A Rapid Review of Image Captioning

Adriyendi*1

1Informatic Group, FEBI IAIN Batusangkar, Indonesia
1suratkudisman@gmail.com
*Corresponding Author

Received 29 April 2021; accepted 22 July 2021

Abstract. Image captioning is an automatic process for generating text based on the content observed in an image. We do a review, create a framework, and build an application model. We review image captioning into 4 categories based on input model, process model, output model, and lingual image caption. The input model is based on criteria caption, method, and dataset. The process model is based on the type of learning, encoder-decoder, image extractor, and metric evaluation. The output model is based on architecture, features extraction, feature mapping, model, and the number of the caption. Lingual image caption based on language model with 2 groups: bilingual image caption and cross-language image caption. We also design a framework with three framework models. Furthermore, we also build applications with three application models. We also provide research opinions on trends and future research that can be developed with image caption generation. Image captioning can be further developed on computer vision versus human vision.

Keyword: Image, Caption, language, model

1 Introduction

Image captioning is a popular new research project in image analysis and text analysis. It is an automatic process for creating image captions. The caption displays natural language text based on the image [1]. It is defined as the process of producing a textual description for an image [2]. In other words, it is creating a description for an image that is inputted into natural language.

Starting with computer vision which is used to identify objects, attributes, and their relationships, then, natural language processing is used to monitor syntax and semantics, and finally, machine learning is used to produce text [3]. In another words, it is a technology that produces textual descriptions of images for semantic indexing. The object recognition from an image, the relationship between recognized objects to understand events in the image, and produces analysis to the image benefits from this technology [4]. Image captioning arises from the need to translate between two different modalities (multimodal), but usually in pairs.

Image capture based on content (content-based information retrieval) can be done by image captioning [5]. The application of image captioning is very broad and significant, especially in the field of human and computer interaction. Apart from this, practical image captioning helps disabled people interact with normal people [6]. It is also used to create multimedia content descriptions, assist e-commerce companies with digital marketing, and create online news stories for news content narratives [7]. Image captioning can also help people with visual disabilities understand image content. In the medical field, image captioning can also provide a diagnosis or medical assistance based on image content analysis [8]. Image captioning also plays a very important role in interaction and textual communication on social media. Text is an important component in posting images with captions. Text can increase user engagement in
campaigns and advertisements. Apart from social media, text for images also plays a role in making sentences that better reflect user activity, activity on business sites with image captions, captions for digital organization, and company profiles. The main contribution of this paper are summarized as follows:

1) Categorization model for the image to text synthesis based on input model, process model, output model, lingual image caption model.
2) Design framework model as the conceptual model for image captioning generation. Application models as design applications to implemented image captioning.
3) Computer vision, natural language processing, machine learning as the main component for the image to text synthesis.

We do a rapid review based on articles/papers in IEEE Xplore, arXiv, ACM, Science Direct, and Elsevier publisher. We select articles based on keyword “image captioning”. We find 12 articles are survey paper and review papers, 9 articles are technical papers based on the keyword “image lingual caption”, and 9 articles are technical paper based on keyword “image lingual caption plus multi-modal”.

The rapid review model based on the categorization model in this study is underexplored on the topic of image captioning. Image captioning is an applicative and futuristic development model in the image to text synthesis.

2 Review
2.1 Input Model
The input model based on criteria caption, method, and the dataset is shown in Table 1.

| Type of Caption   | Method     | Dataset     |
|-------------------|------------|-------------|
| Template Base     | Attention  | Flickr8k    |
| Retrieval Base    | Novel Object | Flickr30k   |
| Novel Base        | Semantic   | Ms. COCO    |

The image caption-based template approach has an improved template with several empty slots for generating text. In this approach, detecting objects, attributes, actions and then fill the empty spaces in the template [5]. Text for request image is selected from candidate text set. Novel base captions are carried out by analyzing the visual content of an image and producing a description of the image using a language model [5]. The image caption is generated by syntax and semantic in a restricted process. In the retrieval-based image caption approach, the text is taken from a series of existing texts from a predetermined set of sentences [10]. This method is used to find images that are visually similar to the candidate text from the training dataset. In image captioning applications, it is possible to find novel bases that are not in the predetermined vocabulary with sentence data in paired images. There is no need to retrain the entire system when multiple drawings with new concepts emerge [10].

