International Journal of Information System & Technology
Akreditasi No. 36/E/KPT/2019 | Vol. 5, No. 3, (2021), pp. 292-298

Image Captioning Approach for Household Environment Visual Understanding

Dhomas Hatta Fudholi

Department of Informatics, Universitas Islam Indonesia, Yogyakarta, Indonesia

hatta.fudholi@uii.ac.id

Abstract

Image captioning task generates a description from an image in a form of a sentence. It holds various critical usage in different applications and domains, such as indexing in retrieval system, by capturing semantic information within the images. In the area of live quality support system, an image captioning model may be used by visual impaired people to achieve visual understanding of their surroundings. In this paper, we present a novel image captioning model that aims to give visual understanding on household environment. To develop the model, we use five different household objects (sinks, chairs, tables, beds, couches) from MS COCO datasets. We create three new captions, in Bahasa Indonesia, for the selected data. The captions describe the name, the color, the position/location, the size, and the type/characteristic of the related object and its close surrounding. InceptionV3 and LSTM architecture is used to train the model with GloVe as the word embedding. In this study, our developed image captioning model can generate caption well and achieved BLEU-1 score of 0.502033, BLEU-2 score of 0.312539, BLEU-3 score of 0.193333, BLEU-4 score of 0.106111, METEOR score of 0.183193, ROUGE-L score of 0.358339, and CIDEr score of 0.348903.

Keywords: *Image Captioning, Deep Learning, Household Environment, Visual Understanding*

1. Introduction

Image captioning is an Artificial Intelligent (AI) task to understand an image by giving a natural language description. Detecting and recognizing the objects within an image is entailed in the image captioning process [1]. The application of image captioning is wide. In the information or image retrieval area, image captioning could give more accurate content-based retrieval. Moreover, for people with visual impairment, such as blindness, image captioning can deliver a visual understanding of their surroundings [2].

In favor of the emerging deep learning methodology, the development of image captioning model is larger than before. Encoder-decoder framework is the basis of most deep learning-based method in image captioning. A Convolutional Neural Network (CNN)-based architecture is utilized in the encoder side to extract image features. Different kinds of CNN architecture have been utilized in various image captioning works such as VGG [3], ResNet [4], and InceptionV3 [5]. On the decoder side, a word-by-word caption is generated by language modeling framework based on Recurrent Neural Network (RNN) architecture, such as LSTM [3].

Due to both the vast application of image captioning and the emerging deep learning technologies. studies in the development of image captioning model in Bahasa Indonesia has started to emerge. The general image captioning in Bahasa Indonesia can be found in [6,7]. These works used deep learning architecture, such as CNN, ResNet, and LSTM.

In this research, we initialized the development of a novel model of image captioning that aims to support visual understanding of household environment in Bahasa Indonesia. A portion of few household objects images is used as the dataset. These images are taken from MS COCO dataset [8]. InceptionV3 and LSTM are used as the base architecture in

ISSN : 2580-7250

the encoder-decoder framework. Finally, we evaluated the model using language translation evaluation metrics, such as BLEU [9], METEOR [9], ROGUE [10], and CIDEr [11].

The rest of the paper is structured as follows. Section 2 elaborates related works of the study. The section presents recent image captioning studies that use deep learning methodology in building the model. We are more focused on the work of generating image captioning model for Bahasa Indonesia. Section 3 gives the detail of the research methodology, including the architecture used and the model evaluation methods. Section 4 presents the result of the study. Finally, section 5 concludes the paper.

2. Research Methodology

Deep learning methodology and technology is undoubtedly accelerating the research and implementation of AI in various domain. In term of image captioning, the complexities and challenges within the area can be handled well by deep learning technologies. The survey work in [1] capturing more than 40 image captioning works, globally, from 2014 to 2018. Various deep learning architecture has been used in both feature extraction process on the encoder part and caption generation process on the decoder part. Few different works with different image feature extraction architecture that are used in the works are VGG [3], Inception-V3 [5], and ResNet [4], GoogleNet [12], and AlexNet [13]. On the caption generation part, few different works use different architectures as well, such as RNN [14] and LSTM [3]. In term of dataset, most of the studies surveyed in [1] uses MS COCO [8] and Flickr [15,16] dataset. The common evaluation metrics uses in the studies are BLEU [9], METEOR [9], ROUGE [10], and CIDEr [11].

The studies of image captioning in Bahasa Indonesia have given broader application of the image captioning task. Two of them are [6] and [7]. The work in [6] initiates Indonesian image captioning model in a view that most research in the area generates English caption and every language has uniqueness. Moreover, Indonesia is the largest country in southeast Asia and Indonesian language has been taught in other countries. In the research, MS COCO and Flickr30k dataset is used as the base images and the captions are translated to Bahasa Indonesia. ResNet101 architecture is used in the encoder side and LSTM architecture is used in the decoder side. The generated model achieved BLUE-1 score of 0.678. Extending the works in modeling image captioning in Bahasa Indonesia, the work [7] uses FEEH-ID dataset in the making of the model and achieve a quite promising result with BLEU-1 score achieved equals to 50.0.

