Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 7, No. 1, February, Pp. 91-98

A study on visual understanding image captioning using different word embeddings and CNN-based feature extractions

Dhomas Hatta Fudholi*[1], Annisa Zahra[2], Royan Abida N. Nayoan[3]
Department of Informatics, Universitas Islam Indonesia, Indonesia[1]
Bachelor Program in Informatics, Universitas Islam Indonesia, Indonesia[2]
Master Program in Informatics, Universitas Islam Indonesia, Indonesia[3]

Article Info
Keywords:
BLEU, CNN-based, Deep Learning, Feature Extraction, Image Captioning, LSTM, Word Embedding

Article history:
Received: January 19, 2022
Accepted: February 25, 2022
Published: March 10, 2022

Cite:
D. H. Fudholi, A. Zahra, and R. A. N. Nayoan, "A Study on Visual Understanding Image Captioning using Different Word Embeddings and CNN-Based Feature Extractions", *KINETIK*, vol. 7, no. 1, pp. 91-98, Feb. 2022.
https://doi.org/10.22219/kinetik.v7i1.1394

*Corresponding author.
Dhomas Hatta Fudholi
E-mail address:
hatta.fudholi@uii.ac.id

Abstract
Image captioning is a task that provides a description of an image in natural language. Image captioning can be used for a variety of applications by giving such visual understanding, such as image indexing and virtual assistants. Since there are many different Deep Learning architecture and setup, we tried to highlight few named architectures and find the best setup in the area. In this research, we compared the performance of three different word embeddings, namely, GloVe, Word2Vec, FastText and six CNN-based feature extraction architectures such as, Inception V3, InceptionResNet V2, ResNet152 V2, EfficientNet B3 V1, EfficientNet B7 V1, and NASNetLarge which then will be combined with LSTM as the decoder to perform image captioning. We used ten different household objects (bed, cell phone, chair, couch, oven, potted plant, refrigerator, sink, table, and tv) that were obtained from MSCOCO dataset to develop the model. Then, we created five new captions in Bahasa Indonesia for the selected images. The captions contain details about the name, the location, the color, the size, and the characteristics of an object and its surrounding area. In our 18 experimental models, we used different combination of the word embedding and CNN-based feature extraction architecture, along with LSTM to train the model. As the result, the model that used the combination of Word2Vec + NASNetLarge performed better in generating captions based on BLEU-4 metric.

1. Introduction
The task for providing a description of an image in natural language is called image captioning [1]. In image captioning, a description generation model should not only capture the objects/scenes present in an image, but also be capable of depicting how those objects/scenes relate to each other [2]. There are several applications of image captioning, including recommendations in editing applications, usage with virtual assistants, image indexing, for people with visual impairments and also for social media, and many other natural language processing-based applications [3]. This task also can be helpful to enhance the accuracy of search engines, develop and enhance new image datasets, optimize the operation of Google Photos and other systems, and to improve self-driving vehicles' optical system analysis [4].

Bahasa Indonesia is the official language of Indonesia. Since Indonesia has the fourth largest population in the world, Bahasa Indonesia is one of the world's most widely spoken languages [5]. Therefore, it is essential to generate image captions in Bahasa Indonesia. Several studies on image captioning using Bahasa Indonesia have been carried out previously. [6] used translated Flickr30K in Bahasa Indonesia with pre-trained Inception V3 stacked with Gated Recurrent Unit (GRU) as the experimental model.

Many deep learning-based image captioning methods use encoder-decoder frameworks. In order to extract image features, a Convolutional Neural Network (CNN)-based architecture is used on the encoder side. There are various types of CNN architectures used for image captioning tasks, like Inception [7] and NASNet [8]. While for the decoder, the caption can be generated by using Long Short-Term Memory (LSTM) method [9].

In this study, we attempt to explore image captioning in Bahasa Indonesia by using several household objects images from the MSCOCO dataset [10]. We use several different word embeddings, such as GloVe, Word2Vec, and FastText to represent the words. Here, we use Long Short-Term Memory (LSTM) as the decoder to which then will be combined with a variety of deep learning architectures that will work in extracting features. The deep learning models we use are namely, Inception V3, InceptionResNet V2, ResNet152 V2, EfficientNet B3 V1, EfficientNet B7 V1, and NASNetLarge. To evaluate the model, we use BLEU-n, one of the popular language translation evaluation metrics.

