

Một trình tự CNN để phân loại các chữ số viết tay

Mạng nơron tích chập (còn gọi là ConvNet / CNN) là một thuật toán Deep Learning có thể lấy hình ảnh đầu vào, gán độ quan trọng (các trọng số - weights và độ lệch - bias có thể học được) cho các đặc trưng/đối tượng khác nhau trong hình ảnh và có thể phân biệt được từng đặc trưng/đối tượng này với nhau. Công việc tiền xử lý được yêu cầu cho mạng nơron tích chập thì ít hơn nhiều so với các thuật toán phân loại khác. Trong các phương thức sơ khai, các bộ lọc được thiết kế bằng tay (hand - engineered), với một quá trình huấn luyện để chọn ra các bộ lọc/đặc trưng phù hợp thì mạng nơron tích chập lại có khả năng tự học để chọn ra các bộ lọc/ đặc trưng tối ưu nhất.

Kiến trúc của nơron tích chập tương tự như mô hình kết nối của các nơron trong bộ não con người và được lấy cảm hứng từ hệ thống vỏ thị giác trong bộ não (visual cortex). Các nơ-ron riêng lẻ chỉ phản ứng với các kích thích trong một khu vực hạn chế của trường thị giác được gọi là Trường tiếp nhận (Receptive Field). Một tập hợp các trường như vậy chồng lên nhau để bao phủ toàn bộ khu vực thị giác.

Tại sao ConvNet mà không phải là mạng nơron đa lớp (MLP)?

|  |  |
| --- | --- |
| Làm phẳng (flattening) một ma trận 3x3 thành một véc tơ 9x1 | Một hình ảnh không gì khác là một ma trận các giá trị pixel. Vậy tại sao chúng ta không làm phẳng hình ảnh (ví dụ: ma trận hình ảnh  3×3 thành một vectơ 9×1) và đưa nó vào một mạng nơron đa lớp cho mục đích phân loại? Không thực sự đơn giản như vậy...  Trong trường hợp của hình ảnh nhị phân cơ bản (binary images), phương pháp nơron đa lớp có thể cho ra độ chính xác trung bình khi phân loại nhưng sẽ không còn chính xác khi phân tích các hình ảnh phức tạp - nơi mà các pixel có một mức độ phụ thuộc lẫn nhau. |

Phương pháp nơron tích chập có năng lực để hiểu được các mức độ phụ thuộc không gian và tạm thời giữa các pixel trong dữ liệu ảnh thông qua các bộ lọc ứng dụng đặc trưng. Kiến trúc này có hiệu suất tốt hơn cho tập dữ liệu dạng hình ảnh do làm giảm được số lượng tham số liên quan và khả năng tái sử dụng (reusability) các trọng số. Nói cách khác, mô hình mạng nơron tích chập có thể được huấn luyện để hiểu được sự tinh tế của hình ảnh tốt hơn các mô hình khác.

|  |  |
| --- | --- |
|  | Ở hình bên, chúng ta có một hình ảnh RGB được phân tách theo 3 mặt phẳng (kênh) màu - Đỏ, Xanh lục và Xanh dương. Có một số không gian màu khác cho hình ảnh như Grayscale, RGB, HSV, CMYK, v.v.  Bạn có thể thấy chi phí tính toán sẽ lớn như thế nào khi hình ảnh đạt đến kích thước, giả sử 8K (7680 pixel × 4320 pixel). Vai trò của nơron tích chập là giảm chiều hình ảnh thành một dạng dễ xử lý hơn, mà không đánh mất đi các đặc trưng quan trọng của hình ảnh để có được một dự đoán tốt (good prediction). Điều này rất quan trọng khi chúng ta thiết kế một kiến ​​trúc không chỉ tốt về việc học tập các đặc trưng mà còn có thể tương thích với các bộ dữ liệu có kích thước lớn. |

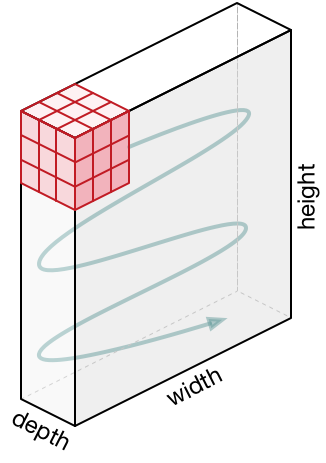
|  |  |
| --- | --- |
|  | Kích thước hình ảnh = 5(Chiều cao) x 5 (Chiều rộng) x 1 (Số lượng kênh, ví dụ: RGB).  Ở hình bên, phần màu xanh lục là hình ảnh đầu vào 5×5×1 của chúng ta, ta gọi hình ảnh đầu vào này là *I*  Phần tử liên quan đến việc thực hiện thao tác tích chập trong phần đầu tiên của lớp tích chập được gọi là Bộ lọc (Kernel / Filter) , K, được thể hiện bằng màu vàng. Chúng ta chọn  K là một ma trận 3×3×1. |

