International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 12, No.1 (Aug-2022)

https://dx.doi.org/10.12785/ijcds/120138

The Role of Transformer-based Image Captioning for Indoor Environment Visual Understanding

Dhomas Hatta Fudholi[1] and Royan Abida N. Nayoan[2]

[1]*Department of Informatics, Universitas Islam Indonesia, Yogyakarta, Indonesia*
[2]*Master Program in Informatics, Universitas Islam Indonesia, Yogyakarta, Indonesia*

Received 12 Sep. 2021, Revised 3 Jul. 2022, Accepted 15 Jul. 2022, Published 6 Aug. 2022

Abstract: Image captioning has attracted extensive attention in the field of image understanding. Image captioning has two natural parts; image and language expressions that combines computer vision and NLP to generate caption. Image captioning focuses on making the model to be able to get the description of the image as accurate as the ground-truth captions delivered by humans. Image captioning can be applied into different scenarios, such as helping the visually impaired people to get a better visual understanding of their surroundings environment through generated image caption that can be translated to speech. In this paper, we present a novel image captioning approach in Bahasa Indonesia, using Transformer, to enable visual understanding of indoor environments. We use our own modified MSCOCO dataset. Here, we used ten different indoor objects from MSCOCO datasets namely, beds, sinks, chairs, couches, tables, televisions, refrigerators, house plants, ovens, and cellphones. We modified the captions by creating three new captions in Bahasa Indonesia that includes the objects name, color, position, size, characteristics, and its close surrounding. We use Transformer architecture, which is then compared with merged encoder-decoder architecture model with different hyperparameter tunings. Both model architectures used InceptionV3 in extracting image features. The result of our experiment shows that the Transformer model with a batch size of 64, number of attention heads of 4, and a dropout of 0.2 outperforms other models with a BLEU-1 score of 0.527565, BLEU-2 score of 0.353696, BLEU-3 score of 0.227728, BLEU-4 score of 0.146192, METEOR score of 0.184714, ROUGE-L score of 0.377379, and CIDEr score of 0.393117. Finally, the inference result shows that the generated captions could give indoor environment understanding.

Keywords: : Image Captioning, Bahasa Indonesia, Transformer, Visual Understanding, Indoor Environment

1. Introduction

Image captioning has been very popular in the field of artificial intelligence that helps in generating description of the image. Image captioning generation combines computer vision, Natural Language Processing (NLP), and machine learning. Image captioning is crucial for various reasons and can be applied into different scenarios like adding subtitles to video, video question answering, image searching [1], and assistive application for the blind. For the blind and impaired people, image captioning could play a huge role in helping them and get a better sense of what is happening around.

Due to the rapid development of deep learning, image captioning has now gotten better and better. The first approach of image captioning based on deep learning is the retrieval-based method. The recent advances in image captioning architectures can be divided into several categories: encoder-decoder methods, attention-based methods, semantic-based methods, and transformer-based methods.

Most image captioning methods usually use encoder-decoder framework that consists of two simple parts [2], [3]. The first part is the encoder. CNN (Convolutional Neural Network) is usually utilized as an encoder to encode the images and turn them into embedding vectors. The second part is to generate the caption word by word and RNN (Recurrent Neural Network) is usually used as the decoder. Encoder-decoder model were used in previous works [4], [5] by employing LSTM to generate high-quality image captions and CNN as the encoder to mapped image features into embedding vector representation. Figure 1 shows the illustration of common encoder-decoder architecture in image captioning.

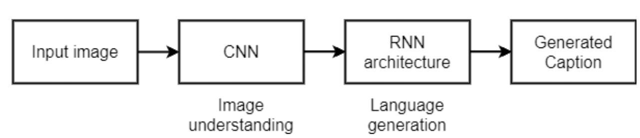

Figure 1. Common encoder-decoder architecture for image captioning.

Attention-based methods are becoming popular after its first introduction in the paper “Show, Attend and Tell” by

E-mail address: hatta.fudholi@uii.ac.id, royan.nayoan@students.uii.ac.id

Xu et al. [6]. The paper explained the two attention types: a soft attention mechanism and a hard attention mechanism. Between both attentions, hard attention is slightly better. Hard attention outperformed other models like Google NIC [6], MS Captivator [7], and Log Bilinear [8]. Even so, this model has a drawback in capturing high-level information since it utilizes image features from the lower CNN layer to focus on most relevant image regions during generator.

Semantic-based image captioning works by selectively attend to semantic concept proposals. Study [9] created a captioning model using concept tokens that produces rich semantics. The study employed Concept Token Network (CTN) that is composed of Meshed-Memory transformer blocks. The model shows by incorporating semantic captions, model is able to improve the CIDEr and BLEU-4 score and benefits the captioning task. Study [10] works by fusing the semantic concept proposals into hidden states and outputs of recurrent neural networks (RNN). Their work managed to outperform other state-of-the-art models on different evaluation metrics. Other model such as [11] and [12] also incorporated semantic approach and their results exceed other models on MSCOCO benchmark datasets.