Various studies on image captioning have been carried out by researchers using various datasets and different methods. MS COCO [11], [12] and Flickr [13], [14] are two English datasets that are widely used in previous studies.

Some studies even used both MS COCO and Flickr datasets [15], [16]. While several research used the translated version of MS COCO and Flickr to other language, such as Bahasa Indonesia [6], [17].

In both encoder and decoder parts, different deep learning architectures have been employed for feature extraction and caption generation. Inception-v3 [7], NASNet [8], VGG-16 [18], and ResNet50 [9] are the examples of some feature extraction architectures that have been used in earlier research. The work in [19] compared the performance of VGG19 and ResNet101 as encoders using the same image captioning model. As the results, image captioning model with ResNet got higher BLEU-4 score and by using ResNet, the model could achieve a comparable score with the VGG-based model with less training epochs. In terms of caption generation, several different works have utilized different architectures as well, including GRU [6] and LSTM [20]. Gaining vector representations can be done using some word embedding methods, such as Glove [21], FastText [22], and Word2Vec [23]. The work in [24] compared Glove and Word2Vec and the results showed that in that case, GloVe embeddings are more suitable than Word2Vec, but both of them were succeeded in improving the quality of the model. While for the evaluation, some common metrics that are usually applied are BLEU [13], METEOR [25], and CIDEr [17].

In this study, we attempt to compare several different word embeddings and deep learning-based feature extraction architectures for image captioning task. We use dataset which consist of some images from the MS COCO dataset for ten different household items (bed, cell phone, chair, couch, oven, potted plant, refrigerator, sink, table, tv) and are then captioned manually in Bahasa Indonesia. Each image in our dataset is given five captions and the five captions are different sentences. Previous study has also used some different household objects from MS COCO and were also captioned manually using Bahasa Indonesia, three captions are added for each image [26]. The study applied Inception-v3 and LSTM architecture, along with GloVe as the word embedding to train the model. As the results, their model was able to generate caption well.

2. Research Method

The methodology used in this study is a sequence of data collection, data preparation, image captioning model, and model evaluation. Each step is explained as follows.

2.1 Data Collection

In this study, we use data from the Microsoft Common Objects in Context (MS COCO) dataset. MS COCO is a dataset that detects and segments everyday objects in the natural environment [10]. At this point, we will use ten common household objects to develop our image captioning model. These ten objects are bed, cell phone, chair, couch, oven, potted plant, refrigerator, sink, table, and tv. The total images that we selected are 773 images on all ten object categories (80 cell phone, 78 potted plant, 80 oven, 56 refrigerator, 80 tv, 80 table, 80 sink, 80 couch, 79 chair, and 80 bed). The examples of the selected images are shown in Figure 1.


Image_1


Image_2

Figure 1. Examples of Selected Images

Instead of using the captions provided by MS COCO, we added five new captions in Bahasa Indonesia for each image. Each of the sentences are written to simulate how different persons describes the images. The captions may contain details regarding object's name, location, color, size, distict characteristics or its surrounding area. The examples of the caption for our collected images are presented in Table 1.

Table 1. An Example of Table Caption

Image	Caption	Translated Caption
Image_1	'Di depan terdapat sofa besar berwarna abu-abu', 'Sebuah meja kayu berukuran sedang berada di depan sofa',	'In the front there is a large gray sofa', 'A medium-sized wooden table is in front of the sofa',

	'Di samping kiri sofa terdapat pot berwarna putih dengan tanaman di dalamnya', 'Di depan terdapat sofa panjang berwarna abu-abu dengan meja dari kayu yang rendah', 'Di bagian kiri terdapat tanaman dan lampu tinggi di samping rak buku'	'On the left side of the sofa there is a white pot with a plant in it', 'In the front there is a long gray sofa with a low wooden table', 'On the left are plants and a tall lamp beside the bookshelf'
Image_2	'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', 'Di bagian kanan meja bertaplak merah terdapat tumpukan piring berwarna putih', 'Di bagian kiri meja bertaplak merah terdapat tumpukan cangkir plastik'	'There are berry cakes, biscuits and grapes on the table', 'The round table with the red cloth has a lot of food on it', 'Food utensils, plates, glasses and knives are on the red-clothed table', 'On the right side of the table with the red cloth there is a pile of white plates', 'To the left of the table with the red cloth is a pile of plastic cups'

2.2 Preprocessing

In this process, all images are resized, and the resizing size follows the required input size according to the architecture used. For the caption side, the captions are all lowercased. Each caption is also given a startseq and endseq to indicate its beginning and ending caption.

