Thai-IC: Thai Image Captioning based on CNN-RNN Architecture

Pakpoom Mookdarsanit[1*], Lawankorn Mookdarsanit[2]

[1*]Computer Science, Faculty of Science
Chandrakasem Rajabhat University
Bangkok, Thailand
pakpoom.m@chandra.ac.th
[2]Business Computer, Faculty of Management Science
Chandrakasem Rajabhat University
Bangkok, Thailand
lawankorn.s@chandra.ac.th

Abstract—The trend of news are represented in an image with a short description (or caption) and quickly shared on social media. Most short captions in many languages (e.g., English, Indonesian, Myanmar, Chinese, Arabic, etc.) are manually written by human. Instead of human labor, the visual objects within an image have enough information to autonomously generate the caption, called image captioning. Thai image captioning (Thai-IC) is such a new problem in Thai natural language processing (Thai-NLP) to make the model understand the image. This paper proposes an end-to-end deep learning model to generated Thai image caption. The model consists of encoding stage by convolutional neural network (CNN) and decoding stage by recurrent neural network (RNN). Visual geometry group in 16-layers (VGGNet-16) is used to extract visuals from an image as CNN-encoder. The visuals are used to generate Thai captions by Long-short-term memory (LSTM) as RNN-decoder. Thai captioning corpus is constructed by secondary and primary data that has 10,732 images. This Thai-IC is evaluated by Bilingual Evaluation Understudy (BLEU) on the 10-fold cross validation. ***(Abstract)***

Keywords-Thai image captioning; Thai caption generation; Visual semantic information; Thai image description; Visual relation

I. INTRODUCTION

Million images are posted and publicly shared on social media (e.g., Facebook) within minute. The news publication trends to rely on the concept "*short communication*" (as well as tweeting on Twitter) that makes the readers take only short time [1] to get the main information [2], e.g., the real-time traffic news during the car running via *JS-100 Radio* app [3]. Most short information of an image shared on the page is still manually described by human. Since a digital image contains story [4-5], the semantic information can be autonomously generated by visual objects [6], without the help of textual information. As well as a famous quote from the old age "*…An image says more than thousand words…*" that is possible to generate the image caption (IC) using their own contents [7].

IC is the meeting between computer vision (CV) and natural language processing (NLP) [8]. Most recent works are related to deep learning. However, those methods are proposed to English image captioning. Some other languages are also introduced, e.g., Indonesian [9], Myanmar [10], Chinese [11], Arabic [12], etc.

Figure 1. Thai image captioning (Thai-IC), a view from Chao Phraya River – a popular cultural tourist attraction in Bangkok, Thailand

Thai image captioning (Thai-IC) can be categorized as a field of Thai natural language processing (Thai-NLP) that has been being very long history [13-14] with *Thailand's National Electronics and Computer Technology Center (NECTEC)* [15-16]. Some local well-known examples from *AI for Thai* [17] are Thai text to speech engine (*Vaja*: วาจา) [18] and Thai AI anchors (*Suthichai AI*: สุทธิชัย นักข่าวเอไอ) [19].

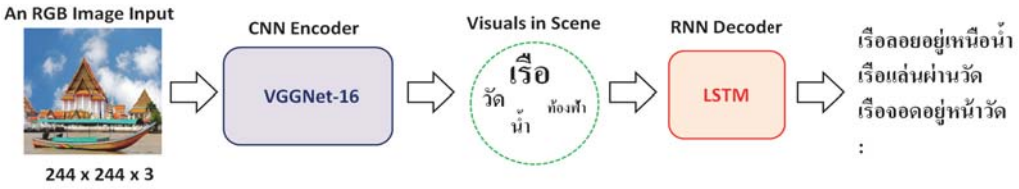

Figure 2. The proposed Thai image captioning based on CNN-RNN Architecture

Thai-NLP researches concerns not only speech but also image contents [20]. Most Thai-NLP applications with images are only related to Thai handwriting recognition [21] or Thai optical character recognition (Thai-OCR) [22], known as image to text conversion (a.k.a. Thai-IMG2TXT) [23].

