Automatic image captioning in Thai for house defect using a deep learning-based approach

Thitirat Siriborvornratanakul[1*], Manadda Jaruschaimongkol[1†], Krittin Satirapiwong[1†], Kittipan Pipatsattayanuwong[1†], Suwant Temviriyakul[1†] and Ratchanat Sangprasert[1†]

[1*]Graduate School of Applied Statistics, National Institute of Development Administration, 148 Serithai Road, Klong-Chan, Bangkapi, 10240, Bangkok, THAILAND.
[1]Graduate School of Applied Statistics, National Institute of Development Administration, 148 Serithai Road, Klong-Chan, Bangkapi, 10240, Bangkok, THAILAND.

*Corresponding author(s). E-mail(s): thitirat@as.nida.ac.th;
Contributing authors: manadda.jar@stu.nida.ac.th; krittin.sat@stu.nida.ac.th; kittipan.pip@stu.nida.ac.th; suwant.tem@stu.nida.ac.th; ratchanat.san@stu.nida.ac.th;
†These authors contributed equally to this work.

Abstract

A house inspection enables you to make a more informed choice regarding the house you are thinking about purchasing. Making an inspection report, an inspector typically takes between one and two hours, depending on the size of the house and the number of defects by inserting all defect images in excel and manually captioning what defects should be resolved. This report represents property 's risk includes environmental, financial, or physical factors. We tried to find research related to this work but still not found. However, there were some studies and techniques that can be used for house inspections. Many researchers used encoder-decoder which is a deep learning-based model to generate captions. So, we applied this approach in our work by using VGG16, MobileNet, InceptionV3 for an encoder, GRU for a decode, and added one more technique which is an additive attention mechanism, Bahdanau, to enhance model performance. The experimental results showed VGG16

takes the least time to train a model, while MobileNet gets the highest BLEU-1 to BLEU-4 score of 0.866, 0.850, 0.823, and 0.728 respectively, and loss value is the same for all models. By the way, we suggested to use InceptionV3 in this work because of the best performance from attention plotted when compared to VGG16 and MobileNet and the difference of BLEU score of three models were not significant. There are some challengs in our works. Firstly, our dataset varies and is unique (long-tailed class), which affects imbalanced traning data. Second, some images have many captions, but our model can predict only one caption. The future works can improve our model by gathering more training data for some classes or using the techniques that can be dealt with imbalanced dataset and change the model architecture to generate more than one caption.

Keywords: Thai image captioning, house defect, attention based image captioning

1 Introduction

A house inspection is an observation and reporting on the real estate property's condition and always happens when it is in the selling state in the market. The house inspectors who are engineers or professionals evaluate the condition of the property. They also look for evidence of damage or any problems that may affect the property's value.

There are three important factors in construction projects including time, cost, and quality. The contractor has to balance these factors during the construction process. However, time and cost are important factors to focus on because reducing project duration can increase the cost and a delay in one activity leads to increase project duration.

For house construction, project quality may decrease project duration and cost. It will cause some major and minor defects in house construction. To address these problems, the buyer has to hire an engineer or professional who will identify major and minor defects within a house before purchasing. If a buyer is not comfortable with the finding of defects, a buyer can cancel the offer to buy. After the house inspection, the buyer will receive an inspection report including defects and captions that ask the seller for repairs defects that were found in the house. On the other hand, having an inspection done for sellers before they are putting their houses on the market may increase sales.

The duration of making an inspection report can vary depending on house size and number of defects and takes time between one and two hours. For preparing the inspection report, the inspector has to insert many pictures and describes how are the problems and use this report to inform the seller to fix these problems. To improve the process of preparing the inspection report, there is a need to develop a method that can help inspectors improve the process to prepare an inspection report.

Automatic image captioning, which is the procedure for producing a textual description of a picture using both Natural Language Processing (NLP) and Computer Vision, can provide the information that describes the content of an image automatically. Therefore, the use of image captioning can help the house inspector to improve the process of preparing house inspection report such as reducing the time and can use to train junior inspector or staff with less technical skill. Following are the main contribution of our study:

1. We developed deep learning model using image captioning with attention machanism to generate Thai caption for house defect images. It can be used for the process of inspection report to reduce time and cost.
2. We shared the house defect images with Thai caption for further study in https://github.com/manadda-j/deep-learning/tree/main/04ImageCaptioning/Dataset which has never been shared before.

