**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA HỆ THỐNG THÔNG TIN**

Ảnh có chứa biểu tượng, Đồ họa, hình mẫu, Phông chữ

Mô tả được tạo tự động

**mÔN DỮ LIỆU LỚN**

**SỬ DỤNG CNN ĐỂ DỰ ĐOÁN XU HƯỚNG BÁN BẤT ĐỘNG SẢN**

**Lớp:** IS405.O11.HTCL

**GVHD:** ThS.Nguyễn Hồ Duy Tri

**Nhóm:** 7

**Số thứ tự:** 14

Nguyễn Thị Thu Thủy 20520797

Nguyễn Cẩm Tú 20520837

Nguyễn Thanh Sơn 20521847

Đặng Trần Tuấn Anh 20521058

**Thành phố Hồ Chí Minh, tháng 12 năm 2023**

# NHẬN XÉT

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

# MỤC LỤC

[NHẬN XÉT 2](#_Toc154035185)

[MỤC LỤC 3](#_Toc154035186)

[PHỤ LỤC HÌNH ẢNH 4](#_Toc154035187)

[NỘI DUNG 5](#_Toc154035188)

[CHƯƠNG 1: GIỚI THIỆU 5](#_Toc154035189)

[1.1. Giới thiệu đề tài 5](#_Toc154035190)

[1.2. Mô tả bộ dữ liệu 5](#_Toc154035191)

[CHƯƠNG 2: TRỰC QUAN HÓA & TIỀN XỬ LÝ DỮ LIỆU 7](#_Toc154035192)

[2.1. Trực quan hóa dữ liệu 7](#_Toc154035193)

[2.2. Tiền xử lý dữ liệu 13](#_Toc154035194)

[CHƯƠNG 3: THUẬT TOÁN KHAI THÁC DỮ LIỆU: CNN 17](#_Toc154035195)

[3.1. Mô tả bài toán 17](#_Toc154035196)

[3.2. Định nghĩa 17](#_Toc154035197)

[3.3. Cấu trúc mạng CNN 17](#_Toc154035198)

[3.4. Triển khai 21](#_Toc154035199)

[CHƯƠNG 4: KẾT QUẢ 24](#_Toc154035200)

[4.1. Kết quả 24](#_Toc154035201)

[4.2. Đánh giá 24](#_Toc154035202)

[CHƯƠNG 5: KẾT LUẬN 25](#_Toc154035203)

[5.1. Ưu điểm 25](#_Toc154035204)

[5.2. Nhược điểm 25](#_Toc154035205)

[5.3. Hướng phát triển 25](#_Toc154035206)

[TÀI LIỆU THAM KHẢO 27](#_Toc154035207)

# PHỤ LỤC HÌNH ẢNH

[Hình 1.1. Dữ liệu gốc 6](#_Toc154035135)

[Hình 2.1 Import thư viện 7](#_Toc154035136)

[Hình 2.2 Đọc dữ liệu từ file excel 7](#_Toc154035137)

[Hình 2.3 Liệt kê các cột trong bộ dữ liệu 7](#_Toc154035138)

[Hình 2.4. Liệt kê các thuộc tính của cột trong bộ dữ liệu 7](#_Toc154035139)

[Hình 2.5 Kiểm tra số lượng phần tử null của từng cột 8](#_Toc154035140)

[Hình 2.6 Loại bỏ các phần từ null 8](#_Toc154035141)

[Hình 2.7 Chuyển đổi cột Date Recorded từ string qua date 8](#_Toc154035142)

[Hình 2.8 Mô tả bộ dữ liệu với các giá trị count, mean, stddev, min, max của từng cột 8](#_Toc154035143)

[Hình 2.9 Mô tả cột ‘Sale Amount’ với các giá trị count, mean, stddev, min, max 8](#_Toc154035144)

[Hình 2.10 Thống kê số lượng tài sản 9](#_Toc154035145)

[Hình 2.11 Đếm số lượng của từng loại nhà ở 9](#_Toc154035146)

[Hình 2.12 Thống kê và đếm số lượng từng nhà ở 10](#_Toc154035147)

[Hình 2.13 Thống kê loại tài sản và sắp xếp từ cao đến thấp 10](#_Toc154035148)

[Hình 2.14 Trực quan hóa thống kê số lượng bán của từng loại tài sản 11](#_Toc154035149)

