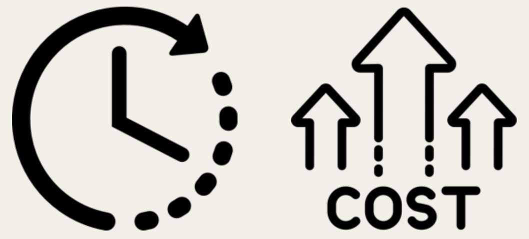
Automatic image captioning in Thai for house defects using a deep learning-based approach

Present by
6220421004 Suwant Temviriyakul
6220421005 Ratchanat Sangprasert
6220421007 Manadda Jaruschaimongkol
6220422017 Kittipan Pipatsattayanuwong
6220422046 Krittin Satirapiwong

Because



It takes time and costs to do the inspection report

House inspector can benefit from image captioning that can help to improve the process of preparing inspection report

Automatic image captioning in Thai for house defects can use to train junior inspector or staff with less technical skill

Having an inspection done for sellers before they are putting their houses on the market may increase sales

To improve the process of preparing the inspection report,
automatic image captioning in Thai for house defects will help.

Scope

- Develop a model that help to generate image captioning in Thai for house defects using a deep learning-based approach

16 classes

- The model can generate image captioning for 16 classes of house defects

Limitation

- Insufficient dataset
- 1 image 1 caption

Work Process

Data Preparation

- Acquire the dataset
- Data labeling/grouping
- Data Augmentation
- Text Tokenization
- Text Preprocessing

Image Captioning

- Encoder using VGG16, MobileNetV1, InceptionV3
- Decoder using GRU
- Bahdanau Attention

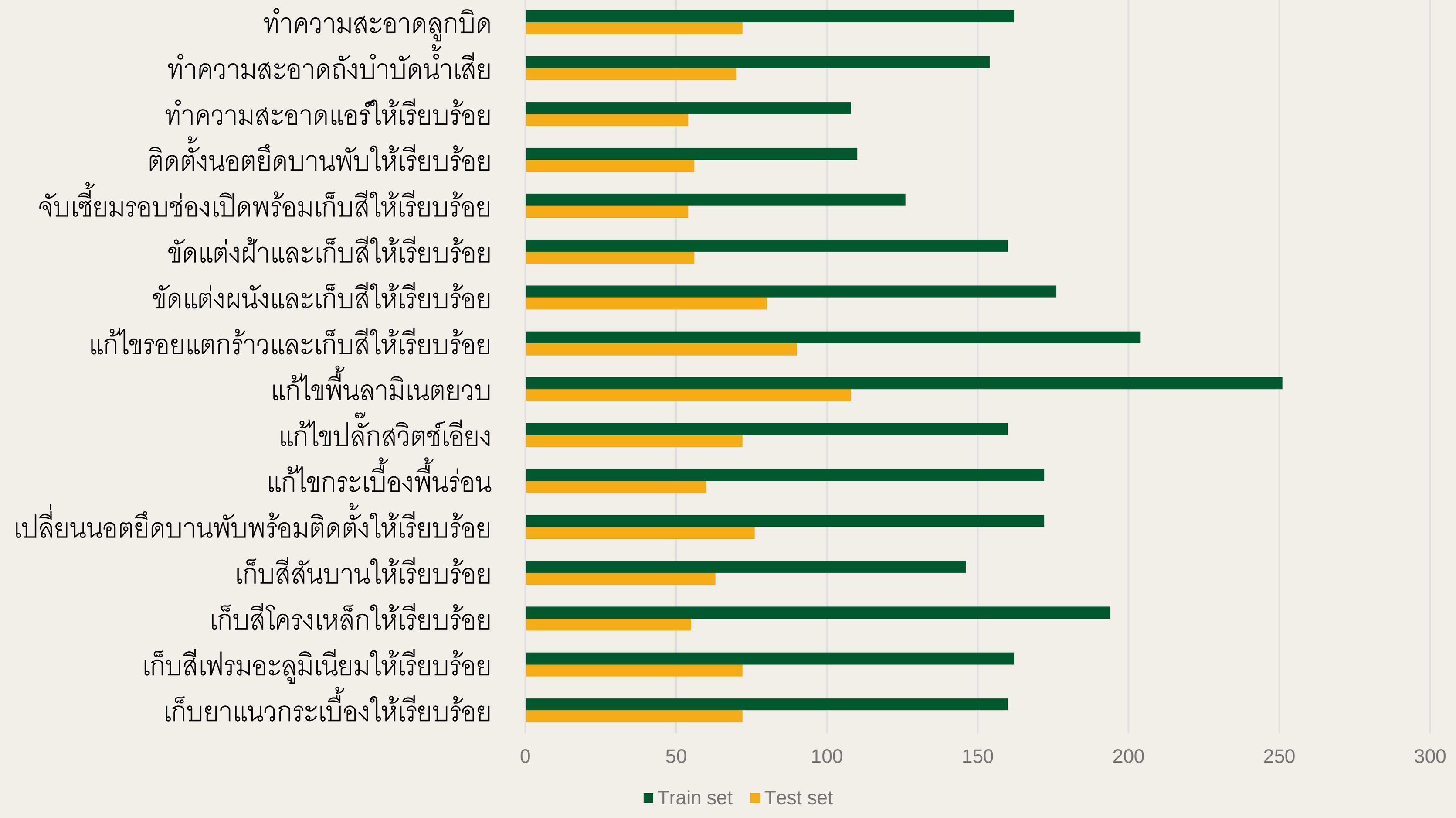
Model Evaluation

- Training Loss
- Training Time
- BLEU (BiLingual Evaluation Understudy)