2 Related works

In this section, we review the related works on automatic image captioning. Many studies on automatic image captioning techniques include encoder-decoder architecture-based, attention-based, unified encoder-decoder and other techniques to improve model performance.

An encoder-decoder architecture-base method used a Convolutional Neural Network (CNN) to extract the features from images and Recurrent Neural Networks (RNNs) to generate image captions. For example, Pakpoom and Lawankorn (2020) [8] proposed a Thai image captioning (Thai-IC) which is a deep learning model that generate Thai image caption. The model consists of encoding stage by CNN using VGG-16 with pre-trained on ImageNet and decoding stage by using LSTM. The Bilingual Evaluation Understudy (BLEU) metric was used to evaluate on the 10-fold cross validation and get the average BLEU-4 score of 0.2719. Seshadri et al., (2020) [10] proposed an Inception injected encoder model using InceptionV3 as image encoder and LSTM as decoder on the Flickr8k and Google's Conceptual Captions dataset, the model achieved 0.13, 0.14, and 0.18 on BLEU, METEOR and ROUGE. Geetha et al. (2020) [5] proposed a model using pre-trained VGG-19 on ImageNet data as encoder and GRU as decoder for labeling the satellite picture to describe atmospherical conditions and caption of land cover or land use.

The attention-based image captioning method enabled the model to selectively focus on the relevant part of images while ignoring others in the model and also have a better result than encoder-decoder architecture-based. For example, Chun et al. (2021) [4] developed the model that can detect the status of damage and generate sentences for bridge images by using CNN and GRU with the attention mechanism. The dataset has 3,118 bridge photos that have been taken during bridge inspection work. The results show that the model with the attention mechanism given the better performance which has BLEU-4 score of 0.693 and the percentage of correct completed sentences is 69.3%.

Khan et al., (2022) [7] developed the deep learning model by using 4 pre-trained CNNs including InceptionV3, DenseNet169, ResNet101, and VGG16 as an encoder and used GRU with Bahdanau attention as a decoder to increase the model performance. They have compared their models with state-of-art models on the MS COCO dataset. The results show that their models have better performance than state-of-art models which BLEU, Rouge, CIDER, and Meteor have 0.37, 0.59, 1.109, and 0.29 respectively. Chu et al., (2020) [3] proposed Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention (AICRL) which is an architecture consisting of one encoder and one decoder with soft attention. The AICRL model gave the BLEU-4, METEOR, and CIDEr scores of 0.326, 0.261, and 0.872 on the MS COCO 2014 dataset.

To create a new pre-training method, Zhou et al., (2020) [14] proposed a unified encoder-decoder model named the Vision-Language Pre-training (VLP). For encoding and decoding, they use a shared multi-layer transformer network, pre-trained on images with captioning, and optimized for both bidirectional and sequence to sequence masked language prediction. In this paper, they experimented with Visual Question Answering (VQA) and image captioning tasks on MS COCO, Flickr30k, and VQA dataset. The results show that using unified VLP outperforms on MS COCO has scored 35.5, 28.5, 118.0, and 21.6 for BLEU-4, METEOR, CIDER, and SPICE metrics, got 67.4 on overall accuracy on VQA2.0, and all of metrics on Flickr30k has scored 29.7, 23.8, 69.1, and 17.6 for the same metrics as MS COCO. This model is more generally relevant and meaningful vision-language image that easy to fine-tuned for tasks like VQA and image captioning and reach state-of-the-art for both tasks using a single model architecture.

Other works also explored many techniques to improve model performance. For example, Chang et al., (2022) [2] enhanced image captioning using the color recognition method. They purposed a ROS-based image captioning model which uses VGG16 as an encoder for features extraction and uses LSTM with attention as a decoder for semantic language processing. Furthermore, to do object detection and color recognition, Mask R-CNN with OpenCV is applied. For their dataset, they used the MS COCO dataset and self-made traffic signal. The algorithm will first recognize the object, then the segmented image extracted the color of each pixel and converted it into HSV. The results show this model can apply for both images that contain unique objects and similar objects. The model can better describe the image mood and add the color detail to the recognized object resulting in a better understanding of the image. Singh et al., (2022) [12] proposed a novel show, attend, and tell model (ATM) for medical image captioning and used a Strength Pareto Evolutionary Algorithm-II (SPEA-II) to choose the initial values of the model's parameter. This technique can improve over other models at 1.178% of medical image captioning. Atliha and Šešok (2022) [1] proposed a model-compression technique for image captioning to reduce size of model and can be used the model on mobile devices. The result showed that the model can reach 127.4 CIDEr and 21.4 SPICE and the model size was reduced from 791.8 MB to 34.8 MB.

