MÔ HÌNH HỌC SÂU SONG TUYẾN CHO NHẬN DẠNG ĐỐI TƯỢNG

Nguyễn Đăng Bình

Khoa Công nghệ Thông tin, Trường Đại học Khoa học, Đại học Huế

Email: ndbinh@husc.edu.vn

Ngày nhận bài: 13/12/2022; ngày hoàn thành phản biện: 20/12/2022; ngày duyệt đăng: 26/6/2023

TÓM TẮT

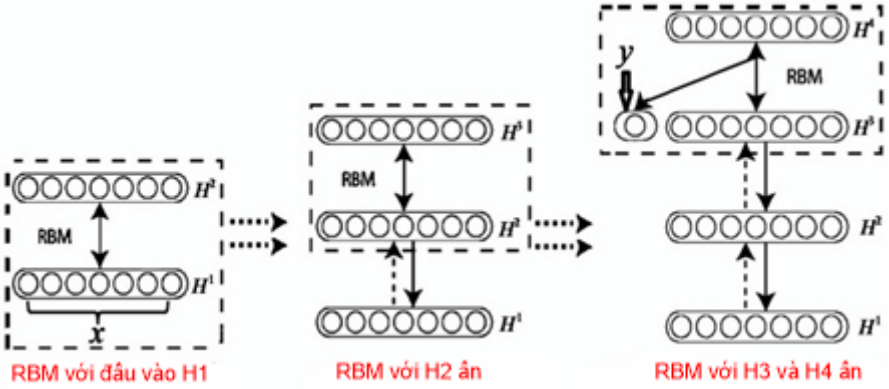
Nhận dạng đối tượng là một trong những vấn đề cơ bản trong thị giác máy tính. Bài báo này giới thiệu mô hình học sâu được gọi là mạng niềm tin sâu song tuyến tính (BDBN) cho bài toán nhận dạng đối tượng ảnh. Để thích nghi nhận dạng đối tượng trong thế giới thực, bài báo nghiên cứu phát triển BDBN trên phương pháp học bán giám sát giúp cho mô hình học sâu được tốt hơn khi mẫu huấn luyện được gán nhãn không đủ. Thực nghiệm và lượng hóa kết quả so sánh với các phương pháp khác trên các bộ dữ liệu dành cho cộng đồng nghiên cứu cho kết quả khá tốt có phần vượt trội.

Từ khóa: Học sâu, nhận dạng đối tượng, phép chiếu song tuyến tính.

1. MỞ ĐẦU

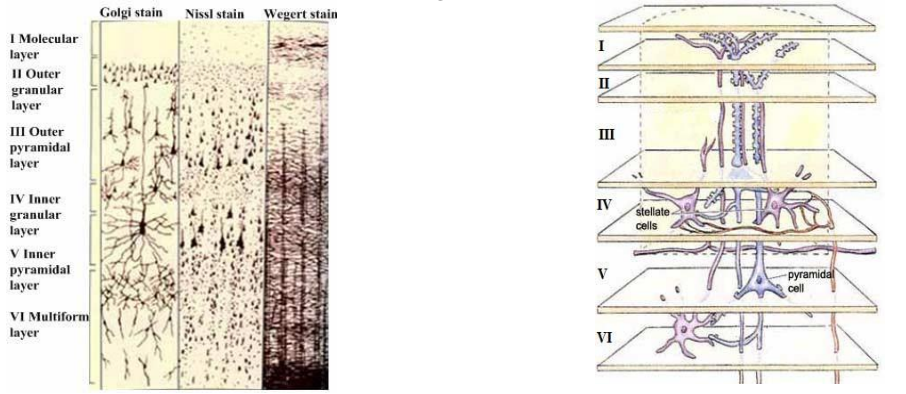
Nhận dạng đối tượng là một phần quan trọng của thị giác máy tính vì nó liên quan chặt chẽ đến sự thành công của nhiều ứng dụng thị giác máy tính và công nghệ hiện nay. Một số mô hình, thuật toán và hệ thống nhận dạng đối tượng đã được đề xuất trong một thời gian dài để giải quyết vấn đề này. Tuy nhiên trong từng ứng dụng thực tế thì kết quả còn khiêm tốn. Để cải thiện độ chính xác của nhận dạng đối tương thì học sâu là một hướng được nhiều nhà nghiên cứu quan tâm hiện nay. Có hai cách tiếp cận đó là học sâu có giám sát và học sâu không có giám sát. Học sâu có giám sát phục vụ mục đích phân lớp mẫu, thông thường bằng cách đặc trưng hóa phân phối sau của các lớp dựa trên dữ liệu quan sát được [1]. Dữ liệu đích luôn có sẵn ở dạng trực tiếp hay gián tiếp cho các dạng học có giám sát. Loại này còn có tên gọi là mạng học sâu dự báo. Trong số các loại mạng học sâu không giám sát, phổ biến nhất là loại mô hình học sâu dựa trên mức năng lượng [6]. Một đề xuất một khối chuyển đổi song tuyến tính sâu (DBT) mới, khối này có thể được tích hợp vào các mạng thần kinh tích chập sâu. DBT tận dụng lợi thế của thông tin ngữ nghĩa và có thể thu được các tính năng song tuyến tính một cách hiệu quả bằng cách tính toán tương tác theo cặp trong một nhóm ngữ nghĩa [11]. Sử dụng các phép chiếu song tuyến tính để thay thế các

phép chiếu đầy đủ phi cấu trúc truyền thống để tối ưu hóa mạng nơ ron học sâu [12]. Hiệu năng của học sâu được chú ý, đặc biệt là sau sự ra đời của mô hình mạng tin sâu (DBN) [10]. Đây là mô hình học không giám sát được tạo bởi nhiều lớp biến ngẫu nhiên và lớp biến ẩn. Hai lớp trên cùng có những liên kết gián tiếp, đối xứng giữa chúng. Các lớp thấp hơn nhận kết nối trực tiếp từ những lớp trên. Các thủ tục học của DBN có thể được chia thành hai giai đoạn: trừu tượng hóa thông tin theo từng tầng và tinh chỉnh toàn bộ mạng sâu với mục tiêu huấn luyện cuối cùng [10].



Hình 1. Kiến trúc mạng tin sâu (DBN)

Một DBN với một đầu vào lớp H^1, ba lớp ẩn H^2, H^3, H^4, trong khi x là véc tơ diễn tiến của dữ liệu đầu vào, và y là các mục tiêu cần học. Trong giai đoạn đầu tiên, DBN bắt cặp mỗi lớp tiến về phía trước với một lớp phản hồi để tái tạo lại các lớp dữ liệu đầu vào từ các lớp đầu ra. Trong Hình 1, xây dựng lại lớp khôn ngoan sẽ xảy ra giữa H^1 và H^2, H^3 và H^3, H^3 và H^4 được thực hiện bởi một (RBMs). Sau khi từng cặp của các lớp được học không giám sát tham lam, các đặc trưng ở lớp thấp hơn sẽ kết hợp thành đại diện cấp cao gọn hơn. Kiến trúc sâu là giống với nhiều lớp cấu trúc vật lý của vỏ não giống chức năng thị giác của con người [9]. Vỏ não là nơi liên kết với nhiều khả năng nhận thức, có một hệ thống phân cấp lớp đa phức tạp. Các lớp cấu trúc vỏ não được xếp thành phiến hình minh họa trong Hình 2.



(a) Kiến trúc xếp thành phiến (b) Kiến trúc nhiều lớp

Hình 2. Kiến trúc nhiều lớp của vỏ não mới [9].

