**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC NÔNG LÂM TP HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

A green circle with white text and tree branches and a book

Description automatically generated

**LUẬN VĂN TỐT NGHIỆP**

**NGHIÊN CỨU VISION TRANSFORMER ỨNG DỤNG NHẬN DIỆN SẢN PHẨM TRONG THANH TOÁN Ở CỬA HÀNG TIỆN LỢI**

**Ngành :** CNTT

**Niên khoá :** 2020 – 2024

**Lớp :** DH20DTB

**Sinh viên thực hiện:** Nguyễn Hà Phước Hậu

Nguyễn Ngọc Huy

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC NÔNG LÂM TP HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

A green circle with white text and tree branches and a book

Description automatically generated

**LUẬN VĂN TỐT NGHIỆP**

**NGHIÊN CỨU VISION TRANSFORMER ỨNG DỤNG NHẬN DIỆN SẢN PHẨM TRONG THANH TOÁN Ở CỬA HÀNG TIỆN LỢI**

**Giảng viên hướng dẫn: Sinh viên thực hiện:**

TS. Nguyễn Văn DũNguyễn Hà Phước Hậu– 20130254

Nguyễn Ngọc Huy – 20130282

TP.HỒ CHÍ MINH, tháng 09 năm 2024

Nghiên cứu Vision Transformer ứng dụng nhận diện sản phẩm trong

thanh toán ở cửa hàng tiện lợi

Năm

2024

**CÔNG TRÌNH HOÀN TẤT TẠI**

**TRƯỜNG ĐẠI HỌC NÔNG LÂM TP HCM**

Cán bộ hướng dẫn: TS. Nguyễn Văn Dũ

Cán bộ phản biện: TS. Nguyễn Thị Phương Trâm

Luận văn cử nhân được bảo vệ tại HỘI ĐỒNG CHẤM LUẬN VĂN CỬ NHÂN TRƯỜNG ĐẠI HỌC NÔNG LÂM TP HCM ngày 05 tháng 09 năm 2024

*Nhận xét của Cán bộ hướng dẫn:*

*Nhận xét của Cán bộ phản biện:*

|  |  |
| --- | --- |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO TRƯỜNG ĐẠI HỌC NÔNG LÂM TPHCM** | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc Lập – Tự Do – Hạnh Phúc** |

**NHIỆM VỤ LUẬN VĂN CỬ NHÂN**

Họ tên sinh viên: **NGUYỄN HÀ PHƯỚC HẬU** Phái: Nam

Ngày tháng năm sinh: 04/10/2002 Nơi sinh: Bà Rịa

Chuyên ngành: Công Nghệ Thông Tin

Ngành: Công Nghệ Thông Tin

Họ tên sinh viên: **NGUYỄN NGỌC HUY** Phái: Nam

Ngày tháng năm sinh: 02/02/2002 Nơi sinh: Bình Định

Chuyên ngành: Công Nghệ Thông Tin

Ngành: Công Nghệ Thông Tin

I. TÊN ĐỀ TÀI: **Nghiên cứu Vision Transformer ứng dụng nhận diện sản phẩm trong thanh toán ở cửa hàng tiện lợi**

II. NHIỆM VỤ VÀ NỘI DUNG

- Nhiệm vụ: Xây dựng website bán hàng tích hợp Vision Transformer (ViT) để phân loại sản phẩm và in hóa đơn cho khách hàng.

- Nội dung:

* Nghiên cứu về mô hình Vision Transformer.
* Xây dựng mô hình máy học phân loại hình ảnh sản phẩm.
* Xây dựng phần mềm thanh toán bằng nhận diện sản phẩm.

III. NGÀY GIAO NHIỆM VỤ: **20/03/2024**

IV. NGÀY HOÀN THÀNH NHIỆM VỤ: **30/08/2024**

V. HỌ VÀ TÊN CÁN BỘ HƯỚNG DẪN: **TS. Nguyễn Văn D**

Ngày / /

KHOA CNTT

(Ký và ghi rõ họ tên)

Ngày / /

CÁN BỘ HƯỚNG DẪN

(Ký và ghi rõ họ tên)

Ngày / /

CÁN BỘ PHẢN BIỆN

(Ký và ghi rõ họ tên)

LỜI CẢM ƠN

Chúng em xin chân thành cảm ơn các thầy cô khoa Công nghệ thông tin trường Đại học Nông Lâm TP.Hồ Chí Minh, với những kiến thức quý báu và sự tận tâm, nhiệt huyết mà thầy cô đã truyền đạt cho chúng em trong suốt những năm đại học.

Đặc biệt, chúng em xin chân thành cảm ơn người Thầy Nguyễn Văn Dũ đã tận tình hướng dẫn, chỉ bảo và giúp đỡ chúng em trong suốt quá trình thực hiện đề tài nghiên cứu này. Cảm ơn Thầy vì Thầy đã truyền tải cho chúng em không chỉ về kiến thức chuyên ngành mà còn cả về cách sống, cách ứng xử, cách suy nghĩ giúp chúng em trưởng thành hơn.

Bên cạnh đó, em xin cám ơn tất cả những bạn bè đã chia sẻ kiến thức và tận tình giúp đỡ chúng tôi hoàn thành đề tài này. Trong quá trình thực hiện đề tài nghiên cứu, mặc dù chúng em đã có những cố gắng nỗ lực thực hiện nhưng chúng em không thể tránh được những sai sót nhất định. Kính mong sự thông cảm và tận tình chỉ bảo của quý Thầy Cô. Xin chân thành cảm ơn mọi người.

***Sinh viên thực hiện***

Nguyễn Hà Phước Hậu

Nguyễn Ngọc Huy

**DANH SÁCH CHỮ VIẾT TẮT**

CNN **C**onvolutional **N**eural **N**etwork

Mạng Neural tích chập

ViT **V**ision **T**ransformer

Biến đổi Thị giác

DL **D**eep **L**earning

Học sâu

ML **M**achine **L**earning

Học máy

COCO **C**ommon **O**bjects in **C**ontext

Common Objects in Context

**DANH MỤC CÁC HÌNH**

[Hình 1 . Kiến trúc Vision Transformer 4](#_Toc17405)

[Hình 2 . Chia hình ảnh ra thành các patch 6](#_Toc27439)

[Hình 3 . Tại sao phải thêm Position embeddings vào embedding của từng patch 7](#_Toc27681)

[Hình 4 . Hình ảnh minh họa cho công thức Self-attention 9](#_Toc4835)

[Hình 5 . Công thức Self-attention 10](#_Toc24696)

[Hình 6 . So sánh khi Fine-Tuning 2 model trên cùng 1 dataset 12](#_Toc30225)

[Hình 7 . Hình ảnh minh họa chức năng nhận diện sản phẩm(Bước 1) 24](#_Toc30487)

[Hình 8 . Hình ảnh minh họa chức năng nhận diện sản phẩm(Bước 2) 25](#_Toc144)

[Hình 9 . Hình ảnh minh họa chức năng nhận diện sản phẩm(Bước 3) 25](#_Toc2211)

[Hình 10 . Hình ảnh minh họa chức năng thành toán(Bước 1) 26](#_Toc20295)

[Hình 11 . Hình ảnh minh họa chức năng thành toán(Bước 2) 27](#_Toc4451)

[Hình 12 . Hình ảnh minh họa chức năng thành toán(Bước 3) 27](#_Toc31205)

[Hình 13 . Hình ảnh minh họa chức năng thành toán(Bước 4) 28](#_Toc26777)

[Hình 14 . Hình ảnh minh họa chức năng thêm sản phẩm(Bước 1) 29](#_Toc9426)

[Hình 15 . Hình ảnh minh họa chức năng thêm sản phẩm(Bước 2) 29](#_Toc531)

[Hình 16 . Hình ảnh minh họa chức năng thêm sản phẩm(Bước 3) 30](#_Toc28517)

[Hình 17 . Hình ảnh minh họa chức năng thêm sản phẩm(Bước 4) 31](#_Toc28157)

[Hình 18 . Hình ảnh minh họa chức năng sửa sản phẩm(Bước 1) 31](#_Toc4695)

[Hình 19 . Hình ảnh minh họa chức năng sửa sản phẩm(Bước 2) 32](#_Toc2809)

[Hình 20 . Hình ảnh minh họa chức năng sửa sản phẩm(Bước 3) 33](#_Toc17015)

[Hình 21 . Hình ảnh minh họa chức năng sửa sản phẩm(Bước 4) 33](#_Toc7794)

[Hình 22 . Hình ảnh minh họa chức năng xóa sản phẩm(Bước 1) 34](#_Toc11168)

[Hình 23 . Hình ảnh minh họa chức năng xóa sản phẩm(Bước 2) 34](#_Toc18220)

[Hình 24 . Hình ảnh minh họa chức năng xóa sản phẩm(Bước 3) 35](#_Toc6120)

[Hình 25 . Hình ảnh minh họa chức năng xóa sản phẩm(Bước 4) 35](#_Toc32288)

[Hình 26 . Hình ảnh minh họa chức năng quản lý đơn hàng(Bước 1) 36](#_Toc5915)

[Hình 27 . Hình ảnh minh họa chức năng quản lý đơn hàng(Bước 2) 36](#_Toc6536)

[Hình 28 . Hình ảnh minh họa chức năng xem đơn hàng chi tiết 37](#_Toc14952)

[Hình 29 . Lược đồ tuần tự chức năng nhận diện sản phẩm 42](#_Toc1590)

[Hình 30 . Lược đồ tuần tự chức năng thêm sản phẩm 43](#_Toc14425)

[Hình 31 . Lược đồ tuần tự chức năng sửa sản phẩm 44](#_Toc11049)

[Hình 32 . Lược đồ tuần tự chức năng xóa sản phẩm 45](#_Toc3337)

[Hình 33 . Lược đồ tuần tự chức năng xem đơn hàng(Quản lý đơn hàng) 47](#_Toc11106)

[Hình 34 . Sơ đồ Hệ thống 48](#_Toc27281)

[Hình 35 . Hình ảnh giới thiệt dataset 49](#_Toc31220)

[Hình 36 . Sơ đồ Phân tích phân phối dữ liệu huấn luyện 50](#_Toc3523)

[Hình 37 . Sơ đồ quy trình sử lý dữ liệu 52](#_Toc29787)

[Hình 38 . Cấu trúc mô hình VIT 54](#_Toc19227)

**DANH MỤC CÁC BẢNG**

[Bảng 1 . So sánh VIT và CNN 14](#_Toc1702)

**TÓM TẮT**

Trong những năm gần đây, thương mại điện tử đã có sự tăng trưởng ấn tượng. Đặc biệt, thị trường bán lẻ trong các siêu thị như Cửa hàng tiện lợi cũng đang phát triển mạnh mẽ. Để đáp ứng nhu cầu của người bán hàng trong việc cải thiện trải nghiệm mua sắm của khách hàng, việc tự động hóa quá trình nhận diện sản phẩm và thanh toán trở nên cực kỳ quan trọng. Đề tài này tập trung vào nghiên cứu và ứng dụng Vision Transformer (ViT) để nhận diện hình ảnh sản phẩm trên website bán hàng của Cửa hàng tiện lợi. Công nghệ này không chỉ giúp tự động phân loại sản phẩm mà còn hỗ trợ quá trình in hóa đơn và thanh toán một cách nhanh chóng và hiệu quả. Chúng tôi đã tích hợp ViT vào website bán hàng của Cửa hàng tiện lợi, giúp người bán hàng cải thiện quy trình kinh doanh và nâng cao chất lượng dịch vụ.

**ABSTRACT**

In recent years, e-commerce has shown impressive growth, especially in the retail market of supermarkets like Bach Hoa Xanh. To meet the needs of sellers in improving the shopping experience for customers, automating the process of product recognition and payment becomes crucial. This thesis focuses on researching and applying Vision Transformer (VIT) for product image recognition on Bach Hoa Xanh's e-commerce website. This technology not only helps automatically classify products but also supports the process of billing and payment quickly and efficiently. We have integrated ViT into Bach Hoa Xanh's sales website, helping sellers improve their business processes and enhance service quality.

**MỤC LỤC**

[CHƯƠNG 1. MỞ ĐẦU 1](#_Toc174442638)

[**1.1 Lý do chọn đề tài** 1](#_Toc174442639)

[**2.1 Mục đích, đối tượng và phạm vi nghiên cứu** 1](#_Toc174442640)

[**2.1.1 Mục tiêu nghiên cứu** 1](#_Toc174442641)

[**2.1.2 Đối tượng nghiên cứu** 1](#_Toc174442642)

[**2.1.3 Phạm vi nghiên cứu** 2](#_Toc174442643)

[**3.1 Kết quả cần đạt** 2](#_Toc174442644)

[CHƯƠNG 2: NỘI DUNG NGHIÊN CỨU 3](#_Toc174442645)

[**2.1. Tổng quan về Vision Transformer** 3](#_Toc174442646)

[**2.1.1. Giới thiệu về Vision Transformer** 3](#_Toc174442647)

[**2.1.2. Kiến trúc của Vision Transformer** 3](#_Toc174442648)

[**2.1.2.1. Embedding** 5](#_Toc174442649)

[**2.1.2.2. Transformer Encoder** 8](#_Toc174442650)

[2.1.2.3. MLP Head 10](#_Toc174442651)

[**2.1.3. Sự khác biệt giữa Vision Transformer và các mô hình CNN truyền thống** 11](#_Toc174442652)

[**2.1.4. Lợi ích và hạn chế của Vision Transformer** 15](#_Toc174442653)

[**2.1.4.1 Lợi ích của Vision Transformer** 15](#_Toc174442654)

[**2.1.4.2 Hạn chế của Vision Transformer** 16](#_Toc174442655)

[**2.2. Ứng dụng của Vision Transformer trong nhận diện sản phẩm** 16](#_Toc174442656)

[**2.2.1. Giới thiệu về nhận diện sản phẩm** 16](#_Toc174442657)

[**2.2.2. Các phương pháp nhận diện sản phẩm truyền thống** 17](#_Toc174442658)

[**2.2.3. Ứng dụng Vision Transformer trong nhận diện sản phẩm** 18](#_Toc174442659)

[**2.2.4. Các nghiên cứu và ứng dụng liên quan** 18](#_Toc174442660)

[CHƯƠNG 3. NGHIÊN CỨU VÀ PHÁT TRIỂN 21](#_Toc174442661)

[**3.1. Mô tả yêu cầu và chức năng của hệ thống** 21](#_Toc174442662)

[**3.1.1. Mô tả các chức năng** 21](#_Toc174442663)

[**3.1.1.1. Nhận diện sản phẩm** 21](#_Toc174442664)

[**3.1.1.2. Thanh toán** 23](#_Toc174442665)

[**3.1.1.3. Quản lý sản phẩm** 25](#_Toc174442666)

[**3.1.1.3. Quản lý đơn hàng** 32](#_Toc174442667)

[**3.1.2. Công nghệ sử dụng** 34](#_Toc174442668)

[**3.1.2.1. Frontend** 34](#_Toc174442669)

[**3.1.2.1. Backend** 35](#_Toc174442670)

[**3.1.3. Lược đồ tuần tự (Sequence Diagram)** 36](#_Toc174442671)

[**3.1.3.1. Chức năng nhận diện sản phẩm** 36](#_Toc174442672)

[**3.1.3.2. Chức năng thêm sản phẩm(Quản lý sản phẩm)** 37](#_Toc174442673)

[**3.1.3.3. Chức năng sửa sản phẩm(Quản lý sản phẩm)** 38](#_Toc174442674)

[**3.1.3.4. Chức năng xóa sản phẩm(Quản lý sản phẩm)** 39](#_Toc174442675)

[**3.1.3.5. Chức năng xem đơn hàng(Quản lý đơn hàng)** 40](#_Toc174442676)

[**3.1.4. Sơ đồ hệ thống** 41](#_Toc174442677)

[**3.2. Xây dựng mô hình Vision Transformer** 41](#_Toc174442678)

[**3.2.1. Giới thiệu dataset** 41](#_Toc174442679)

[**3.2.2. Thu thập và tiền xử lý dữ liệu** 43](#_Toc174442680)

[**3.2.3. Phân tích phân phối dữ liệu huấn luyện** 46](#_Toc174442681)

[**3.2.4. Xây dựng và huấn luyện mô hình** 46](#_Toc174442682)

[**3.2.5. Kết quả thực nghiệm và đánh giá** 49](#_Toc174442683)

[CHƯƠNG 4. ỨNG DỤNG VÀ TRIỂN KHAI 50](#_Toc174442684)

[**4.1 Ứng dụng trong hệ thống thanh toán tại cửa hàng tiện lợi** 50](#_Toc174442685)

[**4.1.1. Tích hợp hệ thống nhận diện sản phẩm vào quy trình thanh toán** 50](#_Toc174442686)

[**4.1.2. Kiểm thử và triển khai hệ thống** 52](#_Toc174442687)

[**4.2. Đánh giá hiệu quả hệ thống** 54](#_Toc174442688)

[**4.2.1. Hiệu quả về thời gian** 54](#_Toc174442689)

[**4.2.2. Hiệu quả về chi phí** 54](#_Toc174442690)

[**4.2.3.** **Độ chính xác và độ tin cậy của hệ thống** 55](#_Toc174442691)

# CHƯƠNG 1. MỞ ĐẦU

## **1.1 Lý do chọn đề tài**

Cách mạng Công nghiệp 4.0 đã đánh dấu một bước tiến quan trọng

trong sự phát triển của xã hội, với sự tích hợp mạnh mẽ của công nghệ số vào các lĩnh vực khác nhau. Trong bối cảnh này, nhu cầu của xã hội đối với các hệ thống phần mềm thông minh ngày càng tăng lên, đặc biệt là trong việc xử lý và phân tích dữ liệu hình ảnh.

Vì vậy đề tài này được chọn để thâm nhập sâu vào việc áp dụng công nghệ nhận diện sản phẩm trực tiếp vào quá trình thanh toán. Mục tiêu là tạo ra một phương pháp thanh toán tiện lợi, nhanh chóng và an toàn hơn thông qua việc nhận diện sản phẩm trực tiếp thay vì các phương pháp truyền thống như quét mã QR. Điều này sẽ mang lại sự tiện ích cao cho người dùng và thúc đẩy sự phát triển của thương mại điện tử thông minh.

## **2.1 Mục đích, đối tượng và phạm vi nghiên cứu**

### **2.1.1 Mục tiêu nghiên cứu**

- Nghiên cứu về mô hình Vision Transformer.

- Xây dựng mô hình máy học phân loại hình ảnh sản phẩm.

- Xây dựng phần mềm thanh toán bằng nhận diện sản phẩm.

### **2.1.2 Đối tượng nghiên cứu**

- Chuẩn bị và tiền xử lý dữ liệu để xây dựng mô hình.

- Phân tích cấu trúc và hoạt động: Nghiên cứu chi tiết cách mà Vision Transformer xử lý thông tin hình ảnh, tập trung vào cơ chế self-attention và việc ánh xạ các phần nhỏ của hình ảnh.

- Hiệu suất của Vision Transformer: Đánh giá khả năng của mô hình trong việc nhận diện và phân loại hình ảnh trên các bộ dữ liệu thử nghiệm khác nhau.

- Tối ưu hóa và ổn định hóa: Tìm hiểu cách tối ưu hóa mô hình, điều chỉnh tham số để cải thiện độ chính xác và ổn định hóa quá trình huấn luyện.

- Xây dựng phần mềm thanh toán bằng nhận diện sản phẩm demo.

### **2.1.3 Phạm vi nghiên cứu**

- Ứng dụng này sẽ được áp dụng tại các cửa hàng tiện ích hoặc là siêu thị nhỏ.

- Sử dụng camera điện thoại kết nối với máy tính. camera được đặt sao cho có thể chụp từ trên xuống mặt phẳng sản phẩm một cách hiệu quả.

- Nhận diện được các loại như chai các chai, lon, bàn chải ... không hỗ trợ các thực phẩm tươi như rau, củ, quả, thịt…

- Các sản phẩm phải nằm trên bề mặt phẳng, nhãn hiệu phải hướng lên trên và không bị chồng và che bởi các vật khác, hình ảnh phải rõ nét không bị mờ.

- Không xem xét đến kích thước của sản phẩm.

## **3.1 Kết quả cần đạt**

- Phần mềm thanh toán bằng nhận diện sản phẩm từ thành quả nghiên cứu trên bao gồm các chức năng:

+ Nhận diện sản phẩm thông qua hình ảnh khi thanh toán.

+ Tính ra số tiền từ sản phẩm được nhận dạng và hiển thị hóa đơn cho khách hàng.

- Tài liệu báo cáo về Vision Transformer để xử lý dữ liệu đầu vào của hình ảnh, giải quyết các vấn đề về nhận diện.

# CHƯƠNG 2: NỘI DUNG NGHIÊN CỨU

## **2.1. Tổng quan về Vision Transformer**

### **2.1.1. Giới thiệu về Vision Transformer**

Vision Transformer (ViT) là một kiến trúc mạng nơ-ron đột phá, tái hiện cách chúng ta xử lý và hiểu hình ảnh. Mô hình ViT được giới thiệu vào năm 2021 trong bài báo nghiên cứu tại hội nghị ICLR 2021 với tiêu đề “An Image is Worth 16\*16 Words: Transformers for Image Recognition at Scale”. Lấy cảm hứng từ thành công của các mô hình Transformer trong xử lý ngôn ngữ tự nhiên, ViT giới thiệu một cách mới để phân tích hình ảnh bằng cách chia chúng thành các mảng nhỏ và tận dụng các cơ chế tự chú ý (self-attention). Điều này cho phép mô hình nắm bắt cả các mối quan hệ cục bộ và toàn cục trong hình ảnh, dẫn đến hiệu suất ấn tượng trong các nhiệm vụ thị giác máy tính khác nhau.

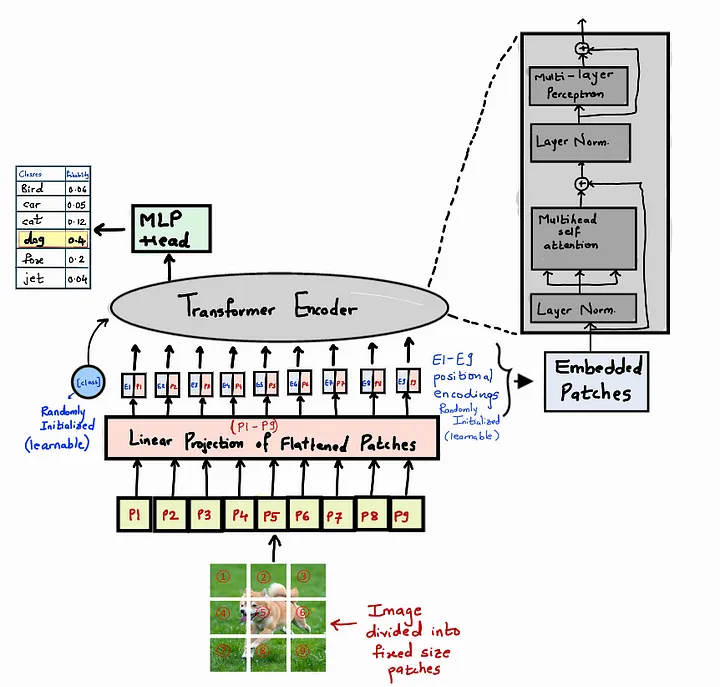
Vision Transformer đã trở thành một đối thủ cạnh tranh cho mạng neural tích chập (CNN), đang là công nghệ hàng đầu trong lĩnh vực thị giác máy tính và được sử dụng rộng rãi cho các nhiệm vụ nhận dạng hình ảnh khác nhau.