[Hình 2.15 Thống kê loại nhà ở và sắp xếp từ cao đến thấp 11](#_Toc154035150)

[Hình 2.16 Trực quan hóa thống kê số lượng bán của từng loại tài sản 12](#_Toc154035151)

[Hình 2.17 Thống kê tỷ lệ xuất hiện của loại tài sản và nhà ở 12](#_Toc154035152)

[Hình 2.18 Chọn 3 cột cần thiết cho thuật toán 13](#_Toc154035153)

[Hình 2.19 Tổng hợp 2 đặc trưng thành vector 14](#_Toc154035154)

[Hình 2.20 Chuẩn hóa các đặc trưng bằng StandardScaler 14](#_Toc154035155)

[Hình 2.21 Chọn cột đã chuẩn hóa dữ liệu và cột để dự đoán 15](#_Toc154035156)

[Hình 2.22 Chuyển đổi giá trị dự đoán thành dạng nhị phân 15](#_Toc154035157)

[Hình 2.23 Chuyển dữ liệu thành mảng 16](#_Toc154035158)

[Hình 3.1 Kiến trúc mô hình CNN 18](#_Toc154035159)

[Hình 3.2 Công thức ReLU 18](#_Toc154035160)

[Hình 3.3 Ví dụ Convolution 19](#_Toc154035161)

[Hình 3.4 Ví dụ Max Pooling 19](#_Toc154035162)

[Hình 3.5 Fully Connected Layer 20](#_Toc154035163)

[Hình 3.6 Hàm Sigmoid 20](#_Toc154035164)

[Hình 3.7 Chia train test 21](#_Toc154035165)

[Hình 3.8 Hiển thị X\_train, y\_train, X\_test, y\_test 21](#_Toc154035166)

[Hình 3.9 Hàm ReLu 21](#_Toc154035167)

[Hình 3.10 Hàm tích chập 21](#_Toc154035168)

[Hình 3.11 Hàm max pooling 21](#_Toc154035169)

[Hình 3.12 Hàm sigmod 21](#_Toc154035170)

[Hình 3.13 Mô hình CNN 22](#_Toc154035171)

[Hình 3.14 Hàm loss 22](#_Toc154035172)

[Hình 3.15 Train model 22](#_Toc154035173)

[Hình 3.16 Huấn luyện Model 23](#_Toc154035174)

[Hình 3.17 Đánh giá mô hình 23](#_Toc154035175)

[Hình 3.18 Hàm dự đoán 23](#_Toc154035176)

[Hình 4.1 Nhập thông tin cần dự đoán 24](#_Toc154035177)

[Hình 4.2 Kết quả dự đoán 24](#_Toc154035178)

# NỘI DUNG

## GIỚI THIỆU

### Giới thiệu đề tài

Dự báo giá bất động sản là một thách thức quan trọng trong lĩnh vực kinh tế, đòi hỏi sự hiểu biết về nhiều yếu tố đa dạng. Trong nghiên cứu này, chúng ta sẽ sử dụng dữ liệu lớn từ Real Estate Sales 2001-2020 để dự báo giá bất động sản. Phương pháp được áp dụng là Convolutional Neural Network (CNN), một thuật toán học sâu thích hợp cho xử lý dữ liệu không gian như hình ảnh.

Dự án này sẽ tập trung vào việc phân tích các yếu tố như đặc trưng của tài sản, vị trí địa lý, thời gian, tình hình kinh tế, chính sách, xu hướng thị trường, nhu cầu và cung cấp để xây dựng một mô hình dự báo chính xác. CNN sẽ giúp chúng ta khám phá mối quan hệ phức tạp giữa các yếu tố này và giúp nâng cao khả năng dự báo.

Kết quả của nghiên cứu này có thể cung cấp thông tin hữu ích cho nhà đầu tư, nhà quản lý, và người tiêu dùng trong việc đưa ra quyết định về bất động sản. Tuy nhiên, cần lưu ý rằng dự báo giá bất động sản vẫn mang tính dự đoán và có thể bị ảnh hưởng bởi những biến động không dự đoán khác trong thực tế.

