computation process during training and increase the learning rate in the model. Additionally, batch normalization can reduce the gradient on initialization values, allowing for higher learning rates without the risk of divergence [13]. The batch

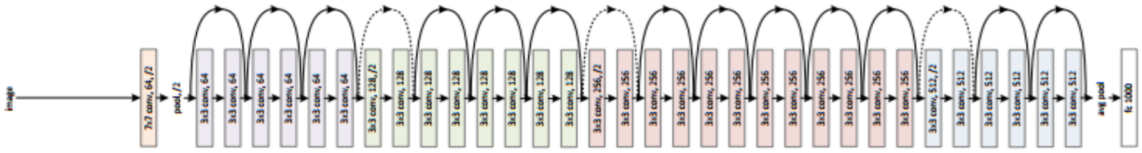

**Figure 3:** ResNet Network

normalization calculation by Ioffe & Szegedy [**?**] is formulated as:

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{2.3}$$

$$\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2 \tag{2.4}$$

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} \tag{2.5}$$

where $\mu_B$ is the batch mean value, $m$ is the number of training datasets, $\varepsilon = 10^{-8}$, and $\gamma$ and $\beta$ are parameters during training (commonly initialized with $\gamma = 1$ and $\beta = 0$).

### 2.5.5 Pooling Layer

The pooling layer is an architecture in CNN located between the convolution layer and the ReLu activation function, with the main function of reducing resolution or reducing the dimensions of each feature map of an image to obtain a broad overview while still retaining important information from the feature map [**?**]. The pooling layer output calculation is:

$$O = \frac{I - K}{S} + 1 \tag{2.6}$$

where $O$ is the pooling layer output, $I$ is the input matrix, $K$ is the kernel matrix, and $S$ is the number of strides.

### 2.5.6 Fully Connected Layer

The fully connected layer is a feed-forward type artificial neural network. The fully connected layer is divided into 3 parts: input layer, hidden layer, and output layer. It is called a fully-connected layer because neurons in each layer are fully connected to neurons before and after them [**?**]. The equation for the fully connected layer is:

$$z_i = \sum_{j=1}^{c} w_{i,j}^T x_i + b_j \tag{2.7}$$

where $z_i$ is the output value, $x_i$ is the input value from extraction result, $w_{i,j}$ is the weight, and $b_j$ is the bias.

## 2.6 Deep Residual Network (ResNet)

ResNet is a type of deep network based on residual learning and also an architecture of Convolutional Neural Network. Residual Network is a residual network with the deepest network, consisting of 152 layers. ResNet-50 itself is a variant of Residual Network with 50 layers [**?**].

## 2.7 Gated Recurrent Unit (GRU)

Gated Recurrent Unit (GRU) was first introduced by Cho et al. in 2014 [**?**]. GRU has a computationally simpler architecture compared to LSTM but has similar or even quite effective accuracy compared to

LSTM and the ability to handle vanishing gradient like LSTM. This is because GRU does not store information using a cell state like LSTM; instead, GRU uses the hidden state to store information [**?**].

The equations and processes performed in GRU are as follows:

$$r_t = \sigma(W_{rh}h_{t-1} + W_{rx}x_t + b_r) \tag{2.8}$$

$$z_t = \sigma(W_{zh}h_{t-1} + W_{zx}x_t + b_z) \tag{2.9}$$

$$\tilde{h}_t = \tanh\big(W_{\tilde{h}h}(r_t \circ h_{t-1}) + W_{\tilde{h}x}x_t + b_{\tilde{h}}\big) \tag{2.10}$$

$$h_t = (1 - z_t) \circ h_{t-1} + z_t \circ \tilde{h}_t \tag{2.11}$$

where $\circ$ denotes element-wise multiplication.

The next process is the fully connected process:

$$Y = Wx + b \tag{2.12}$$

## 2.8 Attention Mechanism

As in neural networks, which simulate human brain actions in a simplified way, this attention mechanism is applied based on human intuition to selectively concentrate on the most relevant thing. It can be described as a mapping from a query and a set of key-value pairs to an output. The attention mechanism is used after the LSTM layer. This mechanism is used to improve the quality of the generated text by selecting parts of the original texts [**?**].

## 2.9 Bilingual Evaluation Understudy (BLEU)

BLEU (Bilingual Evaluation Understudy) is an algorithm that functions to evaluate the quality of a model that produces machine translations from one natural language to another. The closer the machine translation result, the better the value produced. The BLEU score is obtained by measuring the average value of the modified $n$-gram precision score between the automatic machine translation result and the actual translation (real caption) using a constant called brevity penalty in BLEU [**?**].

