

A study on attention-based deep learning architecture model for image captioning

Dhomas Hatta Fudholi, Umar Abdul Aziz Al-Faruq, Royan Abida N. Nayoan, Annisa Zahra

Department of Informatics, Universitas Islam Indonesia, Yogyakarta, Indonesia

Article Info

Article history:

Received Nov 20, 2022

Revised Mar 17, 2023

Accepted Mar 27, 2023

Keywords:

Attention

Encoder-decoder

Evaluation metrics

Image captioning

Transformer

ABSTRACT

Image captioning has been widely studied due to its ability in a visual scene understanding. Automatic visual scene understanding is useful for remote monitoring system and visually impaired people. Attention-based models, including transformer, are the current state-of-the-art architectures used in developing image captioning model. This study examines the works in the development of image captioning model, especially models that are developed based on attention mechanism. The architecture, the dataset, and the evaluation metrics analysis are done to the collected works. A general flow of image captioning model development is also presented. The literature search process carried out on Google Scholar. There are 36 literatures used in this study, including a specific image captioning development in Indonesian. It is done to take one point of view of image captioning development in a low resource language. Studies using transformer model generally achieves higher evaluation metric scores. In our finding, the highest evaluation scores on the consensus-based image description evaluation (CIDEr) c5 and c40 metrics are 138.5 and 140.5 respectively. This study gives a baseline on future development of image captioning model and brings the general concept of the image captioning development process including a picture of the development in low resource language.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Dhomas Hatta Fudholi

Department of Informatics, Universitas Islam Indonesia

Kaliurang St. Km. 14.5 Yogyakarta 55584, Indonesia

Email: hatta.fudholi@uii.ac.id

1. INTRODUCTION

Image captioning is the ability to describe the contents of an image in the form of sentences [1]. This ability requires methods from two fields of artificial intelligence, namely computer vision to understand the content of a given image and natural language processing (NLP) to convert the content in the image into sentence form. Due to the advancement of deep learning models in these two fields, image captioning has received a lot of attention in recent years. On the computer vision side, the improvement of the convolution neural network (CNN) architecture and object detection contributed to the improvement of the image captioning system. On the NLP side, more advanced sequential models, such as attention-based recurrent networks, result in more accurate text generation.

Most successful image captioning uses an encoder-decoder approach inspired by the sequence-to-sequence model for machine translation. This framework uses CNN + recurrent neural network (RNN), where CNN is used as an image encoder that extracts region-based visual features from the input image and RNN is used as a caption decoder to make sentences [2]. With the development of machine translation, a new architecture emerged, namely attention, which has become a state-of-the-art method in the NLP field and

makes image captioning models. Transformer [3] is an attention-based architecture that largely being adopted in the field.

This study aims to discover state-of-the-art method in image captioning, especially attention-based architecture, including transformer. We also take Indonesian as one of the low resource languages as a case study to give a picture of how mature the development of image captioning model in such language. An analysis of the model and architecture used, as well as the results of the study was carried out. Additionally, we develop a general concept that indicate what needs to be done, what techniques and data are typically used in image captioning research. The results of this study are expected to be a reference for the image captioning method in future research and provide recommendations for the best method for doing image captioning.

2. METHOD

2.1. Research flow

The systematic literature review presented in this paper has two main steps in its methodology. The first step is to formulate and conduct a search for related literatures. The second step is to formulate, conduct, and discuss the literature analysis from several points of view. Details of the methodological steps carried out will be explained in the next section.

2.2. Searching method

The searching process for related literatures was carried out in three parts. The process uses the help of the Google Scholar search engine. The searching process will select literatures that meets the following criterias: i) the scientific publication literature uses the attention mechanism model as the basis for doing image captioning; and ii) the literature provides details on the implementation of the method in image captioning task.

The first part of the searching process uses three main keywords: *image captioning*, *transformer models*, and *attention mechanism*. The scientific publication taken in this part is publications within the time frame between 2017 and 2020. This time frame is taken to depicts the growth and the development of the attention-based mechanism in image captioning task. The second part of the searching process is carried out with the same keywords as the first part, but the time frame is between 2021 and 2022. In this part, only top four cited open access publication is taken. The aim of the second part is to get some knowledge how attention-based model in image captioning still can be improved. The last part is slightly different from the previous parts. The third part focuses on the study of image captioning in Indonesian. *Indonesian image captioning* is used as the keyword. This part intends to search studies using Indonesian language datasets from 2017 to 2022 that use attention-based model including transformer.

2.3. Analytical method

Selected literatures are analyzed from several points of view. There are five analyzes to be carried out, namely architectural analysis, method analysis, dataset analysis, metric evaluation analysis, and general modeling model analysis of image captioning. The analysis was carried out with the aim of obtaining information that could support the growth of research in this study. The results of the analysis are presented in tabular form with a detailed explanation of the analysis.

3. RESULTS AND DISCUSSION

Using the keywords given in the literature search, 27 literatures were obtained in the first step. Of the 27 literatures, 25 literatures met the search criteria, and two literatures did not meet the search criteria. One literature does not meet the specified time range between 2017 to 2020 and one last literature does not use the attention mechanism model. In the second search process, we selected 4 literatures from 2021, and 4 others from 2022 that are on the first five pages of search results and can be accessed. While our third search resulted in 4 literatures that match our criteria.

Table 1 (see in Appendix) shows the comparison between the selected literatures according to the search criteria. The table provides a comparison between the architecture, the methods, and the evaluation metrics results. The evaluation metrics presented are bilingual evaluation understudy (BLEU), metric for evaluation for translation with explicit ordering (METEOR), recall-oriented understudy for gisting evaluation-longest common subsequence (ROUGE-L), consensus-based image description evaluation (CIDEr), and semantic propositional image caption evaluation (SPICE). The higher the score, the more accurate the predicted caption based on the referenced caption. The elaboration of the architecture, the *dataset*, and the evaluation metrics analysis are presented in the following section. The literature comparison for papers that used Indonesian dataset is presented in Table 2.

Table 2. Literature review result for Indonesian datasets

Literature	Architecture	Method	Evaluation Score
[34]	Attention	ResNet101 and LSTM	BLEU-1,2,3,4: 67.8, 51.2, 37.5, 27.4; CIDEr: 99.0
[35]	Attention	EfficientNet	Average of the BLEU train score: 73.39; Average of the BLEU validation score: 24.51
[36]	Transformer	Pretrained ResNet	BLEU-1,2,3,4: 56.00, 41.17, 29.42, 20.57; METEOR: 19.50; ROUGE-L: 44.16; CIDEr: 57.26
[37]	Transformer	Multi-head attention and CNN	Average BLEU-1,2,3,4: 78.05, 68.21, 61.89, 52.09

3.1. Architecture analysis

In this section, the elaboration of different approaches in developing image captioning model are presented. The analysis of the elaboration is grouped into two. The first one is the vanilla attention mechanism, and the second one is the transformer which uses a more specific attention mechanism called self-attention.

