

Model Caption Generator Using Visual Geometry, Residual, and Inception Architecture

Agus Nursikuwagus
School of Electrical Engineering and
Informatics,
Institut Teknologi Bandung
agus.nursikuwagus@students.itb.ac.id

Rinaldi Munir
School of Electrical Engineering and
Informatics,
Institut Teknologi Bandung
rinaldi@informatika.org

Masayu Leylia Khodra
School of Electrical Engineering and
Informatics,
Institut Teknologi Bandung
masayu@stei.itb.ac.id

Abstract—Image extraction is to be an essential task upon image classification. One of the challenges of these topics is an improves the extraction model. After that, they combined it with a recurrent neural network to generate a word fitting to an image's area. Interpretation of geology images takes a long time. It needs many geologists, especially in the description of the content of image rocks. Based on these problems, this study proposed a model that can be conducted for the geologist tasks. It enabled to make a caption for an image of geology rocks. The study uses VGG16, ResNet, and InceptionV3 concatenate to LSTM and word2vec that successfully captioned images of the foreground object like cars, people, animals, and many others. Even though the model can extract the image, the outcome does not align with the research objective. The study confirmed that the outcome has a value BLEU score of 4-gram of 0.367, 0.344, and 0.273, respectively. The study outcomes still mistake identifying objects of background and do not correctly caption relate to rock contents. The study concluded that the proposed new model is to be open challenges to achieve a result precisely to geologist descriptions.

Keywords—Geological Rocks Image, Visual Geometry, Residual, InceptionV3, Convolutional Neural Network, Recurrent Neural Network (keywords)

I. INTRODUCTION

Based on the activity of the geologists who annotate the name of rock in the field, it is to be a fundament to engine the intelligence system within caption modeling. One interesting mind is how to create a system that can give output similar to a geologist's captions about the rocks image.

Leveraging two disciplines task between computer vision and natural language processing, it is a challenge to create a captioning model for the geology of rocks. Identification of rocks is to be a primary task in the models. Following the step for captioning model, the task must be detecting an object an generate a sentence similar to geologist annotation [1]–[3].

The development of the image captioning task is inseparable from the deep network architecture. The most famous architecture for image captioning is to follow the encoder-decoder proposed by [4]. This image captioning task has separated into several tasks such as encoder using convolution neural network (CNN) to determine vector features, and decoder using recurrent neural network (RNN) to generate a caption [4].

There are three approaches to producing image captions [5]. The first approach uses a template where this method emphasizes object detection and mapping the results to the language structure [6]. The second approach is to use a modeling language. This method is more expressive and can

overcome the shortcomings of the template method. This method is known as deep learning architecture using LSTM based on RNN [4]. The third approach is to take a caption from the training data with the object's proximity in the image used as a template, then use a language model for the caption on the test data image [7].

Image captioning is used by various researchers, mostly based on deep network architecture. Many use the convolutional neural network (CNN) architecture in the encoder architecture and recurrent deep network architecture in the architecture of automatic text generators. Several image extraction and detection architectures have been developed such as ImageNet [8]–[11], YOLO [12], GoogleNet [13], VGGNet [11], Resnet [9], and InceptionV3 [14]. This model is often used as a baseline model to develop new architectures with various kinds of captioning, such as based on visual attention and semantic attention.

A caption is an essential target of image captioning. The two tasks that needed to process the image and the text such as image extraction and word embedding. These tasks support the learning process in captioning that involve two models such as computer vision and natural language processing [15]. It can be observed that these algorithms evolved the structure based on CNN and RNN. Variations of the CNN layer and convolution matrices such as 1x1, 3x3, 5x5, and 7x7 strongly influence the image extraction results. Using the ReLU function (reactivation linear unit) and the SoftMax function as a determinant of the probability of each word becomes the power of generating expert-appropriate captions. Some results acquired the number of timesteps based on the number of words in the caption strongly influences the prediction of words that appear [16].

