Xử lý ảnh số

Tên đề tài:

Nhận dạng chữ viết

Giáo viên hướng dẫn:
Lã Thế Vinh

Sinh viên thực hiện:
Nguyễn Đức Tuấn

Tháng 2 - 2013

Mục lục

Bài toán nhận dạng chữ viết

- Bài toán: từ một ảnh chụp chứa một đoạn text nào đó, máy tính phải nhận ra được nội dung đó
- Bài toán nhận dạng được phát triển từ nhiều năm nay:
  - Từ 1954, máy nhận dạng chữ viết được phát triển bởi J. Rainbow để đọc chữ in hoa
  - 1967, IBM thương mại hóa hệ thống nhận dạng chữ viết
  - Từ 1990 – nay, kỹ thuật nhận dạng phát triển cùng với việc phát triển dựa trên phương pháp luận của lĩnh vực học máy
- Ứng dụng
  - Google docs
  - Các phần mềm chuyển đổi văn bản scan sang văn bản text : Abbyy finereader, tesseract-ocr, VNDocr, vietocr,...
  - Các phần mềm nhập liệu dữ liệu bằng tay trên thiếtt bị cảm ứng,..

Hệ thống nhận dạng chữ viết

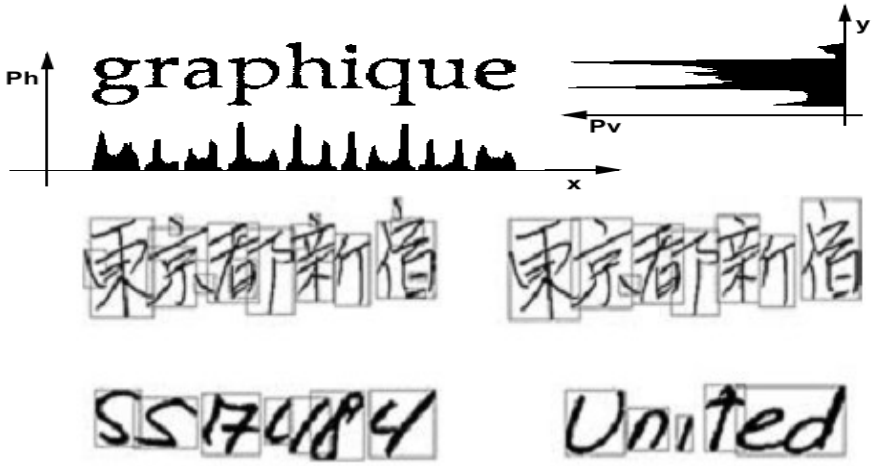
- Một hệ thống nhận dạng chữ viết thường trải qua các bước sau:
  1. Tiền xử lý ảnh
  2. Phân tách ký tự
  3. Trích xuất đặc trưng
  4. Hậu xử lý

Tiền xử lý ảnh

- Tập hợp các kỹ thuật nhằm tăng cường chất lượng ảnh cho quá trình nhận dạng
  - Lọc bỏ nhiễu
  - Nhị phân hóa ảnh
  - Khử độ nghiêng chữ

Hệ thống nhận dạng chữ viết

- Phân tách ký tự
  - Mục đích: trích chọn các ký tự từ hình ảnh đã được tăng cường
  - Một số kỹ thuật
    - Phân tách ký tự dựa trên phép chiếu (projection analysis)
    - Phân tách ký tự dựa trên phân tích thành phần kết nối (connected component analysis)



Hệ thống nhận dạng chữ viết

- Trích chọn đặc trưng
  - Mục đích: trích xuất các đặc trưng cho việc huấn luyện và nhận dạng
  - Các loại đặc trưng:
    - Đặc trưng thống kê: các đặc trưng dựa trên việc thống kê như histogram, giao điểm và khoảng cách,...
    - Đặc trưng cấu trúc: Đặc trưng dựa trên hình thái hình học của chữ viết như điểm giao, vòng lặp,..
    - Đặc trưng toàn cục và biến đổi chuỗi: các đặc trưng hình thành dựa trên phép biến đổi ảnh như Fourier,...

Hệ thống nhận dạng chữ viết

- Phân loại:
  - Sử dụng các đặc trưng đã được trích chọn, áp dụng các kỹ thuật phân loại và huấn luyện để nhận dạng chữ
  - Một số kỹ thuật phân loại:
    - K-nearest neighbor
    - Mạng neuron
    - SVM
    - Markov ẩn

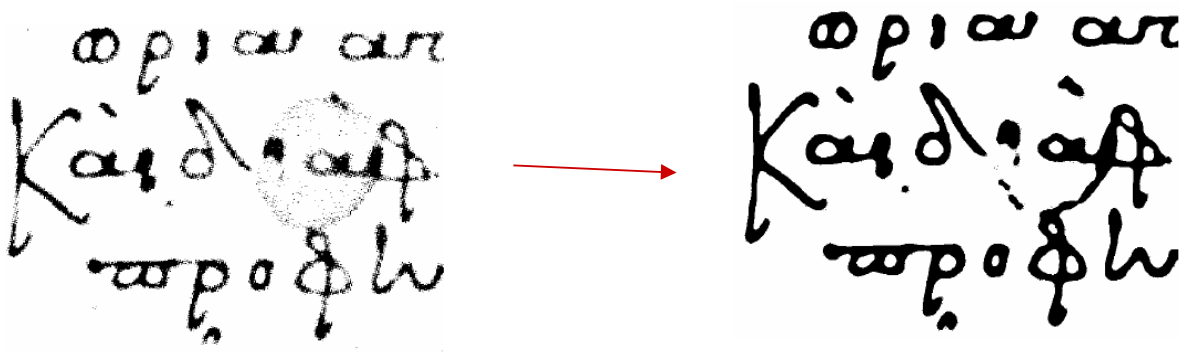
Một số kỹ thuật tiền xử lý

- Thông thường các ảnh đem vào nhận dạng thường có chất lượng không tốt như bị nhiễu, bị nghiêng,...

