TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

─────── \* ───────

ĐỒ ÁN

**TỐT NGHIỆP ĐẠI HỌC**

NGÀNH CÔNG NGHỆ THÔNG TIN

**PHÂN VÙNG ẢNH MÀU SỬ DỤNG THUẬT TOÁN K-MEANS**

Sinh viên thực hiện: **Trần Văn Thành**

Lớp CNTT2.01 – K58

Giáo viên hướng dẫn: **PGS. Phạm Văn hải**

HÀ NỘI 1 - 2019

# Phiếu giao nhiệm vụ đồ án tốt nghiệp

*1. Thông tin về sinh viên*

Họ và tên sinh viên: Trần Văn Thành

Điện thoại liên lạc: 0978779701

Email: thanhtranvan.hust@gmail.com

Lớp: CNTT2.01 - K58 Hệ đào tạo: Đại học chính quy

Đồ án tốt nghiệp được thực hiện tại: Viện CNTT&TT - Đại học Bách Khoa Hà Nội

Từ ngày 31/ 08/ 2018 đến ngày 28/ 12/ 2018

*2. Mục đích nội dung của ĐATN*

Phần 1: Thuật toán K-means clustering và ứng dụng cho bài toán phân vùng ảnh.

Phần 2: Phát hiện đám cháy trong ảnh và video thời gian thực.

*3. Các nhiệm vụ cụ thể của ĐATN*

1. Tìm hiểu và cài đặt các ứng dụng của K-means cho các bài toán phân loại chữ số viết tay, nén ảnh và phát hiện bất thường.
2. Cải tiến thuật toán K-means clusstering, áp dụngcho bài toán phân vùng ảnh màu, đánh giá và so sánh kết quả với các công trình nghiên cứu khác.
3. Xử lý ảnh phát hiện đám cháy trong video thời gian thực.

*4. Lời cam đoan của sinh viên*

Tôi – Trần Văn Thành - cam kết các kết quả nêu trong ĐATN là trung thực, không phải là sao chép toàn văn của bất kỳ công trình nào khác.

|  |  |
| --- | --- |
|  | *Hà Nội, ngày tháng năm 2019*  *Trần Văn Thành* |

*5. Xác nhận của giáo viên hướng dẫn về mức độ hoàn thành và cho phép bảo vệ*

|  |  |
| --- | --- |
|  | *Hà Nội, ngày tháng năm 2019*  *PSG. Phạm Văn Hải* |

# Lời cảm ơn

Lời đầu tiên, em xin chân thành cảm ơn tới thầy Phạm Văn Hải giảng viên bộ môn HTTT và thầy Đỗ Phan Thuận giảng viên bộ môn KHMT đã chấp nhận đề tài và tận tình chỉ bảo em hoàn thành đồ án này. Thầy đã cung cấp kiến thức, hướng dẫn tận tình, chỉ bảo cho em từ những bước đầu tiếp cận đề tài và trong suốt quá trình thực hiện đề tài này. Qua sự hướng dẫn của thầy em không chỉ học được nhiều kiến thức bổ ích mà còn học được những phương pháp nghiên cứu khoa học, thái độ và tinh thần nghiêm túc hết mình trong công việc.

Mặc dù em đã có những cố gắng hoàn thành đề tài này, xong do thời gian có hạn nên em cũng không thể tìm hiểu kỹ và đầy đủ hết các khía cạnh mà đề tài đề cập đến. Trong thời gian sắp tới em rất mong nhận được sự chỉ bảo tận tình của thầy hơn nữa để giúp em có thể hoàn thiện được đề tài hơn và áp dụng đề tài vào thực tế để đem lại lợi ích thiết thực theo đúng định hướng và mục tiêu mà đề tài đã đưa ra.

Em xin trân trọng cảm ơn thầy!

Mục lục

[Phiếu giao nhiệm vụ đồ án tốt nghiệp 1](#_Toc528541201)

[Lời cảm ơn 2](#_Toc528541202)

[Danh sách các từ viết tắt 4](#_Toc528541203)

[Danh sách các ký hiệu và thuật ngữ dùng trong đồ án 5](#_Toc528541204)

[Danh sách bảng 6](#_Toc528541205)

[Danh sách hình vẽ 7](#_Toc528541206)

[Giới thiệu chung 8](#_Toc528541207)

[Chương 1. Tổng quan đề tài 9](#_Toc528541208)

[1.1. Giới thiệu đề tài 9](#_Toc528541209)

[1.2. Các công việc liên quan 9](#_Toc528541210)

[1.3. Đóng góp của đồ án 11](#_Toc528541211)

[Chương 2. Kiến thức cơ sở về RNA 12](#_Toc528541212)

[2.1. Cấu trúc bậc hai 12](#_Toc528541213)

[2.2. Các bài toán liên quan 13](#_Toc528541214)

[Chương 3. Kiến thức cơ sở về học sâu 14](#_Toc528541215)

[3.1. Tổng quan về học sâu 14](#_Toc528541216)

[3.2. Convolutional network 17](#_Toc528541217)

[Chương 4. Xây dựng mô hình 19](#_Toc528541218)

[4.1. Biểu diễn dữ liệu 19](#_Toc528541219)

[4.2. Xây dựng mô hình học sâu 20](#_Toc528541220)

[Chương 5. Kết quả thử nghiệm 23](#_Toc528541221)

[5.1. Mô tả dữ liệu 23](#_Toc528541222)

[5.2. Kết quả của 5-fold cross validation 25](#_Toc528541223)

[5.3. Kết quả trên dữ liệu test 28](#_Toc528541224)

[Chương 6. Kết luận và công việc tương lai 30](#_Toc528541225)

[Tài liệu Tham Khảo 31](#_Toc528541226)

# Danh sách các từ viết tắt

|  |  |
| --- | --- |
| RNA | Ribonucleic acid |
| A | Adenine |
| C | Cytosine |
| G | Guanine |
| U | Uracil |
| miRNA | microRNA |
| pre-miRNA | Precursor microRNA |
| ConvNet | Convolutional network |
| CNN | Convolutional neural netwok |
| LSTM | Long short term memory |

# Danh sách các ký hiệu và thuật ngữ dùng trong đồ án

|  |  |
| --- | --- |
| L | Chiều dài của chuỗi RNA |
| human | Tập dữ liệu RNA của con người |
| cross-species | Tập dữ liệu RNA của tập hợp nhiều loài |
| new | Tập dữ liệu RNA của tập hợp nhiều loài mới |
| pairing matrix | Ma trận ghép nối |

# Danh sách bảng

[Bảng 1. Thống kê của các tập dữ liệu 24](#_Toc528541227)

[Bảng 2. Kết quả trên tập dữ liệu human 26](#_Toc528541228)

[Bảng 3. Kết quả trên tập dữ liệu cross-species 27](#_Toc528541229)

[Bảng 4. Kết quả trên tập dữ liệu new 27](#_Toc528541230)

# Danh sách hình vẽ

[Hình 1. Cấu trúc bậc hai và cách biểu diễn dữ liệu 12](#_Toc528541235)

[Hình 2. Các bài toán liên quan miRNA 13](#_Toc528541236)

[Hình 3. Ví dụ một mạng nơron đơn giản 14](#_Toc528541237)

[Hình 4. Lịch sử deep learning (Nguồn: Deep Learning 101) 15](#_Toc528541238)

[Hình 5. Thống kê về số các bài báo nộp vào hội nghị ICML 2018 16](#_Toc528541239)

[Hình 6. Mạng Densely [13] 16](#_Toc528541240)

[Hình 7. Mạng Seq2seq [12] 17](#_Toc528541241)

[Hình 8. Ví dụ mạng nơron convolutional 18](#_Toc528541242)

[Hình 9. Mô hình ConvNet đề xuất 22](#_Toc528541243)

[Hình 10. Thống kê dữ liệu 24](#_Toc528541244)

[Hình 11. Biểu đồ các chỉ số đánh giá trên bộ dữ liệu 5-fold và training 29](#_Toc528541245)

# CHƯƠNG I: GIỚI THIỆU

# 1.1. Tổng quan

Những năm gần đây, AI - Artificial Intelligence (Trí Tuệ Nhân Tạo), và cụ thể hơn là Machine Learning (Học Máy) nổi lên như một bằng chứng của cuộc cách mạng công nghiệp lần thứ tư 4.0 (1 - động cơ hơi nước, 2 - năng lượng điện và động cơ đốt trong, 3 – khoa học máy tính và internet). Trí Tuệ Nhân Tạo đang xuất hiện trong mọi lĩnh vực của đời sống mà có thể chúng ta không nhận ra. Xe tự lái của Google và Tesla, hệ thống tự gắng thẻ (tag) khuôn mặt người dùng của Facebook, trợ lý ảo Siri của Apple, hệ thống gợi ý sản phẩm của Amazon, hệ thống gợi ý phim của Netflix, máy chơi cờ vây AlphaGo của Google DeepMind, …, đó chỉ là một vài trong vô vàn những ứng dụng của AI/Machine Learning.

Học máy (machine learning - ML) là một tập con của AI. Nói đơn giản, Machine Learning là một lĩnh vực nhỏ của Khoa Học Máy Tính, nó có khả năng tự học hỏi dựa trên dữ liệu đưa vào mà không cần phải được lập trình cụ thể.

Sự phát triển của các hệ thống tính toán cùng với đó, lượng dữ liệu khổng lồ được thu thập bởi các hãng công nghệ lớn đã giúp Machine Learning tiến thêm một bước dài. Một lĩnh vực mới được ra đời gọi là học sâu (deep learning - DL). Deep Learning đã giúp máy tính thực thi những việc tưởng chừng như không thể vào 10 năm trước: phân loại cả ngàn vật thể khác nhau trong các bức ảnh, tự tạo chú thích cho ảnh, bắt chước giọng nói và chữ viết của con người, giao tiếp với con người, hay thậm chí cả sáng tác văn hay âm nhạc.

Lĩnh vực machine learning và deep learning là cực kỳ rộng lớn và có nhiều nhánh nhỏ. Chúng ta luôn nhớ rằng điều đơn giản nhất là điều quan trọng nhất. Khi bắt tay vào giải quyết một bài toán machine learning hay bất cứ một bài toán nào, chúng ta nên bắt đầu từ những thuật toán đơn giản nhất. Ta không nên nghĩ rằng chỉ có những thuật toán phức tạp mới có thể giải quyết được vấn đề. Những thuật toán phức tạp thường yêu cầu độ tính toán cao và nhạy cảm với cách chọn tham số đầu vào. Thay vào đó, những thuật toán đơn giản giúp chúng ta sớm có một mô hình tổng quá cho bài toán. Các thuật toán đơn giản thường được gọi là baseline, sẽ giúp chúng ta có cái nhìn ban đầu về sự phức tạp của mỗi bài toán. Việc cải thiện kết quả sẽ được dần cải thiện ở các bước sau.

Từ những lý do trên, em quyết định sử dụng một trong những thuật toán cơ bản nhất trong Unsupervised learning - thuật toán K-means clustering (phân cụm K-means) để giải quyết bài toán.

**1.2. Mục đích đồ án**

Nghiên cứu về bài toán phân vùng ảnh (Image Segmentation) và ứng dụng của thuật toán phân cụm K-means để áp dụng nó vào giải quyết các bài toán xử lý ảnh trong thực tế. Đồng thời điều chỉnh các tham số thích hợp cho mô hình để đạt được kết quả có độ chính xác tốt nhất.

**1.3. Phát biểu bài toán**

Phần 1: Xử lý ảnh vệ tinh để tính toán được sự thay đổi của đô thị hóa, hay sự sói mòn đất, nạn chặt phá rừng, mức độ thiệt hại của thiên tai. Theo phương pháp truyền thống từ xưa dùng các bản vẽ, dựa vào bản đồ và các chuyến đi thực nghiệm, nhưng không phải nơi đâu trên trái đất con người cũng có thể đặt chân đến để kiểm tra đo đạc. Vì vậy ngày nay với các công cụ khoa học hiện đại, từ vệ tinh con người có thể thu được các hình ảnh từ bất cứ nơi đâu trên trái đất. Qua các ảnh vệ tinh chụp được một khu rừng, sau khi xử lý phân vùng ảnh thành các vùng như cây, đất, nước… Ta tính được tỷ lệ phần trăm diện tích của các vùng trong tấm ảnh, qua đó tính được diện tích các vùng, và so sánh kết quả từ đó kết luận được độ che phủ rừng, hay sự đô thị hóa… thay đổi qua các mốc thời gian.

Bài toán đặt ra: xây dựng mô hình học máy cho bài toán phân vùng ảnh vệ tinh (Image Segmentation) sử dụng thuật toán K-means clustering và một số ứng dụng cho các bài toán thực tế như : phân nhóm chữ số viết tay, tách vật thể trong ảnh, nén ảnh.

Phần 2: Phát hiện đám cháy trong ảnh và video thời gian thực. Hiện nay, các hệ thống báo cháy tự động trong các tòa nhà cao tầng, văn phòng,… là các cảm biến khói – thiết bị dùng để phát hiện khói – một thành phần khởi đầu của sự cháy. Có 2 loại cảm biến khói là quang điện và ion. Các cảm biến này thưởng gây ra các hiện tượng báo cháy giả và có thể làm tốn kém tiền của, khiến người dùng thiếu đi sự tin tưởng, dẫn đến hậu quả nghiêm trọng.

Bài toán đặt ra: Tách vùng cháy trong ảnh, phát hiện đám cháy trong video thời gian thực lấy từ các camera báo cháy, kết hợp với các cảm biến khói để giảm thiểu báo cháy giả.

**1.4. Dữ liệu đầu vào**

Tập dữ liệu MNIST chứa 70 000 hình ảnh về chữ số viết tay. Chi tiết về dữ liệu đầu vào ta có thể tham khảo mục 3.1 trong đồ án.

Tập dữ liệu ảnh vệ tinh được thu thập từ <https://earthengine.google.com/timelapse/>

**1.5. Dữ liệu đầu ra**

Đối với bài toán phân vùng ảnh vệ tinh: ảnh đầu ra sau khi được phân vùng sẽ được gán nhãn và tỷ lệ % diện tích các vùng ảnh, đánh giá và so sánh kết quả mô hình.

Với bài toán phát hiện đám cháy trong ảnh và video realtime: đầu ra sẽ là 1 ảnh hoặc tập các ảnh là các frame được cắt từ video đầu vào sau đó được gán nhãn là cháy hoặc không cháy.

**1.6. Công cụ lựa chọn**

Thuật toán sử lý ảnh được sử dụng để giải quyết bài toán là thuật toán phân cụm K-means, được cài đặt và triển khai trên môi trường hệ điều hành Windows. Ngôn ngữ để triển khai bài toán là python v2.7 và matlab v2015b cùng thư viện sử dụng là OpenCV, thư viện này sẽ giúp ta triển khai các hàm toán học và xử lý ảnh một cách dễ dàng hơn.