Thai-IC is still a new Thai-NLP paradigm that is to generate the caption to an unknown image using their visual information. Thai local culture or architecture as semantic visuals/objects can be used to generate Thai caption: Thai folk-gesture [24-25], Buddhism [26-28], Thai national animal [29] and food [30-31]. A previous research on Thai-IC is original and good beginning [32] that generates global scene graph generation (SGG) [33] by English caption. Later, those node relations on SGG are translated into Thai as an indirect approach. However, this method seems to be such a word-level translation that is not enough for natural Thai novel captioning because it is not directed Thai captioning and limited to the given images in database.

To magnify the previous Thai-IC, this paper proposes end-to-end model based on deep learning that constructs a Thai novel caption for an unknown image. The proposed Thai-IC is based on deep learning that uses convolutional neural network (CNN) to extract visual features from an image, called *encoder*. And Thai text understanding (based on word embedding) is modeled by recurrent neural network (RNN), called *decoder*. For the architecture, CNN is implemented by pre-trained and fine-tuned Visual Geometry Group 16 layers (VGGNet-16) [34] and RNN is Long-short-term memory (LSTM) [35]. Thai image captioning corpus consists of secondary and primary data that contains 10,732 images with Thai captions.

This paper is organized into 5 sections. The proposed Thai image captioning framework is in section 2. The section 3 talks about Thai image captioning corpus. Experiments are described in section 4. The conclusion is finally summarized in section 5.

II. Proposed Thai Image Captioning Framework

The proposed Thai image captioning (Thai-IC) framework is shown in Fig.2. The first stage is called "*encoder*" that makes visual understanding from an image by convolutional neural network (CNN). The second one is "*decoder*" that makes language understanding by recurrent neural network (RNN).

A. Encoder

Convolutional neural network (CNN) is successful in computer vision applications. As well as [36], Visual Geometry Group (VGGNet) with pre-training has shown the captioning performance in image understanding as *encoder*. Historically, VGGNet [34] was the winner in Large Scale Visual Recognition Challenge 2013 (ILSVRC 2013) [37] competition. VGGNet-16 (16 layers with ImageNet *pre-training*) was shown to be compact but great performance. An (RGB) image input is resized into 224x224x3. Each image has 4,096 feature size. The output is a 1,000-sized vector as the representation of an image, as shown in Fig.3.

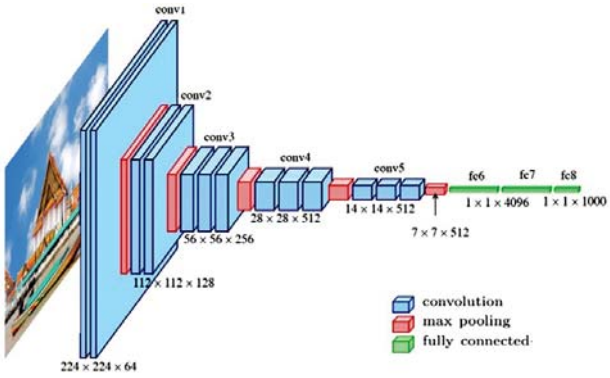

Figure 3. VGGNet-16 architecture as *CNN-encoder*

As well as ImageNet [38] *pre-training*, some local images with some specific visuals/objects taken from any Thai events are also cropped and trained to the model (as *fine-tuning*), as in Fig.4.