3.1.1. Attention mechanism

Of the image captioning methods that use the attention mechanism, many use a CNN encoder and an RNN or long short-term memory (LSTM) decoder. Among them are the hierarchical attention network (HAN) [7] which pays attention to semantic features of various levels that make it easy to predict different words based on different features, while the multivariate residual module (MRM) makes it easy to extract relevant relations from various features. There are also other methods that also use the same encoder and decoder, namely, attention on attention (AoA) [8], scene graph auto-encoder (SGAE) [12], Adaptive attention through visual sentinel [20], policy optimization gradient SPIDER [22], and recurrent fusion network (RFNet) [15]. In addition, research that utilizes spatial and semantic information using attention mechanisms, namely the graph convolutional network-long short-term memory (GCN-LSTM) [17], and spatial and channel wise attentions in a CNN (SCA-CNN) [21] which can score high when compared to state-of-the-art models. LSTM-P [5] uses an RNN-based language model that presents a novelty, namely the exploitation of the pointer mechanism to accommodate dynamic word generation through an RNN-based language model and word copying from the object being studied. hierarchy parsing (HIP) [6] which integrates a hierarchical structure into an image encoder. HIP functions as a feature refiner producing a rich and multi-level representation of the image. Unsupervised image captioning [16] uses an encoder, generator, and descriptor. In this method, CNN will encode the input image into a feature representation, then the LSTM as a generator will decoding the image representation into a sentence that describes the image content, while the LSTM discriminator is tasked with distinguishing the original sentence from the sentence generated by the model. The combination of top-down and bottom-up attention mechanisms [19] where bottom-up based on Faster R-CNN processes the image area and converts it into feature vectors, while top-down determines feature weighting. Research [23] overcomes the variation and ambiguity of image descriptions with the convolution technique. Research on updating the long short-term memory (LSTM-A) architecture [25] has succeeded in integrating attributes in CNNs + RNN framework image captioning. Having a proven performance in generating meaningful sentences and being very successful in advancing state-of-the-art, the attention mechanism is still being used in recent research [33]. Prophet attention [33] was introduced for calculating the ideal attention weights towards image region by using the future information.

According to our search, attention is rarely used in Indonesian image captioning from 2017 as we only found two papers that matched our criteria. To produce the next word, adaptive attention [34] was used to determine when and at which part of the image should be focused on by using translated Microsoft Common Objects in Context (MS COCO) and Flickr30k datasets. Research [35] applied visual attention mechanism to their model to produce a caption for the image that makes greater sense. As the result, their model was able to give a sensible and detailed caption in the local tourism domain.

3.1.2. Transformer

Transformer works as an encoder-decoder architecture that uses an attention mechanism. The captioning transformer (CT) study [18] was developed to overcome the problem of image captioning which is often developed using an LSTM decoder. Although good at remembering sequentially, LSTM has a complicated sequential problem in terms of timing. CT only has an attention module without a time dependency, so this model not only remembers sequence dependencies, but can also be trained in parallel.

Research [1] uses a spatial graph encoding transformer layer with a modified encoding transformer arrangement and an implicit decoding transformer layer which has a decoder layer and an LSTM layer in it to overcome the structure of the semantic unit of the image and each word in a different sentence. Research [4] improves image encoding and text prediction by using meshed transformer with memory to get low- and high-level features so that it can predict images that are not even in the training data. Multimodal transformer [2] is composed of an image encoder and a text decoder simultaneously capturing intra and inter-modal interactions

such as the relationship of words, objects, words in attention blocks that can produce captions accurately. Boosted transformer [9] utilizes semantic concepts (CGA) and visual features (VGA) to enhance the description of the resulting image. Personality-captions [13] uses TransResNet and a dataset that supports personality differentiation to produce image descriptions that are closer to humans. Conceptual dataset [14] was also developed using Inception-ResNetv2 as a feature extractor and transformer to perform image captioning. Another study with encoder and decoder transformer using object spatial relationship model [10] was built by explicitly including spatial relationship information between input objects detected from attention geometric. EnTangled transformer [11] was developed to exploit all semantic and spatial information from images.

Since various transformer-based models have achieved promising success on the image captioning task [31], recent research has still widely used it. dual-level collaborative transformer [26] was proposed to complement region and grid features for image captioning by applying intra-level fusion via comprehensive relation attention (CRA) and dual-way self attention (DWSA). Global enhanced transformer (GET) [27] makes it possible to obtain a more comprehensive global representation, which guide the decoder in creating a high-quality caption. Caption transformer (CPTR) [28], as a full transformer model, is capable of modeling global context information throughout encoder at every layer. Transformer-based semi-autoregressive model for image captioning, which keeps the autoregressive property in global and non-autoregressive property in local, tackles the heavy latency during inference issue that is caused by adopting autoregressive decoders [29]. Spatial and scale-aware transformer (S^2 transformer) [30] explores both low-level and high-level encoded features simultaneously in a scale-wise reinforcement module and learns pseudo regions by learning clusters in a Spatial-aware Pseudo-supervised module. Relational transformer (ReFormer) [31] was proposed to improve the quality of image captions by generating features that have relation information embedded, as well as explicitly expressing pair-wise relationships between images and their objects. While research [32] used a transformer-based architecture called attention-reinforced transformer to overcome the problem of cross entropy limiting diversity in image captioning.

For research with Indonesian dataset, we only found two paper that use transformer in their study, [36] and [37]. The result of research [37] showed that the implementation of the transformer architecture significantly exceeded the results of existing Indonesian image captioning research. In addition, the use of EfficientNet model obtains better results than InceptionV3. Research [36] has different approach, which use ResNet family as the base of visual feature extraction.

3.2. Dataset analysis

From the analysis conducted on existing studies, five main datasets were generally used for image captioning, namely conceptual captions, MS COCO, Flickr8K, Flickr30K, and a specially made dataset for local tourism domain. The composition of the four datasets and studies that utilize these datasets can be seen in Table 3. On Table 3, we can see the detail of the number of images in the training, the validation, and the testing dataset. Furthermore, we can see the number of annotations for each image in different dataset.