Mặc dù mạng neural tích chập đã thống trị lĩnh vực thị giác máy tính (Computer vision) trong nhiều năm, nhưng các mô hình Vision Transformer mới đã cho thấy khả năng đáng kinh ngạc, đạt được hiệu suất tương đương hoặc thậm chí vượt trội hơn so với CNN trên nhiều bài toán. Khi được huấn luyện trước trên một lượng lớn dữ liệu và chuyển sang các bộ chuẩn nhận diện hình ảnh có kích thước vừa hoặc nhỏ (như ImageNet, CIFAR-100, VTAB, v.v.), Vision Transformer (ViT) đạt được kết quả xuất sắc so với các mạng tích chập tiên tiến trong khi yêu cầu ít tài nguyên tính toán hơn đáng kể để huấn luyện.

### **2.1.2. Kiến trúc của Vision Transformer**

Vision Transformer có ứng dụng rộng rãi trong các bài toán nhận dạng hình ảnh phổ biến như object detection, image segmentation, image classification và action recognition. Ngoài ra, ViTs còn được áp dụng trong các mô hình generative và các bài toán đa mô hình như visual grounding, visual-question answering, and visual reasoning,…

Các mô hình ViT chia nhỏ hình ảnh thành các phân vùng (patches), sau đó chuyển các phân vùng này thành các chuỗi (sequences) và đưa chúng vào một mô hình transformer. Các phân vùng này được biến đổi thành các vectơ thông qua một lớp tuyến tính trước khi được đưa vào mô hình Transformer. Mô hình Transformer sau đó sẽ học cách ánh xạ các vectơ đầu vào này thành các vectơ đầu ra tương ứng với các đặc trưng quan trọng trong hình ảnh.



Hình 1. Kiến trúc Vision Transformer

Dựa vào kiến trúc trên ta có thể tóm gọn cách hoạt động của ViT như sau:

1. Embedding
2. Transformer Encoder
3. MLP Head

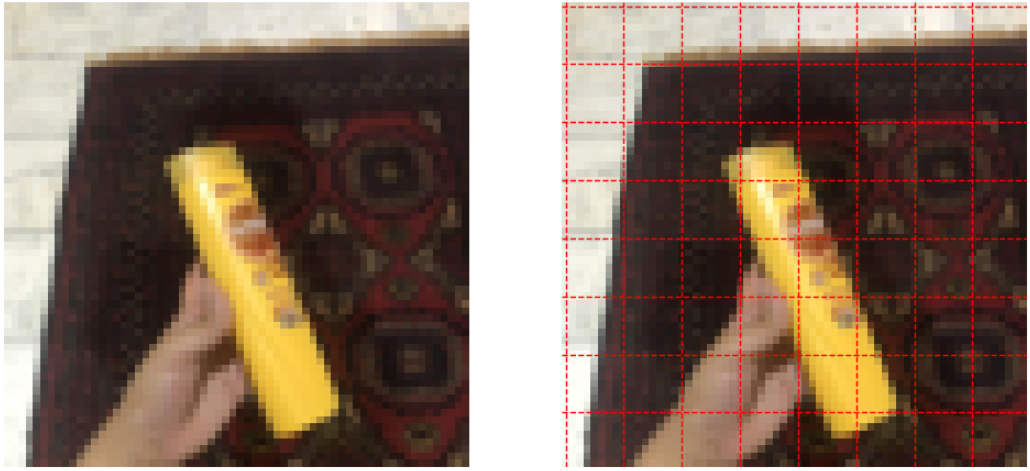
#### **2.1.2.1. Embedding**

Embedding đóng vai trò quan trọng trong kiến trúc Vision Transformer (ViT) và là bước đầu tiên trong quá trình xử lý hình ảnh. Đây là giai đoạn chuyển đổi hình ảnh từ dạng pixel thành các vector nhúng có thể xử lý được bởi mô hình transformer. Các bước chi tiết của quá trình embedding trong ViT bao gồm:

1. **Chia nhỏ hình ảnh**Việc chia nhỏ hình ảnh thành các mảnh nhỏ hơn (patches) là một bước quan trọng trong Vision Transformer (ViT), giúp mô hình xử lý hình ảnh theo cáchhiệu quả và linh hoạt hơn.  
   Đầu tiên, hình ảnh đầu vào được chia thành một lưới các mảnh nhỏ (patches), mỗi mảnh có kích thước bằng nhau. Điều này tương tự như việc chia nhỏ một bức tranh thành các mảnh ghép nhỏ để dễ dàng quản lý và xử lý. Giả sử chúng ta có một hình ảnh kích thước **H×W** và chúng ta muốn chia hình ảnh này thành các mảnh nhỏ kích thước **P×P**.

Số patches = **H/P \* W/P**

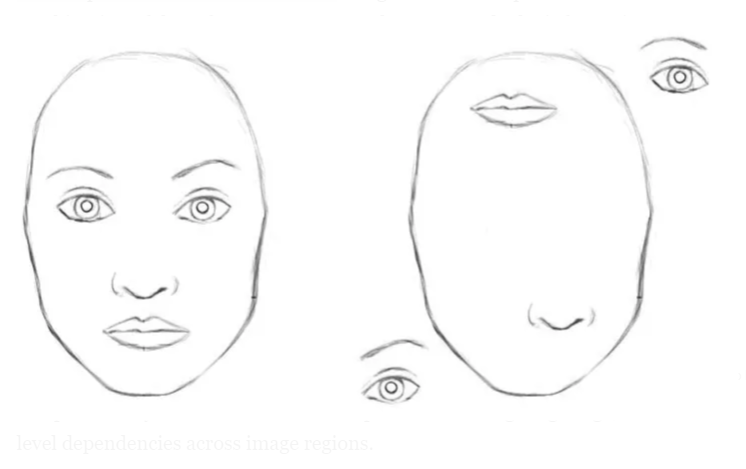
Ví dụ, một hình ảnh 224x224 có thể được chia thành 196 mảnh nhỏ, mỗi mảnh kích thước 16x16. Quá trình này giúp đơn giản hóa việc xử lý hình ảnh bằng cách tạo ra các đơn vị nhỏ hơn và dễ xử lý hơn.



Hình 2. Chia hình ảnh ra thành các patch

1. **Nhúngcác mảnh hình ảnh:**   
   Mỗi mảnh hình ảnh sau đó được biến đổi thành một vector có độ chiều cao cố định. Quá trình này tương tự như việc chuyển đổi các từ trong xử lý ngôn ngữ tự nhiên thành các vector nhúng. Các mảnh hình ảnh được flatten (trải phẳng) và sau đó được chiếu vào không gian vector thông qua một lớp Dense. Điều này giúp chuyển đổi thông tin từ dạng pixel thành dạng vector có thể xử lý bởi mô hình transforme.
2. **Thêm position embeddings:**

Trong ViT nó đóng vai trò quan trọng vì chúng cung cấp thông tin về vị trí của các patch trong hình ảnh, điều mà mô hình transformer vốn không tự động nắm bắt. Mô hình transformer, vốn được phát triển cho các bài toán xử lý ngôn ngữ tự nhiên, không có khả năng nắm bắt thông tin về thứ tự hoặc vị trí của các token đầu vào vì nó hoạt động dựa trên cơ chế attention. Trong ngữ cảnh văn bản, thứ tự của các từ là quan trọng, nhưng các mô hình transformer không tự động nhận diện điều này mà không có thêm thông tin bên ngoài. Khi áp dụng transformers vào hình ảnh, hình ảnh đầu vào thường được chia thành các patch nhỏ (ví dụ, 16x16 pixel). Mỗi patch được chuyển đổi thành một vector nhúng, nhưng các mô hình transformer không tự hiểu được vị trí của từng patch trong toàn bộ hình ảnh. Thông tin về vị trí của các patch là quan trọng vì mối quan hệ không chỉ giữa các patch mà còn giữa các vị trí cụ thể trong hình ảnh cũng ảnh hưởng đến việc hiểu và phân tích hình ảnh.



Hình 3. Tại sao phải thêm Position embeddings vào embedding của từng patch

Để khắc phục điều này, ViT thêm các position embeddings vào các embeddings của các patch. Position embeddings là các vector đặc biệt được gán cho từng vị trí của patch trong hình ảnh và được cộng vào các embeddings của patch. Điều này cho phép mô hình học được thông tin về vị trí tương đối và tuyệt đối của các patch trong hình ảnh, từ đó nắm bắt được mối quan hệ không gian giữa các phần của hình ảnh. Việc này rất quan trọng để đảm bảo rằng mô hình hiểu được ngữ cảnh không gian của các mảnh hình ảnh, cho phép nó nắm bắt được cả thông tin địa phương và toàn cục của hình ảnh.

1. **Patch Embeddings với Linear Projection**: Mỗi patch được xử lý bởi một lớp chiếu tuyến tính (linear projection) để chuyển đổi thành một embedding vector. Sau khi thêm các position embeddings, các vector này được xếp thành một chuỗi, tạo nên đầu vào cho transformer. Điều này giúp chuẩn hóa các vector và đảm bảo rằng chúng có thể được xử lý hiệu quả bởi các transformer encoder.

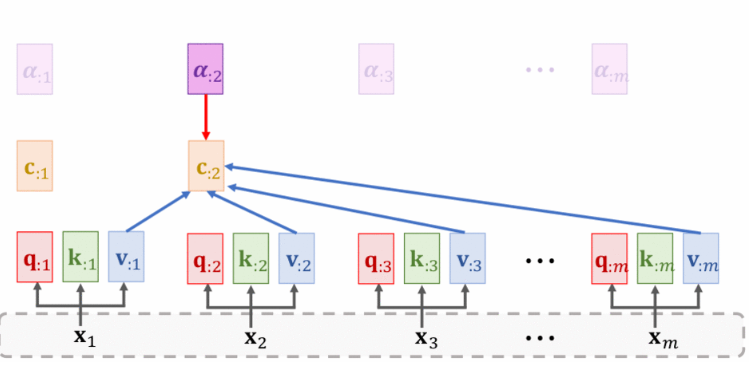
Vai trò của embedding trong Vision Transformer là chuyển đổi hình ảnh từ dạng pixel thô thành các vector nhúng, giữ lại thông tin không gian cần thiết và chuẩn bị dữ liệu cho các bước xử lý tiếp theo. Quá trình này là nền tảng cho việc sử dụng mô hình transformer trong thị giác máy tính, cho phép mô hình nắm bắt và phân tích các thông tin từ hình ảnh một cách hiệu quả và chính xác. Embedding không chỉ đơn thuần là bước chuẩn bị dữ liệu mà còn đóng vai trò quan trọng trong việc đảm bảo rằng mô hình có thể hiểu và xử lý đúng các thông tin từ hình ảnh.

#### **2.1.2.2. Transformer Encoder**

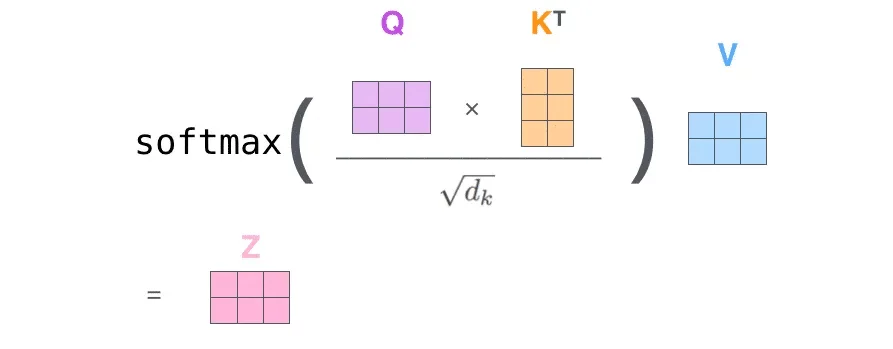
Phần cốt lõi của Vision Transformer bao gồm nhiều lớp transformer encoder, mỗi lớp chứa hai thành phần chính: multi-head self-attention và mạng nơ-ron feedforward.

Cơ chế self-attention nắm bắt các mối quan hệ giữa các mảnh khác nhau trong chuỗi đầu vào. Đối với mỗi mảnh nhúng (patch embedding), self-attention tính toán tổng có trọng số của tất cả các mảnh nhúng, trong đó trọng số được xác định bởi mức độ liên quan của mỗi mảnh đối với mảnh hiện tại. Cơ chế này cho phép mô hình tập trung vào các mảnh quan trọng trong khi xem xét cả ngữ cảnh địa phương và toàn cục. Multi-head attention sử dụng nhiều bộ tham số học được (attention heads) để nắm bắt các loại mối quan hệ khác nhau.

Công thức self-attention:



Hình 4. Hình ảnh minh họa cho công thức Self-attention

Trong mỗi patch ta có được một khóa K, giá trị V và truy vấn Q. Truy vấn Q sẽ tìm kiếm trong các giá trị khóa K của tất cả các patch khác để tìm ra ứng viên cung cấp ngữ cảnh cho nó (thông qua tích vô hướng). Cuối cùng, ta có ma trận đầu ra của lớp attention:



Hình 5. Công thức Self-attention

Trong đó:

* Q (query): vector truy vấn.
* K (key): vector khóa.
* V (value): vector giá trị.
* dK: kích thước của vector khóa.

Một patch có thể có nhiều kết nối với các patch khác theo những cách rất đa dạng, do đó mỗi patch có thể có nhiều bộ Q-K-V liên kết với nó. Kiến trúc này có tên gọi là multi-headed attention. Mỗi ma trận self-attention được tính toán riêng rẽ trên toàn bộ để học ra những tầng ngữ nghĩa khác nhau trước khi được kết hợp với nhau bởi một ma trận trọng số.

Sau self-attention, đầu ra từ cơ chế self-attention của mỗi mảnh được chuyển qua một mạng nơ-ron feedforward. Mạng này thường bao gồm một lớp fully connected tiếp theo là một hàm kích hoạt như ReLU (Rectified Linear Unit) hoặc GELU (Gaussian Error Linear Unit) . Mục đích của mạng feedforward là giới thiệu tính phi tuyến tính và cho phép mô hình học các mối quan hệ phức tạp giữa các mảnh.

Cả đầu ra của cơ chế self-attention và mạng feedforward đều được theo sau bởi layer normalization và residual connections. Layer normalization giúp ổn định và tăng tốc độ huấn luyện bằng cách chuẩn hóa các đầu vào cho mỗi lớp con. Residual connections, còn được gọi là skip connections, thêm các nhúng đầu vào ban đầu vào đầu ra của mỗi lớp con. Điều này giúp dòng gradient trong quá trình huấn luyện và ngăn chặn vấn đề vanishing gradient.

#### 2.1.2.3. MLP Head

MLP Head (Multi-Layer Perceptron Head) là một cấu trúc mạng neural đa lớp, thường được sử dụng để chuyển đổi đầu ra của mô hình transformer thành dự đoán cuối cùng. MLP Head bao gồm một hoặc nhiều lớp fully connected (dense layers) với các chức năng kích hoạt (activation functions) nhằm học các đặc trưng phân loại từ embeddings.

Các Thành Phần Chính của MLP Head:

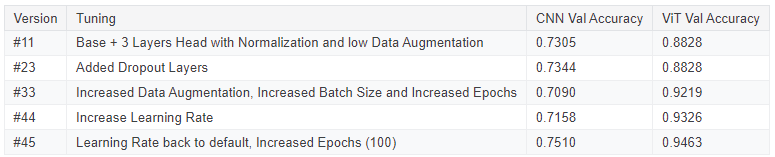
* Fully Connected Layers: MLP Head thường bao gồm một hoặc nhiều lớp fully connected. Mỗi lớp fully connected được nối với tất cả các neuron của lớp trước đó và học các trọng số để thực hiện phân loại.
* Activation Functions: Sau mỗi lớp fully connected, thường có một lớp kích hoạt ví dụ như relu hoặc gelu để thêm tính phi tuyến vào mô hình, giúp cải thiện khả năng học.
* Output Layer: Lớp đầu ra của MLP Head là một lớp fully connected với kích thước bằng số lượng lớp phân loại. Lớp này sẽ cung cấp các logits hoặc giá trị không chuẩn hóa cho các lớp phân loại.

### **2.1.3. Sự khác biệt giữa Vision Transformer và các mô hình CNN truyền thống**

Khi so sánh Vision Transformer (ViT) với Convolutional Neural Networks (CNN), chúng ta có thể thấy rõ sự khác biệt trong cách mà mỗi mô hình xử lý hình ảnh và học từ dữ liệu.

**Convolutional Neural Networks (CNN)** đã được sử dụng rộng rãi trong nhiều bài toán phân loại hình ảnh trong nhiều năm. CNN hoạt động bằng cách sử dụng các lớp tích chập để trích xuất đặc trưng từ hình ảnh. Các lớp này áp dụng các bộ lọc nhỏ qua toàn bộ hình ảnh để học các đặc trưng địa phương như cạnh và kết cấu. Bằng cách kết hợp các lớp pooling, CN có thể giảm kích thước không gian của hình ảnh, giữ lại các đặc trưng quan trọng và xây dựng các hierarchies của đặc trưng từ thấp đến cao. Cuối cùng, các đặc trưng này được chuyển đến các lớp fully connected để thực hiện phân loại. CNN thường yêu cầu ít tài nguyên tính toán hơn và có thể đạt được hiệu suất cao khi tinh chỉnh cho các bài toán cụ thể. Tuy nhiên, CNN có thể gặp khó khăn khi cần nắm bắt các mối quan hệ toàn cục trong hình ảnh, đặc biệt là khi xử lý hình ảnh có kích thước lớn hoặc yêu cầu hiểu biết toàn cục sâu hơn.

Ngược lại, **Vision Transformer (ViT)** là một cách tiếp cận mới hơn, sử dụng các lớp transformer để xử lý hình ảnh. Thay vì áp dụng bộ lọc cục bộ, ViT chia hình ảnh thành các patch nhỏ và ánh xạ chúng thành các vector nhúng. Các vector này sau đó được xử lý qua các lớp transformer, trong đó cơ chế attention giúp học các mối quan hệ toàn cục giữa các patch. Điều này cho phép ViT nắm bắt các đặc trưng toàn cục của hình ảnh tốt hơn. Mặc dù ViT có khả năng mở rộng tốt hơn và có thể làm việc hiệu quả với dữ liệu lớn, nó yêu cầu nhiều tài nguyên tính toán hơn, đặc biệt đối với các mô hình lớn và phức tạp. ViT có thể vượt trội trong các bài toán phân loại hình ảnh khi có lượng dữ liệu lớn và tài nguyên tính toán đủ.



Hình 6. So sánh khi Fine-Tuning 2 model trên cùng 1 dataset

Ta so sánh khi Fine-Tuning 2 model trên cùng 1 dataset:

Dựa trên bảng so sánh, chúng ta có thể thấy rõ sự vượt trội của Vision Transformer (ViT) so với Convolutional Neural Networks (CNN) trong nhiều điều kiện khác nhau.

Khi bắt đầu với phiên bản cơ bản, ViT đã thể hiện ưu thế rõ rệt với độ chính xác đạt 0.8828, cao hơn đáng kể so với CNN với độ chính xác 0.7305. Điều này cho thấy ViT có khả năng học và phân tích hình ảnh tốt hơn ngay từ đầu, ngay cả khi không áp dụng nhiều kỹ thuật tinh chỉnh.

Khi chúng tôi thêm các lớp dropout vào CNN ở phiên bản #23, mặc dù độ chính xác của CNN đã tăng lên một chút, đạt 0.7344, ViT vẫn duy trì độ chính xác ổn định ở mức 0.8828. Điều này chứng tỏ rằng, mặc dù cải tiến nhỏ có thể nâng cao hiệu suất của CNN, nhưng ViT vẫn giữ được lợi thế vượt trội về tổng thể.

Tiến đến phiên bản #33, khi áp dụng các kỹ thuật tối ưu hóa như tăng cường dữ liệu, tăng kích thước batch và số lượng epochs, độ chính xác của CNN lại giảm xuống còn 0.7090, trong khi ViT đạt độ chính xác 0.9219. Điều này cho thấy các thay đổi lớn hơn trong quá trình huấn luyện có thể làm giảm hiệu suất của CNN, trong khi ViT vẫn duy trì hiệu suất cao.

Ở phiên bản #44, với việc tăng learning rate, CNN đạt được một mức độ chính xác cải thiện lên 0.7158. Tuy nhiên, ViT tiếp tục cho thấy sự vượt trội với độ chính xác 0.9326, cho thấy rằng việc điều chỉnh learning rate có tác động tích cực đến CNN nhưng vẫn không thể sánh với sự hiệu quả của ViT.

Cuối cùng, khi chúng tôi điều chỉnh learning rate trở lại giá trị mặc định và tăng số lượng epochs lên 100 ở phiên bản #45, độ chính xác của CNN tăng lên 0.7510. Mặc dù có sự cải thiện này, ViT vẫn đạt được hiệu suất ấn tượng nhất với độ chính xác lên tới 0.9463. Điều này chứng tỏ rằng, ngay cả khi tối ưu hóa các tham số và tăng cường huấn luyện, ViT vẫn giữ được ưu thế đáng kể về hiệu suất.

Nhìn chung, bảng so sánh cho thấy rõ rằng ViT không chỉ thể hiện hiệu suất vượt trội hơn so với CNN mà còn duy trì sự ổn định và hiệu quả cao trong nhiều

điều kiện khác nhau. Điều này nhấn mạnh rằng ViT có thể là lựa chọn ưu việt hơn cho các bài toán phân loại hình ảnh, đặc biệt khi các điều kiện và kỹ thuật tinh chỉnh được áp dụng.