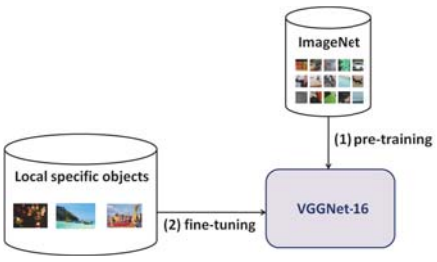

Figure 4. VGGNet-16 training: *pre-training* and *fine-tuning*

B. Decoder

Thai caption is generated as a result of the previous visual understanding in encoder, as if the encoder informs the recurrent neural network (RNN) what visuals are included in the image. RNN is used in word-level embedding as *decoder*. Long-short-term memory (LSTM) is used to sequentially generate Thai words or phrases by computing the probability from image and the previous words/phrases as a sequence-to-sequence (seq-to-seq) model. The embedding weights for Thai are learnt during the model training. The language understanding is trained to predict the next word within a Thai caption. For catching up the overfitting problem (in Fig.5), dropout and batch normalization are added to perform before the soft-max layer.

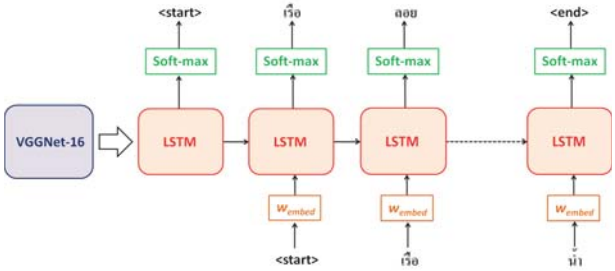

Figure 5. LSTM architecture as *RNN-decoder*

For working, the Thai-IC model based on CNN-RNN architecture can be described as following steps:

Step 1: CNN architecture (based on pre-trained VGGNet-16) is used to visual extraction from an input RGB image (I_{RGB}) that consists of convolution, Max pooling and Batch normalization, by (1).

$$x_0 = VGG_{16}(I_{RGB}) \tag{1}$$

Step 2: The sequential input of RNN architecture (based on LSTM) can be computed from the sequence of words (s_t), by (2).

$$x_t = w_{embed} s_t \tag{2}$$

Step 3: Next words probability can be iteratively computed by LSTM with the previous hidden memory (h_{t-1}), by (3).