→ Phải thực hiện tiền xử lý để tăng cường ảnh

- Một số kỹ thuật tiền xử lý
  - Lọc bỏ nhiễu
  - Nhị phân hóa ảnh
  - Phát hiện và loại bỏ độ nghiêng

Một số kỹ thuật tiền xử lý



- Lọc bỏ nhiễu:
  - Hạn chế ảnh hưởng của nhiễu đến kết quả nhận dạng
  - Sử dụng bộ lọc thông thấp để làm mịm ảnh

Một số kỹ thuật tiền xử lý

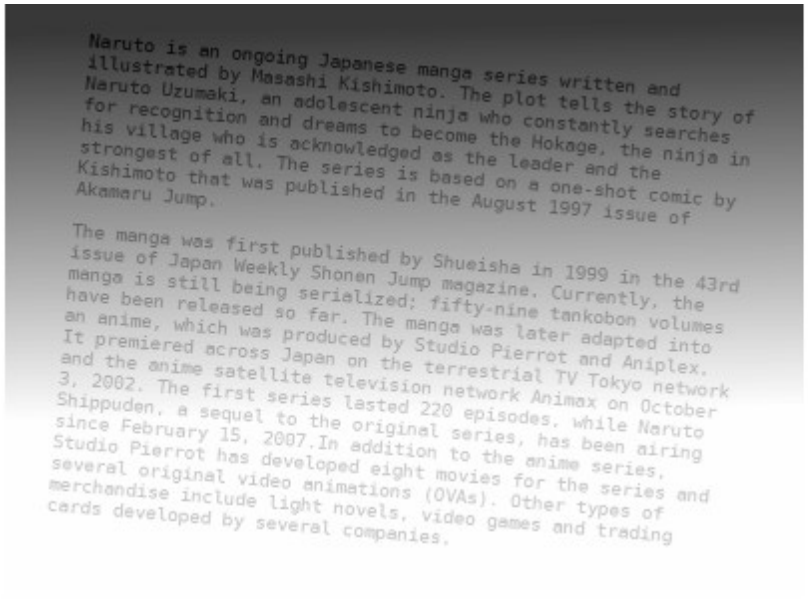
- Nhị phân hóa ảnh
  - Ảnh đầu vào có thể là ảnh màu hoặc ảnh đa mức xám

    → Cần chuyển thành ảnh nhị phân:
    - Bit 1: pixel thuộc chữ
    - Bit 0: pixel thuộc nền
  - Chuyển đổi từ ảnh màu sang ảnh đa mức xám:
    - gray(x,y) = R(x,y) * 0.2 + G(x,y) * 0.72 + 0.08 * B(x,y)
    - Trong đó:
      - gray(x,y): giá trị độ xám tương ứng tại (x,y) trong ảnh đa mức xám
      - R(x,y), G(x,y), B(x,y) tương ứng là giá trị màu sắc trong ảnh RGB
  - Chuyển đổi ảnh đa mức xám sang ảnh nhị phân
    - Tách ngưỡng toàn cục (global gray level threshold)
    - Tách ngưỡng cục bộ (local gray level threshold)
    - Tách ngưỡng sử dụng đặc trưng cụ bộ (loal feature threshold)

Một số kỹ thuật tiền xử lý

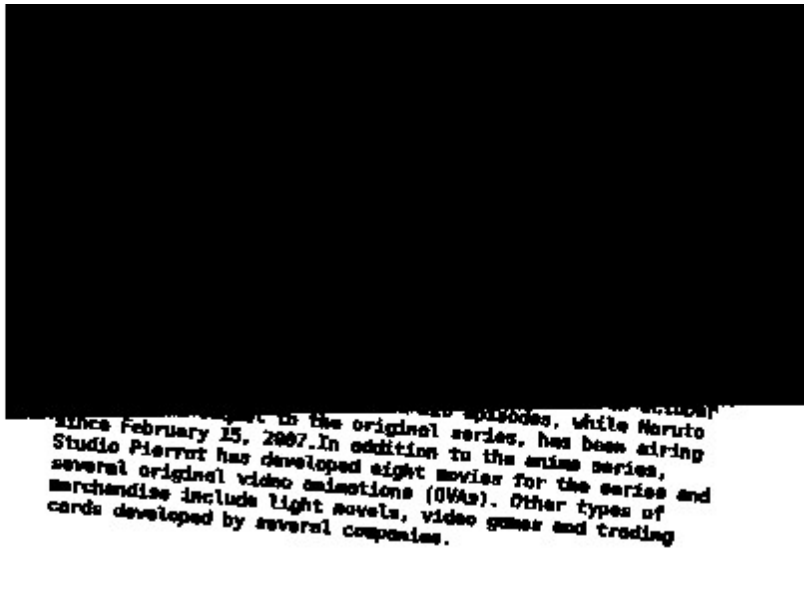
- Tách ngưỡng toàn cục
  - Sử dụng một ngưỡng toàn cục duy nhất
    - Xét một ngưỡng L:
      - Gray < L → pixel thuộc chữ
      - Gray >= L → pixel thuộc nền
  - Chọn ngưỡng:
    - Chọn bằng tay hoặc sử dụng giả thuật Otsu
  - Hạn chế: do chỉ sử dụng với một ngưỡng duy nhất nên với ảnh biến đổi mạnh khó thu được kết quả tốt

Một số kỹ thuật tiền xử lý ảnh



Naruto is an ongoing Japanese manga series written and illustrated by Masashi Kishimoto. The plot tells the story of Naruto Uzumaki, an adolescent ninja who constantly searches for recognition and dreams to become the Hokage, the ninja in his village who is acknowledged as the leader and the strongest of all. The series is based on a one-shot comic by Kishimoto that was published in the August 1997 issue of Akamaru Jump.