Mainly contribution to this study can present as follows:

- The objects concerned lie in the background. The separated object between foreground and background is the primary process for captioning.
- CNN based on visual geometry, residual, and inceptionV3 had different outcomes and evaluations. Applying the model to the image of geological rocks impacts geology caption, however, still some mistakes in identifying the rock image content.
- CNN architecture in the form of VGG16, ResNet50, and InceptionV3 combined with a LSTM model to produce a word, Argmax log-likelihood for $P(I, S|\theta)$, where I is the pixel arrangement (x,y) and $S = \{s_1, s_2, \dots, s_3\}$ is the

This research was funded by Ministry of Research, Technology and Higher Education, Republic of Indonesia, grant number 083/E5/PG.02.00.PT/2022.

word in the image that has a strong influence in producing captions with the same language arrangement.

- Construction of the model consisting of CNN, channels, filter, pooling, ReLU, and SoftMax function has a different outcome, mainly time execution, train parameters accurate, and predicted word.
- Beam Search $K=3$ has structured a sequence of word from log-likelihood function that align with Indonesian language structures.

II. DATA COLLECTIONS

A. Image Acquisitions

The geological image dataset used results from field investigations regarding geological rocks. The image was acquired from a geologist exclusively and is not shareable. In the experiments, the dataset acquired was 121 images and 675 captions from geologists [17]. This study divides the image into two datasets such as dataset training and dataset testing. Dataset training is 101 images and dataset testing is 20 images.

B. Data Pre-processing

Following the pipeline in Fig. 2, the study starts with data pre-processing. The process evolves tasks such as resizing, cropping, and reformatting the color. The image originally has a different size and resizes into a uniform size of 224x224 pixels. Each image has a parameter such as width, length, and channel. The color format was originally in red, green, and blue (RGB). At Fig. 1 is an example geological image rocks and its color histogram,

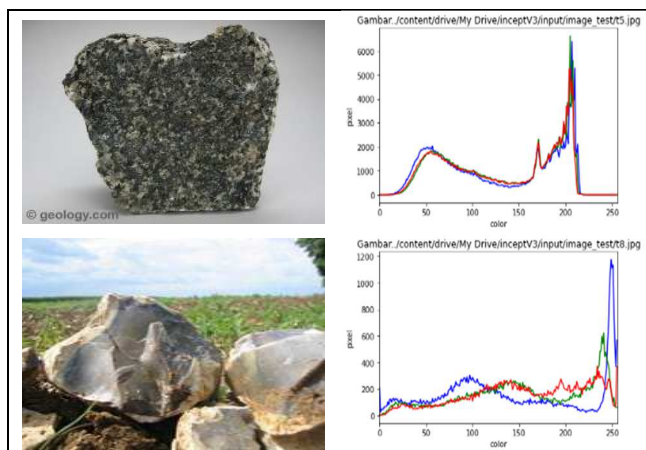


Fig. 1. Geological Rocks Image and Histogram.

III. METHODS

A. Proposed Methods

Several deep learning architectures for image extraction have been proposed, such as CitraNet [9]–[11], YOLO [12], GoogleNet [18], and VGGNet [11]. This model is often used as a baseline model to develop new architectures with various kinds of captioning, such as visual attention-based.

Fig. 2 is a pipeline for image extraction and word extraction. At the last layer of VGG16 is pruned to be fully connected (FC) and remove the classification layer. The process just needs the FC layer that produces 4096 units, and

merge into 256 units output using the ReLU activation function. In carrying on CNN, the output concatenates with 256 units LSTM output. Operation process used element-wise matrix operation between CNN dense unit and LSTM unit. After concatenation, the process going to merge using ReLU activation. This operation obtains word predictions that appear by relying on the input vector image descriptors and LSTM vector values.

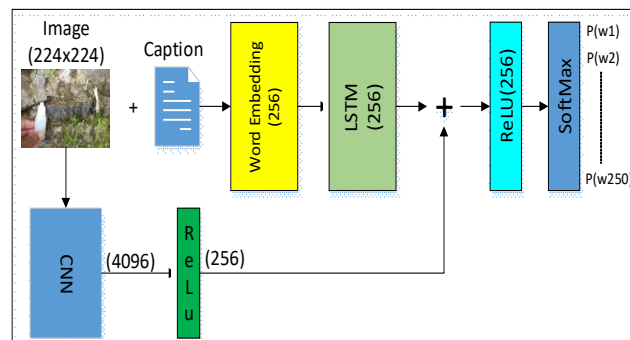


Fig. 2. Image captioning pipeline process.