$$h_t = LSTM(x_t, h_{t-1}) \tag{3}$$

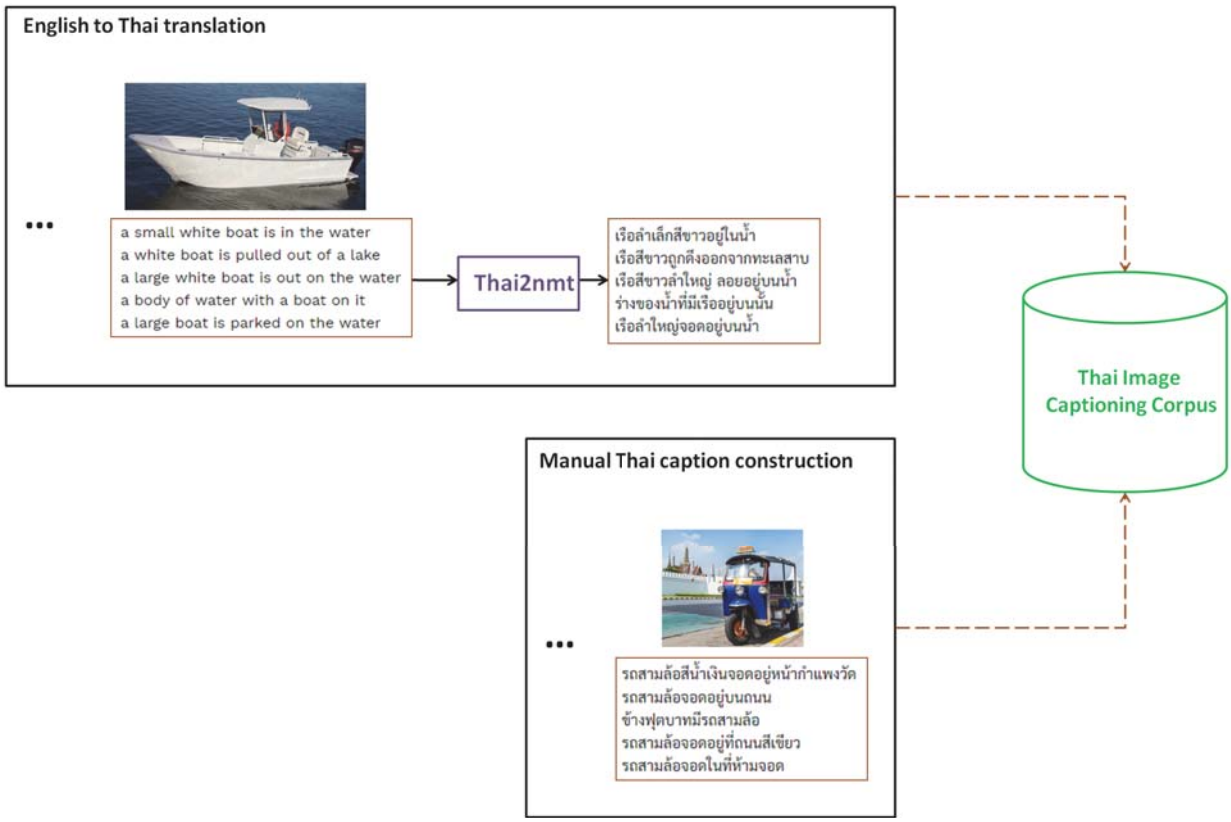

Figure 6. Construction of Thai image captioning corpus

(a) English: *Many Buddha images are in front of the great Thai painting wall*

(b) English: *A few krathongs are floating on the water*

(c) English: *A candle parade is moving on the road*

(d) English: *Many lights are soaring into the sky.*

Figure 7. Images with final caption generations

Note that the loss function for CNN-RNN architecture can be computed by probability and predicted word at time (t) defined by (4).

$$Loss(I_{RGB}, S) = -\sum_{t=1}^{n} \log p_t(s_t) \quad (4)$$

III. Thai Image Captioning Corpus

For constructing Thai image caption corpus (as in Fig.6), both secondary and primary data (10,732 images) is applied to the experiment. This section can be categorized into 1) English to Thai translation and 2) manual Thai caption construction.

A. English to Thai Translation

Firstly, Flickr8k dataset [39-40] (as secondary data) contains 8,902 images with English image captions. Each image has 5 different captions. Those English captions are directly translated into Thai texts by *VISTEC thai2nmt* [41]. This machine translation based on Transformer totally reduces the time for manual captioning.

B. Manual Thai Caption Contruction

Another one is to manually correct images in local events as primary data: Loi Krathong (Thai: ลอยกระทง), Songkran (Thai: สงกรานต์), Royal Barge Procession (Thai: กระบวนพยุหยาตราชลมารค), Chinese New Year (Thai: ตรุษจีน), Royal Ploughing (Thai: แรกนาขวัญ) and other Thai events. The primary data contains 1,830 images that are manually captioned in different 5 Thai texts.

IV. Experiments

This section talks about the experimental details that can be categorized into 1) experimental environment and metric and 2) results.

A. Experimental Environment and Metric

The proposed Thai image captioning was trained on K80 GPU cloud machine using Keras library. The learning model was set as 10 epoch. For the evaluation, the Bilingual Evaluation Understudy (BLEU) metric was use to evaluate the correctness of translation in each generated Thai caption. The BLEU value ranged from 0 (min) to 1 (max). The higher value was better. The BLEU formulation could be computed by (5).

$$BLEU = \min\left(1, \frac{length_{output}}{length_{reference}}\right)\left(\prod_{i=1}^{4} precision_i\right)^{\frac{1}{4}} \quad (5)$$

B. Results

The dataset was divided into training and testing set. According to 10-fold cross validation, all 10,732 images were randomly grouped into 10 partitions. The first 9 partitions were used to train the proposed CNN-RNN model that each partition had 1,073 images. And another one had 1,075 images that were used to for model testing. The BLEU scores computed by (5) in each fold are shown in Table 1.