2.3 Image Captioning Model

We apply the merge architecture for image captioning and make some experimental setups. In this study, we use three different pre-trained word embedding models separately. The models are GloVe[1] with a vector size of 50, Word2Vec[2] with a vector size of 400, and FastText[3] with a vector size of 300. For the image feature extraction, several different CNN-based architectures are also used separately in different experimental setups, thus we can get the feature vector from the images. The CNN-based architectures we use in this work are namely, InceptionV3, ResNet152V2, InceptionResNetV2, EfficientNetB3V1, EfficientNetB7V1, and NASNetLarge.

The Inceptionv3 model was utilized by TensorFlow in extracting or classifying image features. Paper on this model shows that Inception-v3 has significant impact on improving the performance and efficiency of deep learning neural networks [27]. Previous study was using Inception-v3 to develop flower classifier and the result shows that the model can be used to significantly improve the model accuracy [28]. While InceptionResNetv2 is other variation of the Inception-v3 model, which is significantly deeper than Inception-v3 and has significantly improved recognition performance. The InceptionResNetv2 architecture is shown to be more accurate than previous state-of-the-art models [29].

EfficientNet offers far greater accuracy and efficiency than previous ConvNets. In particular, EfficientNet-B7 achieves to be the state-of-the-art model. EfficientNet-B7 also achieves state-of-the-art on various transfer learning datasets. While model EfficientNet-B3 achieves higher accuracy than ResNeXt101 [30]. Model Residual Networks (Resnet) introduces a structure called Residual Learning Unit that has the main advantage of improving accuracy without increasing the complexity of the model. Resnet152 is selected as it achieves the best accuracy among Resnet family members [31]. Whereas NasNetLarge model outperforms other state-of-the-art approaches such as DenseNet, moreover NasNet also works splendidly on MS COCO datasets and surpasses other models as well [32].

The merge architecture for this study is shown in Figure 2. We set the maximum length of the caption to 27 as presented in input_3 and will then be fed into the embedding layer. In embedding layer, the words are mapped to the certain embedding, GloVe, Word2Vec or FastText. The word embedding that is used in Figure 2 is Word2Vec. Next, we add a dropout layer of 0.5 to prevent overfitting. The output from the dropout layer is then fed into the LSTM layer with 256 nodes to be processed. While input_2 contains the image feature vector that is previously extracted using certain CNN-based architecture. The CNN-based architecture used in Figure 2 is EfficientNetB3V1. This layer is also followed by a dropout layer of 0.5 and the output will then be fed into a dense layer.

The next step is concatenating the output from LSTM layer and dense layer to be fed into another dense layer with relu as the activation function. The output from this dense layer is then fed into the last dense layer with softmax activation function. Finally, we use the two algorithms, Greedy Search (an algoritm that generates caption by choosing one of the best candidate at each step and using argmax function to select word with the highest probability) and BEAM

[1] https://github.com/irfanhanif/Mira
[2] https://www.kaggle.com/bhimantoros/pretrained-word2vec-indonesia?select=wiki.id.case.vector
[3] https://github.com/indobenchmark/indonlu

Search (a greedy search algorithm based on heuristics that selects multiple alternatives word instead of one) to generate Indonesian captions using the value of 3 as index to predict the caption for the test set [33].