The manga was first published by Shueisha in 1999 in the 43rd issue of Japan Weekly Shonen Jump magazine. Currently, the manga is still being serialized; fifty-nine tankobon volumes have been released so far. The manga was later adapted into an anime, which was produced by Studio Pierrot and Aniplex. It premiered across Japan on the terrestrial TV Tokyo network and the anime satellite television network Animax on October 3, 2002. The first series lasted 220 episodes, while Naruto Shippuden, a sequel to the original series, has been airing since February 15, 2007.In addition to the anime series, Studio Pierrot has developed eight movies for the series and several original video animations (OVAs). Other types of merchandise include light novels, video games and trading cards developed by several companies.



since February 15, 2007.In addition to the anime series, Studio Pierrot has developed eight movies for the series and several original video animations (OVAs). Other types of merchandise include light novels, video games and trading cards developed by several companies.

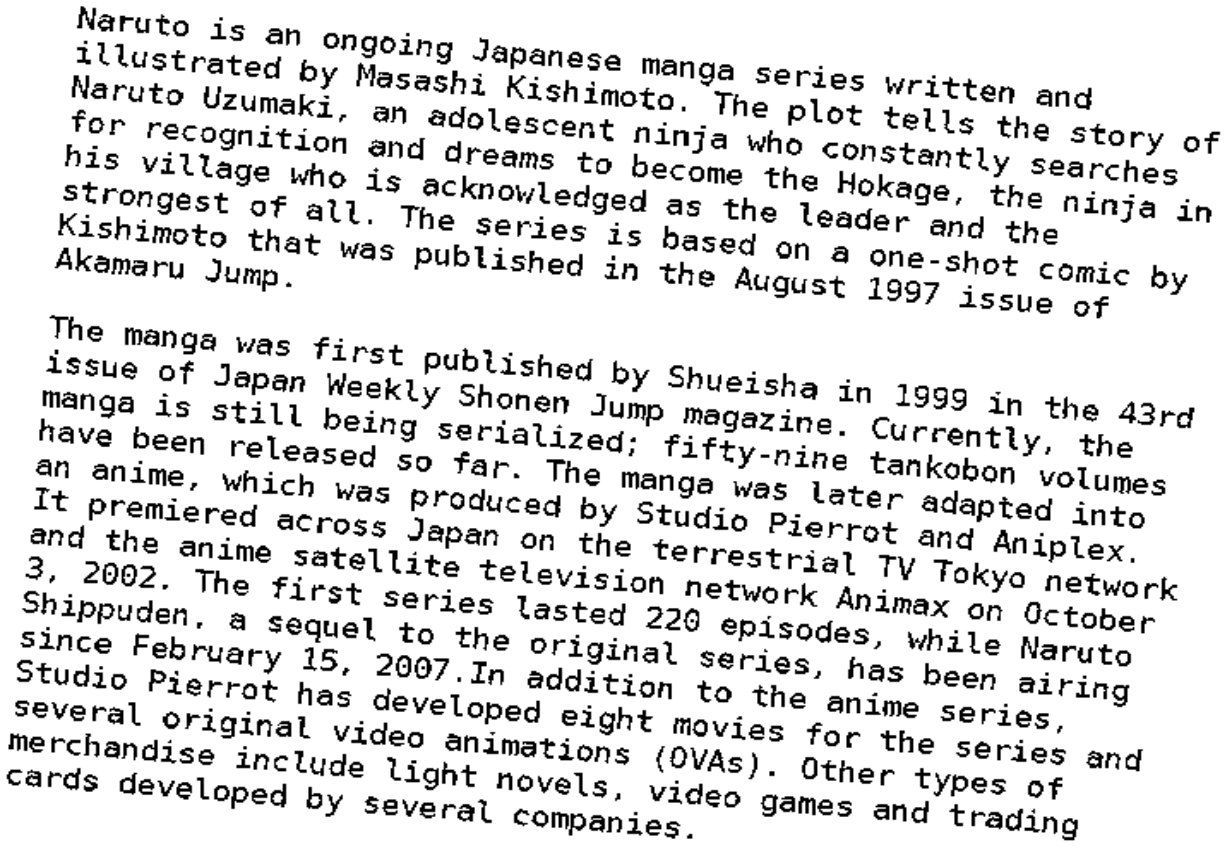
Một số kỹ thuật tiền xử lý

- Nhị phân sử dụng ngưỡng cục bộ
  - Lần lượt xét trên từng vùng cục bộ (5x5,..)
  - Việc xếp pixel thuộc chữ hay nền ngoài phụ thuộc cả vào vùng cục bộ và ngưỡng toàn cục
  - Ví dụ: Giải thuật Bersen
    - Xác định ngưỡng toàn cục global_threshold = 128, constrast = 15
    - Trên mỗi vùng cục bộ, tính các giá trị
      - mid_gray = (max_gray + min_gray)/2
      - local_constrast = max_gray – min_gray
    - Nhị phân hóa ảnh

```
If (local_constrast < constrast)
     Pixel = mid_gray < global_threshold ? Object:
background
Else
     Pixel = pixel < mid_gray: object : background
```

