TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KỲ MÔN**

**XỬ LÝ DỮ LIỆU LỚN**

*Người hướng dẫn*: **ThS. Nguyễn Thành An**

*Người thực hiện*: **Nguyễn Quốc Anh – 52100871**

**Võ Phú Vinh – 52100947**

**Lục Minh Hiếu – 52100889**

**Nguyễn Vũ Tường – 52100944**

Lớp **: 21050301**

Khoá  **: 25**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KỲ MÔN**

**XỬ LÝ DỮ LIỆU LỚN**

*Người hướng dẫn*: **ThS. Nguyễn Thành An**

*Người thực hiện*: **Nguyễn Quốc Anh – 52100871**

**Võ Phú Vinh – 52100947**

**Lục Minh Hiếu – 52100889**

**Nguyễn Vũ Tường – 52100944**

Lớp **: 21050301**

Khoá  **: 25**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

**LỜI CẢM ƠN**

Trước hết, chúng em xin bày tỏ lòng biết ơn sâu sắc đến thầy Nguyễn Thành An, giảng viên của môn Xử lý dữ liệu lớn, vì đã hướng dẫn, hỗ trợ và truyền đạt những kiến thức quý báu, giúp chúng em có thể hoàn thành báo cáo này. Sự kiên nhẫn, nhiệt huyết và sự hiểu biết sâu rộng của thầy về lĩnh vực xử lý dữ liệu lớn đã trở thành nguồn cảm hứng lớn, không chỉ cho dự án này mà còn cho sự phát triển học thuật và chuyên môn của chúng em trong tương lai.

Chúng em cũng xin được gửi lời cảm ơn đến các bạn bè và gia đình đã luôn ủng hộ chúng em trong suốt quá trình thực hiện báo cáo này. Sự khích lệ và hỗ trợ của các bạn đã là nguồn động viên tinh thần quan trọng, giúp chúng em vượt qua những thách thức và hoàn thành báo cáo một cách tốt nhất.

Cuối cùng, chúng em muốn nhấn mạnh rằng mọi thành công và kết quả đạt được trong báo cáo này đều không thể thiếu đi sự dẫn dắt và giáo dục từ thầy Nguyễn Thành An. Chúng em mong rằng báo cáo này không chỉ thể hiện được sự tiến bộ trong học thuật mà còn phản ánh được lòng biết ơn và sự trân trọng mà chúng em dành cho thầy.

*TP. Hồ Chí Minh, ngày 31 tháng 03 năm 2024*

*(Ký tên và ghi rõ họ tên)*

*Nguyễn Vũ Tường*

*Võ Phú Vinh*

*Nguyễn Quốc Anh*

*Lục Minh Hiếu*

# 

# TÓM TẮT

Đề tài giữa kỳ môn Xử lý dữ liệu lớn yêu cầu sinh viên thực hiện một dự án nhóm (từ 3 đến 5 người) trong thời gian 3 tuần, tập trung vào việc xử lý và phân tích tập tin baskets.csv, chứa dữ liệu mua hàng của người dùng. Các yêu cầu cụ thể bao gồm:

**Câu 1 (5 điểm)**:

**Thành phần Lý thuyết:**

* Hiểu và triển khai Tự động hóa việc tính toán đạo hàm (Autograd) trong Học Sâu:
  + Phần giới thiệu về Autograd: Giải thích chi tiết về tự động hóa việc tính toán đạo hàm (AD) và tầm quan trọng của nó trong quá trình đào tạo các mô hình học sâu.
  + So sánh giữa phương pháp tiến và phương pháp ngược trong Autograd: Bàn luận về ưu và nhược điểm của cả hai phương pháp.
  + Vai trò của đồ thị tính toán trong Autograd và lan truyền ngược: Giải thích cách đồ thị tính toán hỗ trợ trong việc tính toán đạo hàm tự động và lan truyền ngược.