### Mô tả bộ dữ liệu

Nguồn gốc của dataset Real Estate Sales 2001-2020 được thu thập từ **Connecticut Department of Revenue Services**, một cơ quan chính phủ của tiểu bang Connecticut, Mỹ. Dữ liệu này bao gồm hơn 2 triệu giao dịch bất động sản tại Connecticut từ năm 2001 đến 2020, với nhiều thông tin chi tiết về giá, diện tích, số phòng, loại tài sản, vị trí, thời điểm, v.v. Dữ liệu này được cập nhật hàng năm và được công bố miễn phí trên trang web của cơ quan. Dữ liệu này được sử dụng cho nhiều mục đích như nghiên cứu, giáo dục, kinh doanh, hoặc giải trí.

Nguồn: [Real Estate Sales 2001-2020 (kaggle.com)](https://www.kaggle.com/datasets/derrekdevon/real-estate-sales-2001-2020)

Dữ liệu gồm 997213 dòng và 11 cột.

Ảnh có chứa văn bản, Phông chữ, đen và trắng, tài liệu

Mô tả được tạo tự động

Hình .. Dữ liệu gốc

**Bảng mô tả thuộc tính**

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên thuộc tính | Mô tả | Kiểu dữ liệu |
| 1 | Serial Number | Một chuỗi số duy nhất được sử dụng để xác định mỗi giao dịch. | Int |
| 2 | List Year | Năm mà tài sản cụ thể được đưa ra bán. | Int |
| 3 | Date Recorded | Ngày mà giao dịch được hoàn tất, tức là năm mà tài sản được mua. | String |
| 4 | Town | Thị trấn nơi tài sản này được đặt | String |
| 5 | Address | Địa chỉ cụ thể của tài sản. | String |
| 6 | Assessed Value | Đây là giá trị mà tài sản được đánh giá nói chung. | Int |
| 7 | Sale Amount | Là số tiền thực tế mà tài sản được bán. | Double |
| 8 | Sales Ratio | Tỉ lệ này đo lường mức độ gần gũi của giá bán của tài sản so với giá định giá của nó. | Double |
| 9 | Property Type | Mô tả loại tài sản là gì, ví dụ: nhà, căn hộ, đất đai, v.v. | String |
| 10 | Residential Type | Nếu tài sản là nhà ở, loại nhà ở là gì, ví dụ: nhà phố, chung cư, biệt thự, v.v. | String |
| 11 | Years until sold | Số năm trôi qua trước khi tài sản cuối cùng được bán. | Int |

## TRỰC QUAN HÓA & TIỀN XỬ LÝ DỮ LIỆU

### Trực quan hóa dữ liệu

Ảnh có chứa văn bản, Phông chữ, ảnh chụp màn hình

Mô tả được tạo tự động

Hình . Import thư viện

Ảnh có chứa văn bản, Phông chữ, ảnh chụp màn hình, biên lai

Mô tả được tạo tự động

Hình . Đọc dữ liệu từ file excel

Ảnh có chứa văn bản, Phông chữ, ảnh chụp màn hình

Mô tả được tạo tự động

Hình . Liệt kê các cột trong bộ dữ liệu

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động

Hình .. Liệt kê các thuộc tính của cột trong bộ dữ liệu

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, hàng

Mô tả được tạo tự động

Hình . Kiểm tra số lượng phần tử null của từng cột

Ảnh có chứa văn bản, Phông chữ, ảnh chụp màn hình, màu trắng

Mô tả được tạo tự động

Hình . Loại bỏ các phần từ null

Ảnh có chứa văn bản, Phông chữ, hàng, ảnh chụp màn hình

Mô tả được tạo tự động

Hình . Chuyển đổi cột Date Recorded từ string qua date

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, hàng

Mô tả được tạo tự động

Hình . Mô tả bộ dữ liệu với các giá trị count, mean, stddev, min, max của từng cột