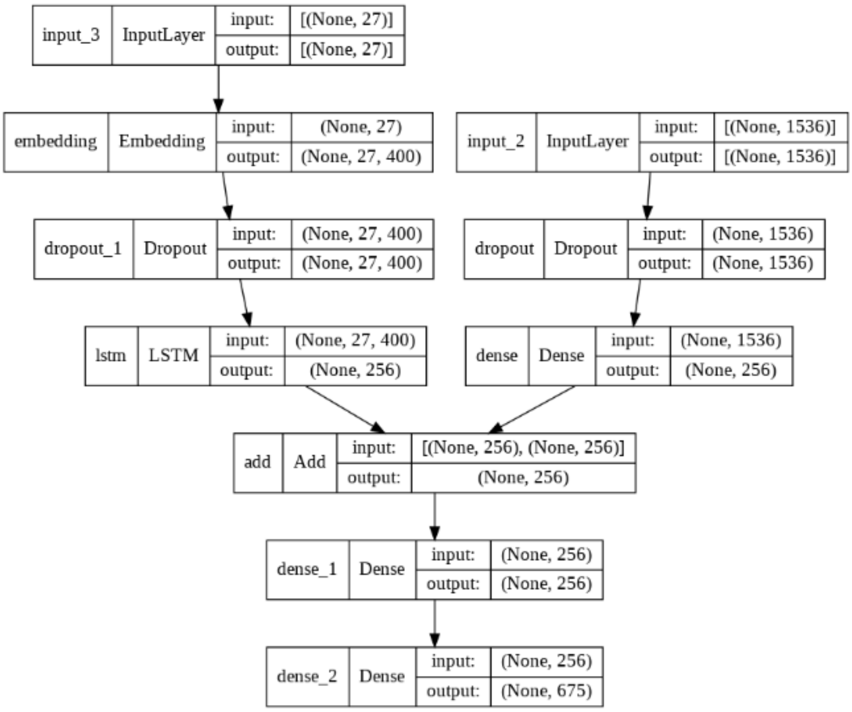

Figure 2. Image Captioning Model Architecture based on EfficientNet B3 V1

2.4 Evaluation

We use BLEU-n (BLEU-1, BLEU-2, BLEU-3, BLEU-4) to evaluate the generated captions. Bilingual Evaluation Understudy (BLEU) is commonly used for evaluating Natural Language Processing (NLP) systems that generate language, especially in natural language generation and machine translation [34]. The number in BLEU indicates the n-gram that the BLEU evaluates. In BLEU, the highest number of n-gram is 4. This metric calculates the similarity score between generated and target text that ranges from 0 to 1, where 1 means similar and 0 is not similar [17].

3. Results and Discussion

We trained the 773 images from our dataset using Adam as the optimizer, batch size value of 8, and 100 epochs. We picked 10 images from Google to be used as the test set. Here we have 18 models with different word embedding and architectures. These 18 models and their model loss are presented in Table 2. From the table, whichever the word embedding is used, Inception V3, InceptionResNet V2 and ResNet152 V2 have higher loss scores compared to other models such as EfficientNet B3 V1, EfficientNet B7 V1, NASNetLarge that almost share the same lesser loss score. Model 1 has the highest loss score of 1.0337 by combining GloVe, Inception V3 and LSTM, while Model 12 scored the least loss score of 0.3330 with a combination of Word2Vec NASNetLarge and LSTM.

Table 2. Models' Loss

Model	Word Embedding	Experimental Model	Loss
Model 1	GloVe	Inception V3 + LSTM	1.0337
Model 2		InceptionResNet V2 + LSTM	0.8743
Model 3		ResNet152 V2 + LSTM	0.7424
Model 4		EfficientNet B3 V1 + LSTM	0.6681
Model 5		EfficientNet B7 V1 + LSTM	0.6503
Model 6		NASNetLarge + LSTM	0.6533
Model 7	Word2Vec	Inception V3 + LSTM	0.5745
Model 8		InceptionResNet V2 + LSTM	0.4877
Model 9		ResNet152 V2 + LSTM	0.3960
Model 10		EfficientNet B3 V1 + LSTM	0.3476

Model 11		EfficientNet B7 V1 + LSTM	0.3476
Model 12		NASNetLarge + LSTM	0.3330
Model 13	FastText	Inception V3 + LSTM	0.5912
Model 14		InceptionResNet V2 + LSTM	0.4994
Model 15		ResNet152 V2 + LSTM	0.4000
Model 16		EfficientNet B3 V1 + LSTM	0.3609
Model 17		EfficientNet B7 V1 + LSTM	0.3571
Model 18		NASNetLarge + LSTM	0.3517

We evaluate our models using BLEU-1,2,3,4 and the results are presented in Table 3. As can be seen in the table, Model 2 by combining GloVe, InceptionResNet V2 and LSTM reached the highest BLEU-1 using greedy search and BLEU-2 score using BEAM search. Model 8 by combining Word2Vec, InceptionResNet V2 and LSTM reached the highest BLEU-2 score using greedy search and BLEU-1 score using BEAM search. Model 10 reached the highest BLEU-3 score using BEAM search by combining Word2Vec, EfficientNet B3 V1 and LSTM. Model 12 get the highest BLEU-4 score using greedy search by combining FastText, NASNetLarge, and LSTM. Model 18 by combining FastText, NASNetLarge, and LSTM obtained the highest BLEU-3 using greedy search and BLEU-4 score using BEAM search.