**Dựa vào thông tin trên ta có được so sánh trực quan về ViT và CNN:**

|  |  |  |
| --- | --- | --- |
| **Tiêu chí** | **CNN** | **ViT** |
| **Khả năng học đặc trưng** | Tốt trong việc học các đặc trưng cục bộ nhờ các lớp tích chập. Khả năng học các hierarchies đặc trưng từ thấp đến cao. | Tốt trong việc học các mối quan hệ toàn cục nhờ cơ chế attention, giúp nắm bắt các đặc trưng toàn diện hơn. |
| **Yêu cầu dữ liệu** | Hiệu quả với tập dữ liệu vừa và nhỏ. Cần ít dữ liệu để đạt được hiệu suất tốt, nhờ vào khả năng trích xuất đặc trưng mạnh mẽ từ các lớp tích chập. | Cần nhiều dữ liệu hơn để đạt được hiệu suất tối ưu. Hiệu quả hơn với các tập dữ liệu lớn nhờ khả năng học toàn cục. |
| **Tài nguyên tính toán** | Thường yêu cầu ít tài nguyên tính toán hơn so với ViT, đặc biệt đối với các mô hình nhỏ và trung bình. | Yêu cầu nhiều tài nguyên tính toán hơn do số lượng lớn các tham số và tính toán attention. |
| **Khả năng mở rộng** | Có thể gặp khó khăn khi mở rộng với hình ảnh có kích thước lớn hoặc yêu cầu hiểu biết toàn cục phức tạp. | Tốt hơn trong việc mở rộng với các tập dữ liệu lớn và hình ảnh có kích thước lớn nhờ khả năng học toàn cục. |
| **Khả năng tổng quát** | Hiệu suất tốt với các bài toán phân loại hình ảnh truyền thống và có thể đạt kết quả cao với các mô hình đã tinh chỉnh tốt. | Cung cấp khả năng tổng quát tốt hơn khi có lượng dữ liệu lớn, giúp mô hình học các đặc trưng phức tạp hơn. |
| **Đặc điểm học tập** | Học đặc trưng cục bộ nhanh chóng và hiệu quả với số lượng lớp và bộ lọc hạn chế. | Cần nhiều epoch và tài nguyên tính toán hơn để học các mối quan hệ toàn cục phức tạp. |

Bảng 1. So sánh VIT và CNN

**Dưới đây là một số điểm cần lưu ý về Vision Transformers (ViTs):**

1. ViT đã chứng minh hiệu quả của mình đối với các nhiệm vụ thị giác máy tính; các mô hình vision transformer đã nhận được sự chú ý đáng kể và làm suy yếu sự thống trị của CNNs trong lĩnh vực thị giác máy tính.
2. Vì Transformers yêu cầu một lượng lớn dữ liệu để đạt độ chính xác cao, quá trình thu thập dữ liệu có thể kéo dài thời gian dự án. Trong trường hợp có ít dữ liệu, CNNs thường cho kết quả tốt hơn so với Transformers.
3. Thời gian huấn luyện của Transformer nhanh hơn so với CNNs. So sánh theo hiệu suất tính toán và độ chính xác, Transformers có thể được lựa chọn nếu thời gian huấn luyện mô hình bị giới hạn.
4. Cơ chế tự chú ý (self-attention) có thể mang lại nhận thức sâu hơn cho mô hình. Rất khó hiểu được các điểm yếu của mô hình CNNs, nhưng attention maps có thể được trực quan hóa và giúp tìm ra cách cải tiến mô hình. Quá trình này khó hơn đối với các mô hình CNNs.
5. Mặc dù có một số framework cho Transformers, triển khai các mô hình CNN vẫn ít phức tạp hơn.
6. Sự xuất hiện của Vision Transformers cũng cung cấp một nền tảng quan trọng cho việc phát triển các mô hình computer vision. Mô hình vision lớn nhất là ViT-MoE của Google, có 15 tỷ tham số, đã lập kỷ lục mới trong việc phân loại ImageNet-1K.

ViT đã trở thành một trong những tiến bộ quan trọng trong lĩnh vực xử lý hình ảnh và tiếp tục được nghiên cứu và phát triển để cải thiện hiệu suất và khả năng ứng dụng trong tương lai.

### **2.1.4. Lợi ích và hạn chế của Vision Transformer**

Khi nói về Vision Transformer (ViT), chúng ta có thể thấy rõ ràng những lợi ích và hạn chế của mô hình này, đặc biệt là khi so sánh với các phương pháp học sâu truyền thống như Convolutional Neural Networks (CNN).

#### **2.1.4.1 Lợi ích của Vision Transformer**

Một trong những điểm mạnh nổi bật của Vision Transformer là khả năng nắm bắt các mối quan hệ toàn cục trong hình ảnh. Khác với các CNN, vốn chủ yếu tập trung vào các đặc trưng cục bộ qua các lớp tích chập, ViT sử dụng cơ chế attention để học các mối liên hệ giữa các patch của hình ảnh. Điều này giúp mô hình hiểu rõ hơn về cấu trúc tổng thể của hình ảnh. Ví dụ, trong các bài toán phân loại hình ảnh phức tạp như nhận diện cảnh vật hoặc phân tích chi tiết các đối tượng trong bức tranh lớn, ViT có thể cung cấp sự hiểu biết sâu hơn và chính xác hơn về toàn bộ bức tranh.

Thêm vào đó, ViT cho thấy hiệu suất cao khi được đào tạo trên các tập dữ liệu lớn. Nghiên cứu của Google Brain cho thấy ViT đạt hiệu suất ấn tượng trên các bộ dữ liệu quy mô lớn như ImageNet khi có đủ tài nguyên và dữ liệu để huấn luyện. Điều này cho thấy rằng ViT có khả năng khai thác tối đa thông tin từ lượng dữ liệu phong phú.

ViT cũng có khả năng mở rộng tốt với kích thước mô hình và dữ liệu. Khi mô hình và dữ liệu lớn hơn, ViT thường tiếp tục cải thiện hiệu suất, điều này khác biệt so với các CNN truyền thống, nơi việc mở rộng kích thước mô hình có thể gặp nhiều thách thức. Chẳng hạn, khi áp dụng ViT cho các bài toán phân tích hình ảnh trong các ứng dụng yêu cầu độ phân giải cao hoặc số lượng lớp lớn, mô hình này cho thấy sự gia tăng hiệu suất rõ rệt.

#### **2.1.4.2 Hạn chế của Vision Transformer**

Tuy nhiên, Vision Transformer không phải là không có hạn chế. Một trong những thách thức lớn nhất là yêu cầu về tài nguyên tính toán. ViT yêu cầu nhiều tài nguyên tính toán hơn so với các CNN, đặc biệt khi xử lý các mô hình lớn. Điều này có thể tạo ra khó khăn trong việc triển khai trên các hệ thống có tài nguyên hạn chế, như các thiết bị di động hoặc máy tính cá nhân. Ví dụ, việc huấn luyện ViT trên các GPU cần cấu hình cao hoặc các cụm máy chủ mạnh mẽ có thể làm tăng chi phí và thời gian huấn luyện.

Bên cạnh đó, ViT cũng cần một lượng dữ liệu lớn để đạt hiệu suất tối ưu. Trong các tình huống với dữ liệu hạn chế, ViT có thể không đạt được kết quả tốt như mong muốn, điều này trái ngược với các CNN có thể hoạt động tốt hơn trên các tập dữ liệu nhỏ hơn nhờ khả năng học các đặc trưng cục bộ hiệu quả hơn.

Cuối cùng, mặc dù ViT có kiến trúc đơn giản hơn so với các CNN truyền thống, điều này không đồng nghĩa với việc nó luôn dễ dàng trong việc triển khai hoặc tối ưu hóa. ViT có thể cần sự điều chỉnh tinh vi và tối ưu hóa đặc biệt để đạt hiệu suất tốt nhất trong các ứng dụng thực tế.

Nhìn chung, ViT cung cấp một cách tiếp cận mới mẻ và mạnh mẽ cho các bài toán phân loại hình ảnh, nhưng nó cũng đi kèm với các yêu cầu về tài nguyên và dữ liệu cần được cân nhắc khi áp dụng trong thực tế.

## **2.2. Ứng dụng của Vision Transformer trong nhận diện sản phẩm**

### **2.2.1. Giới thiệu về nhận diện sản phẩm**

Nhận diện sản phẩm là một lĩnh vực quan trọng trong công nghệ thông tin và thương mại điện tử, nhằm mục đích xác định và phân loại các sản phẩm từ hình ảnh hoặc video.

Quá trình này không chỉ giúp tăng cường hiệu quả của các hệ thống quản lý kho hàng mà còn nâng cao trải nghiệm mua sắm của khách hàng bằng cách cung cấp thông tin chi tiết và chính xác về sản phẩm.

Đặc biệt trong bối cảnh sự phát triển mạnh mẽ của thương mại điện tử, nhận diện sản phẩm trở thành một phần thiết yếu trong việc tối ưu hóa quy trình bán hàng, từ việc tự động hóa kiểm kê hàng hóa đến cải thiện hệ thống gợi ý sản phẩm, từ đó giúp tăng cường sự hài lòng và giữ chân khách hàng.

### **2.2.2. Các phương pháp nhận diện sản phẩm truyền thống**

Các phương pháp nhận diện sản phẩm truyền thống thường dựa vào các kỹ thuật học máy và thị giác máy tính, bao gồm nhiều phương pháp khác nhau:

* **Phát hiện và mô tả đặc trưng:** Những kỹ thuật như SIFT (Scale-Invariant Feature Transform) và SURF (Speeded-Up Robust Features) đã được sử dụng để phát hiện các điểm đặc trưng trong hình ảnh và mô tả chúng một cách hiệu quả. Tuy nhiên, những phương pháp này thường yêu cầu quá trình tiền xử lý và điều chỉnh tham số tốn thời gian.
* **Phân loại dựa trên các đặc trưng:** Sau khi phát hiện, các đặc trưng được sử dụng để huấn luyện các mô hình học máy như SVM (Support Vector Machine) hoặc k-NN (k-Nearest Neighbors) nhằm phân loại sản phẩm. Mặc dù những phương pháp này đã đạt được một số thành công nhất định, nhưng chúng thường không đủ mạnh mẽ để xử lý các hình ảnh phức tạp trong môi trường thực tế.
* **Deep Learning với CNNs (Convolutional Neural Networks):** Mạng nơ-ron tích chập đã trở thành tiêu chuẩn vàng trong nhận diện hình ảnh. CNNs tự động học các đặc trưng từ dữ liệu hình ảnh mà không cần phải thiết kế thủ công, giúp tăng cường độ chính xác và khả năng tổng quát của mô hình. Tuy nhiên, CNNs vẫn gặp một số hạn chế trong việc nắm bắt các mối quan hệ không gian phức tạp giữa các đối tượng trong hình ảnh.

### **2.2.3. Ứng dụng Vision Transformer trong nhận diện sản phẩm**

Vision Transformer (ViT) là một kiến trúc mạng nơ-ron mới, được phát triển dựa trên khái niệm Transformer, vốn được áp dụng thành công trong lĩnh vực xử lý ngôn ngữ tự nhiên.

ViT xử lý hình ảnh bằng cách chia chúng thành các patch nhỏ và xem xét chúng như một chuỗi, điều này cho phép mô hình nắm bắt được các mối quan hệ không gian phức tạp hơn so với các phương pháp truyền thống.

ViT đã chứng minh khả năng vượt trội trong việc nhận diện hình ảnh, đặc biệt là trong các nhiệm vụ yêu cầu xử lý hình ảnh lớn và đa dạng.

**Ví dụ:** Trong một dự án nhận diện sản phẩm, quy trình áp dụng ViT có thể được thực hiện như sau:

1. **Chuẩn bị dữ liệu:** Hình ảnh sản phẩm được chia thành các patch nhỏ, với kích thước thường là 16x16 pixels. Điều này cho phép mô hình phân tích chi tiết từng phần của hình ảnh và phát hiện các đặc trưng quan trọng.
2. **Huấn luyện mô hình:** ViT sẽ được huấn luyện trên một tập dữ liệu lớn gồm nhiều hình ảnh sản phẩm khác nhau, nhằm học các đặc trưng quan trọng và nhận diện các mẫu trong dữ liệu. Quá trình này thường bao gồm việc tối ưu hóa các tham số của mô hình để đạt được độ chính xác cao nhất.
3. **Nhận diện:** Sau khi huấn luyện, mô hình có thể sử dụng để dự đoán nhãn sản phẩm cho hình ảnh đầu vào mới. Điều này không chỉ giúp tăng cường độ chính xác mà còn giảm thời gian xử lý trong các ứng dụng thực tế.

### **2.2.4. Các nghiên cứu và ứng dụng liên quan**

Nhiều nghiên cứu và ứng dụng đã chỉ ra khả năng và tiềm năng to lớn của Vision Transformer (ViT) trong lĩnh vực nhận diện sản phẩm. Dưới đây là một số nghiên cứu tiêu biểu và các ứng dụng liên quan:

1. **Nghiên cứu của Dosovitskiy et al. (2020):**

* Bài báo này trình bày Vision Transformer, một kiến trúc mới kết hợp giữa các khối Transformer và hình ảnh, cho thấy ViT có thể đạt được hiệu suất cao hơn hoặc tương đương với các mô hình CNN truyền thống trên nhiều tập dữ liệu khác nhau. Nghiên cứu này mở ra hướng đi mới trong việc sử dụng các mô hình không gian hình ảnh phức tạp.
* **Liên kết:** [An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale](https://arxiv.org/abs/2010.11929)

1. **ViT for Fine-Grained Image Classification:**

* Nghiên cứu này áp dụng Vision Transformer vào lĩnh vực phân loại hình ảnh tinh vi, trong đó các đối tượng cần nhận diện thường có những khác biệt rất nhỏ về đặc trưng. Kết quả cho thấy ViT có thể học và phân loại các đặc trưng chi tiết của sản phẩm, điều này rất quan trọng trong thương mại điện tử khi phân loại các sản phẩm tương tự.
* **Liên kết:** [Fine-Grained Image Classification Using Vision Transformers](https://arxiv.org/abs/2105.09443)

1. **Ứng dụng Vision Transformer trong thương mại điện tử:**

* Bài báo này nghiên cứu về việc các công ty lớn như Amazon và Alibaba đã tích hợp Vision Transformer vào hệ thống gợi ý sản phẩm và quy trình kiểm kê hàng hóa. Việc áp dụng ViT không chỉ cải thiện độ chính xác trong nhận diện sản phẩm mà còn giúp tối ưu hóa quy trình quản lý tồn kho.
* **Liên kết:** [Transforming E-Commerce with Vision Transformers](https://arxiv.org/abs/2106.08032)

1. **Vision Transformer for Object Detection:**

* Nghiên cứu này giới thiệu cách ViT có thể được áp dụng cho nhiệm vụ phát hiện đối tượng, rất quan trọng trong việc nhận diện sản phẩm trên các kệ hàng trong cửa hàng bán lẻ. Việc sử dụng ViT giúp cải thiện độ chính xác và khả năng phát hiện các đối tượng trong các bối cảnh phức tạp.
* **Liên kết:** [End-to-End Object Detection with Transformers](https://arxiv.org/abs/2005.12872)

1. **Ứng dụng trong bán lẻ:**

* Một bài báo nghiên cứu đã chỉ ra cách các kỹ thuật học sâu, bao gồm Vision Transformer, có thể được sử dụng để theo dõi và quản lý hàng hóa trong môi trường bán lẻ. Việc này giúp tối ưu hóa quy trình kiểm kê và nâng cao trải nghiệm của khách hàng.
* **Liên kết:** [Deep Learning for Retail Inventory Management](https://www.sciencedirect.com/science/article/pii/S2352914821002068)

1. **Sử dụng ViT trong lĩnh vực y tế:**

* Mặc dù nghiên cứu này không hoàn toàn tập trung vào nhận diện sản phẩm, nhưng nó cho thấy tiềm năng của Vision Transformer trong nhận diện hình ảnh y tế, từ đó mở ra khả năng áp dụng trong các lĩnh vực khác như kiểm tra chất lượng sản phẩm trong sản xuất.
* **Liên kết:** [Vision Transformers for Medical Image Analysis](https://arxiv.org/abs/2102.10873)

Những nghiên cứu và ứng dụng này không chỉ chứng minh được hiệu quả của Vision Transformer trong lĩnh vực nhận diện sản phẩm mà còn mở ra nhiều cơ hội và hướng đi mới cho việc phát triển công nghệ trong tương lai.

Việc tham khảo các tài liệu này sẽ giúp bạn có cái nhìn sâu sắc hơn về cách ViT đang thay đổi cách chúng ta nhận diện và phân loại sản phẩm trong nhiều lĩnh vực khác nhau.

# CHƯƠNG 3. NGHIÊN CỨU VÀ PHÁT TRIỂN

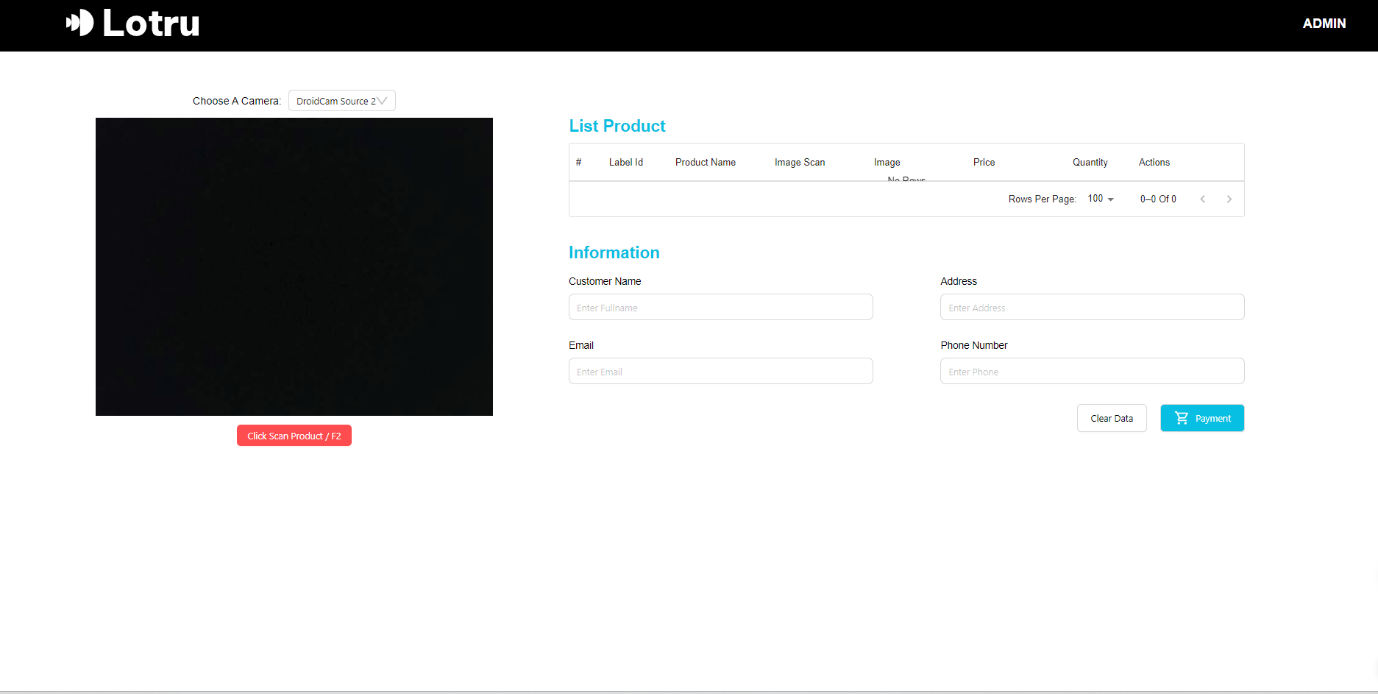
## **3.1. Mô tả yêu cầu và chức năng của hệ thống**

### **3.1.1. Mô tả các chức năng**

Hệ thống bao gồm các chức năng chính như nhận diện sản phẩm, thanh toán, quản lý sản phẩm, và xem các đơn hàng. Các chức năng này được chi tiết như sau:

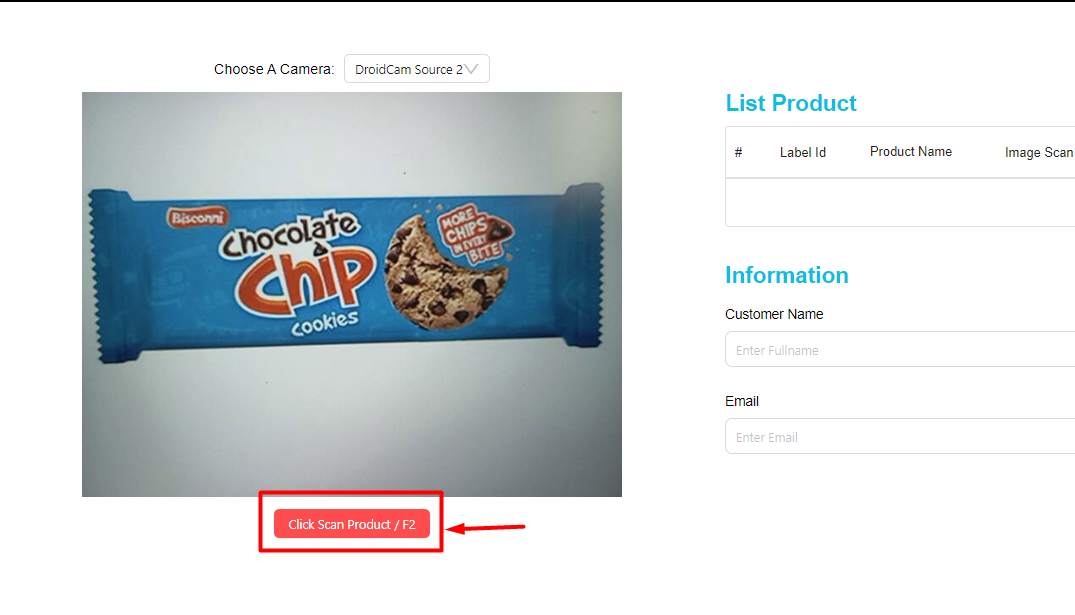
#### **3.1.1.1. Nhận diện sản phẩm**

* Mô tả: Chức năng này cho phép hệ thống tự động nhận diện sản phẩm thông qua hình ảnh. Người dùng có thể chụp ảnh sản phẩm hoặc tải lên hình ảnh từ thư viện để hệ thống nhận diện.
* Quy trình: (Chụp ảnh bỏ vào khi hoàn tất giao dien)

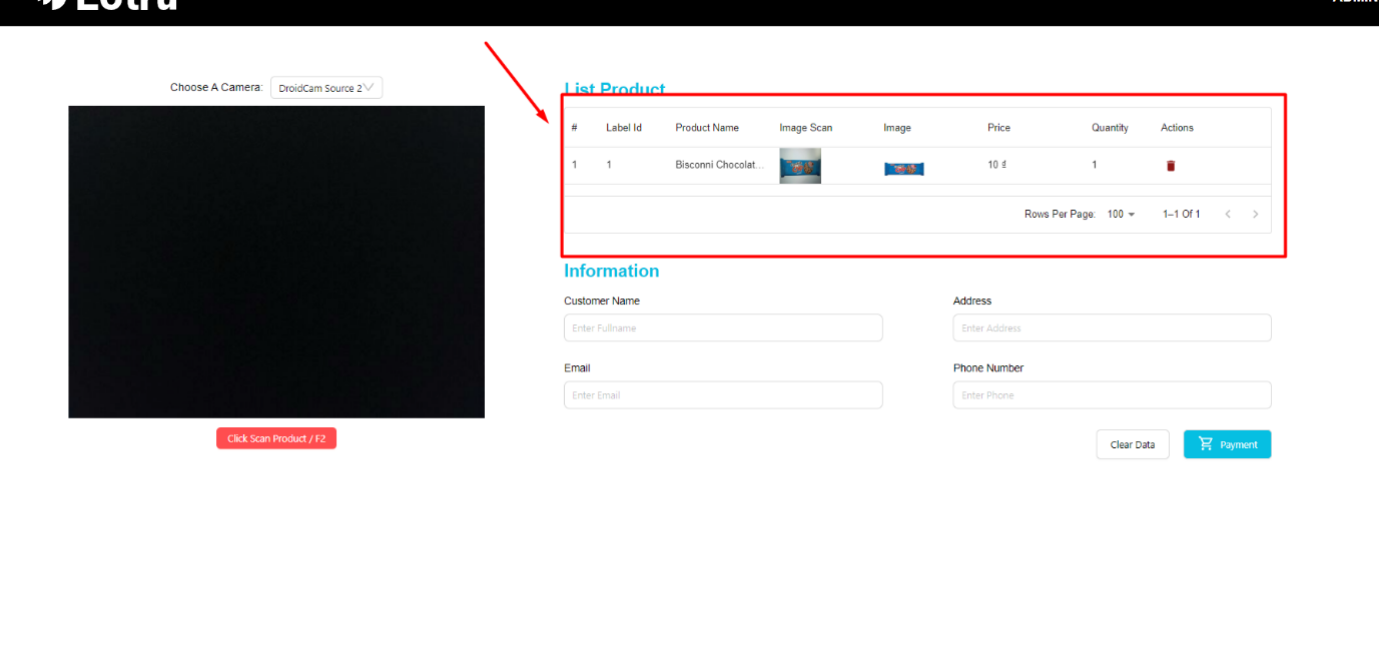
Bước 1: Chủ cửa hàng truy cập vào website và truy cập vào phần nhận diện sản phẩm ở home page

Hình 7. Hình ảnh minh họa chức năng nhận diện sản phẩm(Bước 1)

Bước 2: chụp hình ảnh của sản phẩm.