Kernel/Filter,

K = 1 0 1

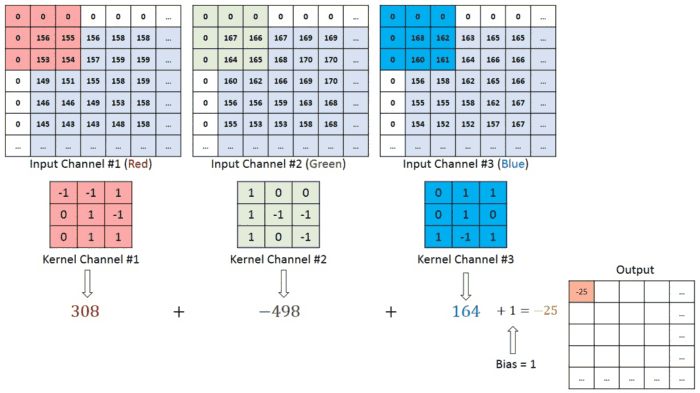
0 1 0

1 0 1



Cách di chuyển của bộ lọc

Bộ lọc di chuyển sang phải với Giá trị trượt cố định cho đến khi hoàn thành việc quét theo chiều rộng. Tiếp tục, nó nhảy xuống phía đầu bên trái của hình ảnh với cùng Giá trị trượt và lặp lại quá trình cho đến khi toàn bộ hình ảnh được duyệt qua.



Phép tính tích chập của một hình ảnh ma trận MxNx3 với một bộ lọc 3x3x3

Trong trường hợp hình ảnh có nhiều kênh (ví dụ: RGB), Bộ lọc (Kernel) có cùng độ sâu với hình ảnh đầu vào. Phép nhân ma trận được thực hiện giữa Kn và In

([K1,I1];[K2,I2];[K3,I3]) và tất cả các kết quả được cộng với độ lệch để cung cấp cho chúng ta một kênh một chiều Đầu ra đặc trưng tích chập (Convoluted Feature Output).

|  |  |
| --- | --- |
| Một phép tính chập với độ dài trượt bằng 2 | Mục tiêu của phép tính tích chập là trích xuất các đặc trưng cấp cao như các cạnh (edges), từ hình ảnh đầu vào. Mạng nơron tích chập không nhất thiết chỉ giới hạn trong một lớp tích chập. Thông thường, lớp tích chập đầu tiên chịu trách nhiệm nắm bắt các đặc trưng cấp thấp như màu sắc (colors), hướng dốc (gradient orientation), v.v. Với các lớp tích chập được thêm vào, mô hình cũng nắm bắt các đặc trưng cấp cao, mang đến cho chúng ta một mạng lưới nơron tích chập có sự hiểu biết toàn diện về hình ảnh trong bộ dữ liệu, tương tự như cách chúng ta - con người hiểu về hình ảnh. |

Có hai loại kết quả cho phép tính chập - một loại trong đó kết quả tích chập bị giảm chiều so với đầu vào và loại khác trong đó chiều của kết quả tích chập được tăng hoặc giữ nguyên. Điều này được thực hiện bằng cách áp dụng phép Đệm hợp lệ (Valid Padding) trong trường hợp trước, hoặc phép Đệm tương tự (Same Padding) trong trường hợp sau.

|  |  |
| --- | --- |
|  | Khi chúng ta tăng (đệm) hình ảnh có kích thước 5×5×1 thành hình ảnh 6×6×1 và sau đó áp dụng bộ lọc 3×3×1 lên nó, chúng ta thấy rằng ma trận tích chập đầu ra có kích thước 5×5×1. Đó là phép Đệm tương tự (same padding).  Mặt khác, nếu chúng ta thực hiện cùng một thao tác mà không cần đệm, chúng ta sẽ có một ma trận tích chập đầu ra có kích thước 3×3×1bằng chính bộ lọc - Đó là Đệm hợp lệ (valid padding). |