Vỏ não có thể được tạm chia thành sáu lớp chức năng khác biệt với lớp phân tử I đến đa dạng lớp VI [9]. Do đó, hàng chục lớp trên vỏ não liên kết với nhau tạo ra các tế bào nhận thức đơn giản nhất. Cách thức dữ liệu được phân phối trong một kiến trúc sâu là mô phỏng tốt các thông tin lan truyền ngược trong vỏ não. Có nhiều lý do để tin rằng các hệ thống nhận thức có nhiều lớp mô hình sinh sản,trong đó các liên kết từ trên xuống có thể được sử dụng để tạo ra các tính năng cấp thấp của hình ảnh từ các đại diện cấp cao, và các liên kết từ dưới lên có thể được sử dụng để suy ra đại diện cấp cao tạo ra một bộ quan sát các tính năng cấp thấp. Việc lưu giữ các ô nhớ và các kết nối qua lại giữa các khu vực trong vỏ não đã đưa ra một hệ thống các tính năng dần dần phức tạp hơn, trong đó mỗi lớp có thể ảnh hưởng đến các lớp ở dưới nó. Một số mô hình kiến trúc nổi bật là máy Boltzmann hạn chế (RBM) [5] [7], mạng tin sâu (DBN), mạng nơ ron sâu (DNN) [2], Mạng nơ ron hồi quy (RNNs) [3], Mạng nơ ron hồi quy (RNNs) có thể được xét như là một phân lớp khác của mạng học sâu không giám sát, loại mà chiều sâu có thể tăng cùng với chiều dài của chuỗi dữ liệu đầu vào [8].

Nội dung nghiên cứu tập trung vào việc trình bày một mô hình kiến trúc sâu thích hợp và tương ứng với thuật toán học cho các nhiệm vụ nhận dạng đối tượng ảnh. Các kết quả nghiên cứu mới nhất và kết quả từ khoa học thần kinh đã chỉ ra rằng mô hình sâu phù hợp với cấu trúc vật lý, quá trình tiến hóa của trí thông minh, và phương pháp lan truyền ngược thông tin trong vỏ não giống quá trình nhận thức của con người. Do đó, nó cho thấy tiềm năng rất lớn để cung cấp hệ thống có thể nhận thức giống như con người khi phân tích các nội dung đa phương tiện trong hệ thống thị giác máy tính.

Trong bài báo này trình bày mô hình học sâu song tuyến tính. Cuối cùng là trình bày thực nghiệm, đánh giá kết quả phương pháp xây dựng, từ đó thảo luận một số hướng nghiên cứu tiếp theo trong nhận dạng đối tượng.

2. MÔ HÌNH HỌC SÂU SONG TUYẾN TÍNH

Từ nhận xét trong các lĩnh vực khoa học thần kinh, mô hình sâu được chọn trình bày trong bài báo này này cho việc nhận dạng đối tượng. Để thích ứng tốt hơn với các dữ liệu hình ảnh và các ứng dụng phân loại hình ảnh, mô hình lý thuyết học sâu được gọi là mạng song sâu (BDBN) [9] với một kiến trúc sâu mới và một thuật toán học sâu mới. Kiến trúc sâu của BDBN được thiết kế bằng cách tham khảo hệ thống nhận thức con người nhận thức. Trong vỏ não ban đầu, cách thức tiếp nhận thông tin qua các cơ quan cảm giác đến vị trí dây thần kinh là một tín hiệu trực tiếp từ một góc trong mắt, giống như một ma trận. Vì vậy, các lớp đầu vào và tất cả các lớp ẩn trong BDBN được xây dựng bởi một tập hợp mặt phẳng bậc 2, mà cũng phù hợp với cơ cấu căng tự nhiên của hình ảnh. Tất cả mặt phẳng được kết nối đầy đủ với những mặt phẳng kế cận cho đến lớp ra, đó là một véc tơ được gán nhãn của những hình ảnh.

Dựa trên kiến trúc sâu mới, một thuật toán học sâu với ba giai đoạn: khởi tạo song tuyến phân biệt, xây dựng lại lớp tham lam thông minh, và tinh chỉnh toàn cục. Lý giải cho việc học tập ba giai đoạn xuất phát từ hiện tượng hai đỉnh kích hoạt trong các khu vực trên vỏ não. Đối với việc nhận dạng các đối tượng, đỉnh trước đó là liên quan đến sự kích hoạt của một "dự đoán ban đầu" dựa trên những kiến thức phân biệt đã được kế thừa lại, trong khi đó đỉnh sau phản ánh những khái niệm liên quan đến các đối tượng đã được nhận dạng [9]. Trong hầu hết các mô hình học sâu hiện tại, chuyển kích hoạt được mô hình hóa bởi các giai đoạn tinh chỉnh, nhưng quá trình "dự đoán ban đầu" bị bỏ quên. Trong mô hình mới này, hai đỉnh sẽ được kích hoạt và lan truyền thông tin trong vỏ não được nhận thức một cách trung thực.

Mô hình kích hoạt đỉnh trên cùng của "dự đoán ban đầu" bằng cách giữ các thông tin phân biệt của dữ liệu được gán nhãn mức lớn nhất. Hầu hết các mô hình sâu hiện tại khởi tạo các không gian tham số ngẫu nhiên và tiến dần về cực trị địa phương. Thật không may, nếu không gian khởi tạo tham số ban đầu không tốt có thể dẫn đến cực trị địa phương không tốt và sẽ ảnh hưởng nghiêm trọng việc học sau này. Để giải quyết vấn đề này, sử dụng một chiến lược song tuyến phân biệt để xây dựng lớp thứ hai theo thứ tự từ lớp thấp hơn [9]. Các liên kết trọng số đối xứng giữa hai lớp liền kề được sử dụng như là không gian tham số ban đầu cho việc học tập sau này. Hơn nữa, "dự đoán ban đầu" phân biệt mang lại một lợi thế bổ sung cho kiến trúc này. Hiện nay, số lượng các tế bào thần kinh trong mỗi lớp là cố định và được xác định trước bằng trực giác. Trong mô hình này, kích thước của các kiến trúc sâu được xác định dựa vào kích thước tối ưu cho việc giữ lại các thông tin phân biệt.

2.1. Mạng niềm tin sâu song tuyến tính

Gọi X là một tập hợp các dữ liệu mẫu $X = [X_1, X_2, ... X_k, ..., X_K]$, với $\mathbf{X}_k$ là một dữ kiện mẫu trong không gian ảnh $R^{I \times J}$, K là số lượng dữ liệu mẫu.

Gọi Y là tập các nhãn tương ứng với X, $Y = [y_1, y_2, ..., y_k]$, và y là vector nhãn của $\mathbf{X}_k$ trong không gian R^c, trong đó C là số lớp.

$$y_k^c = \begin{cases} 1 \text{ nếu } X_k \text{ thuộc lớp } c \\ 0 \text{ nếu } X_k \text{ không thuộc lớp } c \end{cases}$$

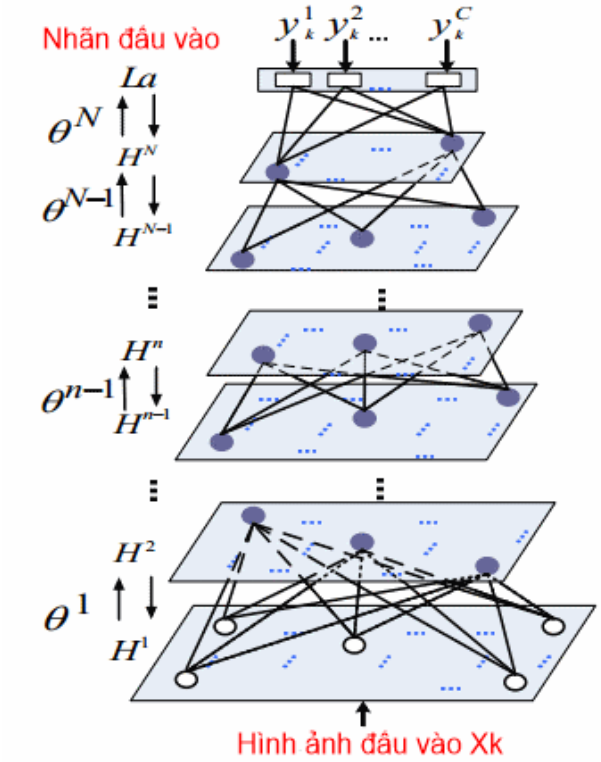
Dựa trên tập huấn luyện đã cho, mục tiêu trong phân lớp đối tượng là học bằng một hàm ánh xạ từ các hình ảnh tập X đến nhãn tập Y và sau đó phân loại các dữ liệu mới theo các hàm phân biệt đã được huấn luyện.