**1.7. Tóm tắt nội dung đồ án**

Đồ án bao gồm phụ lục và 5 chương:

Chương 1: Giới thiệu: Chương này trình bày về nhiệm vụ cần thực hiện trong khuôn khổ đồ án tốt nghiệp, trình bày định hướng giải quyết các vấn đề đặt ra. Giới thiệu tóm tắt về cơ sở lý thuyết và công cụ được lựa chọn để giải quyết các vấn đề đặt ra.

Chương 2: Unsupervised Learning và K-means Clustering: Chương này trình bày những lý thuyết cơ bản về Unsupervised Learning , cung cấp một cách nhìn tổng quát, và những vấn đề về học không giám sát, thuật toán phân cụm K-means Clustering.

Chương 3: Image Segmentation: Chương này chủ yếu trình bày việc áp dụng thuật toán phân cụm K-means cho bài toán phân vùng ảnh, và một số bài toán khác như phân loại chữ số viết tay, tách vật thể trong ảnh, nén ảnh, phát hiện bất thường.

Chương 4: Phát hiện đám cháy trong ảnh và video realtime: Chương này tình bày các quy tắc để phát hiện 1 ảnh có chứa đám cháy hay không và áp dụng cho bài toán thực tế là phát hiện đám cháy trong video thời gian thực qua webcam trên laptop.

Chương 5: Kết luận và định hướng phát triển: Ở chương này, chúng ta đưa ra kết luận tổng kết các kết quả ban đầu, khó khăn và thuận lợi của đồ án. Các định hướng phát triển tiếp theo.

Phụ lục: Ở phần này sẽ nói về cách cài đặt, chạy chương trình giải quyết bài toán.

**Chương 2: Unsupervised Learning và K-means Clustering**

**2.1. Học không giám sát - Unsupervised Learning**

Supervised learning (học có giám sát)

Chúng ta cần nhắc lại về học có giám sát Supervised learning. Như chúng ta đã biết Supervised learning là thuật toán dự đoán đầu ra (outcome) của một dữ liệu mới (new input) dựa trên các cặp (input, outcome) đã biết từ trước. Cặp dữ liệu này còn được gọi là (data, label), tức (dữ liệu, nhãn). Supervised learning là nhóm phổ biến nhất trong các thuật toán Machine Learning. Về mặt toán học, Supervised learning là khi chúng ra có một tập hợp biến đầu vào **X =**  và một tập hợp nhãn tương ứng **Y =** , trong đó **,**  là các vector. Các cặp dữ liệu biết trước được gọi là tập *training data* (dữ liệu huấn luyện). Từ tập traing data này, chúng ta cần tạo ra một hàm số ánh xạ mỗi phần tử từ tập **X** sang một phần tử (xấp xỉ) tương ứng của tập **Y**:

**f(), i 1, 2, …, N**

Mục đích là xấp xỉ hàm số **f** thật tốt để khi có một dữ liệu **x** mới, chúng ta có thể tính được nhãn tương ứng của nó **y = f(x)**.

Thuật toán supervised learning còn được tiếp tục chia nhỏ ra thành hai loại chính:

Classification (Phân loại)

Một bài toán được gọi là *classification* nếu các *label* của *input data* được chia thành một số hữu hạn nhóm. Với **y** chỉ nhận giá trị từ một tập rời rạc, chẳng hạn {cá, cây, quả, mèo}. Ví dụ: Gmail xác định xem một email có phải là spam hay non\_spam.

Regression (Hồi quy)

Nếu *label* không được chia thành các nhóm mà là một giá trị thực cụ thể. Ví dụ: một căn nhà rộng **x** , có **y** phòng ngủ và cách trung tâm thành phố **z km** sẽ có giá là bao nhiêu? Với **y** nhận giá trị là số thực.

Unsupervised Learning (Học không giám sát)

Trong thuật toán này, chúng ta không biết được *outcome* hay nhãn *label* mà chỉ có dữ liệu đầu vào. Thuật toán unsupervised learning sẽ dựa vào cấu trúc của dữ liệu để thực hiện một công việc nào đó, ví dụ như phân nhóm (clustering) hoặc giảm số chiều của dữ liệu (dimension reduction) để thuận tiện trong việc lưu trữ và tính toán.

Về mặt toán học, Unsupervised learning có nghĩa là khi chúng ta chỉ có dữ liệu vào **X** mà không biết nhãn **Y** tương ứng. Máy tính cần học một hàm **y = f(x)** từ tập học cho trước . Y có thể là các cụm dữ liệu hoặc các cấu trúc ẩn.

Những thuật toán loại này được gọi là Unsupervised learning vì không giống như Supervised learning, chúng ta không biết câu trả lời chính xác cho mỗi dữ liệu đầu vào. Giống như khi ta học, không có thầy cô giáo nào chỉ cho ta biết đó là chữ A hay chữ B. Cụm không giám sát được đặt tên theo nghĩa này.

Các bài toán Unsupervised learning được tiếp tục chia nhỏ thành ba loại:

Clustering (phân nhóm)

Một bài toán phân nhóm toàn bộ dữ liệu X thành các nhóm nhỏ dựa trên sự liên quan giữa các dữ liệu trong mỗi nhóm. Ví dụ: phân nhóm khách hàng dựa trên hành vi mua hàng. Điều này cũng giống như việc ta đưa cho một đứa trẻ rất nhiều mảnh ghép với các hình thù và màu sắc khác nhau, ví dụ tam giác, vuông, tròn với màu xanh và đỏ, sau đó yêu cầu trẻ phân chúng thành từng nhóm. Mặc dù không cho trẻ biết mảnh nào tương ứng với hình nào hoặc màu nào, nhiều khả năng chúng vẫn có thể phân loại các mảnh ghép theo màu hoặc hình dạng.

Community detection: phát hiện các cộng đồng trong mạng xã hội.

Trends detection: phát hiện xu hướng thị yếu.

**2.2. K-means Clustering**

**2.2.1. Giới thiệu**

Trong thuật toán K-means clustering, chúng ta không biết nhãn *label* của từng điểm dữ liệu. Mục đích là làm thế nào để phân dữ liệu thành các cụm *cluster* khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau.

Ý tưởng đơn giản nhất về *cluster* là tập hợp các điểm ở gần nhau trong một không gian nào đó (không gian này có thể có rất nhiều chiều trong trường hợp thông tin về một điểm dữ liệu là rất lớn). Hình bên dưới là một ví dụ về 3 cụm dữ liệu trong không gian 2 chiều.



**2.2.2. Phân tích toán học**

Mục đích cuối cùng của thuật toán phân nhóm này là: từ dữ liệu đầu vào và số lượng nhóm chúng ta muốn tìm, hãy chỉ ra *center* của mỗi nhóm và phân các điểm dữ liệu vào các nhóm tương ứng. Giả sử rằng mỗi điểm dữ liệu chỉ thuộc vào đúng một nhóm.

Giả sử có **N** điểm dữ liệu là **X =**  và **K N** là số cluster chúng ta muốn phân chia. Chúng ta cần tìm các center **, , …,** và label của mỗi điểm dữ liệu.

Với mỗi điểm dữ liệu đặt **=** là label vector của nó, trong đó nếu được phân vào *cluster* **k** thì **= 1** và **= 0**, . Điều này có nghĩa là có đúng một phần tử của vector là bằng 1 (tương ứng với cluster của ), các phần tử còn lại bằng 0. Ví dụ: nếu một điểm dữ liệu có *label* vector là thì nó thuộc vào *cluster* 1, nếu là thì nó thuộc vào *cluster* 2, .. Ràng buộc của có thể viết dưới dạng toán học như sau:

Hàm mất mát và bài toán tối ưu

Nếu ta coi center là center (hoặc representative) của mỗi *cluster* và ước lượng tất cả các điểm được phân vào *cluster* này bởi , thì một điểm dữ liệu được phân vào cluster ***k*** sẽ bị sai số là . Ta mong muốn sai số này có trị tuyệt đối nhỏ nhất nên sẽ tìm cách để đại lượng sau đây đạt giá trị nhỏ nhất:

Hơn nữa, vì được phân vào *cluster* ***k*** nên **= 1, = 0,** . Khi đó, biểu thức bên trên sẽ được viết lại là:

Sai số cho toàn bộ dữ liệu sẽ là:

Trong đó **Y =** , **M =** lần lượt là các ma trận được tạo bởi *label* vector của mỗi điểm dữ liệu và *center* của mỗi *cluster*. Hàm số mất mát trong bài toán K-means clustering là hàm với ràng buộc như được nêu trong phương trình .

Ta cần tối ưu bài toán sau:

Ta cần tối ưu 2 biến số Y và M để hàm mất mát *đạt giá trị nhỏ nhất*.

Thuật toán tối ưu hàm mất mát

Bài toán là một bài toán khó tìm điểm tối ưu vì nó có thêm các điều kiện ràng buộc. Bài toán này thuộc loại *mix-integer programming* (điều kiện biến là số nguyên) - là loại rất khó tìm nghiệm tối ưu toàn cục (*global optimal point*, tức nghiệm làm cho hàm mất mát *đạt giá trị nhỏ nhất* có thể). Tuy nhiên, trong một số trường hợp ta vẫn có thể tìm được phương pháp để tìm được nghiệm gần đúng hoặc điểm cực tiểu. Điểm cực tiểu ở đây chưa chắc đã phải là điểm làm cho hàm số đạt giá trị nhỏ nhất (cực tiểu toàn cục), mà nó có thể là điểm cực tiểu địa phương (điểm cực tiểu cục bộ).

Một cách đơn giản để giải bài toán là xen kẽ giải Y và M khi biến còn lại được cố định làm tham số.

Cố định M và tìm Y

Giả sử đã tìm được các centers, hãy tìm các label vector để hàm mất mát đạt giá trị nhỏ nhất. Điều này tương đương với việc tìm cluster cho mỗi điểm dữ liệu.

Khi các centers là cố định, bài toán tìm label vector cho toàn bộ dữ liệu có thể được chia nhỏ thành bài toán tìm label vector cho từng điểm dữ liệu như sau:

Vì chỉ có một phần tử của *label* *vector* bằng 1 nên bài toán có thể tiếp tục được viết dưới dạng đơn giản hơn:

Vì chính là bình phương khoảng cách tính từ điểm đến *center* , ta có thể kết luận rằng mỗi điểm thuộc vào *cluster* có *center* gần nó nhất! Từ đó ta có thể dễ dàng suy ra *label vector* của từng điểm dữ liệu.

Cố định Y và tìm M

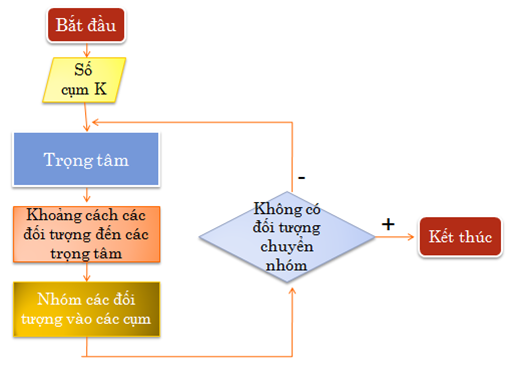
Giả sử đã tìm được cluster cho từng điểm, hãy tìm center mới cho mỗi cluster để hàm mất mát đạt giá trị nhỏ nhất. Một khi chúng ta đã xác định được label vector cho từng điểm dữ liệu, bài toán tìm center cho mỗi cluster được rút gọn thành:

Lúc này, ta có thể tìm nghiệm bằng phương pháp giải phương trình đạo hàm bằng 0, vì hàm cần tối ưu là một hàm liên tục và có đạo hàm xác định tại mọi điểm. Và quan trọng hơn, hàm này là hàm lồi (convex) theo nên chúng ta sẽ tìm được giá trị nhỏ nhất và điểm tối ưu tương ứng.

Giải phương trình đạo hàm bằng 0 ta có:

Ở đây mẫu số chính là phép đếm số lượng các điểm dữ liệu trong cluster ***j***. Còn tử số chính là tổng các điểm dữ liệu trong cluster ***j***. Nghĩa là chính là trung bình cộng của các điểm trong cluster ***j***.

**2.2.3. Thuật toán K-means**



Đầu vào: Dữ liệu X và số cụm cluster cần tìm K.

Đầu ra: Các center M và label vector cho từng điểm dữ liệu Y.

1. Chọn K điểm bất kỳ làm các center ban đầu.
2. Phân mỗi điểm dữ liệu vào cluster có center gần nó nhất.
3. Nếu việc gán dữ liệu vào từng cluster ở bước 2 không thay đổi so với vòng lặp trước nó thì ta dừng thuật toán.
4. Cập nhật center cho từng cluster bằng cách lấy trung bình cộng của tất các các điểm dữ liệu đã được gán vào cluster đó sau bước 2.
5. Quay lại bước 2.

Thuật toán sẽ dừng lại sau một số hữu hạn vòng lặp (luôn hội tụ sau một số hữu hạn lần lặp). Vì hàm mất mát là một số dương và sau mỗi bước 2 hoặc 3, giá trị của hàm mất mát bị giảm đi. Số lượng cách phân nhóm cho toàn bộ dữ liệu là hữu hạn nên đến một lúc nào đó, hàm mất mát sẽ không thể thay đổi, và khi đó thuật toán sẽ dừng.

**Chương 3: Image Segmentation**

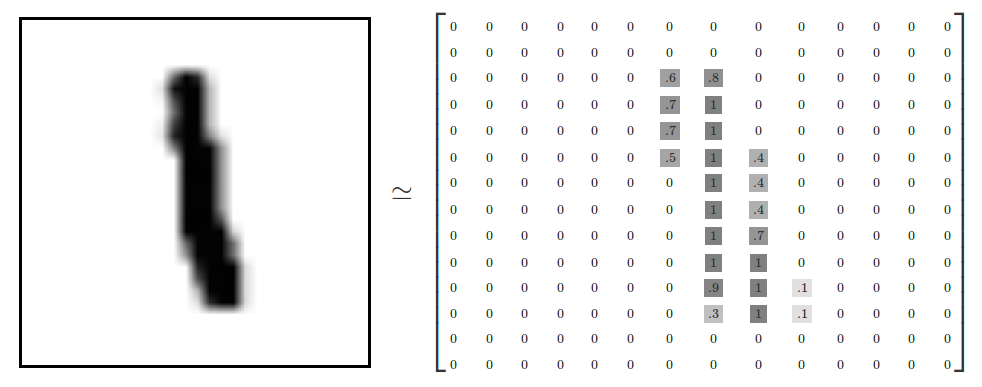
**3.1. Phân nhóm chữ số viết tay**

**3.1.1 Feature extraction**

Trước hết ta đi vào bài toán Image Classification cho bộ cơ sở dữ liệu chữ số viết tay MNIST gồm hai tập con: tập dữ liệu huấn luyện (*training\_set*) có tổng cộng 60000, và tập dữ liệu thử nghiệm (*test\_set*) có tổng cộng 10000 ví dụ khác nhau. Tất cả đều đã được gán nhãn.

