TRƯỜNG ĐẠI HỌC XÂY DỰNG HÀ NỘI

KHOA CÔNG NGHỆ THÔNG TIN

-----🙞🙞&🙜🙜-----

Icon

Description automatically generated

***Đề tài:***

**TIỂU LUẬN:**

**XÂY DỰNG MÔ HÌNH CNN ĐỂ XÁC ĐỊNH LOẠI QUẢ**

Nhóm sinh viên thực hiện: Nhóm 2

Thành viên:

1. Lý Hải Yến MSSV: 1557164

2. Nguyễn Minh Tùng MSSV: 218464

3. Nguyễn Thị Toàn MSSV: 1549964

HÀ NỘI 1-2022

**Nội Dung**

[**Lời cảm ơn** 2](#_Toc95255894)

[**Lời mở đầu** 4](#_Toc95255895)

[**1.** **Mạng nơ ron tích chập (CNN)** 5](#_Toc95255896)

[***1.1.*** ***Lớp tích chập*** *(****Convolutional layers)*** 5](#_Toc95255897)

[***1.2.*** ***Lớp ReLU*** 6](#_Toc95255898)

[***1.3.*** ***Lớp gộp (pooling layers)*** 6](#_Toc95255899)

[***1.4.*** ***Lớp kết nối đầy đủ (Fully connected layers)*** 6](#_Toc95255900)

[**2.** **Tập dữ liệu sử dụng: Fruits-360 data set** 6](#_Toc95255901)

[**3.** **Cấu trúc của mạng nơ-ron được sử dụng.** 9](#_Toc95255902)

[**4.** **Kết quả thực nghiệm** 10](#_Toc95255903)

[**Tài liệu tham khảo** 11](#_Toc95255904)

# **Lời cảm ơn**

Lời đầu tiên, nhóm xin gửi lời cảm ơn chân thành đến thầy Nguyễn Đình Quý. Trong quá trình học tập và tìm hiểu bộ môn Học máy, chúng em đã nhận được sự giúp đỡ, hướng dẫn rất tận tình, tâm huyết của thầy. Nhờ những kiến thức mà thầy truyền tải, chúng em càng hiểu thêm về các ứng dụng của thuật toán học máy trong cuộc sống. Thông qua bài tiểu luận này, chúng em xin trình bày những gì mà nhóm đã tìm hiểu về “xây dựng mô hình CNN để xác định loại quả” gửi đến thầy.

Vì kiến thức còn hạn chế, trong quá trình hoàn thành tiểu luận chúng em không tránh khỏi những thiếu sót, kính mong nhận được những ý kiến đóng góp từ thầy để bài tiểu luận của nhóm được hoàn thiện hơn.

Chúng em xin chân thành cảm ơn thầy!

# **Lời mở đầu**

Mạng neural nhân tạo hay gọi ngắn gọn là mạng neural (tiếng anh là Artifical Neural Network) đang là vấn đề khá nổi trong cuộc cách mạng công nghệ 4.0. Các nhà nghiên cứu, các nhà khoa học, tổ chức đã và đang ứng dụng khá nhiều công nghệ liên quan đến mạng neural vào đời sống hằng ngày. Đơn giản là việc phân loại, nhận dạng các đối tượng trên camera. Chúng góp phần không nhỏ cho việc quản lý, nhận dạng các đối tượng nhằm giảm thiểu sự khó khăn trong công việc của con người.

Ở nước ta, nền nông nghiệp được chú trọng phát triển. Mỗi đợt thu hoạch nông sản, người nông dân phải làm thao tác phân loại nông sản bằng các biện pháp thủ công nên rất mất thời gian, công sức. Đây là công việc đơn giản nhưng lại ảnh hưởng rất lớn đối với quy mô phát triển nông nghiệp.

Hiện nay nhờ sự phát triển mạnh mẽ của công nghệ trong lĩnh vực học máy. Việc sử dụng các mạng neural nhân tạo trong các máy móc phục vụ nông nghiệp khá là hay, luôn hấp dẫn các nghiên cứu sinh nghiên cứu về nó. Do đó, nhờ sự tìm hiểu và sự giúp đỡ từ thầy hướng dẫn, em chọn đề tài: “Xây dựng mô hình CNN để xác định loại quả”.

Với tập dữ liệu tập, chất lượng cao về hình ảnh có chứa trái cây có trên Kaggle. Nhóm sẽ học tập, xây dựng và trình bày kết quả của một số thí nghiệm đào tạo mạng nơ-ron nhận diện trái cây.

