



Birzeit University  
Faculty of Engineering & Technology  
Department of Electrical & Computer Engineering

## Project Title

Prepared By:  
Student 1  
Student 2  
Student 3

Supervised By:  
Dr. Ahmad Ahmad

A Graduation Project submitted to the Department of Electrical and  
Computer Engineering in partial fulfillment of the requirements for the  
degree of B.Sc. in Computer Engineering

Birzeit  
April, 2022

# Abstract

Write here

# المستخلص

أكتب هنا

# Table of Contents

English Abstract	I
Arabic Abstract	II
Table of Contents	III
List of Tables	V
List of Figures	VI
<b>1 Introduction and Motivation</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Problem Statement . . . . .	1
1.3 Methodology . . . . .	1
1.4 Contribution . . . . .	1
1.5 Report Outline . . . . .	1
<b>2 Background</b>	<b>2</b>
2.1 Transformer . . . . .	2
2.1.1 Model Architecture . . . . .	2
2.1.2 Scaled Dot-Product Attention . . . . .	2
2.1.3 Multi-Head Attention . . . . .	2
2.1.4 Self-Attention and Multi-Head Self-Attention . . . . .	2
2.1.5 Feed Forward Network . . . . .	2
2.2 Vision Transformer (ViT) . . . . .	2
2.3 Lightweight ViT . . . . .	2
<b>3 Literature Review—ViT Acceleration Techniques</b>	<b>3</b>
3.1 Pruning . . . . .	3
3.2 Quantization . . . . .	3

3.3	Low-Rank Approximation . . . . .	3
3.4	Knowledge Distillation . . . . .	3
3.5	Lightweight ViT . . . . .	3
3.6	Transformer Acceleration on Hardware . . . . .	3
<b>4</b>	<b>Proposed Work</b>	<b>4</b>
<b>5</b>	<b>Project Plan</b>	<b>5</b>
<b>6</b>	<b>Conclusion and Future Work</b>	<b>6</b>
	<b>Bibliography</b>	<b>7</b>

# List of Tables

# List of Figures

# Chapter 1

## Introduction and Motivation

### Contents

---

1.1	Motivation . . . . .	1
1.2	Problem Statement . . . . .	1
1.3	Methodology . . . . .	1
1.4	Contribution . . . . .	1
1.5	Report Outline . . . . .	1

---

### 1.1 Motivation

### 1.2 Problem Statement

### 1.3 Methodology

### 1.4 Contribution

### 1.5 Report Outline



# Chapter 2

## Background

### Contents

---

<b>2.1</b>	<b>Transformer . . . . .</b>	<b>2</b>
2.1.1	Model Architecture . . . . .	2
2.1.2	Scaled Dot-Product Attention . . . . .	2
2.1.3	Multi-Head Attention . . . . .	2
2.1.4	Self-Attention and Multi-Head Self-Attention . . . . .	2
2.1.5	Feed Forward Network . . . . .	2
<b>2.2</b>	<b>Vision Transformer (ViT) . . . . .</b>	<b>2</b>
<b>2.3</b>	<b>Lightweight ViT . . . . .</b>	<b>2</b>

---

## 2.1 Transformer

### 2.1.1 Model Architecture

### 2.1.2 Scaled Dot-Product Attention

### 2.1.3 Multi-Head Attention

### 2.1.4 Self-Attention and Multi-Head Self-Attention

### 2.1.5 Feed Forward Network

## 2.2 Vision Transformer (ViT)

## 2.3 Lightweight ViT

# Chapter 3

## Literature Review—ViT Acceleration Techniques

### Contents

---

3.1	Pruning . . . . .	3
3.2	Quantization . . . . .	3
3.3	Low-Rank Approximation . . . . .	3
3.4	Knowledge Distillation . . . . .	3
3.5	Lightweight ViT . . . . .	3
3.6	Transformer Acceleration on Hardware . . . . .	3

---

### 3.1 Pruning

### 3.2 Quantization

### 3.3 Low-Rank Approximation

### 3.4 Knowledge Distillation

### 3.5 Lightweight ViT

### 3.6 Transformer Acceleration on Hardware

## Chapter 4

### Proposed Work

# Chapter 5

## Project Plan

## Chapter 6

### Conclusion and Future Work

# Bibliography