TABLE I. The 10-Fold Cross Validation

Fold	BLEU-1	BLEU-2	BLEU-3	BLEU-4
1	70.5	51.3	41.9	29.4
2	61.6	47.5	44.3	21.7
3	65.7	52.7	39	28.5
4	65.2	45.7	43.8	22.8
5	68	55.3	35.7	34
6	67.2	46.8	41.4	30.4
7	62	54.7	43	24.3
8	58.7	47.2	38.9	31.1
9	67.2	50	36.2	26.7
10	61.6	47	42.2	23
Mean	**64.77**	**49.82**	**40.64**	**27.19**

The proposed Thai-IC modeled accurately the relationship between visuals and correctly generated Thai caption in Fig.7(a). In Fig.7(b) and Fig.7(c), the significant visuals were well-summarized with the correct captioning, although the Fig.7(c) seemed to be similar to Thai Royal Barge Procession (Thai: กระบวนพยุหยาตราชลมารค) – an official Thai royal event in the river. On the contrary, the model

wrongly generated caption in Fig.7(d) "*Many lights are soaring into the sky*" that the ground-truth information was the lights were hung on the wires – an image taken from Chinese New Year Event (Thai: เทศกาลตรุษจีน). This CNN-RNN model's weights and parameters can be transferred to a newer model with more data.

V. CONCLUSION

The proposed Thai image caption (Thai-IC) relates to the solution of manual image captions by human labor. This paper introduces the first end-to-end Thai-IC based on CNN-RNN architecture as a new problem of Thai natural language processing (Thai-NLP). The secondary images from Flicklr8k and primary ones were used to construct Thai-IC corpus in this experiment. For *encoder*, the visual features from an image were extracted by Visual geometry group 16-layers (VGGNet-16). VGG-16 was pre-trained on ImageNet and fine-tuned on the local visuals and objects from Thai events. For *decoder*, Thai caption were generated to visual features by Long-short-term memory (LSTM). To construct Thai-IC corpus, some secondary images with captions were crawled from Flickr8k dataset and those captions were translated to Thai. Other primary images with Thai texts were manually captioned. With the machine translation moving forward, the corpus can be grown by global large-scale dataset and multi-language captioning. To post many images on social media, Thai caption automatically generated by their visuals within an image as the trend of real-time short communication.

VI. ACKNOWLEDGEMENT

Toward the short message communication in social media, Thai caption can be automatically generated by itself using the semantic visuals/objects within an image. This pre-trained model can be requested to continue more Thai-IC works or compare the experiments by the authors' email. All computational hardware and software were supported by Chandrakasem Rajabhat University.

REFERENCES

[1] K. Shuster, S. Humeau, H. Hu, A. Bordes and J. Weston, "Engaging Image Captioning via Personality," The 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 12508-12518.

[2] I. Laina, C. Rupprecht and N. Navab, "Towards Unsupervised Image Captioning With Shared Multimodal Embeddings," The 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, South Korea, 2019, pp. 7413-7423.

[3] "Live Traffic Reports from JS100 Radio Added to Nostra Map App," [Online]. Available: https https://www.nationthailand.com/business/30228564 [Accessed: 8 July 2020].

[4] L. Soimart and P. Mookdarsanit, "Name with GPS Auto-tagging of Thai-tourist Attractions from An Image," The 2017 Technology Innovation Management and Engineering Science International Conference, Nakhon Pathom, Thailand, 2017, pp. 211-217.

[5] P. Mookdarsanit and L. Mookdarsanit, "Contextual Image Classification towards Metadata Annotation of Thai-tourist Attractions," in ITMSoc Transactions on Information Technology Management, vol.3, no.1, pp. 32-40, 2018.

[6] A. Olaode and G. Naghdy, "Review of the application of machine learning to the automatic semantic annotation of images," in IET Image Processing, vol. 13, no. 8, pp. 1232-1245, 2019.