A summary of used deep learning-based approach for image captioning shown in Table 1. To compare these related works, there is still room for improvement in field of image captioning. Existing researches show their performance on open datasets such as MS COCO, Flickr30K, VQA and these model performances achieve on general images. However, image captioning for house defect cannot be generated through existing models because the open dataset does not train on house defect context. Therefore, our study aims to contribute a training dataset for house defect and build a deep learning model for image captioning in Thai language that can help house inspector working with inspection report efficiency and use BLEU metric which is good for short sentence for measuring image captioning performance.

Table 1: Summary of used deep learning-based approach for image captioning

Reference	Image Encoder	Language Model	Dataset	Evaluation Metrics
Chu et al. (2020) [3]	ResNet50	LSTM with Soft Attention	MS COCO, Flickr30k	BLEU, METEOR, CIDEr
Geetha et al. (2020) [5]	VGG-19	GRU	Satellite Images	Accuracy
Pakpoom and Lawankorn (2020) [8]	VGG-16	LSTM	Flickr8k, Thai event images	BLEU
Seshadri et al. (2020) [10]	InceptionV3	LSTM	Flickr8k, Google's Conceptual Captions	BLEU, METEOR, ROUGE
Zhou et al. (2020) [14]	Transformer, BERT, VLP	Transformer, BERT, VLP	MS COCO, Flickr30k, VQA	BLEU, METEOR, CIDEr, SPICE
Chun et al. (2021) [4]	InceptionV3	GRU with Attention	Bridge Defects	BLEU
Atliha and Šešok (2022) [1]	ResNet101, MobileNetV2, EfficientDet	LSTM	MS COCO	BLEU, METEOR, ROUGE-L, CIDEr, SPICE
Chang et al. (2022) [2]	VGG-16	LSTM with Attention	MS COCO	N/A
Khan et al. (2022) [7]	InceptionV3, DenseNet169, ResNet101, VGG-16	GRU with Bahdanau Attention	MS COCO	BLEU, METEOR, ROUGE, CIDEr
Singh et al. (2022) [12]	DT, MLP, RF, SVM, ANN	LSTM	Biomedical Image	MSE, Accuracy, F-measure, Specificity, Sensitivity, Kappa Statistics

3 Proposed method

In this paper, we proposed the encoder-decoder recurrent neural network to automatically generate Thai captions for house defects. Moreover, we added an

attention mechanism to our proposed models to enhance their ability to process information. The details are shown and discussed in the following subsections.

3.1 Pre-trained Image Model

We used three pre-trained image models which are VGG-16, MobileNet, and InceptionV3. These models are trained on several images from the ImageNet dataset. We used the different model architectures because we compare the results from each model. And for the reason why we chose these models as our encoders, we will talk about it in the next part of each model.

1. VGG-16 is a Convolutional Neural Network (CNN) that was submitted to the ILSVRC (ImageNet) Challenge in 2014, and it won the competition in localization task while bagging 2nd position in the classification task. However, it was not the only reason we chose this model, but also its simple architecture and support for the notion of deeper CNNs for enhanced performance [11]. We started with downloaded its architecture using Keras, then excluded the 3 fully connected layers at the top of the network by setting include top to False and used pre-training on ImageNet. We froze all trainable parameters and then removed classification layers.
2. MobileNet is a lightweight deep neural network based on a streamlined architecture that uses depth-wise separable convolutions. And we can develop our model for mobile applications in the future by using this architecture [6]. The step to use this architecture same as VGG-16 that is to download the model using Keras, exclude fully-connected layers at the top of the network, freeze all trainable parameters and then removed the softmax layers.
3. InceptonV3 was developed by Google which is a convolutional neural network architecture that makes several improvements from the Inception family. The reason that we selected this model is it has proven to be more computationally effective than VGGNet [13]. By the way, e applied the same steps from both VGG-16 and MobileNet to prepare this model.