Table 3. Dataset composition

Dataset	Training	Validation	Testing	Annotations for each image	Literature
Conceptual	3,318,333	15,840	12,559	5	[14]
MS COCO	113,287	5,000	5,000	5	All literature that used English datasets
Flickr30K	29,783	1,000	1,000	5	[13], [20], [21], [24]
Flickr8K	6,000	1,000	1,000	5	[21]
Specially made (Local Tourism Domain)	1,356	340	-	1	[35]

3.2.1. Conceptual captions

Conceptual captions [14] were created using the Flume pipeline. This pipeline processes billions of internet pages in parallel. From these web pages, extraction, filtering, and pairing processes (images, descriptions) were carried out. The filtering and processing steps are divided into four, namely image-base filtering, text-based filtering, image & text-based filtering, and text transformation with hypernymization.

3.2.2. MS COCO

MS COCO [38] is a dataset from Microsoft COCO which is a large dataset containing object detection, segmentation, and captioning. This dataset has 328,000 images with a total of 2,500,000 labels, 80

object categories, and 91 object categories. This dataset is divided into training, validation, and testing data, as in Table 3, using Karpathy's splits [39].

The image captions dataset in MS COCO consists of two dataset collections. The first dataset, MS COCO c5, has five text references for each image in the MS COCO dataset training, validation, and testing. The second dataset, MS COCO c40, has 40 text references and randomly selects 5,000 images from the MS COCO testing dataset. MS COCO c40 builds on the many automated evaluation metrics that give results that achieve higher correlations than human judgments when given more references [40].

3.2.3. Flickr30K and flickr8K

Flickr30K [41] is a popular dataset used as a benchmark for text generation and retrieval. Flickr30k has 31,783 images focused on humans and animals, as well as a total of 158,915 English subtitles for these images to reference. While Flickr8K [42] has a total of 8000 images collected from Flickr. Each image in the dataset has five human-annotated captions.

3.2.4. Indonesian datasets

To translate the English MS COCO dataset into Indonesian, two methods were used: Google Translate [36], [37] and manual translation [34]. However, the results of the google translate translation are not very good, so research [34] used both of those two methods in their study to get a good Indonesian dataset. Study with a specific domain, requires a specially made dataset because it has not been available before. Research [35] collected a total of 1,696 local tourism-related images from Google search engines.

3.3. Evaluation metrics analysis

From the analyzed literature, five different evaluation metrics are commonly used to evaluate image captioning: BLEU, METEOR, ROUGE, CIDEr and SPICE. These metrics mainly measure the similarity between generated and reference captions through word overlap. Especially in SPICE, it uses "scene graph" to measure the similarity [43]. BLEU, METEOR, and ROUGE-L are evaluation models originally developed to assess the performance of Machine Translation. In recent years CIDEr and SPICE were developed specifically to evaluate image captions and showed more success than the previous ones [44].

The evaluation is done to measure the quality of a candidate caption c_i given a set of reference captions $S_i = \{s_{i1}, s_{i2}, \dots, s_{ij}\} \in S$. When the sentences are represented using sets of n -grams, $\omega_k \in \Omega$ is a set of one or more ordered words. For the candidate sentence $c_i \in C$, $h_k(s_{ij})$ or $h_k(c_i)$ denotes the number of times an n -gram ω_k occurs in a sentence s_{ij} .

BLEU calculates the n -gram overlap between candidate and reference texts to evaluates candidate texts. The BLEU score is calculated by the geometric mean of the modified n -gram score accuracy. The score is multiplied by a short penalty factor to give a "punishment" to short sentences so that the evaluation results are more representative [44]. The clipped n -gram precision between sentences is computed at the corpus level as shown in (1) [40]. In this case, k represents the set of possible n -grams of length n . While CP_n favors short sentences as it's a precision score. It is also used a brevity penalty as in (2) to favor a short sentence. Here, l_C represents the total length of candidate sentences c_i 's and l_S represents the corpus-level effective reference length. For the brevity penalty, we use the closest reference length whenever multiple references exist for a candidate sentence. To calculate the overall BLEU score, a weighted geometric mean of the individual n -gram precision is applied as in (3) where the values of N are 1, 2, 3, 4 and w_n is usually constant for all n .

$$CP_n(C, S) = \frac{\sum_i \sum_k \min(h_k(c_i), \max_{j \in m} h_k(s_{ij}))}{\sum_i \sum_k h_k(c_i)} \quad (1)$$

$$b(C, S) = \begin{cases} 1, & \text{if } l_C > l_S \\ e^{1-l_S/l_C}, & \text{if } l_C \leq l_S \end{cases} \quad (2)$$

$$BLEU_N(C, S) = b(C, S) \exp(\sum_{n=1}^N w_n \log CP_n(C, S)) \quad (3)$$

Metric for Evaluation for Translation with Explicit Ordering (METEOR) evaluates the candidate text based on overlapping unigrams between candidate and reference texts. This corresponds to a unigram based on meanings, exact and stemmed forms [44]. During calculating the alignment between the words in the candidate and reference sentences, the number of contiguous and identically ordered chunks of tokens in the sentence pair (ch) is minimized. This evaluation is conducted using the default parameters γ , α and θ . So, based on a set of alignments (m), the METEOR score is derived from the harmonic mean of precision (P_m) and recall (R_m) between the candidate and reference with the best score [40], see (4)-(8).

$$Pen = \gamma \left(\frac{ch}{m} \right)^\theta \quad (4)$$

$$F_{mean} = \frac{P_m R_m}{\alpha P_m + (1-\alpha) R_m} \quad (5)$$

$$P_m = \frac{|m|}{\sum_k h_k(c_i)} \quad (6)$$

$$R_m = \frac{|m|}{\sum_k h_k(s_{ij})} \quad (7)$$

$$METEOR = (1 - Pen) F_{mean} \quad (8)$$

ROUGE-L gives an evaluation by automatically comparing generated summaries with human reference summaries based on longest common subsequences (LCS). The LCS is between the generated summaries and the human result summaries. If high similarity is shown then the summary system quality is considered good [45]. Considering $l(C_i, S_{ij})$ is the length of the LCS between two sentences, $ROUGE_L$ is obtained by calculating the F-measure [40]. In (9)-(11) are used to calculate the metrics. R_l is recall and P_l is precision of LCS. While β is typically set to favor recall ($\beta = 1.2$).