RNN has been used as a decoder in image captioning tasks. However, RNN has a hard time to maintain long-term dependencies and is slow to train. In 2017, Vaswani et al. [13] introduces Transformer that offers a solution and fixes the drawbacks of RNN. Since then, different breakthrough models based on Transformer are developed such as BERT [14] and GPT [15]. This shows that Transformers by employing self-attention gives superior results compared to other RNN models. Transformer has then gained popularity and used as the standard architecture for various language understanding tasks, including image captioning as a sequence-to-sequence problem and their results are very promising [16], [17].

There are not many Indonesian captioned datasets to support Indonesian image captioning [18]. To get the model that will only generate a good and natural caption, the dataset that is used must be a proper translated dataset. The previous Indonesian image captioning papers use Google translate engine or a professional English-Indonesian translator to translate English captioned dataset such as MSCOCO or Flickr [18], [19].

Motivated by the idea of enabling the visual understanding for people in need, in this work, we created a Transformer model to describe indoor space surroundings using Bahasa Indonesia to achieve visual understanding. We build a model to generate captions from the images. We use a Transformer model and fine-tune the model. We also compare the Transformer models to a merged encoder-decoder model to get the best result. This model contributes in identifying indoor objects to achieve visual understanding in an indoor space. Study [20] says that visually impaired people spend their most 80%-90% of the time inside a building. Hence, the image captioning model can be useful in helping the visually impaired people to get a visual understanding of their surrounding environment better through generated image caption that can be translated to speech.

The dataset used to create image captioning in this paper is the images provided by MSCOCO with its original captions dropped. We created our own Indonesian captions that may include object's name, color, position/location (viewer's point of view), characteristics, and its close surrounding. The remaining of the paper is outlined as follows. Section 2 gives an overview of related works in image captioning. Section 3 elaborates our methods to create a Transformer image captioning model starting from the dataset, preprocessing steps, architectures, and evaluation metrics to evaluate our model. Section 4 presents our result and discussion. Lastly, conclusion and future works are presented in Section 5.

2. Related Works

In recent years, studies in image captioning is emerging. The studies in the area are mostly use deep learning to extracts features automatically from the training set. Deep learning is known for its ability to handle a large and diverse set of images or videos [21]. Moreover, deep learning also works best in overcoming the complexities of image captioning. In image captioning, convolutional neural network (CNN) as the encoder is usually used to extract the features and followed by recurrent neural network (RNN) decoder to generate captions. The drawback of using recurrent network models for generating texts is that the model doesn't have the ability to maintain long-term dependencies between the generated words [22].

There are images captioning models that utilize attention mechanisms to their CNN encoder and RNN decoder. Hierarchical attention network (HAN) [23] is one of the said models that paid attention to semantic features in different level that helps in predicting different word depending on the semantic feature while the multivariate residual module (MRM) helps in extracting relevant relation from various features. There are other methods that also utilized attention mechanism to their encoder-decoder methods, such as Attention on Attention (AoA) [24], Auto-Encoder Scene Graph (SGAE) [25], Adaptive attention via visual sentinel [26], gradient policy optimization of SPIDEr [27], and Recurrent Fusion Network (RFNet) [28]. Another research using attention mechanism is HIerarchy Parsing (HIP) [29] that integrated hierarchical structures into an image encoder. HIP helps in filtering features that result in a rich and multi-level image representation.

A new architecture, Transformer, was introduced as one of many breakthroughs in language understanding tasks and easily gained popularity as it fixed the drawback of recurrent models [13]. Transformer is an encoder-decoder model that uses attention (a concept to help in improving the performance of machine translation) to boost the speed. This model has then been adopted by researchers in image captioning to get the best description of images. Image captioning uses an encoder-decoder framework, which is also widely used in attention mechanism and transformer models.

Different research on Transformer improved the image encoding and the generated texts using meshed transformer with memory (M2M) to get the low- and high-level feature that helps in predicting the captions [30]. Another work by [17] created a boosted transformer that utilized semantic concepts (CGA) and visual features (VGA) to improve the model ability in predicting image's description. Personality-captions [31] uses TransResNet and dataset that supported in differentiating personalities to generate image descriptions that are closer to human. In [32], a combination of Inception-ResNetv2 in extract image features and a Transformer model for sequence modeling achieves good result on a conceptual captions dataset (a developed dataset that represents a wider variety of images and caption styles).

In this work, we aim to generate textual description of an image to achieve visual understanding in an indoor space. Our main contribution lies in presenting the evaluation of Transformer architecture on Indonesian language image captions which are different from the common datasets such as MSCOCO [33] or Flickr30k [34] datasets. We dropped the original captions from MSCOCO and replaced them with our own captions that may include object's name, color, position/location (viewer's point of view), characteristics, and its close surrounding. We propose a deep learning architecture using Transformer model. We fine-tuned our model and compared them to another deep learning model; a merged encoder-decoder model, to get the best model in generating captions.