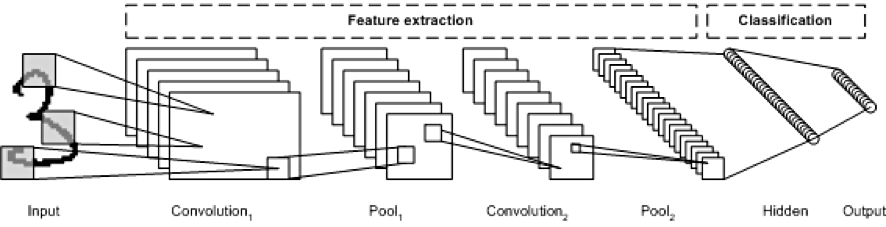


Mỗi bức ảnh là một ảnh đen trắng (có 1 channel), có kích thước 28x28 pixel. Khi đó mỗi bức ảnh sẽ là 1 vector 1x784, mà mỗi phần tử của vector là một pixel mang một giá trị là một số tự nhiên từ 0 đến 255. Các pixel màu đen có giá trị bằng 0, các pixel càng trắng thì có giá trị càng cao. Tất cả các chữ số được đặt trên nền đen, và chữ số màu sáng. Dưới đây là một ví dụ về chữ số 7 và giá trị các pixel của nó (ảnh nền đã được thay đổi thành nền trắng).



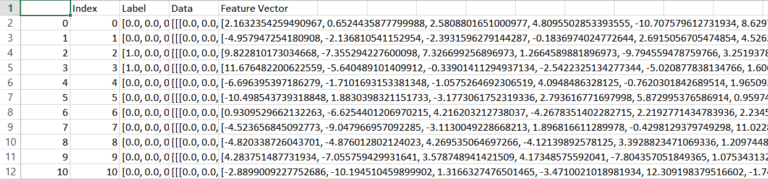
Ta coi mỗi bức ảnh là một điểm dữ liệu. Và mỗi điểm dữ liệu là 1 vector có kích thước 1x784 biểu diễn 1 bức ảnh chữ số.

Ở đây chúng ta sẽ sử dụng các vector đặc trưng được trích xuất bởi các mô hình CNN để phân cụm dữ liệu MNIST:



Sơ đồ kiến trúc mạng LeNet

Trích xuất đặc trưng từ ảnh thu được các vector đặc trưng và lưu vào file .csv có dạng như sau:



**3.1.2. K-Means clustering of MNIST data**

Dự đoán số cụm K

Trước khi áp dụng thuật toán K-means clustring cho bộ dữ liệu ta cần dự đoán số cụm K thích hợp.

Phương pháp khuỷu tay (Elbow): ý tưởng là đối với một phạm vi các giá trị thích hợp của K, ta tính sai số của tổng lỗi trong cụm (wcss - within-cluster sum of errors) để tìm ra giá trị K tối ưu nhất. Ta có thể tìm thấy K ở vị trí khuỷu tay, tức là trước khuỷu tay lỗi giảm nhanh dần và sau khi qua khuỷu tay lỗi giảm không đáng kể.

Phương pháp Average Silhouette: phương pháp này thể hiện tính nhất quán trong các *cluster*, giá trị *silhouette* là thước đo mức độ tương tự của một đối tượng với cụm (sự gắn kết) của chính nó so với các cụm khác (sự tách biệt) hay mức độ phân tách giữa các cụm. Silhouette nằm trong phạm vi −1 đến +1, trong đó giá trị cao biểu thị rằng đối tượng được kết hợp tốt với cụm của chính nó và tách rời với các cụm lân cận. Nếu hầu hết các đối tượng có giá trị cao, thì cấu hình phân cụm là phù hợp. Nếu nhiều điểm có giá trị thấp hoặc âm, thì cấu hình phân cụm có thể có quá nhiều hoặc quá ít cụm. Silhouette được tính như sau:

Tính khoảng cách trung bình từ tất cả các điểm dữ liệu trong cùng một cụm ().

Tính khoảng cách trung bình từ tất cả các điểm dữ liệu trong cụm gần nhất ().

Tính hệ số:

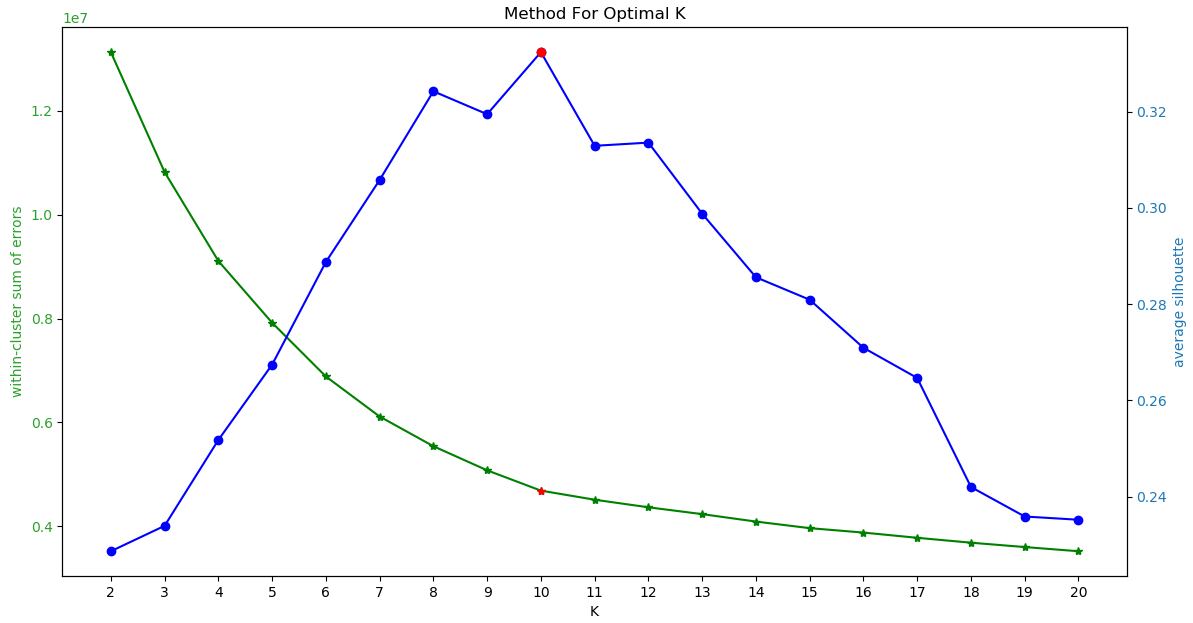
Hoặc là

Nếu mẫu i rất gần với các cụm lân cận.

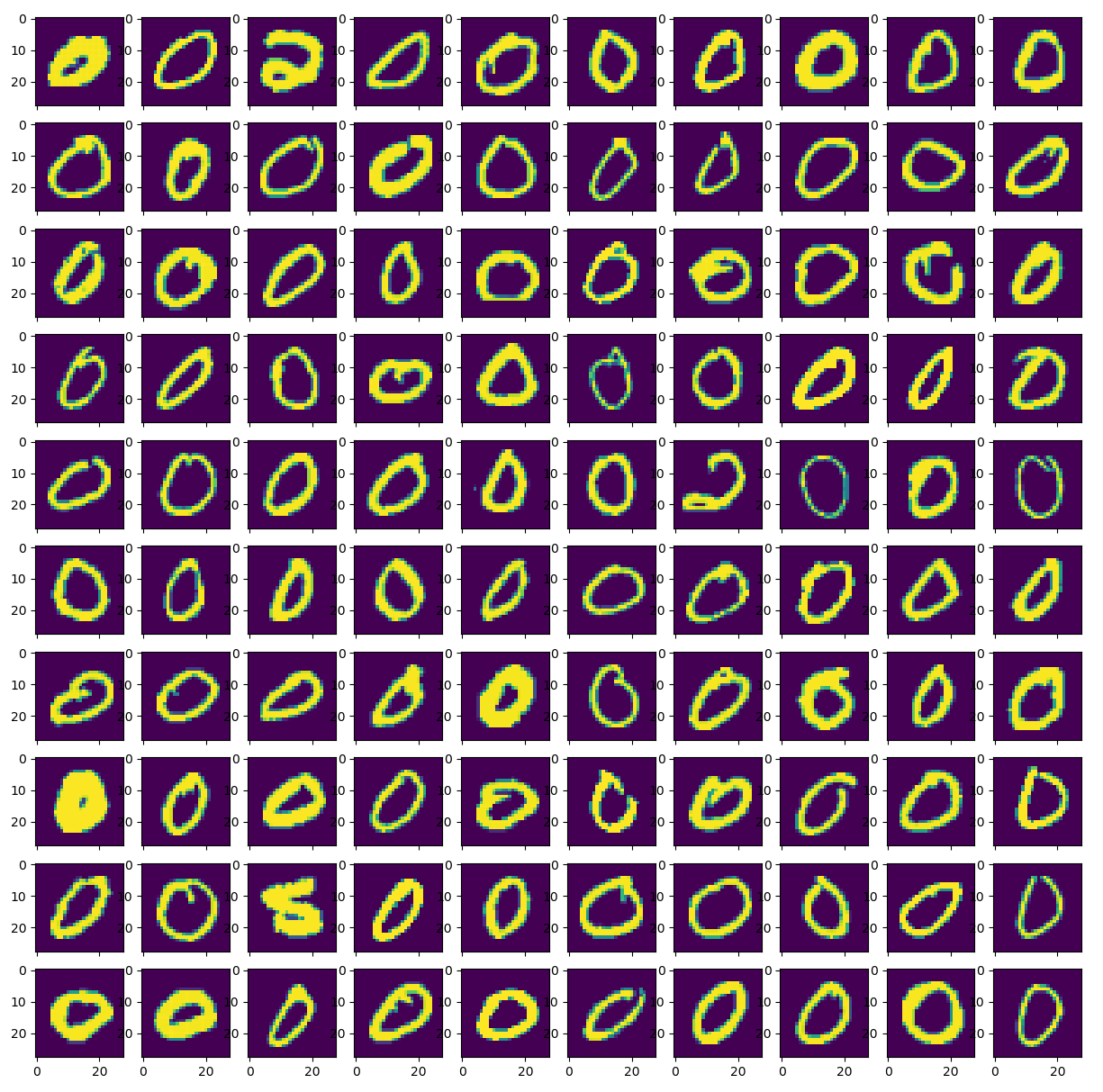
Nếu mẫu nằm cách xa các cụm lân cận.

Nếu mẫu được gán cho các cụm sai.

Dưới đây là kết quả khi thực hiện tối ưu số cụm K, sử dụng 2 phương pháp trên, kết quả khá rõ khi chọn số cụm K = 10 là tốt nhất.



Triển khai thuật toán K-means clustering cho bộ dữ liệu features matrix ta được kết quả:



Đánh giá độ chính xác

Trong đó: là số chữ số giống nhau nhiều nhất trong cụm *j*

K là số cụm, N tổng số bức ảnh, là số bức ảnh thuộc cụm *j*.

Tức là ta chia số lượng những chữ số được phân đúng (số lượng nhiều nhất trong 1 cụm) chia cho tổng các chữ số trong cụm đó. Trong hình trên, ta hiển thị 1 cụm có kích thước 100 chữ số, gồm có 97 chữ số 0 và 3 chữ số bị phân nhóm sai là 2 chữ số 2 và 1 chữ số 5. Như vậy độ chính xác của cụm này là 97%. Dưới đây là kết quả phân loại trên toàn bộ tập dữ liệu:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Accuracy (%) | 94.68 | 93.54 | 80.69 | 91.56 | 96.71 | 92.66 | 90.10 | 86.82 | 94.84 | 85.47 |
| Average (%) | 90.71 | | | | | | | | | |
| Time  (s) | 4.640 | | | | | | | | | |

Nhận thấy kết quả ở mức khá tốt khi thực hiện phân loại chữ số viết tay bằng K-meas clustering, so với các kết quả từ các bài báo khác sử dụng mô hình CNN sẽ có độ chính xác cao từ 95-98% với cùng 1 tập dữ liệu MNIST.

**3.2. Image Segmentation và cải tiến K-means**

**3.2.1. Bài toán phân vùng ảnh (image segmentation)**

**3.2.2. Cải tiến thuật toán K-means**

Ý tưởng cơ bản để phân đoạn hình ảnh bằng thuật toán phân cụm K-means là tạo các cụm dựa trên giá trị màu của mỗi pixel, vì thế độ chói (hay độ sáng) của mỗi pixel sẽ bị bỏ qua, ta chỉ quan tâm đến các thành phần về mầu của pixel. Chính vì thế nên trong phần này, ta áp dụng phân cụm K-means cho ảnh trên không gian mầu Lab. Trong không gian màu L\*a\*b ta loại bỏ thành phần về độ sáng L, và chỉ giữ lại 2 thành phần về mầu là kênh mầu a và b. Khi đó ta cẩn reshape lại kích thước của ma trận ảnh từ ma trận 3 chiều sang ma trận 2 chiều. Lúc này kích thước dữ liệu đầu vào giảm, giúp cho thuật toán chạy nhanh hơn và phân đoạn ảnh chính xác hơn. Công thức chuyển đổi từ hệ màu RGB sang hệ màu L\*a\*b:

Trong đó:

và

Hiệu năng của thuật toán K-means bị ảnh hưởng rất nhiều ở bước khởi tạo trung tậm K cụm đầu tiên. Tùy vào các center ban đầu mà thuật toán có thể có tốc độ hội tụ rất chậm, hoặc thậm chí cho chúng ta nghiệm không chính xác (chỉ là local minimum - điểm cực tiểu địa phương - mà không phải giá trị nhỏ nhất). Đó là một nhược điểm lớn của K-means cho bài toán phân đoạn ảnh nói riêng và phân cụm nói chung. Vì thế ta sẽ đi khắc phục nhược điểm này để có được khởi tạo trung tâm K ban đầu tốt nhất.

Một nhược điểm nữa ta sẽ cần khác phục đó là K-means cho kết quả không tốt đối với các hình ảnh bị nhiễu, hay có chất lượng thấp. Vì thế trước khi thực hiện phân cụm ảnh với K-means, ta sẽ nâng cao chất lượng ảnh đầu vào, tăng cường đặc trưng cho các điểm ảnh.