$$R_l = \max_j \frac{l(c_i, s_{ij})}{|s_{ij}|} \quad (9)$$

$$P_l = \max_j \frac{l(c_i, s_{ij})}{|c_i|} \quad (10)$$

$$ROUGE_L(c_i, S_i) = \frac{(1+\beta^2)R_l P_l}{R_l + \beta^2 P_l} \quad (11)$$

Consensus-based image description evaluation (CIDEr) gives measurements to the consensus between candidate and reference texts using n-gram matching. Term frequency inverse document frequency (TF-IDF) weighting is calculated for n-grams that are common in all texts [38]. The frequency of occurrence of n-grams ω_k in the reference sentence for the candidate sentence c_i is denoted by $h_k(s_{ij})$ or $h_k(c_i)$. CIDEr calculates the TF-IDF weighting $g_k(s_{ij})$ for each n-gram ω_k by using [40] (12). I represents the set of all images in the dataset and Ω represents the vocabulary of all n-grams. The first term calculates the TF of each n-gram ω_k , while the second term calculates the rarity of ω_k by using its IDF. To calculate the $CIDEr_n$ score for n-grams of length n , we use the mean cosine similarity between candidate sentences and reference sentences. considering precision and recall the calculation is as in (13). The vector $g^n(c_i)$ represents all n-grams of length n and is formed by $g_k(c_i)$, while $\|g^n(c_i)\|$ represents their magnitude. Likewise, for $g^n(s_{ij})$. Grammatical properties and richer semantics were captured by longer n-grams. The scores from n-grams of various lengths were combined as (14). $w_n = 1/N$ is used for the uniform weights, with 4 as the value of N .

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{\omega_l \in \Omega} h_l(s_{ij})} \log \left(\frac{|I|}{\sum_{l \in I} \min(1, \sum_q h_k(spq))} \right) \quad (12)$$

$$CIDEr_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{g^n(c_i) \cdot g^n(s_{ij})}{\|g^n(c_i)\| \|g^n(s_{ij})\|} \quad (13)$$

$$CIDEr(c_i, S_i) = \sum_{n=1}^N w_n CIDEr_n(c_i, S_i) \quad (14)$$

Semantic propositional image caption evaluation (SPICE) estimates text quality by converting candidate and reference texts into semantic representations called “scene graphs” that encode objects, attributes, and relationships found in the text [44]. In this evaluation, we first define the parsing captions’ subtask to scene graphs. We parse a caption c into a scene graph, given a set of attribute types A , a set of object classes C , and a set of relation types R , as (15) [46]. $O(c) \subseteq C$ is the set of objects named in c , $E(c) \subseteq O(c) \times R \times O(c)$ is a hyper-edge set that represents the relationship between objects, and $K(c) \subseteq O(c) \times A$ is the set of attributes related to the object. For the second step, we calculate the F-score. We view the scene graph semantic relationships as a conjunction of logical propositions or tuples, to compare how closely two scene graphs, resemble one another. So, we have a function T that reads the scene graph and returns a logical

tuple as (16). Each tuple consists of one, two, or three components that, accordingly, represent the objects, attributes, and relations. We define the binary matching operator \otimes as the function that returns matching tuples in two scene graphs by looking at the semantic propositions in the scene graph as a set of tuples. Next, we define *SPICE*, recall *R*, and precision *P* as (17)-(19).

$$G(c) = \langle O(c), E(c), K(c) \rangle \quad (15)$$

$$T(G(c)) \triangleq O(c) \cup E(c) \cup K(c) \quad (16)$$

$$P(c, S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(c))|} \quad (17)$$

$$R(c, S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(S))|} \quad (18)$$

$$SPICE(c, S) = F_1(c, S) = \frac{2 \cdot P(c, S) \cdot R(c, S)}{P(c, S) + R(c, S)} \quad (19)$$

Of all the metrics used, namely BLEU-n, METEOR, ROUGE-L, CIDEr and SPICE, we selected four evaluation metrics that are used in almost all literature: BLEU-4, ROUGE-L, CIDEr, and METEOR. Table 4 shows the average value obtained in the evaluation metric of the transformer and attention architecture used in the literature discussed. However, calculations in Table 4 only include literature that contains c5 and c40 scores. It can be seen from the table that the transformer model gets a higher score for each evaluation metric than the vanilla attention mechanism model.

Table 4. Average score for each evaluation metric

Evaluation metric/Architecture	BLEU-4		ROUGE-L		CIDEr		METEOR	
	C5	C40	C5	C40	C5	C40	C5	C40
Transformer	39.68	72.00	59.27	74.58	129.49	131.59	29.32	38.76
Attention	36.41	66.95	56.94	71.98	116.06	118.14	27.48	36.53

3.3. General concept

At this stage of analysis, the steps for modeling the image captioning model that are commonly found are formulated. The general model formulation factors are taken from the dataset, preprocessing, architecture, and methods side, as well as the evaluation used. Figure 1 shows a generic flow chart based on the general modeling of the literature review.

Many of the datasets used are open-source. The dataset is available along with the distribution of the dataset for image captioning, such as Flickr30k with data totaling 30 thousand images, Flickr8k with data totaling 8 thousand images, MS COCO 180 thousand images, and Conceptual 3.3 million images. To see the distribution of the dataset, the number of images and the literature can be seen in Table 3.

Preprocessing is mostly done by following the steps of Karpathy [39]. To perform the preprocessing stage of the MS COCO dataset, you can follow the available source code [47]. Karpathy's preprocessing stage includes tokenizing, lowercase the text, then changing the word to "UNK" (unknown) or deleting words whose frequency is less than 5. In addition, some literatures also set varying caption lengths for MS COCO and Flickr30k [12], [20]. These words are then represented using GloVe or one-hot-encoding [2], [25].

The architecture and methods studied show various transformer and attention models. Many models rely on encoder and decoder frameworks because they are considered flexible [48]. The flexibility is not only in the designing the model architecture, but also the flexibility in implementing such architecture to different domain, for instance, molecular image captioning [49]. The role of the encoder is to extract features from the input image. While the decoder is useful for generating grammatically appropriate words. In this study, the most widely used encoder is the CNN variation model, while the most widely used decoders are LSTM, CNN, and RNN [2], [5], [7], [8], [42]. The method used is a modification of transformers and attention. The discussion of the method can be seen in section 3.1. The methods studied were obtained from 36 literatures. These methods are evaluated using the original dataset or from Karpathy [8].

The evaluations that are widely used are the BLEU, METEOR, ROUGE-L, CIDEr, and SPICE evaluation metrics. The evaluation score is widely used to evaluate the results of image captioning. To evaluate much of the literature we reviewed, we used the source code available from MS COCO [50] to calculate the evaluation metric. From all the literature that we reviewed, the evaluation metrics that are often used are BLEU-4, Rouge-L, CIDEr, and METEOR which are popular and known to have a strong correlation with human judgment [51], [52].