3.2 Attention based Architecture

We adopted the attention mechanisms because we want to pay attention to a few keywords rather than a single vector containing information about the entire sentence. The attention will help our decoder determines how much attention should be placed on each word in the input by using attention weights which help the translation better. Bahdanau attention was applied in our paper. It is an additive attention that combines the decoder states and the encoder states linearly, and learns to translate and align at the same time. We can observe which elements of the image the model emphasizes as it creates a caption from its results. Our model is shown in the Figure 1.

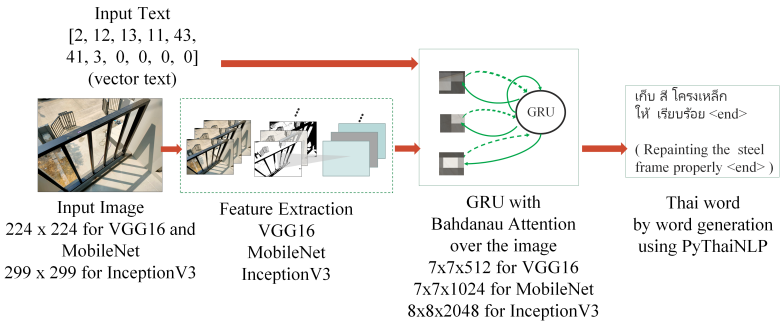


Fig. 1: Architecture overview

4 Data Preparation

4.1 Dataset

The dataset of house defect images were created by house inspectors during inspection work in Bangkok from 2018 to 2021 about 40 projects and type of residential buildings included individual houses, town houses and condominiums. The image captioning that describe house defect was collected from inspection report which have one caption for each image. An example of dataset is shown in Figure 2. The dataset has 4000 images and 4000 captions. Our dataset is long-tail distribution shown in Figure 3, so we select type of defects that more than 20 images for this study. Therefore, we have only 840 images and 840 captions for 16 type of house defects and then we randomly split the dataset 70% into a training set and 30% into a test set. The Figure 4 shown the number of type of house defects in training set and test set

After we split the dataset into training set and test set, our dataset is imbalance. Theregfore, we applied data augmentation techniques to increase the amount of dataset. The types of data augmentation included image rotations, image shifts, image flips, image brightness and image zoom shown in Figure 5. Finally, we have 2,617 training images and 1,110 test images and shown the number of type of house defects in training set and test set in Figure 6.

4.2 Data Preprocessing

Data Preprocessing is separated into two stages.

1. Image Pre-Processing

   - Grouping the images belong to house defect captions.
   - Apply data augmentation to images including image rotations, image shifts, image flips, image brightness and image zoom.
   - Resized image that suit for each feature extraction model for VGG16 resized images to 224x224, MobileNet resized images to 224x224 and InceptionV3 resized images to 299x299.

(a) Clean the wall and repainting properly

(b) Clean the septic tank properly

(c) Replater the crack and repainting properly

(d) Install the screws properly

(e) Repaint the steel frame properly

(f) Stick back the laminate

Fig. 2: Examples of dataset

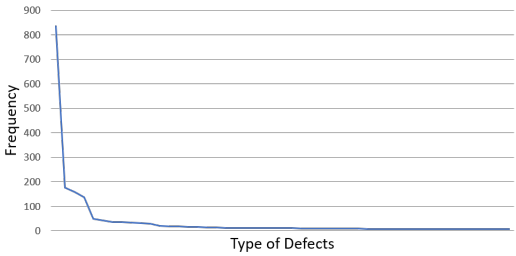


Fig. 3: Defect Frequency Distribution

2. Text Pre-Processing

- Tokenized the captions to obtain a unique vocabulary using PythaiNLP [9] with deep cut engine.
- Added start and end tags for every caption to let the model understand where the start and end of each caption.
- Covert text to vector and padding all the sequence to the same length as longest sentence using tf.keras.layers.TextVectorization