Công việc cải tiến phương pháp cho bài toán phân vùng ảnh màu sử dụng K-means clusstering:

1. Đọc ảnh màu RGB.
2. Chuyển đổi hình ảnh từ định dạng RGB sang Lab.
3. Định hình lại ảnh thành ma trận hai chiều. Vì ảnh ban đầu ở không gian 3D, ta chuyển về không gian 2D (Trong không gian L\*a\*b loại bỏ kênh độ sáng L, chỉ lấy 2 kênh màu a\* và b\*).
4. Áp dụng phương pháp Elbow và Silhouette để dự đoán số cụm K từ ma trận dữ liệu.
5. Sắp xếp ma trận dữ liệu hình ảnh theo thứ tự không giảm, chia toàn bộ dữ liệu hình ảnh được sắp xếp thành K phần (K đã được dự đoán ở bước 4).
6. Tìm giá trị trung bình của từng phần và giá trị trung bình đó trở thành trọng tâm cho cụm tương ứng và lưu lại trung tậm cụm ban đầu.
7. Áp dụng thuật toán phân cụm K-means cho ma trận dữ liệu hình ảnh với trung tâm cụm ban đầu và số cụm K đã dự đoán ở trên.

**3.2.3. Kết quả và đánh giá**

Các thuật toán phân cụm có thể khó đánh giá mà không có trước thông tin hoặc giả định về dữ liệu.

Nội dung đồ án bao gồm các phần chính:

# Tổng quan đề tài

## Giới thiệu đề tài

Axit ribonucleic (RNA) là các phân tử sinh học có liên quan đến nhiều quá trình sinh học. RNA được lắp ráp như một chuỗi các monome gọi là nucleotide. Có bốn loại nucleotide RNA phục vụ như là các khối xây dựng được sắp xếp tùy ý của chuỗi nucleotide: adenine (A), cytosine (C), guanine (G) và uracil (U). MicroRNA (miRNA) là một loại phân tử RNA không mã hóa, thường dài khoảng 20-23 nucleotide. Trong ba thập kỷ qua, chúng đã được phát hiện trong một số lượng lớn các sinh vật như con người [1] và thực vật [2], cũng như trong các virus [3]. Chúng liên kết với các RNA thông tin (mRNA) để ức chế sự dịch mã các mRNA thành các protein [4]. Tầm quan trọng của chúng trong việc điều hòa gen đóng một phần trong các bệnh như ung thư [5], [6], [7], và chúng là những mục tiêu tốt cho các dấu hiệu bệnh và trị liệu [7]. Do đó, việc xác định miRNA là một nhiệm vụ quan trọng trong điều trị y khoa. Nhưng rất khó để xác định miRNA trực tiếp bởi vì chúng rất ngắn. Hầu hết các nghiên cứu đều tập trung vào các phương pháp tính toán để xác định các miRNA tiền thân (pre-miRNAs). Các chuỗi chính được gọi là pri- miRNA được xử lý để tạo thành các pre-miRNA và sau đó đến các miRNA trưởng thành. Việc xác định các pre-miRNA dễ dàng hơn so với các miRNA vì các pre-miRNA dài hơn rất nhiều (khoảng 80 nucleotide) và chúng có cấu trúc vòng kẹp tóc với nhiều đặc trưng cấu trúc hơn. Nhận dạng pre-miRNA là một nhiệm vụ phân loại, cho ra kết quả “positive” hoặc “negative”. Nó có khả năng là một nhiệm vụ khó khăn vì số lượng lớn các chuỗi có thể được tạo ra bằng cách sử dụng 4 loại nucleotide A, C, G và U. Hơn nữa, số lượng các pre-miRNA đã được khám phá nhỏ hơn nhiều so với số lượng chuỗi RNA có cấu trúc kẹp tóc giả (tức là các RNA có cấu trúc vòng kẹp tóc tương tự với các pre-miRNA nhưng không chứa các miRNA trưởng thành). Do đó, chúng ta phải đối phó với một vấn đề mất cân bằng lớp (class inbalance). Mặt khác, RNA có nhiều tính năng cấu trúc và sinh học và chúng ta không biết cái nào thực sự cần thiết cho việc nhận dạng pre-miRNA. Do đó, học máy là một cách tiếp cận thích hợp để cân nhắc các tính năng tự động.

## Các công việc liên quan

Công việc trước đây về miRNA và nhận dạng pre-miRNA dựa trên các quy tắc thủ công (MIReNA [8])hoặc học máy. Phương pháp học máy ngày càng phổ biến trong thập kỷ qua và được chứng minh là hứa hẹn nhất, với các công cụ như HuntMi [9], miRBoost [10], CSHMM [11], microPred [12], miPred [13], triplet-SVM [14], Mirann [15], DP-miRNA [16], deepMiRGene [17]. Họ dự đoán cấu trúc bậc hai (tương tác ghép đôi cơ sở trong chuỗi RNA, ví dụ xem hình. 1AB) bằng các phương pháp chuẩn như RNAfold [18], Gtfold [19], và CyloFold [20]. Sau đó, họ trích xuất nhiều đặc trưng thủ công, chẳng hạn như số cặp nucleotide Watson-Crick (AU, CG), chiều dài vòng lặp [16]; [10]), độ dài chuỗi [16], tần số cặp dinucleotide [16] [21] [10] [12] [22], tần số cặp trinucleotide (cấu thành 64 tính năng) [16] [21], nhiệt độ nóng chảy [16], năng lượng tự do tối thiểu [23] [16] [24]. Các đặc trưng này được sử dụng làm đầu vào cho các phương pháp học máy như máy support vector machine (SVM) ( [10] [12] [14], random forest [21] , neural network [15] [16] [17] [21] và các mô hình hidden Markov model [11]. Ví dụ: Ref. [21] sử dụng 98 trong số các đặc trưng nói trên cho đầu vào của mạng nơron thiết kế bởi họ. DP-miRNA [16] sử dụng mạng nơron với 58 đặc trưng được trích xuất dựa trên thành phần trình tự và cấu trúc bậc hai được dự đoán bằng phần mềm RNAfold, các biện pháp gấp bao gồm các công thức khác nhau của năng lượng vật lý. Một trong những phương pháp giàu đặc trưng nhất là Ref. [24], sử dụng khoảng 900 đặc trưng. Các tác giả đã chỉ ra rằng yếu tố quan trọng nhất trong nhận dạng pre-miRNA là cấu trúc bậc hai. Chúng ta có thể thấy cấu trúc bậc hai là quan trọng và nhiều đặc trưng có thể được trích xuất từ ​​chúng. Phương pháp tối tân deepMiRGene [17] sử dụng mã hóa one-hot để biểu diễn cấu trúc bậc hai của RNA làm đầu vào cho mạng nơron. Vấn đề với biểu diễn này là thông tin chứa trong đó khá “vướng víu”: Không có mục nhập dữ liệu nào cho biết hai nucleotide {i, j} có được ghép nối với nhau hay không. Tìm hiểu xem chúng có được ghép nối hay không đòi hỏi phải phân tích cú pháp phần lớn toàn bộ biểu diễn cấu trúc bậc hai. Việc phân tích cú pháp này đạt được bằng cách huấn luyện một mạng nơron với memory và các cơ chế attention. Tuy nhiên, "gia công" như vậy một sự mất ý nghĩa thông tin có ý nghĩa cho việc học tập thường dẫn đến sự mất cân bằng tối ưu và một nhiệm vụ học tập tổng thể khó khăn hơn [25].Ưu điểm lớn nhất của deepMiRGene là chúng không cần bất kỳ đặc trưng thủ công nào [17]. Học sâu có thể trích xuất các đặc trưng ẩn của đầu vào rất tốt và có hiệu suất tốt trên nhiều nhiệm vụ như phân loại hình ảnh, nhận diện khuôn mặt và tiếng nói [26], dự đoán phân tử [27] [28], protein thứ cấp [29] và dự báo cấu trúc bậc ba [30], dự báo tiếp xúc protein [31]. Lấy cảm hứng từ thành công của việc học sâu và tầm quan trọng của cấu trúc bậc hai để nhận dạng pre-miRNA, chúng tôi đề xuất một phương pháp học sâu từ đầu đến cuối bằng cách sử dụng chuỗi đầu vào đã cho và cấu bậc hai của nó. Đối với dữ liệu chuỗi, mọi người thường sử dụng mạng nơron LSTM [32] vì chúng có thể xử lý thông tin nằm trong chuỗi một cách hiệu quả. Tuy nhiên, trong bài báo này, chúng ta sẽ sử dụng một mạng nơron tích chập (convolutional, ConvNet) để phân loại các chuỗi RNA có phải là pre-miRNA hay không. Chúng tôi sử dụng một tầng downscaling (downscaling layer) để củng cố các tính năng và cho phép đầu vào của ConvNet có kích thước biến thiên (variable-sized input) nhưng vẫn dẫn đến một dự đoán có kích thước lớp cố định. Chìa khóa của cách tiếp cận của chúng tôi là mã hóa cấu trúc bậc hai trong định dạng ma trận ghép nối (*pairing matrix*, xem [33] [34] [35] [36]). Trong *pairing matrix*, chúng tôi xác định năng lượng tự do tối thiểu của chuỗi và tương tác trực tiếp giữa các nucleotide. Kết quả cho thấy thuật toán của chúng tôi với kỹ thuật mới này hoạt động tốt hơn so với các phương pháp hiện đại khác trên các bộ dữ liệu tiêu chuẩn, cụ thể là *human*, *cross-species*, và *new* [17] [10]. Khi chúng tôi so sánh kết quả của chúng tôi khi sử dụng và không sử dụng *pairing matrix*, chúng tôi thấy rằng kết quả tệ hơn nhiều nếu không có *pairing matrix*.

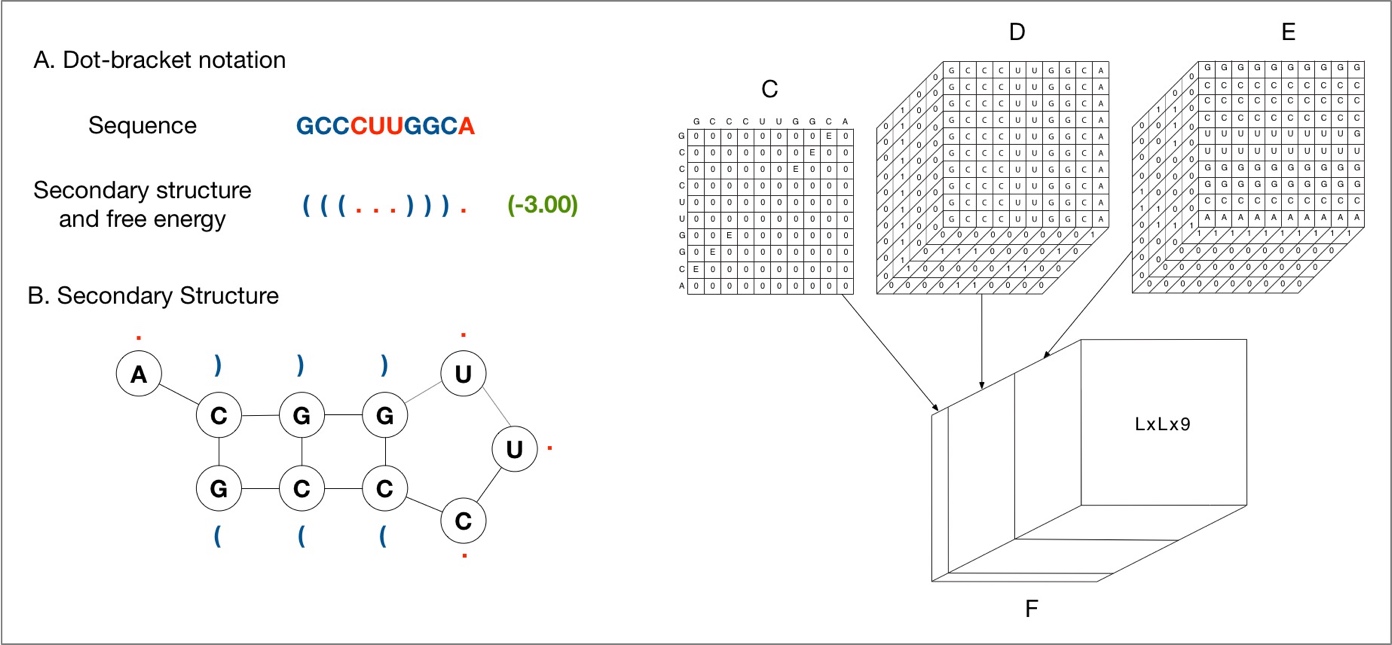
## Đóng góp của đồ án

* Một cách kết hợp nhiều kênh 2 chiều của chuỗi RNA, cấu trúc bậc hai, năng lượng giải phóng tối thiểu, ngoài ra không sử dụng đặc trưng được trích trọn bằng tay nào khác.
* Một mạng convolutional network thích hợp cho cách biểu diễn đầu vào trên và đạt được kết quả vượt trội so với các phương pháp khác.
* So sánh giữa hai cách biểu diễn đầu vào cho mạng ConvNet: variable-sized và fixed-sized.