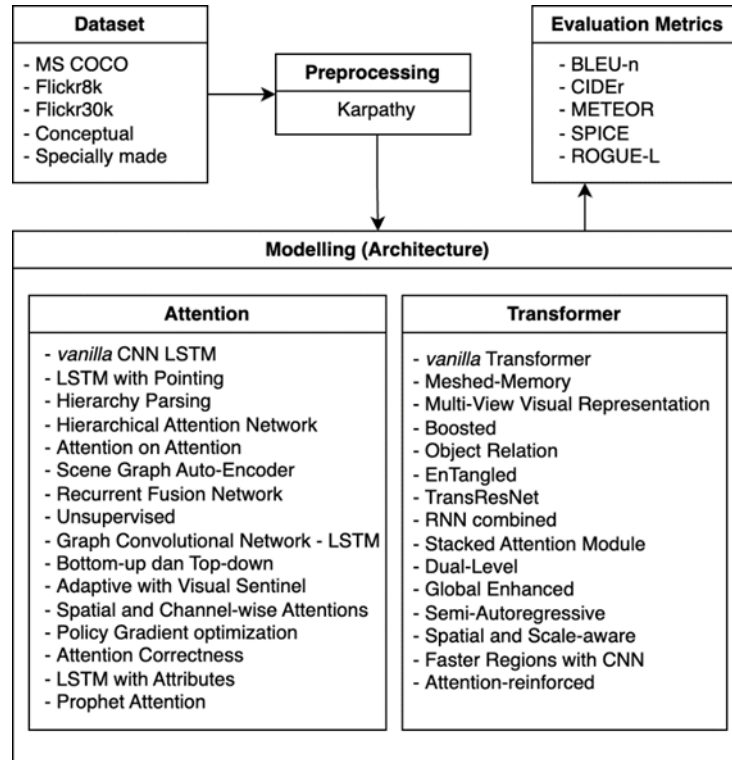


Figure 1. General flow diagram of image captioning model and concept

4. CONCLUSION

From this study on attention-based deep learning architecture model for image captioning, 36 literatures were found. With all models are attention-based architecture, half of the works listed in this study uses transformer. Five types of evaluation metrics were used across the works: BLEU, CIDEr, METEOR, ROUGE-L, and SPICE. BLEU still become the most used evaluation metrics for image captioning. From the analysis, it is known that on average, the transformer model obtains higher evaluation metric score at BLEU-4, CIDEr, ROUGE-L, and METEOR than the works with vanilla attention mechanism model. In the Indonesian language domain, as one example of a low resource language, only few works are found and most of them still rely on the common MSCOCO dataset as the base. However, there is an effort to create a novel dataset which is great to capture local culture in the caption. Finally, this study provides a foundation for the future development of the image captioning model and presents a general understanding of the process of developing image captions, including a representation of the process in low resource languages.

ACKNOWLEDGEMENT

The study is sponsored by SAME (Scheme for Academic Mobility and Exchange) 2022 program (Decree No. 3253/E4/DT.04.03/2022) from Directorate of Resources, Directorate General of Higher Education, Research and Technology, Ministry of Education, Culture, Research and Technology, Republic of Indonesia.

APPENDIX

Table 1. Literature review result

Literature	Architecture	Method	Result on MSCOCO online testing server
[1]	Transformer	Image transformer	BLEU-1,4(c5, c40): (81.2, 95.4), (39.6, 71.5); METEOR(c5, c40): (29.1, 38.4); ROUGE-L(c5, c40): (59.2, 74.5); CIDEr(c5, c40): (127.4, 129.6); SPICE(c5, c40): -
[4]	Transformer	Meshed-memory transformer	BLEU-1,2,3,4(c5, c40): (81.6, 96.0), (66.4, 90.8), (51.8, 82.7), (39.7, 72.8); METEOR(c5, c40): (29.4, 39.0); ROUGE-L(c5, c40): (59.2, 74.8); CIDEr(c5, c40): (129.3, 132.1); SPICE(c5, c40): -
[5]	Attention	Long short-term with pointing	BLEU-1,2,3,4: - ; METEOR: 23.4; ROUGE-L: - ; CIDEr: 88.3; SPICE: 16.6

Table 1. Literature review result (*continue*)