In contrast with the aforementioned works, the aim of this paper is to give a visual understanding of household environment through a novel image captioning model. From the MS COCO datasets, we choose several household objects and give them new captions, in Bahasa Indonesia. Each caption describes the descriptive information of the related object. The model is trained using InceptionV3 and LSTM architecture with GloVe as the word embedding.

From collecting data to preparing the data, processing it, to modeling, and then evaluating the outcome of the research, these are the steps we underwent. In the following section, each step in the methodology is elaborated.

2.1. Data Collection

To develop our image captioning model for household environment, we use Microsoft Common Objects in COntext (MS COCO) dataset. MS COCO is a big dataset from Microsoft COCO that consist of object detection, segmentation, and captioning [8]. The dataset has 328,000 images with 91 different types of objects that can easily be recognized by a four-year-old. Eighty-two (82) of these objects containing over 5,000 labeled instances and created a total of 2,500,000 labeled instances.

As an initial approach in developing our image captioning model, we take five household objects at this point. The five objects are *sinks*, *chairs*, *tables*, *beds*, and *couches*. For each object, we take 50 images. Therefore, our selected household dataset has a total of 250 images. Figure 1 shows example of the selected images within the selected object.

image_1 image_2 image_3

Figure 1. Example of Selected Object Images

For each selected image in all five objects, we dropped its original caption and added three new captions in Bahasa Indonesia. The three new captions for each image are three different sentences. The different captions reflect the different way people of describing the image. Each caption may include related information on the name, the color, the position/location (from the viewer point of view), the size, and the type/characteristic of the object and its close surrounding. Table 1 shows the example of the caption for images in Figure 1.

Table 1. Example of the Image Caption

Image	Caption	Translated Caption (English)
image_1	‘wastafel berwarna putih berada di depan’, ‘tempat sampah kecil berwarna biru berada di bgian kanan wastafel’, ‘di bagian kanan wastafel terdapat kloset duduk’	'the white sink is in the front of you', 'the little blue trash can is to the right of the sink', 'to the right of the sink there is a toilet seat'
image_2	‘di bagian kanan terdapat meja bundar kecil bertaplak hitam’, ‘di bagian kanan terdapat beberapa pria yang mengitari meja makan yang memiliki banyak kue di atasnya’, ‘di bagian kiri terdapat banyak pria dan wanita yang sedang berdiri’	‘on the right side, there is a small round table with a black cloth', 'on the right side there are several men around the dining table which have a lot of cakes on it', 'to the left side there are many men and women standing'
image_3	‘di depan terdapat kasur berukuran kecil dengan corak kotakkotak berwarna biru dan putih’, ‘sepasang sepatu berwarna hijau berada di atas kasur’, ‘rak kecil dengan lampu meja di atasnya terletak di samping kanan kasur’	'In the front, there is a small bed with blue and white checkered patterns', 'a pair of green shoes are on the bed', ‘a small shelf with a table lamp on it is located to the right of the bed’

2.2. Data Preparation

During the data preparation process, all images is resized to 299x299 pixels. The resizing size follows the required input size in the InceptionV3 architecture. In the image caption side, all captions are lower cased. In addition, to indicate the beginning and the end of the caption sentence, we add *startseq* and *endseq* in each caption.

2.3. Modelling

A merge architecture for caption generation is used in this study. In the merge architectures, the image features are extracted separately using CNN-based architecture and the RNN-based architecture only handles the caption prefix, or, in other words, handles only linguistic information [17]. A separate multimodal layer is then created after the RNN subnetwork is constructed to combine the image vector with the prefix vectors. By concatenating two vectors, for instance, the two can be merged. The RNN-based architecture is trained to encode only the prefix, and the mixture is handled in a subsequent feedforward layer.

Figure 2 shows the merge architecture that we used to develop the image caption model. From the figure, input_4 take the caption sentences as the input. The word embedding model that we used in this study is an Indonesian GloVe model (https://github.com/irfanhanif/Mira). The RNN-based architecture that we used in this architecture is LSTM with one layer and 256 nodes. On the other hand, the image feature extractor architecture used is InceptionV3. InceptionV3 is one of the state-of-the art CNN architecture. The extracted image vector enters through the input_3 block (as seen in Figure 2). Finally, *Greedy search* method is used to choose the prediction of the caption words.