3. Method

In this section we explain in detail, the datasets, methods, and the steps needed to create image captioning model. We firstly collect the data from the large dataset MSCOCO and add our own captions for each of the images that we used. The next step is preprocessing. We preprocessed the texts and the images before feeding them to the model. The third step is feeding the training set to the transformer and merged encoder-decoder architecture. Here we elaborate both architectures in detail. The last step is evaluation, where we explained the evaluation metrics that we used in this work, to evaluate our image captioning model.

A. Data Collection

We are developing image captioning model which can deliver a mechanism of visual understanding inside an indoor space. The captions that we used are different from the common and popular synthetic datasets such as MSCOCO [29] or Flickr30k [30]. In this study, the data captions need to be modified from the original MSCOCO to fit our goal to enable visual understanding of indoor environments. For the image dataset, we use the images provided by MSCOCO, a large-scale dataset with high-quality visual datasets for computer vision that are consisted of object detection, segmentation, and captioning published by Microsoft. The dataset itself was developed with the goal of advancing image recognition. MSCOCO has 1.5 million object instances; 80 object categories that include things like person, chair, etc.; and 91 stuff categories that include things that have no boundaries. The datasets we took from MSCOCO are ten indoor objects, namely, beds, sinks, chairs, couches, tables, ovens, cellphones, televisions, refrigerators, and house plants. Each object is consisted around 70 to 80 images considering the limited images in MSCOCO that are taken in an indoor space. Figure 2 shows the examples of the images.

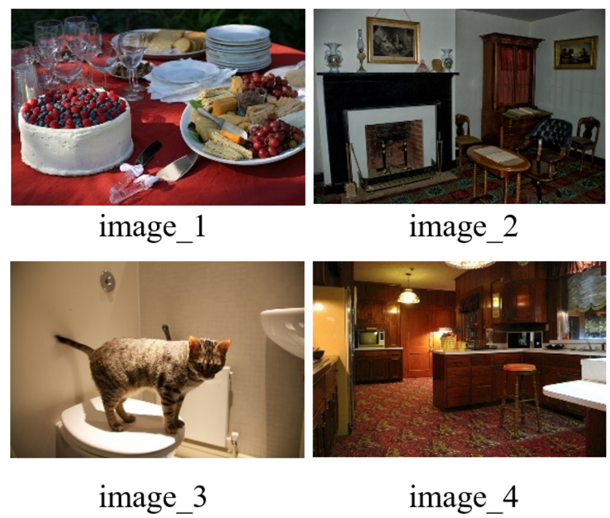

Figure 2. Examples of our selected images in the dataset.

Instead of using MSCOCO's available captions, we created our own captions in Bahasa Indonesia that include object's name, color, position/location (viewer's point of view), characteristics, and its close surrounding. Each of the images are given 3 different captions that mimics the way people describe the images differently. Table I shows the caption and translated caption of each example images with respect to images in Figure 2. Hence, we have a total of 771 images and 2313 captions.

Inspired by MSCOCO [33], we make a few rules in writing the images captions. (1) Since our goal is to describe indoor space surroundings to achieve visual understanding, we added the location information of each object whether the objects are located on the left side/right side/in front of the room and information of their surrounding objects. (2) We only describe the main part of the scene by describing the main objects within view. (3) In describing the objects within view, we also mention the color of the objects and their characteristics since it could be beneficial to help distinguish each object [35].

Our dataset is randomly split into two datasets, train and test dataset. The dataset is divided into 80% for the train dataset and 20% for the test dataset to obtain the best result

[36]. The train dataset has 617 images and 1851 captions, and the test dataset has 154 images and 462 captions in total.

B. Preprocessing

We preprocessed the images and the captions before feeding them to both architectures; Transformer and merged encoder-decoder. Here, we have two different preprocessing for each architecture. Each of the preprocessing are elaborated as follows.

1) Preprocessing for Transformer architecture

We preprocess the image by resizing them into 299x299 size before feeding them to InceptionV3, one of the state-of-the-art pre-trained model [37] that we use as the feature extractor. Since InceptionV3 is only used as a feature extractor and not to classify the images, we remove the last softmax layer.

For the natural language preprocessing, the captions are cleaned from punctuation, single character, and numeric values to obtain clean captions. $\langle start \rangle$ and $\langle end \rangle$ tags is added to the clean caption to make the model understand the beginning and end of the caption. Next, the texts will be tokenized to build a vocabulary. Word in captions that does not exist in the vocabulary will be flagged with an $\langle unk \rangle$ tag. $\langle pad \rangle$ tag is also added to fit a caption that is less than the maximum length of words. In our research, we use the length of 25 words.

2) Preprocessing for merged encoder-decoder architecture

As for the preprocessing in the merged encoder-decoder model, the step is quite simple. Similar to the preprocessing in the Transformer model, we resize the images into a smaller size of 299x299 before feeding them into InceptionV3 without the last layer, to extract image features. For text preprocessing, the step is also similar to the Transformer in cleaning the punctuation, single character, and numeric values. After getting the cleaned text, the last step is adding $\langle startseq \rangle$ and $\langle endseq \rangle$ in each caption.