Literature	Architecture	Method	Result on MSCOCO online testing server
[6]	Attention	Hierarchy parsing	BLEU-1,2,3,4(c5, c40): (81.6, 95.9), (66.2, 90.4), (51.5, 81.6), (39.3, 71.0); METEOR(c5, c40): (28.8, 38.1); ROUGE-L(c5, c40): (59.0, 74.1); CIDEr(c5, c40): (127.9, 130.2); SPICE(c5, c40): -
[2]	Transformer	Multimodal Transformer with multi-view Visual representation	BLEU-1,2,3,4(c5, c40): (81.7, 95.6), (66.8, 90.5), (52.4, 82.4), (40.4, 72.2); METEOR(c5, c40): (29.4, 38.9); ROUGE-L(c5, c40): (59.6, 75.0); CIDEr(c5, c40): (130.0, 130.9); SPICE(c5, c40): -
[7]	Attention	Hierarchical attention network	BLEU-1,2,4(c5, c40): (80.4, 94.5), (63.8, 87.7), (36.5, 66.8); METEOR(c5, c40): (27.4, 36.1); ROUGE-L(c5, c40): (57.3, 71.9); CIDEr(c5, c40): (115.2, 118.2); SPICE(c5, c40): -
[8]	Attention	Attention on attention	BLEU-1,2,3,4(c5, c40): (81.0, 95.0), (65.8, 89.6), (51.4, 81.3), (39.4, 71.2); METEOR(c5, c40): (29.1, 38.5); ROUGE-L(c5, c40): (58.9, 74.5); CIDEr(c5, c40): (126.9, 129.6); SPICE(c5, c40): -
[9]	Transformer	Boosted transformer	BLEU-1,2,3,4(c5, c40): (80.5, 94.7), (65.2, 88.8), (50.6, 80.1), (38.6, 69.6); METEOR(c5, c40): (28.8, 37.9); ROUGE-L(c5, c40): (58.5, 73.5); CIDEr(c5, c40): (125.0, 126.8); SPICE(c5, c40): -
[10]	Transformer	Object relation transformer	BLEU-1,4: 80.5, 38.6; METEOR: 28.7; ROUGE-L: 58.4; CIDEr: 128.3; SPICE: 22.6
[11]	Transformer	EnTangled transformer	BLEU-1,2,3,4(c5, c40): (81.2, 95.0), (65.5, 89.0), (50.9, 80.4), (38.9, 70.2); METEOR(c5, c40): (28.6, 38.0); ROUGE-L(c5, c40): (58.6, 73.9); CIDEr(c5, c40): (122.1, 124.4); SPICE(c5, c40): -
[12]	Attention	Scene graph auto-encoder	BLEU-4(c5, c40): (38.5, 69.7); METEOR(c5, c40): (28.2, 37.2); ROUGE-L(c5, c40): (58.6, 73.6); CIDEr(c5, c40): (123.8, 126.5); SPICE(c5, c40): -
[13]	Transformer	TransResNet	BLEU-1,4: 79.3, 36.4; METEOR: -; ROUGE-L: 57.5; CIDEr: 124.0; SPICE: 21.2
[14]	Transformer	RNN and transformer	BLEU-1,2,3,4: -; METEOR: -; ROUGE-L: 0.336; CIDEr: 1.676; SPICE: 0.257
[15]	Attention	Recurrent fusion network	BLEU-1,2,3,4(c5, c40): (80.4, 95.0), (64.9, 89.3), (50.1, 90.1), (38.0, 69.2); METEOR(c5, c40): (28.2, 37.2); ROUGE-L(c5, c40): (58.2, 73.1); CIDEr(c5, c40): (122.9, 125.1); SPICE(c5, c40): -
[16]	Attention	unsupervised learning	BLEU-1,2,3,4: 58.9, 40.3, 27.0, 18.6; METEOR: 18.6; ROUGE-L: 43.1; CIDEr: 54.9; SPICE: 11.1
[17]	Attention	Graph convolutional networks-LSTM	BLEU-2,3,4(c5, c40): (65.5, 89.3), (50.8, 80.3), (38.7, 68.7); METEOR(c5, c40): (28.5, 37.6); ROUGE-L(c5, c40): (58.5, 73.4); CIDEr(c5, c40): (125.3, 126.5); SPICE(c5, c40): -
[18]	Transformer	Stacked attention modules	BLEU-1,2,3,4: 73.0, 56.9, 43.6, 33.3; METEOR: -; ROUGE-L: 54.8; CIDEr: 108.1; SPICE: -
[19]	Attention	Bottom-up and top-down attention	BLEU-1,2,3,4(c5, c40): (80.2, 95.2), (64.1, 88.8), (49.1, 79.4), (36.9, 68.5); METEOR(c5, c40): (27.6, 36.7); ROUGE-L(c5, c40): (57.1, 72.4); CIDEr(c5, c40): (117.9, 120.5); SPICE(c5, c40): (21.5, 71.5)
[20]	Attention	Adaptive attention model with visual sentinel	BLEU-1,2,3,4(c5, c40): (0.748, 0.920), (0.584, 0.845), (0.444, 0.744), (0.336, 0.637); METEOR(c5, c40): (0.264, 0.359); ROUGE-L(c5, c40): (0.550, 0.705); CIDEr(c5, c40): (1.042, 1.059); SPICE(c5, c40): -
[21]	Attention	Spatial and channel-wise attentions in a CNN	BLEU-1,2,3,4(c5, c40): (71.2, 89.4), (54.2, 80.2), (40.4, 69.1), (30.2, 57.9); METEOR(c5, c40): (24.4, 33.1); ROUGE-L(c5, c40): (52.4, 67.4); CIDEr(c5, c40): (91.2, 92.1); SPICE(c5, c40): -
[22]	Attention	Policy gradient optimization of SPIDEr	BLEU-1,2,3,4: 0.743, 0.578, 0.433, 0.322; METEOR: 0.251; ROUGE-L: 0.544; CIDEr: 1.000; SPICE: -
[23]	Attention	Convolutional image captioning	BLEU-1,2,3,4(c5, c40): (0.715, 0.896), (0.545, 0.805), (0.408, 0.693), (0.304, 0.582); METEOR(c5, c40): (0.246, 0.333); ROUGE-L(c5, c40): (0.525, 0.673); CIDEr(c5, c40): (0.910, 0.914); SPICE(c5, c40): -
[24]	Attention	Attention correctness	BLEU-3,4: 38.0, 28.1; METEOR: 23.01; ROUGE-L: -; CIDEr: -; SPICE: -
[25]	Attention	Long short-term memory with attributes	BLEU-1,2,3,4(c5, c40): (78.7, 93.7), (62.7, 86.7), (47.6, 76.5), (35.6, 65.2); METEOR(c5, c40): (27.0, 35.4); ROUGE-L(c5, c40): (56.4, 70.5); CIDEr(c5, c40): (116.0, 118.0); SPICE(c5, c40): -
[26]	Transformer	Dual-level collaborative transformer	BLEU-1,2,3,4(c5, c40): (82.4, 96.6), (67.4, 91.7), (52.8, 83.8), (40.6, 74.0); METEOR(c5, c40): (29.8, 39.6); ROUGE-L(c5, c40): (59.8, 75.3); CIDEr(c5, c40): (133.3, 135.4); SPICE: -
[27]	Transformer	Global enhanced transformer	BLEU-1,2,3,4(c5, c40): (81.6, 96.1), (66.5, 90.9), (51.9, 82.8), (39.7, 72.9); METEOR(c5, c40): (29.4, 38.8); ROUGE-L(c5, c40): (59.1, 74.4); CIDEr(c5, c40): (130.3, 132.5); SPICE: -
[28]	Transformer	Caption transformer	BLEU-1,2,3,4(c5, c40): (81.8, 95.0), (66.5, 89.4), (51.8, 80.9), (39.5, 70.8); METEOR(c5, c40): (29.1, 38.3); ROUGE-L(c5, c40): (59.2, 74.4); CIDEr(c5, c40): (125.4, 127.3); SPICE: -
[29]	Transformer	Semi-autoregressive transformer	BLEU-1,2,3,4(c5, c40): (80.3, 94.5), (64.4, 87.9), (49.2, 78.2), (37.0, 67.2); METEOR(c5, c40): (28.2, 37.0); ROUGE-L(c5, c40): (57.8, 72.6); CIDEr(c5, c40): (121.5, 124.1); SPICE: -
[30]	Transformer	Spatial and scale-aware transformer	BLEU-1,2,3,4(c5, c40): (82.2, 96.5), (67.0, 91.4), (52.4, 83.3), (40.1, 73.5); METEOR(c5, c40): (29.6, 39.3); ROUGE-L(c5, c40): (59.5, 75.0); CIDEr(c5, c40): (132.6, 135.0); SPICE: -
[31]	Transformer	Faster regions with convolutional neural networks	BLEU-1,4(c5, c40): (82.0, 96.7), (40.1, 73.2); METEOR(c5, c40): (29.8, 39.5); ROUGE-L(c5, c40): (59.9, 75.2); CIDEr(c5, c40): (129.9, 132.8); SPICE: -