Table 3. BLEU Scores

Model	Greedy Search				BEAM Search			
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Model 1	0.41961	0.29425	0.29975	0.33799	0.34193	0.30616	0.41130	0.46063
Model 2	**0.49126**	0.33662	0.33550	0.35047	0.39822	**0.40567**	0.43434	0.43758
Model 3	0.35022	0.25854	0.30533	0.34417	0.31341	0.23742	0.33425	0.37108
Model 4	0.38274	0.32848	0.37918	0.41366	0.24706	0.27905	0.36609	0.39548
Model 5	0.40316	0.25708	0.32561	0.38294	0.31423	0.28405	0.40160	0.44605
Model 6	0.42099	0.29679	0.35304	0.35797	0.37985	0.25600	0.35379	0.39330
Model 7	0.36391	0.24879	0.25484	0.31511	0.32030	0.28111	0.41560	0.47111
Model 8	0.46055	**0.39029**	0.35563	0.34659	**0.41383**	0.38563	0.42219	0.44655
Model 9	0.36482	0.27980	0.34699	0.40519	0.28031	0.24636	0.32430	0.35714
Model 10	0.39297	0.31092	0.34483	0.36756	0.34439	0.38461	**0.44549**	0.47005
Model 11	0.47597	0.34760	0.37819	0.41065	0.33493	0.32684	0.40868	0.45603
Model 12	0.37901	0.29151	0.37688	**0.42338**	0.37933	0.35502	0.44044	0.45422
Model 13	0.43281	0.32979	0.31204	0.34826	0.28107	0.27594	0.39908	0.45828
Model 14	0.46116	0.33752	0.30684	0.33652	0.34095	0.32260	0.42406	0.46138
Model 15	0.34272	0.29610	0.33017	0.36971	0.21916	0.28389	0.36828	0.41444
Model 16	0.42183	0.31727	0.34637	0.37844	0.27315	0.26360	0.34677	0.38899
Model 17	0.44348	0.29807	0.32796	0.35553	0.38109	0.27337	0.36856	0.40942
Model 18	0.33936	0.32872	**0.38076**	0.41061	0.29817	0.32837	0.43948	**0.49002**

We tested these 18 image captioning models on our test set that is consisted of 10 images that we collected from Google. Due to limitation, we show only a few samples of Indonesian generated caption for models with the highest BLEU scores (Model 2, Model 8, Model 10, Model 12 and Model 18) along with the English translation in Table 4. From the table, it can be seen that the models are able to generate captions that are barely out of context from the given images by using both Greedy and Beam search. We selected 4 models (Model 2, Model 8, Model 10, Model 12, and Model 18) since other models performed poorly in generating captions and to see if the model's performance matched the BLEU score obtained. Among these 4 models, Model 12 shows a good performance and works better in generating Indonesian captions that correspond to the given images.

Model 12 is able to generate good captions for 7 given pictures including object's name, location ("di samping kiri" / "on the left"), color ("laptop putih" / "white laptop") and characteristics ("komputer yang menyala" / "turned-on computer"). Model 2 is also able to generate sufficient captions for 6 given pictures. But, compared to Model 12, Model 2 struggles in generating the correct object's name and failed to include object's color. Whereas, other models such as Model 8, Model 10, and Model 18, although having the highest BLEU scores, these models performed poorly in generating the right caption for the given images.This can be the case where a high BLEU scores does not necessarily mean that the quality of the generated text is good [35].

From our test set that is consisted of 10 images, most models are capable in distinguishing & generating captions of kitchen room images, laptops, sinks, and bed rooms. On the other hand, most models also find difficulty in generating correct captions for images such as getting the shape or characterstics of dining table images and naming random objects on a table. Compared to the other models with high BLEU scores, Model 12 with a combination of Word2Vec, NASNetLarge & LSTM turns out has better ability to distinguish the random objects on top of a table.