C. Transformer Architecture

The Transformer architecture was firstly proposed in the paper "Attention is All You Need" by Vaswani et al. [13]. As the title indicates in the paper, Transformer utilizes attention-mechanism to boost the speed. Attention was once introduced to mimic the human mind, which is to selectively focus on a relevant matter and ignoring the other. Attention works by selectively looking for important sequences at each step in the input sequence.

Transformer is one of Seq2Seq architectures with the help of two parts; encoder and decoder, but it differs from the usual Seq2Seq architecture since Transformer doesn't require any Recurrent Neural Networks (RNN). Instead, Transformer is a transduction model that entirely relies on self-attention to compute the representations of its input and output. The encoder and decoder in Transformer are made of a stacked encoder and decoder. In the paper, the

TABLE I. EXAMPLES OF IMAGE CAPTIONS AND THE ENGLISH TRANSLATED CAPTION

Image	Caption (Indonesian)	Translated Caption
Image1	'di atas meja tersedia aneka kue berry, biskuit dan buah anggur', 'meja bundar bertaplak merah memiliki banyak makanan di atasnya', 'peralatan makanan piring, gelas dan pisau berada di atas meja bertaplak merah'	'there are various berry cakes, biscuits, and grapes on top of the table','a lot of foods are placed on top of a round table with a red tablecloth', 'cutleries such as plates, glasses, and knifes, are placed on top of a red tableclothed round table'
Image2	'di depan terdapat perapian dengan foto menggantung di atasnya', 'di samping kanan perapian terdapat bufet tinggi yang berada di pojok ruangan', 'terdapat kursi kecil di sisi-sisi bufet'	'at the front there is a fireplace with a picture frame hanging above', 'on the right side of the fireplace is a tall showcase cabinet placed in the corner of the room', 'there are small chairs on each side of the tall showcase cabinet'
Image3	'terdapat kucing yang berdiri di atas kloset duduk yang tertutup cover', 'wastafel berbentuk oval berada di bagian kanan', 'kloset duduk berada di bagian kiri dengan cover tertutup'	'a cat stand on top of a toilet seat ', 'an oval-shaped sink is on the right side', 'a toilet seat is on the left side'
Image4	'di depan bagian kanan terdapat kursi bundar yang tinggi', 'terdapat meja konter yang ada di setiap sisi ruangan', 'di langit-langit ruangan terdapat dua buah lampu gantung yang letaknya berjauhan'	'at the right front there is a tall round chair', 'there are a counter tables on each side of the room', 'on the ceiling of the room there are two chandeliers hanging that are located far apart'

Transformer model consists of 6 encoders and 6 decoders stacked on top of each other. Encoder block consists of one layer of a multi-head attention (MHA) and a layer of feed forward. The decoder block has a similar layer to the encoder, but decoder has one more extra masked MHA placed between the layers. According to the paper, MHA allows the model to look at other positions in the input that lead to a better encoding for the word. Since no recurrent network is used to remember the sequences, Transformer has a positional encoding to give every word or part their relative or absolute information position.

In this work, we applied Transformer architecture by following the original paper, without significant architectural model modification. We fine-tune our model by setting the hyper-parameter to get the highest result in image captioning. Figure 3 shows the Transformer architecture that we use to train our image captioning model. The experiment and the fine-tuned model results are elaborated in Section 4.

In experimenting the Transformer architecture, we changed the batch size, the attention heads, and the dropout. The modified hyper-parameter for all the models can be seen in Table II. As for the batch size, Transformer uses the typical size of $|Bk| \in \{32,64,\cdots,512\}$ [38]. Model #1, #2 and #3, use a batch size of 64, 128, and 32, respectively. Dropout is also applied to the Transformer layer to reduce over-fitting. A dropout value is ranging from 1.0 to 0.0, where 1.0 means no dropout, and low values of dropout mean more dropout [39]. The original paper uses a dropout of p= 0.1, and also experimented using p=0.2 and p=0.3 for big Transformer model [13]. To note that Model #1 follows the hyper-parameter setting in the original paper. All Transformer models run in 40 epochs and use sparse categorical as the loss function. The model used Adam optimizer with $\beta 1$=0.9 and $\beta 2$=0.98. The learning rate is varied over the course of training, the formula used is Equation 1 [13]. Based on Equation 1, the learning rate is varied by increasing and decreasing the number of learning rate. $d_{m}odel$ on Equation 1 denotes the number used as input in the encoder/decoder. $Step_{n}um$ denotes the total number of training steps, and $warmup_{s}teps$ is the number set at the beginning of the training to reduce the impact of deviating the model. The warmup steps used is taken from the original paper, which is 4000. We also vary the value of attention heads by following the attention heads values suggested in the original paper [13]. The Transformer baseline in the paper uses the attention head value of 8 (Model #1) (see Table II). Here, we experimented the value of attention head of 4 in Model #2 and 16 in Model #3.