**Thành phần Triển khai:**

* Sử dụng PyTorch để triển khai một mô hình mạng nơ-ron đơn giản:
  + Cách sử dụng khả năng Autograd của PyTorch để triển khai mô hình.
  + Định nghĩa một hàm mất mát tùy chỉnh và khám phá cách Autograd tính toán đạo hàm cho hàm này.
  + Thử nghiệm với các thuật toán tối ưu hóa khác nhau (ví dụ: SGD, Adam) và quan sát ảnh hưởng của chúng đến quá trình hội tụ và hiệu suất đào tạo.

**Phân tích Thí nghiệm::**

* Phân tích hiệu quả và ổn định số của việc tính toán đạo hàm dựa trên Autograd.
* Nghiên cứu ảnh hưởng của kích thước batch, tốc độ học, và kiến trúc mạng đến việc tính toán đạo hàm và động lực đào tạo.

**Câu 2 (5 điểm)**:

* Triển khai mô hình CNN-LSTM cho bài toán Tạo chú thích cho ảnh:
  + Triển khai ít nhất hai phiên bản: một sử dụng cơ chế chú ý trong quá trình giải mã và một không sử dụng.
  + Huấn luyện mô hình trên môi trường Colab sử dụng GPU và chỉ CPU.

# MỤC LỤC

[**TÓM TẮT 1**](#_30j0zll)

[**MỤC LỤC 3**](#_3znysh7)

[**I. Yêu cầu 1: 4**](#_vpfq3oitx9ar)

[1. Giới thiệu 4](#_23xvgwudze3p)

[2. Một số phương pháp tính đạo hàm trong Deep Learning 6](#_a886p0js1gs0)

[2.1 Numerical Differentiation using Finite Differences 6](#_t7pt5j5i43b0)

[2.2 Symbolic Differentiation 7](#_gg93d4aqj1cv)

[3. Automatic differentiation (AD) 9](#_lvlcpgm7i9sf)

[3.1 Chain rule 11](#_2n5zkjlimk8h)

[3.2 Chi tiết Forward mode và Reverse mode trong AD 12](#_t0rv18p1z6md)

[3.2.1 Forward mode 12](#_crgzug8keo3d)

[3.2.2 Reverse mode 14](#_pjr6cxeqc6dw)

[3.3 Computational graph 15](#_d9za3zdobbpm)

[**II. Yêu cầu 2: 17**](#_ychu1w5o7bqw)

[1. Lời mở đầu 17](#_gjdgxs)

[1.1 Định nghĩa vấn đề Image Captioning 17](#_sstanziil6oa)

[1.2 Giới hạn và rào cản của Image Captioning 18](#_ct7wpnh84fdm)

[2. Cơ Sở Lý Thuyết 18](#_wvhpum1sqz66)

[2.1 CNN-LSTM (encoder-decoder based image captioning methods) 18](#_yvved2jwdjfy)

[2.2 Ý nghĩa của attention trong LSTM với bài toán Image Captioning 20](#_cgjglns3d36n)

[3. Mô Hình CNN-LSTM Sử Dụng Attention 22](#_1fob9te)

[3.1 Tiền Xử Lý Hình Ảnh: 22](#_pi4pag7h6wou)

[3.2 Dataset và Dataloader: 22](#_1m4lj2mo51ss)

[3.3 Encoder - CNN: 23](#_kpmfbor45c7j)

[3.4 Attention: 23](#_wdl3hwevvdro)

[3.5 Decoder - RNN: 23](#_gqqffphkmosc)

[3.6 Huấn Luyện Mô Hình: 23](#_235g8fhko21u)

[3.7 Mô Hình EncoderDecoder: 24](#_cu3mvq70hcyy)

[4. Mô Hình CNN-LSTM Không Sử Dụng Attention 24](#_sceup3x065rt)

[4.1 EncoderCNN 24](#_iev78jppaa87)

[4.2 DecoderRNN 24](#_a8q4h2r0ddwx)

[4.3 EncoderDecoder 25](#_40qo3a3y8z3b)

[4.4 Quá Trình Huấn Luyện và Sinh Mô Tả 25](#_xkw5hdrcpd7r)

# 