# Kiến thức cơ sở về RNA

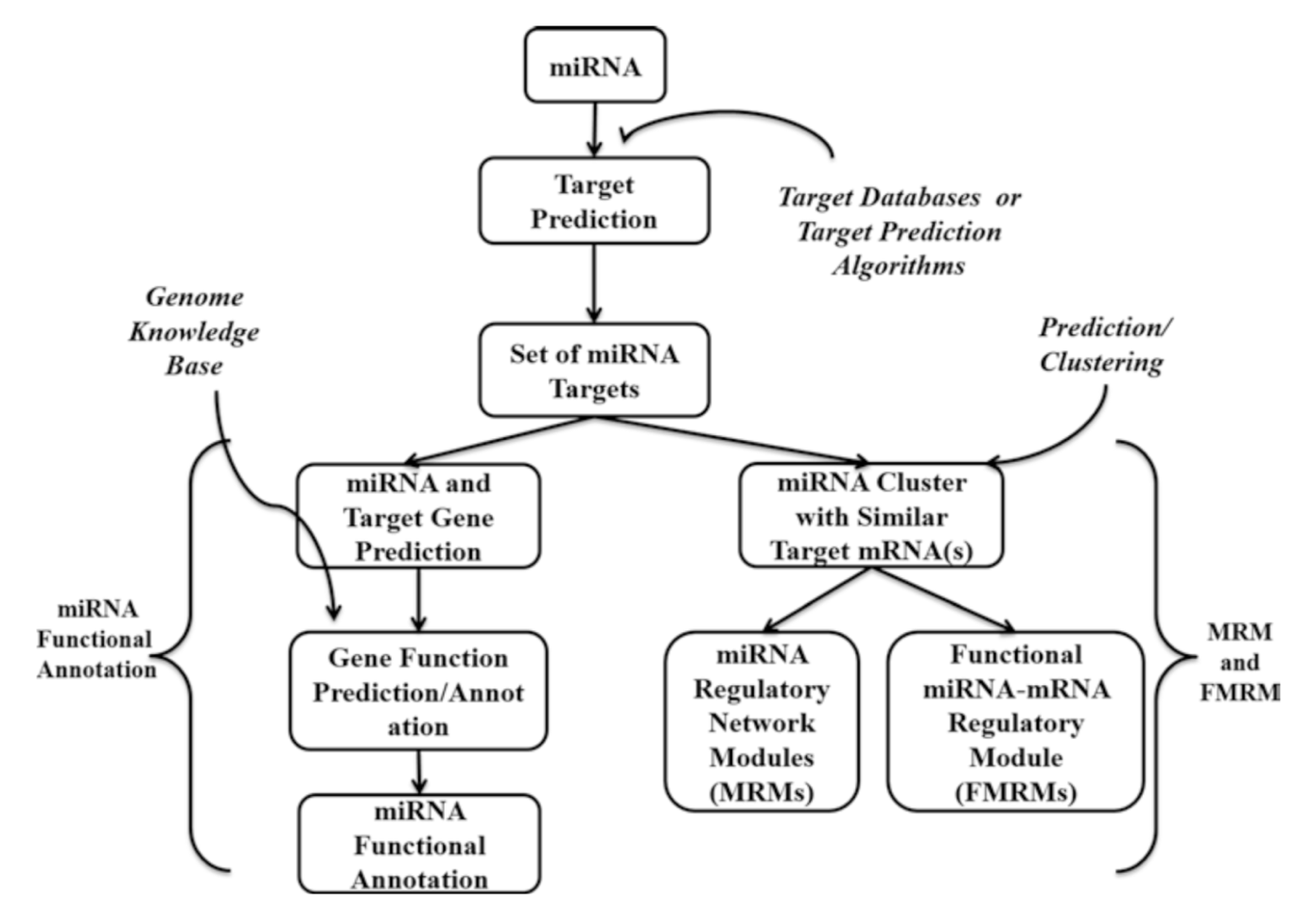
## Cấu trúc bậc hai

Cấu trúc bậc hai của một bản sao RNA biểu diễn các tương tác ghép nối cơ sở trong chuỗi RNA đó. Cấu trúc bậc hai thông thường của một pre-miRNA được thể hiện trong Hình 1AB, cho thấy rằng một pre-miRNA là một chuỗi xoắn kép ghép đôi chứ không phải là một sợi đơn, và sự ghép đôi này là một trong những đặc trưng nổi bật nhất cho nhận dạng pre-miRNA. Cấu trúc bậc hai của một chuỗi nhất định có thể được dự đoán bằng các công cụ được sử dụng rộng rãi như RNAfold. Nó xây dựng cấu trúc bậc hai dựa trên phương pháp ổn định nhiệt động lực học từ một chuỗi RNA đã cho bằng cách tính toán năng lượng tự do tối thiểu và các cặp cơ sở có thể xảy ra. Tuy nhiên, việc xác định pre-miRNA đáng tin cậy đòi hỏi các đặc trưng khác với cấu trúc thứ cấp được xem xét, vì các mẫu negative có thể bị nhận nhầm thành positive do các hạn chế của thuật toán dự đoán cấu trúc và khả năng đoán trước vốn có của các cấu trúc này.



Hình 1. Cấu trúc bậc hai và cách biểu diễn dữ liệu

## Các bài toán liên quan



Hình 2. Các bài toán liên quan miRNA

Ngoài nhận biết miRNA**,** bên cạnh miRNA, chúng ta còn rất nhiều bài toán quan trọng khác liên quan, ví dụ: dự đoán đích của miRNA (miRNA target prediction), dự đoán chức năng của miRNA (miRNA functional preduction) [37] Hình 2.

Dự đoán đích của miRNA là dự đoán đoạn mRNA mà miRNA sẽ bám vào để thực hiện chức năng điều hoà gen, quá trình này sẽ ngăn chặn quá trình hình thành protein. Vì thế, đó là một bài toán quan trọng, kết quả của nó tham gia vào nhiều quá trình sinh học và khám phá nhiều loại bệnh tật khác nhau.

Như đã nói ở trên, miRNA sẽ bám vào mRNA, ngăn chặn quá trình hình thành protein trong quá trình dịch mã. Quá trình ngăn chặn protein sẽ sinh ra nhiều loại bệnh tật, ví dụ như ung thư phổi, ung thư vú, vân vân. Vì vậy, bài toán dự đoán chức năng của miRNA cũng quan trọng không kém các bài toán kể trên.

Trong đồ án này, em sẽ tập trung vào giải quyết bài toán cốt lõi, làm nền tảng cho các bài toán khác liên quan đến miRNA, đó là nhận biết miRNA. Điểm nhấn trong đồ án này là cách biểu diễn chuỗi miRNA dưới dạng ma trận nhiều chiều dựa trên cấu trúc bậc hai của nó. Cách biểu diễn này có thể áp dụng cho các bài toán liên quan kể trên.

# Kiến thức cơ sở về học sâu

## Tổng quan về học sâu

Học sâu (Deep Learning) là một kỹ thuật trong Machine Learning, liên quan đến các thuật toán lấy cảm hứng từ cấu trúc và hoạt động của bộ não động vật gọi là mạng nơron nhân tạo (artificial neural networks). Mạng nơron đơn giản nhất bao gồm ba lớp: input, hidden, và output như sau:



Hình 3. Ví dụ một mạng nơron đơn giản

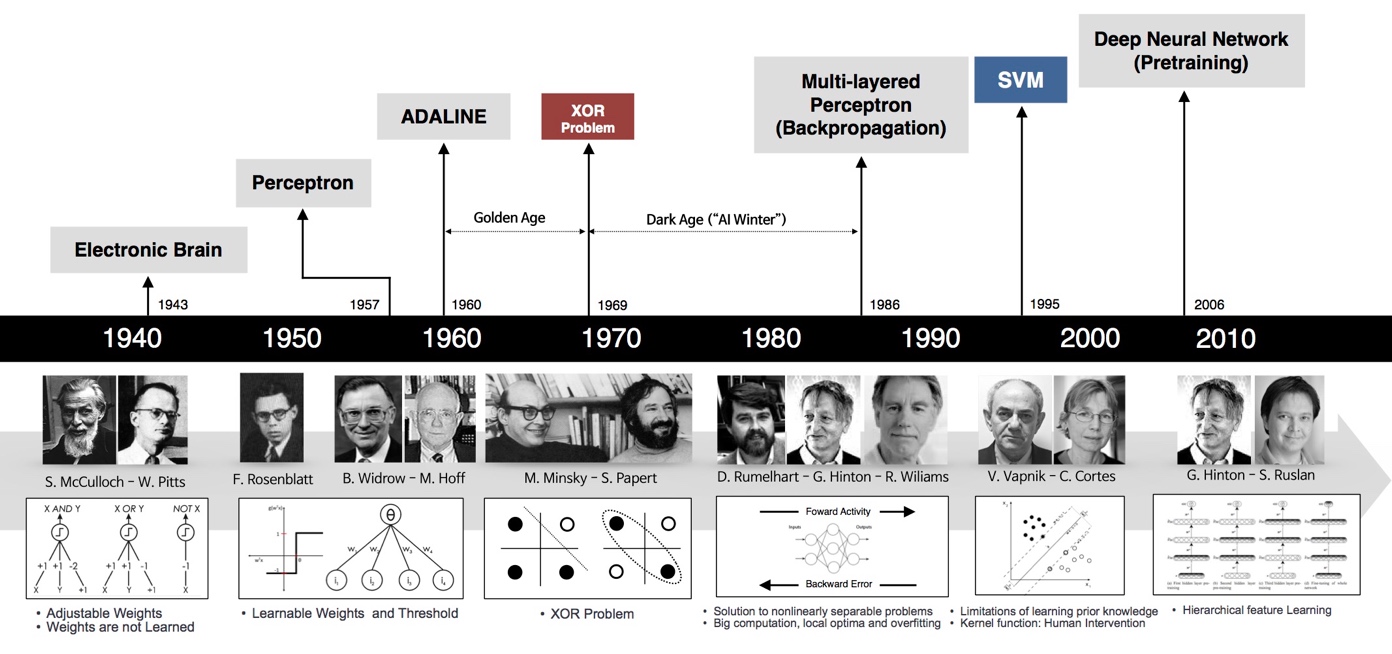
Mỗi lớp bao gồm nhiều nơron, được kết nối với nhau bởi các hàm actication, chẳng hạn hàm sigmoid, rectified linear, tanh.

,

,

,

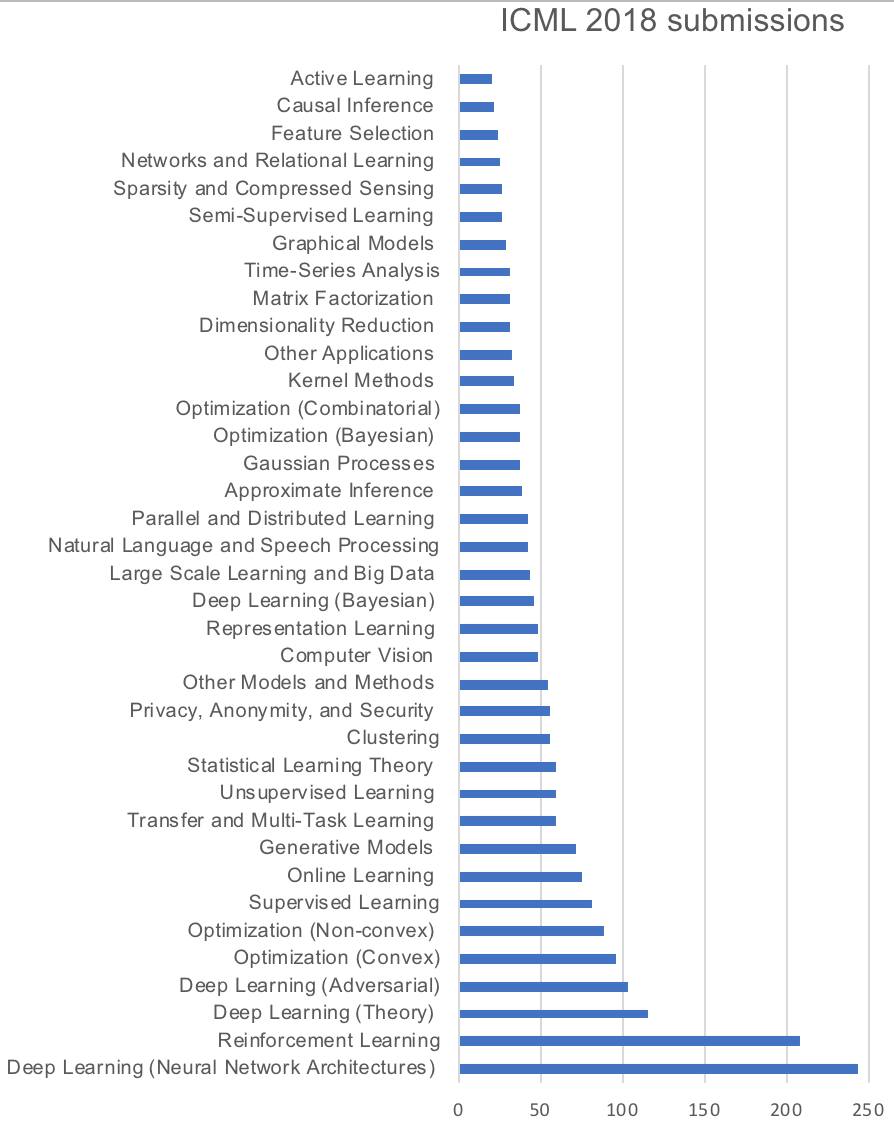
các hàm này có giá trị trả về nằm trong khoảng [0, 1], tượng trưng cho các kết nối, gọi là các khớp thần kinh (synapses), thường nối từ các axon tới các tế bào tua gai thần kinh (dendrite) trong bộ não con người.



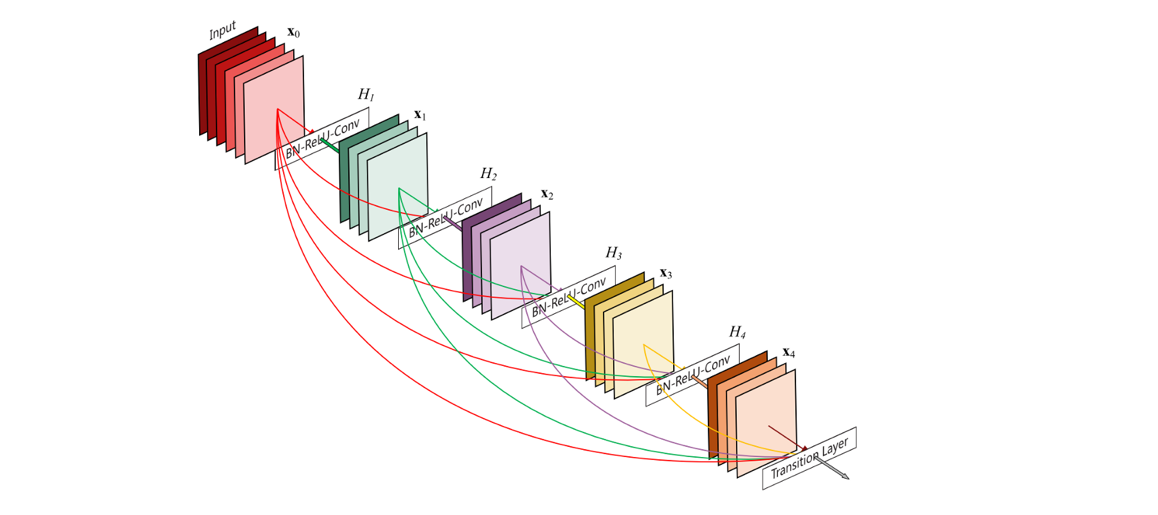
Hình 4. Lịch sử deep learning (Nguồn: Deep Learning 101)

Lịch sử phát triển của mạng nơron được thể hiện qua Hình 4. Từ những năm 40-50 của thế kỷ 20, mạng nơron được sáng tạo ra với tên là Electronic Brain và phát triển từ từ cho đến năm 2010. Trong vòng một thập kỷ gần đây, mạng nơron cùng với học sâu bùng nổ bởi khả năng tính toán của máy tính tăng cao, đặc biệt là tính toán hiệu năng cao trên GPU. Điều đó dẫn tới học sâu đang dần thống trị trong thế giới trí tuệ nhân tạo gần đây. Học sâu được các nhà nghiên cứu quan tâm nhiều hơn, điều đó được thể hiện số bài báo về học sâu được nộp vào hội nghị ICML 2018, một trong những hội nghị lớn nhất trên thế giới về học máy Hình 5.