Để giải quyết vấn đề phân loại hình ảnh, một mạng tin sâu kết nối với nhau bao gồm các lớp đầu vào H^1, ẩn lớp H^2, ..., H^n và một lớp nhãn Lα ở lớp trên cùng. Lớp đầu vào H^1 có I x J đơn vị và kích thước này tương đương với kích thước của các đặc trưng đầu vào [9]. Trong mô hình này, sử dụng các giá trị điểm của tập mẫu $\mathbf{X}_k$ là các đặc trưng đầu vào ban đầu. Trên cùng, lớp nhãn có C đơn vị tương ứng với số lượng của

các lớp. Việc tìm kiếm hàm ánh xạ từ X vào Y chuyển thành tìm kiếm các không gian tham số tối ưu θ cho kiến trúc sâu.

Quy trình học của mô hình BDBN

1. Chiến lược BDBN được sử dụng để xây dựng ánh xạ dữ liệu gốc vào không gian con song tuyến phân biệt.
2. Các liên kết đối xứng ban đầu được xây dựng giữa các lớp liền kề dựa trên các thông tin phân biệt. Kích thước của các kiến trúc sâu sẽ được tính tự động dựa trên kích thước tối ưu để giữ lại các thông tin phân biệt.
3. Sau khi kiến trúc của lớp tiếp theo được xác định, không gian tham số được tinh chỉnh bằng cách tái cấu trúc các lớp khôn ngoan tham lam sử dụng RBMs như các khối.
4. Lặp lại bước thứ 3 cho đến khi không gian tham số θ trong tất cả các lớp được tái cấu trúc.
5. Trong quá trình kích hoạt sau, toàn bộ các mô hình sâu đều được tinh chỉnh để giảm thiểu các lỗi phân loại dựa trên phương pháp lan truyền ngược.



Hình 3. Kiến trúc mạng song sâu BDBN

2.2. Khởi tạo song tuyến phân biệt

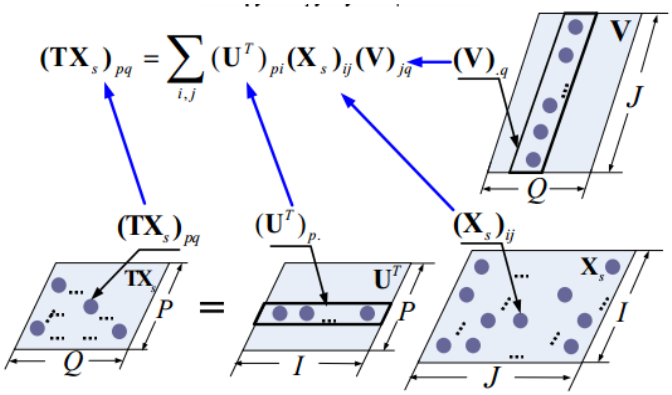
Phép chiếu song tuyến sử dụng trích xuất các thông tin phân biệt từ các tập dữ liệu hình ảnh ban đầu, với các tập dữ liệu huấn luyện được gán nhãn $X_1, X_2, X_3, ...X_n \in R^{IxJ}$ không gian mở, dữ liệu đầu vào là kiểu véc tơ, Phép chiếu song tuyến tìm thấy 2 ma trận chiếu $U \in R^{JxP}$ và $V \in R^{JxQ}$ với $TX_1 = U^T X_1 V$ (s = 1, …., L) giống như Hình 4 . Các $TX_1, TX_2, …, TX_t \in R^{IxJ}$ (I và J là kích thước ma trận ban đầu, P và Q là kích thước ma trận lúc sau).

Ta có công thức (u, v là véc tơ riêng và λ là giá trị riêng của từng véc tơ):

$$D_v u=\lambda u, \text{ với } Dv = \sum_{st} E_{st}(X_s - X_t)^T V V^T (X_s - X_t)^T$$

$$D_u v=\lambda v, \text{ với } Du = \sum_{st} E_{st}(X_s - X_t)^T U U^T (X_s - X_t)$$

Chúng ta có thể tối ưu hóa luân phiên U (với một V cố định) và V (với U cố định). Các bước trên tăng đơn điệu J(U,V) và khi đó hàm được bao bọc phía trên sẽ hội tụ về một điểm quan trọng với ma trận biến đổi U, V.



Hình 4. Phép chiếu song tuyến.

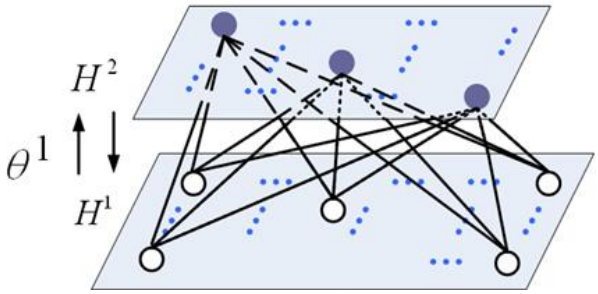
Các kích thước của P và Q được xác định bởi số lượng các giá trị đặc trưng tích cực trong D_V và D_U, tương ứng, khi bổ sung các véc tơ riêng tương ứng với giá trị riêng không dương sẽ không tăng J(U, V). Kết quả là kích thước ban đầu IxJ sẽ tự động giảm vào PxQ.

2.3. Tái cấu trúc lớp thông minh tham lam

Các tập mẫu dữ liệu X được nhập cho các kiến trúc sâu như các lớp đầu vào H^1 để xây dựng một RBM với lớp ẩn đầu tiên H^2.

Tất cả các mô hình học sâu hiện tại đều xác định cấu trúc dựa vào kích thước của các lớp ẩn hoặc trực giác. Trong mô hình này sẽ cung cấp một kiến trúc ý nghĩa hơn bằng cách tích hợp các thông tin mang tính quyết định từ dữ liệu ghi trên nhãn. Để tích hợp các thông tin thu được từ phép chiếu song tuyến phân biệt (DBP) để phân loại,

chúng ta có hai thủ tục: xác định kích thước của lớp ẩn và tính toán các liên kết trọng số đối xứng phân biệt.