Hình 8. Hình ảnh minh họa chức năng nhận diện sản phẩm(Bước 2)

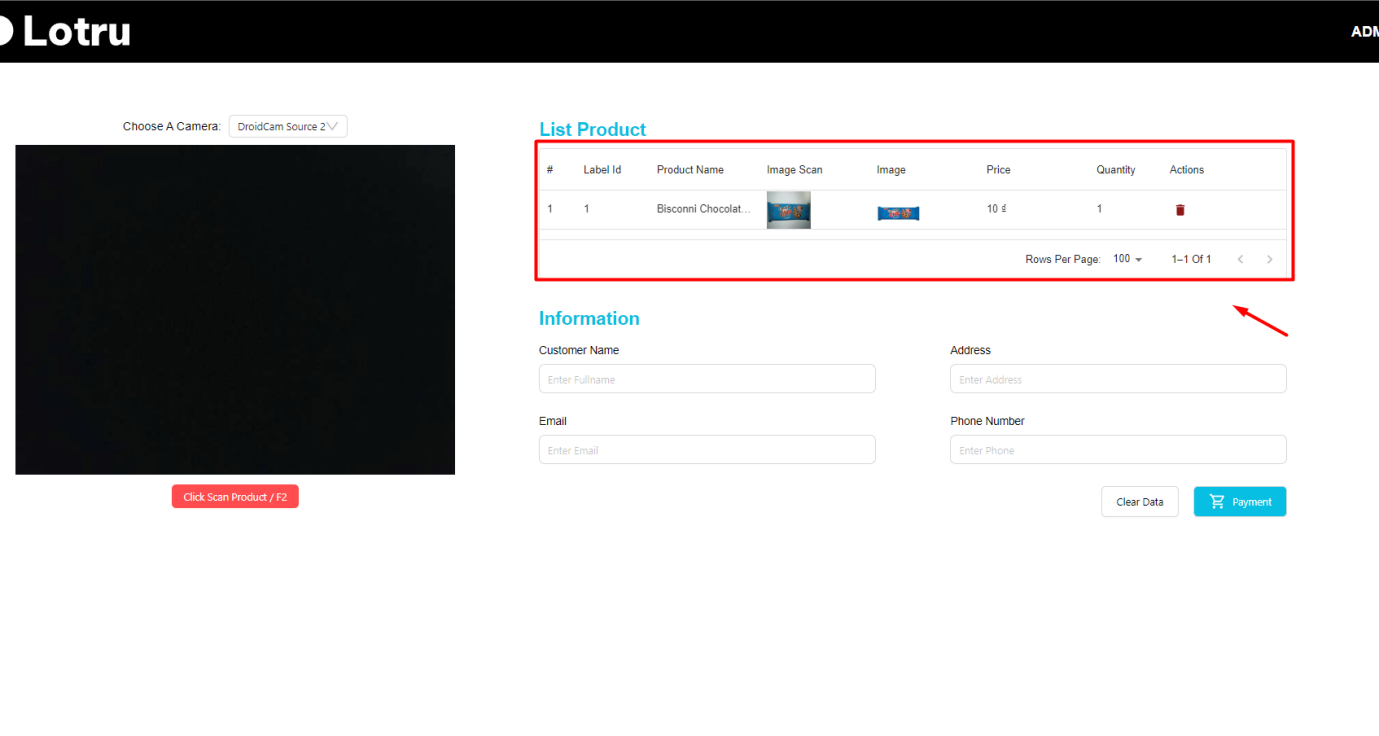
Bước 3: Nhấn vào nút click scan product/f2. Hệ thống xử lý và phân tích hình ảnh thông qua mô hình Vision Transformer (ViT).

Hình 9. Hình ảnh minh họa chức năng nhận diện sản phẩm(Bước 3)

* Kết quả nhận diện được hiển thị, bao gồm tên sản phẩm, giá cả, và thông tin chi tiết khác. Những thông tin này sẽ được hiển thị bên trong bảng trong trang để dễ dàng tiến hành thanh toán
* Kỹ thuật: Sử dụng mô hình ViT để xử lý và phân tích hình ảnh. Mô hình này đã được huấn luyện trên một tập dữ liệu lớn của các sản phẩm từ cửa hàng tiện lợi..

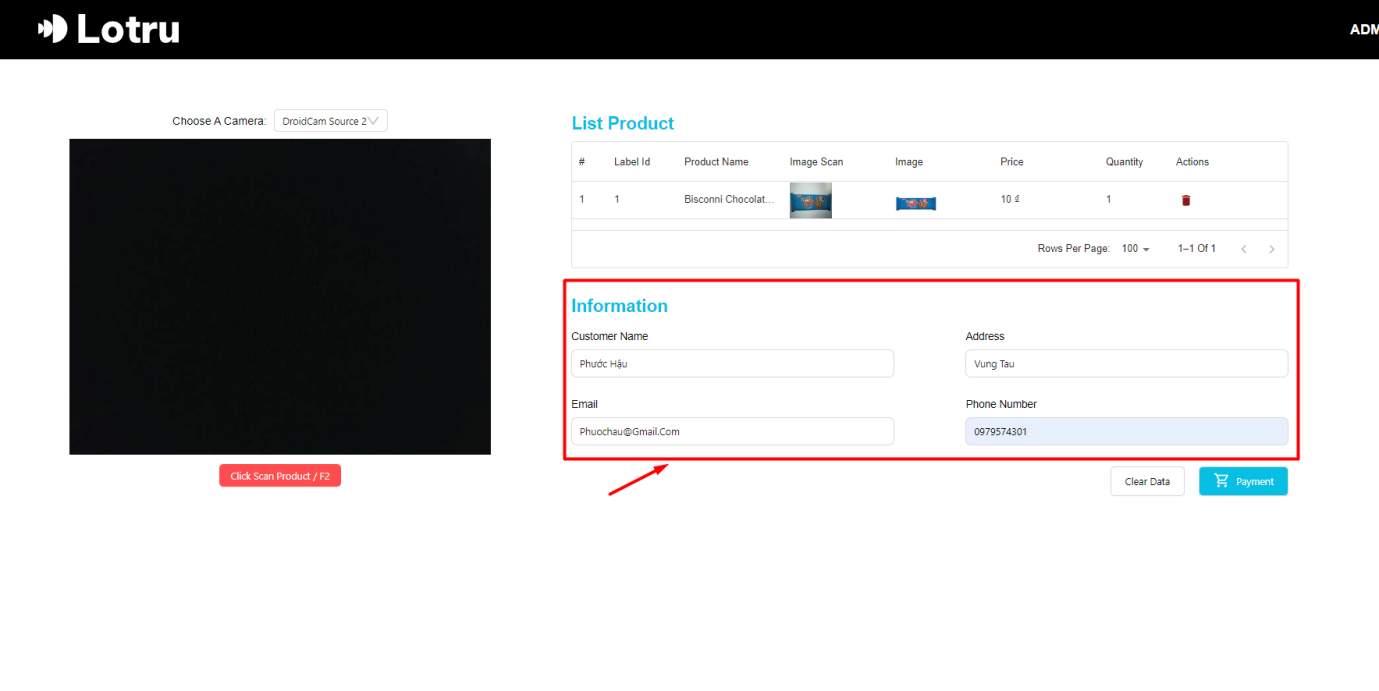
#### **3.1.1.2. Thanh toán**

* Mô tả: Chức năng thanh toán cho phép người bán hàng có thể hoàn tất quá trình mua sắm thông qua website. Người dùng có thể xem lại các sản phẩm đã nhận diện và thanh toán
* Quy trình:

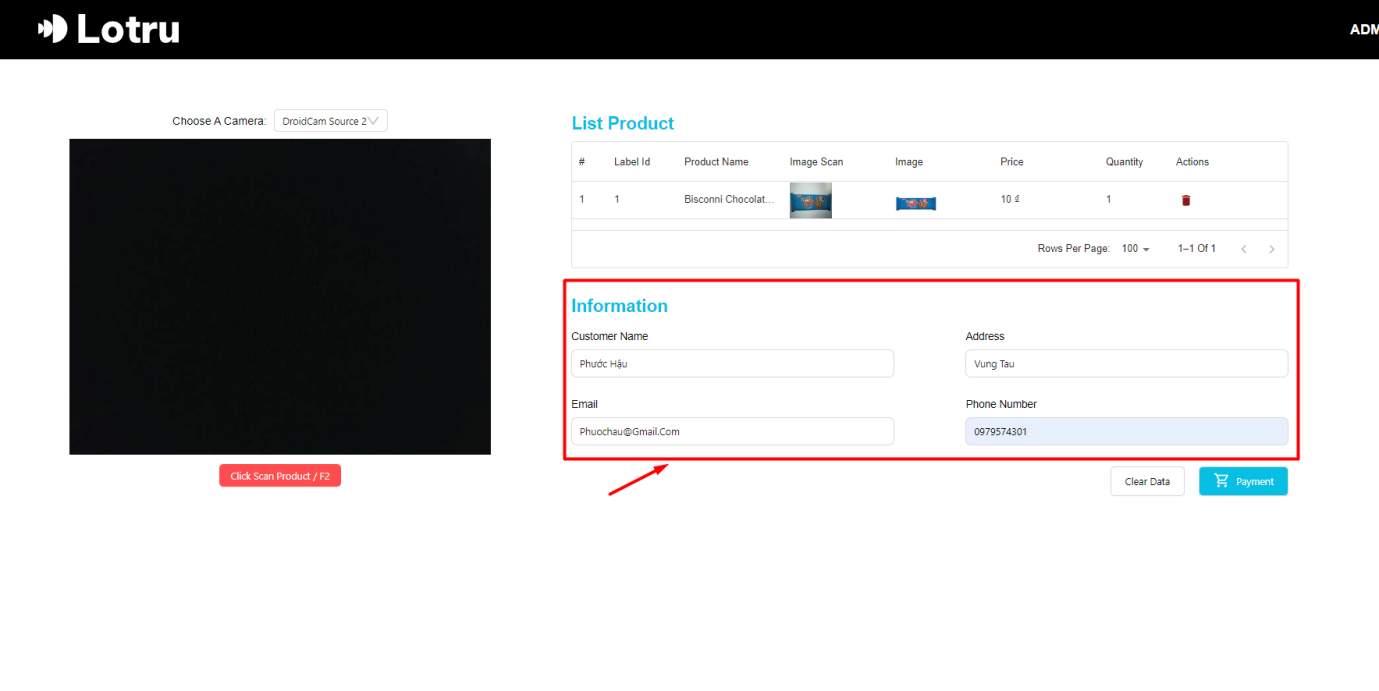
Bước 1: Người dùng xem lại các sản phẩm đã nhận diện.

Hình 10. Hình ảnh minh họa chức năng thành toán(Bước 1)

Bước 2: Điền đầy đủ thông tin của khách hàng vào form

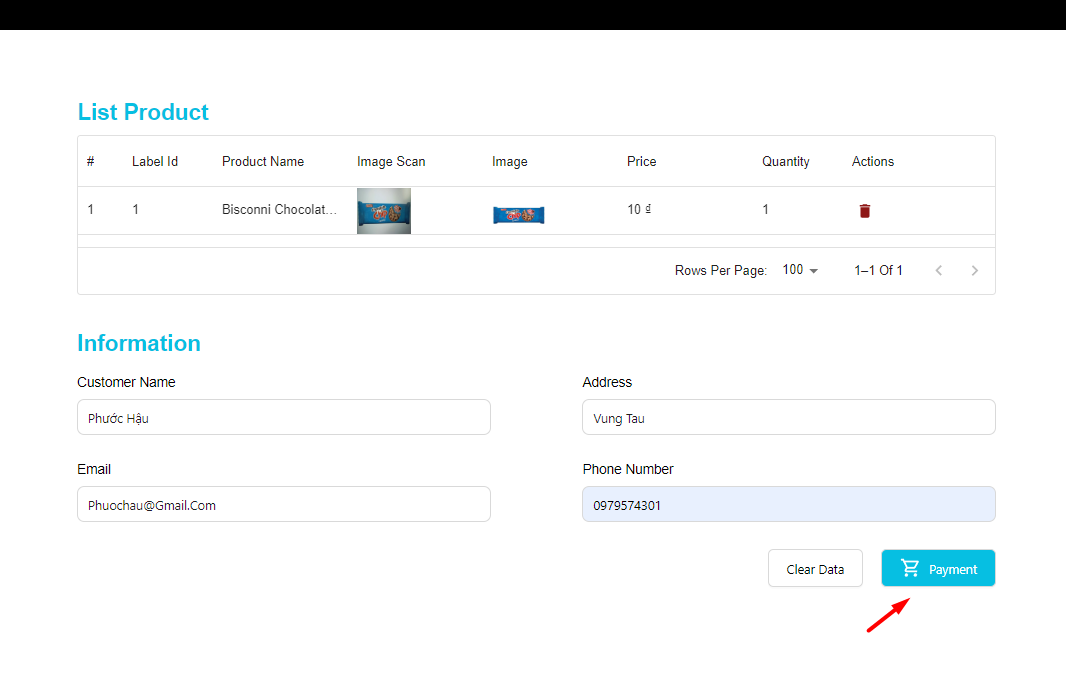
Hình 11. Hình ảnh minh họa chức năng thành toán(Bước 2)

Bước 3: Điền đầy đủ thông tin của khách hàng vào form



Hình 12. Hình ảnh minh họa chức năng thành toán(Bước 3)

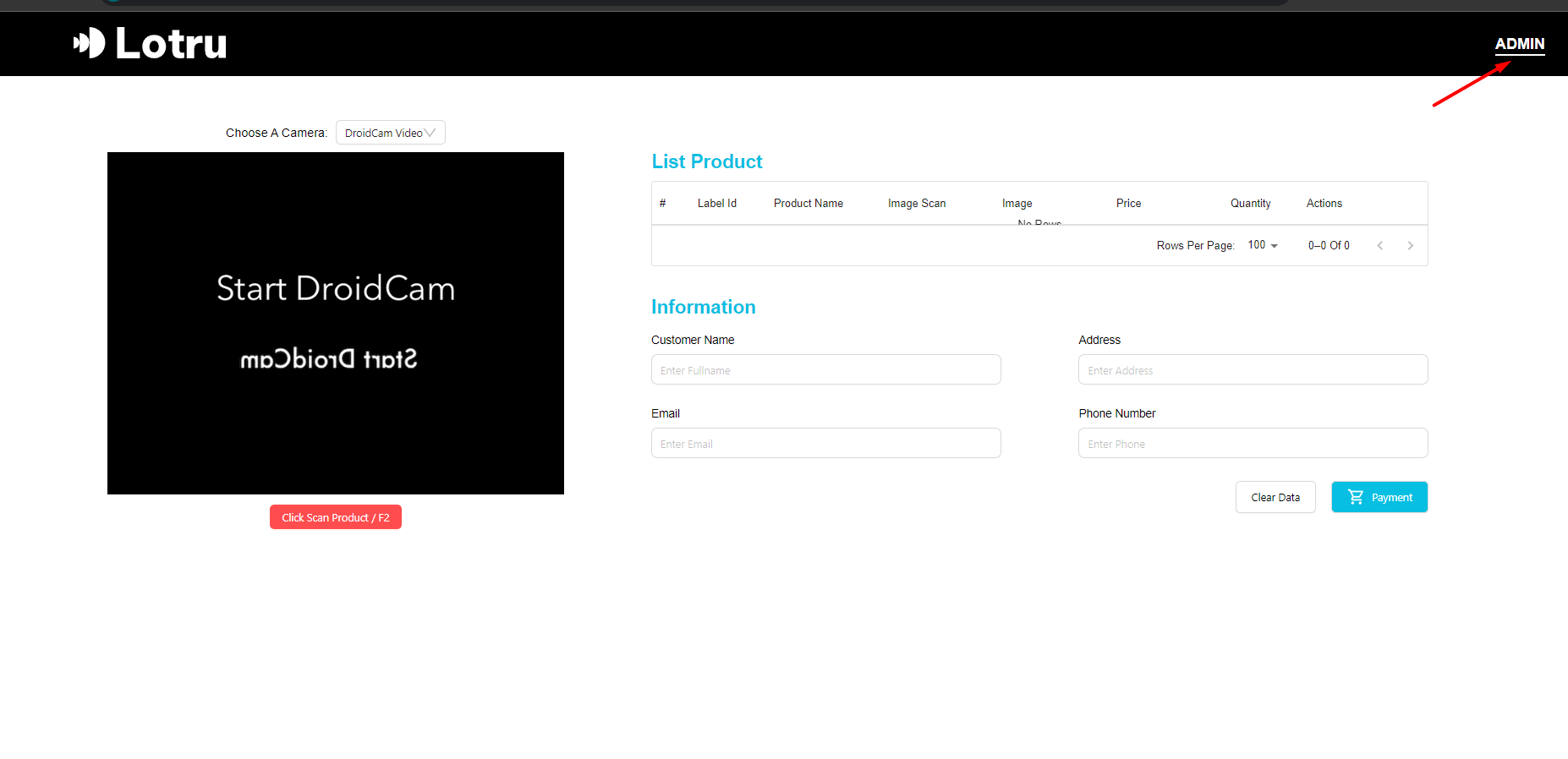
Bước 4: Nhấn vào Nút Payment bên dưới form để tiến thành thanh toán.

Hình 13. Hình ảnh minh họa chức năng thành toán(Bước 4)

#### **3.1.1.3. Quản lý sản phẩm**

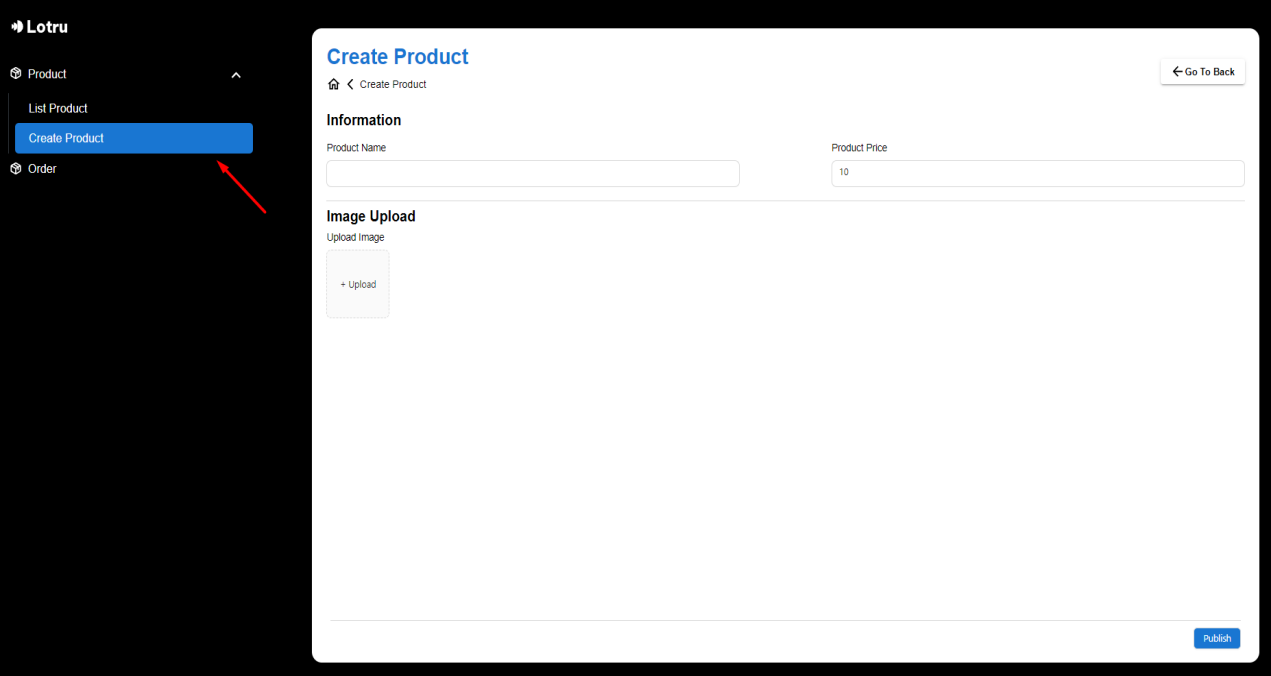
* Mô tả: Chức năng quản lý sản phẩm dành cho quản trị viên của hệ thống, cho phép họ thêm, sửa, hoặc xóa sản phẩm khỏi cơ sở dữ liệu. Dưới đây là mô tả chi tiết về từng phần trong quá trình thêm, sửa, và xóa sản phẩm:
* Thêm mới sản phẩm
* Mô tả: Chức năng này cho phép quản trị viên thêm mới các sản phẩm vào cơ sở dữ liệu. Quản trị viên cần nhập đầy đủ thông tin sản phẩm và tải lên hình ảnh minh họa.
* Quy trình:

Bước 1: Truy cập vào trang web. Trên thanh menu nhấn vào menu admin.