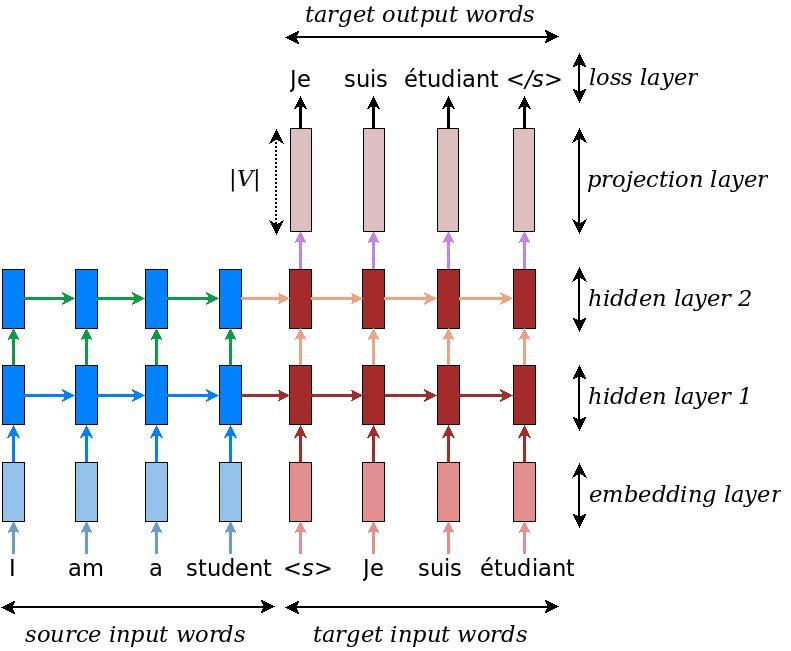
Ngày nay, xuất hiện nhiều mạng nơron phức tạp hơn (ví dụ: Densely network Hình 6, Hình 7), chúng bao gồm nhiều kỹ thuật mới, chẳng hạn convolutional, residuall block [38], long-short term memory (LSTM) [39].



Hình 5. Thống kê về số các bài báo nộp vào hội nghị ICML 2018



Hình 6. Mạng Densely [40]

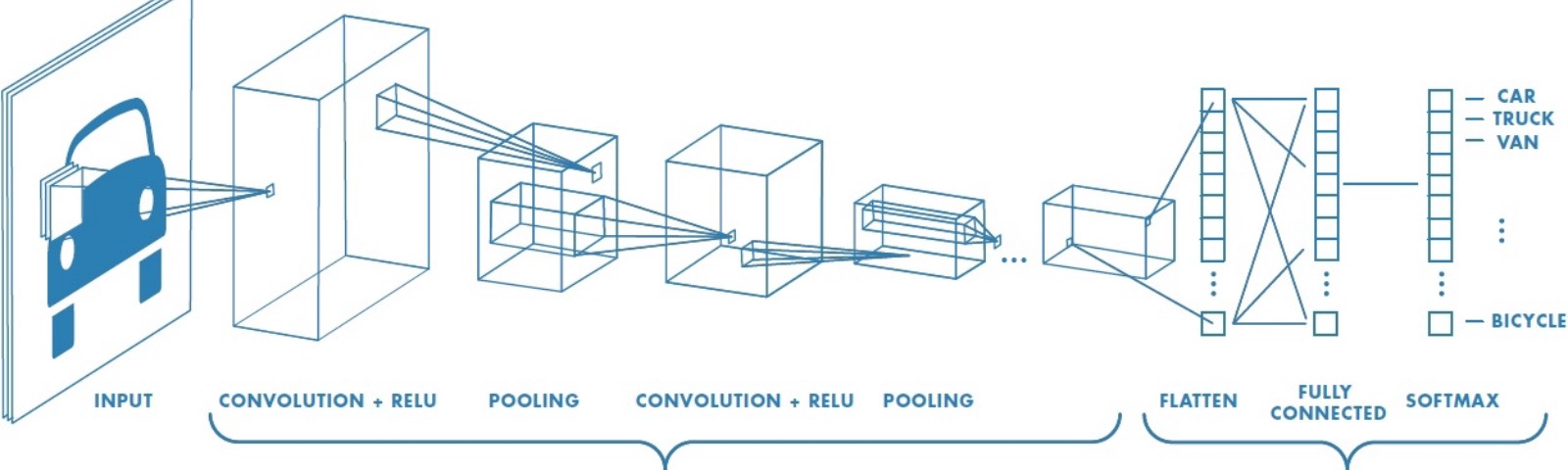


Hình 7. Mạng Seq2seq [39]

## Convolutional network

Ngày nay, các mạng nơron convolutional (CNN) là các mô hình học máy phổ biến nhất trong học sâu nói chung và khoa học nói riêng. CNN là các mạng chuyển tiếp được lấy cảm hứng từ sinh học bởi não người. CNN chứa nhiều lớp tích chập để nó có thể tổng quát hoá dữ liệu tốt. Bằng cách tự động trích xuất các đặc trưng cấp cao cần thiết để phân loại hoặc tập hợp, CNN đã đạt được nhiều kết quả ấn tượng trong nhiều tác vụ như nhận diện khuôn mặt, dịch máy, thử thách nhận diện vật thể Imagenet. Do đó, chúng tôi quyết định thể hiện chuỗi đầu vào dưới định dạng nhiều kênh và sau đó áp dụng CNN cho chúng.

Hình 8 là một ví dụ trực quan cho cấu trúc của một mạng convolutional cơ bản. Các khối 3D tượng trưng cho từng tầng convolution. Chúng trích trọn đặc trưng dựa trên cơ chế dùng cửa sổ trượt và hàm tích chập trong của sổ trượt đó. Của sổ trượt trượt toàn bộ không gian đặc trưng và biến không gian đặc trưng này thành không gian đặc trưng khác có số kênh (channel) nhỏ hơn hoặc lớn hơn, tuỳ thuộc vào từng bài toán và bộ nhớ tính toán. Trong các khối convolution, con người thường sử dụng thêm các tầng pooling để cô đọng số chiều của không gian đặc trưng đồng thời tăng tốc tốc độ tính toán.

****

Hình 8. Ví dụ mạng nơron convolutional

# Xây dựng mô hình

Cách tiếp cận của chúng tôi bao gồm hai bước. Thứ nhất, chúng tôi biểu diễn chuỗi RNA đã cho ở định dạng đa kênh 2D dựa trên mã hóa one-hot, và ma trận ghép nối (*pairing matrix*) của cấu trúc bậc hai của nó. Sử dụng biểu diễn đầu vào này, chúng tôi huấn luyện một mạng ConvNet để xác định pre-miRNA. Mạng ConvNet được đề xuất của chúng tôi được thiết kế để thích ứng với kích thước đầu vào biến đổi (variable-sized input) bằng cách sử dụng một tâng downscaling ở giữa các tầng tích chập và tầng fully connected. Để so sánh, chúng tôi cũng xây dựng một đầu vào có kích thước cố định (fixed-sized input) ConvNet bằng cách chèn 0 vào các đầu vào để chúng có cùng kích thước.

## Biểu diễn dữ liệu

Cấu trúc bậc hai là một đặc trưng rất có lợi cho việc nhận dạng pre-miRNA [21] [24]. Trong đồ án này, chúng tôi biểu diễn cấu trúc bậc hai bằng cách sử dụng định dạng ma trận ghép nối, xem Hình 1C, cho mỗi cặp nucleotide: 0 là viết tắt của không tương tác, giá trị khác 0 là tương tác.

Mỗi nucleotide A, C, G, U của chuỗi đầu vào được biểu diễn bằng cách sử dụng mã hoá one-hot, cái là một vector nhị phân của 4 chiều. Vì vậy, chúng ta có từ điển {A → (1,0,0,0), C → (0,1,0,0), G → (0, 0, 1, 0), U → (0, 0, 0, 1)}. Từ đó, với mỗi chuỗi đầu vào độ dài L, ta có thế ánh xạ ra một mảng L × 4 đại diện cho toàn bộ chuỗi trong mã hóa một lần. Mỗi chuỗi đầu vào RNA có độ dài L được biểu diễn như sau (xem Hình 1 để biết ví dụ):

* Sử dụng phần mềm RNAfold [18] để dự đoán cấu trúc bậc hai và năng lượng tự do tối thiểu.
* Biễu diễn cấu trúc bậc hai đã dự đoán như một ma trận ghép nối L × L, trong đó vị trí (i, j) cho biết nucleotide i có tương tác với nucleotide thứ j hay không.
* Định hình lại chuỗi mã hóa một lần L × 4 thành L × 1 × 4 và sao chép nó theo chiều ngang để để được một ma trận mới có kích thước L × L × 4.
* Định hình lại chuỗi mã hóa nóng một lần L × 4 thành 1 × L × 4 và sao chép nó theo chiều dọc để được một ma trận mới có kích thước L × L × 4.
* Nối ba ma trận nói trên theo kênh thứ 3, sinh ra L × L × (1 + 4 + 4).

Bằng cách sử dụng biểu diễn dữ liệu này, chúng tôi đảm bảo rằng mỗi 1 × 1 × 9 "pixel" (i, j) của mảng L × L × 9 chứa toàn bộ thông tin có sẵn về cặp nucleotide (i, j), cụ thể là loại nucleotide tại vị trí i và tại vị trí j cũng như liệu hai nucleotide có được ghép nối hay không. Hơn nữa, vùng lân cận của “pixel” đó chứa toàn bộ thông tin có sẵn về các vùng lân cận tuần tự của các vị trí i và j và sự ghép đôi giữa các vùng lân cận này. Sự hiện diện và sắp xếp của các mẫu lân cận như vậy là đặc trưng của một số loại RNA nhất định. Một mạng convolutional thích hợp cho việc trích xuất thông tin mức cao từ các mẫu cục bộ (như được biểu diễn cho các đại diện L × L của protein [31]), và bằng cách sử dụng các tầng gộp / downscaling củng cố thông tin trên toàn bộ phân tử. Hình 1 là một ví dụ về bước tiền xử lý của chúng tôi cho chuỗi “GCCCUUGGCA”, trong đó Hình 1F là đầu ra của bước này.

## Xây dựng mô hình học sâu

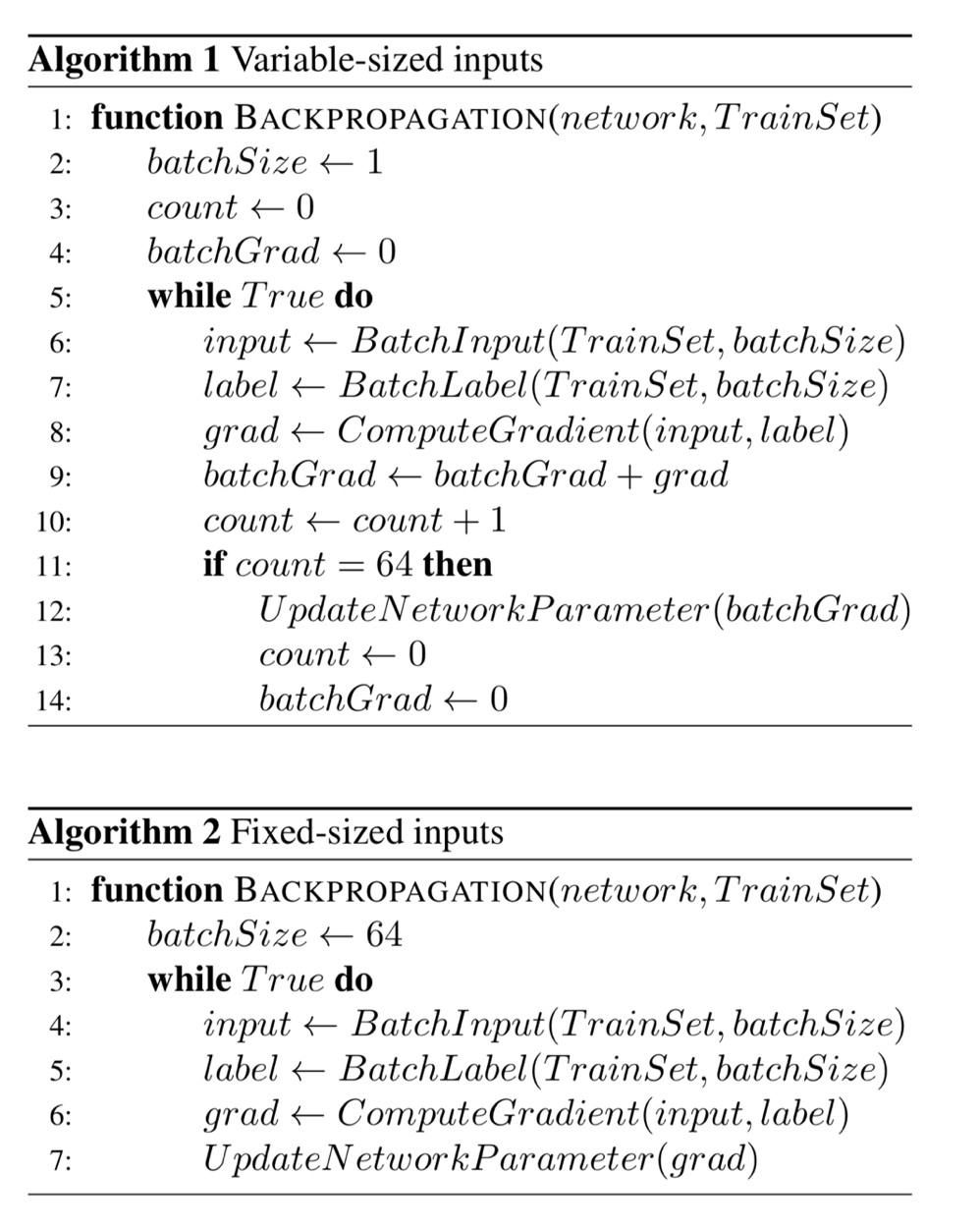
Kiến trúc mạng của chúng tôi được thể hiện trong Hình 9. Cả hai kiến ​​trúc đều chứa ba tầng tích chập convolution theo sau bởi hai tầng kết nối hoàn toàn. Các lý do để sử dụng kiến ​​trúc như vậy kết hợp với biểu diễn đầu vào của chúng tôi được mô tả trong Phần 1.2. Để so sánh *variable-sized input* (Thuật toán 1, Hình 9A) với *fixed-sized input* (Thuật toán 2, Hình 9B), chúng tôi đã chèn 0 và ma trận đầu vào L × L × 9 để nó có kích thước 400 × 400 × 9 trước khi cung cấp làm đầu vào ConvNet. Đối với các đầu vào có kích thước thay đổi (*variable-sized input* Thuật toán 1), vì độ dài biến L của chuỗi đầu vào, chúng ta đặt batch size là 1 và sử dụng một tầng downscaling sau các tầng convolution để thu được các đặc trưng với kích thước cố định, cái mà sẽ được đưa vào các tầng kết nối đầy đủ (*fully connected layers*).