Concatenation results will be processed by ReLU and SoftMax function. The SoftMax function is a function that gives the probability value of each word class of 250 words. The result of this SoftMax function will choose the highest probability value as the word it generates.

The algorithm for image captioning can be written as presented by [19] :

Image captioning algorithm:

1. The setting of CNN, by omitting the last layer and the SoftMax layer, thus getting 4096-dimensional vectors that describe the global content of the image.
2. Transfer the weights of the prepared VGG16 on IMAGENET.
3. Generate Word:
 - a. Mapping each object into word by Recurrent Process.
 - b. Compute probabilistic distribution by $P(S|I)$, where S is sequence of word $\langle w_1, w_2, \dots, w_n \rangle$ at the image region of I .
 - c. Maximize posterior at the dataset by probabilistic distribution:

$$P(w_1, w_2, \dots, w_{|S|} | I) = \prod_{t=1}^{|S|} P(w_t | I, w_{1:t-1})$$
 Assumptions: generating a word depends on image area I and previous $w_{1:t-1}$.
 - d. Iterate probabilistic distribution for LSTM.
 - e. When time = t , LSTM (h_{t-1}, c_{t-1}), using formula (7), (8), (9), (10), (11), and (12).

4. Maximize *log-likelihood* by loss function =

$$L = \sum_{I, S \in X} \log P(S|I; \theta) = \sum_{I, S \in X} \sum_{t=1}^{|S|} \log P(w_t | w_{1:t-1}, I; \theta)$$

B. CNN Architectures

1) Visual Geometric Group 16 (VGG16)

VGG (Visual Geometric Group) Fig.3 has a minimal convolution architecture with a size of 3x3 or 1x1 convolution layer [20]. For max-pooling size 2x2 with the

exact dimensions on each layer max pooling. VGG implemented with a layer count of 16 or 19 [11].

Karpathy uses the top 19 objects detected from the observed image and calculates the representation of each box that is the top of the object using the equation $v = W_m[CNN_{\theta_c}(Ib)] + b_m$, where $CNN_{\theta_c}(Ib)$ is a transform function of the pixels in the Ib box become a layer with 4096 dimensions. $_c$ is a CNN parameter with 60 million parameters. W_m is a matrix with a size $h * 4096$, and h is a value ranging from 1000 to 1600. Furthermore, the image that a box has bounded is represented as $\{v_i | i = 1..20\}$ [3].

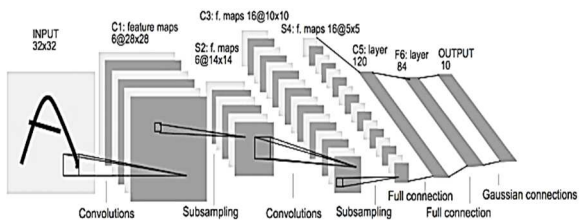


Fig. 3. VGG16 architectures [20].

2) ResNet

In Fig. 4, the residual mapping is formally a nonlinear fit layer of equation $F(x) = H(x) - x$. $H(x)$ is a residual mapping when it is reversed to the original mapping $H(x) = F(x) + x$. Optimization is carried out on x as a residual whose value is pushed towards zero compared to fitting the value of x on a nonlinear layer stack. The value of the function $F(x) + x$ is a shortcut connection concept. This concept allows the process to jump on one layer at a time.

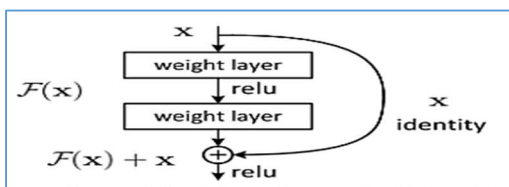


Fig. 4. Residual Learning.

$$y = F(x, \{W_i\}) + x \quad (1)$$

Symbol of y and x is an output and input respectively, corresponding to residual mapping $F(x, \{W_i\})$. Fig. 4 has two layers weight that containing of $F = W_2\sigma(W_1x)$. σ as ReLU function and bias. An $F + x$ operation constructs by *shortcut connection* and *element-wise* summation. The process adopts a non-linearity summation [9]. Experiment about residual learning can be constructed two or more layers from F residual function [9]. If F is a single layer, then the function will be a linear function $y = W_1x + x$. $F(x, \{W_i\})$ shows a convolutional layer with element-wise operation at the feature map for each channel.