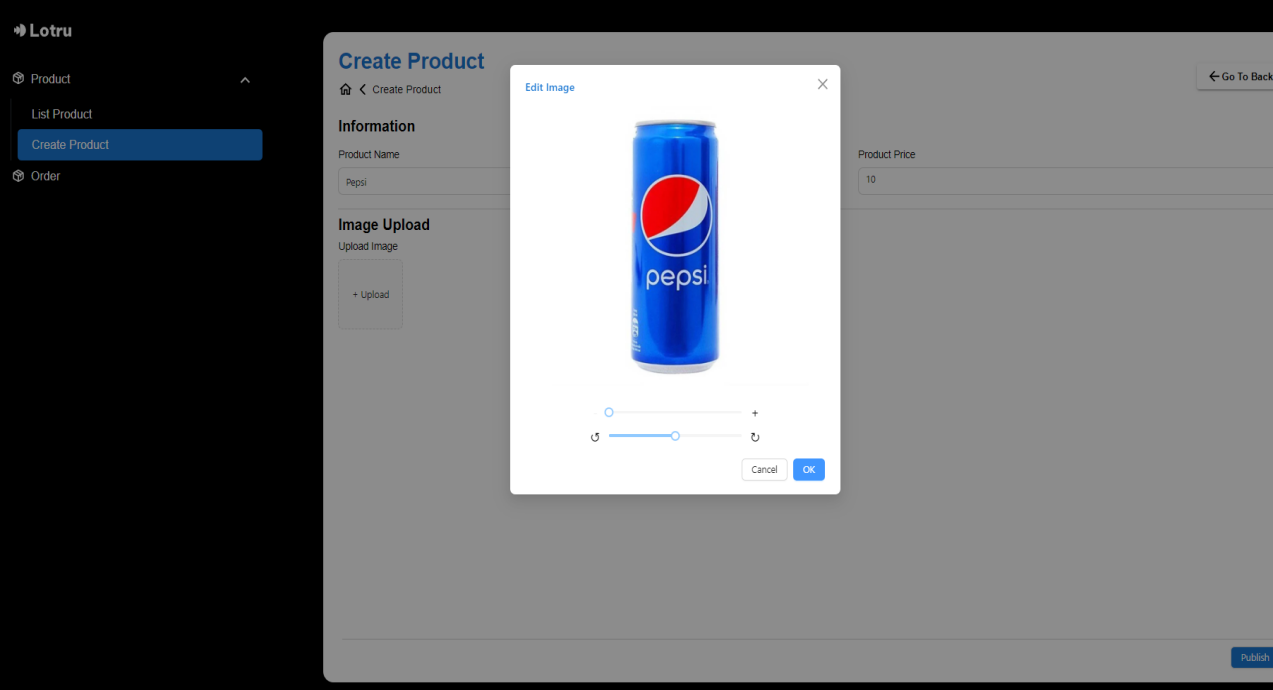
Hình 14. Hình ảnh minh họa chức năng thêm sản phẩm(Bước 1)

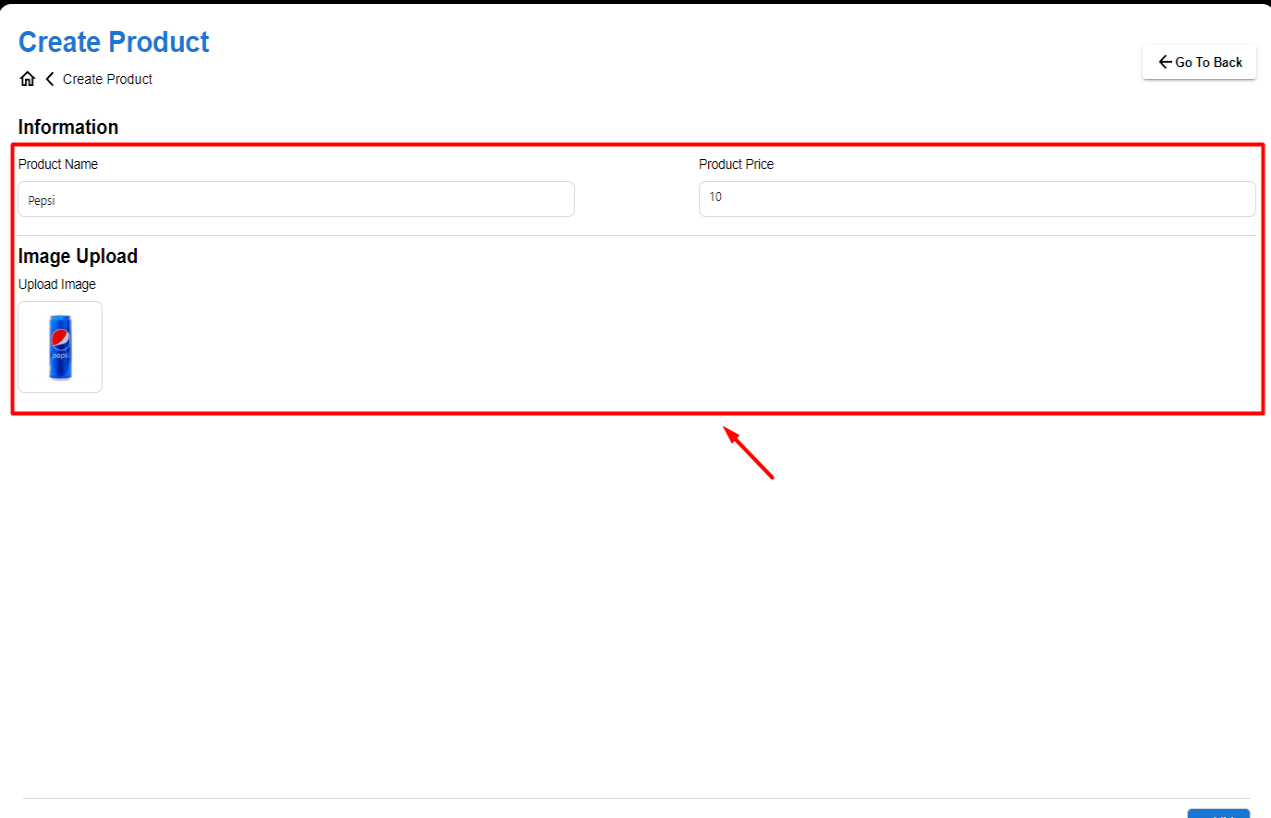
Bước 2: Giao diện page admin sẽ hiện ra. Nhấn chọn vào Product bên trong sidebar bên trái -> nhấn vào menu create product trên sidebar



Hình 15. Hình ảnh minh họa chức năng thêm sản phẩm(Bước 2)

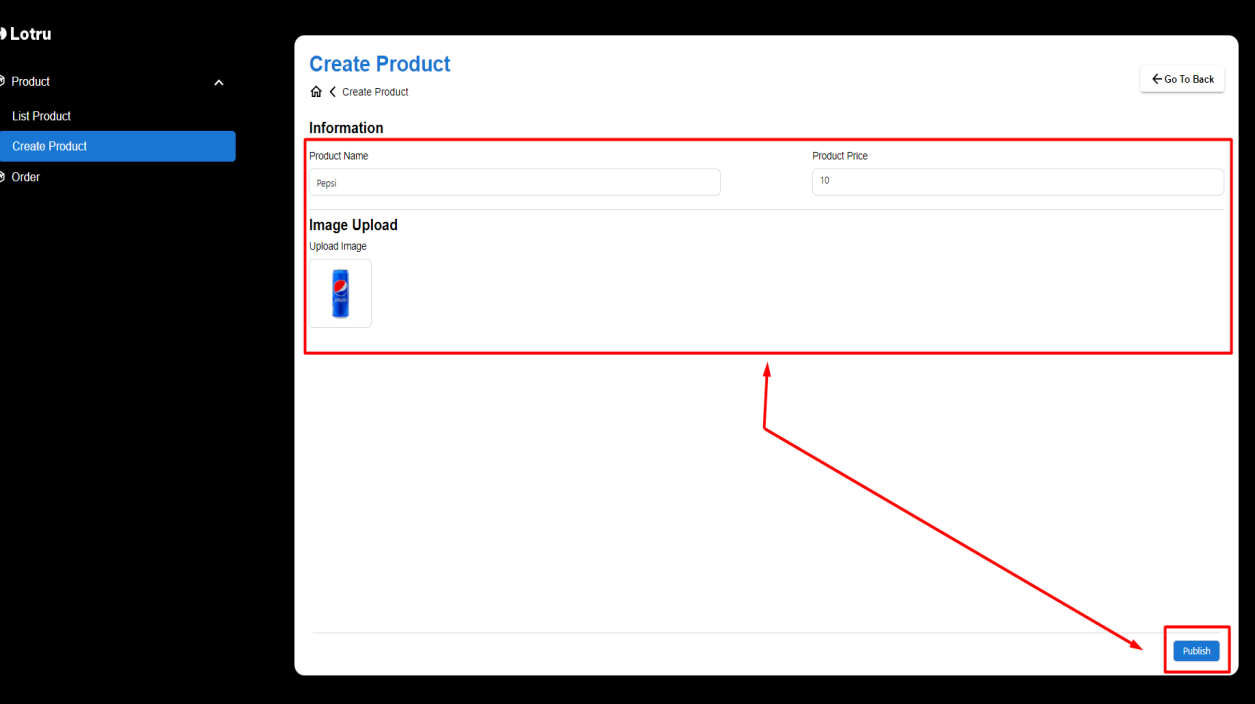
Bước 3: Nhập đầy đủ thông tin của product bao gồm tên product, hình ảnh…





Hình 16. Hình ảnh minh họa chức năng thêm sản phẩm(Bước 3)

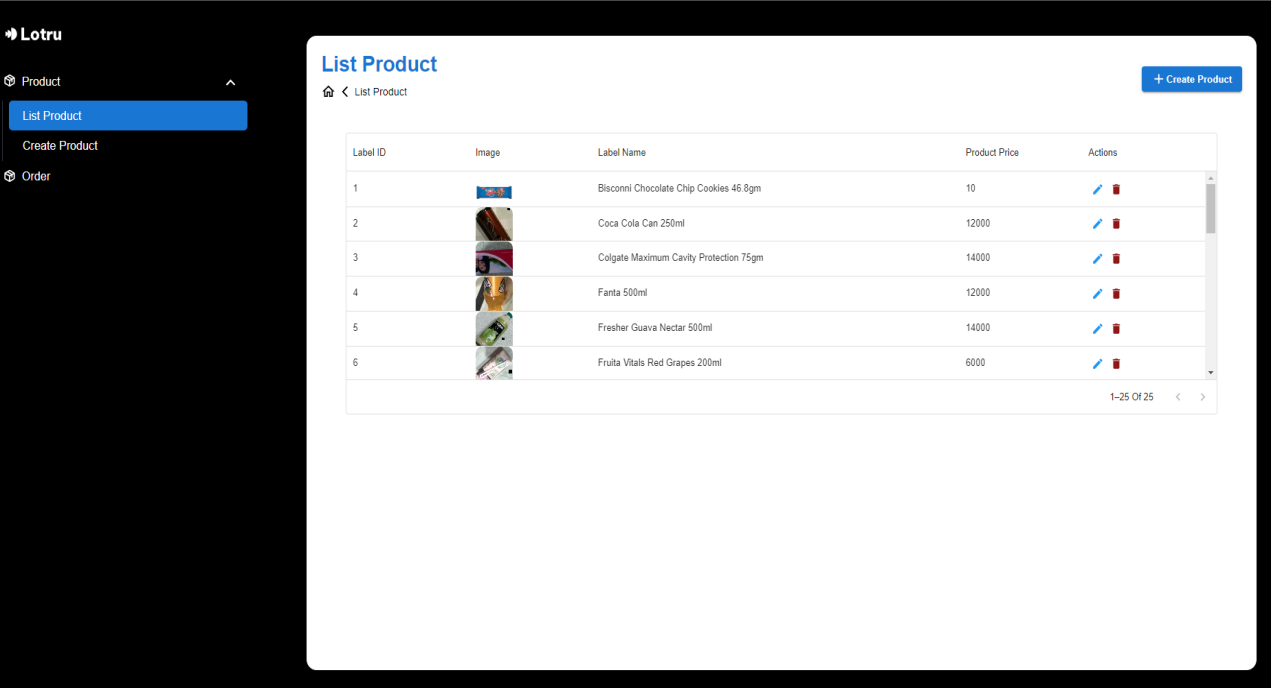
Bước 4: Nhấn vào nút “Pushlish” để thêm sản phẩm.



Hình 17. Hình ảnh minh họa chức năng thêm sản phẩm(Bước 4)

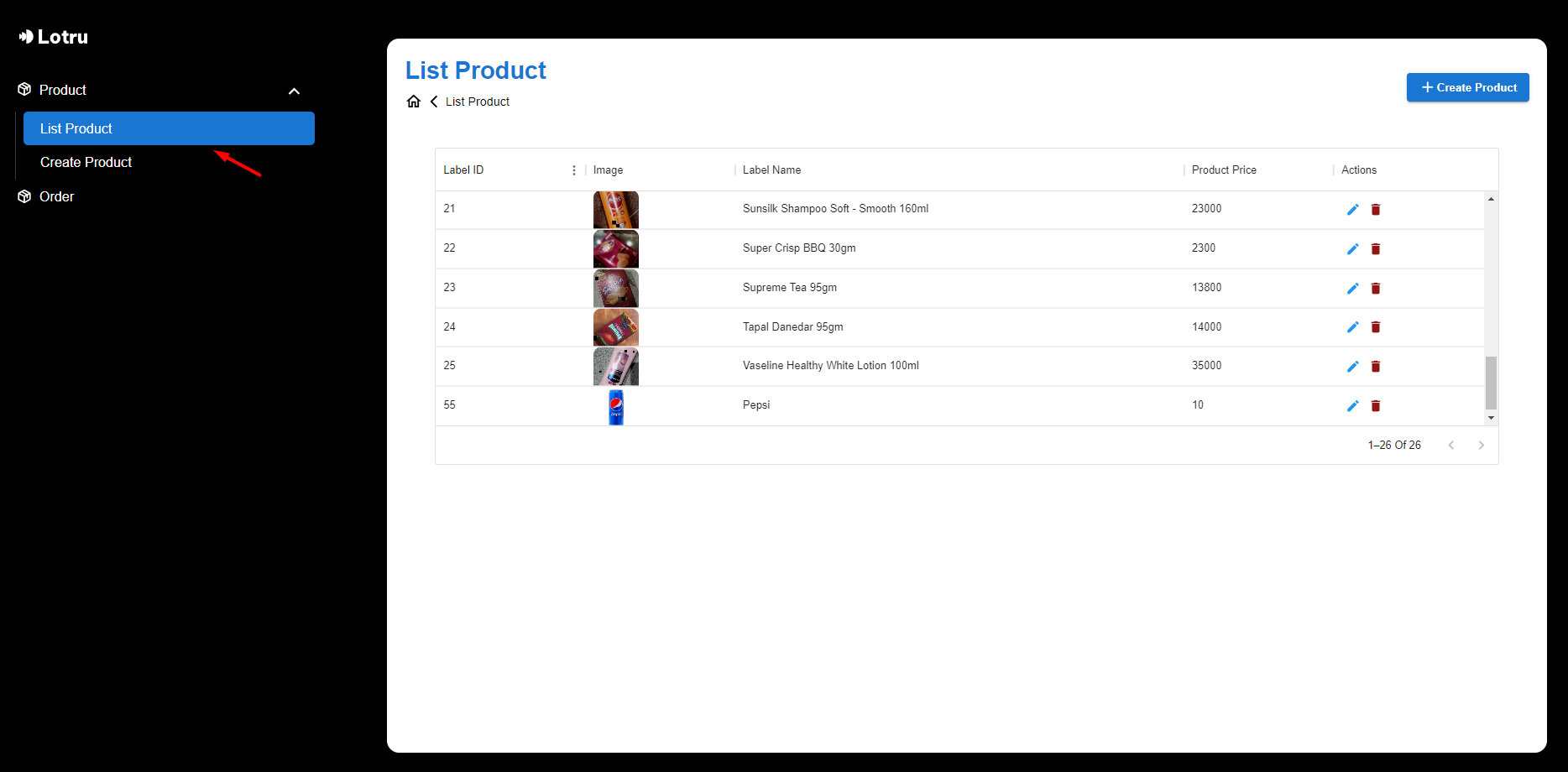
* Sửa thông tin sản phẩm
* Mô tả: Chức năng này cho phép quản trị viên chỉnh sửa thông tin của các sản phẩm đã có trong cơ sở dữ liệu.
* Quy trình:

Bước 1: Truy cập vào trang web. Trên thanh menu nhấn vào menu admin.



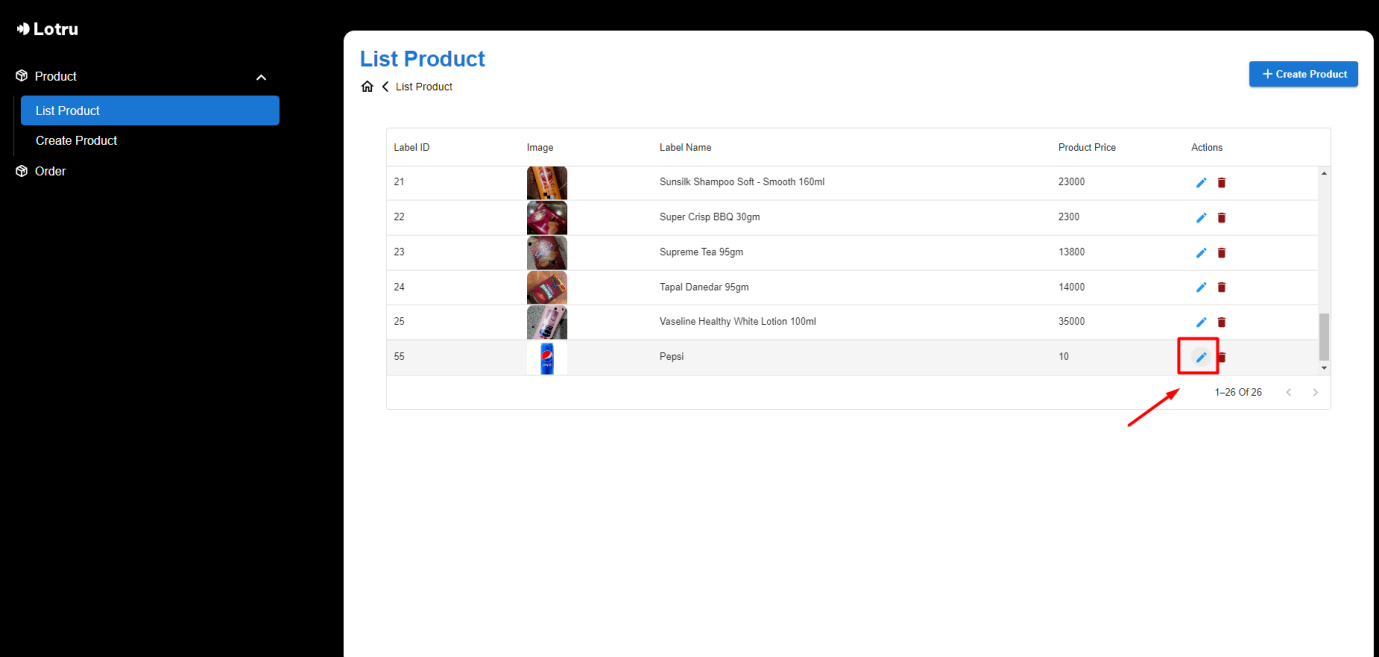
Hình 18. Hình ảnh minh họa chức năng sửa sản phẩm(Bước 1)

Bước 2: Giao diện page admin sẽ hiện ra. Nhấn chọn vào Product bên trong sidebar bên trái -> nhấn vào menu “List Product” trên sidebar



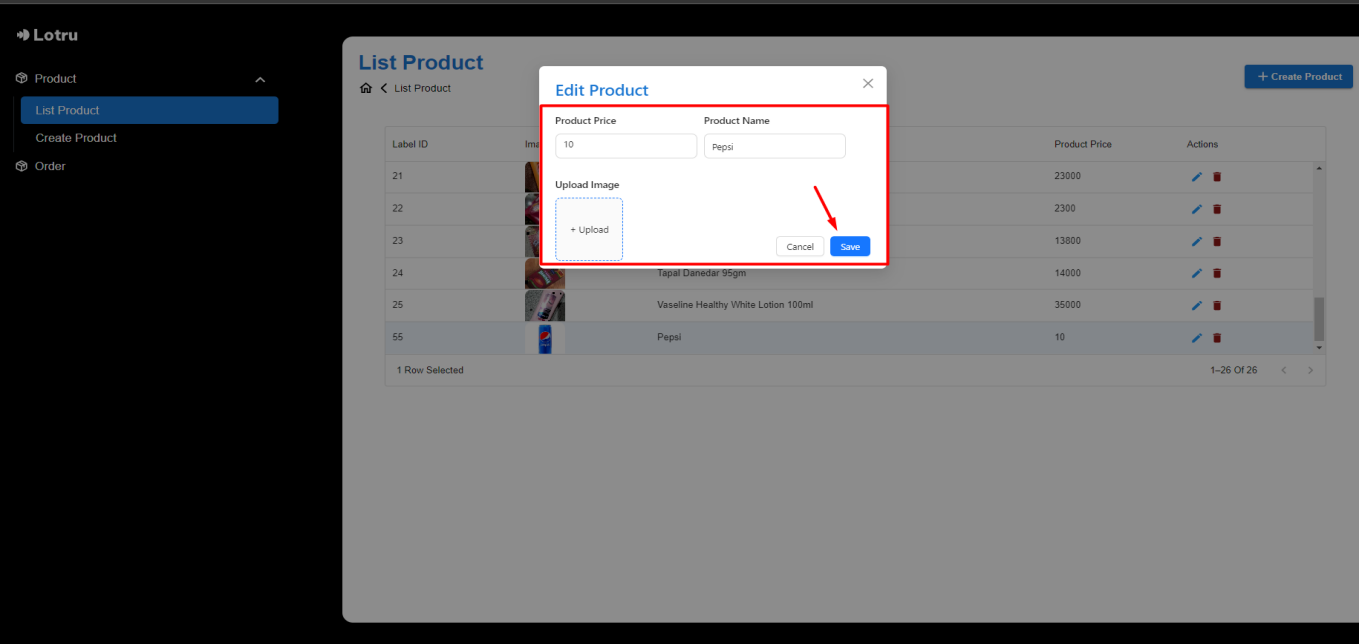
Hình 19. Hình ảnh minh họa chức năng sửa sản phẩm(Bước 2)

Bước 3: Ở đây sẽ hiển thị tất cả những sản phẩm có trong cửa hàng trong bảng. Chọn vào biểu tượng trên cột “Actions” sản phẩm nào muốn chỉnh sửa.