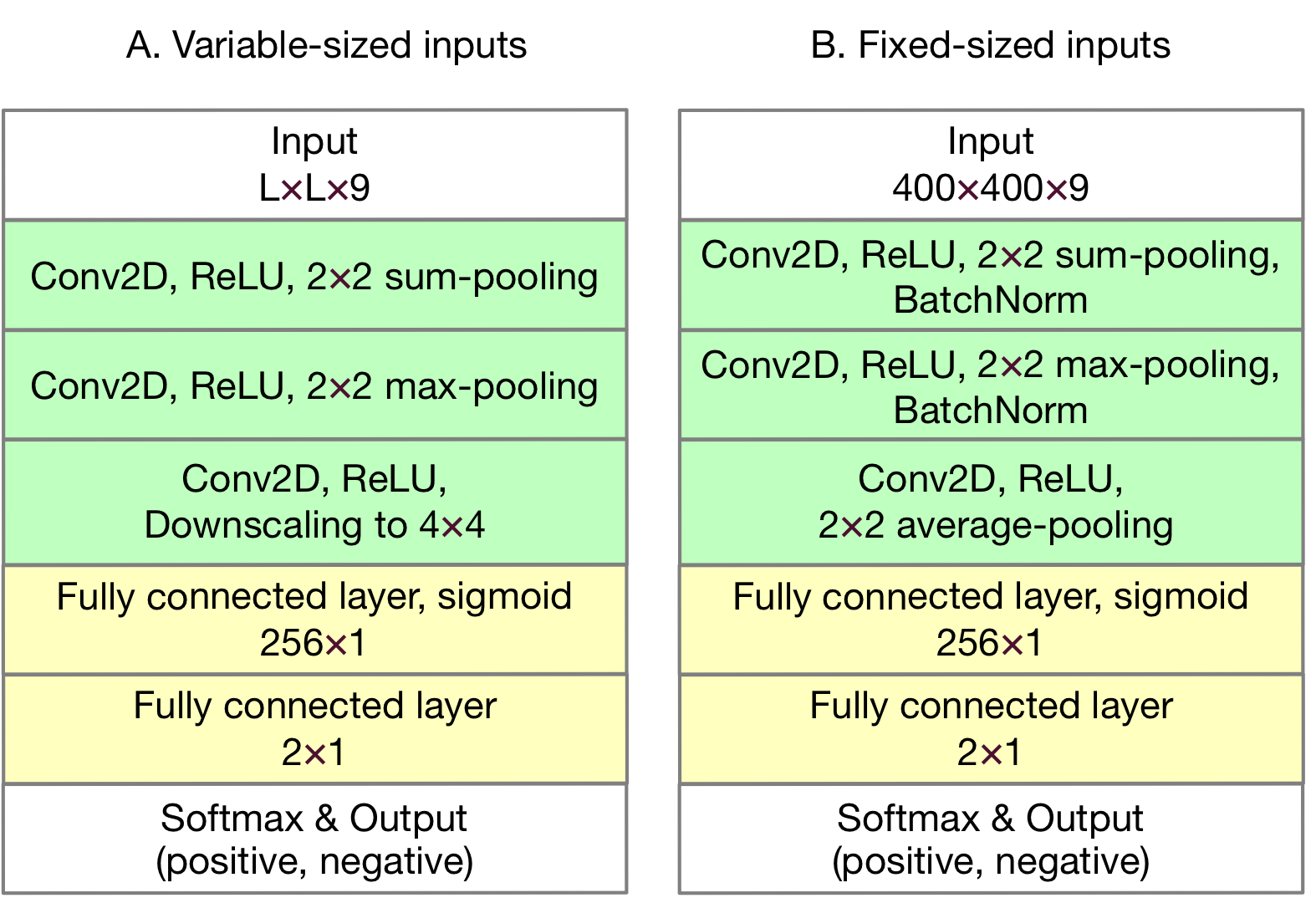
Chúng tôi chuẩn hóa các đầu ra nằm trong khoảng bằng cách sử dụng một tầng softmax ở cuối. Nó được tính như sau, trong đó 2 là số lượng nhãn:

Mô hình này được huấn luyện (*train*) bằng cách giảm thiểu hàm mất mát cross entropy:

trong đó liệt kê các mẫu được huyến luyện (training set), là nếu là mẫu *“positive”*, nếu nó là mẫu *“negative”* ( là trọng số giải quyết sự mất cân bằng trong lớp (class imbalance); , trong đó, tương ứng là số lượng mẫu *“positive”* và *“negative”*), là đầu ra của tầng softmax.

Chúng tôi đã huấn luyện mạng ConvNet bằng phương pháp *adaptive subgradient (AdaGrad)* [41] với *learning rate* là 0,001. Trong các thử nghiệm thay thế, chúng tôi cũng đã thử thay mã hóa *one-hot* bằng một dạng mã hoá mềm mỏng hơn như là hoặc cho chuỗi đầu vào thay vì , trình tối ưu hóa *Adam* [42], kiến trúc mạng *AlexNet* [43] và *ResNet* [38], kích thước *batch size* khác nhau, *dropout* [44], và trọng lượng [45], nhưng kết quả là tương đương trong một số trường hợp thậm chí tệ hơn. Việc sử dụng các tầng convolution ít hơn cũng mang lại kết quả tệ hơn. Mô hình của chúng tôi có thể khớp (fit) với dữ liệu training, được chứng minh bằng accuracy cao và loss function thấp, Hình 11F. Tuy nhiên, khi tách dữ liệu đào tạo thành 5-fold cross validation, các chỉ số thay đổi hướng sau khoảng 40 epochs (Hình 11A – E). Vì vậy, chúng tôi dừng lại sau 40 epochs khi huấn luyện (training) để tránh overfitting.





Hình 9. Mô hình ConvNet đề xuất

# Kết quả thử nghiệm

## Mô tả dữ liệu

Trong các thí nghiệm của chúng tôi, chúng tôi sử dụng các tập dữ liệu được sử dụng trong [17] [10], được lấy từ miRbase [46], NCBI (<https://www.ncbi.nlm.nih.gov/>), fRNAdb [33], NONCODE [47] và snoRNA-LBME-db [48]. Các bộ dữ liệu được gọi là *human, cross-spieces, new.*

Bảng 1 cho thấy thống kê số lượng mẫu “positive” và “negative” trong mỗi bộ dữ liệu. Tổng cộng, có 3230 mẫu “positive” và 23934 mẫu “negative”, dẫn đến sự mất cân bằng trong các lớp. Để đối phó với sự mất cân bằng của lớp, chúng tôi gán trọng số nhiều hơn cho các mẫu “positive” trong hàm mất mát cross entropy. Hình 10 cho thấy biểu đồ độ dài chuỗi trong các tập dữ liệu. Trong khi chiều dài trung bình của chuỗi RNA là 96, thì vẫn có chuỗi với hơn 300 nucleotide. Trình tự RNA dài nhất và ngắn nhất tương ứng là 398 và 45 nucleotide.

Trong giai đoạn xây dựng mô hình, 90% bộ dữ liệu *human* và *cross-spieces* được sử dụng. Chúng tôi chọn kiến ​​trúc của mô hình và hyperparameters (tức là số epochs, learning rate, batch size) dựa trên kết quả của 5-fold cross validation (cross validation được chạy trên 90% của tập *human* và trên 90% của tập *cross-spieces*, nhưng chúng tôi đã tìm thấy các hyperparameters nói trên hoạt động tốt cho cả hai), đó là quy trình chuẩn để điều chỉnh hyperparameters trong tài liệu nhận dạng miRNA trước [46] [10].

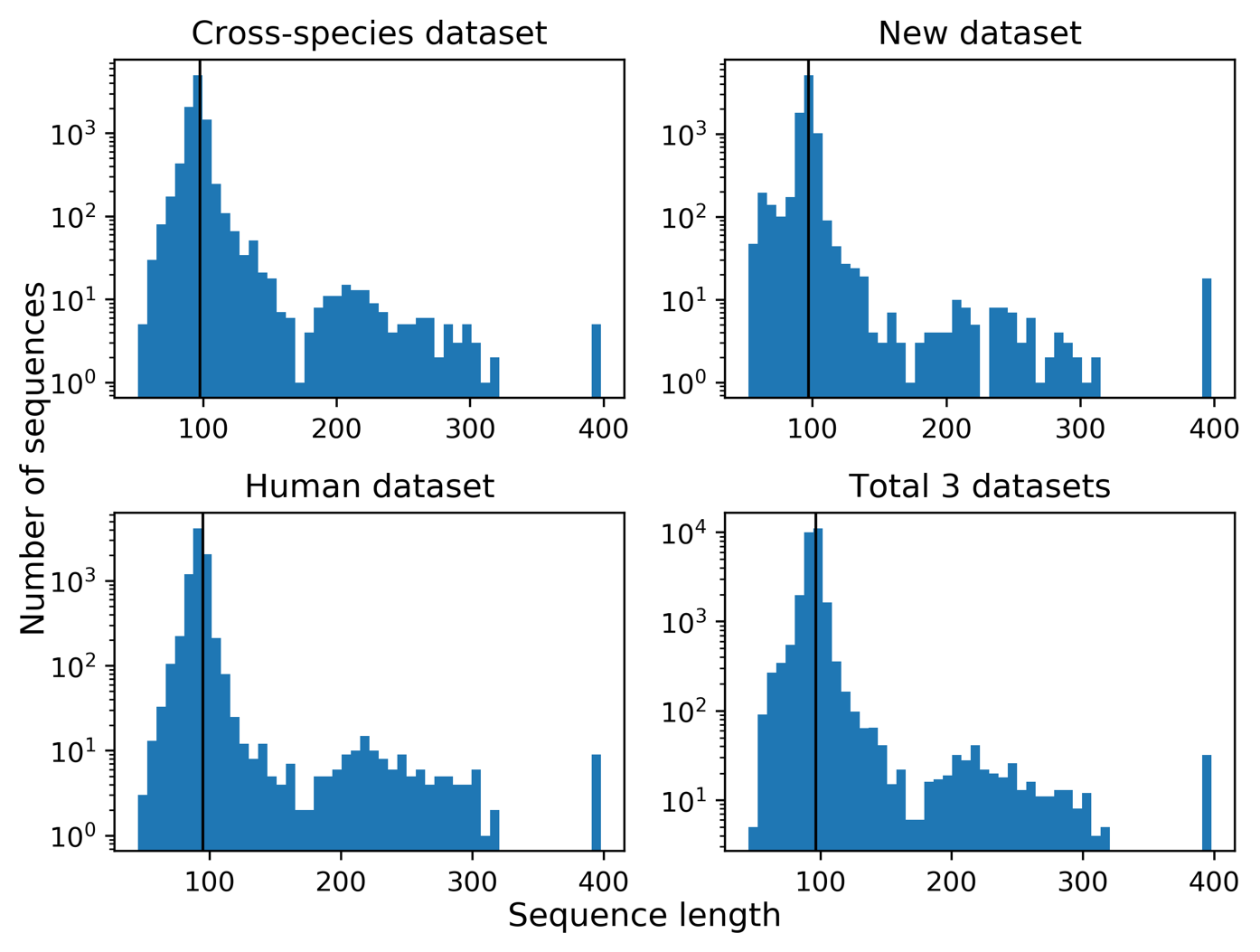
Sau khi thực hiện xong bước cross validation, chúng tôi sử dụng ba thủ tục đánh giá (testing) tiêu chuẩn:

1. Train trên 90% trên bộ *human*, test 10% còn lại của bộ dữ liệu *human*, như được thực hiện trong [17].
2. Train trên 90% trên bộ *cross-spieces*, test 10% còn lại của tập *cross-spieces*, như đã hoàn thành trong [17].
3. Train trên toàn bộ dữ liệu cross-spieces, test toàn bộ tập dữ liệu *new*, như đã đề xuất.

Chúng tôi đã không sử dụng thử nghiệm 10% bộ dữ liệu *human* và *cross-spieces*, cũng như bất kỳ mẫu nào từ tập dữ liệu *new* tại bất kỳ thời điểm nào trong quá trình training cũng như tìm kiếm hyperparameters.

|  |  |  |  |
| --- | --- | --- | --- |
|  | human | Cross-species | new |
| Positive samples | 863 | 1677 | 690 |
| Negative samples | 7422 | 8266 | 8246 |

Bảng 1. Thống kê của các tập dữ liệu



Hình 10. Thống kê dữ liệu

Chúng tôi đánh giá và so sánh kết quả của mô hình với các nghiên cứu trước đó trong lĩnh vực này. Như đã mô tả ở trên, chúng tôi đã thử nghiệm trên 10% bộ dữ liệu *human, cross-spieces*, và tất cả các tập dữ liệu *new*. Chúng tôi cũng báo cáo kết quả 5-fold cross validation, cũng là một đánh giá phổ biến trong học máy. Để so sánh, chúng tôi sử dụng các chỉ số sensitivity (SE), specificity (SP), positive predictive value (PPV), F-score, g-mean, area under the receiver operating characteristic (AUROC), and area under the precision-recall curve (AUPR). Các kết quả được thể hiện trong Bảng 2 Bảng 3 Bảng 4. Các kết quả của miRBoost, CSHMM, triplet-SVM, microPred, MIReNA và deepMiRGene được lấy từ [17]. Chúng tôi cũng chạy lại phương pháp deepMiRGene với source code các tác giả cung cấp, sau đó cũng đưa vào trong Bảng 2 Bảng 3 Bảng 4 để đánh giá. Các chỉ số như đã được được mô tả bên trên được tính như sau: TP: true positive, TN: true negative, FP: false positive, FN: false negative, SE = TP / (TP + FN), SP = TN / (TN + FP), PPV = TP / (TP + FP), F-score = 2TP / (2TP + FP + FN), g-mean = . Ngưỡng quyết định là 0.5.