Table 1. Literature review result (*continue*)

Literature	Architecture	Method	Result on MSCOCO online testing server
[32]	Transformer	Attention-reinforced transformer	BLEU-1,2,3,4(c5, c40): (83.4, 97.4), (68.8, 93.0), (54.3, 85.6), (42.0, 76.1); METEOR(c5, c40): (30.6, 40.4); ROUGE-L(c5, c40): (60.8, 76.4); CIDEr(c5, c40): (138.5, 140.5); SPICE: -
[33]	Attention	Prophet attention	BLEU-1,2,3,4(c5, c40): (81.8, 96.3), (66.5, 91.2), (51.9, 83.2), (39.8, 73.3); METEOR(c5, c40): (29.6, 39.3); ROUGE-L(c5, c40): (59.4, 75.1); CIDEr(c5, c40): (130.4, 133.7); SPICE: -





REFERENCES

- [1] S. He, W. Liao, H. R. Tavakoli, M. Yang, B. Rosenhahn, and N. Pugeault, "Image captioning through image transformer," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12625 LNCS, pp. 153–169, 2021, doi: 10.1007/978-3-030-69538-5_10.
- [2] J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal transformer with multi-view visual representation for image captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4467–4480, 2020, doi: 10.1109/TCSVT.2019.2947482.
- [3] A. Vaswani *et al.*, "Attention is all you need," in *Proceedings of the 31st Conference on Neural Information Processing Systems*, Dec. 2017, pp. 5998–6008, doi: 10.48550/arXiv.1706.03762.
- [4] 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. 10575–10584, 2020, doi: 10.1109/CVPR42600.2020.01059.
- [5] Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei, "Pointing novel objects in image captioning," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 12489–12498, 2019, doi: 10.1109/CVPR.2019.01278.
- [6] 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-Oct., pp. 2621–2629, 2019, doi: 10.1109/ICCV.2019.00271.
- [7] 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, doi: 10.1609/aaai.v33i01.33018957.
- [8] 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-Oct., pp. 4633–4642, 2019, doi: 10.1109/ICCV.2019.00473.
- [9] J. Li, P. Yao, L. Guo, and W. Zhang, "Boosted transformer for image captioning," *Applied Sciences (Switzerland)*, vol. 9, no. 16, pp. 1–15, 2019, doi: 10.3390/app9163260.
- [10] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," *Advances in Neural Information Processing Systems*, vol. 32, no. NeurIPS, pp. 1–11, 2019.
- [11] 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-Oct., no. C, pp. 8927–8936, 2019, doi: 10.1109/ICCV.2019.00902.
- [12] 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. 10677–10686, 2019, doi: 10.1109/CVPR.2019.01094.
- [13] 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. 12508–12518, 2019, doi: 10.1109/CVPR.2019.01280.
- [14] P. Sharma, N. Ding, S. Goodman, and R. Soicrut, "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, doi: 10.18653/v1/p18-1238.
- [15] W. Jiang, L. Ma, Y. G. Jiang, W. Liu, and T. Zhang, "Recurrent fusion network for image captioning," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11206 LNCS, pp. 510–526, 2018, doi: 10.1007/978-3-030-01216-8_31.
- [16] Y. Feng, L. Ma, W. Liu, and J. Luo, "Unsupervised image captioning," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 4120–4129, 2019, doi: 10.1109/CVPR.2019.00425.
- [17] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11218 LNCS, pp. 711–727, 2018, doi: 10.1007/978-3-030-01264-9_42.
- [18] X. Zhu, L. Li, J. Liu, H. Peng, and X. Niu, "Captioning transformer with stacked attention modules," *Applied Sciences (Switzerland)*, vol. 8, no. 5, 2018, doi: 10.3390/app8050739.
- [19] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, 2018, doi: 10.1109/CVPR.2018.00636.
- [20] 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-Jan., pp. 3242–3250, 2017, doi: 10.1109/CVPR.2017.345.
- [21] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," *Proceedings-30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Jan., pp. 6298–6306, 2017, doi: 10.1109/CVPR.2017.667.
- [22] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of SPIDEr," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-Oct., pp. 873–881, 2017, doi: 10.1109/ICCV.2017.100.
- [23] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 5561–5570, 2018, doi: 10.1109/CVPR.2018.00583.
- [24] 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, doi: 10.1609/aaai.v31i1.11197.
- [25] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-October, pp. 4904–4912, 2017, doi: 10.1109/ICCV.2017.524.