3) InceptionV3

Fig. 5 shows the inceptionV3 architecture uses a 1x1 convolution to eliminate the boundaries of computation, ignore information reduction, and maintain the model's performance in terms of accuracy and loss. The technique with the CNN region model from [9] can solve trials such as using color and texture from object locations and employing

CNN to identify objects' categories at that location. Detection of objects based on this area was developed with multi-box predictions for recall of bounding-box objects, then employing this to enhance the categorization process [18].

C. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is a recurrent module that allows long-term learning. The LSTM unit has an additional hidden state as a nonlinear mechanism that allows a state to propagate back without any modification, change, or reset. Learning in LSTM uses simple function gates that have the ability to learn speech recognition and language translation models.

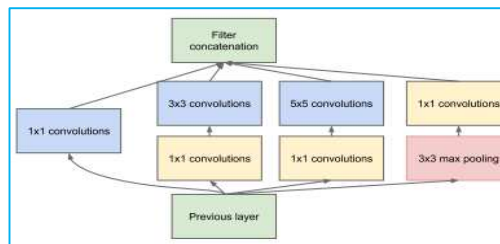


Fig. 5. InceptionV3 architectures [18].

RNN (LSTM) is a common generator caption and many use by scholars. RNN language has a task to generate word that build from three layers, and compute value by following the formula at bellows:

$$x(t) = [w(t) \cdot r(t-1)] \quad (2)$$

$$r(t) = (U \cdot x(t)) \quad (3)$$

$$y(t) = g(V \cdot r(t)) \quad (4)$$

$$m(t) = g_m(V_w \cdot w(t) + V_r \cdot r(t) + V_i \cdot I) \quad (5)$$

The symbol shows at the equation assign to input word layer (w), recurrent layer (r), output layer (y), $f(\cdot)$ and $g(\cdot)$ correspond to a variable nonlinear function, U and V are a weighted matrix learning, $m(t)$ is an activation vector layer for multimodal compute.

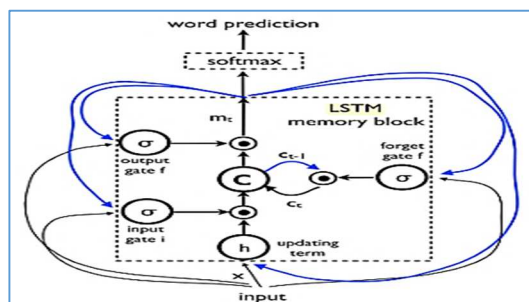


Fig. 6. LSTM architectures [4].

Fig. 6 is a simplicity of recurrent neural networks (RNN). RNN runs like dynamic temporal model that map a sequence input into *hidden states*, and continues to output following the equation bellows:

$$h_t = g(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \quad (6)$$

$$z_t = g(W_{hz}h_t + b_z) \quad (7)$$

The letter symbol assign to *element-wise non-linearity* (g) for sigmoid or tangent hyperbolic function, *input* (x_t), *hidden state* (h_t) with N *hidden units* $\in \mathbb{R}^N$, and *output* (z_t) at time t . For each *input sequence* $\langle x_1, x_2, \dots, x_T \rangle$ with length

To proceed by calculate each word input and ignore h_0 , $h_1, z_1, h_2, z_2, \dots, h_T, z_T$. Sigmoid function $\sigma(x) = (1 + e^{-x})^{-1}$ is a non-linear function with real value and encourage between $[0,1]$. On the other hand, hyperbolic function $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = 2\sigma(2x) - 1$ is a function that has a limitation value between $[-1,1]$. Updating LSTM is aligning with time t for each input x_t , h_{t-1} , and c_{t-1} following with function as follows:

$$i_t = \sigma(W_{ix}x_t + W_{im}m_{t-1}) \quad (8)$$

$$f_t = \sigma(W_{fx}x_t + W_{fm}m_{t-1}) \quad (9)$$

$$o_t = \sigma(W_{ox}x_t + W_{om}m_{t-1}) \quad (10)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot h(W_{cx}x_t + W_{cm}m_{t-1}) \quad (11)$$

$$m_t = o_t \odot c_t \quad (12)$$

$$p_{t+1} = \text{Softmax}(m_t) \quad (13)$$

D. Beam Search

Beam search is an algorithm to construct the proper caption. The work follows argmax function that selected value with the highest probabilistic value for each word generated. Beam search set into $K=3$ with 22 length words. The process rehased by following with approximate S value as an $\text{argmax}_{S'} = p(S^i|I)$ function [3], [21].