The attention mechanism is based on the signal from the input image by frame into encoder-decoder [10]. This model is made to replicate the way people pay attention to objects in an image and the way people make annotations between objects in the image [11]. People can choose information, ignore primary information, and secondary information. This ability to choose is called attention. The novel object caption can produce object descriptions in text and images that are not in the dataset. Steps: (1) Separate lexical classifications and language models. (2) In-text models are trained on paired image text data. (3) The two models are combined to practice together in producing texts for novel objects. The semantic concept-based method selectively results in the extraction of an image. Steps: (1) Encode image features using by the encoder. (2) The image feature is inputted to the language model input. (3) Added to the various hidden states by semantic concept in the language model. (4) Semantic
implementation can design a clear way to feed sentiment into the image captioning system [11]. An important aspect of the attention mechanism is how to decide what to describe and in what order in the image captioning process [13].

Flickr8K is a dataset containing thousands of images extracted from Flickr. Flickr contains thousands of photos (mainly humans and animals), thousands of training images, thousands of image verifications, and thousands of image tests [1]. This dataset has 5 reference captions for each image [5]. The Flickr30K is a dataset contains thousands of images provided by human and for each image with 5 reference captions. This dataset also contains detectors for common objects with color classification [5]. This dataset is an extended dataset from the Flickr8k dataset. The pictures in this dataset are mainly of daily human activities. Based on crowdsourcing services, the caption is annotated by humans for each image [10]. The images in this dataset come from Yahoo’s photos. These images can be used for training, testing, and validation in caption [12]. Ms. COCO (Microsoft Common Object in Context) is a very large data set for image recognition, object segmentation, and text writing. Microsoft COCO contains thousands of images, millions of instances, dozens of object categories, and five texts per image [5]. This dataset for scene understanding and capturing images of complex daily events. This dataset is more challenging because the images contain many objects, messy backgrounds, and complex semantic relationships. Other datasets for image captioning are BBC News, Pascal UIUC, and SBU. PASCAL 1K, AI Challenger Dataset, STAIR Captions, IAPR TC-12, Stock3M, FlickrStyle10K, and Visual Genome [12].

2.2 Process Model

Process model based on type of learning, encoder-decoder, image extractor, and evaluation metric. The process model is shown in Table 2.

| The Type of Learning | Encoder-Decoder | Image Extractor | Evaluation Metric |
|----------------------|------------------|-----------------|------------------|
| Supervised Learning  | CNN              | VGG-Net         | BLEU             |
| Unsupervised Learning| RNN              | Res-Net         | CIDE             |
| Reinforcement Learning| LSTM            | Alex-Net        | METEOR           |
|                     | GRU              | Google-Net      | ROUGE            |
|                     | -                | Dense-Net       | SPICE            |

In supervised learning, training data is used to produce the desired output called labels, while unsupervised learning deals with data that are not label-ed. Reinforcement learning is a type approach to finding data and/or labels through exploration and given signals.

GRU (Gate Recurrent Unit) uses to control the flow of information [5]. CNN (Convolutional Neural Network) is used to understand image content. LSTM (Long Short-Term Memory) focuses on modelling text data, combining context information and images, and predicting word distribution [12]. RNN (Recurrent Neural Network) is used for the encoder-decoder process (the encoder is a CNN and the decoder is RNN).

Commonly used image feature extractors are VGG-Net and Res-Net. VGG-Net has a simple and powerful model. Res-Net is most efficient compared to all other extractors. Other alternative image feature extractors that can be used are Alex-Net, Google-Net, and Dense-Net. Google-Net is also called Inception-X Net [11].