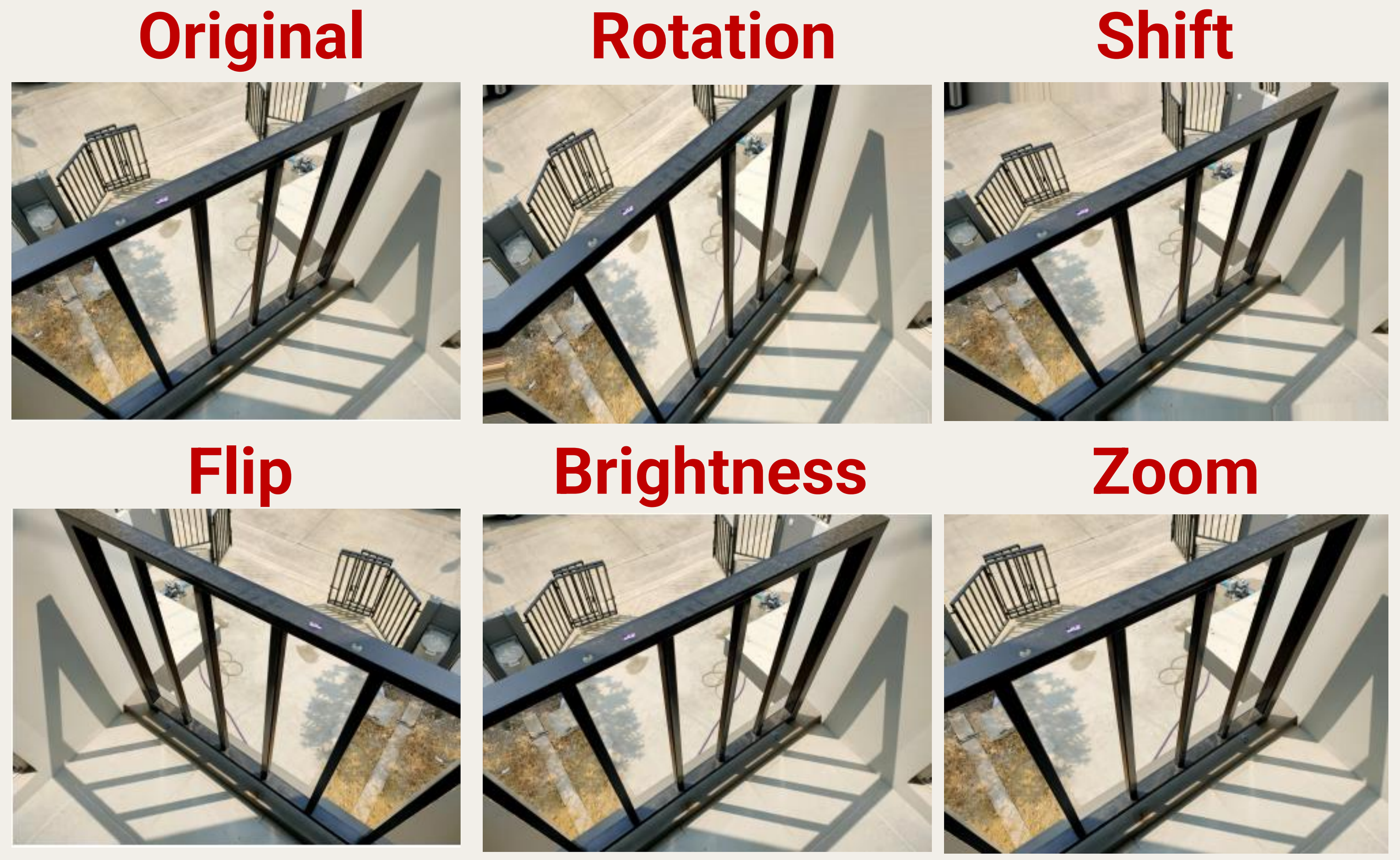
Data Preparation

Image Pre-processing

1. Manual labeling and grouping images for each class

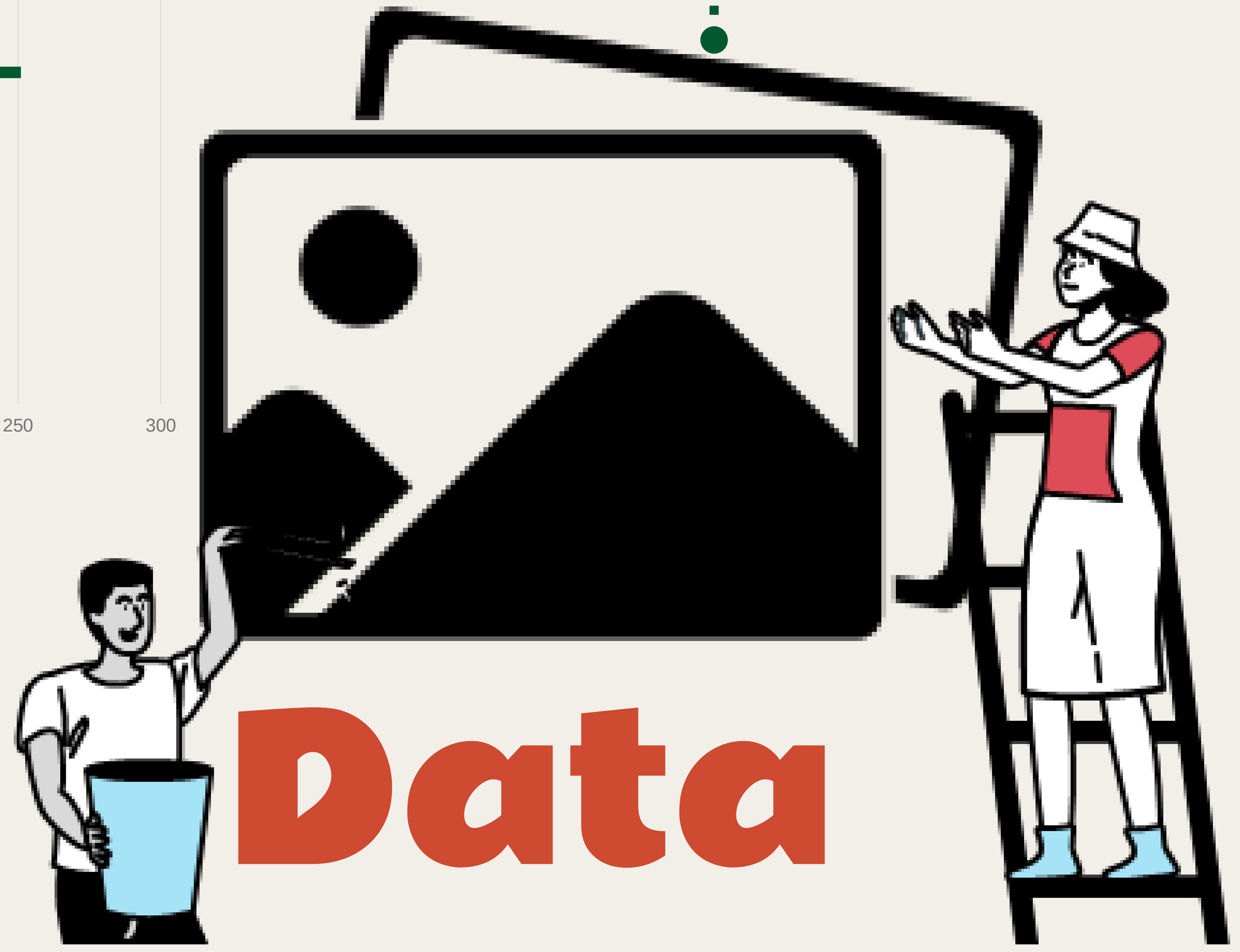


2. Data Augmentation



3. Resized images that suit for each model

- VGG16 224 x 224
- MobileNet 224 x 224
- InceptionV3 299 x 299



Text Pre-processing

1. Tokenized the captions
using PyThaiNLP with deep cut engine

เก็บสีโครงเหล็กให้เรียบร้อย
เก็บ สี โครงเหล็ก ให้ เรียบร้อย

2. Added start and end tags for every caption
to help model understands the start and end of each caption

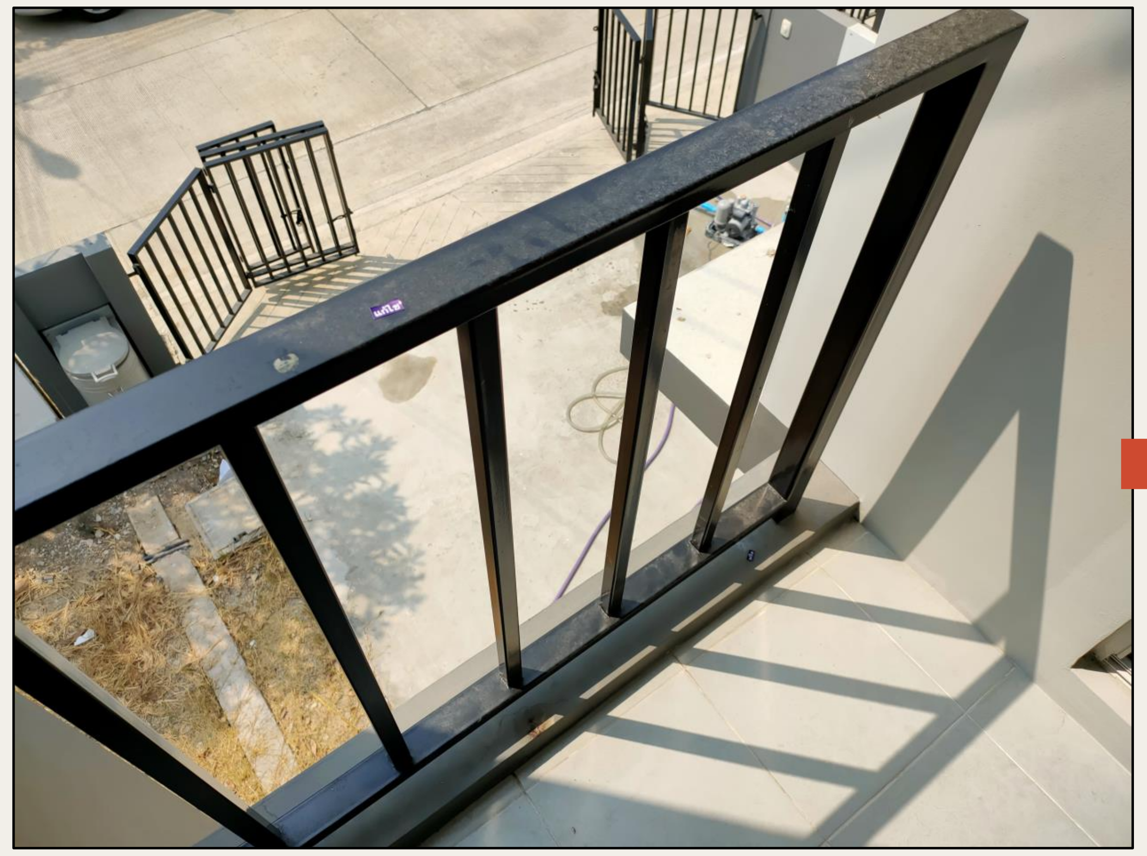
<start> เก็บ สี โครงเหล็ก ให้ เรียบร้อย <end>