E. Metric

For assessment of the model, the study uses a metric that evaluate the precision between caption generated and caption referenced from ground-truth [22]. The algorithm run first to calculate of precision modification p_n , for all of testing corpus.

$$p_n = \frac{\sum_{c \in \{\text{candidate}\}} \sum_{n\text{-gram} \in c} \text{Count}_{\text{clip}}(n\text{-gram})}{\sum_{c \in \{\text{candidate}\}} \sum_{n\text{-gram} \in c} \text{Count}_{\text{clip}}(n\text{-gram})} \quad (14)$$

Second run to calculate BLEU score by:

$$\text{BLEU} = \text{BP} * e^{(\sum_{i=1}^N w_n \log p_n)} \quad (15)$$

$$\text{BP} = \begin{cases} 1, & \text{jika } c > r \\ e^{(1-r/c)}, & \text{jika } c \leq r \end{cases} \quad (16)$$

BP, N , P_n , and W_n assign to brevity penalty, n -gram candidate, precision score, weighted modification of precision, respectively. Default value for those parameters is $N=4$, $W_n = 1/4$ or 0.25.

IV. RESULT AND DISCUSSION

A. Computation Environment

The experiment uses Google Collab Pro and Python version 3.6. The hardware set GPU NVIDIA T4 or P100 and 25Gb RAM. The python library that occupies such as NumPy, pandas, string, pickle, os, keras 2.3.0, and TensorFlow 1.x. For extracting the image, the study used VGG16 [9], [11], and InceptionV3 [18]. Besides that, construction of proper a word leveraged model [4].



B. Result

Table 1 shows the comparison of results with ground truth. Several results were shown bias, implying it does not match the caption it should. It occurs because the dataset

used still does not meet the formed vocabulary. However, it is still essential to prepare a sufficient number of datasets in conducting training. What must be considered is the caption that has the same semantics.

The log-likelihood optimization method has shown the results of the calculation of loss obtained based on (19). The loss obtained that there is an optimization difference between the argmax of the parameters and the resulting model. This difference occurs due to the empirical distribution defined in training set with the probability distribution of the resulting model. The noticeable difference is in the InceptionV3 model and shows approximately 26% [23].

TABLE I
RESULT COMPARISON FROM EACH MODEL

Rock Images	Caption
	VGG16 + LSTM (translate) quartzite rock is brown ResNet50 + LSTM (translate) granite InceptionV3 + LSTM (translate) granite Ground-Truth (translate) Gabbro is a coarse-grained, dark-colored intrusive igneous rock, usually black or dark green
	VGG16 + LSTM (translate) outcrop of clastic sedimentary rock with indistinct layering planes massive fractured limestone mixed with weathered ResNet50 + LSTM (translate) clastic sedimentary rock with unclear bedding planes of partially crushed and weathered sandstone InceptionV3 + LSTM (translate) clastic sedimentary rock with unclear bedding planes of partially crushed and weathered sandstone Ground-Truth Flint-igneous flint is primarily known for its high hardness and for providing sparks when struck

$$CE = -\sum_i^C t_i \log(f(s)_i) \quad (17)$$

CE, C , t_i , s_i , and $f(s_i)$ assign to cross entropy, total label class, numeric embedding of word by ground-truth, numeric value from LSTM as a prediction word, SoftMax function at (20), respectively.

$$f(s)_i = \frac{e^{s_i}}{\sum_j^C e^{s_j}} \quad (18)$$

C. Discussion

This study uses two language model that is LSTM and bi-LSTM [4]. The difference is on folding at recurrent state, bi-LSTM uses twice to predict the next word and LSTM only one recurrent state. Process combines of image descriptors and feature text on LSTM or bi-LSTM to predict a word through concatenate process, it will be created a precise word even longer time. The true captions are a target of this study, and try to re-engineering the CNN with the true output. Image extraction and text extraction were supported by

preparation of sequence the vector to reach the objective caption.