Một số kỹ thuật tiền xử lý

- Nhị phân hóa bằng tách ngưỡng cục bộ

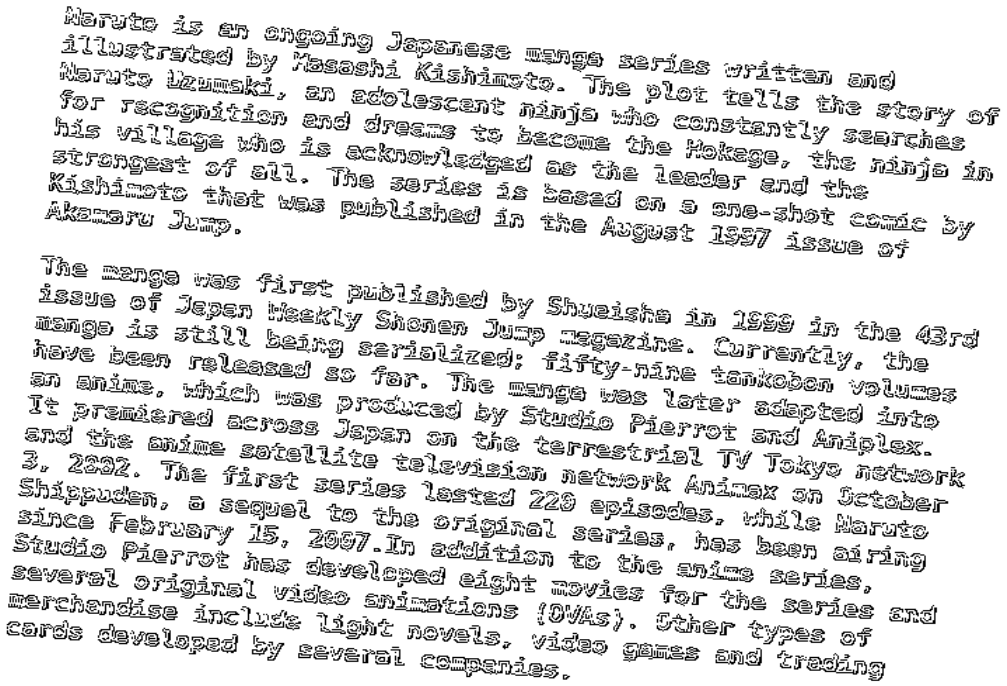


Naruto is an ongoing Japanese manga series written and illustrated by Masashi Kishimoto. The plot tells the story of Naruto Uzumaki, an adolescent ninja who constantly searches for recognition and dreams to become the Hokage, the ninja in his village who is acknowledged as the leader and the strongest of all. The series is based on a one-shot comic by Kishimoto that was published in the August 1997 issue of Akamaru Jump.

The manga was first published by Shueisha in 1999 in the 43rd issue of Japan Weekly Shonen Jump magazine. Currently, the manga is still being serialized; fifty-nine tankobon volumes have been released so far. The manga was later adapted into an anime, which was produced by Studio Pierrot and Aniplex. It premiered across Japan on the terrestrial TV Tokyo network and the anime satellite television network Animax on October 3, 2002. The first series lasted 220 episodes, while Naruto Shippuden, a sequel to the original series, has been airing since February 15, 2007.In addition to the anime series, Studio Pierrot has developed eight movies for the series and several original video animations (OVAs). Other types of merchandise include light novels, video games and trading cards developed by several companies.

Một số kỹ thuật tiền xử lý

- Nhị phân sử dụng đặc trưng cục bộ
  - Sử dụng các đặc trưng cục bộ để nhị phân hóa ảnh
  - Ví dụ: Sử dụng các bộ phát hiện đường biên

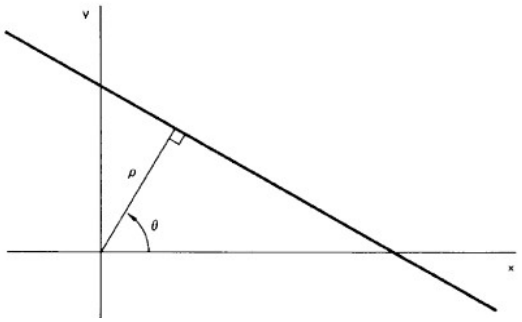


Naruto is an ongoing Japanese manga series written and illustrated by Masashi Kishimoto. The plot tells the story of Naruto Uzumaki, an adolescent ninja who constantly searches for recognition and dreams to become the Hokage, the ninja in his village who is acknowledged as the leader and the strongest of all. The series is based on a one-shot comic by Kishimoto that was published in the August 1997 issue of Akamaru Jump.

The manga was first published by Shueisha in 1999 in the 43rd issue of Japan Weekly Shonen Jump magazine. Currently, the manga is still being serialized; fifty-nine tankobon volumes have been released so far. The manga was later adapted into an anime, which was produced by Studio Pierrot and Aniplex. It premiered across Japan on the terrestrial TV Tokyo network and the anime satellite television network Animax on October 3, 2002. The first series lasted 220 episodes, while Naruto Shippuden, a sequel to the original series, has been airing since February 15, 2007.In addition to the anime series, Studio Pierrot has developed eight movies for the series and several original video animations (OVAs). Other types of merchandise include light novels, video games and trading cards developed by several companies.