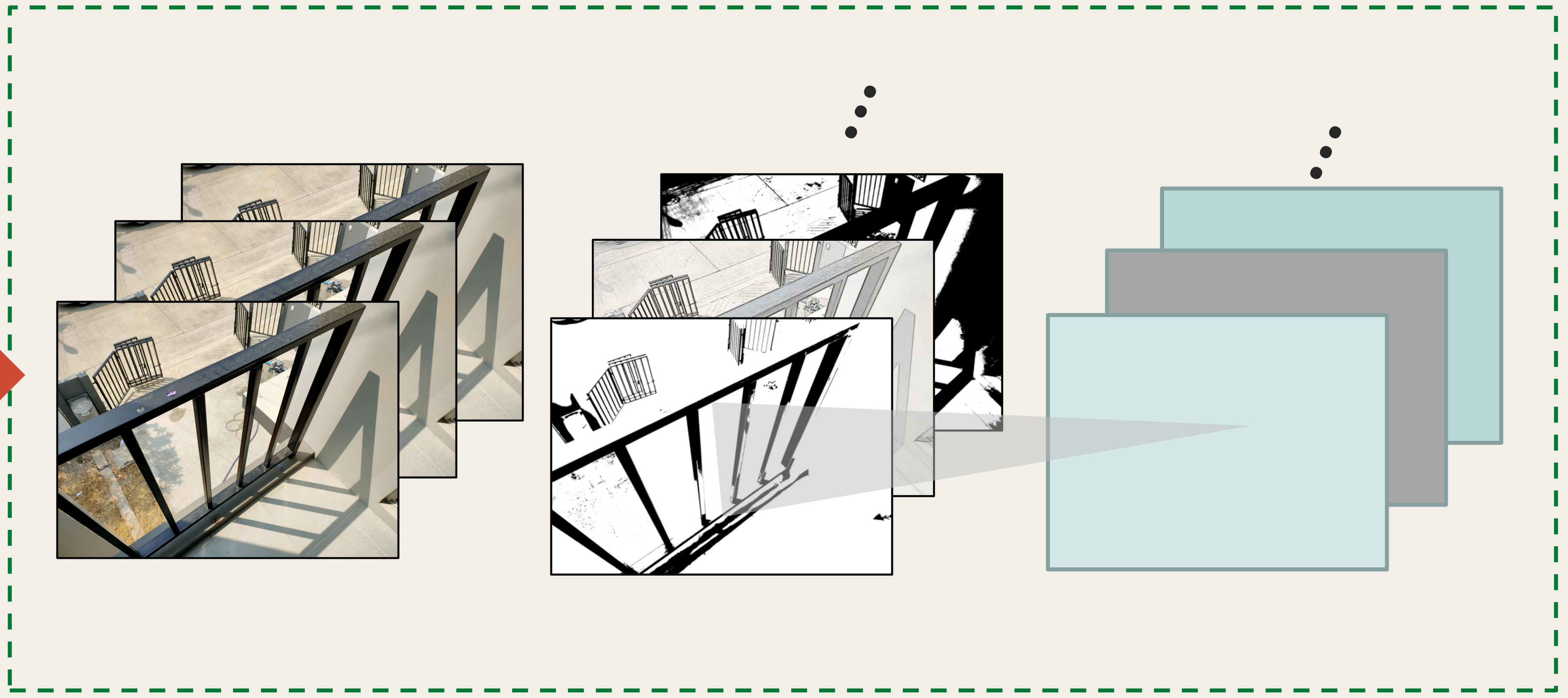
3. Covert text to vector and padding all the sequence to the same length as longest sentence
using tf.keras.layers.TextVectorization

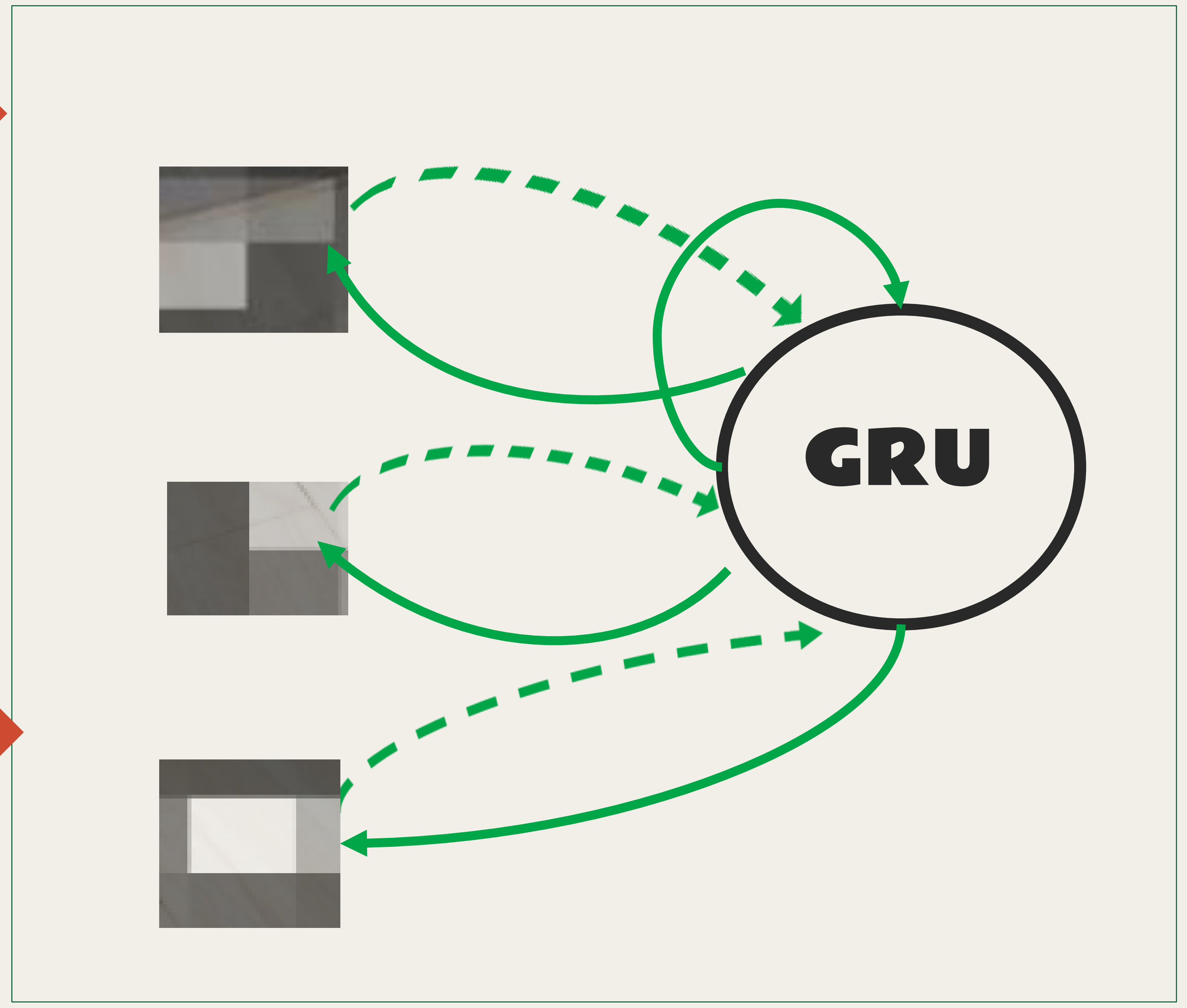
[2, 12, 13, 11, 43, 41, 3, 0, 0, 0, 0]