Table 2 shows various methods among VGG16, ResNet, and InceptionV3 and combines with word2vec. The VGG16 architecture does not leverage normalization, nevertheless, the extraction process maintains the number of parameters processed. It can be stated from Table 2 that the use of normalization and time can influence the run time and parameter created. InceptionV3+word2vec is the model better than VGG16+word2vec and ResNet50+word2vec. It is evident because of the use of normalization, which gives the process space for the sample to be divided into smaller ones. Moreover, the normalization on ResNet50 and InceptionV3 can speed up outcomes when generating the caption. The experiments use 224x224 and 229x229 input shape, MaxPooling, and ReLU activation, and 4096 and 2048 dense units.

TABLE II
SETTING PARAMETERS OF CNN WHEN EXECUTED

Image Descriptors	Layer	Filter	Parameters	Normalization	Time
VGG16	16	1x1, 3x3, 5x5	134,260,544	-	92
ResNet50	50	1x1, 3x3	23,534,592	Batch Normalization Size (1024, 256)	70
InceptionV3	42	3x3, 5x5	21,768,352	Batch Normalization size (96, 192, 288, 384, 480, 576, 1152, 1344)	45

In Table 3, columns BLEU-N (B@N), the calculation of precision caption results is presented using the testing dataset. It can be concluded that VGG16+word2vec+LSTM has a higher BLEU score for B@1 to B@4. Measuring BLEU score can be viewed from the precision and calculating used (14), (15), and (16) of caption compares to ground truth. This study focuses on B@4 because the advantage and weakness can be observed when generating the word. Each word was generated or predicted by the model always the precise caption that can be achieved. Several results probable mistakes when generating the word, pairwise between area image and the text still shifted from the original text or ground truth.

TABLE III
CAPTION BENCHMARK FROM EACH MODEL

Model	B@1	B@2	B@3	B@4
VGG16+Word2Vec+LSTM	41.3	41.5	42.5	36.7
ResNet50+Word2Vec+LSTM	39.9	38.3	40.3	34.4
InceptionV3+Word2Vec+LSTM	33.2	31.2	33.0	27.3
VGG16 + One-Hot+bi-LSTM	39.9	38.3	40.5	34.7
InceptionV3 +One-Hot+bi-LSTM	37.6	34.1	35.4	29.6
ResNet50 +One-Hot bi-LSTM	40.3	39.2	41.5	35.9

The score of B@1, Table 3, in the test dataset can still be taken into account because the similarity per word is possible. Nevertheless, when B@2, B@3, and B@4, the score results will not be decisive because word pairs are rarely found in the corpus. When viewed from the side of the Brevity Penalty (BP) calculation, the caption results will

always be calculated by comparing the result caption with the ground truth caption. If the generated word length is smaller, the BLEU score will be lower [22].

If it reviews the image extraction model and the language model in Table 2 and Table 3, it can be confirmed that the three models are still relevant for the caption. However, the results of this model have not been related to the specifics of the image read, such as:

- Similarity of area objects that only rely on color, texture, and object size [21].
- The class used for object detection does not yet reference the geological image of the rocks. In fact, the classes in the VGG16, ResNet50, and InceptionV3 models are taken from ImageNet weights. Classes defined in ImageNet are classes that recognize various objects such as humans, objects, vehicles, and others. Meanwhile, background objects such as land surface, mountains, rocks in mountains and rivers have not been annotated [9].
- The mapping of image features and caption features in making the model does not pay attention to the semantics and relations between the objects being read. In the caption, geological rocks are very important in expressing the right word relations. This word relation is to strengthen the description of the geological image of the rocks in order to approach the ground-truth [24], [25].
- The use of bounding-box in InceptionV3 is empirically compatible with the case raised. But InceptionV3 still get high Loss and low precision. This is evidenced by the results in Table 2 and Table 3. The order of words that follow the standard arrangement of the descriptions of geologists is very important to note [17].