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động

Hình . Mô tả cột ‘Sale Amount’ với các giá trị count, mean, stddev, min, max

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động

Hình . Thống kê số lượng tài sản

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động

Hình . Đếm số lượng của từng loại nhà ở

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

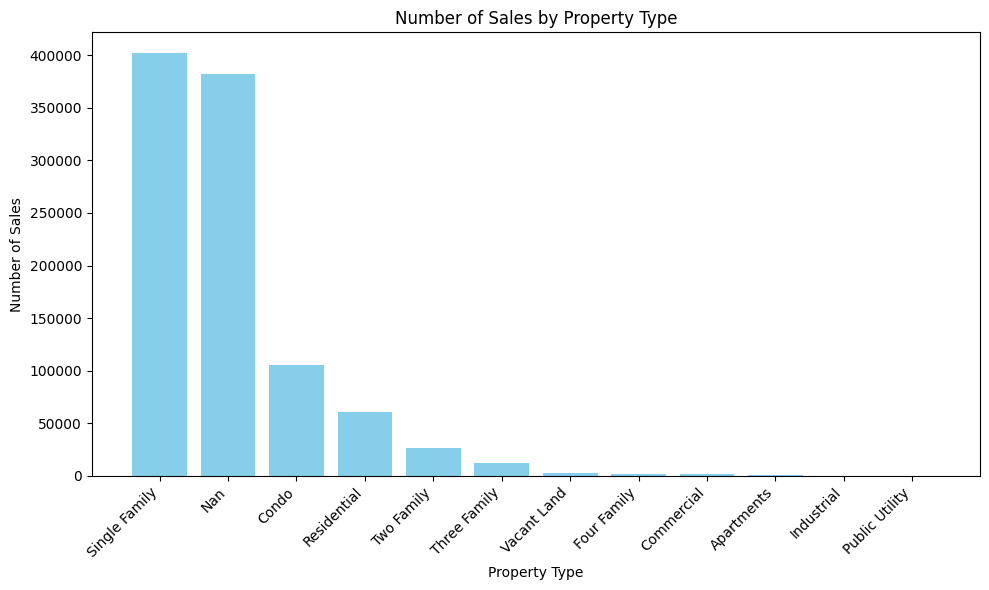
Mô tả được tạo tự động

Hình . Thống kê và đếm số lượng từng nhà ở

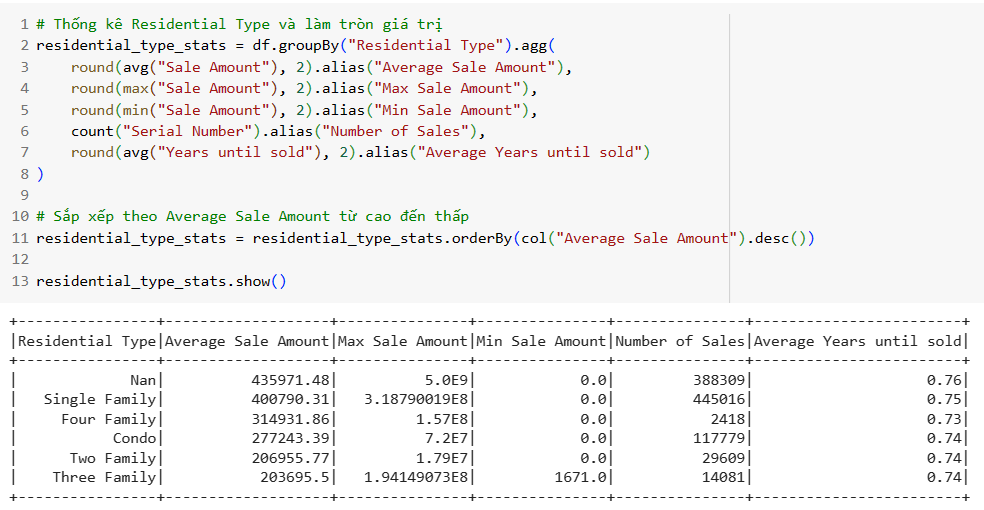
Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, thực đơn

Mô tả được tạo tự động

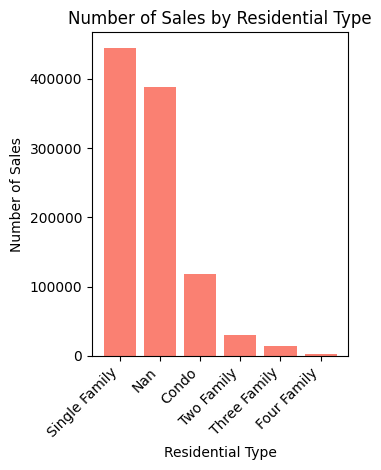
Hình . Thống kê loại tài sản và sắp xếp từ cao đến thấp



Hình . Trực quan hóa thống kê số lượng bán của từng loại tài sản



Hình . Thống kê loại nhà ở và sắp xếp từ cao đến thấp



Hình . Trực quan hóa thống kê số lượng bán của từng loại tài sản

Ảnh có chứa văn bản, Phông chữ, ảnh chụp màn hình, đen và trắng

Mô tả được tạo tự động

Hình . Thống kê tỷ lệ xuất hiện của loại tài sản và nhà ở

### Tiền xử lý dữ liệu

Bài được sử dụng 2 kĩ thuật tiền xử lý dữ liệu: StandardScaler để chuẩn hóa dữ liệu số và chuyển đổi giá trị thành nhị phân. Các kĩ thuật trên hỗ trợ các mô hình phụ thuộc vào khoảng giá trị như hồi quy truyến tính, SVM, Neural Network, CNN....