1. **Mạng nơ ron tích chập (CNN)**

Mạng nơ-ron tích chập là một phần của mô hình học sâu. Một mạng như vậy có thể bao gồm các lớp tích chập, lớp gộp, lớp ReLU, lớp kết nối đầy đủ. Trong một kiến ​​trúc CNN điển hình, mỗi lớp chập được theo sau bởi một lớp ReLU, sau đó là lớp gộp rồi một hoặc nhiều lớp chập và cuối cùng là một hoặc nhiều lớp được kết nối đầy đủ. Một đặc điểm giúp CNN khác biệt với mạng nơ-ron thông thường là tính đến cấu trúc của hình ảnh trong khi xử lý chúng. Lưu ý rằng mạng nơ-ron thông thường chuyển đổi đầu vào trong mảng một chiều, điều này làm cho bộ phân loại được đào tạo ít nhạy cảm hơn với các thay đổi vị trí.

* 1. ***Lớp tích chập*** *(Convolutional layers)*

Lớp tích chập được đặt tên theo phép toán tích chập. Trong toán học, phép tích chập là một phép toán trên 2 hàm số, kết quả cho ra 1 hàm số thứ 3 là phiên bản sửa đổi (tích chập) của một trong các hàm ban đầu. Một lớp chập bao gồm các nhóm tế bào thần kinh tạo nên ma trận lọc (**Filter, Kernel** hay **Feature Detector** đều là cách gọi của ma trận lọc). Các nơ-ron từ ma trận lọc được kết nối với một vùng nhỏ của đầu vào, được gọi là trường tiếp nhận (receptive field), bởi vì việc liên kết tất cả các nơ-ron với tất cả các đầu ra trước đó rất kém hiệu quả trong trường hợp đầu vào có kích thước cao như hình ảnh. Ví dụ: một hình ảnh 100 x 100 có 10000 pixel và nếu lớp đầu tiên có 100 tế bào thần kinh, nó sẽ dẫn đến 1000000 tham số. Thay vì mỗi nơ-ron có trọng số cho kích thước đầy đủ của đầu vào, một nơ-ron giữ trọng số cho kích thước của đầu vào hạt ma trận lọc. Các hạt ma trận lọc trượt qua chiều rộng và chiều cao của đầu vào, áp dụng phép tích vô hướng để tính toán, tạo ra bản đồ kích hoạt 2 chiều (**Convolved Feature**, **Activation Map** hay **Feature Map** là đầu ra của ảnh khi cho bộ lọc chạy hết bức ảnh với phép tích vô hướng). Bước trượt mà tại đó hạt ma trận lọc trượt được đưa ra dưới dạng một tham số. Đầu ra của một lớp tích chập được thực hiện bằng cách xếp chồng các bản đồ kích hoạt kết quả, đến lượt nó được sử dụng để xác định đầu vào của lớp tiếp theo.

Áp dụng một lớp chập trên một hình ảnh có kích thước 32 X 32 dẫn đến một bản đồ kích hoạt có kích thước 28 X 28. Nếu chúng ta áp dụng nhiều lớp tích chập hơn, kích thước sẽ giảm hơn nữa và kết quả là kích thước hình ảnh bị giảm đáng kể dẫn đến việc mất thông tin và vấn đề đạo hàm biến mất. Để sửa lỗi này, chúng tôi sử dụng padding. Padding làm tăng kích thước của dữ liệu đầu vào bằng cách điền vào các hằng số xung quanh dữ liệu đầu vào. Trong hầu hết các trường hợp, hằng số này bằng 0 nên thao tác được đặt tên là không đệm. Đệm tương tự (same padding) có nghĩa là bản đồ đối tượng đầu ra có cùng kích thước không gian với bản đồ đối tượng đầu vào. Thao tác này cố gắng chèn đều bên trái và bên phải, nhưng nếu số lượng cột được thêm vào là số lẻ, nó sẽ thêm một cột bổ sung vào bên phải. Đệm hợp lệ tương đương với không có phần đệm.

Các bước trượt khiến ma trận lọc bỏ qua các pixel trong hình ảnh và không đưa chúng vào đầu ra. Khi ma trận lọc đang trượt đầu vào, nó sử dụng độ dài dải trượt để xác định khoảng cách dịch chuyển của ma trận lọc sau mỗi lần tính.

* 1. ***Lớp ReLU***

Lớp ReLU áp dụng chức năng kích hoạt max (0, x), có tác dụng đưa các giá trị âm về 0. Nó không làm giảm kích thước của ảnh và không có thêm bất kì tham số nào. Mục đích của lớp ReLU là đưa một mức ngưỡng, ở đây là 0. Để loại bỏ các giá trị âm không cần thiết mà có thể ảnh hưởng cho việc tính toán các layer sau đó.