Several errors were confirmed when generating a word. The errors that arise from the caption results are:

- 1) *Biased word generation, as in:*
 - VGG16 + LSTM: quartzite rock is brown;
 - ResNet50 + LSTM: granite rock;
 - InceptionV3 + LSTM: granite rock;
 - Ground-Truth: Gabbro is a coarse-grained, dark intrusive igneous rock, usually black or dark green.
- 2) *Generating semantic appropriate captions but low BLEU score:*
 - VGG16 + LSTM: marble or green marble;
 - ResNet50 + LSTM: gray schist rock;
 - InceptionV3 + LSTM: gray schist rock;
 - Ground-truth: Topaz contact pneumatolytic metamorphic rock.

The results given by VGG16 are close to that they are both metamorphic. However, the caption results have not been able to show a good score. Machine failures are in ResNet50 and InceptionV3. Unlike the VGG16 model, it is still better at producing captions. It is understandable reasoning because ResNet50 and InceptionV3 include an image reduction task at each convolution stage.

V. CONCLUSION

The study can be concluded from the experimental results are: 1) the arrangement of captions that follow the standard

rules in providing geological descriptions; 2) each caption sometimes has adjacent semantics according to its types, such as sedimentary rock, igneous rock, and metamorphic rock; 3) the number of vocabularies in the corpus must have a sufficient amount so that when testing, the results are not too biased; 4) Color gradations that characterize rocks that are rather dark and rocks are sharp, it is necessary to prepare a color corpus as identification of the color of the rocks; 5) Determination of the word relation, it must be recognized from the image whether it is stacked, crossed, scattered, or otherwise. It also needs to be followed up with a corpus regarding the word relation.

There needs to be an in-depth study of the image. The need for image analysis includes: Preprocessing tasks, such as augmentation tasks, image sizes, and image captions. In the caption, it is necessary to pay attention to the use of punctuation marks. Sometimes the image description uses the "-" sign to make adjectives or derivatives of rock types. Feature extraction task that handles feature extraction from the image. This extraction must be able to recognize parameters such as color, texture, and object size. The most important thing is that this extraction must be able to directly classify the rock's name, color, and texture. The mention of the name of the rocks can be directly accompanied by their nature, such as carbonate mudstone, clay sandstone, and others. Task classification and interpretation, this task is to help compose captions based on image feature extraction and text feature extraction.

ACKNOWLEDGMENT

I would like to express my very great appreciation to Dr. Joko Wahyudiono and Mrs. Fitriani Agustin for they valuable and constructive suggestions during the planning and development of this research work. His willingness to give his time so generously has been very much appreciated and materials image and caption of geological rocks.