Evaluation metrics used to measure linguistic quality and semantic correctness. This is necessary for comparing text, images analysis, and generating sentences. Evaluation metrics include BLEU, ROUGE, METEOR, SPICE, and CIDE. BLEU (Bi-
Lingual Evaluation Understudy) measures the same number of words as the base adverb [3], which is used to measure the similarity between two sentences. BLEU is a metric that compares candidate sentences with reference sentences [10]. In this metric, the process is the same as the machine translation process. Difference between the reference sentence and the candidate sentence by using BLEU metric. BLEU can also be used to perform sentence-level analysis [12].

CIDE (Consensus-based Image Description Evaluation) is a metric to measure the similarity of the resulting sentence to the correct sentence. In this metric, encoding the n-gram frequency in the candidate sentence to toward in the reference sentence. This metric is used to evaluate sentences in terms of grammaticality, salience, importance, and accuracy [10]. CIDE is specially designed for image annotation problems and measures image annotation consistency. CIDE is a metric that measures the similarity of a resulting sentence against a set of root sentences that are true [12].

Metric for Evaluation of Translation with Explicit Ordering or METEOR is another type metric for machine translation. This metric evaluates the candidate's sentences and reference sentences, then calculates a score based on the matching results. The calculations involve the accuracy, memory, and alignments of the matching words. This metric can be used to overcome the weakness of the BLEU metric. Apart from searching for exact word matches, this metric can also search for synonyms of words and look for correlations at the segment level or sentence level [3]. This metric is based on the n-gram precision-matched [10]. METEOR is a metric with harmonic averages of precision and recall, and paraphrasing between reference sentences and candidate sentences. METEOR shows a more accurate evaluation especially for a small number in reference sentences [12].

ROUGE (Recall-Oriented Understanding for Gist Evaluation) is designed to use the longest general order between candidate sentences and a set of reference sentences. The longest common order between two sentences requires only consecutive word matching, and the words that match do not have to be in order. ROUGE is used to evaluate text summarization algorithms. If ROUGE score is high, then the performance is better [1].

Semantic Propositional Image Caption Evaluation or SPICE is used to measure restores objects, attributes, and relationships effectively. SPICE can be able to capture human judgments about model subtitles [1]. This metric places the adverb similarity based on the scene graph tuple (candidate sentences and reference sentences) [12].

BLEU is great for evaluating small sentences. ROUGE can be used to evaluate different types of text. ROUGE is precise for automatic summaries, and CIDE and SPICE are precise for captions. BLEU and METEOR are precise for machine translation, ROUGE is precise for automatic summaries, and CIDE and SPICE are precise for captions [1]. METEOR can perform evaluations on various segments of information. SPICE is better at understanding the semantic breakdown of text compared to other metrics [5]. The most widely used evaluation metrics are BLEU and ROUGE, but they are weakly correlated with human judgment. For correlation with human assessment, METEOR and CIDE are better [12].

### 2.3 Output Model

Output model based on the architecture, feature extraction, feature mapping, model, and the number of the caption. The output model is shown in Table 3.

| Architecture | Feature Extraction | Feature Mapping | Number of Caption | Model   |
|--------------|--------------------|-----------------|-------------------|---------|
| Composition  | Handcraft          | Visual Space    | Dense Caption     | Classic |
The Composition architecture-based method consists of several independent functional building blocks: (1) Extract semantic from images by CNN. (2) Generate a set of candidate caption texts by the language model. (3) This candidate text is repeated using RNN [5]. Encoders are a framework for the transformation process from one representation to another. The encoder network encodes the input into context vectors and the decoder network translates the context vectors to produce the output [9]. In this encoder-decoder process, first, encode an image into an intermediate representation by using encoders, next, the decoder changes the intermediate representation as input, and generates word by word [10]. An encoder is a process of “reading” an image (inputting an image), then extracting the image with a high-level feature representation. A decoder is a process of “producing words” (the output is words) as an image representation of the encoder (an encoded image). This process produces words to represent images with correct sentences syntactically and semantically [11].