The BLEU score equations are as follows:

$$BP = \begin{cases} 1, & c > r \\ \exp(1 - r/c), & c \leq r \end{cases} \tag{2.13}$$

$$P_n = \frac{\sum_{C \in corpus} \sum_{n\text{-gram} \in C} count_{clip}(n\text{-gram})}{\sum_{C \in corpus} \sum_{n\text{-gram} \in C} count(n\text{-gram})} \tag{2.14}$$

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2.15}$$

According to Google Cloud, Figure **??** can serve as a guide for interpreting the BLEU score expressed in percent.

## 2.10 Gated Recurrent Unit (GRU)

# 3 Results and Discussion

## 3.1 Testing Results

The BLEU score results performed on the testing dataset for each image are shown in Table 1 below with predetermined weight calculations. A comparison of the original caption and the machine's predicted caption is made against the images in the testing data.

Next, model evaluation uses the BLEU score. The BLEU score is a metric for evaluating generated sentences (predicted) against existing reference sentences. The BLEU score compares the machine's caption results with reference sentences in several $n$-grams as explained in the previous chapter

| BLEU Score | Interpretation |
|---|---|
| < 10 | Almost useless |
| 10 - 19 | Hard to get the gist |
| 20 - 29 | The gist is clear, but has significant grammatical errors |
| 30 - 40 | Understandable to good translations |
| 40 - 50 | High quality translations |
| 50 - 60 | Very high quality, adequate, and fluent translations |
| > 60 | Quality often better than human |

The following color gradient can be used as a general scale interpretation of the BLEU score:

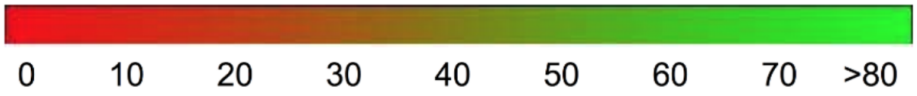


**Figure 4:** Penafsiran BLEU score (https://cloud.google.com/translate/automl/docs/evaluate)

**Table 1:** BLEU Score for Generated Captions

| Image (Source) | Caption | Prediction | BLEU Score |
|---|---|---|---|
|  https://visitingjogja.jogjaprov.go.id/... | Pengunjung sedang berfoto di puncak mangunan (visitors are taking photos at Puncak Mangunan) | Seorang pengunjung berfoto di puncak mangunan (a visitor is taking photos at Puncak Mangunan) | BLEU-1: 83.3<br>BLEU-2: 70.7<br>BLEU-3: 63.2<br>BLEU-4: 53.7 |
|  https://instagram.com/... | Seorang pengunjung dengan tas nya terlihat berada di gapura makam raja mataram kotagede (a visitor with their bag is seen at the gate of the Mataram King's Tomb, Kotagede) | Seorang pengunjung berfoto di makam raja mataram kotagede (a visitor is taking photos at the Mataram King's Tomb, Kotagede) | BLEU-1: 46.8<br>BLEU-2: 37.8<br>BLEU-3: 29.6<br>BLEU-4: 22.8 |

in the form of BLEU-1, BLEU-2, BLEU-3, and BLEU-4. The BLEU score produces a score in the range of 0–1 but is formatted as a percentage by multiplying by 100. This evaluation uses the Natural Language Toolkit Library (NLTK) to facilitate sentence comparison. Below is the calculation of the average BLEU score for BLEU-1, BLEU-2, BLEU-3, and BLEU-4 with predetermined weights for each $n$-gram.

## 3.2 Discussion

Based on the model testing results, the model built using Convolutional Neural Network (CNN) with ResNet-50 feature extraction, Gated Recurrent Unit (GRU), and Attention Mechanism produces words that are quite meaningful in sentences, although there are still grammatical errors in several images where the grammar is less precise. However, in some incorrect descriptions, the predicted words are still accurate in determining the location of the tourist spot. The amount of data influences performance in generating image captions. The more varied and larger the dataset, the better the expected performance. From the performance testing results of the Convolutional Neural Network (CNN) model with ResNet-50 feature extraction, Gated Recurrent Unit (GRU), and Attention Mechanism that was built, the scores for BLEU-1, BLEU-2, BLEU-3, and BLEU-4 are {42.64, 32.56, 18.55, 16.27}.