Hình 5. Sơ đồ tái cấu trúc từ lớp H^1 tới lớp H^2

Năng lượng của trạng thái (h^1, h^2) trong RBM đầu tiên là:

$$E(h^1, h^2; \theta^1) = -(h^1 A^1 h^2 + b^1 h^1 + c^1 h^2) \quad (1)$$

Với

- $\theta^1 = (A^1, b^1, c^1)$ là các mô hình tham số giữa các lớp đầu vào H^1 và lớp ẩn đầu tiên H^2 .
- $A^1_{ij,pq}$ là tương tác đối xứng giữa đầu vào (i, j) trong H^1 và đầu ra ẩn (p, q) trong H^2 . b^1_{ij} là (i, j)th độ dốc của lớp H^1 và c^1_{pq} là (p,q)th độ dốc của lớp H^2 .
- IxJ là số đơn vị trong H^1 trong khi P^2 x Q^2 là số đơn vị trong H^2 .

$$P(h^1) = \frac{1}{Z} \sum_{h^2} e^{-E(\boldsymbol{h^1}, \boldsymbol{h^2}, \boldsymbol{\theta^1})} = \frac{\sum_{h^2} e^{-E(\boldsymbol{h^1}, \boldsymbol{h^2}, \boldsymbol{\theta^1})}}{\sum_{\boldsymbol{h^1}} \sum_{\boldsymbol{h^2}} e^{-E(\boldsymbol{h^1}, \boldsymbol{h^2}, \boldsymbol{\theta^1})}} \quad (2)$$

Và hàm khả năng (log-likelihood) của $P(h^1)$ là:

$$\text{Log } P(\mathbf{h}^1) = \log \sum_{\boldsymbol{h^2}} \boldsymbol{e}^{-E(\boldsymbol{h^1}, \boldsymbol{h^2}, \boldsymbol{\theta^1})} - \log \sum_{\boldsymbol{h^1}} \sum_{\boldsymbol{h^2}} \boldsymbol{e}^{-E(\boldsymbol{h^1}, \boldsymbol{h^2}, \boldsymbol{\theta^1})} \quad (3)$$

Áp dụng thuật toán ước lượng xấp xỉ để cập nhật không gian tham số cho bước thứ nhất.

$$\frac{\boldsymbol{\partial log P(h^1(0))}}{\boldsymbol{\partial A^1}} = \langle h^1(0) h^2(0) \rangle_{data} - \langle h^1(1) h^2(1) \rangle_{recon} \quad (4)$$

Với

- $\langle . \rangle_{data}$ biểu thị kỳ vọng với việc phân phối dữ liệu
- $\langle . \rangle_{recon}$ biểu thị tái cấu trúc phân phối dữ liệu sau bước thứ nhất.

Phép chiếu song tuyến có thể tự động giảm kích thước ban đầu từ IxJ thành PxQ thông qua việc chuyển đổi ma trận U^1 và V^1. Kết quả là số lượng tế bào thần kinh trong lớp H^2 được xác định bởi các hàng và cột kích thước của ma trận biến đổi U^1 và V^1.

$$\mathbf{P}^2 = \text{row}(\mathbf{U}^1) \text{ , } \mathbf{Q}^2 = \text{column}(\mathbf{V}^1)$$

Trong các mô hình học sâu hiện tại, trọng số của các liên kết đối xứng A khởi tạo giá trị ngẫu nhiên nhỏ nhất được chọn từ số zero-mean Gaussian với độ lệch chuẩn trong khoảng 0,01. Với mô hình này, việc khởi tạo các trọng số cho liên khác đối xứng ban đầu thông qua phương trình.

$$\boldsymbol{A^1_{ij,pq}}\,(0) = (\boldsymbol{U^1_{ip}})^T \boldsymbol{V^1_{jq}} \qquad (5)$$

Với $\boldsymbol{A^1_{ij,pq}}$ là tương tác đối xứng giữa đầu vào (i,j) trong H^1 và đầu ra ẩn (p,q). Lớp trừu tượng khôn ngoan tham lam cho lớp đầu tiên H^1 với lớp kế cận H^2 và qui trình này sẽ được lặp lại ở các lớp cao hơn tiếp theo.

Tinh chỉnh toàn bộ

Việc sử dụng thuật toán tham lam theo lớp để học trong mô hình học sâu với sự hỗ trợ của các thông tin thu được từ phép chiếu song tuyến phân biệt. Bài báo dùng lan truyền ngược cho toàn bộ mô hình sâu để tinh chỉnh các tham số θ=[A,b,c] để tái cấu trúc tối ưu [9]. Trong giai đoạn thông tin trừu tượng theo lớp tham lam, việc tìm kiếm toàn cục đã được thực hiện tốt và hợp lý trong không gian tham số. Trước khi tiến hành

$$\theta^* = \arg\min_{\theta}[-\sum_{l} \mathbf{y}_l \log \hat{\mathbf{y}}_l] \qquad (6)$$

tinh chỉnh, bài báo xây dựng mô hình dữ liệu tốt chính xác, sau đó lan truyền ngược được sử dụng để tinh chỉnh toàn bộ mạng học sâu bằng cách tìm cách thông số tối ưu $\theta^* = [A^*, b^*, c^*]$ để phân loại dữ liệu một cách hiệu quả. Thuật toán được sử dụng để giảm thiểu tối đa các phân loại lỗi $[-\sum_t \boldsymbol{y_l}\, \boldsymbol{log \hat{y}_l}]$ với y_l và $\hat{\mathbf{y}}_l$ là các nhãn chính xác và các nhãn đầu ra thuộc dữ kiện X_l thuộc không gian X^L .

2.4. Thuật toán học sâu song tuyến tính

Đầu vào:

Tập huấn luyện X, tập mẫu đã được gán nhãn X^L trong X;

Tương ứng với tập nhãn Y;

N là số lớp, E là số vòng lặp;

L số dữ liệu đã được gán nhãn, tham số α;

Trọng số giữa 2 lớp $\mathbf{B}_{st}$, trọng số trong lớp $W_{st;}$

Tham số độ dốc ban đầu **b** và **c;**

Động lượng ϑ và mức học ε_A, ε_b, $\varepsilon_{e;}$

Trọng số cân bằng α $\boldsymbol{\epsilon}$ [0,1].

Đầu ra:

Không gian tham số tối ưu $\theta^* = [\mathbf{A}^*, \mathbf{b}^*, \mathbf{c}^*]$

Giải thuật:

Lặp n = 1 đến N

  Lặp e = 1 đến E

    Nếu n = 1

      $T^n = X^L$

    Ngược lại

      Lặp l = 1 đến L **do**

        $\boldsymbol{T_l^n} = \sigma(\boldsymbol{T_l^{n-1}A^{n-1}} + \boldsymbol{c^{n-1}})$

      Kết thúc lặp

    Kết thúc nếu

    Lặp không là ma trận đường chéo (thuật toán QR tìm trị riêng)

      Dv là ma trận đặc trưng theo cột;

      $\mathbf{D}_U$ là ma trận đặc trưng theo dòng;

      Cố định **V**, tính **U** bằng $D_v u = \lambda u$ (u là véc tơ riêng của D_v với giá trị riêng λ);

      Cố định **U**, tính **V** bằng $D_u v = \lambda v$ (v là véc tơ riêng của D_u với giá trị riêng λ);

    Kết thúc lặp

    Xác định kích thước của lớp tiếp theo

    $P^{n+1} = \text{row}(U^n)$, $Q^{n+1} = \text{column}(V^n)$

    Tính trọng số khởi tạo của liên kết

    $\boldsymbol{A_{ij,pq}^n}(0) = (\boldsymbol{U_{ip}^n})^T \boldsymbol{V_{jq}^n}$

    Cập nhật trọng số và độ dốc

    $\boldsymbol{A_{ij,pq}^n} = \boldsymbol{\vartheta A_{ij,pq}^n} + \varepsilon_A(<\boldsymbol{h_{ij}^n}(0)\ \boldsymbol{A_{pq}^{n+1}}(0)>_{data} - <\boldsymbol{h_{ij}^n}(1)\ \boldsymbol{h_{pq}^{n+1}}\ (1)>_{recon})$

    $\boldsymbol{b_{ij}^1} = \vartheta \boldsymbol{b_{ij}^1} + \varepsilon_b\ (\boldsymbol{h_{ij}^1}(0) - \boldsymbol{h_{ij}^1}(1))$ (là độ dốc của lớp H^1)

    $\boldsymbol{c_{pq}^1} = \vartheta \boldsymbol{b_{pq}^1} + \varepsilon_c\ (\boldsymbol{h_{pq}^2}(0) - \boldsymbol{h_{pq}^2}(1))$ (là độ dốc của lớp H^2)

  Kết thúc lặp

Kết thúc lặp

Tính không gian tham số tối ưu bằng phương pháp lan truyền ngược

$\theta^* = \boldsymbol{argmin}_{\boldsymbol{\theta}}[- \sum_l \boldsymbol{y}_l \log \widehat{\boldsymbol{y}}_l]$

Kết thúc thuật toán.

3. THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

Trong phần này trình bày về mô phỏng thuật toán và tiến hành các thực nghiệm lượng hóa kết quả. Giới thiệu việc tạo lập một bộ dữ liệu đầu vào cho quá trình huấn luyện và một bộ dữ liệu cho phương pháp học sâu. Thiết kế chương trình huấn luyện dữ liệu và học bằng MATLAB 2014R và chương trình mô phỏng bằng ngôn ngữ lập trình C# sử dụng công cụ Visual Studio 2013. Minh họa cho mô hình mạng song sâu.

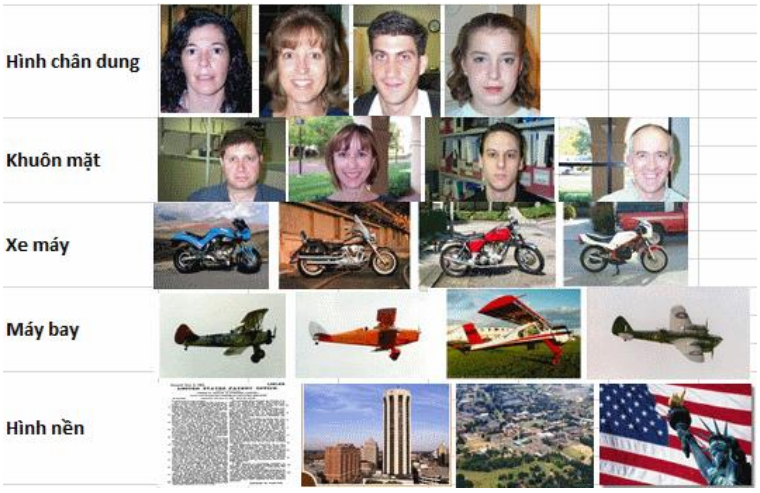
3.1. Các bộ dữ liệu thực nghiệm

1) ***Bộ dữ liệu CMU và PIE*** tại https://www.ri.cmu.edu/publications/the-cmu-pose-illumination-and-expression-pie-database/. Bộ dữ liệu này chứa 68 đối tượng với tổng số 41.368 hình ảnh khuôn mặt, các hình ảnh khuôn mặt được chụp đồng bộ bởi 13 máy ảnh với 21 đèn flash với các tư thế khác nhau (C05, C07, C09, C27, C29) như Hình 6. Với cách chọn này, đã chọn được 170 hình có độ phân giải 32x32 cho mỗi người.

Hình 6. Các mẫu hình được chọn để huấn luyện.

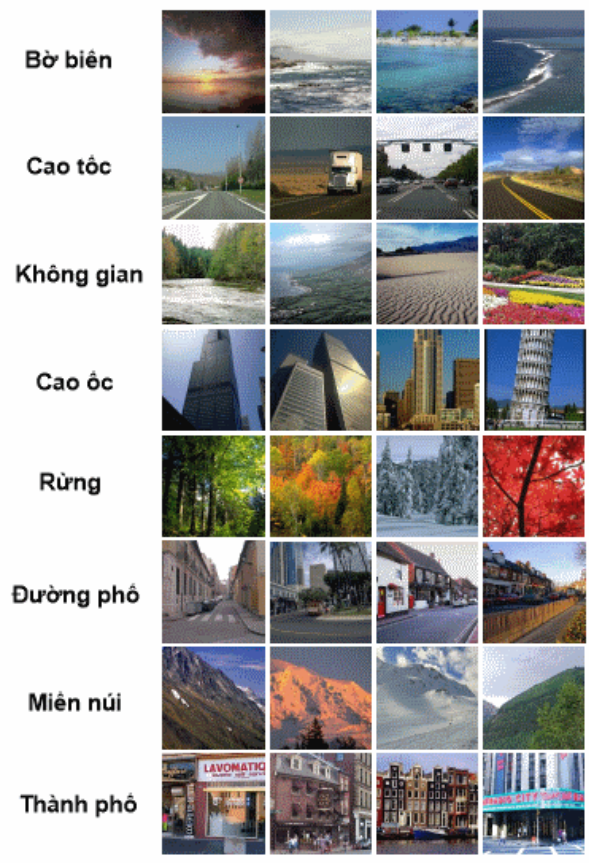
2) ***Bộ dữ liệu MNIST*** tại http://yann.lecun.com/exdb/mnist/ bao gồm 60.000 mẫu huấn luyện và 10.000 mẫu khác để nhận dạng với các hình ảnh chữ số viết tay từ 0 tới 9, mỗi mẫu là một ảnh kích thước 28x28.

3) ***Bộ dữ liệu Caltech101*** bao gồm các đối tượng hình ảnh thuộc 101 chủ đề, mỗi chủ đề có khoảng 40 tới 800 hình. Mỗi hình có kích thước 300x200. Đặt tại địa chỉ https://data.caltech.edu/records/mzrjq-6wc02. Trong thực nghiệm này, chọn 2935 hình với 5 chủ đề bao gồm: 435 hình khuôn mặt, 435 hình chân dung, 798 hình xe máy, 467 hình nền từ google và 800 hình máy bay.



Hình 7. Các mẫu hình từ Caltech101.

4) Bộ dữ liệu Urban & Natural Scene tại http://cvcl.mit.edu/database.htm với 2688 hình ảnh màu với 8 chủ đề: biển và bờ biển, đường cao tốc, không gian mở, nhà cao tầng, rừng, đường, miền núi, trung tâm thành phố. Mỗi chủ đề sẽ lấy ngẫu nhiên 50 hình.



Hình 8. Các mẫu từ Urban & Natural Scene.

3.2. Phân tích, thiết kế cài đặt chương trình

3.2.1. Phân tích

Thông số được cài đặt trong chương trình rất thông dụng trong các mô hình học sâu với trọng số cân bằng α là 0.5, số vòng lặp epoch được cố định là 30, cấp độ học η là 0.1, giá trị khởi tạo ban đầu là 0.5. Sau vòng lặp thứ 5, giá trị khởi tạo được thiết đặt là 0.9. Đến giai đoạn tinh chỉnh, phương pháp độ dốc liên hợp được dùng cho mỗi vòng lặp cho đến khi hội tụ. Bộ dữ liệu sau khi chọn lọc như tiêu chí trong phần 3.1 được chia làm 3 tập: tập huấn luyện, tập kiểm tra, tập giám sát sau đó được huấn luyện và kiểm tra dữ liệu.

Số lượng nơ ron của tầng H^1 bằng với kích thước của hình đưa vào trong phần 3.1 là 32x32. Số nơ ron tương ứng với tầng H^2, H^3 và H^4 lần lượt là 24x24, 21x21 và 19x19. Các kiến trúc sâu trước đó số lượng nơ ron tương ứng với tầng H^2, H^3, H^4 là 500, 500 và 2000.