Handcraft features based on the maximum estimate and the greatest chance. This feature studies the visual detectors and language models from the image captioning dataset. This feature is used for images analysis, detect objects, and generate descriptions. Deep learning features can be used to generate text for image caption. The RNN is used as an attention mechanism in the image caption. This feature is achieved good results in language model [1]. It can handle large and varied image sets. CNN is widely used for learning features such as soft-max which is used for classification. CNN is used to image understanding and RNN is used for text generation [5].

Text generated by using mapping in visual space. Also, it can be used in multimodal space. The visual space-based method performs an explicit mapping from images to caption. In contrast, the multimodal space-based method combines vision and implicit mapping [5]. In dense captioning, the text is generated for each scene area in an image. Area-based descriptions are objective and detail. This area is better for local image descriptions than global image descriptions based on region descriptions. Region-based descriptions are dense captioning. The dense localization layer processes the image efficiently, implicitly predicting a set of regions of interest. Another method is whole scene captioning which generates text for the entire scene such as encoders, composition, attention, semantics, novel objects, and deep learning. This method generates single or multiple texts for the entire scene [5]. The classic model is building a natural language from visual content. Visual content based on visual-to-text technical patterns falls into four distinct categories: (1) language rules or template-based scripting, (2) taking descriptions from other visual content, (3) supplemental descriptions with visual recognition, and (4) create a language model to produce descriptions. Advanced model is the creation of the latest image captions from image to text. This model encodes an image to a text decoder by translating images into sentences. This model includes text visual embedding and attention mechanisms [12].

### 2.4 Lingual Model

| No | Caption | Description |
|----|---------|-------------|
| 1  | Vietnamese Image Captioning [7] | Microsoft COCO dataset with manual text and annotation tool to construct dataset. Image caption generation from English text to the Vietnamese Language. |
| 2  | Chinese Image Captioning [15] | Supervised learning, feature extraction using deep learning, dense caption, and attention for robot rehabilitation used to Autism Syndrome Disorder. Image caption generation from English to the Mandarin Language. |
| 3  | Myanmar Image Captioning [16] | CNN encoder to LSTM decoder and corpus caption in Myanmar Language. Image caption generation from English text to the Myanmar Language. |
The bilingual image caption is shown in Table 4. Cross-lingual image caption shows in Table 5. The making of cross-lingual languages depends on their ability to support the use of non-English languages with model trends annotations. Cross-lingual generates graphic text from English to non-English (1 source language to more than 1 target language). Cross-lingual caption based on language model and cross-modal based on multimodal (text, image, speech, sound, audio, and video). Lingual models for image captions are divided into 2 groups, namely: bilingual (image to text) and cross-language image text (image to text plus multimodal). Models for image captioning given the scarcity of data label-ed non-English. Machine translation is given more benefits from large data (bilingual and multilingual contexts). Bilingual captions
generate graphic text from English to non-English (1 source language to 1 target language).

3 Discussion
3.1 Framework Model

The framework created as a model for building image captioning can be seen in Figure 1, Figure 2, and Figure 3.

![Figure 1. Framework 1](image)

Framework 1 is used to classify image captioning based on category (input, process, and output) and group (lingual caption). Categories are used to determine the design and model of image captioning. The groups are divided into 2 groups, namely: bilingual (image to text) and cross-lingual image caption (image to text plus multimodal).

![Figure 2. Framework 2](image)

Framework 2 is used to classify image captioning based on input, process, and output. Input using computer vision, a process using machine learning, and output using natural language processing.

Framework 3 displays image captioning based on input, process, and output. First, input image with encoder process using image transform and image taken from Flickr8k dataset. Next, do the encoder process using the CNN Model using image features and embedded image feature vectors. Next, do the decoder process with the LSTM Model, RNN Model, and GRU Model. Finally, displaying text generation output as image captioning using natural language processing.

Framework 1, Framework 2, Framework 3 are conceptual models. This framework is popularly used to a developed image-to-text synthesis for image captioning generation.
3.2 Application Model

Figure 4 shows CNN as an Encoder to understand images (Image Understanding) and RNN as a Decoder to decode words (Text Generating). CNN is used to encode the image and RNN is used to decode a sentence from embedding. The process of converting image to text to produce output (Image Captioning). The function of $x_t$ is input $x$ at time $t$ to calculate $h_t$. We multiply $h_t$ with matrix $w$ to predict for $y_t$. RNN makes prediction for the word based on the hidden state in the previous time-step and current input.