Model **Architecture**

1.
Input Text
[2, 12, 13, 11, 43, 41, 3, 0, 0, 0, 0]
vector text







เก็บ
สี
โครงเหล็ก
ให้
เรียบร้อย
<end>

1.
Input Image
224 x 224 for VGG16 and MobileNet
299 x 299 for InceptionV3

2.
Feature Extraction
VGG16
MobileNet
InceptionV3

3.
GRU with
Bahdanau Attention
over the image
7x7x512 for VGG16
7x7x1024 for MobileNet
8x8x2048 for InceptionV3

4.
Thai word
by word
generation

Weight=ImageNet
Epoch=30
Optimizer=Adam
Loss Function=Sparse Categorical Cross Entropy

Image, Caption, And Attention Results

Real Caption:
เก็บ ยา แนว กระเบื้อง ให้ เรียบร้อย

VGG16

Predicted Caption

เก็บ

ยา

แนว

กระเบื้อง

ให้

เรียบร้อย

Inception V3

Predicted Caption

เก็บ

ยา

แนว

กระเบื้อง

ให้

เรียบร้อย

MobileNet

Predicted Caption

เก็บ

ยา

แนว

กระเบื้อง

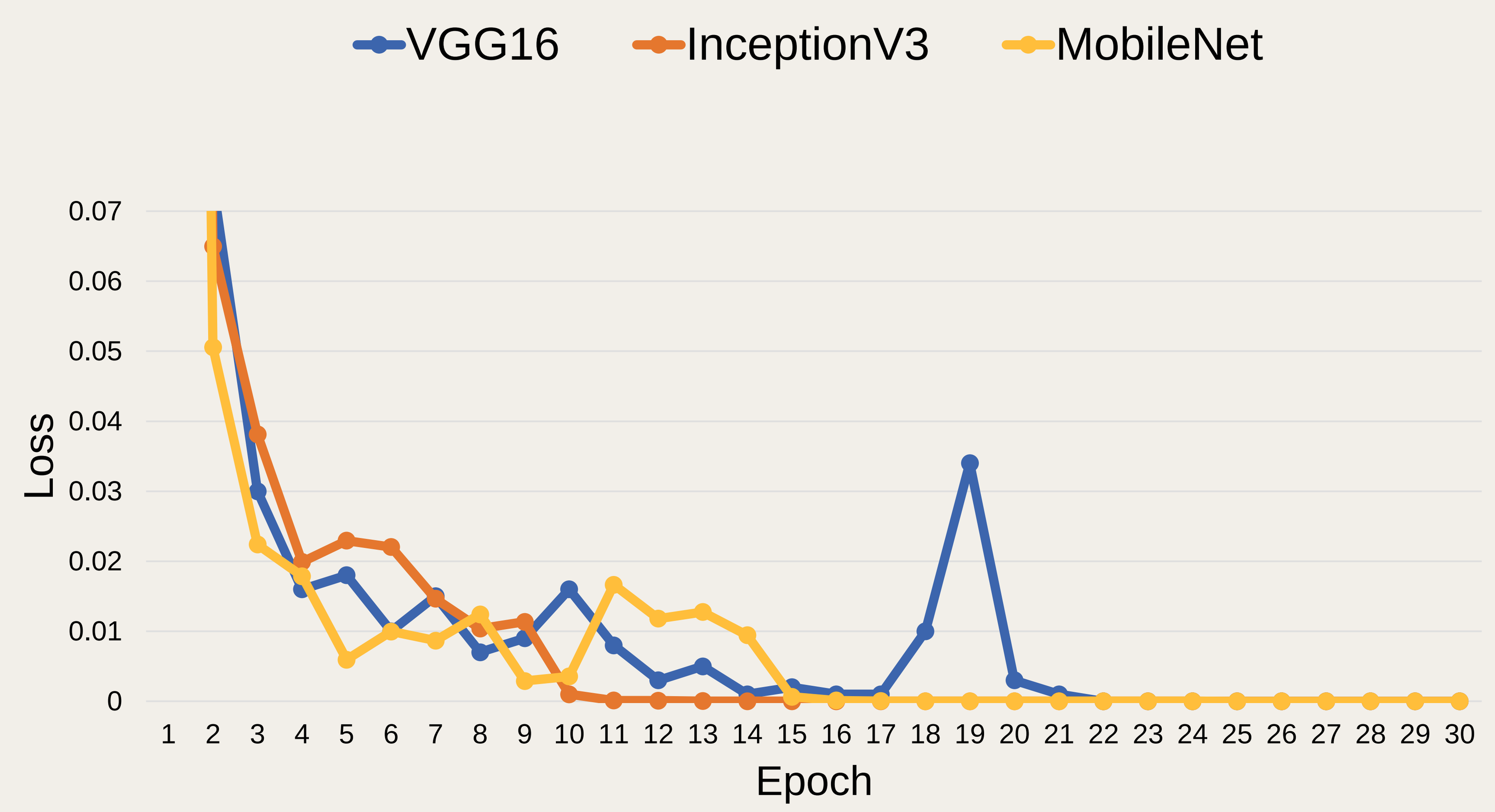
ให้

เรียบร้อย

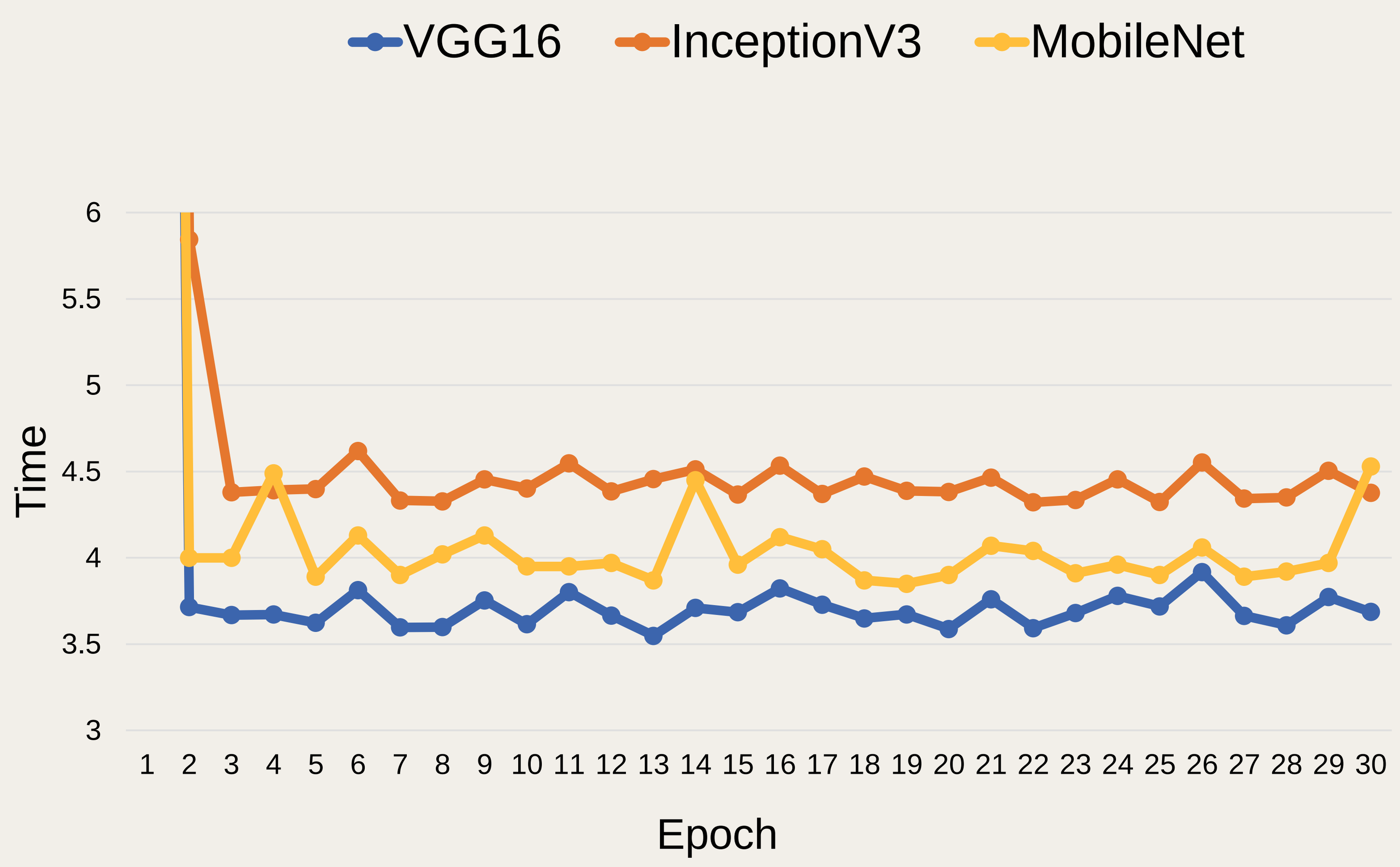
Model Evaluation

Training Performance

Loss

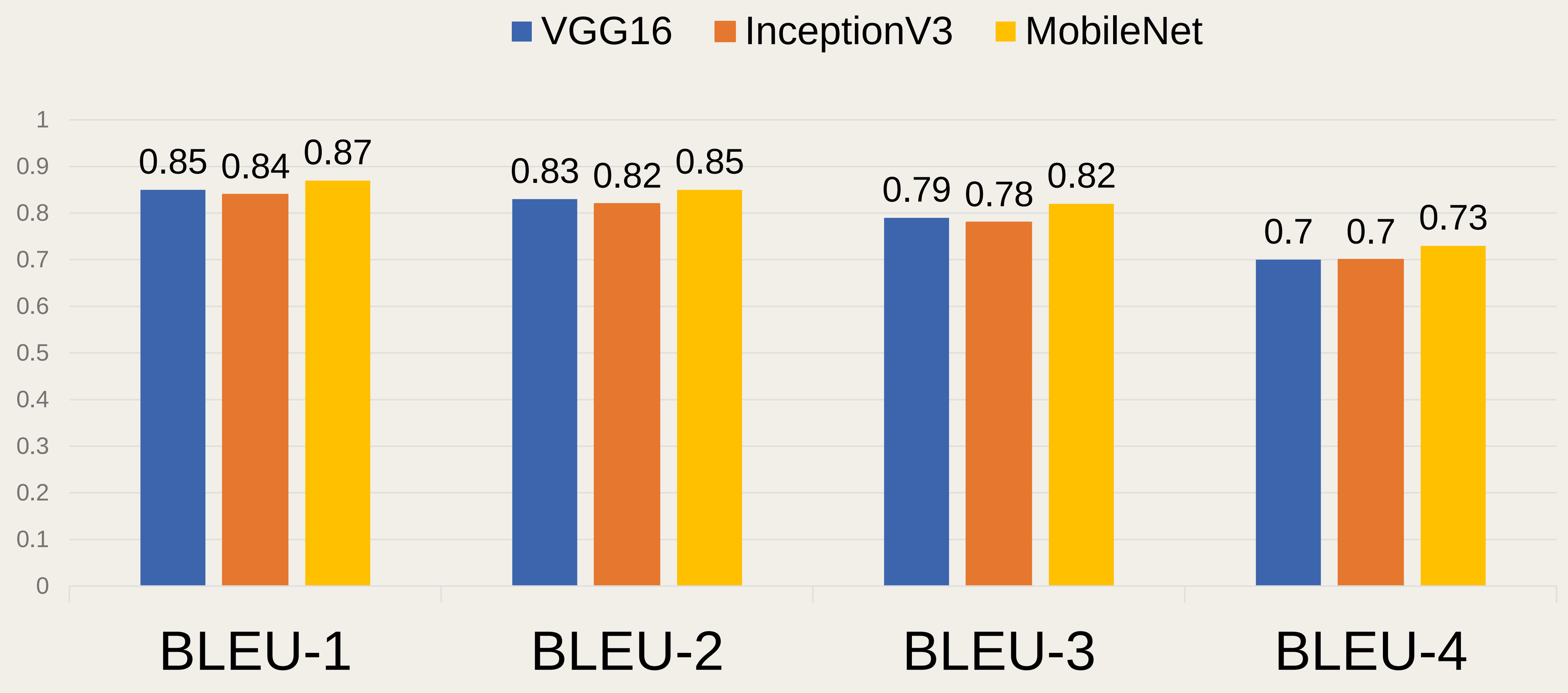


Time



ImageCaptioning Performance

BLEU

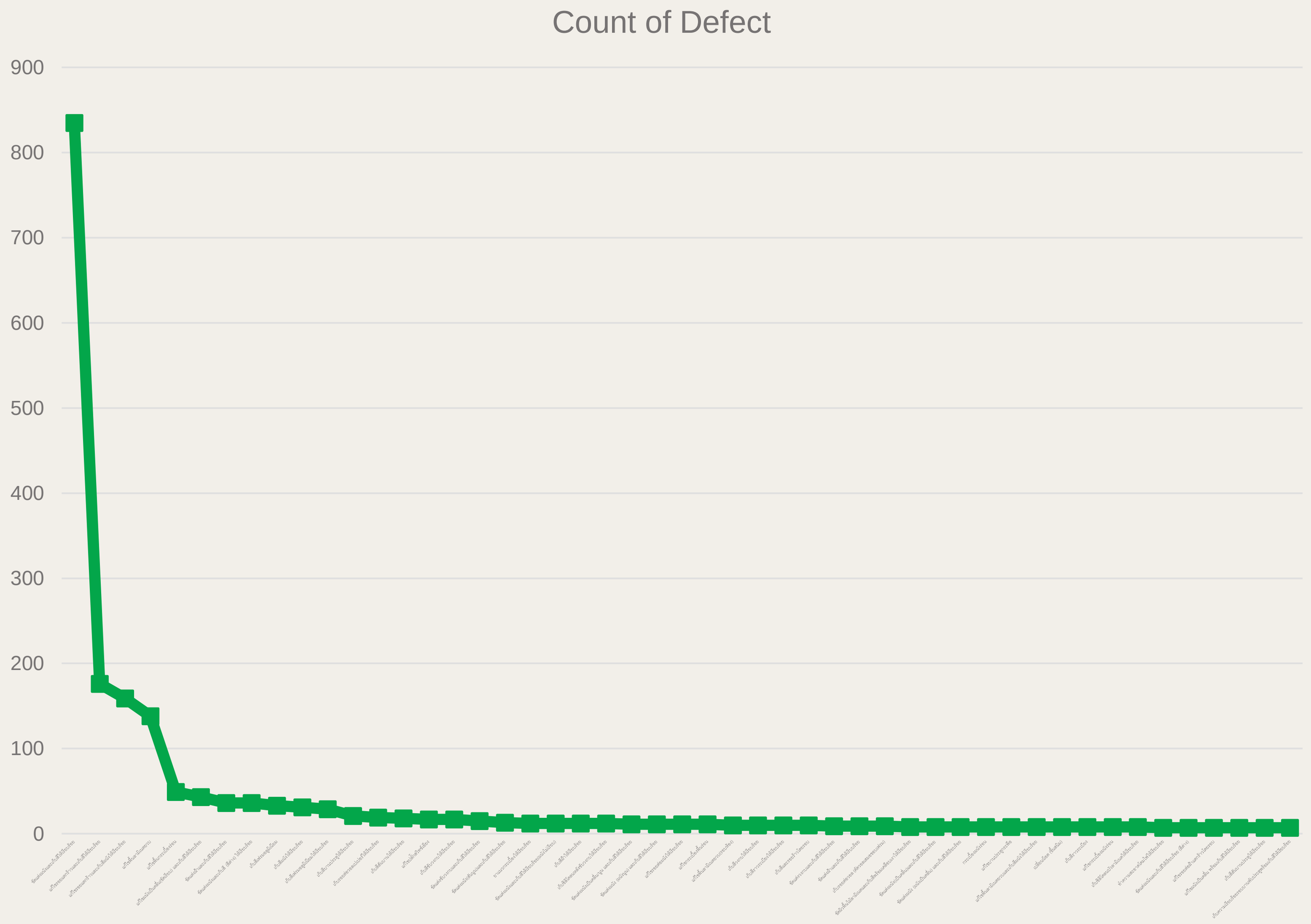


VGG16 takes the least time to train a model, while MobileNet get the highest BLEU score