For word embedding, Word2Vec generated better caption when it's combined with NASNetLarge & LSTM. While GloVe seems to work better when it's combined with InceptionResNet V2 & LSTM. Although it doesn't perform as well as the first one, the later still generate sufficient and within context caption. From the three word embeddings, FastText has poorer performance and struggled in generating the correct captions.

Table 4. Model Generated Captions

No.	Image	Model	Generated Caption	
			Greedy Search	BEAM Search
1.		Model 2	di depan terdapat meja wastafel dengan sikat gigi di atasnya	di depan terdapat meja wastafel yang panjang dengan wastafel di tengahnya
			in front there is a sink table with a toothbrush on it	in front there is a long sink table with a sink in the middle
		Model 8	di depan terdapat wastafel dengan lemari cermin di atasnya	di depan terdapat wastafel dengan lemari cermin di atasnya
			in front there is a sink with a mirror cupboard on it	in front there is a sink with a mirror cupboard on it
		Model 10	di depan terdapat seorang pria yang sedang memegang gelas yang memasak	di bagian kiri terdapat wastafel berwarna putih
			in front there is a man holding glass while cooking	on the left there is a white sink
		Model 12	di samping kiri terdapat wastafel yang berada di meja konter dapur	di samping kiri terdapat wastafel yang berada di meja konter dapur
			on the left side there is a sink on the kitchen counter	on the left side there is a sink on the kitchen counter
		Model 18	di bagian kanan terdapat kompor oven dan teflon	terdapat dua handle faucet yang berada di atas meja
			on the right side there is an gas stove and a teflon	there are two faucet handles on the table
2.		Model 2	di depan terdapat seorang pria yang sedang duduk di dekat kiri dan laptop di meja depan	di atas meja terdapat komputer yang menyala
			in front there is a man sitting near the left and a laptop at the front desk	there is a turned-on computer on the table
		Model 8	di depan terdapat laptop berwarna hitam yang di atas meja susun	di depan terdapat laptop berwarna hitam dan berwarna hitam di atas meja berwarna putih
			in front there is a black laptop on a stacked-table	in front there is a black and black laptop on a white table
		Model 10	di depan terdapat seorang pria yang sedang duduk di atas meja kayu berwarna cokelat	di depan terdapat seorang pria yang sedang duduk di atas meja kayu berwarna cokelat
			in front there is a man sitting on a brown wooden table	in front there is a man sitting on a brown wooden table
		Model 12	di depan terdapat laptop berwarna putih yang menyala	di depan terdapat laptop berwarna putih dengan layar menyala berada di atas meja berwarna cokelat
			in front there is a white turned-on laptop	in front there is a white laptop with a lit screen on a brown table
		Model 18	di depan terdapat banyak perangkat elektronik dan laptop	di depan terdapat banyak perangkat elektronik dan laptop di atas meja
			in front there are many electronic devices and laptops	in front there are many electronic devices and laptops on the table

No.	Image	Model	Generated Caption	
			Greedy Search	BEAM Search
3.		Model 2	di atas meja terdapat banyak gelas wine dan gelas minuman	di atas meja terdapat peralatan kamera dan laptop
			On the table there are many wine glasses and drinking glasses	on the table there are camera equipment and a laptop
		Model 8	di depan terdapat seorang pria yang sedang memegang makanan	di depan terdapat seorang pria yang sedang minuman dari botol kaca ke gelas wine
			in front there is a man holding food	in front there is a man who is drinking from glass bottle to wine glass
		Model 10	di depan terdapat seorang pria yang menggunakan dan	di depan terdapat seorang pria yang memegang ponsel genggam
			in front there is a man who uses dan	in front there is a man holding a mobile phone
		Model 12	di depan terdapat meja makan dengan beberapa gelas kaca besar	di atas meja terdapat beberapa gelas dan gelas kaca
			In front there is a dining table with several large glass glasses	on the table there are some glasses and glass cups
		Model 18	di depan terdapat seorang pria yang sedang memegang ponsel untuk berkomunikasi	seorang pria sedang memegang ponsel untuk berkomunikasi
			in front there is a man holding a cell phone to communicate	a man holding a cell phone to communicate