[7] S. Li, Z. Tao, K. Li and Y. Fu, "Visual to Text: Survey of Image and Video Captioning," in IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 3, no. 4, pp. 297-312, 2019.

[8] MD. Z. Hossain, F. Sohel, M. F. Shiratuddin and H. Laga "A Comprehensive Survey of Deep Learning for Image Captioning," in ACM Computing Surveys, vol. 51, no. 6, pp.1-36, 2019.

[9] A. A. Nugraha, A. Arifianto and Suyanto, "Generating Image Description on Indonesian Language using Convolutional Neural Network and Gated Recurrent Unit," The 2019 International Conference on Information and Communication Technology (ICoICT), Kuala Lumpur, Malaysia, 2019, pp. 1-6.

[10] S. P. P. Aung, W. P. Pa and T. L. New, "Automatic Myanmar Image Captioning using CNN and LSTM-based Language Model," The 2020 Joint Workshop on Spoken Language Technologies for Under-resourced languages and Collaboration and Computing for Under-Resourced Languages (SLTU-CCURL), Marseille, France, 2020, pp.139-143.

[11] C. Zhang, Y. Dai, Y. Cheng, Z. Jia and K. Hirota, "Recurrent Attention LSTM Model for Image Chinese Caption Generation," The 2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS), Toyama, Japan, 2018, pp. 808-813.

[12] J. Zakraoui, S. Elloumi, J. M. Alja'am and S. Ben Yahia, "Improving Arabic Text to Image Mapping Using a Robust Machine Learning Technique," in IEEE Access, vol. 7, pp. 18772-18782, 2019.

[13] H. Thaweesak Koanantakool, T. Karoonboonyanan and C. Wutiwiwatchai, "Computers and the Thai Language," in IEEE Annals of the History of Computing, vol. 31, no. 1, pp. 46-61, 2009.

[14] C. Tapsai, P. Meesad and H. Unger, "An Overview on the Development of Thai Natural Language Processing", in Information Technology Journal, vol. 15, no. 2, pp. 45-52, 2019.

[15] M. Jotisakulratana, N. Koomgun, R. Virojrid and B. Udomsaph, "Applying International Patent Classification (IPC) to strategic planning processes of an R&D organization: The case of NECTEC, Thailand," The PICMET '13: Technology Management in the IT-Driven Services (PICMET), San Jose, CA, 2013, pp. 1913-1918.

[16] C. Wutiwiwatchai, V. Chunwijitra, S. Chunwijitra, P. Sertsi, S. Kasuriya, P. Chootrakool and K. Thangthai, "The NECTEC 2015 Thai Open-Domain Automatic Speech Recognition System," International Symposium on Natural Language Processing (SNLP), Ayutthaya, Thailand, 2016, pp. 124-136.

[17] "AI for Thai," [Online]. Available: https://www.nectec.or.th/research/research-project/aiforthai-digitaltransformation.html [Accessed: 8 July 2020]. [in Thai].

[18] "VAJA Text-to-Speech Engine," [Online]. Available: https://www.nectec.or.th/innovation/innovation-mobile-application/vaja.html [Accessed: 8 July 2020]. [in Thai].

[19] "Thailand's first AI journalist named 'Suthichai AI' unveiled," [Online]. Available: https://www.thaipbsworld.com/thailands-first-ai-journalist-named-suthichai-ai-unveiled/ [Accessed: 8 July 2020].

[20] C. Tanprasert and S. Sae-Tang, "Thai type style recognition," The 1999 IEEE International Symposium on Circuits and Systems (ISCAS), Orlando, FL, 1999, pp. 336-339.

[21] P. Mookdarsanit and L. Mookdarsanit, "ThaiWrittenNet: Thai Handwritten Script Recognition Using Deep Neural Networks," in Azerbaijan Journal of High Performance Computing, vol. 3, no. 1, pp. 75-93.