TABLE II. IMAGE CAPTIONING HYPER-PARAMETER SETTING

Model	Architecture	Batch Size	Drop out	Attention Heads
#1	Transformer (Original paper)	64	0.1	8
#2	Transformer	128	0.2	4
#3	Transformer	32	0.2	16
#4	Merged Encoder & Decoder	3	0.5	-

$$lrate = d_{model}^{-0.5} \cdot min((step_{num})^{-0.5}, step_{num} \cdot (warmup_{steps})^{-1.5}) \quad (1)$$

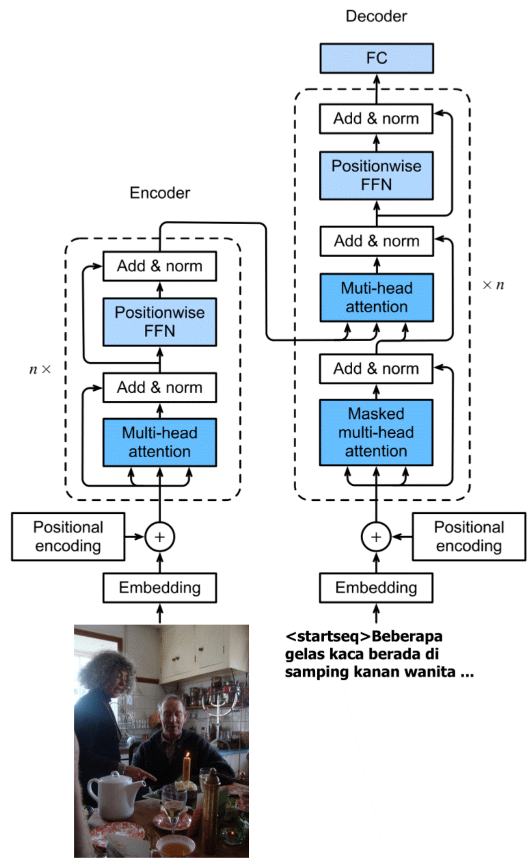

Figure 3. Transformer Architecture Illustration.

D. Merged Encoder-Decoder

As mentioned before, we would like to compare the performance of the Transformer-based architecture with other architecture to show the role of Transformer in creating image captioning model. The chosen architecture to compare is the merged encoder-decoder architecture. The idea behind merged encoder-decoder [40] is to merge image vectors with the prefix outside RNN architecture before feeding them to the feed forward layer. This means that the merged encoder-decoder is used to keep the image out of RNN architecture. This architecture has two parts to handle images and the language separately. Figure 4. shows the conceptual views of a merged encoder-decoder.

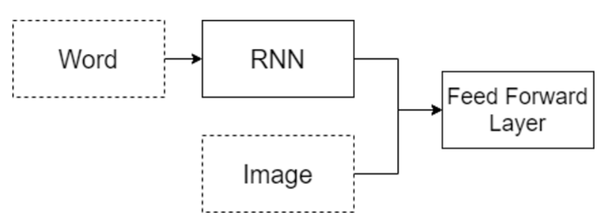

Figure 4. Illustration on merged encoder-decoder.

We use LSTM architecture in training the language and InceptionV3 to extract the image feature. Figure 5 shows the merged encoder-decoder architecture. As seen in Figure 5, the merged encoder-decoder model takes linguistic input in input_4, while the extracted image features are taken in input_3. The model uses keras embedding layer. Both texts and images applied dropout of 0.5 [39] to reduce overfitting. The word vectors are then passed to an LSTM, while the images are passed to a dense layer. Both are then concatenated in the Add layer before passing them to the next fully connected layer.

For the merged encoder-decoder, we experimented using a mini batch of 3. The commonly used value for mini batch is between m=2 and m=32 [41]. The dropout we applied to the merged encoder-decoder architecture is also p= 0.5 [39]. The hyper-parameter chosen for each model can be seen in Table II.

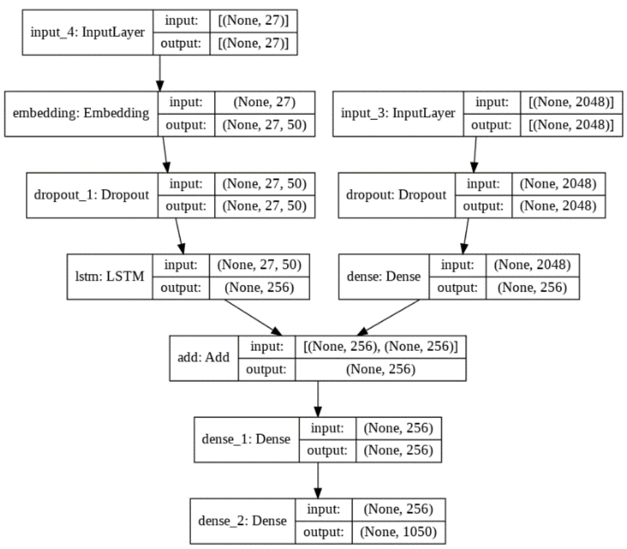

Figure 5. Merged encoder-decoder architecture.