4. Conclusion

In this study, we created an image captioning using various household objects such as bed, cell phone, chair, couch, oven, potted plant, refrigerator, sink, table and tv that are collected from the MSCOCO dataset. We created 18 experimental models that compared the performance of three word-embedding techniques (GloVe, Word2Vec, FastText) combined with several CNN-based architectures (InceptionV3, ResNet152V2, InceptionResNetV2, EfficientNet B3 V1, EfficientNet B7 V1, and NASNetLarge) along with LSTM as decoder to get the best image captioning model. From these combinations we found that our model showed better performance in generating Indonesian captions than other models when word embeddings Word2Vec is combined with the CNN-based model NASNetLarge. We also found out that models with high BLEU scores doesn't guarantee that models will generate a good caption that correspond to the given image.

References

[1] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," *Adv. Neural Inf. Process. Syst.*, vol. 32, no. NeurIPS, pp. 1–11, 2019. https://doi.org/10.48550/arXiv.1906.05963

[2] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting Image Captioning with Attributes," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-Octob, pp. 4904–4912, 2017. https://doi.org/10.1109/ICCV.2017.524

[3] L. Srinivasan and D. Sreekanthan, "Image Captioning-A Deep Learning Approach," *Int. J. Appl. Eng. Res.*, vol. 13, no. 9, pp. 7239–7242, 2018.

[4] U. Bhoga, V. Aravind, G. Sreeja, and M. Arif, "Image Caption Generation Using CNN and LSTM," *JAC A J. Compos. Theory*, vol. XIV, no. VII, pp. 257–263, 2021.

[5] E. Cahyaningtyas and D. Arifianto, "Development of under-resourced Bahasa Indonesia speech corpus," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Dec. 2017, vol. 2018-Febru, no. December, pp. 1097–1101. https://doi.org/10.1109/APSIPA.2017.8282191

[6] A. A. Nugraha, A. Arifianto, and Suyanto, "Generating image description on Indonesian language using convolutional neural network and gated recurrent unit," *2019 7th Int. Conf. Inf. Commun. Technol. ICoICT 2019*, pp. 1–6, 2019. https://doi.org/10.1109/ICoICT.2019.8835370

[7] P. Shah, V. Bakrola, and S. Pati, "Image captioning using deep neural architectures," in *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Mar. 2017, pp. 1–4. https://doi.org/10.1109/ICIIECS.2017.8276124

[8] N. S. B, L. White, and M. Bennamoun, *NNEval : Neural Network Based*, vol. 1. Springer International Publishing. https://doi.org/10.1007/978-3-030-01237-3_3

[9] Y. Chu, X. Yue, L. Yu, M. Sergei, and Z. Wang, "Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention," *Wirel. Commun. Mob. Comput.*, vol. 2020, pp. 1–7, Oct. 2020. https://doi.org/10.1155/2020/8909458

[10] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," 2014, pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[11] C. Sur, "MRRC: multiple role representation crossover interpretation for image captioning with R-CNN feature distribution composition (FDC)," *Multimed. Tools Appl.*, vol. 80, no. 12, pp. 18413–18443, May 2021. https://doi.org/10.1007/s11042-021-10578-9

[12] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga, and M. Bennamoun, "Text to Image Synthesis for Improved Image Captioning," *IEEE Access*, vol. 9, pp. 64918–64928, 2021. https://doi.org/10.1109/ACCESS.2021.3075579