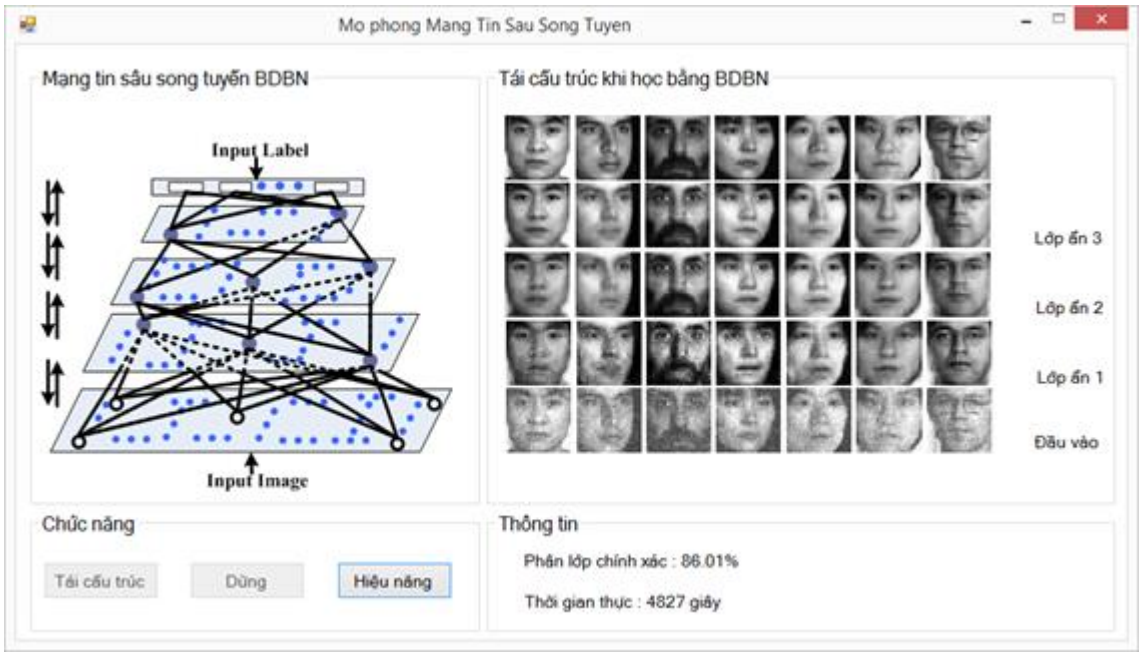
3.2.2. Cài đặt mô phỏng chương trình

Chương trình sử dụng MATLAB 2014R (với công cụ DeepLearnToolbox mã nguồn mở https://github.com/rasmusbergpalm/DeepLearnToolbox) trên giao diện console để huấn luyện và kiểm tra dữ liệu với bộ mã nguồn được tham khảo từ Sheng-hua Zhong, Ruslan Salakhutdinov và Geoff Hinton chạy trên máy tính hệ điều hành WIN8.1 với cấu hình như sau: CPU Core 2 Duo T6600 2.2GHz, RAM 4G, VGA 512MB.

Chương trình mô phỏng quá trình tái cấu trúc hình ảnh được viết bằng ngôn ngữ lập trình được sử dụng là C# trong bộ Visual Studio 2013 của Microsoft. Đây là bộ công cụ lập trình cho phép xây dựng các ứng dụng có giao diện Winform tương đối dễ dàng và tiện dụng. Ngoài ra Visual Studio 2013 cho phép tích hợp các .NET FrameWork. Trong nghiên cứu này sử dụng các .NET Framework như sau:

.NET Framework 4.5 do Microsoft cung cấp các tính năng thực thi trên các môi trường hệ điều hành.

Aforge.NET là một nền tảng cho tính toán trong khoa học và lập trình .NET. Nền tảng này được xây dựng dựa trên ngôn ngữ lập trình C#. Nó là gói công cụ dành cho các nhà nghiên cứu và lập trình viên thuộc lĩnh vực thị giác máy tính và trí tuệ nhân tạo AI dùng để xử lý hình ảnh, cung cấp các công cụ và thư viện hỗ trợ nhiều ứng dụng tính toán khoa học như xử lý dữ liệu thống kê, học máy, nhận dạng mẫu. Ngoài ra gói công cụ này còn cung cấp một số lượng lớn các hàm để tính toán phân bố xác suất, kiểm tra giả thuyết và hỗ trợ gần như hầu hết các kỹ thuật đo lường hiệu suất phổ biến.



Hình 9. Hình ảnh **m**inh họa hoàn tất quá trình tái cấu trúc bằng BDBN.

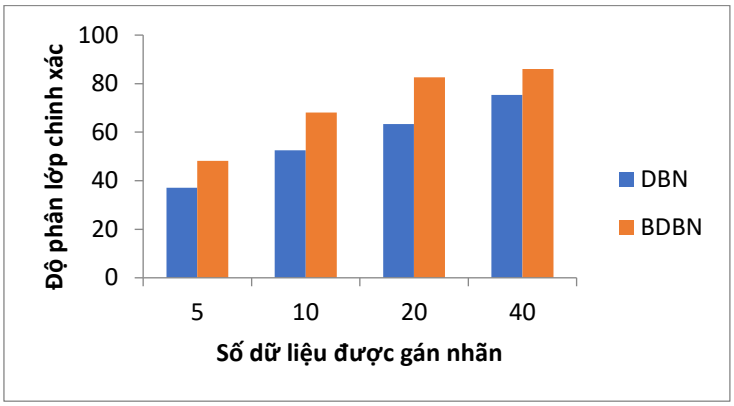
Trong giao diện chính gồm hai khung hình, khung trái mô tả cấu trúc chính của mạng tin song sâu, khung phải mô tả quá trình tái cấu trúc theo từng lớp của BDBN, tính độ chính xác và thời gian thực hiện.

3.2.3. Thực nghiệm, lượng hóa và đánh giá kết quả

1) ***Với bộ dữ liệu CMU PIE:*** Trong số 120 ảnh của mỗi người, số hình ảnh khác nhau được chọn ngẫu nhiên và được gán nhãn trong khi những người còn lại không được gán nhãn. Số lượng dữ liệu được gán nhãn theo mỗi người là 5, 10, 20 và 40. Số lần thử là 10. Số lần lặp là 50. Độ chính xác khi phân loại tăng lên với số lượng dữ liệu được gán nhãn. Kết quả thực nghiệm trên bộ dữ liệu mẫu huấn luyện và phân lớp được mô tả trong Bảng 1 sau:

Bảng 1. Kết quả thực nghiệm nhận dạng đối tượng ảnh với CMU PIE.

	5	**10**	**20**	**40**
DBN	37.10%	52.56%	63.41%	75.30%
BDBN	48.20%	68.10%	82.63%	86.40%



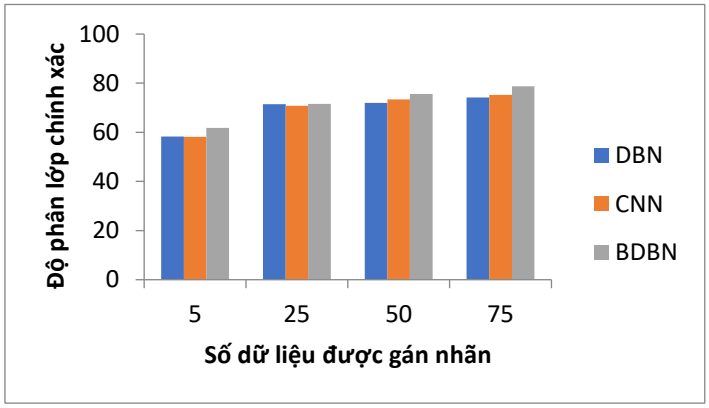
Hình 10. Biểu đồ kết quả thực nghiệm CMU PIE.

Kết quả sau khi thực nghiệm trên bộ dữ liệu CMU PIE từ Bảng 1. Với cùng số dữ liệu được gán nhãn và độ trắng Gaussian trong khoảng [0…0,005] ta thấy sự chính xác trong phân lớp đúng trung bình luôn cao hơn khi sử dụng mô hình học sâu. Vì quá trình tinh chỉnh tái cấu trúc từ tầng H^1 tới tầng H^5, độ nhiễu đã được loại bỏ, khi đến tầng H^5 ảnh gần giống so với ảnh gốc ban đầu giúp quá trình phân loại chính xác hơn.