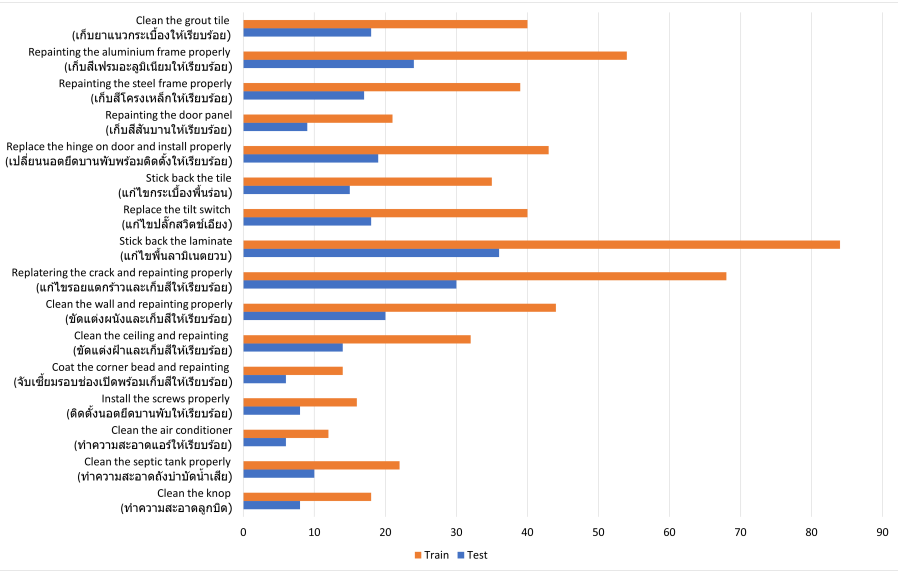
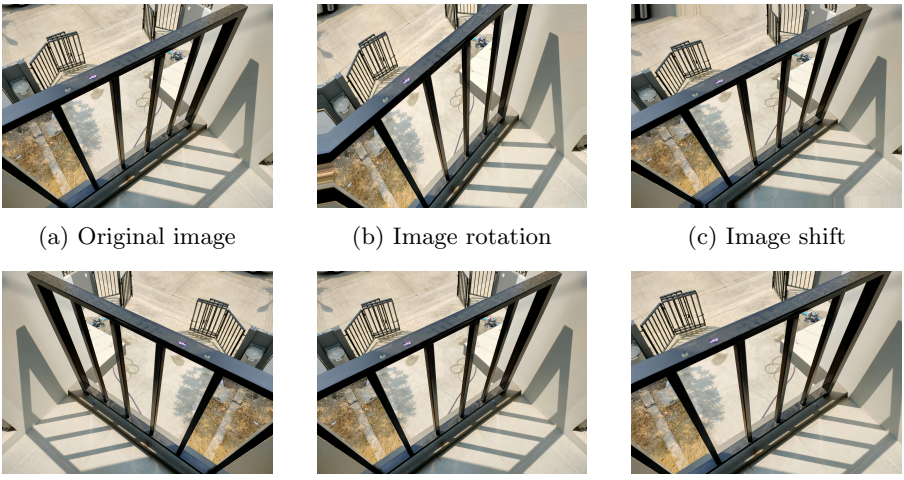


Fig. 4: Total number of house defects by caption



(a) Original image (b) Image rotation (c) Image shift

(d) Image flip (e) Image brightness (f) Image zoom

Fig. 5: Data augmentation on dataset

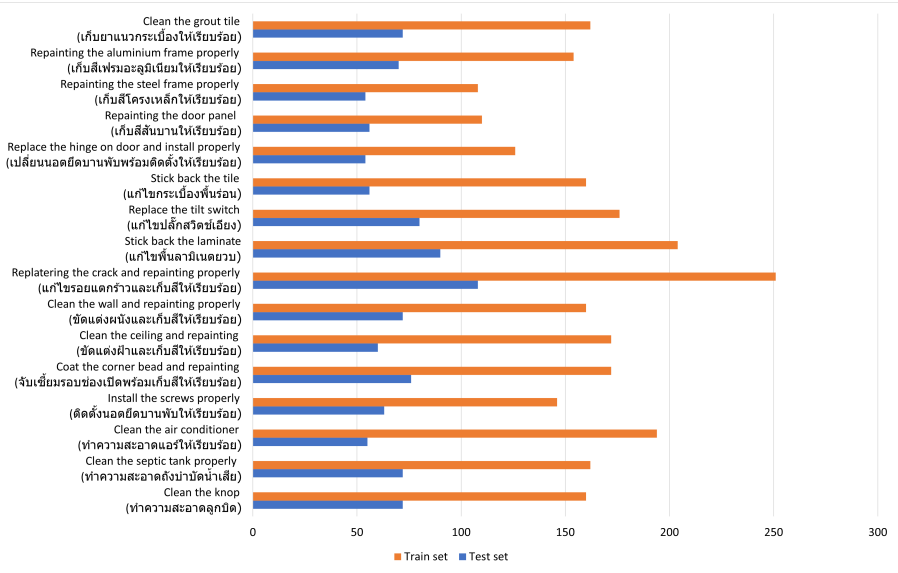


Fig. 6: Total number of house defects by caption after augmentation

5 Experimental Results and Discussion

5.1 Training

In this stage, the encoder models are applied to extract images feature. Before we trained the model, we downloaded the models architecture and ImageNet weight. We fed the dataset after pre-processing to the models.