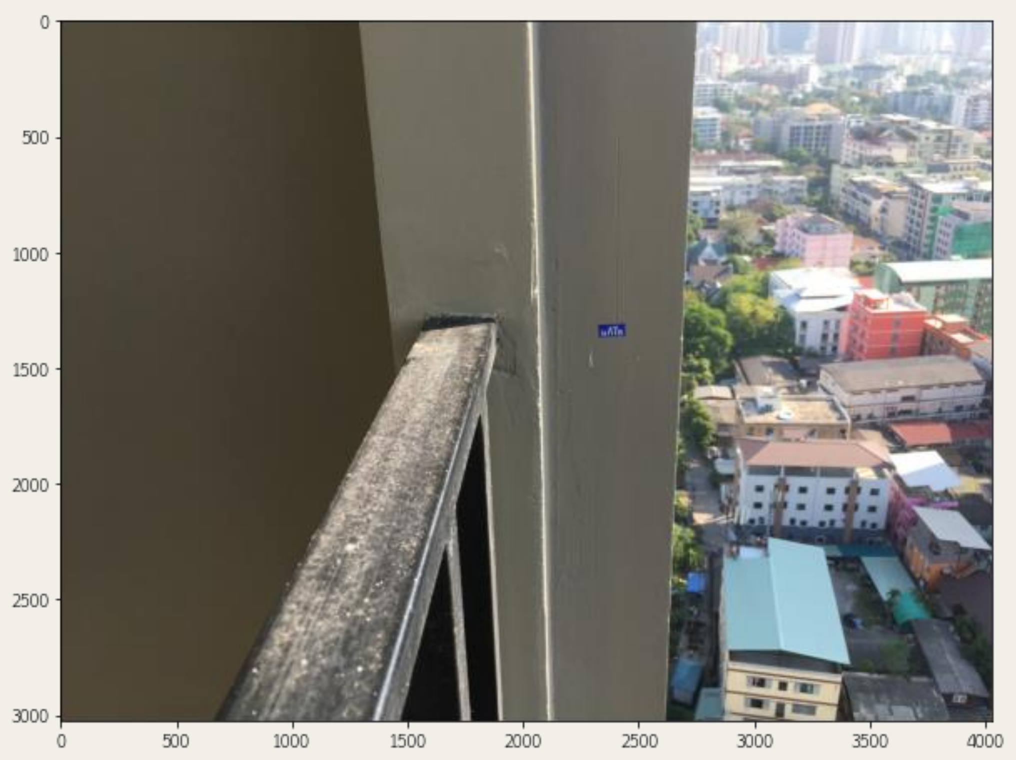
However, loss value is the same for all models

Challenges and Problems

Our dataset
varies and unique
(long-tailed class)