Hình 20. Hình ảnh minh họa chức năng sửa sản phẩm(Bước 3)

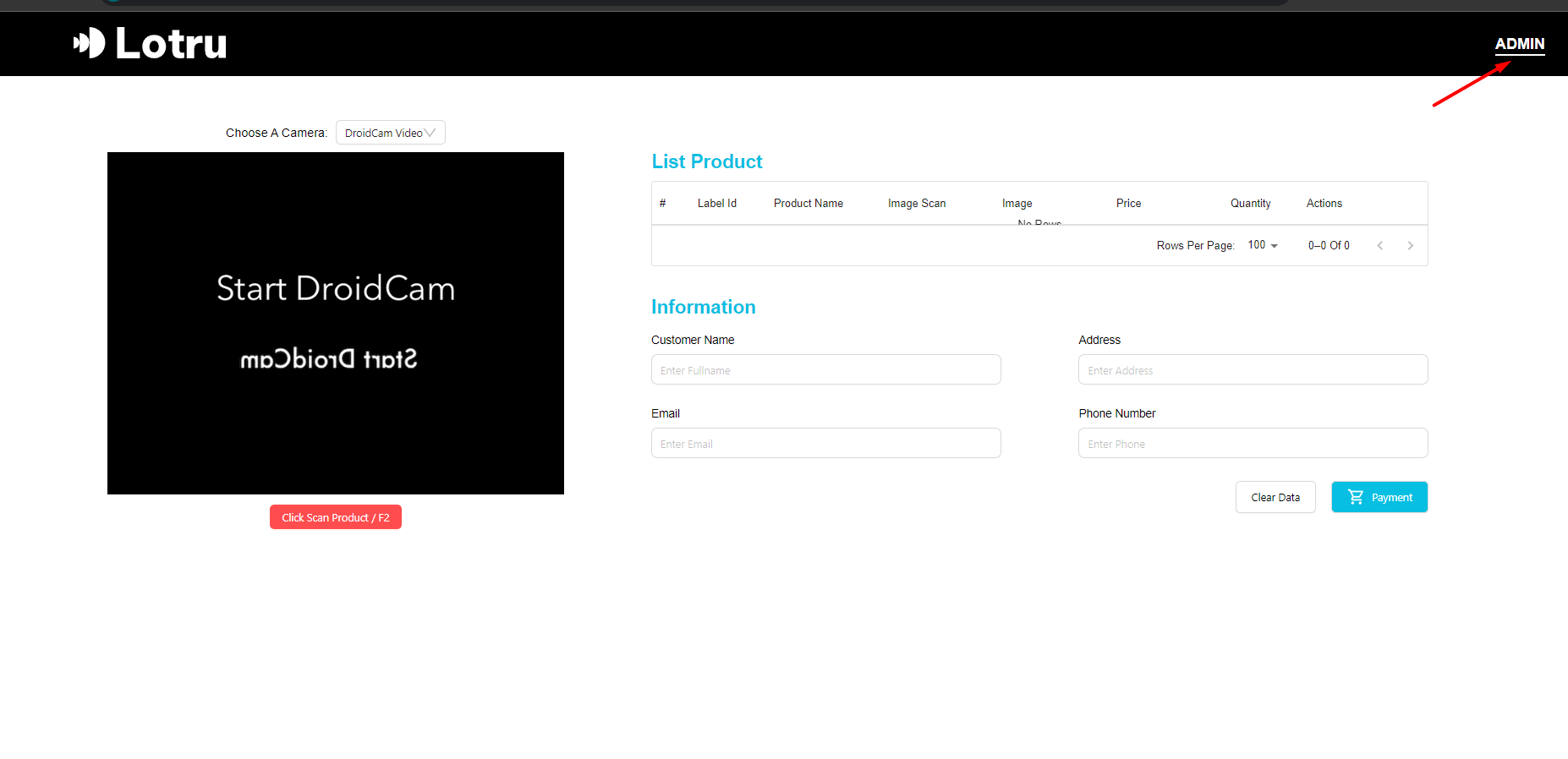
Bước 4: Một cửa sổ sẽ được hiện lên chứa form thông tin của product. Nhập vào cách thông tin cần chỉnh sửa và nhấn vào nút “save” để hoàn tất chỉnh sửa.



Hình 21. Hình ảnh minh họa chức năng sửa sản phẩm(Bước 4)

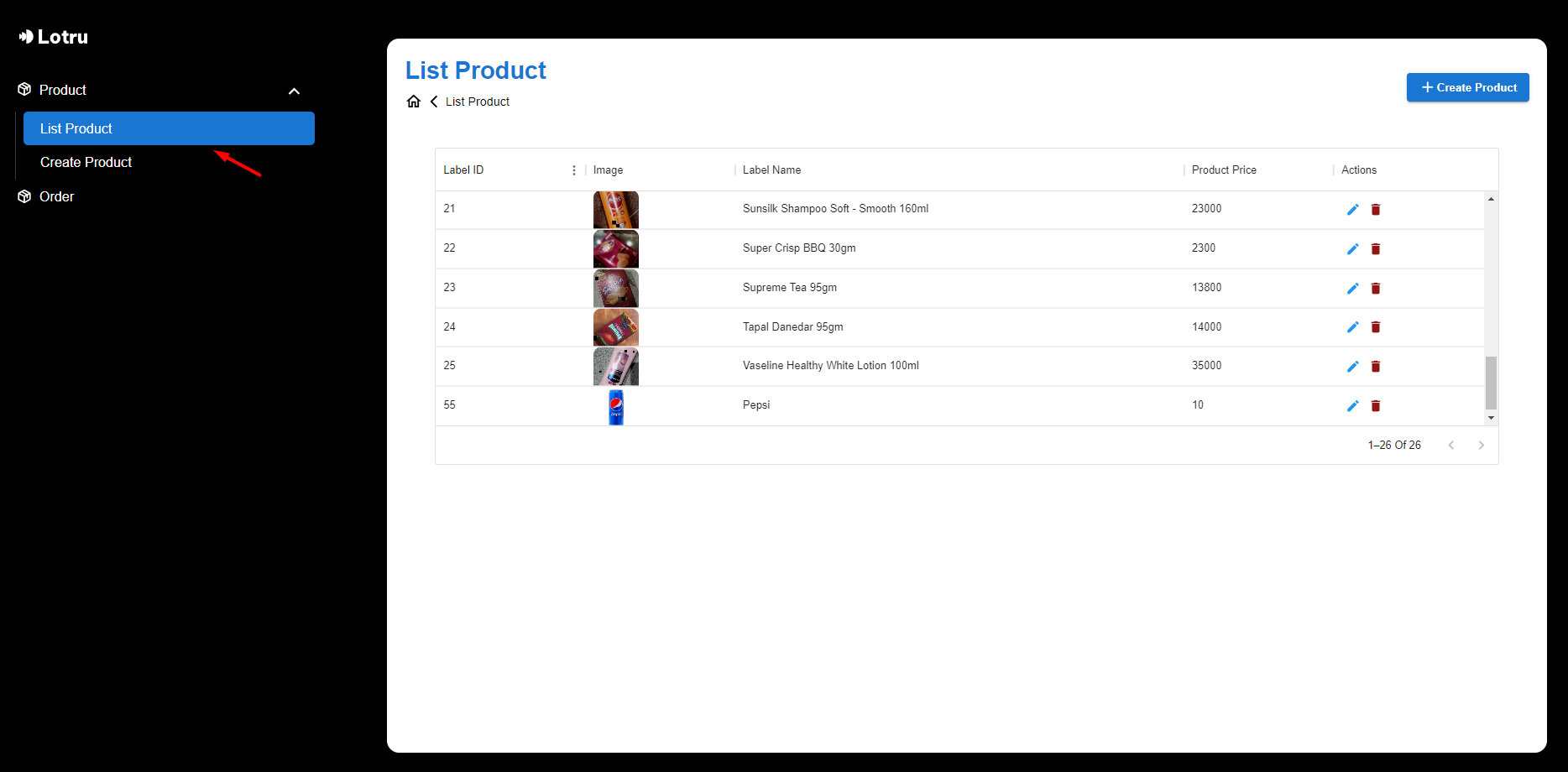
* Xóa sản phẩm
* Mô tả: Chức năng này cho phép quản trị viên xóa các sản phẩm khỏi cơ sở dữ liệu khi không còn cần thiết.
* Quy trình:

Bước 1: Truy cập vào trang web. Trên thanh menu nhấn vào menu admin.



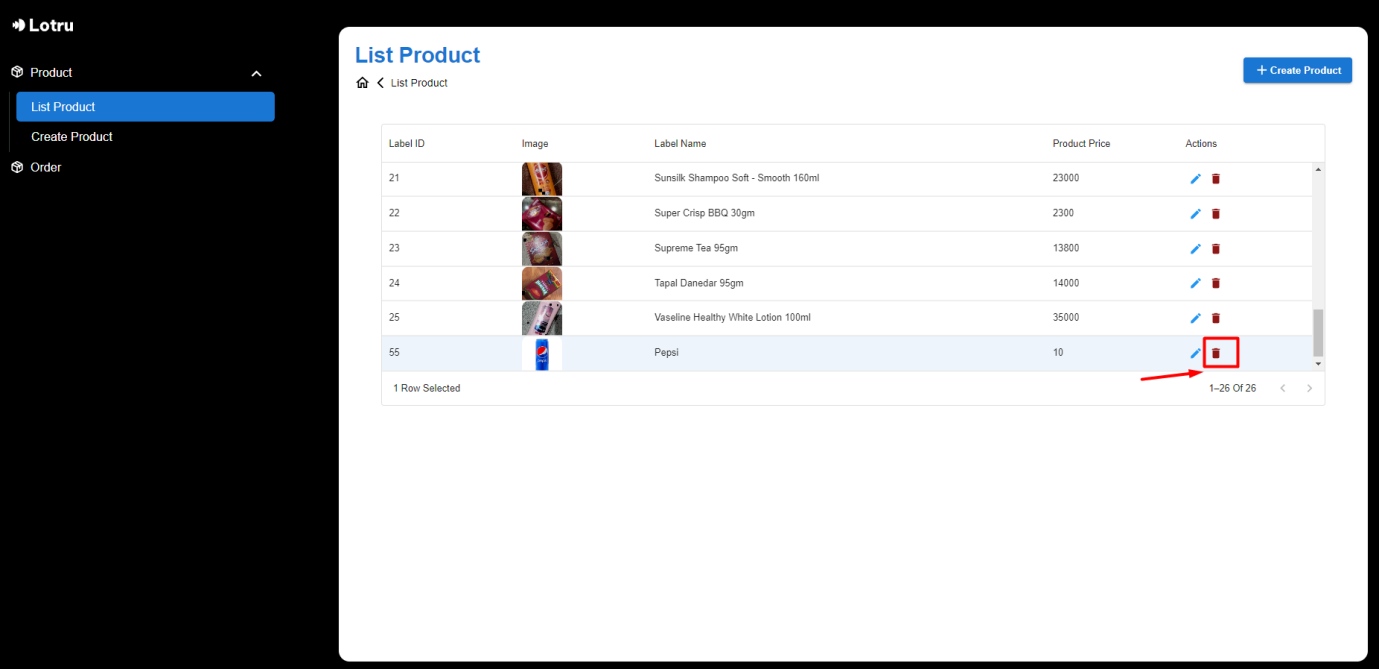
Hình 22. Hình ảnh minh họa chức năng xóa sản phẩm(Bước 1)

Bước 2: Giao diện page admin sẽ hiện ra. Nhấn chọn vào Product bên trong sidebar bên trái -> nhấn vào menu “List Product” trên sidebar



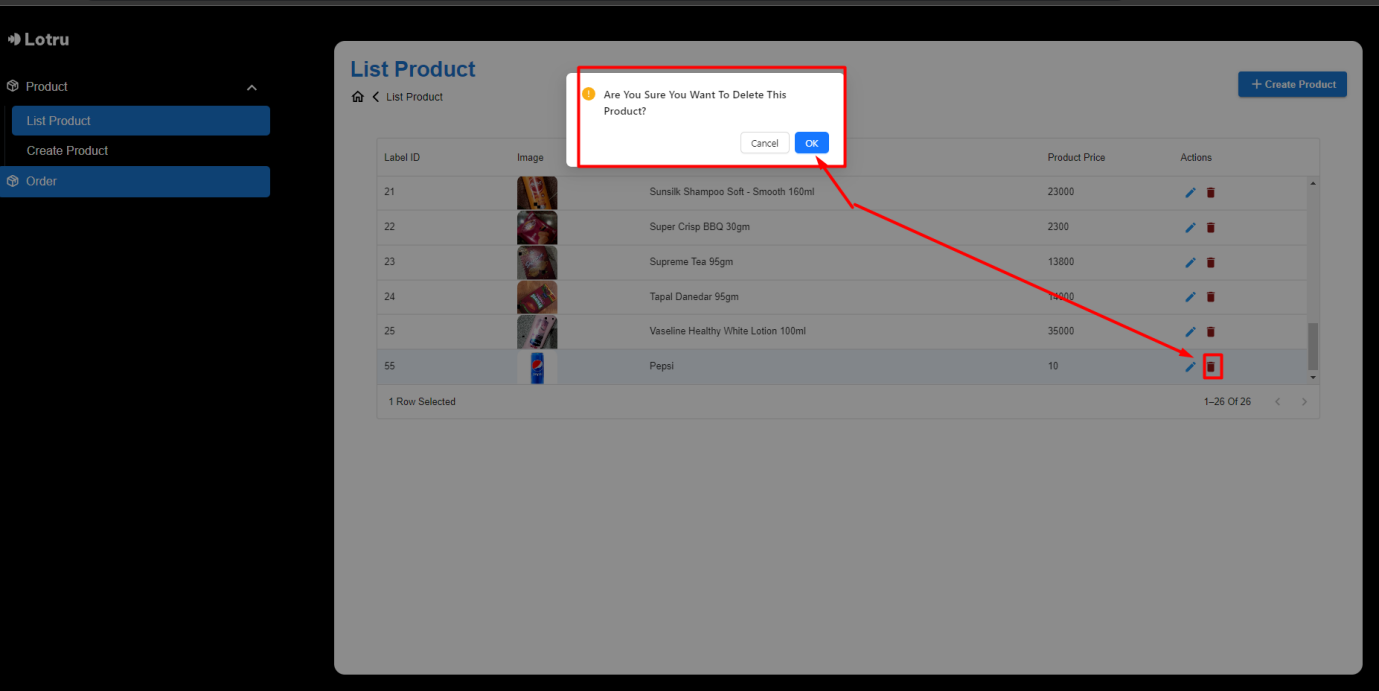
Hình 23. Hình ảnh minh họa chức năng xóa sản phẩm(Bước 2)

Bước 3: Ở đây sẽ hiển thị tất cả những sản phẩm có trong cửa hàng trong bảng. Chọn vào biểu tượng trên cột “Actions” đối với sản phẩm nào muốn xóa.



Hình 24. Hình ảnh minh họa chức năng xóa sản phẩm(Bước 3)

Bước 4: Một cửa sổ sẽ được hiện lên để xác nhận rằng muốn xóa sản phẩm đó hay không. Nếu đồng ý sản phẩm sẽ được xóa ra khỏi cơ sở dữ liệu.

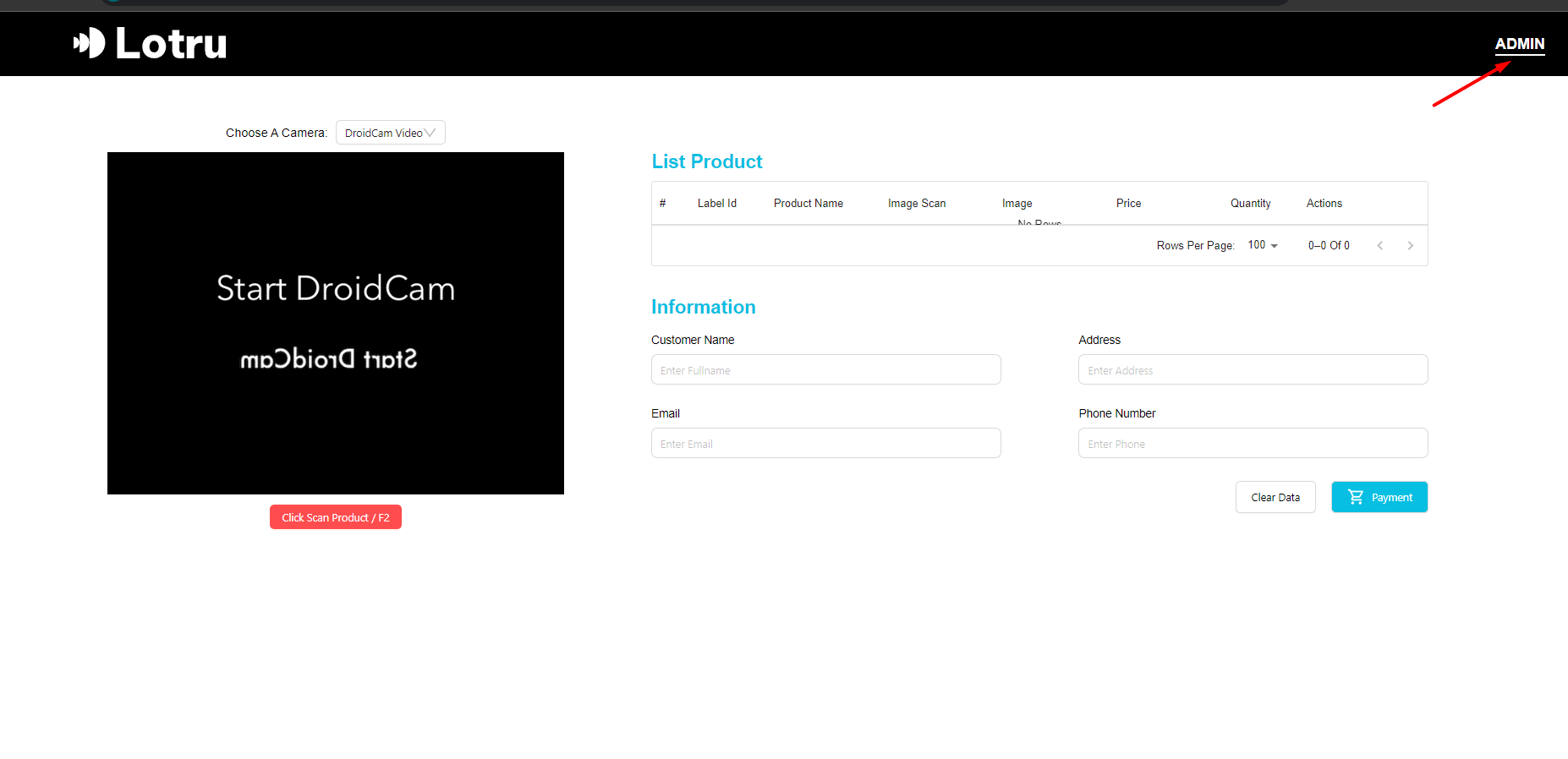


Hình 25. Hình ảnh minh họa chức năng xóa sản phẩm(Bước 4)

#### **3.1.1.3. Quản lý đơn hàng**

* Mô tả: Chức năng này cho phép quản trị viên xem lại các đơn hàng đã đặt. của khách hàng có thể kiểm tra trạng thái đơn hàng, trong khi quản trị viên có thể quản lý và xử lý các đơn hàng.
* Quy trình:

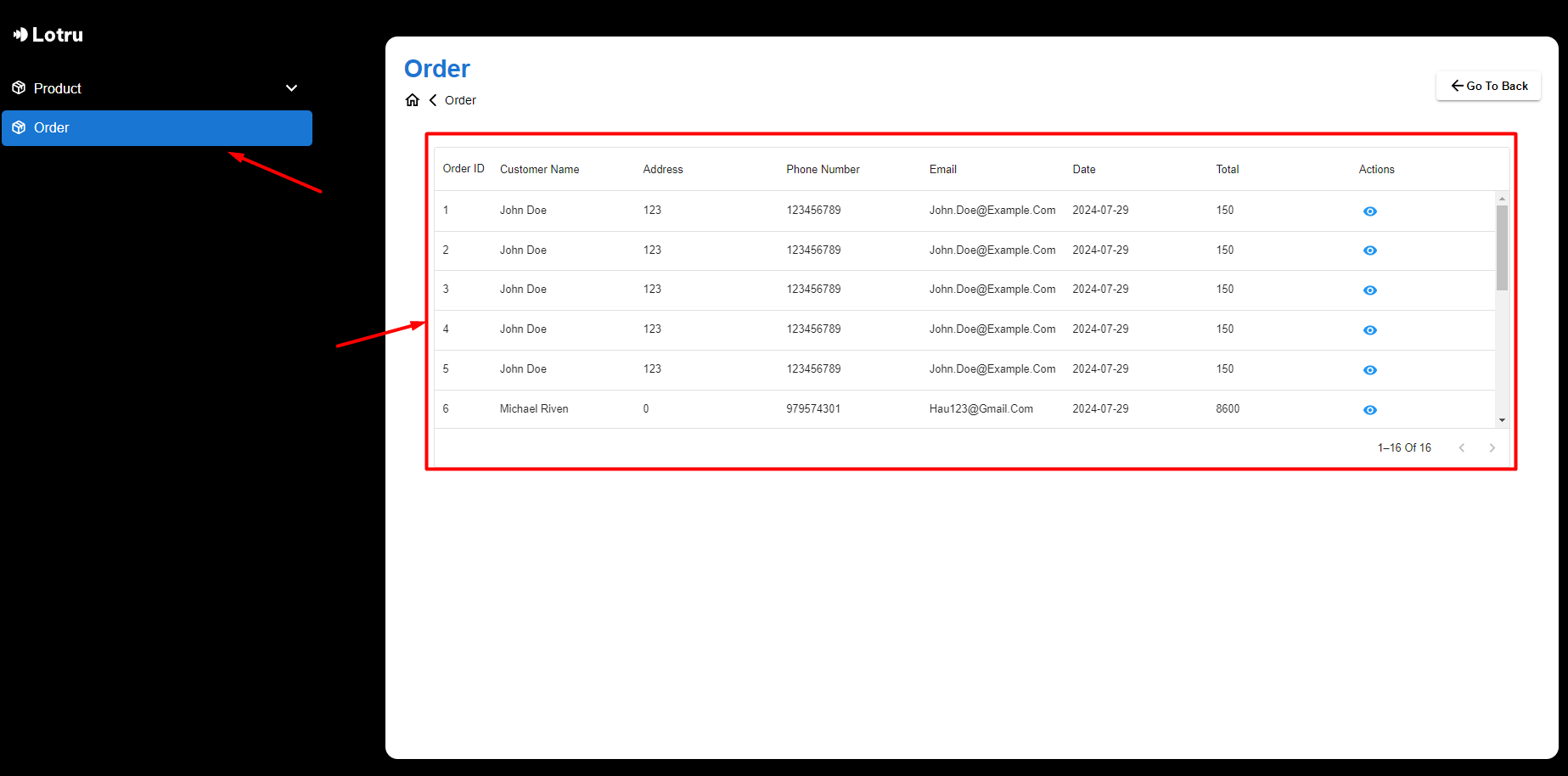
Bước 1: Truy cập vào trang web. Trên thanh menu nhấn vào menu admin.