Bước 1: Lựa chọn các thuộc tính cần thiết cho mô hình CNN: Assessed Value, Sale Amout, Sales Ratio.

Ảnh có chứa văn bản, ảnh chụp màn hình, số

Mô tả được tạo tự động

Hình . Chọn 3 cột cần thiết cho thuật toán

Bước 2: Lựa chọn 2 đặc trưng để đưa vào huấn luyện mô hình: Assessed Value, Sale Amount. Sau đó, tổng hợp 2 đặc trưng thành vector.

Ảnh có chứa văn bản, ảnh chụp màn hình, số, Phông chữ

Mô tả được tạo tự động

Hình . Tổng hợp 2 đặc trưng thành vector

Bước 3: Chuẩn hóa các đặc trưng bằng hàm StandardScaler.

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động

Hình . Chuẩn hóa các đặc trưng bằng StandardScaler

Sau đó chọn những đặc trưng phục vụ cho huấn luyện mô hình là: scaled\_features, Sales Ratio.

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động

Hình . Chọn cột đã chuẩn hóa dữ liệu và cột để dự đoán

Bước 4: Chuyển đổi giá trị dự đoán (Sales Ratio) thành dạng nhị phân.

Ảnh có chứa văn bản, ảnh chụp màn hình

Mô tả được tạo tự động

Hình . Chuyển đổi giá trị dự đoán thành dạng nhị phân

Bước 5: Chọn 2 thuộc tính đưa dô mô hình là scaled\_features, label. Chuyển dữ liệu thành dạng mảng để tiến hành huấn luyện.

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Mô tả được tạo tự động

Hình . Chuyển dữ liệu thành mảng

## THUẬT TOÁN KHAI THÁC DỮ LIỆU: CNN

### Mô tả bài toán

Đồ án phân loại bất động sản sử dụng Convolutional Neural Network (CNN), tập trung vào việc đánh giá và phân loại các bất động sản dựa trên Sales Ratio. Mục tiêu chính của đồ án là xây dựng một mô hình tự động học từ dữ liệu training, giúp dự đoán bất động sản vào các nhóm tương ứng.

Để thực hiện điều này, chúng em đã tổ chức một bộ dữ liệu training chứa thông tin về Assessed Value, Sale Amount và có thể bao gồm thêm các đặc điểm như địa chỉ, thời gian, và những yếu tố khác quan trọng của bất động sản. Mô hình CNN sẽ được huấn luyện trên bộ dữ liệu này để tự động học các đặc điểm đặc trưng và quy luật từ thông tin đánh giá và giá trị bán.

Sau khi huấn luyện, mô hình sẽ có khả năng phân loại các bất động sản mới dựa trên các thông số đầu vào. Đồng thời, sẽ thực hiện quá trình đánh giá mô hình trên tập dữ liệu kiểm thử để đảm bảo tính độ chính xác và khả năng tổng quát hóa của nó.

Qua quá trình này, hy vọng rằng mô hình CNN sẽ giúp tối ưu hóa quá trình phân loại bất động sản, cung cấp một công cụ hữu ích trong lĩnh vực định giá và quản lý bất động sản.

### Định nghĩa

Thuật toán Convolutional Neural Network (CNN), một loại mạng nơ-ron tích chập, để xây dựng một mô hình học sâu có khả năng học được các đặc trưng quan trọng của dữ liệu ảnh, như vị trí, diện tích, kiến trúc, màu sắc, chất lượng, và sử dụng chúng để dự báo giá bất động sản. Đề tài này mong muốn đóng góp vào lĩnh vực kỹ thuật, công nghệ thông tin và bất động sản bằng cách kết hợp dữ liệu lớn và trí tuệ nhân tạo để giải quyết một bài toán thực tế và có giá trị.