Một số kỹ thuật tiền xử lý

- Phát hiện và loại bỏ độ nghiêng
  - Phát hiện và loại bỏ hiện tượng nghiêng chữ do việc chụp, quét ảnh
  - Phát hiện góc nghiêng: Sử dụng biến đổi Hough
    - Coi các dòng là tập hợp các đường thẳng song song
    - Đường thẳng Hough x cos(θ)+ y sin(θ) = r
    - Góc nghiêng cần tìm là θ
    - Tìm cặp (θ, r) mà đi qua nhiều điểm nhất
  - Loại bỏ góc nghiêng: sử dụng phép quay ma trận



$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

```
//canvas
this._super = ocrObj;
this.canvas = ocrObj.canvas;
this.context = ocrObj.context;
this.privCanvas = ocrObj.privC
this.privContext = ocrObj.priv

var index;
```

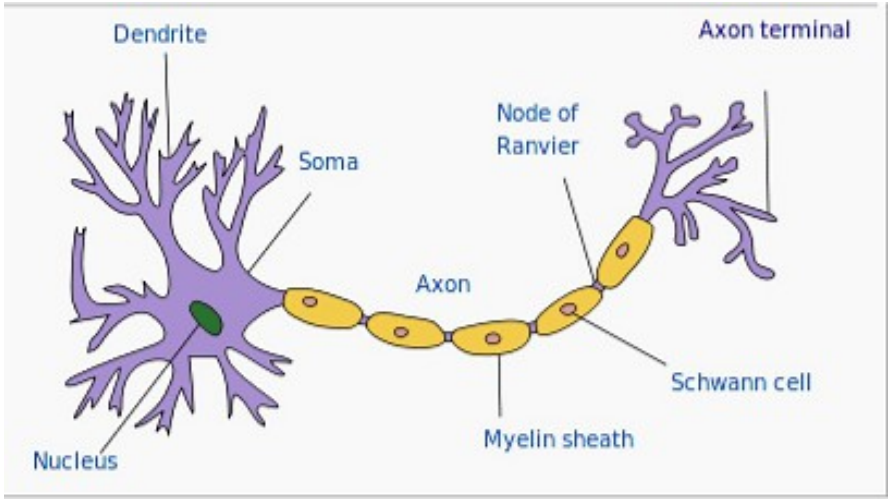
```
//canvas
this._super = ocrObj;
this.canvas = ocrObj.canvas;
this.context = ocrObj.context
this.privCanvas = ocrObj.priv
this.privContext = ocrObj.pri

var index;
```

Mạng Neuron

- Mạng Neuron
- Mạng Kohonen

Mạng Neuron sinh học

- Neuron (tế bào thần kinh) là một loại tế bào có thể xử lý và truyền thông tin qua các tín hiệu điện / hoá học.
- Soma: thành phần trung tâm của neuron, có chứa nhân tế bào.
- Dendrites: các sợi nhánh, nhận các tín hiệu đầu vào của neuron.
- Axon: trục chính, mang các tín hiệu ra từ soma.
- Điểm kết thúc của axon, chứa synapses.