**Lớp gộp (Pooling layer)**

|  |  |
| --- | --- |
| Ma trận gộp 3x3 từ một ma trận tích chập 5x5 | Tương tự như lớp tích chập (Convolutional Layer), lớp gộp (pooling) chịu trách nhiệm để là giảm chiều kết quả tích chập (Convolved Feature). Điều này nhằm mục đích để giảm chi phí tính toán cần phải có để xử lý dữ liệu thông qua việc giảm kích thước tính năng đầu vào. Hơn nữa, nó rất hữu ích để trích xuất các đặc trưng cốt lõi, cái thường bất biến trước các phép xoay và phép trượt, do đó làm cho quá trình huấn luyện mô hình hiệu quả hơn. |

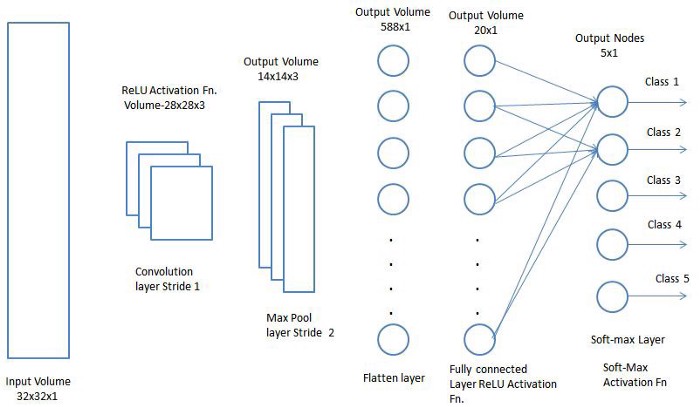
Có hai loại phép gộp: Gộp cực đại (Max Pooling) và Gộp trung bình (Average Pooling). Phép gộp cực đại trả về giá trị lớn nhất từ phần hình ảnh được bao phủ bởi bộ lọc. Trong khi đó, phép gộp trung bình trả về giá trị trung bình của tất cả các giá trị từ phần hình ảnh được bao phủ bởi bộ lọc.

Phép gộp cực đại cũng hoạt động như một công cụ khử nhiễu. Nó loại bỏ các nguồn nhiễu và thực hiện khử nhiễu song song với giảm kích thước. Mặt khác, phép gộp trung bình chỉ đơn giản thực hiện giảm kích thước như một cơ chế khử nhiễu. Do đó, chúng ta có thể nói rằng phép gộp cực đại hoạt động tốt hơn rất nhiều so với phép gộp trung bình.

|  |  |
| --- | --- |
| Các kiểu phép gộp | Lớp tích chập (Convolutional Layer) và lớp gộp (Pooling layer), kết hợp với nhau tạo thành lớp thứ i của mạng nơron tích chập. Tùy thuộc vào độ phức tạp của ảnh, số lượng các lớp như vậy có thể được tăng lên để có thể bắt được các đặc trưng ở mức độ chi tiết hơn nữa, nhưng chi phí cho sức mạnh tính toán cũng sẽ nhiều hơn. |

Sau khi đi qua quá trình trên, chúng ta đã thiết lập thành công mô hình nơron tích chập để học các đặc trưng. Tiếp theo, chúng ta sẽ làm phẳng (flattening) đầu ra cuối và đưa nó vào mạng nơron thông thường cho mục đích phân loại.

**Phân loại – Lớp kết nối đầy đủ (Fully connected layer)**



Sử dụng mạng nơron kết nối đầy đủ là cách làm phổ biến nhất để học các tổ hợp phi tuyến từ các đặc trưng trích xuất từ kết quả ma trận tích chập đầu ra. Mạng nơron kết nối đầy đủ có thể học được các đặc trưng trong không gian phi tuyến này.

Chúng ta đã chuyển đổi hình ảnh đầu vào thành một dạng thích hợp cho mạng nơron đa lớp. Chúng ta sẽ làm phẳng hình ảnh đầu vào này thành một vectơ cột. Vectơ đầu ra đã được làm phẳng sẽ được đưa vào một mạng nơron suy luận tiến (feedforward) và phương pháp truyền ngược (backpropagation) được áp dụng cho quá trình huấn luyện .Qua một loạt lần lặp, mô hình có thể phân biệt giữa các đặc trưng cốt lõi và các đặc trưng không quan trọng trong hình ảnh và phân loại chúng bằng kỹ thuật Phân loại Softmax (softmax classification).

<https://thanhvie.com/tim-hieu-ve-mang-no-ron-tich-chap-convolutional-neural-networks/>