### Cấu trúc mạng CNN

Mạng Neural Convolutional (CNN) là một loại mô hình sâu phổ biến trong lĩnh vực deep learning, thường được sử dụng để xử lý dữ liệu ảnh và video. Mô hình bao gồm các lớp chính như Lớp Tích Chập để tìm ra đặc trưng, Hàm Kích Hoạt ReLU để thêm tính phi tuyến tính, Lớp Max Pooling để giảm kích thước và tăng tính bất biến, và Lớp Fully Connected để kết hợp thông tin và đưa ra dự đoán. Cấu trúc này giúp mô hình tự động học các đặc trưng cấp cao từ dữ liệu đầu vào và làm cho nó phù hợp cho nhiều tác vụ như nhận diện hình ảnh và phân loại.

A diagram of a process

Description automatically generated

Hình . Kiến trúc mô hình CNN

(Nguồn: [CNN의 기본적인 구조 및 원리 - 1 (philipshrimp.github.io)](https://philipshrimp.github.io/posts/neural_networks/cnn_basics/) )

#### Hàm Kích Hoạt ReLU

Hàm kích hoạt ReLU (Rectified Linear Unit) là một thành phần quan trọng trong cấu trúc của mạng Neural Convolutional (CNN). Hàm này được sử dụng để giữ lại các giá trị dương từ đầu ra của lớp tích chập và loại bỏ các giá trị âm. Cơ bản, hàm ReLU được biểu diễn bằng công thức f(x)=max(0,x) , nghĩa là giữ lại giá trị dương nếu x là dương và đặt giá trị âm bằng 0.

A graph of a function

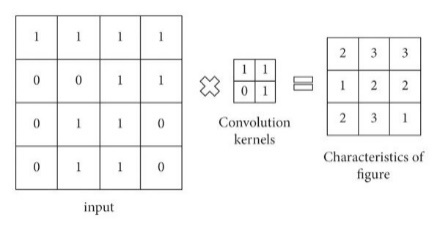
Description automatically generated

Hình . Công thức ReLU

(Nguồn: [A Comprehensive Guide to Understanding CNN Architecture in Deep Learning](https://www.ksolves.com/blog/artificial-intelligence/understanding-convolution-neural-network-architecture))

#### Lớp Tích Chập (Convolution)

Lớp Tích Chập (Convolutional Layer) trong mạng Neural Convolutional (CNN) áp dụng một bộ lọc (kernel) ngẫu nhiên lên dữ liệu đầu vào. Quá trình này giúp mô hình học các đặc trưng không gian cục bộ của ảnh. Sau đó, hàm kích hoạt ReLU được sử dụng để giữ lại các giá trị dương và loại bỏ các giá trị âm, thêm tính phi tuyến tính vào mô hình.



Hình . Ví dụ Convolution

(Nguồn: [The schematic diagram of the convolution layer. | Download Scientific Diagram](https://www.researchgate.net/figure/The-schematic-diagram-of-the-convolution-layer_fig2_359836256) )

#### Max Pooling

Max Pooling được sử dụng để giảm kích thước của đầu ra sau lớp tích chập bằng cách chọn giá trị lớn nhất từ các vùng không chồng lên nhau trong đầu ra. Điều này giúp giảm lượng thông tin và tăng tính bất biến của mô hình.

A diagram of a graph

Description automatically generated

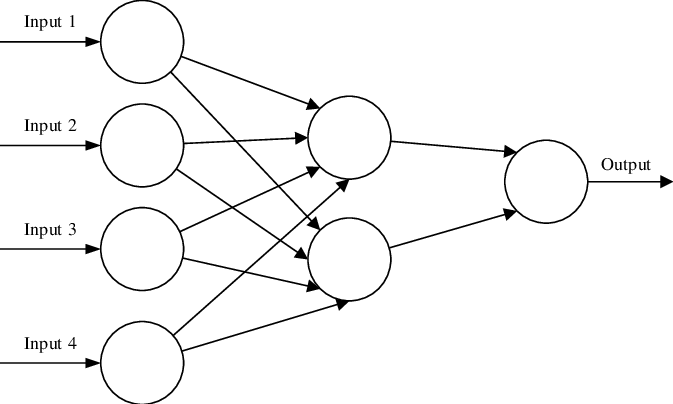
Hình . Ví dụ Max Pooling

(Nguồn: [CNN의 기본적인 구조 및 원리 - 1 (philipshrimp.github.io)](https://philipshrimp.github.io/posts/neural_networks/cnn_basics/) )

#### Fully Connected Layer

Trọng số của Fully Connected Layer được khởi tạo ngẫu nhiên. Mỗi trọng số tương ứng với một kết nối giữa đơn vị nơ-ron trong lớp trước và đơn vị nơ-ron trong lớp đang xử lý.

