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

Accelerated Cloud Native DLRM-based E-commerce Recommendation System

Prepared By:
Ibraheem Alyan
Mohammad Abu-Shelbaia
Nidal Zabadi

Supervised By:
Dr. Ahmed Shawahna

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
December, 2023

Abstract

Write here

المستخلص

أكتب هنا

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction and Motivation

Contents

1.1 Motivation

On average, an intelligent recommender system delivers a 22.66% lift in conversions rates [1] for web products.

1.2 Problem Statement

1.3 Methodology

1.4 Contribution

1.5 Report Outline

Chapter 2

Background

Contents

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

3.2 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

[1] Salesforce Marketing Cloud, "Predictive intelligence benchmark report." https://brandcdn.exacttarget.com/sites/exacttarget/files/deliverables/etmc-predictiveintelligencebenchmarkreport.pdf, 2014.