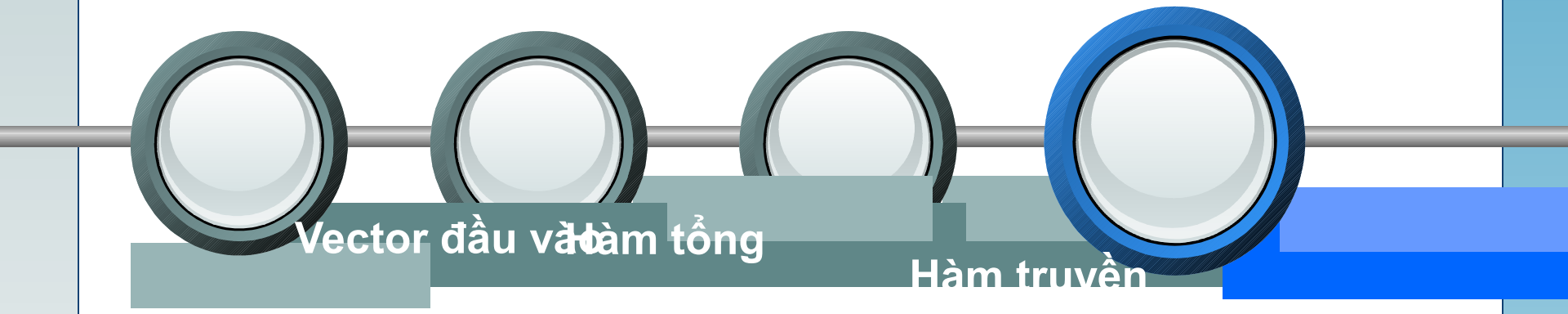
Nhận dạng chữ viết dùng mạng neuron

4.1.1 Mạng Neuron sinh học

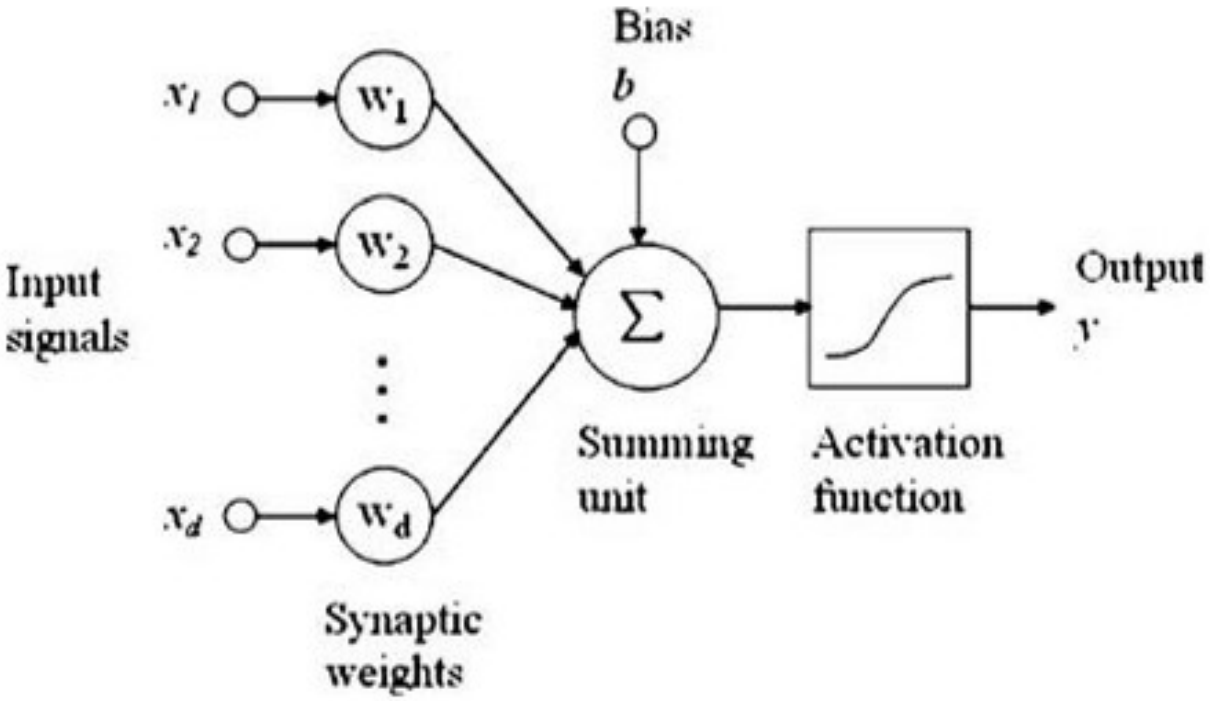
- Mạng neuron là một dãy các neuron liên kết với nhau. Trong đó axon của neuron này liên kết vớidendrite của neuron khác.
- Nếu tín hiệu đầu vào của một neuron đạt giá trị xác định, nó sẽ gửi các tín hiệu tương ứng đến các neuron kết nối với nó thông qua axon.
- Dựa vào cấu trúc của mạng neuron sinh học, người ta đã xây dựng nên mô hình mạng neuron nhân tạo để có thể đạt được sức mạnh tính toán tương tụ.

4.1.2 Mạng Neuron nhân tạo

Trong ANN, neuron (còn được gọi là một unit hay node) là sự mô hình hoá của neuron sinh học. Bao gồm:



4.1.2 Mạng Neuron nhân tạo

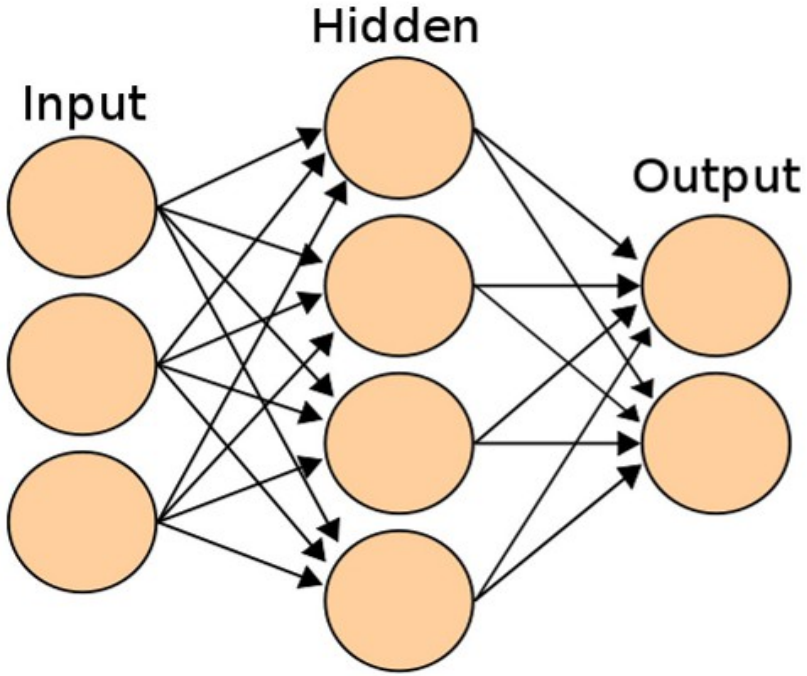


Mô hình neuron nhân tạo

4.1.3 Mạng neuron nhân tạo

Mạng neuron nhân tạo là sự kết nối giữa các neuron nhân tạo với nhau.