The model ran 30 epochs and used batch size as 16. We used Adam for model optimizer with Sparse Categorical Cross Entropy as loss function. Our model was trained on GPU Tesla P100 via Google Colaboratory (Colab) and used Keras library version 2.8.0 for building the model.

Figure 7 shows that after models are trained for 15 epochs, loss value close to 0 until the 30th epoch. The loss values are the same for all models which equal to 0. For training time, we see from Figure 8 and see that VGG16 is the model that used the least time which took 3.5 sec/ep, follow by MobileNet which took around 4.2 sec/ep, and InceptionV3 is the model that uses the most time which is 4.5 sec/ep. However, we assume that if we use MobileNetV3 which is the small model architecture instead of MobileNet that we use in this research, it will be the model that takes the least time. To test our hypothesis, this model can apply in the future work.

5.2 Evaluation

To measure the correctness of generated Thai captions. Our study used the Bilingual Evaluation Understudy (BLEU) as an evaluator. The BLEU score is evaluating between generated sentence and reference sentence. The range of

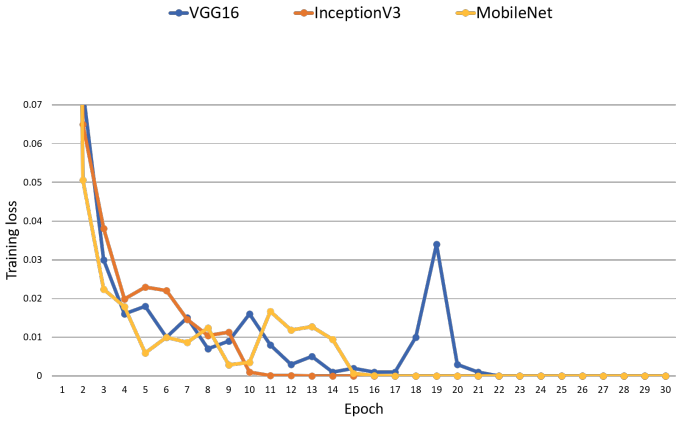


Fig. 7: Training loss

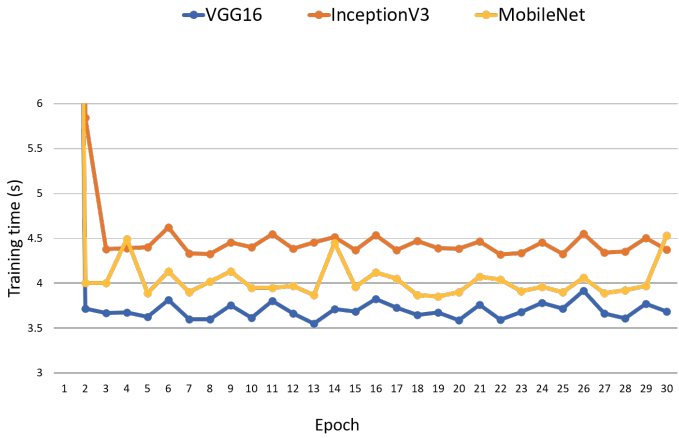


Fig. 8: Training time

scores is 0 to 1. A score closer to 1 mean the perfect match result. The BLEU have the formula as follows:

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

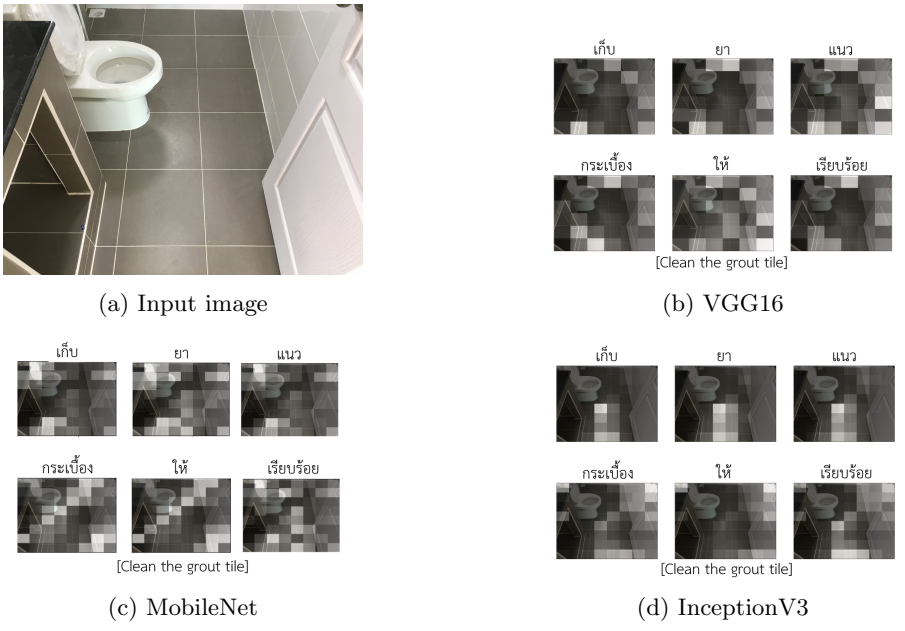
Our research has compared the results of generated Thai caption on the test set from different encoders including VGG16, MobileNet and InceptionV3

with GRU and Bahdanau Attention by using the BLEU score. The results as shown in Table 2.

Table 2: BiLingual Evaluation Understudy (BLEU) score of generated captions on test set

	BLEU-1	BLEU-2	BLEU-3	BLEU-4
VGG16	0.850	0.830	0.792	0.699
InceptionV3	0.842	0.820	0.781	0.697
MobileNet	**0.866**	**0.850**	**0.823**	**0.728**

From Table 2. The model that uses MobileNet as encoder has the highest BLEU score around 0.82 followed by VGG16 and InceptionV3 which are 0.79. Although our BLEU scores were high, they were not significant among the three models. We also plotted the attention plot to see which part of the image that model focused on predicted Thai words. In attention plot shown in Figure 9 10 11, the performance of encoder using InceptionV3 focused on a particular point of images was better when compared to VGG16 and MobileNet which were scatter-focused on the image.



(a) Input image

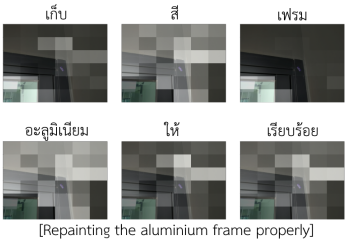
(b) VGG16

(c) MobileNet

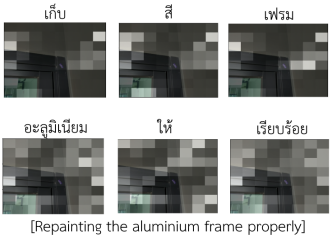
(d) InceptionV3

Fig. 9: Ground Truth Caption "Clean the grout tile"

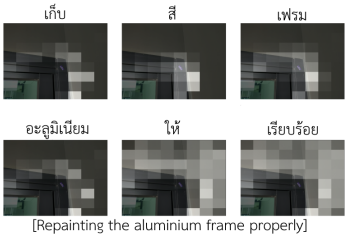
(a) Input image



(b) VGG16



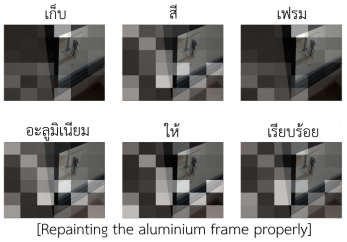
(c) MobileNet



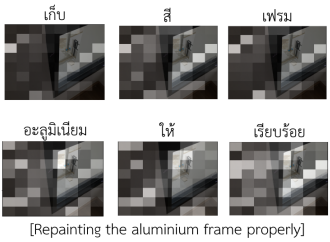
(d) InceptionV3

Fig. 10: Ground Truth Caption "Repainting the Aluminium frame properly"

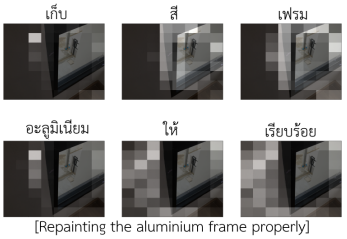
(a) Input image



(b) VGG16



(c) MobileNet



(d) InceptionV3

Fig. 11: Ground Truth Caption "Repainting the Aluminium frame properly"