Figure 5 shows CNN as an Encoder to understand images (Image Understanding) and LSTM as a Decoder to decode words (Text Generating). The process of converting image to text to produce output (Image Captioning). The output layer is a soft-max function. The soft-max function is used to generate the probability distribution vector for each character. This function helps to generate a complete translate word. The ($Sm$) is soft-max is used to predict the probability of being the next caption word. The ($We$) function is Word embedding is used to convert a word index to a vector in a higher-dimensional space.

Figure 6 shows CNN as an Encoder to understand images (Image Understanding) and GRU as a Decoder to decode words (Text Generating). The process of converting image to text to produce output (Image Captioning). GRU is a type of RNN. GRU uses 2 gates (a reset gate and an update gate) instead of 3. We have to transform the captions associated with the image into a token-izer (a list of tokenized words). Function Image model and language model are concatenated by adding and fed into another Fully Connected Layer. The pre-trained word embedding model is used to model training in our embedding layer.
Figure 4. Application Model 1

Figure 5. Application Model 2
4 Conclusion

We have discussed image captioning generation. We do a rapid review, create frameworks, and build application models. This review provides a comprehensive analysis of the aspects of input, process, output, and lingual captions regarding image captioning generation. This framework produces a framework design as an application model. This application can be a technical reference and as an example in building an application about image captioning.

In our view, a lot of improvisation and innovation can be done to develop research on image captioning generation. Image captioning can be further developed based on the synthesis of computer vision vs. human vision, deep learning vs. broad learning as a learning model, and the synthesis of the image to text and text to image by using natural language processing. Image captioning generation has become a multi-disciplinary research model (image analysis, text analysis, and content analysis), integration of applications and models (computer vision, machine learning, and natural language processing), and context systems (multimedia to multimodal) as research are exciting and interesting.

References

1. Wang, H., Zhang, Y., & Yu, X.: An Overview of Image Caption Generation Methods. Computational Intelligence and Neuroscience, 1–13 (2020). https://doi.org/10.1155/2020/3062706
2. Cohan, M., Khan, A., Mahar, M. S., Hassan, S., Ghafoor, A., & Khan, M.: Image Captioning using Deep Learning: A Systematic Literature Review. Internationa Journal of Advanced
Review Networks for Image Captioning in Hindi Language using Convolutional Neural Network and Gated Recurrent Unit. 

Nugraha, A. A., Arifianto, A., & Suyanto, S. (2019). Automatic Myanmar Image Captioning using CNN and LSTM-Based Language Model. In: Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL), 139–143 (2020).

Sur, C.: Gaussian Smoother Semantic Features (GSSF) - Exploring the Linguistic Aspects of Visual Captioning in Indian Languages (Bengali) Using MSCOCO Framework. arXiv:Cs.CL. 1-12 (2020)

Yilmaz, B. D., Demir, A. E., Sönme, E. B., & Yıldız, T.: Image Captioning in Turkish Language. In: Innovations in Intelligent Systems and Applications Conference. 1-5 (2019). https://doi.org/10.1109/ASYP8272.2019.8946358

Nugraha, A. A., Arifianto, A., & Suyanto, S.: Generating Image Description on Indonesian Language using Convolutional Neural Network and Gated Recurrent Unit. In: 7th International Conference on Information and Communication Technology (ICoICT), 1–6 (2019). https://doi.org/10.1109/ICoICT.2019.8835370

Dhir, R., Mishra, S. K., Saha, S., & Bhattacharyya, P.: A Deep Attention based Framework for Image Caption Generation in Hindi Language. Computación y Sistemas, 23(3), 693–701 (2019). https://doi.org/10.13053/CyS-23-3-3269

Al-muzaini, H. A., Al-yahya, T. N., & Benhidour, H.: Automatic Arabic Image Captioning using RNN-LSTM-Based Language Model and CNN. International Journal of Advanced Computer Science and Applications, 9(6), 67–73 (2018). https://doi.org/10.14569/IJACSA.2018.090610

Yoshikawa, Y., Shigeto, Y., & Takeuchi, A.: STAIR Captions: Constructing a Large-Scale Japanese Image Caption Dataset. arXiv:Cs.CL. 1–5 (2017).

Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H.: A Comprehensive Survey of Deep Learning for Image Captioning. ACM Computing Surveys, 51(6), 118:1-118:36 (2019). https://doi.org/10.1145/3295748

Huang, Y., Chen, J., Ouyang, W., Wan, W., & Xue, Y.: Image Captioning With End-to-End Attribute Detection and Subsequent Attributes Prediction. IEEE Transaction on Image Processing, 29, 4013–4026 (2020). https://doi.org/10.1109/TIP.2020.2969330

Lam, Q. H., Le, Q. D., Nguyen, K. V., & Nguyen, N. L.-T.: UIT-ViIC: A Dataset for the First Evaluation on Vietnamese Image Captioning. arXiv:Cs.CL. 1-12 (2020)

Yang, Zhenyu, & Liu, Q.: ATT-BM-SOM: A Framework of Effectively Choosing Image Information and Optimizing Syntax for Image Captioning. IEEE Access, 8, 50565–50573 (2020). https://doi.org/10.1109/ACCESS.2020.2980578

Yang, Zhilin, Yuan, Y., Wu, Y., Salakhutdinov, R., & Cohen, W. W. Review Networks for Caption Generation. In: 30th Conference on Neural Information Processing Systems, 1–9. (2016).

Bai, S., & An, S.: A Survey on Automatic Image Caption Generation. Neurocomputing, 311, 291–304 (2018). https://doi.org/10.1016/j.neucom.2018.05.080

Staniute, R., & Šešok, D.: A Systematic Literature Review on Image Captioning. Applied Science, 9(2024), 1–20 (2019). https://doi.org/10.3390/app9102024

Li, S., Tao, Z., Li, K., & Fu, Y.: Visual to Text: Survey of Convolutional Neural Networks for Image Captioning. IEEE Transaction on Emerging Topics in Computational Intelligence, 3(4), 297–304 (2018).

Yang, Zhenyu, & Liu, Q.: ATTeNSOM: A Framework of Effectively Choosing Image Information and Optimizing Syntax for Image Captioning. IEEE Access, 8, 50565–50573 (2020). https://doi.org/10.1109/ACCESS.2020.2980578

Yang, Zhilin, Yuan, Y., Wu, Y., Salakhutdinov, R., & Cohen, W. W. Review Networks for Caption Generation. In: 30th Conference on Neural Information Processing Systems, 1–9. (2016).

Bai, S., & An, S.: A Survey on Automatic Image Caption Generation. Neurocomputing, 311, 291–304 (2018). https://doi.org/10.1016/j.neucom.2018.05.080

Staniute, R., & Šešok, D.: A Systematic Literature Review on Image Captioning. Applied Science, 9(2024), 1–20 (2019). https://doi.org/10.3390/app9102024

Li, S., Tao, Z., Li, K., & Fu, Y.: Visual to Text: Survey of Convolutional Neural Networks for Image Captioning. IEEE Transaction on Emerging Topics in Computational Intelligence, 3(4), 297–312 (2019). https://doi.org/10.1109/TETCI.2019.2892755

He, S., Liao, W., Tavakoli, H. R., Yang, M., Rosenhahn, B., & Pugeault, N.: Image Captioning through Image Transformer. arXiv:Cs.CV. 1-17 (2020)

Li, X., Xu, C., Wang, X., Lan, W., Jia, Z., Yang, G., & Xu, J.: COCO-CN for Cross-Lingual Image Tagging, Captioning and Retrieval. IEEE Transaction on Multimedia, 7(7), 1–14 (2019). https://doi.org/10.1109/TMM.2019.2896494