E. Evaluation Metrics

We use four different evaluation metrics to evaluate our image captioning model. The evaluation metrics that we used are BLEU-n, ROUGE-L, METEOR, and CIDEr. Those metrics are commonly used metrics in evaluating image captioning. In evaluating the generated captions, candidate and references are used. Candidate refers to model generated caption and references refer to human annotated captions. Evaluation metrics work by comparing the candidate in terms of caption closeness to human generated sentences or semantic correctness [3]. The higher the score, the more related the prediction caption is to the original captions.

1) BLEU

BLEU [42] (Bilingual Evaluation Understudy) is a metric that defines the similarity between the predicted text and the references. BLEU considers n-grams (usually 1-4) instead of words and then matches the occurrence of the n-grams in the predicted caption to the references. The highest number of n-gram is 4 because it is found to be having the highest correlation with human generated captions [43]. In evaluating each text, BLEU doesn't pay any attention to syntactical correctness. If the generated caption is totally like the references, the score is given 1.0, if the generated caption is not at all similar, the score given is 0.0. The BLEU score can be calculated with Equation 2 [42].

$$BLEU = min(1, \frac{output - length}{reference - length})(\prod_{i=1}^{4} precision_i)^{\frac{1}{4}} \quad (2)$$

2) METEOR

METEOR [44] (Metric for Evaluation for Translation with Explicit Ordering) is a metric oriented in single-precision and word recall to address BLEU's flaws. This made METEOR better in semantic correlation and is more relevant to human judgements. METEOR metric calculates the accuracy, recall, and F-mean of each word, stem, and synonym matching. This calculating requires METEOR to use pre-defines set of alignments, specifically, WordNet thesaurus, to take word, stem, and synonyms in consideration.

3) ROUGE-L

ROUGE [45] (Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence) is a metric that matches the basic units such as n-grams, word sequence, and word pairs between the predicted caption and the references for evaluation. ROUGE-L is one of ROUGE's series evaluation methods, the other ROUGE methods are namely, ROUGE-n (n = 1,2,3,4, n represents the number of n-gram), ROUGE-W, ROUGE-L, ROUGE-W, and skip-bigram cooccurrence statistics (ROUGE-S). In this work, we use ROUGE-L that is based on the longest common subsequence (LCS) at sentence level that doesn't require a continuous matching of words.

4) CIDER

CIDEr [46] (Consensus-based Image Description Evaluation) considers each sentence are consisted of n-grams. These n-grams are then encoded, and the weight of each n-grams are calculated using term frequency-inverse document frequencies (T-IDF) between predicted caption and references to calculate cosine similarity score. Instead of treating each word in the sentence equally like BLEU, TF and IDF that work in restricting each other, help CIDEr to only focuses on important and significant words. To evaluate the generated caption, CIDEr changes all words in the predicted and reference sentences into their root or stem forms.

4. Result and Discussion

We trained the model based on the setup presented in Table II. The evaluation metric results can be seen in Table III. The evaluation obtained is the overall score for all test sets. From Table III, it can be concluded that Model #2 outperformed other models in all evaluation metrics.

TABLE III. IMAGE CAPTIONING EVALUATION SCORE USING BLEU-N, METEOR, ROUGE-L, AND CIDER ON TEST SET

Model	BLEU-1	BLEU-2	BLEU-3	Bleu-4	METEOR	ROUGE-L	CIDEr
#1	0.513376	0.323297	0.200862	0.128656	0.184614	0.358689	0.376793
#2	**0.527565**	**0.353696**	**0.227728**	**0.146192**	**0.184714**	**0.377379**	**0.393117**
#3	0.477856	0.282301	0.164416	0.095669	0.163133	0.331824	0.272908
#4	0.485392	0.253145	0.131586	0.065881	0.150266	0.343478	0.353591

TABLE IV. IMAGE CAPTIONING EVALUATION COMPARISON TO PREVIOUS INDONESIAN IMAGE CAPTIONING STUDIES

No	Dataset	Architecture	Total Images	Captions per image	BLEU-1	BLEU-2	BLEU-3	BLEU-4
1	Flickr FEEH-ID [19]	CNN-LSTM	8099	5	50.0	31.4	23.9	13.1
2	Flickr30k-IND version [43]	CNN-GRU	31783	5	36.7	17.8	6.7	2.0
3	MSCOCO and Flick30k [18]	ResNet101-LSTM with adaptive attention	180k	5	67.8	51.2	37.5	27.4
4	**Our modified indoor object dataset**	**Transformer**	**771**	**3**	**52.8**	**35.4**	**22.8**	**14.6**

Not only because Model #2 has the highest score on different evaluation metrics, though some of the images are not described correctly, Model #2 is still able to generate appropriate sentence that corresponds to the given image. Other models such as Model #1 and Model #3 failed to detect objects that are in the given image, while Model #4 generated jumbled of repeated words and failed to describe the image given to the model. Model #2 reaches the BLEU-1 of 0.527565, BLEU-2 score of 0.353696, BLEU-3 score of 0.227728, BLEU-4 score of 0.146192, METEOR score of 0.184714, ROUGE-L score of 0.377379, and CIDEr score of 0.393117. The highest metric scores obtained by Model #2 is resulted from the Transformer architecture with a batch size of 128, attention head number of 4, and a dropout value of 0.2.