[22] T. Emsawas and B. Kijsirikul, "Thai printed character recognition using long short-term memory and vertical component shifting," The 2016 Pacific Rim International Conference on Artificial Intelligence, Phuket, Thailand, 2016, pp. 106-115.

[23] C. Tanprasert and T. Koanantakool, "Thai OCR: a neural network application," The 1996 Digital Processing Applications (TENCON), Perth, WA, Australia, 1996, pp. 90-95.

[24] P. Mookdarsanit and L. Mookdarsanit, "An Automatic Image Tagging of Thai Dance's Gestures," Joint Conference on ACTIS & NCOBA, Ayutthaya, Thailand, 2018, pp. 76-80.

[25] P. Mookdarsanit and L. Mookdarsanit, "A Content-based Image Retrieval of Muay-Thai Folklores by Salient Region Matching," in International Journal of Applied Computer Technology and Information Systems, vol.7, no.2, pp.21-26, 2018.

[26] P. Mookdarsanit and M. Rattanasiriwongwut, "GPS Determination of Thai-temple Arts from a Single Photo," The 11th International Conference on on Applied Computer Technology and Information Systems, Bangkok, Thailand, 2017, pp. 42-47.

[27] P. Mookdarsanit and M. Rattanasiriwongwut, "MONTEAN Framework: A Magnificent Outstanding Native-Thai and Ecclesiastical Art Network," in International Journal of Applied Computer Technology and Information Systems, vol.6, no.2, pp.17-22, 2017.

[28] L. Mookdarsanit, "The Intelligent Genuine Validation beyond Online Buddhist Amulet Market," in International Journal of Applied Computer and Information Systems, vol. 9, no.2, pp. 7-11, 2020.

[29] L. Mookdarsanit and P. Mookdarsanit, "SiamFishNet: The Deep Investigation of Siamese Fighting Fishes," in International Journal of Applied Computer Technology and Information Systems, vol.8, no.2, pp. 40-46, 2019.

[30] L. Soimart and P. Mookdarsanit, "Ingredients estimation and recommendation of Thai-foods," in SNRU Journal of Science and Technology, vol.9, no.2, pp.509-520, 2017.

[31] P. Mookdarsanit and L. Mookdarsanit, "Name and Recipe Estimation of Thai-desserts beyond Image Tagging," in Kasembundit Engineering Journal, vol.8, Special Issue, pp.193-203, 2018.

[32] P. Khuphiran, S. Kajkamhaeng and C. Chantrapornchai, "Thai Scene Graph Generation from Images and Applications," The 2019 International Computer Science and Engineering Conference (ICSEC), Phuket, Thailand, 2019, pp. 361-365.

[33] D. Liu, M. Bober and J. Kittler, "Visual Semantic Information Pursuit: A Survey," in arXiv:1903.05434, 2019.

[34] K. Simonyan and A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition," The 3rd International Conference on Learning Representations, San Diego, CA, 2015.

[35] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko and T. Darrell, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 677-691, 2017.

[36] V. Mullachery and V. Motwani, "Image Captioning," in arXiv: 1805.09137, 2018.

[37] "Large Scale Visual Recognition Challenge 2013 (ILSVRC2013)," [Online]. Available: http://image-net.org/challenges/LSVRC/2013/http://image-net.org/challenges/LSVRC/2013/ [Accessed: 8 July 2020].

[38] J. Deng, W. Dong, R. Socher, L. Li, Kai Li and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," The 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 248-255.

[39] "Flickr8K," [Online]. Available: https://www.kaggle.com/shadabhussain/flickr8k [Accessed: 2 February 2020].

[40] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," The 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 3128-3137.

[41] "vistec-AI/thai2nmt," [Online]. Available: https://github.com/vistec-AI/thai2nmt?fbclid=IwAR2CRfPnlEpEykrBV3h62JrOuUBPnH2tUswI9Vf1x-gCkxeVXogM2pPNlsk [Accessed: 21 July 2020].