[13] K. Arora, A. Raj, A. Goel, and S. Susan, “A Hybrid Model for Combining Neural Image Caption and k-Nearest Neighbor Approach for Image Captioning,” 2022, pp. 51–59. https://doi.org/10.1007/978-981-16-1249-7_6
[14] H. Rampal and A. Mohanty, “Efficient CNN-LSTM based Image Captioning using Neural Network Compression,” Dec. 2020. https://doi.org/10.48550/arXiv.2012.09708
[15] S. Katiyar and S. K. Borgohain, “Image Captioning using Deep Stacked LSTMs, Contextual Word Embeddings and Data Augmentation,” Feb. 2021. https://doi.org/10.48550/arXiv.2102.11237
[16] H. Wei, Z. Li, F. Huang, C. Zhang, H. Ma, and Z. Shi, “Integrating Scene Semantic Knowledge into Image Captioning,” *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 17, no. 2, pp. 1–22, May 2021. https://doi.org/10.1145/3439734
[17] M. R. S. Mahadi, A. Arifianto, and K. N. Ramadhani, “Adaptive Attention Generation for Indonesian Image Captioning,” *2020 8th Int. Conf. Inf. Commun. Technol. ICoICT 2020*, 2020. https://doi.org/10.1109/ICoICT49345.2020.9166244
[18] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, “Deep Reinforcement Learning-Based Image Captioning with Embedding Reward,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 1151–1159. https://doi.org/10.1109/CVPR.2017.128
[19] V. Atliha and D. Sesok, “Comparison of VGG and ResNet used as Encoders for Image Captioning,” in *2020 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*, Apr. 2020, pp. 1–4. https://doi.org/10.1109/eStream50540.2020.9108880
[20] C. Wang, H. Yang, C. Bartz, and C. Meinel, “Image Captioning with Deep Bidirectional LSTMs,” in *Proceedings of the 24th ACM international conference on Multimedia*, Oct. 2016, pp. 988–997. https://doi.org/10.1145/2964284.2964299
[21] Y. Bhatia, A. Bajpayee, D. Raghuvanshi, and H. Mittal, “Image Captioning using Google's Inception-resnet-v2 and Recurrent Neural Network,” in *2019 Twelfth International Conference on Contemporary Computing (IC3)*, Aug. 2019, pp. 1–6. https://doi.org/10.1109/IC3.2019.8844921
[22] M. Humaira, S. Paul, M. Abidur, A. Saha, and F. Muhammad, “A Hybridized Deep Learning Method for Bengali Image Captioning,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 2, 2021. https://dx.doi.org/10.14569/IJACSA.2021.0120287
[23] S. Das, L. Jain, and A. Das, “Deep Learning for Military Image Captioning,” *21st Int. Conf. Inf. Fusion, Cambridge, United Kingdom*, pp. 2165–2171, 2018. https://doi.org/10.23919/ICIF.2018.8455321
[24] V. Atliha and D. Sesok, “Pretrained Word Embeddings for Image Captioning,” in *2021 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*, Apr. 2021, pp. 1–4. https://doi.org/10.1109/eStream53087.2021.9431465
[25] S. Yagcioglu, E. Erdem, A. Erdem, and R. Cakici, “A Distributed Representation Based Query Expansion Approach for Image Captioning,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2015, pp. 106–111. http://dx.doi.org/10.3115/v1/P15-2018
[26] D. H. Fudholi *et al.*, “Image Captioning with Attention for Smart Local Tourism using EfficientNet,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1077, no. 1, p. 012038, Feb. 2021. https://doi.org/10.1088/1757-899X/1077/1/012038
[27] C. Szegedy *et al.*, “Going Deeper with Convolutions,” pp. 1–12, Sep. 2014. https://doi.org/10.48550/arXiv.1409.4842
[28] O. Albatayneh, L. Forslöf, K. Ksaibati, and D. Ph, “Image Retraining Using TensorFlow Implementation of the Pretrained Inception-v3 Model for Evaluating Gravel Road Dust,” vol. 26, no. 2, pp. 1–10, 2020. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000545
[29] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning,” Feb. 2016.
[30] M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” May 2019. https://doi.org/10.48550/arXiv.1905.11946
[31] L. D. Nguyen, D. Lin, Z. Lin, and J. Cao, “Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation,” in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2018, no. May, pp. 1–5. https://doi.org/10.1109/ISCAS.2018.8351550
[32] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, “Learning Transferable Architectures for Scalable Image Recognition,” Jul. 2017. https://doi.org/10.48550/arXiv.1707.07012
[33] S. Takkar, A. Jain, and P. Adlakha, “Comparative Study of Different Image Captioning Models,” in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, Apr. 2021, no. Iccmc, pp. 1366–1371. https://doi.org/10.1109/ICCMC51019.2021.9418451
[34] E. Reiter, “A Structured Review of the Validity of BLEU,” *Comput. Linguist.*, vol. 44, no. 3, pp. 393–401, Sep. 2018. https://doi.org/10.1162/coli_a_00322
[35] M. D. Zakir Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, “A comprehensive survey of deep learning for image captioning,” *ACM Comput. Surv.*, vol. 51, no. 6, 2019. https://doi.org/10.1145/3295748