2) ***Với bộ dữ liệu Caltech101***: do số lượng ảnh trong mỗi chủ đề khác nhau vì vậy trong nghiên cứu này lấy 50 ảnh ngẫu nhiên trong mỗi chủ đề để làm tập kiểm tra và tập huấn luyện. Số hình ảnh được gán nhãn trong dữ liệu lần lượt là 5, 25, 50, 75 trên mỗi chủ đề. Độ nhiễu Gaussian lần lượt được gán là 0.003, 0.005, 0.007, 0.01 và 0.02 . Số lần lặp là 50 nếu chọn số quá lớn sẽ thực nghiệm với thời gian nhiều hơn. Kết quả thực nghiệm trên bộ dữ liệu mẫu huấn luyện và phân lớp được mô tả trong Bảng 2 sau:

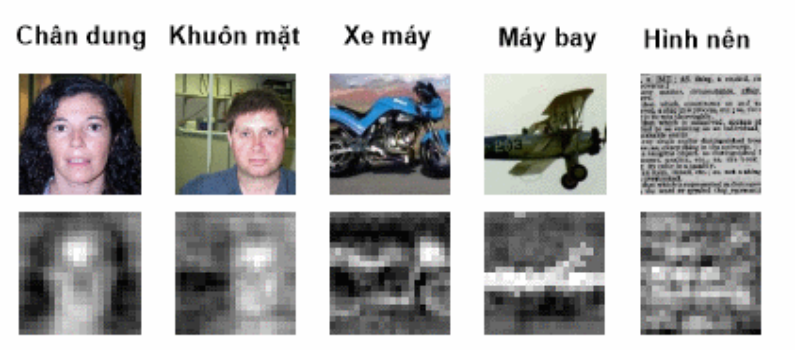
Bảng 2. Kết quả thực nghiệm nhận dạng đối tượng ảnh với Caltech101.

	5	**25**	**50**	**75**
DBN	58.30%	71.40%	72%	74.20%
CNN	58.20%	70.80%	73.40%	75.20%
BDBN	61.80%	71.60%	75.60%	78.80%



Hình 11. Biểu đồ kết quả thực nghiệm trên bộ dữ liệu Caltech101.

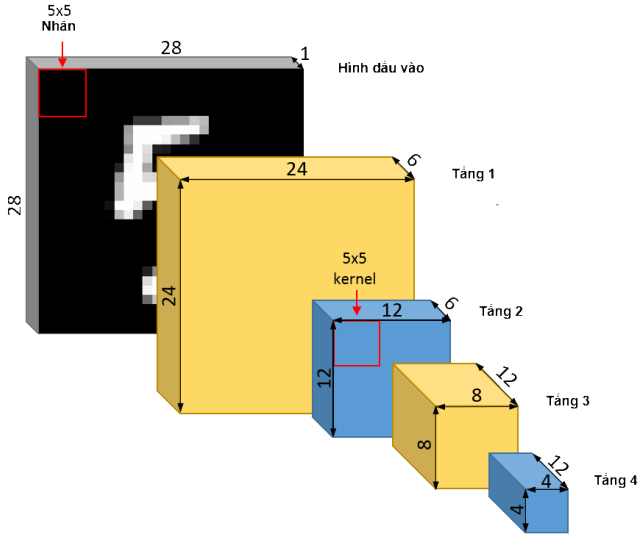
Kết quả sau khi thực nghiệm với bộ dữ liệu Caltech101, mặc dù tất cả đều là mô hình học sâu nhưng BDBN dùng 106 lần lặp so với 290 lần của DBN và CNN sự cải thiện này xuất phát từ phép chiếu song tuyến giúp dự đoán ban đầu tốt hơn khi xây dựng các liên kết trọng số giữa các lớp Hình 12 cho thấy kết quả tầng H^1 bằng trực quan.



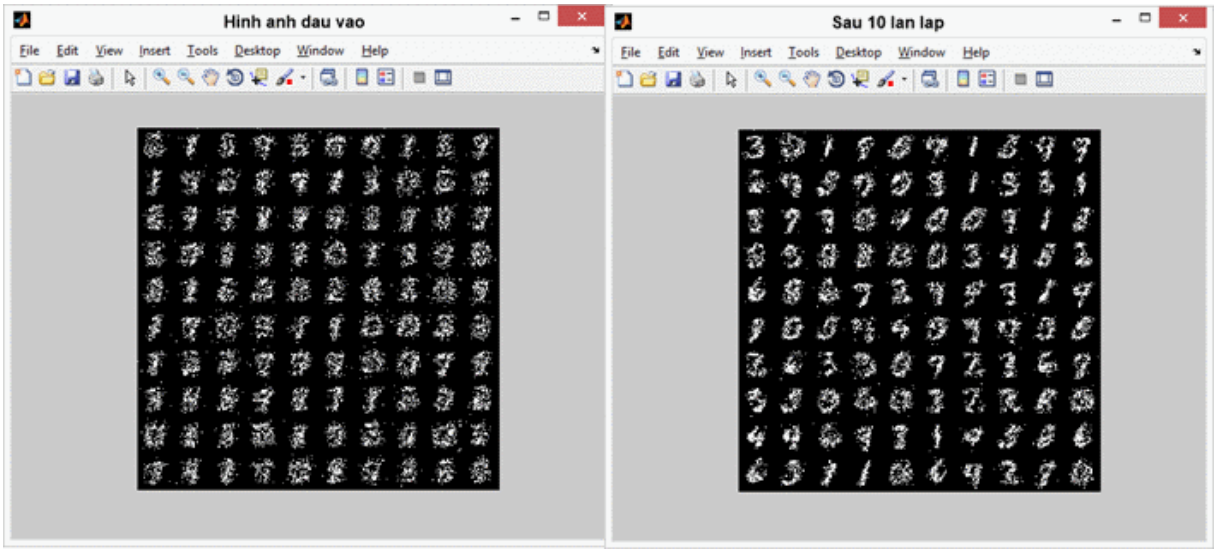
H

Hình 12. Kết quả trực quan ở tầng H1 với Caltech101.

3) ***Với bộ dữ liệu MNIST***: Trong nghiên cứu này chọn tập để huấn luyện 50000 mẫu và 10000 mẫu để kiểm tra với các mẫu đã được cố định kích thước. Trọng số cân bằng α là 1. Số lần lặp là 10. Kết quả thực nghiệm sau 10 lần lặp cho thấy hình ảnh dữ liệu hình ảnh đầu vào đã được cải thiện khi đến tầng H^1 sau quá trình tinh chỉnh tái cấu trúc, Hình 13 và Hình 14 cho thấy kết quả sau từng tầng. Kết quả sẽ cải thiện đáng kể nếu tăng số lần lặp lên 100 với tầng H^1 và 200 với tầng H^2.



Hình 13. Quá trình tái cấu trúc sau từng tầng.

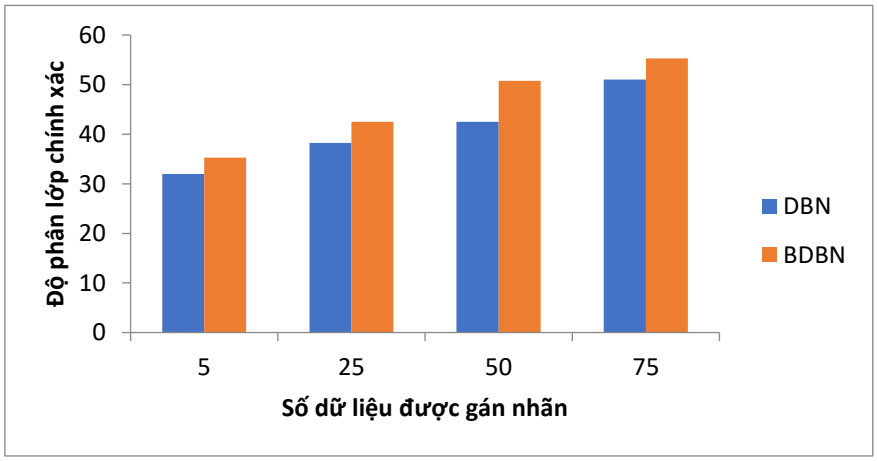


Hình 14. Kết quả sau 10 lần lặp với MNIST.