- [26] Y. Luo *et al.*, “Dual-level collaborative transformer for image captioning,” *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, vol. 3B, no. 3, pp. 2286–2293, doi: 10.1609/aaai.v35i3.16328.
- [27] J. Ji *et al.*, “Improving image captioning by leveraging intra- and inter-layer global representation in transformer network,” *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, vol. 2B, pp. 1655–1663, 2021, doi: 10.1609/aaai.v35i2.16258.
- [28] W. Liu, S. Chen, L. Guo, X. Zhu, and J. Liu, “CPTR: Full transformer network for image captioning,” *Computer Science > Computer Vision and Pattern Recognition*, pp. 1–5, 2021, doi: 10.48550/arXiv.2101.10804.
- [29] Y. Zhou, Y. Zhang, Z. Hu, and M. Wang, “Semi-autoregressive transformer for image captioning,” *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2021-Oct., pp. 3132–3136, 2021, doi: 10.1109/ICCVW54120.2021.00350.
- [30] P. Zeng, H. Zhang, J. Song, and L. Gao, “S2 transformer for image captioning,” *IJCAI International Joint Conference on Artificial Intelligence*, pp. 1608–1614, 2022, doi: 10.24963/ijcai.2022/224.
- [31] X. Yang, Y. Liu, and X. Wang, “ReFormer: the relational transformer for image captioning,” in *Proceedings of the 30th ACM International Conference on Multimedia*, Oct. 2022, pp. 5398–5406, doi: 10.1145/3503161.3548409.
- [32] Z. Wang, S. Shi, Z. Zhai, Y. Wu, and R. Yang, “ArCo: Attention-reinforced transformer with contrastive learning for image captioning,” *Image and Vision Computing*, vol. 128, p. 104570, Dec. 2022, doi: 10.1016/j.imavis.2022.104570.
- [33] F. Liu, X. Ren, X. Wu, W. Fan, Y. Zou, and X. Sun, “Prophet attention: Predicting attention with future attention for image captioning,” *Computer Science > Computer Vision and Pattern Recognition*, 2022, doi: 10.48550/arxiv.2210.10914.
- [34] M. R. S. Mahadi, A. Arifianto, and K. N. Ramadhani, “Adaptive attention generation for Indonesian image captioning,” in *2020 8th International Conference on Information and Communication Technology (ICOICT)*, Jun. 2020, pp. 1–6, doi: 10.1109/ICOICT49345.2020.9166244.
- [35] D. H. Fudholi *et al.*, “Image captioning with attention for smart local tourism using efficientNet,” *IOP Conference Series: Materials Science and Engineering*, vol. 1077, no. 1, p. 012038, 2021, doi: 10.1088/1757-899x/1077/1/012038.
- [36] R. Mulyawan, A. Sunyoto, and A. H. Muhammad, “Automatic Indonesian image captioning using CNN and transformer-based model approach,” *ICOIACT 2022-5th International Conference on Information and Communications Technology: A New Way to Make AI Useful for Everyone in the New Normal Era, Proceeding*, pp. 355–360, 2022, doi: 10.1109/ICOIACT55506.2022.9971855.
- [37] U. A. A. Al-Faruq and D. H. Fudholi, “EfficientNet-Transformer for image captioning in Bahasa,” *Vii International Conference “Safety Problems of Civil Engineering Critical Infrastructures” (Speeci2021)*, vol. 2701, p. 020037, 2023, doi: 10.1063/5.0118155.
- [38] T.-Y. Lin *et al.*, “Microsoft COCO: common objects in context,” in *In: European conference on computer vision*, May 2014, pp. 740–755, doi: 10.1007/978-3-319-10602-1_48.
- [39] 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, 2017, doi: 10.1109/TPAMI.2016.2598339.
- [40] X. Chen *et al.*, “Microsoft COCO captions: Data collection and evaluation server,” *Computer Science > Computer Vision and Pattern Recognition*, 2015, doi: 10.48550/arXiv.1504.00325.
- [41] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, Dec. 2014, doi: 10.1162/tacl_a_00166.
- [42] S. He and Y. Lu, “A modularized architecture of multi-branch convolutional neural network for image captioning,” *Electronics*, vol. 8, no. 12, p. 1417, Nov. 2019, doi: 10.3390/electronics8121417.
- [43] Y. Cui, G. Yang, A. Veit, X. Huang, and S. Belongie, “Learning to evaluate image captioning,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 5804–5812, doi: 10.1109/CVPR.2018.00608.
- [44] N. Sharif, L. White, M. Bennamoun, and S. A. A. Shah, “NNEval: Neural network based evaluation metric for image captioning,” in *ECCV 2018: Computer Vision – ECCV 2018*, 2018, pp. 39–55, doi: 10.1007/978-3-030-01237-3_3.
- [45] A. Cohan and N. Goharian, “Revisiting summarization evaluation for scientific articles,” *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pp. 806–813, 2016, doi: 10.48550/arXiv.1604.00400.
- [46] P. Anderson, B. Fernando, M. Johnson, and S. Gould, “SPICE: Semantic propositional image caption evaluation,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9909 LNCS, pp. 382–398, 2016, doi: 10.1007/978-3-319-46454-1_24.
- [47] Karpathy, “GitHub-karpathy/neuraltalk: NeuralTalk is a Python+numpy project for learning multimodal recurrent neural networks that describe images with sentences,” *GitHub*, 2015. <https://github.com/karpathy/neuraltalk> (accessed May 27, 2023).
- [48] R. Staniute and D. Šešok, “A systematic literature review on image captioning,” *Applied Sciences (Switzerland)*, vol. 9, no. 10, pp. 1–20, 2019, doi: 10.3390/app9102024.
- [49] J. Yi *et al.*, “MICER: a pre-trained encoder-decoder architecture for molecular image captioning,” *Bioinformatics (Oxford, England)*, vol. 38, no. 19, pp. 4562–4572, 2022, doi: 10.1093/bioinformatics/btac545.
- [50] Tylin, “GitHub-tylin/coco-caption,” *GitHub*, 2018. <https://github.com/tylin/coco-caption> (accessed May 27, 2023).
- [51] N. Aafaq, A. Mian, W. Liu, S. Z. Gilani, and M. Shah, “Video description: A survey of methods, datasets, and evaluation metrics,” *ACM Computing Surveys*, vol. 52, no. 6, pp. 1–37, 2019, doi: 10.1145/3355390.
- [52] Y. Graham, “Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE,” *Conference Proceedings-EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, pp. 128–137, 2015, doi: 10.18653/v1/d15-1013.

BIOGRAPHIES OF AUTHORS



Dthomas Hatta Fudholi     is an Associate Professor at the Department of Informatics, Universitas Islam Indonesia, Yogyakarta, Indonesia. He earned his Ph.D. in Computer Science and IT in 2016 from La Trobe University, Melbourne, Australia. He explores data science and deep learning methods to support knowledge base development and various useful applications. He can be contacted at email: hatta.fudholi@uii.ac.id.






Umar Abdul Aziz Al-Faruq    is a freshgraduate from Islamic University of Indonesia. He earned his bachelor's degree in informatics. He gained a strong foundation in computer science and programming. Some of his projects are related to Artificial Intelligent and web development. He can be contacted at email: umaralfaruq597@gmail.com.



Royan Abida Nur Nayoan    is a Research and Development Engineer in the field of Artificial Intelligence, specializing in computer vision and natural language processing. With a Master's degree in Informatics from Universitas Islam Indonesia, she is currently focusing on developing new AI applications to enhance company products. She has contributed to the field of AI through research publications and presentations. With a passion for AI, she is dedicated to staying up-to-date with the latest advancements in the field. She can be contacted at email: royananayoan@gmail.com.



Annisa Zahra    is currently a master's student studying Artificial Intelligence at Universitas Gadjah Mada, Indonesia. She obtained her bachelor's degree in informatics in 2021 from Universitas Islam Indonesia, Indonesia. She has done several machine learning projects in the domain of natural language processing. She can be contacted at email: az.annisazahra@gmail.com.