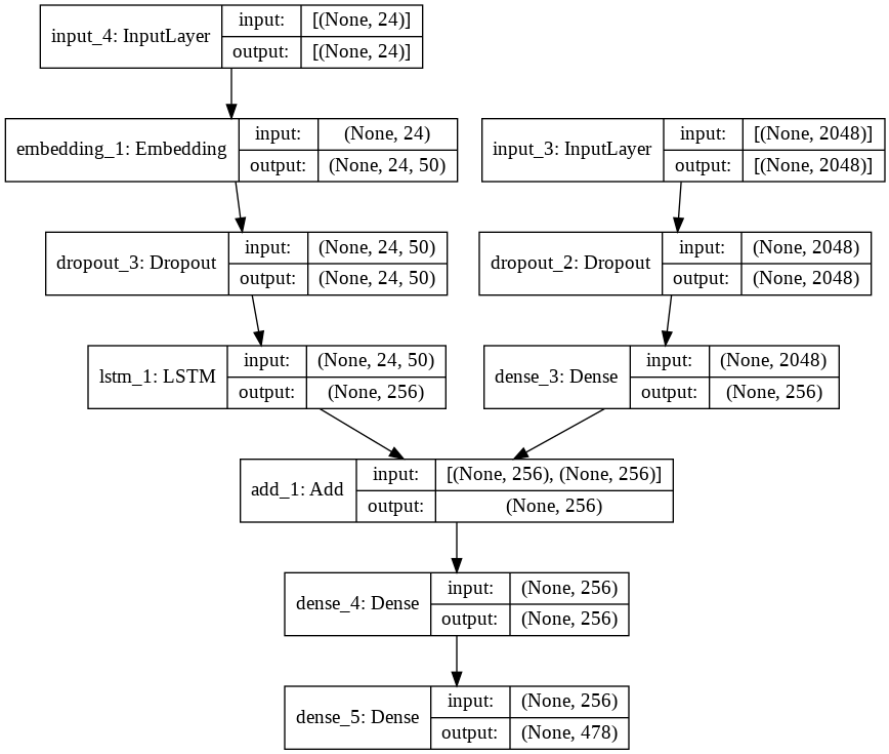


Figure 2. Image captioning model generation architecture

2.4. Evaluation

The evaluation metrics that we used to score the generated caption are BLEU (BLEU-1, BLEU-2, BLEU-3, and BLEU-4), METEOR, ROUGE-L, and CIDEr. BLEU (Bilingual Evaluation Understudy) evaluates candidate texts by calculating the n-gram overlap between candidate and reference texts [31]. METEOR (Metric for Evaluation for Translation with Explicit Ordering) provides an assessment based on overlapping unigrams between candidate and reference texts. This corresponds to a unigram based on meanings, exact forms, and stemmed forms [31]. ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence) compares automatically generated summaries with human summaries based on Longest Common Subsequences

(LCS). In short, the LCS is between automatically generated summaries as well as human results summaries, if both show high similarity then the summary system quality is considered good [32]. CIDEr (Consensus-based Image Description Evaluation) measures the consensus between candidate and reference texts using n-gram matching. N-grams that are common in all texts are weighted by calculating the weighting of the Term Frequency Inverse Document [26].

3. Result and Discussion

During the training, we divided our dataset into 80:20 ratio of train set and test set. The train set has 200 images, and the test set consist of 50 images. In terms of hyperparameters, we used Adam as the optimizer, define batch size value of 3, and run for 60 epochs.

The training process created an image captioning model that may be used in household environment visual understanding. Table 2 shows the evaluation metrics resulted from the testing phase. Our evaluation score is comparable to another Indonesian image captioning model, such as in [6, 7]. In term of the task in generating caption, our model performs quite well. Table 3 shows three test example images with their generated captions. As seen in the generated caption, we are able to generate the name of the object (*meja*/table, *wastafel*/sink), the color of the object (*putih*/white), the position/location of the object (*di depan*/at the front), the size of the object (*berukuran sedang*/medium-sized), and the type/characteristics of the object (*kayu*/wooden). Finally, our model becomes one of the pioneer models that capable in becoming a visual understanding model in household environment domain that might be very useful for monitoring and supporting visual impaired people, especially in Indonesia.