Some images have
many captions,
but our model can
predict only 1 caption



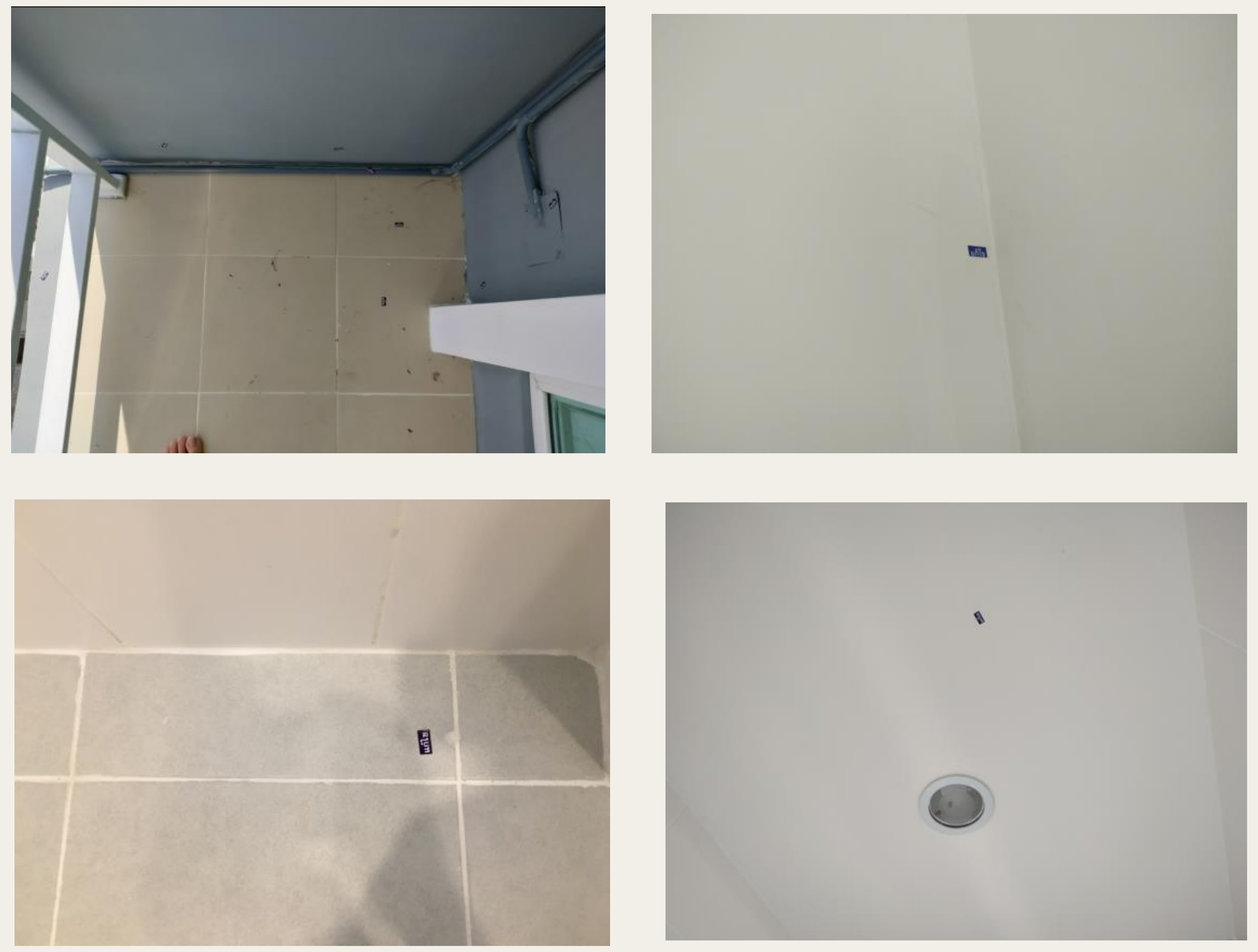
Real Caption

เก็บ สี โครงเหล็ก ให้ เรียบร้อย
ขัดแต่ง ผนัง และ เก็บ สี ให้ เรียบร้อย

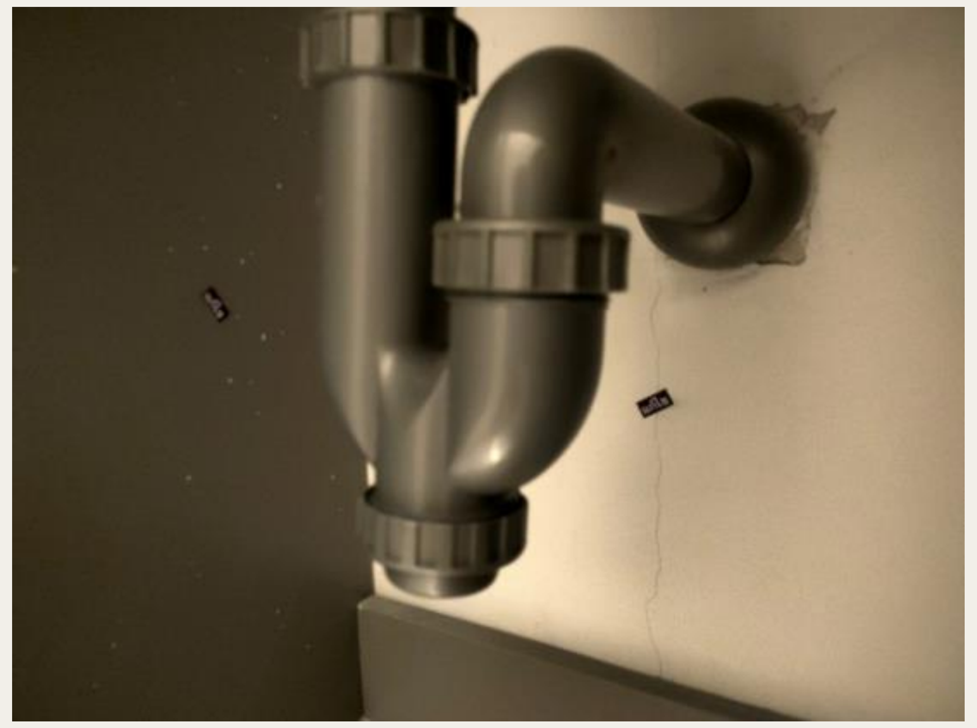
Predicted Caption

ขัดแต่ง ผนัง และ เก็บ สี ให้ เรียบร้อย

Difficult to classify with
similar class



Wrong prediction



Real Caption

แก้ไข รอย แตกร้าว และ เก็บ สี
ให้ เรียบร้อย

Predicted Caption

ทำ ความ สะอาด แอร์ ให้ เรียบร้อย

Future Works

- Gathering more training data for some classes
- Use techniques that can be dealt with imbalanced dataset

Appendix

Dataset

https://github.com/manadda-j/deep-learning/tree/main/04ImageCaptioning/Dataset

Source Code

https://github.com/manadda-j/deep-learning/tree/main/04ImageCaptioning/Source%20Code

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

INTRODUCTION

Background

- It takes time and costs to do the inspection report
- House inspector can benefit from image captioning that can help to improve the process of preparing inspection report
- Having an inspection is done for sellers before they are putting their houses on the market may increase sales
- Automatic image captioning in Thai for house defects can use to train junior inspector or staff with less technical skill

Scope

- Develop a model that helps to generate image captioning in Thai for house defects using a deep learning-based approach
- The model can generate image captioning for 16 classes of house defects

Limitations

- Insufficient dataset
- 1 image 1 caption

Process

Data Preparation
- Acquire the dataset
- Data labeling/grouping
- Data Augmentation
- Text Tokenization
- Text Preprocessing

Image Captioning
- Encoder using VGG16, MobileNet, InceptionV3
- Decoder using GRU Bahdanau Attention

Model Evaluation
- Training Loss
- Training Time
- BLEU (BiLingual Evaluation Understudy)

RESEARCH

Data Preparation

Image Pre-processing

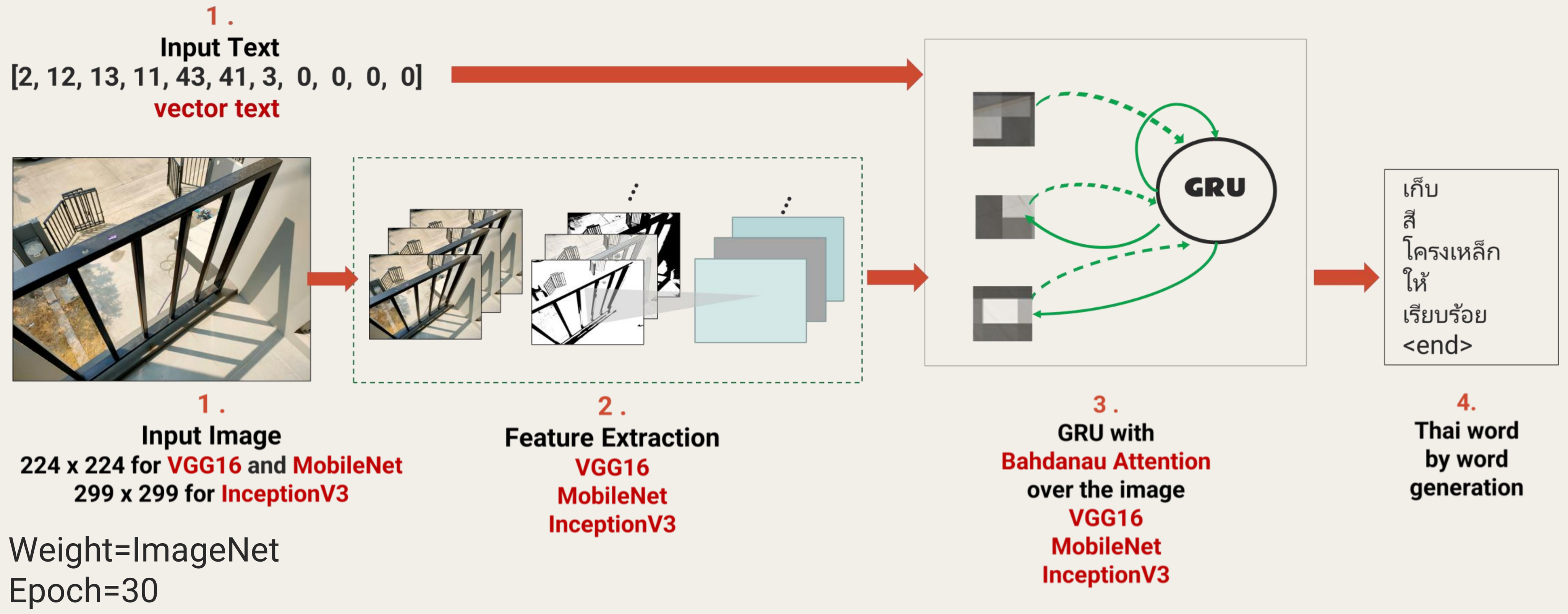
1. Manual labeling and grouping images for each class
2. Data Augmentation using OpenCV with 5 techniques which are rotation, shift, flip, brightness, and zoom
3. Resized images that suit each model 224 x 244 for both VGG16 and MobileNet and 299x299 for inceptionV13

Text Pre-processing

1. Tokenized the captions using PyThaiNLP with deep cut engine
Before: เก็บสีโครงเหล็กให้เรียบร้อย
After: เก็บ สี โครงเหล็ก ให้ เรียบร้อย
2. Added start and end tags for every caption to help model understands the start and end of each caption
<start> เก็บ สี โครงเหล็ก ให้ เรียบร้อย <end>
3. Covert text to vector and padding all the sequence to the same length as longest sentence using tf.keras.layers.TextVectorization
[2, 12, 13, 11, 43, 41, 3, 0, 0, 0, 0]