Mạng MLP (Multilayer Perception - Mạng truyền thẳng nhiều lớp) là một loại mạng neuron có 3 lớp:



4.1.4 Học trong mạng neuron

- Học là quá trình thay đổi hành vi của các vật theo một cách nào đó làm cho chúng có thể thực hiện tốt hơn trong tương lai.
- Với mạng neuron, cho một tập các vector đầu vào X, tập các vector đầu ra tương ứng là Y. Học là việc tìm ra một ánh xạ f : X → Y dựa vào các vector đầu vào/đầu ra đó.
  - Tập X được gọi là tập dữ liệu học hay tập huấn luyện (training set), mỗi phần tử x thuộc X được gọi là một mẫu huấn luyện (training sample).
  - Ánh xạ f được xác định bởi các trọng số của các liên kết trong mạng, chính vì vậy bản chất của việc học chính là điều chỉnh giá trị các trọng số. Trong quá trình học, các trọng số này sẽ dần hội tụ đến giá trị để mạng đạt trạng thái sao cho với mọi $x \in X$ thì f (x) = y (y là đầu ra mong muốn của x).

4.1.4 Học trong mạng neuron

Có 3 phương pháp học:

- Học có giám sát: mẫu huấn luyện sẽ có dạng (x,y) với x là đầu vào, y là đầu ra tương ứng. Mục tiêu là tìm ánh xạ f khớp với các ví dụ.
- Học không giám sát: mẫu huấn luyện chỉ gồm đầu vào x. Nhiệm vụ của học không giám sát là phân chia tập huấn luyện thành các nhóm con, mỗi nhóm chứa các vector đầu vào có đặc trưng giống nhau.
- Học tăng cường (học thưởng phạt): Với học tăng cường, dữ liệu học thường không được cho trước mà được tạo ra trong quá trình hoạt động.

Ánh xạ tự tổ chức

- Ánh xạ tự tổ chức (SOM) là một loại ANN sử dụng phương pháp học không giám sát. SOM khác với các ANN khác ở chỗ nó sử dụng một hàm lân cận (neighborhood function) để bảo toàn trật tự sắp xếp của không gian dữ liệu đầu vào.
- SOM được giới thiệt lần đầu tiên bởi Teuvo Kohonen, vì vậy đôi khi được gọi là mạng Kohonen.

Ánh xạ tự tổ chức

- Tự tổ chức (self-organization):
  - Đầu ra "đúng" không được cung cấp do đó quá trình cập nhật các trọng số là tự động.
  - Với mạng Kohonen, một vector đầu vào được cung cấp mỗi bước. Tất cả các vector đầu vào này tạo nên "môi trường" cho mạng neuron.
  - Mỗi một lần nhận đầu vào là một lần thay đổi mạng để thích nghi với dữ liệu mới. Đến một lúc, mạng neuron có thể trở thành đại diện cho môi trường (tập các vector đầu vào).

Ánh xạ tự tổ chức

- Bảo toàn trật tự sắp xếp (topology-preserving):
  - Các mẫu dữ liệu thuộc không gian nhiều chiều khi ánh xạ thành các mảng neuron với số chiều nhỏ hơn (thường là 2 chiều) được bảo toàn thứ tự sắp xếp không gian của chúng.
  - Với mạng Kohonen, quá trình học là quá trình ánh xạ một neuron tương ứng với một lớp các đầu vào cùng loại. Các neuron lân cận biểu diễn cho những đầu vào với các giá trị "gần nhau".

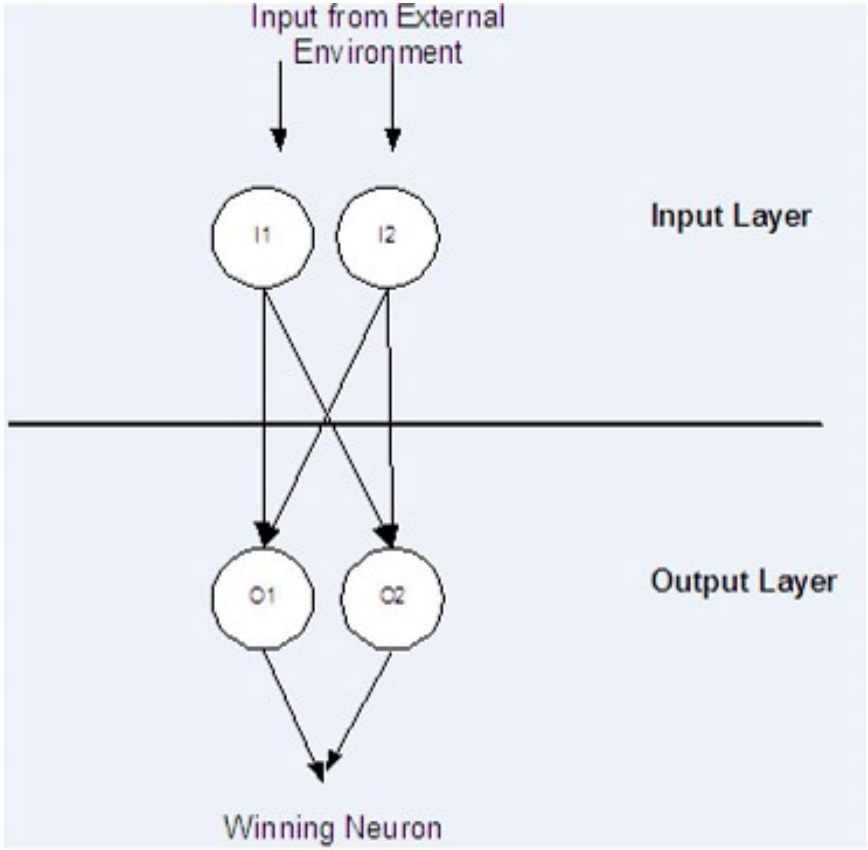
Ánh xạ tự tổ chức

- Bảo toàn trật tự sắp xếp (topology-preserving):
  - Các mẫu dữ liệu thuộc không gian nhiều chiều khi ánh xạ thành các mảng neuron với số chiều nhỏ hơn (thường là 2 chiều) được bảo toàn thứ tự sắp xếp không gian của chúng.
  - Với mạng Kohonen, quá trình học là quá trình ánh xạ một neuron tương ứng với một lớp các đầu vào cùng loại. Các neuron lân cận biểu diễn cho những đầu vào với các giá trị "gần nhau".