6 Conclusion and Future Work

Developing a model to help an inspector create an inspection report is our expectation. To do that we used a deep learning model and benchmark their performance to select the best one. This study uses encoder-decoder model structure with attention mechanism to perform image captioning. We used VGG16, MobileNet, and InceptionV3 for an encoder, GRU for a decode, and added Bahdanau as an attention to enhance model performance. Our models was trained to generate Thai caption for House inspections that can help inspector to generate Thai caption of each tasks to fix in house inspection class instead of manul create caption. After training the models, our models show good performance evaluated by BLEU metric. MobileNet has highest BLEU score followed by VGG16 and InceptionV3. We have investigated the result of incorrect caption. We suggested to use InceptionV3 in this work because of the best performance from attention plotted when compared to VGG16 and MobileNet and the difference of BLEU score of three models were not significant. There are some challengs in our works. Firstly, our dataset varies and is unique (long-tailed class), which affects imbalanced traning data. Another one is our model can predict only one caption but some images have many captions. To improve our model perform better, the next steps are gathering more training data for some classes which will help the model know more about the defects' class and combining more techniques that can be dealt with imbalanced dataset, in case of some defects' class are rarely to find, and the last one is change the model architecture to generate more than one caption. These steps are the future works that will help our model more genralize.

References

[1] Atliha, V., and Šešok, D. (2022). Image-Captioning Model Compression. Applied Sciences (Switzerland), 12(3). https://doi.org/10.3390/app12031638.

[2] Chang, Y. H., Chen, Y. J., Huang, R. H., and Yu, Y. T. (2022). Enhanced image captioning with color recognition using deep learning methods. Applied Sciences (Switzerland), 12(1). https://doi.org/10.3390/app12010209.

[3] Chu, Y., Yue, X., Yu, L., Sergei, M., and Wang, Z. (2020). Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention. Wireless Communications and Mobile Computing, 2020. https://doi.org/10.1155/2020/8909458.

[4] Chun, P., Yamane, T., and Maemura, Y. (2021). A deep learning-based image captioning method to automatically generate comprehensive explanations of bridge damage. Computer-Aided Civil and Infrastructure Engineering, 2021.

[5] Geetha, G., Kirthigadevi, T., Ponsam, G. G., Karthik, T., and Safa, M. (2020). Image Captioning Using Deep Convolutional Neural Networks (CNNs). Journal of Physics: Conference Series, 1712(1). https://doi.org/10.1088/1742-6596/1712/1/012015.

[6] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. http://arxiv.org/abs/1704.04861.

[7] Khan, R., Islam, M. S., Kanwal, K., Iqbal, M., Hossain, M. I., and Ye, Z. (2022). A Deep Neural Framework for Image Caption Generation Using GRU-Based Attention Mechanism. i. http://arxiv.org/abs/2203.01594.

[8] Mookdarsanit, P., and Mookdarsanit, L. (2020). Thai-IC: Thai Image Captioning based on CNN-RNN Architecture. 10(1), 40–45.

[9] Phatthiyaphaibun, W., Chaovavanich, K., Polpanumas, C., Suriyawongkul, A., Lowphansirikul, L., and Chormai, P. (2016). PyThaiNLP: Thai Natural Language Processing in Python. Zenodo. http://doi.org/10.5281/zenodo.3519354.

[10] Seshadri, M., Srikanth, M., and Belov, M. (2020). Image to Language Understanding: Captioning approach. http://arxiv.org/abs/2002.09536.

[11] Simonyan, K., and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 1–14.

[12] Singh, A., Krishna Raguru, J., Prasad, G., Chauhan, S., Tiwari, P. K., Zaguia, A., and Ullah, M. A. (2022). Medical Image Captioning Using Optimized Deep Learning Model. Computational Intelligence and Neuroscience, 2022, 1–9. https://doi.org/10.1155/2022/9638438.

[13] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December, 2818–2826. https://doi.org/10.1109/CVPR.2016.308.

[14] Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J. J., and Gao, J. (2020). Unified vision-language pre-training for image captioning and VQA. AAAI 2020 - 34th AAAI Conference on Artificial Intelligence, 13041–13049. https://doi.org/10.1609/aaai.v34i07.7005.