Hình 26. Hình ảnh minh họa chức năng quản lý đơn hàng(Bước 1)

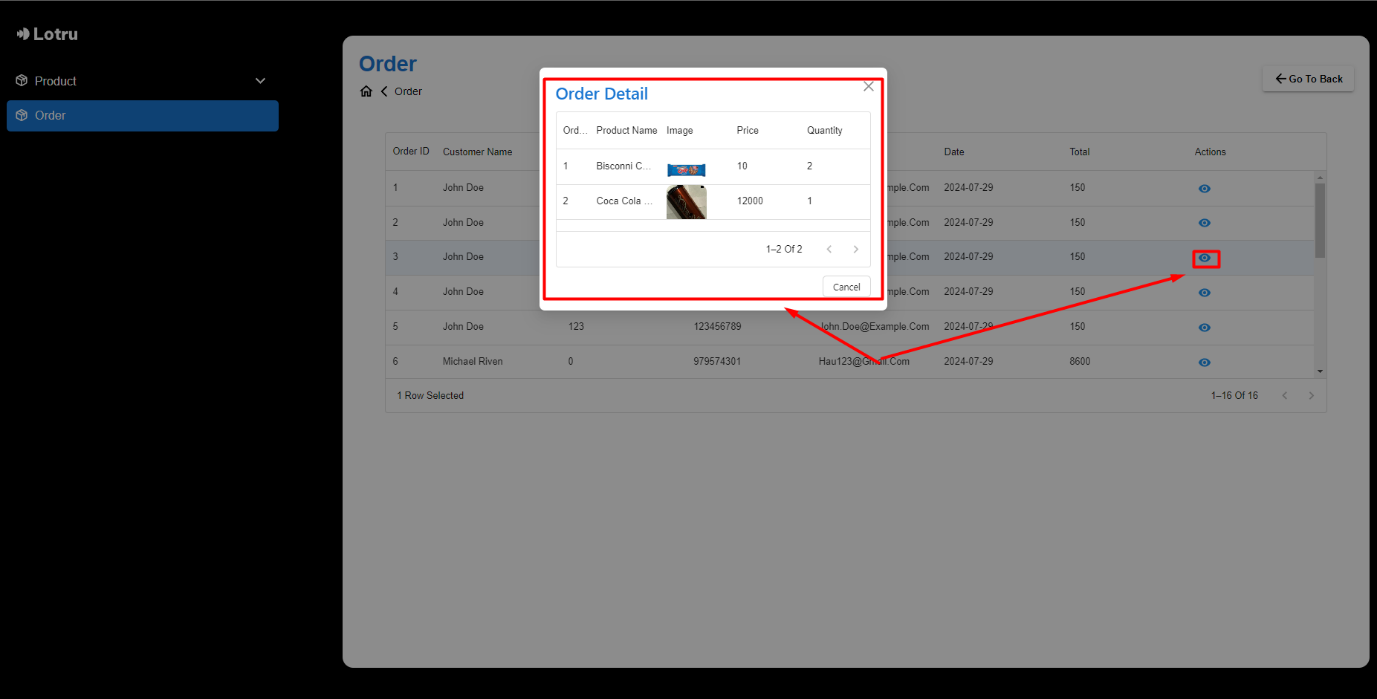
Bước 2: Giao diện page admin sẽ hiện ra. Nhấn chọn vào Order bên trong sidebar

để vào giao diện quản lý đơn hàng



Hình 27. Hình ảnh minh họa chức năng quản lý đơn hàng(Bước 2)

* Ở đây sẽ hiển thị ra các order đã được đặt trước đó. Để xem chi tiết các đơn hàng có thể nhấn vào con mắt  trên 1 hàng chứa đơn hàng bất kì.



Hình 28. Hình ảnh minh họa chức năng xem đơn hàng chi tiết

Hình 3.22. Hình ảnh MINH HỌA

### **3.1.2. Công nghệ sử dụng**

#### **3.1.2.1. Frontend**

* **React**
* Mô tả: React là một thư viện JavaScript dùng để xây dựng giao diện người dùng. Nó được phát triển bởi Facebook và hiện nay được sử dụng rộng rãi trong cộng đồng phát triển web.
* Lý do sử dụng: React cho phép xây dựng các thành phần giao diện tái sử dụng, giúp việc phát triển và bảo trì ứng dụng trở nên dễ dàng hơn. Nó cung cấp cơ chế ảo hóa DOM, giúp tối ưu hóa hiệu suất ứng dụng.
* **TypeScript**
* Mô tả: TypeScript là một ngôn ngữ lập trình phát triển từ JavaScript, bổ sung thêm kiểu dữ liệu tĩnh. Nó được phát triển bởi Microsoft và giúp các lập trình viên viết mã rõ ràng và dễ bảo trì hơn.
* Lý do sử dụng: TypeScript giúp phát hiện lỗi ngay trong quá trình viết code, giảm thiểu lỗi runtime và cải thiện tính minh bạch của mã nguồn. Nó cũng cung cấp tính năng autocompletion và refactoring hiệu quả.
* **Redux Toolkit**
* Mô tả: Redux Toolkit là một bộ công cụ chính thức để quản lý trạng thái ứng dụng với Redux. Nó cung cấp các hàm và tiện ích giúp đơn giản hóa quá trình sử dụng Redux.
* Lý do sử dụng: Redux Toolkit đơn giản hóa quá trình viết mã Redux, giảm bớt boilerplate và cung cấp các công cụ mạnh mẽ. Nó cũng giúp quản lý trạng thái ứng dụng một cách hiệu quả và dễ dàng mở rộng.
* **Cloudinary**
* Mô tả: Cloudinary là một dịch vụ quản lý và tối ưu hóa hình ảnh và video. Nó cung cấp các API mạnh mẽ để tải lên, lưu trữ và xử lý hình ảnh.
* Lý do sử dụng: Cloudinary giúp dễ dàng tải lên, lưu trữ và xử lý hình ảnh một cách hiệu quả. Nó cũng cung cấp các tính năng như nén hình ảnh, chuyển đổi định dạng và tối ưu hóa hiệu suất tải trang.
* **axios**
* Mô tả: axios là một thư viện JavaScript dùng để thực hiện các yêu cầu HTTP. Nó hỗ trợ Promise và tương thích tốt với các trình duyệt hiện đại.
* Lý do sử dụng: axios cung cấp một API đơn giản và dễ sử dụng để tương tác với backend. Nó hỗ trợ các tính năng như hủy bỏ yêu cầu, tự động chuyển đổi dữ liệu và xử lý lỗi.

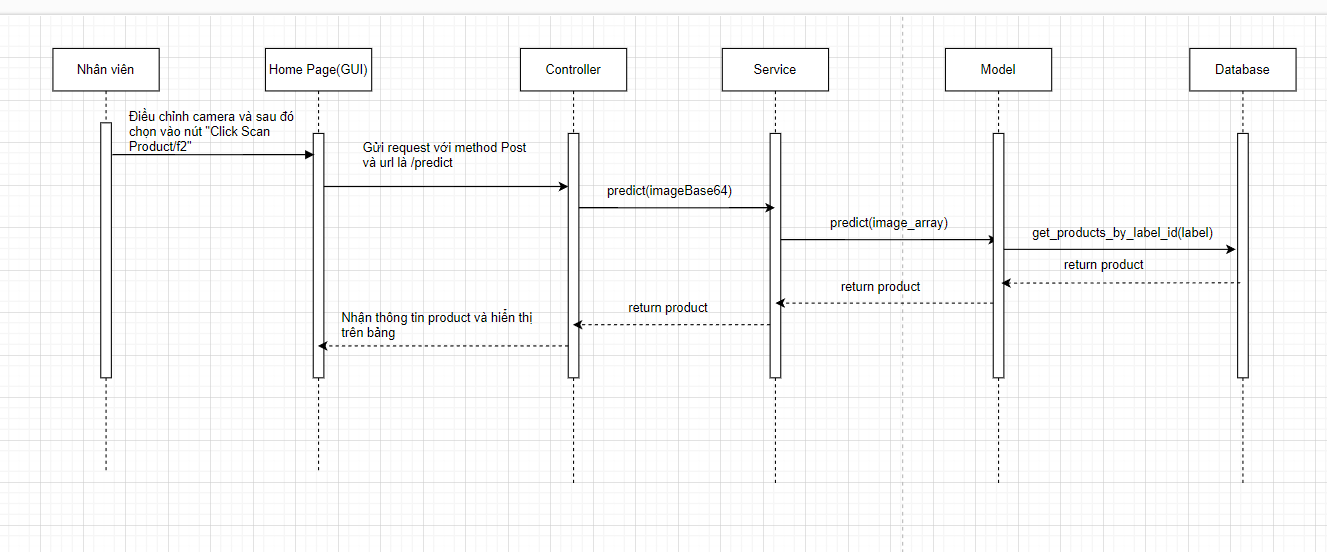
#### **3.1.2.1. Backend**

* **fastapi**
* Mô tả: fastapi là một framework web hiện đại và nhanh chóng cho Python. Nó được thiết kế để dễ dàng xây dựng các API RESTful với hiệu suất cao.
* Lý do sử dụng: fastapi cho phép xây dựng các API hiệu suất cao và dễ dàng, hỗ trợ đầy đủ các tính năng như xác thực, quản lý phiên làm việc và tài liệu tự động. Nó cũng tương thích tốt với các công nghệ async hiện đại.
* **tensorflow**
* Mô tả: tensorflow là một thư viện mã nguồn mở dành cho học máy và học sâu. Nó được phát triển bởi Google và hiện nay là một trong những thư viện học máy phổ biến nhất.
* Lý do sử dụng: tensorflow cung cấp các công cụ mạnh mẽ để xây dựng và triển khai các mô hình học sâu cho việc nhận diện sản phẩm. Nó hỗ trợ nhiều loại mô hình khác nhau và có khả năng mở rộng cao.

Link github: https://github.com/phuochau132/KhoaLuanBE

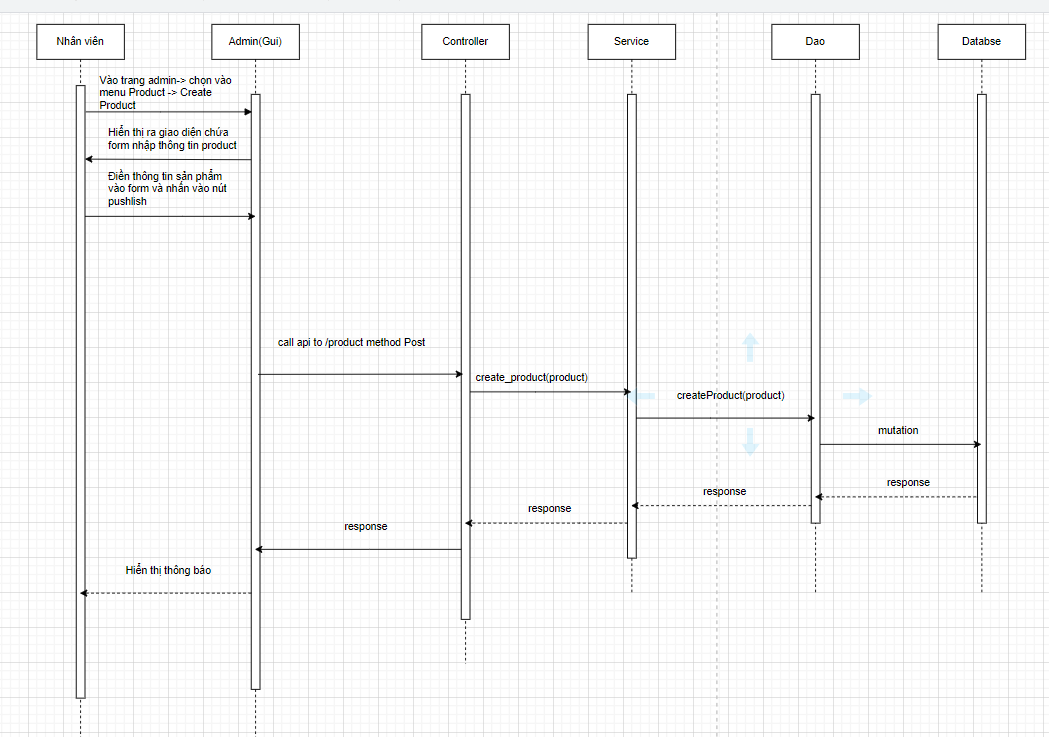
### **3.1.3. Lược đồ tuần tự (Sequence Diagram)**

**3.1.3.1. Chức năng nhận diện sản phẩm**



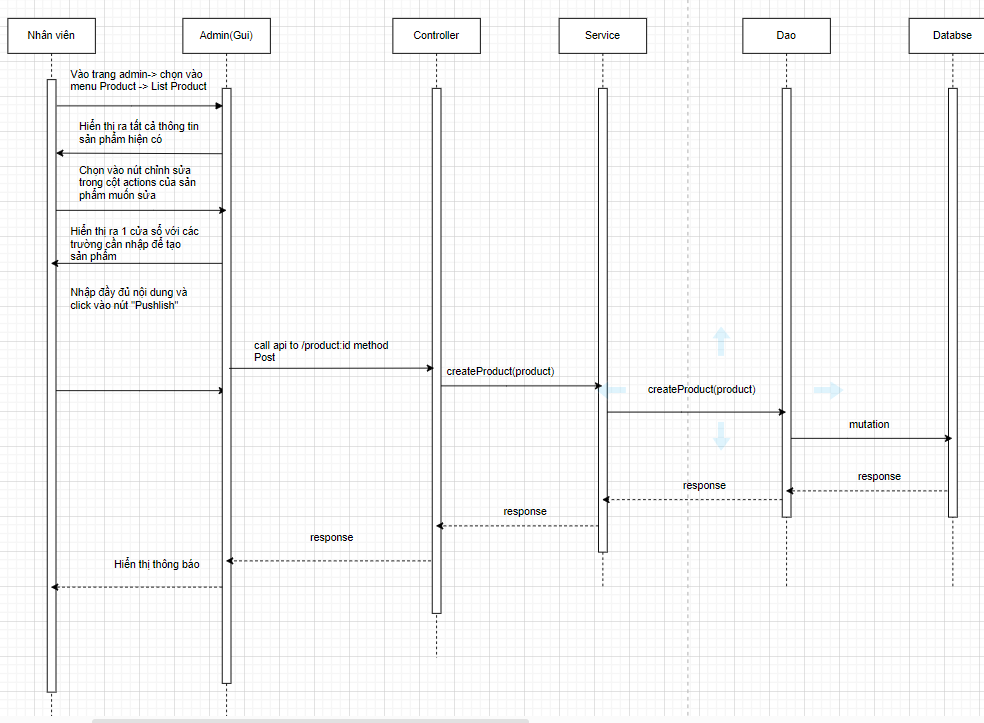
Hình 29. Lược đồ tuần tự chức năng nhận diện sản phẩm

#### **3.1.3.2. Chức năng thêm sản phẩm(Quản lý sản phẩm)**



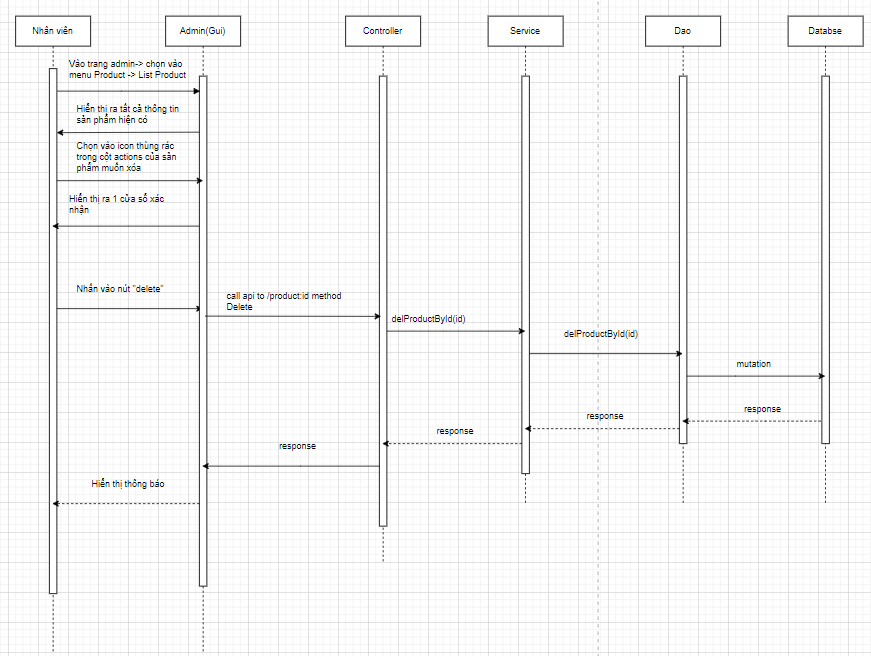
Hình 30. Lược đồ tuần tự chức năng thêm sản phẩm

#### **3.1.3.3. Chức năng sửa sản phẩm(Quản lý sản phẩm)**



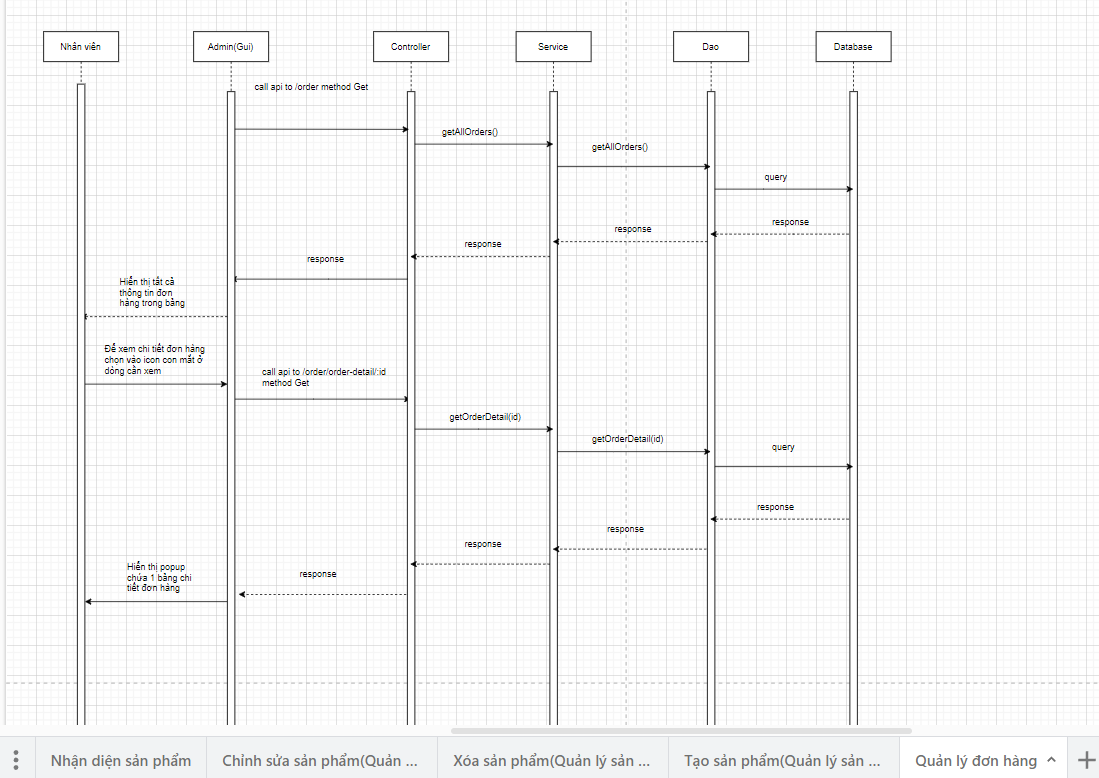
Hình 31. Lược đồ tuần tự chức năng sửa sản phẩm

#### **3.1.3.4. Chức năng xóa sản phẩm(Quản lý sản phẩm)**



Hình 32. Lược đồ tuần tự chức năng xóa sản phẩm

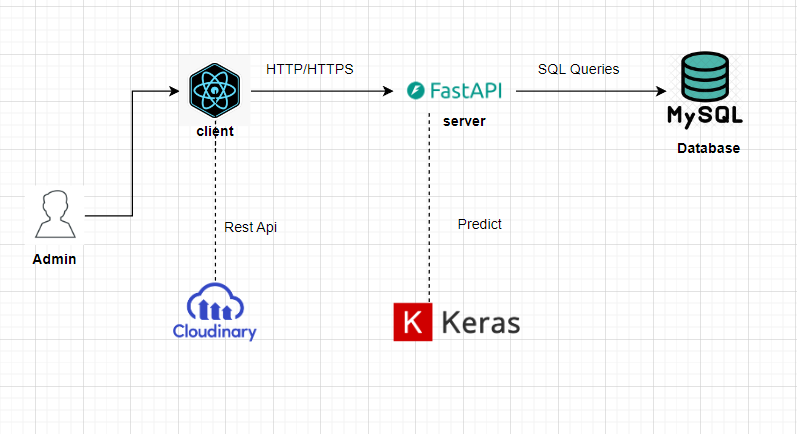
#### **3.1.3.5. Chức năng xem đơn hàng(Quản lý đơn hàng)**



Hình 33. Lược đồ tuần tự chức năng xem đơn hàng(Quản lý đơn hàng)

* **Liên kết:** <https://drive.google.com/file/d/1BTdWwEgLX3J8z1M3sxnmnKf1Vp1r8JRS/view?usp=sharing>

### **3.1.4. Sơ đồ hệ thống**



Hình 34. Sơ đồ Hệ thống

* Link to details: <https://drive.google.com/file/d/1QnbbObXo6NQ8W0ZaayDGKLcheWAYS0Vz/view?usp=sharing>

## **3.2. Xây dựng mô hình Vision Transformer**

### **3.2.1. Giới thiệu dataset**

Trong nghiên cứu này, chúng tôi sử dụng từ nhiều nguồn COCO dataset và tự thu thập thêm data để huấn luyện mô hình Vision Transformer nhận dạng sản phẩm cửa hàng tiện lợi. COCO là một bộ dữ liệu phổ biến trong lĩnh vực thị giác máy tính, cung cấp hình ảnh đa dạng với hơn 1.5 triệu đối tượng được gán nhãn và phân đoạn chi tiết với các mask pixel cho từng đối tượng.

COCO dataset bao gồm các thành phần chính sau:

* Hình ảnh: Chứa thông tin về các hình ảnh như ID, đường dẫn file, kích thước.
* Chú thích (Annotations): Chứa thông tin về các đối tượng trong hình ảnh, bao gồm bounding box, ID hình ảnh, và ID category.
* Danh mục (Categories): Chứa thông tin về các loại đối tượng, bao gồm ID và tên.
* Giấy phép (Licenses): Chứa thông tin về giấy phép sử dụng của hình ảnh.

**Hệ thống thu thập dữ liệu tự động:** Nhóm chúng tôi đã xây dựng một hệ thống tiên tiến nhằm mục đích thu thập dữ liệu nhanh chóng và hiệu quả hơn. Hệ thống này sử dụng máy ảnh để nhận diện và thu thập hình ảnh đầu vào. Khi hình ảnh được nhập vào hệ thống, nó sẽ được xử lý để sinh thêm các biến thể khác nhau, từ đó tạo ra một tập hợp dữ liệu phong phú hơn.

**Quá trình hoạt động của hệ thống bao gồm:**

1. **Nhận diện và thu thập hình ảnh:** Máy ảnh của điện thoại được sử dụng để chụp hình ảnh đầu vào, có thể là từ các nguồn khác nhau .
2. **Xử lý hình ảnh:** Sau khi hình ảnh được thu thập, hệ thống sẽ tiến hành xử lý để sinh thêm các biến thể. Điều này có thể bao gồm việc thay đổi độ sáng, độ tương phản, hoặc áp dụng các bộ lọc khác để tạo ra nhiều phiên bản khác nhau của cùng một hình ảnh.
3. **Tạo và lưu trữ dữ liệu:** Các hình ảnh biến thể được tạo ra sẽ được lưu trữ theo cấu trúc dataset đã được quy định trước đó. Cấu trúc này đảm bảo rằng dữ liệu được tổ chức một cách hệ thống và dễ dàng truy cập, phân tích trong tương lai.

Ở đây ta có 26 label tập trung vào các sản phẩm của một cửa hàng tiện lợi được phân đoạn hình ảnh và gán nhãn cụ thể cho từng sản phẩm ta có được 29689 data train và 2816 data validation và 1433 data dùng để test.