* 1. ***Lớp gộp (pooling layers)***

Các lớp gộp được sử dụng một mặt để giảm kích thước không gian của biểu diễn và giảm độ phức tạp tính toán trong mô hình. Công dụng khác của các lớp gộp là để kiểm soát hiện tượng overfitting. Thông thường, Pool layer có nhiều hình thức khác nhau phù hợp cho nhiều bài toán, tuy nhiên Max Pooling là được sử dụng nhiều vào phổ biến hơn cả với ý tưởng cũng rất sát với thực tế con người đó là: **Giữ lại chi tiết quan trọng** hay hiểu ở trong bài toán này chính giữ lại **pixel có giá trị lớn nhất**. Thông thường max pooling có kích thước là 2 và stride=2. Nếu lấy giá trị quá lớn, thay vì giảm tính toán nó lại làm phá vỡ cấu trúc ảnh và mất mát thông tin nghiêm trọng. Vì vậy mà một số chuyên gia không thích sử dụng layer này mà thay vào đó sử dụng thêm các lớp Conv Layer và tăng số stride lên mỗi lần.

* 1. ***Lớp kết nối đầy đủ (****Fully connected layers)*

Tại lớp mạng này, mỗi một nơ-ron của layer này sẽ liên kết tới mọi nơ-ron của lớp khác. Để đưa ảnh từ các layer trước vào mạng này, buộc phải dàn phẳng bức ảnh ra thành 1 vector thay vì là mảng nhiều chiều như trước. Tại layer cuối cùng sẽ sử dụng 1 hàm kinh điển trong học máy mà bất kì ai cũng từng sử dụng đó là softmax để phân loại đối tượng dựa vào vector đặc trưng đã được tính toán của các lớp trước đó.

1. **Tập dữ liệu sử dụng: Fruits-360 data set**

Hình ảnh được ghi lại bằng cách quay phim các loại trái cây trong khi chúng được quay bằng 1 động cơ, sau đó lấy ra các khung hình. Trái cây được gắn trên trục có tốc độ quay thấp (3 vòng/ phút) và đoạn phim dài 20 giây được ghi lại.

Ảnh có chứa táo, trong nhà

Mô tả được tạo tự động

Phía bên trái là hình ảnh gốc và bên phải và hình ảnh sau khi chuyển về nền trắng và đưa về kích thước 100 x 100 pixels.

Do sự thay đổi trong điều kiện ánh sáng, nền không đồng nhất nên các bức ảnh đã được áp dụng qua một thuật toán chuyên dụng để trích xuất trái cây từ nền. Thuật toán này thuộc loại flood fill type: bắt đầu từ mỗi cạnh của hình ảnh và đánh dấu tất cả các pixel ở đó, sau đó đánh dấu tất cả pixel được tìm thấy trong vùng lân cận của pixel đã được đánh dấu mà khoảng cách giữa các màu nhỏ hơn một giá trị quy định. Lặp lại bước trước cho đến khi không còn pixel nào nữa có thể được đánh dấu. Tất cả các pixel được đánh dấu coi là nền (được tô màu trắng) và phần còn lại được coi là của sự vật. Giá trị lớn nhất cho khoảng cách giữa 2 pixel lân cận là một tham số của thuật toán được tìm ra bằng cách thử sai cho mỗi phim.

Tập dữ liệu được đặt tên là Fruits-360 và có thể được tải trên trang kaggle. Hiện tại (tính đến 2020.05.18) bộ này có 90483 hình ảnh của 131 loại trái cây và rau củ và nó được cập nhật liên tục với hình ảnh của các loại trái cây và rau củ mới.

1. **Cấu trúc của mạng nơ-ron được sử dụng.**

Sử dụng một mạng nơ-ron tích chập (CNN) để phân loại quả. Một đặc điểm tạo nên sự khác biệt giữa CNN và mạng nơ-ron thông thường là sẽ tính đến cấu trúc của ảnh trong khi xử lý chúng. Một mạng nơ-ron thông thường sẽ chuyển đổi đầu vào thành mảng một chiều làm cho bộ phận phân loại được đào tạo ít nhạy cảm hơn với các thay đổi vị trí.