REFERENCES

- [1] Y. U. and T. H. Andrew Shin, 'Image Captioning with Sentiment Terms via Weakly-Supervised Sentiment Dataset', in *British Machine Vision Conference*, 2016, pp. 53.1-53.1, doi: 10.5244/C.30.53.
- [2] J. Aneja and A. G. Schwing, 'Convolutional Image Captioning', *Computer Vision and Pattern Recognition*, pp. 5561-5570, 2017, [Online]. Available: <https://arxiv.org/abs/1711.09151>.
- [3] A. Karpathy and L. Fei-Fei, 'Deep Visual-Semantic Alignments for Generating Image Descriptions', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 664-676, 2015, doi: 10.1109/TPAMI.2016.2598339.
- [4] O. Vinyals and A. Toshev, 'Show and Tell: A Neural Image Caption Generator', in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156-3164.
- [5] V. Kougia, J. Pavlopoulos, and I. Androutsopoulos, 'A Survey on Biomedical Image Captioning', in *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, 2016, pp. 26-36, [Online]. Available: <https://www.aclweb.org/anthology/W19-1803>.
- [6] A. Farhadi *et al.*, 'Every Picture Tells a Story : Generating Sentences from Images', in *European conference on computer vision*, 2010.
- [7] M. Hodosh *et al.*, 'Framing Image Description as a Ranking Task : Data , Models and Evaluation Metrics (Extended Abstract) *', *Journal of Artificial Intelligence Research*, vol. 47, no. Ijcai, pp. 4188-4192, 2015, [Online]. Available: <https://www.jair.org/index.php/jair/article/view/10833/25854>.
- [8] A. Frome *et al.*, 'DeViSE: A Deep Visual-Semantic Embedding Model', in *Neural Information Processing Systems 2013*, 2013, pp. 3128-3137, doi: 10.1016/0921-4534(95)00110-7.
- [9] H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian, 'Deep Residual Learning for Image Recognition', *Computer Vision*, pp. 1-9, 2016.
- [10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, 'ImageNet Classification with Deep Convolutional Neural Networks', *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, vol. 25, pp. 1-9, 2012, doi: 10.1201/9781420010749.
- [11] K. Simonyan and A. Zisserman, 'Very Deep Convolutional Networks for Large-Scale Image Recognition', *arxiv*, pp. 1-14, 2015, [Online]. Available: <http://arxiv.org/abs/1409.1556>.
- [12] J. Redmon and A. Farhadi, 'YOLOv3: An Incremental Improvement', 2018, [Online]. Available: <http://arxiv.org/abs/1804.02767>.
- [13] C. Szegedy, V. Vanhoucke, and J. Shlens, 'Rethinking the Inception Architecture for Computer Vision', in *Computer Vision Foundation*, 2014, pp. 2818-2826.
- [14] Y. Bhatia, A. Bajpayee, D. Raghuvanshi, and H. Mittal, 'Image Captioning using Google's Inception-resnet-v2 and Recurrent Neural Network', in *2019 Twelfth International Conference on Contemporary Computing (IC3)*, 2019, vol. 2019, pp. 1-6.
- [15] S. L. Granizo, A. L. V. Caraguay, L. I. B. Lopez, and M. Hernandez-Alvarez, 'Detection of Possible Illicit Messages Using Natural Language Processing and Computer Vision on Twitter and Linked Websites', *IEEE Access*, vol. 8, pp. 44534-44546, 2020, doi: 10.1109/ACCESS.2020.2976530.
- [16] X. He, B. Shi, X. Bai, G. Xia, and Z. Zhang, 'Image Caption Generation with Part of Speech Guidance', *Pattern Recognition Letters*, vol. 119, no. March, pp. 229-237, 2018, doi: 10.1016/j.patrec.2017.10.018.
- [17] W. Joko, S. Aris, A. Ryandi, and S. Bisma, 'Penelitian Geologi Dan Geofisika Untuk Pengusulan Wilayah Kerja Migas Seram Onshore', Bandung, Indonesia, 2017.
- [18] C. Szegedy, L. Wei, and E. Al, 'Going Deeper with Convolutions', in *Computer Vision*, 2015, pp. 1-9, doi: 10.1002/jctb.4820.
- [19] X. He, B. Shi, X. Bai, G. Xia, and Z. Zhang, 'Image Caption Generation with Part of Speech Guidance', *Pattern Recognition Letters*, vol. 0, pp. 1-9, 2017, doi: 10.1016/j.patrec.2017.10.018.
- [20] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, 'Gradient-Based Learning Applied to Document Recognition', 1998, [Online]. Available: <http://ieeexplore.ieee.org/document/726791/#full-text-section>.
- [21] X. Zhang *et al.*, 'RSTnet: Captioning with adaptive attention on visual and non-visual words', *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 15460-15469, 2021, doi: 10.1109/CVPR46437.2021.01521.
- [22] K. Papineni, S. Roukos, T. Ward, and Z. Wei-Jing, 'BLEU: a Method for Automatic Evaluation of Machine Translation', in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311-318, [Online]. Available: <https://dl.acm.org/citation.cfm?id=1073135>.
- [23] R. Reed and R. J. Marks II, *Neural Smthing*. A Bradford Book, 1999.
- [24] J. Chen *et al.*, 'Computer vision-based limestone rock-type classification using probabilistic neural network', *Geoscience Frontiers*, vol. 7, no. 1, pp. 53-60, 2020, doi: 10.1016/j.gsf.2014.10.005.
- [25] Y. Zhang, M. Li, S. Han, Q. Ren, and J. Shi, 'Intelligent identification for rock-mineral microscopic images using ensemble machine learning algorithms', *Sensors (Switzerland)*, vol. 19, no. 18, 2019, doi: 10.3390/s19183914.