To further analyze the performance of our novel indoor visual understanding image captioning model, we compare our model to other image captioning model in Indonesian language (as seen in Table IV). The metrics that we compare is BLEU, since BLEU-n is often used metrics in image captioning. The score is in percentage form. From the table, we can see that our model is comparable to other models and gives quite well performance with such a small dataset.

Table V shows few images and the generated captions by Model #2. As seen in the table, we can see that our model can generate decent image captions. From the generated caption it can be seen that the models are mostly able to generate captions that are sufficient and within context from the given images. The model is able to include object's name (*rak buku*/shelves, laptop), color (*coklat*/brown, *merah*/red, *hitam*/black), location (*di depan*/at the front, *di kanan*/on the right, *di kiri*/on the left), characteristics (*kaca*/glass), and its close surrounding. However, as seen in Table V, the model is still facing a bit struggle in getting the exact description on some objects. For instance, ”the big brown couch” in Image No. 1 should be single or small size. In image No. 3 model also failed to detect a man sitting on a red chair. This happens since our dataset is still limited for only ten different indoor objects and didn't include human. This causes the model unable to detect the human gender (man/woman).

5. Conclusion

This work is created with a goal of achieving visual understanding in indoor space. We compare two different methods namely Transformer and merged encoder-decoder by setting the hyper-parameter to get the best model. We applied both models on our modified dataset consisting of ten different objects collected from MSCOCO and newly created three Indonesian captions that may include object's name, color, position/location (viewer's point of view), characteristics, and its close surrounding for each of the images. From the results we obtained, we can conclude that Transformer with a batch size of 128, attention head of 4, and a dropout of 0.2 performs better in predicting the image caption with it reaching the highest score in all evaluation metrics; BLEU-1 score of 0.527565, BLEU-2 score of 0.353696, BLEU-3 score of 0.227728, BLEU-4 score of 0.146192, METEOR of 0.184714, ROUGE-L of 0.377379, and CIDEr of 0.393117.

This study aims to identify and provide caption on images taken indoors. For future research, more objects and images can be used, not only limited to indoor but also outdoor objects. This will benefit the captioning model to recognize more objects and help in visual understanding.

TABLE V. CAPTION GENERATION RESULT

No	Image	Generated Captions
1		di depan ada sebuah sofa besar berwarna coklat terletak di samping kanan ruangan. (translation: at the front there is a big brown couch placed on the right side.)
2		di depan terdapat rak buku berukuran sedang dengan banyak rak buku di atasnya. (translation: at the front there is a medium book shelf with a lot of book shelves on top.)
3		di depan ada seorang wanita sedang duduk di kursi berwarna merah. (translation: at the front there is a woman sitting on a red chair.)
4		di samping kiri ada oven kompor berwarna putih. (translation: on the left side there is a white range stove.)
5		di depan terdapat laptop berwarna hitam dengan layar menyala. (translation: at the front there is a black laptop with the screen on.)
6		di depan ada seorang pria sedang berdiri di depan kompor oven. (translation: at the front there is a man standing in front of a range stove.)

References

[1] C. Wang, Z. Zhou, and L. Xu, "An Integrative Review of Image Captioning Research," *Journal of Physics: Conference Series*, vol. 1748, no. 4, 2021.

[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*, vol. 1914, no. 1, 2021.

[3] H. Sharma, M. Agrahari, S. K. Singh, M. Firoj, and R. K. Mishra, "Image Captioning: A Comprehensive Survey," *2020 International Conference on Power Electronics and IoT Applications in Renewable Energy and its Control, PARC 2020*, pp. 325–328, 2020.

[4] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June, pp. 3156–3164, 2015.

[5] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter, pp. 2407–2415, 2015.

[6] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," *32nd International Conference on Machine Learning, ICML 2015*, vol. 3, pp. 2048–2057, 2015.

[7] F. Fang, H. Wang, and P. Tang, "Hanli Wang Department of Computer Science & Technology , Tongji University , Shanghai , P . R . China Key Laboratory of Embedded System and Service Computing , Ministry of Education ,," *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 1278–1282, 2018.

[8] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," *31st International Conference on Machine Learning, ICML 2014*, vol. 3, pp. 2012–2025, 2014.

[9] Z. Fang, J. Wang, X. Hu, L. Liang, Z. Gan, L. Wang, Y. Yang, and Z. Liu, "Injecting semantic concepts into end-to-end image captioning," 12 2021.

[10] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," 3 2016.

[11] P. Cao, Z. Yang, L. Sun, Y. Liang, M. Q. Yang, and R. Guan, "Image captioning with bidirectional semantic attention-based guiding of long short-term memory," *Neural Processing Letters*, vol. 50, pp. 103–119, 8 2019.

[12] L. Guo, J. Liu, J. Tang, J. Li, W. Luo, and H. Lu, "Aligning linguistic words and visual semantic units for image captioning," 8 2019.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.

[14] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, no. Mlm, pp. 4171–4186, 2019.

[15] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[16] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-Linear Attention Networks for Image Captioning," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 10 968–10 977, 2020.