Đầu vào sẽ là các hình ảnh RGB tiêu chuẩn có kích thước 100x100 pixels. Cấu trúc của mạng sử dụng sẽ có dạng như sau:

|  |  |  |
| --- | --- | --- |
| Layer type | Dimensions | Output |
| Convolutional | 5x5x4 | 16 |
| Max pooling | 2x2 S:2 | - |
| Convolutional | 5x5x16 | 32 |
| Max pooling | 2x2 S:2 | - |
| Convolutional | 5x5x32 | 64 |
| Max pooling | 2x2 S:2 | - |
| Convolutional | 5x5x64 | 128 |
| Max pooling | 2x2 S:2 | - |
| Fully connected | 5x5x128 | 1024 |
| Fully connected | 1024 | 256 |
| Softmax | 256 | 131 |

* Lớp đầu tiên (Convolution # 1) là một lớp tích hợp áp dụng 16 bộ lọc 5 x 5. Trên lớp này, chúng tôi áp dụng max pooling với bộ lọc hình dạng 2 x 2 với độ trượt 2 chỉ định rằng các vùng được gộp không chồng chéo (Max-Pool # 1). Điều này cũng làm giảm chiều rộng và chiều cao xuống 50 pixel.
* Lớp tích chập thứ hai (Convolution # 2) áp dụng 32 bộ lọc 5 x 5 bộ lọc xuất ra 32 activation maps. Chúng tôi áp dụng trên lớp này cùng một loại max pooling (Max-Pool # 2) như trên lớp đầu tiên, hình dạng 2 x 2 và độ trượt 2. Điều này cũng làm giảm chiều rộng và chiều cao xuống 25 pixel.
* Lớp tích chập thứ ba (Convolution # 3) áp dụng 64 bộ lọc 5 x 5 cùng với đó là max pooling với bộ lọc hình dạng 2 x 2 với độ trượt 2 (Max-Pool # 3). Điều này cũng làm giảm chiều rộng và chiều cao xuống 12 pixel.
* Lớp tích chập thứ tư(Convolution # 4) áp dụng 128 bộ lọc 5 x 5 cùng với đó là max pooling với bộ lọc hình dạng 2 x 2 với độ trượt 2 (Max-Pool # 4). Điều này cũng làm giảm chiều rộng và chiều cao xuống 6 pixel.
* Lớp thứ năm là lớp fully connected (Fully Connected #1) có 1024 đầu ra.
* Tiếp đến một lớp fully connected khác (Fully Connected #2) với 1024 đầu vào và 256 đầu ra.
* Lớp cuối cùng sẽ là softmax loss (Softmax) với 256 đầu vào, đầu ra là số lượng của các class(131)

1. **Kết quả thực nghiệm**

Áp dụng các thay đổi màu sắc và độ bão hòa ngẫu nhiên trên các hình ảnh RGB đầu vào, lật ngẫu nhiên chúng theo chiều ngang và chiều dọc, sau đó chuyển đổi chúng than không gian màu HSV và thang độ xám và hợp nhất chúng. Sử dụng mạng nơ-ron được mô tả trước đó đã được huấn luyện trong 25 epoch với batch size = 50 hình ảnh được chọn ngẫu nhiên từ tập huấn luyện. Mỗi kỷ nguyên, chúng tôi đều tính toán độ chính xác bằng cách sử dụng cross-validation. Để thử nghiệm, chúng tôi đã chạy mạng được đào tạo trên bộ thử nghiệm.

|  |  |  |
| --- | --- | --- |
|  | Accuracy ontraining set | Accuracy ontest set |
| HSV **+** Grayscale **+** hue**/**saturationchange **+** flips | 99.9% | 95.1% |

Để xác định cấu hình mạng tốt nhất để phân loại hình ảnh trong tập dữ liệu, lấy nhiều cấu hình, sử dụng tập hợp đào tạo để đào tạo chúng và sau đó tính toán độ chính xác của chúng trong tập kiểm tra và huấn luyện

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cấu Hình | | | Độ chính xác trên tập train | Độ chính xác trên tập test |
| Convolutional | 5 x 5 | 16 | 99.9% | 95.4% |
| Convolutional | 5 x 5 | 32 |
| Convolutional | 5 x 5 | 64 |
| Convolutional | 5 x 5 | 128 |
| Fully connected | - | 1024 |
| Fully connected | - | 512 |

Graphical user interface

Description automatically generated with low confidence

Biểu đồ số liệu độ chính xác của tập train và val trong quá trình đào tạo

Graphical user interface

Description automatically generated with low confidence

Biểu đồ số liệu độ hàm mất mát của tập train và val trong quá trình đào tạo

# **Tài liệu tham khảo**

1. Bài giảng môn Học máy – Nguyễn Đình Quý.
2. Fruit recognition from images using deep learning - Horea Muresan, Mihai Oltean