**Table 2:** BLEU Score Evaluation

| BLEU-N | BLEU Validation Score |
|---|---|
| BLEU-1 | 42.64% |
| BLEU-2 | 32.56% |
| BLEU-3 | 18.55% |
| BLEU-4 | 16.27% |

# 4 Conclusion and Suggestions

## 4.1 Conclusion

Based on the research conducted, the results obtained lead to the conclusion that with the determined model using the Convolutional Neural Network method with ResNet-50 feature extraction and Gated Recurrent Unit through data processing, analysis, design, and discussion stages. With a self-made dataset of images and image descriptions totaling 500 images and 1500 image descriptions, the evaluation results using the BLEU score on the model built with CNN ResNet-50 feature extraction and GRU yielded scores for BLEU-1, BLEU-2, BLEU-3, and BLEU-4 of {42.64, 32.56, 18.55, 16.27}. Based on Google Cloud's interpretation, BLEU-1 produces high quality in generating captions. Then for BLEU-2, image descriptions that are understandable are obtained. For BLEU-3 and BLEU-4, the image descriptions produced are still difficult to understand the essence of the sentence. Although it has relatively small BLEU scores, the model is able to produce Indonesian-language image descriptions that match the image objects, although sentences that are less precise are still found. Apart from using the BLEU score, the results of experiments conducted by taking images not in the dataset show that the model is able to generate image descriptions with correct and understandable captions from the image description results.

## 4.2 Suggestions

There are suggestions that can be made with the hope that the research can be developed by future researchers.

1. The data used in this research is only 500 images, divided into 10 predetermined tourist spots. Thus, each tourist spot only has 50 images. Future research can add datasets from other sources so that the model can learn more varied images and is expected to yield better performance.
2. In this research, there are three image descriptions for each image. Future research can add more variations of image descriptions so that the generated captions are more varied.
3. Future work can compare with the addition of other methods besides ResNet-50 or GRU to determine other more optimal methods for improving accuracy.
4. Implement an automation process in model learning so that the model is always updated when it receives new input from users.

# Acknowledgment

# References


[1] P. Shah, V. Bakarola, and S. Pati, "Image captioning using deep neural architectures," arXiv, 2018.

[2] Q. You et al., "6 (sementic objects) Image Captioning with Semantic Attention ," Rbi, no. ref 1053, 2016, [Online]. Available: https://rbi.org.in/scripts/NotificationUser.aspx?Mode=0Id=254.

[3] A. Arnav, H. Jang, and P. Maloo, "Image Captioning Using Deep Learning," pp. 381–395, 2018, doi: 10.4018/978-1-7998-6870-5.ch026.

[4] I. Artyani, "Simulasi Metode Convolutional Neural Network Dan Long Short-Term Memory Untuk Generate Image Captioning Pada Gambar Lalu Lintas Kendaraan Berbahasa Indonesia," , vol. 8, no. 5, p. 55, 2019.

[5] J. Struye and S. Latré, “Hierarchical temporal memory and recurrent neural networks for time series prediction: An empirical validation and reduction to multilayer perceptrons,” Neurocomputing, vol. 396, no. xxxx, pp. 291–301, 2020, doi: 10.1016/j.neucom.2018.09.098.

[6] S. Hochreiter, “Long Short-Term Memory,” vol. 1780, pp. 1735–1780, 1997.

[7] X. Zhu et al., “Attention-based recurrent neural network for influenza epidemic prediction,” BMC Bioinformatics, vol. 20, no. Suppl 18, pp. 1–10, 2019, doi: 10.1186/s12859-019-3131-8.

[8] R. Khan, M. S. Islam, K. Kanwal, M. Iqbal, M. I. Hossain, and Z. Ye, “A Deep Neural Framework for Image Caption Generation Using GRU-Based Attention Mechanism,” no. i, 2022, [Online]. Available: http://arxiv.org/abs/2203.01594.

[9] M. A. Pangestu and H. Bunyamin, “Analisis Performa dan Pengembangan Sistem Deteksi Ras Anjing pada Gambar dengan Menggunakan Pre-Trained CNN Model,” J. Tek. Inform. dan Sist. Inf., vol. 4, pp. 337–344, 2018.

[10] A. Ajit, K. Acharya, and A. Samanta, “A Review of Convolutional Neural Networks,” pp. 1–5, 2020, doi: 10.1109/ic-ETITE47903.2020.049.

[11] P. Wulandari, “Klasifikasi Tingkat Keganasan Kanker Serviks Menggunakan Metode Deep Residual Network,” p. 125, 2019.

[12] N. A. Shafirra and I. Irhamah, “Klasifikasi Sentimen Ulasan Film Indonesia dengan Konversi Speech-to-Text (STT) Menggunakan Metode Convolutional Neural Network (CNN),” J. Sains dan Seni ITS, vol. 9, no. 1, 2020, doi: 10.12962/j23373520.v9i1.51825.

[13] M. Liu, W. Wu, Z. Gu, Z. Yu, F. F. Qi, and Y. Li, “Deep learning based on Batch Normalization for P300 signal detection,” Neurocomputing, vol. 275, pp. 288–297, 2018, doi: 10.1016/j.neucom.2017.08.039.