[17] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-Octob, no. c, pp. 8927–8936, 2019.

[18] 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*, 2020.

[19] 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 - Proceedings*, 2019.

[20] W. Jeamwatthanachai, M. Wald, and G. Wills, "Indoor navigation by blind people: Behaviors and challenges in unfamiliar spaces and buildings," *British Journal of Visual Impairment*, vol. 37, no. 2, pp. 140–153, 2019.

[21] M. D. Zakir Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Computing Surveys*, vol. 51, no. 6, 2019.

[22] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From Show to Tell: A Survey on Image Captioning," pp. 1–22, 2021. [Online]. Available: http://arxiv.org/abs/2107.06912

[23] W. Wang, Z. Chen, and H. Hu, "Hierarchical attention network for image captioning," *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, pp. 8957–8964, 2019.

[24] L. Huang, W. Wang, J. Chen, and X. Y. Wei, "Attention on attention for image captioning," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-Octob, no. Iccv, pp. 4633–4642, 2019.

[25] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 10 677–10 686, 2019.

[26] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 3242–3250, 2017.

[27] C. Liu, J. Mao, F. Sha, and A. Yuille, "Attention correctness in neural image captioning," *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, pp. 4176–4182, 2017.

[28] W. Jiang, L. Ma, Y. G. Jiang, W. Liu, and T. Zhang, *Recurrent Fusion Network for Image Captioning*. Springer International Publishing, 2018, vol. 11206 LNCS.

[29] T. Yao, Y. Pan, Y. Li, and T. Mei, "Hierarchy parsing for image captioning," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-Octob, pp. 2621–2629, 2019.

[30] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 10 575–10 584, 2020.

[31] K. Shuster, S. Humeau, H. Hu, A. Bordes, and J. Weston, "Engaging image captioning via personality," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 12 508–12 518, 2019.

[32] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, vol. 1, pp. 2556–2565, 2018.

[33] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, "Microsoft COCO Captions: Data Collection and Evaluation Server," pp. 1–7, 2015. [Online]. Available: http://arxiv.org/abs/1504.00325

[34] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 74–93, may 2015. [Online]. Available: http://arxiv.org/abs/1505.04870

[35] H. Rashid, A. S. Al-Mamun, M. S. R. Robin, M. Ahasan, and S. M. Reza, "Bilingual wearable assistive technology for visually impaired persons," *1st International Conference on Medical Engineering, Health Informatics and Technology, MediTec 2016*, 2017.

[36] A. Gholamy, V. Kreinovich, and O. Kosheleva, "Why 70 / 30 or 80 / 20 Relation Between Training and Testing Sets : A Pedagogical Explanation," pp. 1–6, 2018.

[37] V. Maeda-Gutiérrez, C. E. Galván-Tejada, L. A. Zanella-Calzada, J. M. Celaya-Padilla, J. I. Galván-Tejada, H. Gamboa-Rosales, H. Luna-García, R. Magallanes-Quintanar, C. A. Guerrero Méndez, and C. A. Olvera-Olvera, "Comparison of convolutional neural network architectures for classification of tomato plant diseases," *Applied Sciences (Switzerland)*, vol. 10, no. 4, 2020.

[38] N. S. Keskar, J. Nocedal, P. T. P. Tang, D. Mudigere, and M. Smelyanskiy, "On large-batch training for deep learning: Generalization gap and sharp minima," *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, pp. 1–16, 2017.

[39] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," Tech. Rep., 2014.

[40] M. Tanti, A. Gatt, and K. P. Camilleri, "Where to put the image in an image caption generator," *Natural Language Engineering*, vol. 24, no. 3, pp. 467–489, 2018.

[41] D. Masters and C. Luschi, "Revisiting Small Batch Training for Deep Neural Networks," pp. 1–18, 2018. [Online]. Available: http://arxiv.org/abs/1804.07612

[42] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method

for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, vol. 371, no. 23. Morristown, NJ, USA: Association for Computational Linguistics, 2001, p. 311.

[43] A. A. Nugraha, A. Arifianto, and Suyanto, "Generating image description on Indonesian language using convolutional neural network and gated recurrent unit," *2019 7th International Conference on Information and Communication Technology, ICoICT 2019*, pp. 1–6, 2019.

[44] A. Lavie and A. Agarwal, "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments," *Proceedings of the Second Workshop on Statistical Machine Translation*, vol. 0, no. June, pp. 228–231, 2007. [Online]. Available: http://acl.ldc.upenn.edu/W/W05/W05-09.pdf{#}page=75

[45] C. Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," 2005. [Online]. Available: https://aclanthology.org/W04-1013

[46] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June, pp. 4566–4575, 2015.

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.

Royan Abida Nur Nayoan was born in Boyolali, Indonesia in 1999. She received her Bachelor of Computer Science in Informatics from Universitas Islam Indonesia, Yogyakarta, in 2019. Currently she is pursuing her master degree and is an assistant researcher at the Department of Informatics in Universitas Islam Indonesia. Her main area of research interests are deep learning and natural language processing.