4) ***Với bộ dữ liệu Urban & Natural Scene:*** Các hình ảnh sẽ được điều chỉnh kích thước 32x32 tương ứng với số lượng nơ ron ở tầng H^1 lần lượt các tầng tiếp theo H^2 , H^3 , H^4 là 24x24 , 21x21 , 19x19 . Số lượng hình ảnh được gán nhãn trong dữ liệu lần lượt là 5, 25, 50 và 75 cho mỗi chủ đề. Số lần lặp là 10. Kết quả thực nghiệm trên bộ dữ liệu mẫu huấn luyện và nhận dạng được mô tả trong Bảng 3 sau:

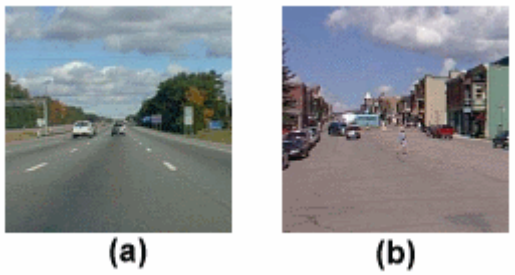
Bảng 3. Kết quả thực nghiệm nhận dạng ảnh với Urban & Natural Scene.

	5	25	50	75
DBN	32.00%	38.25%	42.50%	51.00%
BDBN	35.25%	42.50%	50.75%	55.25%



Hình 15. Biểu đồ kết quả thực nghiệm với Urban & Natural Scene.

Kết quả sau khi thực nghiệm với bộ dữ liệu Urban & Natural Scene, mặc dù độ chính xác có cao hơn so với bình thường, nhưng có các trường hợp nhận dạng không chính xác do cùng một ngữ cảnh.



Hình 16. (a) thuộc chủ đề Cao tốc, (b) thuộc chủ đề Đường.

4. KẾT LUẬN

Bài báo này giới thiệu một mô hình học sâu mới, học sâu song tuyến cho bài toán nhận dạng đối tượng ảnh. Mô hình học sâu song tuyến có một vài đặc tính hấp dẫn, đầu tiên là kiến trúc mới lạ mô phỏng cấu trúc vật lý nhiều lớp của vỏ thị giác và cho phép duy trì cấu trúc vô hướng tự nhiên của hình ảnh đầu vào trong quá trình truyền thông tin. Thứ hai, quá trình học ba giai đoạn của mô hình được thực hiện hóa một cách trung thực quy trình nhận dạng đối tượng của con người. Thứ ba, khả năng học bán giám sát của mô hình làm cho các kỹ thuật sâu được hoạt động tốt với số lượng dữ liệu được gán nhãn không đủ. Các thực nghiệm được cài đặt mô phỏng huấn luyện, kiểm tra và phân lớp dữ liệu có sử dụng mô hình học sâu song tuyến tính trên 4 bộ dữ liệu MNIST, Caltech101, Urban và Natural Scene và CMU PIE, sử dụng bộ công cụ MATLAB 2014R và Visual Studio 2013 với Aforge.net FrameWork. Đây là những công cụ hỗ trợ rất hiệu quả cho chương trình, nó làm cho mức độ chính xác của chương trình được nâng cao. Thời gian tới, hướng nghiên cứu và phát triển theo hai khía cạnh. Thứ nhất là cung cấp đặc tính ngữ nghĩa của hình ảnh, bằng cách gợi ý theo ngữ cảnh trong mô hình học sâu. Thứ hai, sử dụng mô hình học sâu song tuyến cho phân tích nội dung đa phương tiện với tập dữ liệu lớn.

TÀI LIỆU THAM KHẢO

[1]. Yoshua Bengio (2009), "Learning Deep Architectures for AI". *Foundations and Trends in Machine Learning*, Vol II, pp. 13-72.

[2]. Li Deng, Dong Yu (2013), "Deep Learning Methods and Applications". ISSN 1932-8346, Vol 7, pp. 214-241.

[3]. Pooan Safari (2013), "Deep Learning For Sequential Pattern Recognition", *Master thesis*, pp. 24-63.

[4]. Max Mignotte, Pascal Vincent, Yoshua Bengio (2010). "Heuristics for Optimizing Deep Architectures". *ECCV*, pp. 4-24.

[5]. Asja Fischer, Christian Igel, "An Introduction to Restricted Boltzmann Machines", *CIARP 2012*, LNCS 7441, pp. 14-36, 2013.

[6]. Andrew Ng, Jiquan Ngiam, Zhenghao Chen, Pang Wei Koh (2011), "Learning Deep Energy Model", *International Conference on Machine Learning USA* .

[7]. Ruslan Salakhutdinov, Geoffrey Hinton (2009), "Deep Boltzmann Machines", *JMLR*, Vol.5.

[8]. Jason Weston, Frederic Ratle, Hossein Mobahi and Ronan Collobert (2012), "Deep Learning via Semi-supervised Embedding", *Neural Networks: Tricks of the Trade*, LNCS 7700, pp. 639-655.

[9]. Yan Liu, Sheng-hua Zhong, Yang Liu (2011), "Bilinear Deep Learning for Image Classification", *Proceedings of the 19th International Conference on Multimedia*, ACM978-1-4503-0616.

[10]. Geoffrey Hinton (2007), "Deep Belief Nets". *NIPS Tutorial*, University of Toronto, 2007.

[11]. Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, Jiebo Luo (2019), "Learning Deep Bilinear Transformation for Fine-grained Image Representation", Proceeding of 33rd Conference on Neural Information Processing Systems, pp. 1-10

[12]. Litao Yu, Yongsheng Gao, Jun Zhou, Jian Zhang (2020), "Parameter-Efficient Deep Neural Networks With Bilinear Projections", *IEEE Transactions on neural networks and learning systems*, pp. 4075 – 4085.

BILINEAR DEEP LEARNING MODEL FOR OBJECT RECOGNITION

Nguyen Dang Binh

Faculty of Information Technology, University of Sciences, Hue University

Email: ndbinh@husc.edu.vn

ABSTRACT

Object recognition is one of the fundamental problems in computer vision. This paper introduces a deep learning model called bilinear deep learning belief network (BDBN) for image object recognition. In order to adapt the real-world object recognition, the paper develops BDBN on semi-supervised learning method to help bilinear deep learning model to be better when labeled training samples is not enough. Experimental and quantified results compared with other methods on benchmark data sets for the research community giving quite good results, somewhat superior.

Keywords: Bilinear discriminant projection, deep learning, object recognition..

Nguyễn Đăng Bình sinh ngày 08/11/1974 tại Thừa Thiên Huế. Năm 1996, ông tốt nghiệp Đại học ngành Toán - Tin tại trường Đại học Sư pham Huế. Ông nhận bằng thạc sỹ Công nghệ thông tin tại Trường Đại học Bách Khoa Hà Nội năm 2002; bằng Tiến sĩ ngành Công nghệ thông tin tại Viện Công nghệ Kyushu, Nhật Bản năm 2006 và hoàn thành nghiên cứu Sau tiến sĩ tại Viện Thị giác và Đồ họa máy tính tại Đại học Công nghệ Graz, Cộng hòa Áo, năm 2008. Hiện ông công tác tại khoa Công nghệ Thông tin, Trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: Học máy, Thị giác máy tính, Nhận dạng và Xử lý ảnh.