Image Captioning

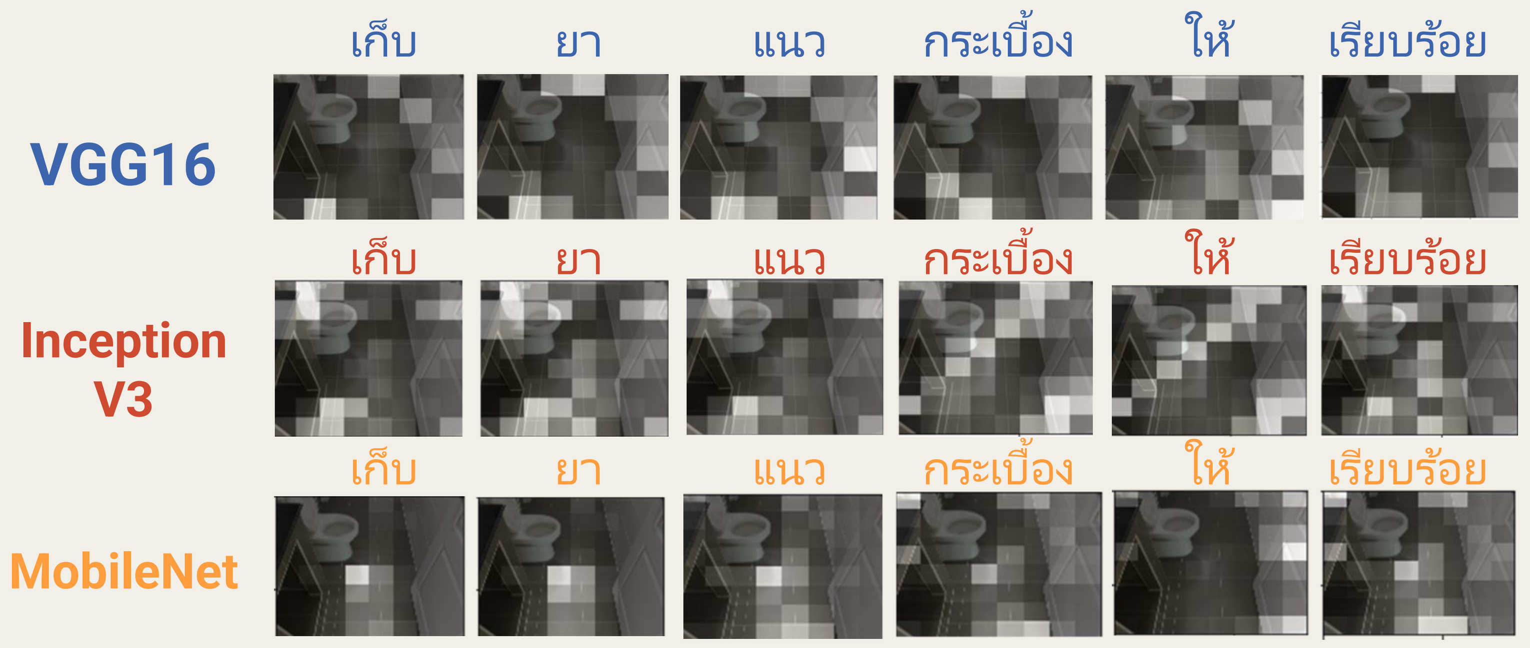
Model Architecture



Weight=ImageNet
Epoch=30
Optimizer=Adam
Loss Function=Sparse Categorical Cross Entropy

Results

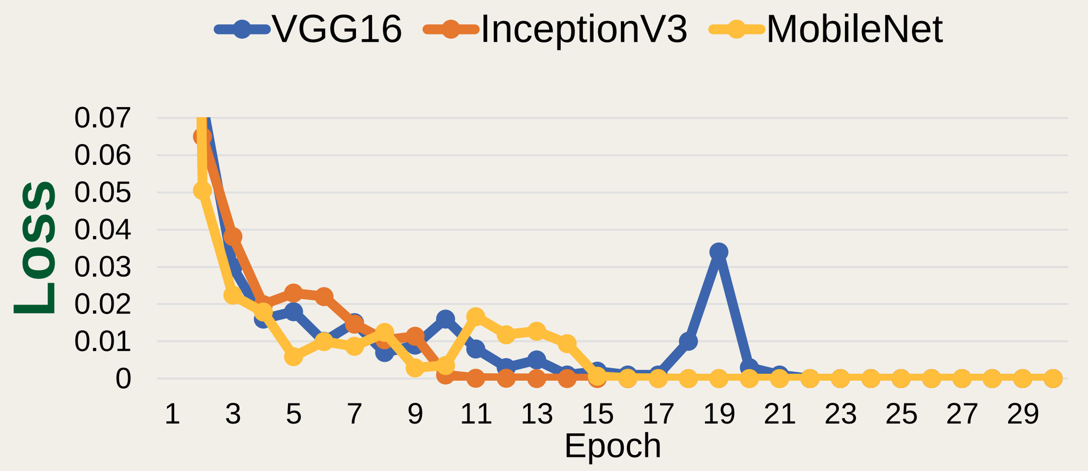
Real Caption:
เก็บ ยา แนว กระเบื้อง ให้ เรียบร้อย



CONCLUSION

Model Evaluation

Training Performance



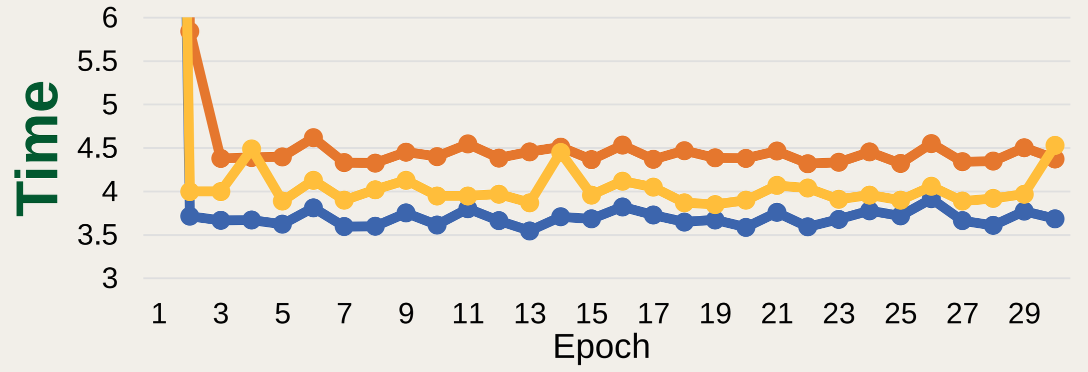
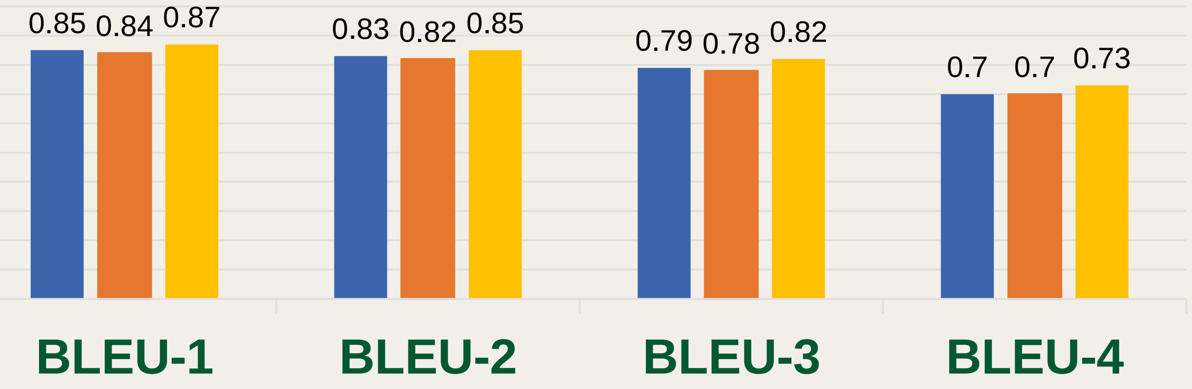


Image Captioning Performance



VGG16 takes the least time to train a model, while MobileNet gets the highest BLEU score However, loss value is the same for all models

Challenges and Problems

- Our dataset varies and unique (long-tailed class)
- Some images have many classes, but our model can predict only 1 caption
- Difficult to classify with similar class
- Wrong prediction

Future Works

- Gathering more training data for some classes
- Use techniques that can be dealt with imbalanced dataset