Các giá trị đầu ra của lớp trước đó (thường là kết quả của một lớp pooling) được nhân với trọng số để tạo ra đầu ra của lớp Fully Connected.

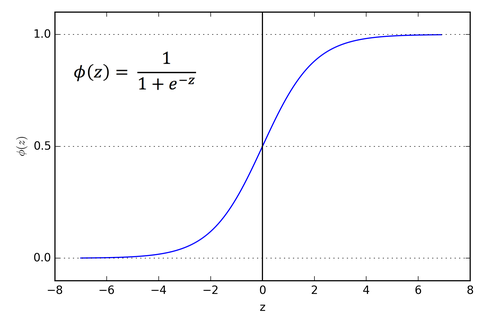


Hình . Fully Connected Layer

(Nguồn: [A simple three layered feedforward neural network (FNN), comprised of a...](https://www.researchgate.net/figure/A-simple-three-layered-feedforward-neural-network-FNN-comprised-of-a-input-layer-a_fig3_285164623))

#### Sigmoid Activation

   Hàm kích hoạt Sigmoid thường được sử dụng ở lớp đầu ra để chuyển đổi giá trị thành xác suất trong khoảng từ 0 đến 1, thích hợp cho các bài toán phân loại nhị phân.



Hình . Hàm Sigmoid

(Nguồn: [7: Sigmoid activation function. | Download Scientific Diagram (researchgate.net)](https://www.researchgate.net/figure/Sigmoid-activation-function_fig12_358233393) )

### Triển khai

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, hàng

Mô tả được tạo tự động

Hình . Chia train test

Ảnh có chứa văn bản, Phông chữ, ảnh chụp màn hình, tài liệu

Mô tả được tạo tự động

Hình . Hiển thị X\_train, y\_train, X\_test, y\_test

Ảnh có chứa văn bản, Phông chữ, ảnh chụp màn hình, màu trắng

Mô tả được tạo tự động

Hình . Hàm ReLu

Ảnh có chứa văn bản, Phông chữ, màu trắng, ảnh chụp màn hình

Mô tả được tạo tự động

Hình . Hàm tích chập

Ảnh có chứa văn bản, Phông chữ, màu trắng, thiết kế

Mô tả được tạo tự động

Hình . Hàm max pooling

Ảnh có chứa văn bản, Phông chữ, màu trắng, hàng

Mô tả được tạo tự động

Hình . Hàm sigmod

Ảnh có chứa văn bản, ảnh chụp màn hình

Mô tả được tạo tự động

Hình . Mô hình CNN

Ảnh có chứa văn bản, Phông chữ, ảnh chụp màn hình, hàng

Mô tả được tạo tự động

Hình . Hàm loss

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động

Hình . Train model

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Mô tả được tạo tự động

Hình . Huấn luyện Model

Ảnh có chứa văn bản, Phông chữ, ảnh chụp màn hình

Mô tả được tạo tự động

Hình . Đánh giá mô hình

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, tài liệu

Mô tả được tạo tự động

Hình . Hàm dự đoán

## KẾT QUẢ

### Kết quả

Ảnh có chứa văn bản, Phông chữ, hàng, ảnh chụp màn hình

Mô tả được tạo tự động

Hình . Nhập thông tin cần dự đoán

Ảnh có chứa văn bản, Phông chữ, màu trắng, biên lai

Mô tả được tạo tự động

Hình . Kết quả dự đoán

### Đánh giá

* Mô hình có độ chính xác trên tập kiểm thử là rất cao (gần 99%).
* Hàm *predict\_property\_category* sử dụng mô hình đã được huấn luyện để dự đoán một bất động sản có thể được phân loại là "Cao" hoặc "Thấp" dựa trên giá trị định giá và giá bán.