Hình 35. Hình ảnh giới thiệt dataset

### **3.2.2. Thu thập và tiền xử lý dữ liệu**

Để xử lý dữ liệu từ nhiều COCO dataset, chúng tôi đã phát triển lớp **COCOParser** với các chức năng chính dọc dữ liệu từ hình từ và label của dataset và thực hiện chuyển đổi label của nó thành label theo quy định riêng của chúng tôi thông qua **JoinDataset** class này phát triền giúp chung tôi có thể chuyển đổi các sản phẩm của các dataset về chung một label mà chung tôi quy định từ trước.

A screenshot of a computer program

Description automatically generated

Hình 37. File mapping các dataset

Bằng cách chúng tôi thêm vào các thư mục chứa dataset của COCO 1 file mapping.csv có tác dụng chuyển đổi các quy định sẳn các label cần đổi lại cho phù hợp với dataset của chúng tôi. Như thế này khi tôi có 3 bộ dataset khác nhau ở dataset đầu tiền sản phẩm Fanta có label là 4 nhưng ở dataset khác Fanta được đánh là 10 thì dựa vào file mapping tôi có thể map Fanta của các bộ dataset về thành label là 8 được tôi quy định.

Bên cạnh nhóm chúng tôi đã thu thập thêm dữ liệu từ thực tế, sử dụng máy ảnh để nhận diện và thu thập hình ảnh đầu vào. Khi hình ảnh được nhập vào hệ thống, nó sẽ được xử lý để sinh thêm các biến thể khác nhau, từ đó tạo ra một tập hợp dữ liệu phong phú hơn.

Quá trình hoạt động của hệ thống bao gồm:

* Nhận diện và thu thập hình ảnh: Máy ảnh được sử dụng để chụp hình ảnh đầu vào, có thể là từ các nguồn khác nhau như tài liệu, sản phẩm, hoặc các đối tượng cần phân tích.
* Xử lý hình ảnh: Sau khi hình ảnh được thu thập, hệ thống sẽ tiến hành xử lý để sinh thêm các biến thể. Điều này có thể bao gồm việc thay đổi độ sáng, độ tương phản, hoặc áp dụng các bộ lọc khác để tạo ra nhiều phiên bản khác nhau của cùng một hình ảnh.
* Tạo và lưu trữ dữ liệu: Các hình ảnh biến thể được tạo ra sẽ được lưu trữ theo cấu trúc dataset đã được quy định trước đó. Cấu trúc này đảm bảo rằng dữ liệu được tổ chức một cách hệ thống và dễ dàng truy cập, phân tích trong tương lai.

A diagram of a diagram

Description automatically generated with medium confidence Hình 37. Sơ đồ quy trình thu thập dữ liệu

* Hình ảnh được chuẩn hóa về kích thước cố định (được chỉ định bởi tham số size).
* Giá trị pixel được chuẩn hóa về khoảng [0, 1].
* Mỗi hình ảnh được gán nhãn là ID của category tương ứng.
* Thực hiện tăng cường dữ liệu giúp đa dạng hóa tập dữ liệu huấn luyện

Việc xử lý dữ liệu này cho phép chúng tôi chuẩn bị một bộ dữ liệu phù hợp để huấn luyện mô hình Vision Transformer, đảm bảo tính nhất quán về kích thước và định dạng của dữ liệu đầu vào.

### **3.2.3. Phân tích phân phối dữ liệu huấn luyện**

Để hiểu rõ hơn về sự phân bố của các lớp trong tập dữ liệu, chúng tôi đã tiến hành phân tích và trực quan hóa số lượng mẫu cho mỗi lớp. Kết quả được thể hiện thông qua biểu đồ cột như sau:

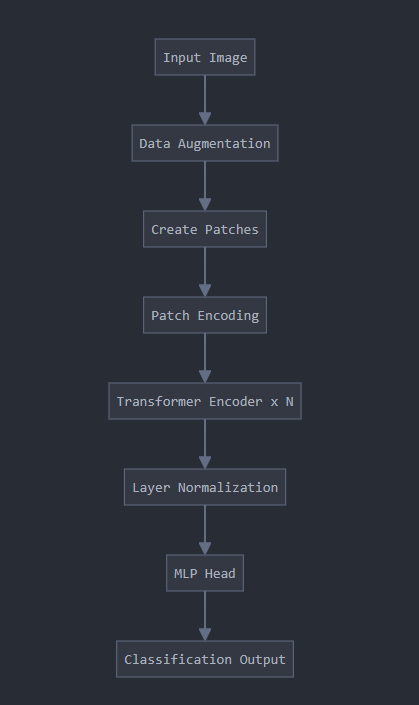
A graph of blue rectangular bars with black text

Description automatically generated

Hình 38. Sơ đồ Phân tích phân phối dữ liệu huấn luyện

### **3.2.4. Xây dựng và huấn luyện mô hình**

Trong nghiên cứu này, chúng tôi đã xây dựng một mô hình Vision Transformer (ViT) tùy chỉnh để nhận dạng sản phẩm cửa hàng tiện lợi. Dưới đây là chi tiết mô hình về cấu trúc và quá trình xây dựng:



Hình 39. Cấu trúc mô hình VIT

Mô hình được biên dịch với các cấu hình sau:

* **Optimizer**: AdamW với learning rate ban đầu là 1e-3 là một biến thể của optimizer Adam, có khả năng tự điều chỉnh learning rate và thường cho kết quả tốt với các mô hình Transformer.
* **Loss function**: SparseCategoricalCrossentropy phù hợp với bài toán phân loại nhiều lớp, đặc biệt khi labels không được one-hot encoded.
* **Metric**: Accuracy là metric trực quan để đánh giá hiệu suất của mô hình phân loại.
* **Callbacks**: Sử dụng ModelCheckpoint và EarlyStopping
  + ModelCheckpoint là một callback dùng để lưu trữ trạng thái của mô hình trong quá trình huấn luyện. Điều này rất hữu ích khi muốn lưu lại các mô hình trong suốt quá trình huấn luyện để có thể tiếp tục huấn luyện hoặc sử dụng mô hình tốt nhất sau khi hoàn tất huấn luyện.
  + EarlyStopping là một callback giúp dừng huấn luyện sớm khi mô hình không còn cải thiện được nữa. Điều này giúp ngăn ngừa hiện tượng overfitting và giảm thời gian huấn luyện không cần thiết.

Trong quá trình phát triển mô hình, chúng tôi đã thử nghiệm với nhiều cấu hình khác nhau của các tham số trên. Một số nhận xét:

* Giảm **patch\_size** giúp mô hình học được các đặc trưng chi tiết hơn, nhưng cũng làm tăng đáng kể độ phức tạp tính toán.
* Điều chỉnh **mlp\_head\_units** và **transformer\_layers** ảnh hưởng đến khả năng học các biểu diễn phức tạp của mô hình.
* Tăng dropout giúp giảm overfitting nhưng nếu quá cao có thể làm giảm khả năng học của mô hình.

Ta thu được kết quả tốt nhất khi model huấn luyện trên dataset trên với các tham số sau:

* **input\_shape** (Kích thước ảnh đầu vào): (64x64x3)
* **patch\_size (**Kích thước của mỗi patch): 4
* **num\_heads** (Số lượng heads trong Multi-Head Attention): 10
* **transformer\_layers** (Số lượng lớp Transformer Encoder ):6
* **mlp\_head\_units** (Số lượng units trong MLP Head): [1024, 512]

### **3.2.5. Kết quả thực nghiệm và đánh giá**

Sau khi hoàn tất quá trình huấn luyện mô hình Vision Transformer, chúng tôi tiến hành đánh giá mô hình dựa trên các chỉ số hiệu suất chính và biểu đồ đường học để hiểu rõ hơn về khả năng hoạt động và hiệu quả của mô hình.

A graph with different colored lines

Description automatically generated

Hình 3.9 Biểu đồ loss và accuracy của mô hình

Kết quả huấn luyện mô hình Vision Transformer cho thấy hiệu suất ấn tượng và cải thiện đáng kể qua các epoch. Trong suốt quá trình huấn luyện, mô hình đã đạt được những chỉ số quan trọng như sau:

* **Hiệu suất Huấn luyện:** Mô hình bắt đầu với độ chính xác khoảng 41.5% và mất mát 1.58 trên tập huấn luyện ở epoch đầu tiên. Qua các epoch tiếp theo, mô hình đã cải thiện đáng kể, với độ chính xác cuối cùng đạt 97.05% và mất mát giảm xuống còn 0.0693. Điều này cho thấy mô hình đã học tốt các đặc trưng của dữ liệu huấn luyện và đạt được hiệu suất cao trong việc phân loại.
* **Hiệu suất Xác thực:** Mô hình cũng cho thấy khả năng tổng quát tốt khi đánh giá trên tập xác thực. Độ chính xác trên tập xác thực bắt đầu từ 55.5% và tăng lên 96.02% ở epoch cuối cùng. Mất mát trên tập xác thực giảm từ 1.1260 xuống 0.14, chứng tỏ mô hình không chỉ học tốt từ dữ liệu huấn luyện mà còn duy trì hiệu suất cao trên dữ liệu chưa thấy.

Đánh giá tổng quát các kết quả trên cho thấy mô hình Vision Transformer có khả năng học tốt và tổng quát mạnh mẽ, với độ chính xác và sự giảm mất mát ổn định qua thời gian. Mô hình đã chứng minh được khả năng phân loại hình ảnh hiệu quả, với hiệu suất cao không chỉ trên tập huấn luyện mà còn trên tập xác thực. Điều này cho thấy mô hình có thể áp dụng thành công vào các bài toán phân loại hình ảnh thực tế.

# CHƯƠNG 4. ỨNG DỤNG VÀ TRIỂN KHAI

## **4.1 Ứng dụng trong hệ thống thanh toán tại cửa hàng tiện lợi**

### **4.1.1. Tích hợp hệ thống nhận diện sản phẩm vào quy trình thanh toán**

Việc tích hợp hệ thống nhận diện sản phẩm vào quy trình thanh toán bao gồm các bước chi tiết và phức tạp, nhằm đảm bảo hệ thống hoạt động hiệu quả và ổn định trong môi trường thực tế. Các bước này bao gồm:

1. Chuẩn bị cơ sở hạ tầng:

* **Cài đặt thiết bị**: Đầu tiên, cần cài đặt các thiết bị cần thiết như camera để chụp hình sản phẩm. Camera cần được đặt ở vị trí phù hợp để có thể chụp rõ nét từng sản phẩm khi khách hàng đưa lên quầy thanh toán. Ngoài ra, cần đảm bảo hệ thống đèn chiếu sáng đủ để hỗ trợ camera chụp hình rõ ràng trong mọi điều kiện ánh sáng.
* **Cài đặt phần mềm**: Hệ thống nhận diện sản phẩm sẽ được cài đặt trên server hoặc sử dụng dịch vụ cloud. Phần mềm này bao gồm các mô-đun nhận diện hình ảnh, xử lý dữ liệu và giao tiếp với hệ thống quản lý đơn hàng.

1. Kết nối phần cứng và phần mềm:

* **Tích hợp camera với phần mềm**: Camera được kết nối với phần mềm nhận diện để tự động chụp và gửi hình ảnh của sản phẩm. Hình ảnh sau đó được xử lý bởi mô-đun nhận diện để xác định sản phẩm.
* **Thiết lập kết nối với hệ thống quản lý đơn hàng**: Sau khi sản phẩm được nhận diện, thông tin về sản phẩm sẽ được gửi đến hệ thống quản lý đơn hàng để xử lý tiếp theo. Quá trình này cần được thực hiện nhanh chóng để không làm gián đoạn quy trình thanh toán.

1. Kết nối phần cứng và phần mềm:

* **Phát triển giao diện nhập thông tin khách hàng**: Sau khi sản phẩm được nhận diện, giao diện sẽ yêu cầu nhân viên nhập thông tin khách hàng như tên, số điện thoại, và địa chỉ. Giao diện này cần được thiết kế trực quan và dễ sử dụng để giúp nhân viên nhập liệu nhanh chóng.
* **Hiển thị thông tin sản phẩm và đơn hàng**: Sau khi thông tin khách hàng được nhập, hệ thống sẽ hiển thị thông tin về sản phẩm và tổng giá trị đơn hàng để nhân viên kiểm tra và xác nhận trước khi tiến hành thanh toán.(cần làm lại)

1. **Tích hợp chức năng thanh toán**:

* **In hóa đơn**: Sau khi thông tin sản phẩm và khách hàng được xác nhận, hệ thống sẽ in hóa đơn cho khách hàng. Hóa đơn sẽ bao gồm các thông tin chi tiết về sản phẩm, tổng giá trị đơn hàng, và thông tin khách hàng.
* **Xử lý thanh toán tại cửa hàng**: Nhân viên sẽ xử lý thanh toán tại cửa hàng theo các phương thức thanh toán như tiền mặt hoặc các phương thức khác mà cửa hàng hỗ trợ. Quá trình thanh toán sẽ diễn ra tương tự như trong các cửa hàng tiện lợi thông thường.

1. Đào tạo nhân viên:

* **Hướng dẫn sử dụng hệ thống**: Nhân viên cần được hướng dẫn chi tiết về cách sử dụng hệ thống mới, bao gồm cách chụp hình sản phẩm, nhập thông tin khách hàng, và xử lý các giao dịch thanh toán.
* **Xử lý tình huống đặc biệt**: Nhân viên cũng cần được đào tạo để xử lý các tình huống khi hệ thống nhận diện không chính xác hoặc gặp sự cố. Các tình huống này cần được xử lý nhanh chóng để không làm gián đoạn quy trình thanh toán.

### **4.1.2. Kiểm thử và triển khai hệ thống**

1. Kiểm thử chức năng:

* **Kiểm tra tính chính xác của hệ thống nhận diện**: Hệ thống cần được kiểm tra với nhiều sản phẩm khác nhau để đảm bảo tính chính xác. Các sản phẩm có thể bao gồm các mặt hàng phổ biến trong cửa hàng tiện lợi như thực phẩm, đồ uống…
* **Đảm bảo hoạt động ổn định**: Hệ thống cần được kiểm tra để đảm bảo hoạt động ổn định trong mọi điều kiện, bao gồm các tình huống có nhiều khách hàng cùng lúc hoặc điều kiện ánh sáng thay đổi.

1. Kiểm thử tích hợp:

* **Kiểm tra kết nối giữa các thành phần**: Hệ thống bao gồm nhiều thành phần khác nhau như camera, phần mềm nhận diện, và hệ thống quản lý đơn hàng. Quá trình kiểm thử cần đảm bảo các thành phần này hoạt động tốt khi tích hợp với nhau.
* **Kiểm thử quy trình từ nhận diện đến thanh toán**: Quá trình từ nhận diện sản phẩm đến thanh toán cần được kiểm tra để đảm bảo không xảy ra lỗi. Điều này bao gồm kiểm tra việc truyền dữ liệu giữa các thành phần và xử lý các giao dịch thanh toán.

1. Kiểm thử tích hợp:

* **Đo lường thời gian xử lý nhận diện sản phẩm**: Thời gian từ khi chụp hình đến khi hiển thị sản phẩm và nhập thông tin khách hàng cần được đo lường để đảm bảo quy trình diễn ra nhanh chóng. Hệ thống cần đảm bảo xử lý nhanh chóng để không làm gián đoạn quy trình thanh toán.
* **Kiểm tra khả năng chịu tải**: Hệ thống cần được kiểm tra để đảm bảo có thể xử lý nhiều giao dịch đồng thời mà không gặp sự cố. Điều này đặc biệt quan trọng trong các giờ cao điểm khi có nhiều khách hàng cùng lúc.

1. Triển khai hệ thống:

* **Cài đặt hệ thống tại các cửa hàng thí điểm**: Ban đầu, hệ thống sẽ được triển khai tại một số cửa hàng thí điểm để kiểm tra hoạt động trong môi trường thực tế. Các cửa hàng thí điểm sẽ cung cấp phản hồi về hoạt động của hệ thống và các vấn đề gặp phải.
* **Giám sát hoạt động và thu thập phản hồi**: Trong quá trình triển khai tại cửa hàng thí điểm, hệ thống sẽ được giám sát liên tục để đảm bảo hoạt động ổn định. Phản hồi từ nhân viên và khách hàng sẽ được thu thập để cải thiện hệ thống.

1. Mở rộng triển khai:

* **Mở rộng triển khai tại các cửa hàng khác**: Sau khi hệ thống thí điểm hoạt động tốt, quá trình triển khai sẽ được mở rộng tại các cửa hàng khác trong chuỗi. Quá trình này cần được thực hiện đồng bộ để đảm bảo tất cả các cửa hàng đều được trang bị hệ thống mới.
* **Đảm bảo triển khai không gây gián đoạn**: Trong quá trình mở rộng triển khai, cần đảm bảo hoạt động của các cửa hàng không bị gián đoạn. Các cửa hàng cần tiếp tục hoạt động bình thường trong khi hệ thống mới được cài đặt và kiểm thử.

## **4.2. Đánh giá hiệu quả hệ thống**

### **4.2.1. Hiệu quả về thời gian**

1. Giảm thời gian xử lý giao dịch:

* **Hệ thống nhận diện tự động**: Hệ thống nhận diện sản phẩm tự động giúp giảm thiểu thời gian cần thiết để quét mã vạch thủ công. Quá trình nhận diện diễn ra nhanh chóng và chính xác, giúp tiết kiệm thời gian cho cả khách hàng và nhân viên.
* **Tốc độ xử lý nhanh**: Thời gian xử lý từ lúc chụp hình đến khi hiển thị sản phẩm và nhập thông tin khách hàng được giảm đáng kể. Điều này giúp quy trình thanh toán diễn ra nhanh chóng, không làm khách hàng phải chờ đợi lâu.

1. Tăng tốc độ phục vụ khách hàng:

* **Giảm thời gian chờ đợi**: Quy trình thanh toán nhanh chóng giúp giảm thời gian chờ đợi của khách hàng tại quầy thanh toán. Khách hàng có thể hoàn tất mua sắm và thanh toán một cách nhanh chóng và tiện lợi.
* **Tăng hiệu suất làm việc của nhân viên**: Nhân viên có thể xử lý nhiều giao dịch hơn trong cùng một khoảng thời gian nhờ vào hệ thống tự động hóa. Điều này giúp nâng cao hiệu suất làm việc và chất lượng dịch vụ.

### **4.2.2. Hiệu quả về chi phí**

1. Giảm chi phí nhân công:

* **Tự động hóa quy trình**: Hệ thống tự động hóa giúp giảm số lượng nhân viên cần thiết tại quầy thanh toán. Nhân viên có thể tập trung vào các nhiệm vụ khác như hỗ trợ khách hàng, sắp xếp hàng hóa, và quản lý kho.
* **Tối ưu hóa sử dụng nhân lực**: Việc giảm số lượng nhân viên tại quầy thanh toán giúp tối ưu hóa sử dụng nhân lực và giảm chi phí tiền lương. Điều này đặc biệt quan trọng trong bối cảnh chi phí lao động ngày càng tăng cao.

2. Giảm chi phí sai sót:

* **Nhận diện chính xác**: Hệ thống nhận diện chính xác giúp giảm thiểu sai sót trong việc quét mã vạch và tính tiền sản phẩm. Điều này giúp giảm thiểu các tình huống trả lại hàng và xử lý các vấn đề liên quan đến thanh toán không chính xác.

### **Độ chính xác và độ tin cậy của hệ thống**

1. Độ chính xác trong nhận diện sản phẩm:

* **Hiệu suất của mô hình Vision Transformer**: Mô hình Vision Transformer được huấn luyện với bộ dữ liệu đa dạng và phong phú, cho phép nó nhận diện chính xác nhiều loại sản phẩm khác nhau. Độ chính xác của hệ thống có thể được đo lường thông qua các chỉ số như độ nhạy (sensitivity) và độ đặc hiệu (specificity).
* **Giảm thiểu sai sót**: Hệ thống nhận diện sản phẩm chính xác giúp giảm thiểu sai sót trong quá trình thanh toán. Việc nhận diện sai sản phẩm hoặc không nhận diện được sản phẩm sẽ được giảm thiểu tối đa, đảm bảo khách hàng được tính đúng giá và thông tin sản phẩm.

2. Độ tin cậy của hệ thống:

* **Khả năng hoạt động liên tục**: Hệ thống cần được thiết kế để hoạt động liên tục và ổn định trong suốt thời gian mở cửa của cửa hàng. Các biện pháp bảo trì và giám sát thường xuyên sẽ giúp phát hiện và khắc phục kịp thời các sự cố, đảm bảo hệ thống luôn sẵn sàng hoạt động.
* **Phản ứng nhanh với thay đổi**: Hệ thống cần có khả năng phản ứng nhanh với các thay đổi, chẳng hạn như việc cập nhật danh mục sản phẩm mới, thay đổi giá cả, hoặc xử lý các tình huống đặc biệt. Điều này đảm bảo hệ thống luôn hoạt động hiệu quả và đáng tin cậy trong mọi điều kiện.

3. Khả năng khắc phục lỗi:

* **Cơ chế phát hiện và sửa lỗi**: Hệ thống cần có cơ chế phát hiện và sửa lỗi tự động, đảm bảo các lỗi phát sinh trong quá trình nhận diện và thanh toán được khắc phục ngay lập tức. Điều này giúp giảm thiểu gián đoạn trong quy trình thanh toán và tăng cường độ tin cậy của hệ thống.
* **Hỗ trợ từ nhân viên kỹ thuật**: Trong trường hợp gặp sự cố phức tạp, hệ thống cần có sự hỗ trợ từ đội ngũ kỹ thuật để giải quyết nhanh chóng. Nhân viên kỹ thuật cần được đào tạo và chuẩn bị sẵn sàng để can thiệp khi cần thiết.

4. Đánh giá định kỳ:

* **Đo lường và báo cáo**: Hệ thống cần được đánh giá định kỳ thông qua các chỉ số đo lường như tỷ lệ nhận diện chính xác, thời gian xử lý, và số lượng lỗi phát sinh. Báo cáo định kỳ sẽ giúp quản lý cửa hàng và đội ngũ kỹ thuật theo dõi hiệu suất của hệ thống và đưa ra các biện pháp cải tiến kịp thời.
* **Cải tiến liên tục**: Dựa trên các báo cáo và phản hồi từ nhân viên và khách hàng, hệ thống cần được cải tiến liên tục để nâng cao độ chính xác và độ tin cậy. Các cải tiến có thể bao gồm cập nhật mô hình nhận diện, tối ưu hóa quy trình xử lý, và nâng cấp phần cứng.

Tổng hợp lại, việc đánh giá hiệu quả hệ thống không chỉ dựa trên hiệu quả về thời gian và chi phí mà còn cần xem xét độ chính xác và độ tin cậy. Hệ thống nhận diện và thanh toán sản phẩm cần đảm bảo hoạt động ổn định, chính xác, và tin cậy để đáp ứng nhu cầu của khách hàng và nâng cao hiệu suất làm việc của nhân viên.

**TÀI LIỆU THAM KHẢO**

<https://en.wikipedia.org/wiki/Vision_transformer>

<https://medium.com/machine-intelligence-and-deep-learning-lab/vit-vision-transformer-cc56c8071a20>

<https://viso.ai/deep-learning/vision-transformer-vit/>

<https://arxiv.org/pdf/2010.11929>

<https://github.com/wangshusen/DeepLearning/blob/master/Slides/10_ViT.pdf>