Table 2. Model Evaluation Result

Metrics	Score
BLEU-1	0.502033
BLEU-2	0.312539
BLEU-3	0.193333
BLEU-4	0.106111
METEOR	0.183193
ROUGE-L	0.358339
CIDEr	0.348903

Table 3. Caption Generation Result

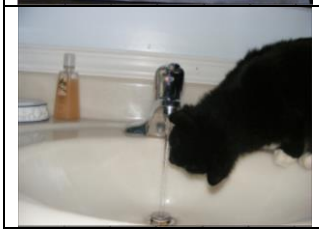
Image	Generated Caption
	*Di depan terdapat meja kayu berukuran sedang dengan sepiring salad dan segelas soda.* (translation: At the front was a medium-sized wooden table with a plate of salad and a glass of soda.)
	*Kucing berbaring di atas wastafel berwarna putih* (translation: Cat lying on the white sink)

Image	Generated Caption
	*Terdapat meja konter dan kabinet mengelilingi dapur* (translation: There is a counter and a cabinet around the kitchen)

5. Conclusion

We introduced an image captioning approach for household environment visual understanding. In this research, we attempt to provide visual comprehension of the household environment. We use five different domestic objects (*sinks, chairs, tables, beds, couches*) from MS COCO datasets to create the model. Three new captions in Bahasa Indonesia are added that mention object's name, color, position/location, size, and type/characteristic, as well as its immediate surroundings. The model is trained using InceptionV3 and the LSTM architecture, with GloVe as the word embedding. Our model can generate captions with the aforementioned aspects of the object. The model achieves BLEU-1 score of 0.502033, METEOR score of 0.183193, ROUGE-L score of 0.358339, and CIDEr score of 0.348903.

References

[1] MD. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A Comprehensive Survey of Deep Learning for Image Captioning", ACM Computing Surveys, 51(6), **(2019)**, pp 1–36.

[2] F. Chen, X. Li, J. Tang, S. Li, and T. Wang, "A Survey on Recent Advances in Image Captioning", Journal of Physics: Conference Series, 1914(1), **(2021)**.

[3] Q. Wu, C. Shen, P. Wang, A. Dick, and A. van den Hengel, "Image Captioning and Visual Question Answering Based on Attributes and External Knowledge". IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(6), **(2018)**, pp 1367–1381.

[4] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-Critical Sequence Training for Image Captioning", 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), **(2017)**.

[5] Z. Li, et. al., "Actor-critic sequence training for image captioning". arXiv preprint arXiv:1706.09601. **(2017)**

[6] M. R. S. Mahadi, A. Arifianto, and K. N. Ramadhani, "Adaptive Attention Generation for Indonesian Image Captioning", 2020 8th International Conference on Information and Communication Technology (ICoICT), **(2020)**.

[7] E. Mulyanto, E. I. Setiawan, E. M. Yuniarno, and M. H. Purnomo, "Automatic Indonesian Image Caption Generation using CNN-LSTM Model and FEEH-ID Dataset", 2019 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), **(2019)**.

[8] T.Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," *Eccv*, [Online], http://arxiv.org/abs/1405.0312, **(2014)**, pp 740-755.

[9] N. Sharif, L. White, M. Bennamoun, and S. A. A. Shah, "NNEval: Neural Network Based Evaluation Metric for Image Captioning", Computer Vision – ECCV 2018, Springer International Publishing, **(2018)**, pp. 39–55.
[10] A. Cohan and N. Goharian, "Revisiting summarization evaluation for scientific articles," *Proc. 10th Int. Conf. Lang. Resour. Eval. Lr. 2016*, **(2016)**, pp. 806–813.
[11] X. Chen *et al.*, "Microsoft COCO Captions: Data Collection and Evaluation Server," [Online], Available: http://arxiv.org/abs/1504.00325, **(2015)**, pp. 1–7.
[12] R. Shetty, M. Rohrbach, L. A. Hendricks, M. Fritz, and B. Schiele, "Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training", 2017 IEEE International Conference on Computer Vision (ICCV), **(2017)**.
[13] S. Ma, and Y. Han, "Describing images by feeding LSTM with structural words", 2016 IEEE International Conference on Multimedia and Expo (ICME), **(2016)**.
[14] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, "Areas of Attention for Image Captioning", 2017 IEEE International Conference on Computer Vision (ICCV), **(2017)**.
[15] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguist.*, vol. 2, **(2014)**, pp. 67–78.
[16] S. He and Y. Lu, "A Modularized Architecture of Multi-Branch Convolutional Neural Network for Image Captioning," *Electronics*, vol. 8, no. 12, **(2019)**.
[17] M. Tanti, A. Gatt, and K. P. Camilleri, "Where to put the image in an image caption generator", Natural Language Engineering, 24(3), **(2018)**, pp 467–489.

Authors

Dhomas Hatta Fudholi is an Assistant Professor at the Department of Informatics, Universitas Islam Indonesia. He earned his Ph.D. in Computer Science and IT in 2016 from La Trobe University, Melbourne, Australia with a full postgraduate scholarship from La Trobe University. Previously, he earned his Master's degree from King Mongkut's Institute of Technology Ladkrabang, Thailand, and his Bachelor's degree from Universitas Gadjah Mada, Indonesia in 2008. His research interests are mainly related to ontology, data science, natural language processing, deep learning, and big data. He explores data science and deep learning methods to support knowledge base development and various useful applications. Outside of academia, he loves photography and cycling.