## KẾT LUẬN

### Ưu điểm

* CNN có khả năng học các đặc trưng phi tuyến từ dữ liệu.
* CNN có thể xử lý dữ liệu đa chiều, nghĩa là dữ liệu có nhiều đặc .
* CNN thường được sử dụng hiệu quả trong các bài toán phân loại nhị phân, giúp dự đoán xác suất hoặc lớp của một quan sát.
* Cấu trúc của CNN giúp tự động học và nhận diện các đặc trưng phi tuyến từ dữ liệu, chia sẻ trọng số để giảm độ phức tạp, áp dụng cấu trúc lặp lại và tạo hierarchy, từ đó tối ưu hóa quá trình dự đoán và phân loại trong lĩnh vực bất động sản.

### Nhược điểm

* Thuật toán CNN được thiết kế để xử lý dữ liệu ảnh, do đó dữ liệu dự đoán khi đưa vào phải ở dạng 2D (ma trận 2 chiều).
* CNN gần như không phù hợp với dữ liệu thay đổi theo thời gian. Vì mô hình không thể nắm bắt được thứ tự trong dữ liệu.
* Chuyển đổi dữ liệu về dạng nhị phân có thể dẫn đến mất mát thông tin so với việc giữ nguyên dữ liệu gốc. Dữ liệu nhị phân có thể không phân biệt được sự thay đổi nhỏ giữa các giá trị, điều này có thể làm mất đi một số đặc trưng quan trọng.
* Các mô hình CNN khá phức tạp và khó điều chỉnh. Các trọng số được học từ dữ liệu có thể khó diễn giải, làm tăng khả năng hiểu và giải thích mô hình.
* Mặc dù đã đưa bài toán vào PySpark để xử lý dữ liệu lớn, CNN vẫn có thể gây quá tải tài nguyên, đặc biệt là trong quá trình huấn luyện.

### Hướng phát triển

* Tích Hợp Dữ Liệu Đa Dạng:

Có thể tích hợp dữ liệu từ nhiều nguồn khác nhau như văn bản, số học, và địa lý, hãy xem xét cách tích hợp tất cả các loại dữ liệu này vào mô hình. Sử dụng lớp kết hợp để học các đặc trưng từ tất cả các nguồn thông tin.

* Kết Hợp CNN với Mô Hình Khác:

Xem xét thử nghiệm việc kết hợp CNN với các mô hình khác như mạng nơ-ron, mô hình dự đoán chuỗi thời gian, hoặc các mô hình phức tạp hơn. Điều này có thể tăng cường khả năng học đặc trưng của mô hình.

* Tuỳ chỉnh Mô hình Cho Nhu Cầu Cụ Thể:

Tùy chỉnh mô hình để đáp ứng nhu cầu cụ thể của bài toán đánh giá bất động sản, ví dụ như điều chỉnh ngưỡng quyết định, thay đổi hàm mất mát, hoặc sử dụng mô hình tương tự để tối ưu hóa kết quả.

# TÀI LIỆU THAM KHẢO

[1]“REAL ESTATE SALES ANALYSIS.” Accessed: Dec. 19, 2023. [Online]. Available: <https://kaggle.com/code/akankshagarg1105/real-estate-sales-analysis>

[2]“7: Sigmoid activation function. | Download Scientific Diagram.” Accessed: Dec. 20, 2023. [Online]. Available: <https://www.researchgate.net/figure/Sigmoid-activation-function_fig12_358233393>

[3]“A Comprehensive Guide to Understanding CNN Architecture in Deep Learning.” Accessed: Dec. 20, 2023. [Online]. Available: <https://www.ksolves.com/blog/artificial-intelligence/understanding-convolution-neural-network-architecture>

[4]“A simple three layered feedforward neural network (FNN), comprised of a... | Download Scientific Diagram.” Accessed: Dec. 20, 2023. [Online]. Available: <https://www.researchgate.net/figure/A-simple-three-layered-feedforward-neural-network-FNN-comprised-of-a-input-layer-a_fig3_285164623>

[5]“CNN의 기본적인 구조 및 원리 - 1.” Accessed: Dec. 20, 2023. [Online]. Available: <https://Philipshrimp.github.io/posts/neural_networks/cnn_basics/>

[6]“The schematic diagram of the convolution layer.,” ResearchGate. Accessed: Dec. 20, 2023. [Online]. Available: <https://www.researchgate.net/figure/The-schematic-diagram-of-the-convolution-layer_fig2_359836256>