Zhang, B., Zhou, L., Song, S., Chen, L., Jiang, Z., & Zhang, J.: Image Captioning in Chinese and Its Application for Children with Autism Spectrum Disorder. In: Proceedings of the 12th International Conference on Machine Learning and Computing, 426–432 (2020). https://doi.org/10.1109/3383972.3384072

Aung, S. P. P., Pa, W. P., & Nwe, T. L.: Automatic Myanmar Image Captioning using CNN and LSTM-Based Language Model. In: Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL), 139–143 (2020)

Sur, C.: Gaussian Smoother Semantic Features (GSSF) - Exploring the Linguistic Aspects of Visual Captioning in Indian Languages (Bengali) Using MSCOCO Framework. arXiv:Cs.CL. 1-12 (2020)

Yilmaz, B. D., Demir, A. E., Sönme, E. B., & Yıldız, T.: Image Captioning in Turkish Language. In: Innovations in Intelligent Systems and Applications Conference. 1-5 (2019). https://doi.org/10.1109/ASYP8272.2019.8946358

Nugraha, A. A., Arifianto, A., & Suyanto, S.: Generating Image Description on Indonesian Language using Convolutional Neural Network and Gated Recurrent Unit. In: 7th International Conference on Information and Communication Technology (ICoICT), 1–6 (2019). https://doi.org/10.1109/ICoICT.2019.8835370

Dhir, R., Mishra, S. K., Saha, S., & Bhattacharyya, P.: A Deep Attention based Framework for Image Caption Generation in Hindi Language. Computación y Sistemas, 23(3), 693–701 (2019). https://doi.org/10.13053/CyS-23-3-3269

Al-muzaini, H. A., Al-yahya, T. N., & Benhidour, H.: Automatic Arabic Image Captioning using RNN-LSTM-Based Language Model and CNN. International Journal of Advanced Computer Science and Applications, 9(6), 67–73 (2018). https://doi.org/10.14569/IJACSA.2018.090610

Yoshikawa, Y., Shigeto, Y., & Takeuchi, A.: STAIR Captions: Constructing a Large-Scale Japanese Image Caption Dataset. arXiv:Cs.CL. 1–5 (2017).
Adriyendi, A Rapid Review of Image Captioning ...169

23. Song, Y., Chen, S., Zhao, Y., & Jin, Q.: Unpaired Cross-lingual Image Caption Generation with Self-Supervised Rewards. In: Proceedings of the 27th ACM International Conference on Multimedia, 1–9 (2019). https://doi.org/ 10.1145/3343031.3350996

24. Thapliyal, A. V., & Soricut, R. Cross-modal Language Generation using Pivot Stabilization for Web-scale Language Coverage. arXiv:Cs.CL2, 1–11 (2020). https://doi.org/2005.00246v1

25. Alikhani, M., Sharma, P., Li, S., Soricut, R., & Stone, M.: Clue: Cross-modal Coherence Modeling for Caption Generation. arXiv:Cs.CL, 1–11 (2020). https://doi.org/2005.00908v1

26. Liu, F., Ren, X., Liu, Y., Lei, K., & Sun, X.: Exploring and Distilling Cross-Modal Information for Image Captioning. ArXiv:Cs.CV, 1–7 (2020)

27. Wei, Q., Wang, X., & Li, X. Harvesting Deep Models for Cross-Lingual Image Annotation. In: Proceedings of CBMI, 1–5 (2017). https://doi.org/10.1145/3095713.3095751

28. Meetei, L. S., Singh, T. D., & Bandyopadhyay, S. WAT2019: English-Hindi Translation on Hindi Visual Genome Dataset. In: Proceedings of the 6th Workshop on Asian Translation, 181–188 (2019)

29. Lan, W., Li, X., & Don, J. Fluency-Guided Cross-Lingual Image Captioning. arXiv:Cs.CL, 1–9 (2017)

30. Miyazaki, T., & Shimizu, N.: Cross-Lingual Image Caption Generation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1780–1790 (2016)