Ánh xạ tự tổ chức

- Mạng Kohonen chỉ 2 lớp:
  - Một lớp đầu vào (với số neuron bằng số chiều vector đầu vào)
  - Một lớp đầu ra (với số neuron bằng số loại mẫu).
  - Không có lớp ẩn.
- Chỉ có một neuron sinh ra kết quả trong mạng neuron, gọi là winning neuron, tương ứng với tập giá trị đầu vào mà có chưa đầu vào hiện tại.

4.2.2 Cấu trúc mạng Kohonen



Một mạng Kohonen 2x2

4.2.3 Giải thuật học

- Bắt đầu: Khởi tạo giá trị ngẫu nhiên cho m vector trọng số liên kết n-chiều của m neuron lớp đầu ra.
- Bước 1: Chọn vector đầu vào X trong tập dữ liệu mẫu.
- Bước 2: Chọn winning neuron, giả sử là k.
- Bước 3: Cập nhật vector trọng số của k và các neuron lân cận thêo công thức:

  $w_i \leftarrow w_i + \eta\theta(i,k)(X - w_i)$

  với η là tốc độ học, θ là hàm lân cận.
- Bước 4: Nếu đã lặp đủ số lần cho trước, hoặc sai số đã đủ nhỏ thì kết thúc. Nếu không thì trở về bước 1.

Áp dụng mạng Kohonen cho bài toán nhận dạng chữ viết

❖Đặc trưng huấn luyện

- Zoning: chia các pixel thành 8x8 vùng, trên mỗi vùng tính độ xám trung bình của các pixel
- Chiếu: thực hiện phép chiếu histogram 2 nửa theo phương thẳng đứng → 10 X 2 đặc trưng
- Profile: Xác định khoảng cách từ đường bao chữ đến biên hình chữ nhật bao chữ → 10 x 4 đặc trưng

→ Tổng cộng: 114 đặc trưng

5. Chương trình demo

- GoogleDocsDemo là một ứng dụng mô phỏng Google Docs của google:
  - Cho phép bôi đen đoạn văn bản chứa trong hình ảnh
  - Nhận dạng nội dung trong phần bôi đen
- Chương trình được viết bằng JavaScript / HTML5, có thể chạy trên bất cứ trình duyệt nào hỗ trợ.

5. Chương trình demo

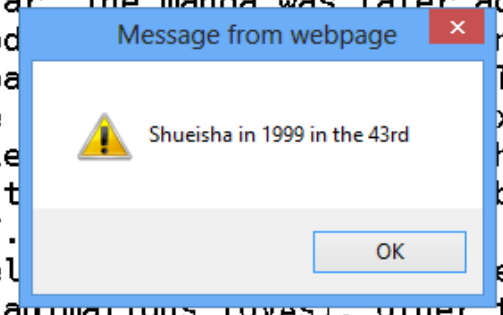
Naruto is an ongoing Japanese manga series written and illustrated by Masashi Kishimoto. The plot tells the story of Naruto Uzumaki, an adolescent ninja who constantly searches for recognition and dreams to become the Hokage, the ninja in his village who is acknowledged as the leader and the strongest of all. The series is based on a one-shot comic by Kishimoto that was published in the August 1997 issue of Akamaru Jump.

The manga was first published by Shueisha in 1999 in the 43rd issue of Japan Weekly Shonen Jump magazine. Currently, the manga is still being serialized; fifty-nine tankobon volumes have been released so far. The manga was later adapted into an anime, which was produced by Studio Pierrot and Aniplex. It premiered across Japan on the terrestrial TV Tokyo network and the anime satellite television network Animax on October 3, 2002. The first series lasted 220 episodes, while Naruto Shippuden, a sequel to the original series, has been airing since February 15, 2007.In addition to the anime series, Studio Pierrot has developed eight movies for the series and several original video animations (OVAs). Other types of merchandise include light novels, video games and trading cards developed by several companies.

5. Chương trình demo

Naruto is an ongoing Japanese manga series written and illustrated by Masashi Kishimoto. The plot tells the story of Naruto Uzumaki, an adolescent ninja who constantly searches for recognition and dreams to become the Hokage, the ninja in his village who is acknowledged as the leader and the strongest of all. The series is based on a one-shot comic by Kishimoto that was published in the August 1997 issue of Akamaru Jump.

The manga was first published by Shueisha in 1999 in the 43rd issue of Japan Weekly Shonen Jump magazine. Currently, the manga is still being serialized; fifty-nine tankobon volumes have been released so far. The manga was later adapted into an anime, which was prod nd Aniplex. It premiered across Japa Tokyo network and the anime satellite x on October 3, 2002. The first serie hile Naruto Shippuden, a sequel to t been airing since February 15, 2007. series, Studio Pierrot has devel e series and several original video animations (OVAs). Other types of merchandise include light novels, video games and trading cards developed by several companies.



Cảm ơn thầy và các bạn đã lắng nghe!