

TABLE OF CONTENTS

COVER	i
CONSENT SHEET	ii
ORIGINALITY SHEET	iii
ABSTRACT	iv
FOREWORD.....	v
ACKNOWLEDGEMENT	vi
TABLE OF CONTENTS.....	vii
IMAGE LIST	ix
TABLE LIST.....	x
APPENDIX LIST	xi
CHAPTER 1 INTRODUCTION	1
1.1. Background	1
1.2. Problem Formulation.....	3
1.3. Purpose and Benefits	4
1.4. Problem Limitations	4
1.5. Research Methods	5
CHAPTER 2 LITERATURE REVIEW	6
2.1. Literature Review	6
2.1.1. Image Captioning	6
2.2. Theoretical Foundations	11
2.2.1. Vision Transformer	11
2.2.2. Distilled Generative Pre-Trained 2	12
2.2.3. Post training Dynamic Quantization	13
2.2.4. Top-k Accuracy	14
2.2.5. Recall-Oriented Understudy for Gisting Evaluation	15
2.2.6. Bilingual Evaluation Understudy	16
CHAPTER 3 SYSTEM DESIGN	17
3.1. System Design Overview.....	17
3.2. Data Acquisition.....	17
3.3. Data Splitting	18
3.4. Data Preprocessing	19

3.4.1.	Image preprocessing	19
3.1.1.	Text preprocessing	21
3.2.	Model Development	24
3.3.	Model Training	26
3.4.	Model Quantization	26
3.5.	Evaluation Metric	27
3.5.1.	Top-k Accuracy	27
3.5.2.	Recall-Oriented Understudy for Gisting Evaluation	27
3.5.3.	Bilingual Evaluation Understudy	27
3.5.4.	Computation Time.....	28
3.5.5.	Model size	28
CHAPTER 4 RESULTS OF EXPERIMENTS AND ANALYSES		29
4.1.	Trial Scenario	29
4.1.1.	Data Preprocessing	29
4.1.2.	Model Training.....	29
4.1.3.	Caption Quality Evaluation.....	29
4.1.4.	Model Performance Comparison	29
4.2.	Experiment Result	30
4.2.1.	Data Preprocessing Results	30
4.2.2.	Model Training Results.....	30
4.2.3.	Caption Quality Results	31
4.2.4.	Model Performance Comparison Results	33
4.3.	Analysis.....	34
4.3.1.	Analysis of Data Preprocessing	34
4.3.2.	Analysis of Model Training.....	34
4.3.3.	Analysis of Caption Quality	35
4.3.4.	Analysis of Model Performance Comparison	35
CHAPTER 5 CONCLUSIONS AND SUGGESTIONS		37
5.1.	Conclusions	37
5.2.	Suggestions	37
BIBLIOGRAPHY		39
APPENDIX		42

IMAGE LIST

Figure 2. 1. Vision transformer architecture	11
Figure 2. 2. Transformers model architecture.	12
Figure 3. 1. System design diagram.....	17
Figure 3. 2. Example of Flickr8k dataset	18
Figure 3. 3. An illustration of the image resizing step.	20
Figure 3. 4. An illustration of the ViT feature extraction process	20
Figure 3. 5. A diagram of model architecture	24
Figure 4. 1. Training and validation loss.....	30
Figure 4. 2. Distribution of generated caption quality	31
Figure 4. 3. A qualitative comparison of generated captions	32
Figure 4. 4. Model size comparison.	33
Figure 4. 5. Model inference time comparison.	34