Vì chúng ta đang xử lý một tập dữ liệu mất cân bằng (có rất nhiều mẫu “negative”), AUPR và AUROC rất quan trọng vì chúng không dễ bị mất cân bằng trong các lớp. AUPR đặc biệt quan trọng đối với chúng tôi vì mục tiêu chính của chúng tôi là phát hiện các mẫu “positive”. Độ chính xác đo lường xác suất phát hiện mẫu “positive” một cách chính xác và không bị ảnh hưởng bởi một số lượng lớn mẫu “negative” trong tập dữ liệu.

## Kết quả của 5-fold cross validation

Nhìn vào kết quả 5-fold cross validation, cách tiếp cận của chúng tôi đã vượt trội so với các phương pháp hiện đại trên bộ dữ liệu *human* và đạt được các kết quả cạnh tranh về tập dữ liệu *cross-spieces*. Cụ thể đối với bộ dữ liệu *human*, SE (0.878) tốt hơn 6% so với kết quả của phương pháp khác (0.818), trong khi F-score, g-mean, AUROC và AUPR thì cũng tốt hơn so với các phương pháp khác. Điều này cho thấy rằng cách tiếp cận của chúng tôi có thể xử lý dữ liệu mất cân bằng và tăng hy vọng cho kết quả tốt trong giai đoạn thử nghiệm (testing).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Methods | SE | SP | PPV | F-score | | g-mean | AUROC | AUPR |
| miRBoost (CV) | 0.803 | 0.988 | 0.887 | 0.843 | 0.891 | | - | - |
| CSHMM (CV) | 0.713 | 0.777 | 0.559 | 0.570 | 0.673 | | - | - |
| triplet-SVM (CV) | 0.669 | 0.986 | 0.851 | 0.749 | 0.812 | | 0.957 | 0.854 |
| microPred (CV) | 0.763 | **0.989** | **0.888** | 0.820 | 0.869 | | 0.974 | 0.890 |
| MIReNA (CV) | 0.818 | 0.943 | 0.624 | 0.708 | 0.878 | | - | - |
| LSTM-ATT-paper (CV) | 0.799 | 0.988 | 0.885 | 0.839 | 0.888 | | 0.984 | 0.915 |
| LSTM-ATT-rerun (CV) | 0.818 | 0.977 | 0.806 | 0.811 | 0.894 | | 0.975 | 0.876 |
| Proposed fixed-sized (CV) | **0.878** | 0.978 | 0.827 | 0.849 | **0.926** | | 0.984 | 0.915 |
| Proposed variable-sized (CV) | 0.835 | 0.985 | 0.868 | **0.851** | 0.907 | | **0.985** | **0.922** |
| Without contact map (CV) | 0.722 | 0.938 | 0.594 | 0.646 | 0.823 | | 0.909 | 0.726 |
| miRBoost (test) | 0.884 | 0.969 | 0.768 | 0.822 | 0.925 | | - | - |
| CSHMM (test) | 0.616 | 0.978 | 0.768 | 0.684 | 0.777 | | - | - |
| triplet-SVM (test) | 0.744 | **0.992** | 0.914 | 0.821 | 0.859 | | 0.947 | 0.830 |
| microPred (test) | 0.779 | 0.988 | 0.882 | 0.827 | 0.877 | | 0.980 | 0.892 |
| MIReNA (test) | 0.826 | 0.941 | 0.617 | 0.706 | 0.881 | | - | - |
| LSTM-ATT-paper (test) | 0.822 | **0.992** | **0.919** | 0.868 | 0.903 | | 0.981 | 0.918 |
| LSTM-ATT-rerun (test) | 0.767 | 0.989 | 0.892 | 0.825 | 0.871 | | 0.981 | 0.900 |
| Proposed fixed-sized (test) | **0.930** | 0.984 | 0.870 | **0.899** | **0.957** | | 0.983 | **0.946** |
| Proposed variable-sized (test) | 0.884 | 0.991 | 0.916 | **0.899** | 0.936 | | **0.986** | 0.934 |
| Without contact map (test) | 0.628 | 0.956 | 0.621 | 0.624 | 0.775 | | 0.892 | 0.669 |

Bảng 2. Kết quả trên tập dữ liệu human

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Methods | SE | SP | PPV | F-score | g-mean | AUROC | AUPR |
| miRBoost (CV) | 0.861 | 0.977 | 0.884 | 0.872 | 0.917 | - | - |
| CSHMM (CV) | 0.826 | 0.576 | 0.533 | 0.564 | 0.524 | - | - |
| triplet-SVM (CV) | 0.735 | 0.967 | 0.819 | 0.775 | 0.843 | 0.943 | 0.869 |
| microPred (CV) | 0.825 | 0.975 | 0.875 | 0.848 | 0.897 | 0.970 | 0.873 |
| MIReNA (CV) | 0.766 | 0.952 | 0.765 | 0.765 | 0.854 | - | - |
| LSTM-ATT-paper (CV) | 0.886 | **0.982** | **0.911** | **0.898** | 0.933 | **0.985** | 0.927 |
| LSTM-ATT-rerun (CV) | 0.883 | 0.970 | 0.861 | 0.871 | 0.926 | 0.981 | 0.929 |
| Proposed fixed-sized (CV) | **0.903** | 0.978 | 0.894 | **0.898** | **0.940** | **0.985** | **0.936** |
| Proposed variable-sized (CV) | 0.881 | 0.981 | 0.906 | 0.893 | 0.930 | 0.983 | **0.936** |
| Without contact map (CV) | 0.785 | 0.928 | 0.704 | 0.737 | 0.853 | 0.926 | 0.828 |
| miRBoost (test) | 0.856 | 0.844 | 0.526 | 0.651 | 0.850 | - | - |
| CSHMM (test) | 0.749 | 0.960 | 0.791 | 0.769 | 0.848 | - | - |
| triplet-SVM (test) | 0.760 | 0.977 | 0.870 | 0.812 | 0.862 | 0.952 | 0.908 |
| microPred (test) | 0.814 | 0.985 | 0.919 | 0.863 | 0.896 | 0.963 | 0.906 |
| MIReNA (test) | 0.796 | 0.950 | 0.764 | 0.780 | 0.870 | - | - |
| LSTM-ATT-paper (test) | 0.900 | 0.983 | 0.913 | 0.906 | 0.940 | 0.984 | **0.955** |
| LSTM-ATT-rerun (test) | 0.862 | 0.984 | 0.917 | 0.889 | 0.921 | 0.984 | 0.954 |
| Proposed fixed-sized (test) | **0.904** | 0.982 | 0.910 | **0.907** | **0.942** | 0.983 | 0.951 |
| Proposed variable-sized (test) | 0.880 | **0.988** | **0.936** | **0.907** | 0.933 | **0.985** | 0.950 |
| Without contact map (test) | 0.695 | 0.955 | 0.758 | 0.725 | 0.815 | 0.914 | 0.798 |

Bảng 3. Kết quả trên tập dữ liệu cross-species

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Methods | SE | SP | PPV | F-score | g-mean | AUROC | AUPR |
| miRBoost | **0.921** | 0.936 | 0.609 | 0.733 | 0.928 | - | - |
| CSHMM | 0.536 | 0.069 | 0.046 | 0.085 | 0.192 | - | - |
| triplet-SVM | 0.721 | **0.981** | 0.759 | 0.740 | 0.841 | 0.934 | 0.766 |
| microPred | 0.728 | 0.970 | 0.672 | 0.699 | 0.840 | 0.940 | 0.756 |
| MIReNA | 0.450 | 0.941 | 0.392 | 0.419 | 0.650 | - | - |
| LSTM-ATT-paper | 0.917 | 0.964 | 0.682 | 0.782 | 0.941 | **0.981** | 0.808 |
| LSTM-ATT-rerun | 0.899 | 0.968 | 0.700 | 0.787 | 0.933 | 0.980 | 0.798 |
| Proposed fixed-sized | 0.917 | 0.967 | 0.696 | 0.792 | **0.942** | 0.797 | **0.864** |
| Proposed variable-sized | 0.859 | **0.981** | **0.779** | **0.817** | 0.918 | 0.979 | 0.818 |
| Without contact map | 0.855 | 0.938 | 0.535 | 0.658 | 0.896 | 0.958 | 0.798 |

Bảng 4. Kết quả trên tập dữ liệu new

## Kết quả trên dữ liệu test

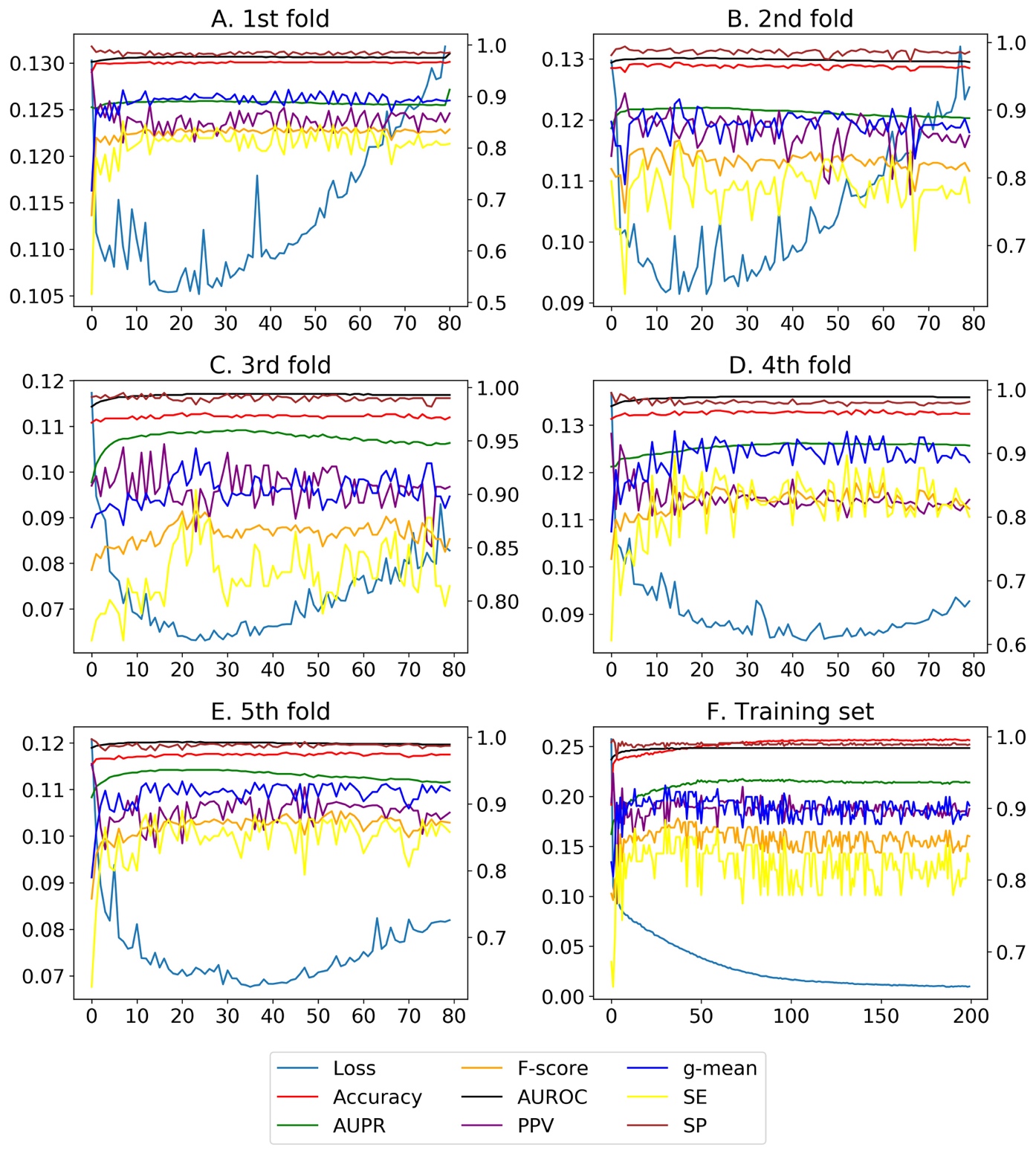
Phương pháp được đề xuất tốt hơn các phương pháp trước khá ổn định. Trong bộ dữ liệu *human*, kết quả kiểm tra của mỗi số liệu (ngoại trừ SP và PPV) đều tốt hơn tất cả các phương pháp khác. Đối với tập dữ liệu *cross-spieces*, phương pháp của chúng tôi cho thấy các kết quả có thể so sánh với các phương pháp khác về SP, PPV và F-score và AUROC. Trên tập dữ liệu *new*, nó đạt được điểm số cao nhất về SP, PPV, F-score và AUPR. Ngoài ra, trong tất cả các tập dữ liệu, chúng tôi đạt được điểm số F-score (0.899 cho *human*, 0.907 đối với *cross-spieces* và 0.817 đối với *new*). Ngay cả trong các số liệu mà phương pháp của chúng tôi không có điểm số tốt nhất, nó là rất gần với tốt nhất trong hầu như tất cả các trường hợp. Trong các bộ dữ liệu *human* và *new*, có thể thấy rằng phương pháp của chúng tôi một sự cải thiện đáng kể so với deepMiRGene.

Chúng ta cũng có thể thấy cách tiếp cận học tập sâu mà không có các đặc trưng thủ công (handcrafted features), cụ thể là deepMiRGene và mô hình của chúng tôi, tốt hơn so với các phương pháp học máy trước đó. Các phương pháp mới có kết quả tốt hơn trong tập dữ liệu *new* trong mỗi số liệu ngoại trừ tính SP, trong đó triplet-SVM cho cùng một điểm như deepMiRGene. Đối với cả hai phần cross validation và test của *cross-spieces*, các phương pháp học sâu đưa ra kết quả tốt hơn tất cả các phương pháp trước đây. Đối với tập dữ liệu *new*, miRBoost cho kết quả tốt nhất về mặt SE, và triplet-SVM có cùng SP như mô hình được đề xuất của chúng tôi, nhưng trong mọi phương diện khác, phương pháp học sâu có kết quả tốt hơn. Xem xét tất cả các tập dữ liệu và kết quả, chúng ta có thể kết luận rằng phương pháp học sâu tốt hơn các giải pháp học máy trước đây.

Khi so sánh deepMiRGene với phương pháp của chúng tôi, cách tiếp cận của chúng tôi cho kết quả tốt hơn trong tất cả các tập dữ liệu thử nghiệm. Đối với tập dữ liệu *human*, cách tiếp cận của chúng tôi cung cấp các cải tiến về mọi chỉ số ngoại trừ SP và PPV. Đối với tập dữ liệu *cross-spieces*, phương pháp của chúng tôi cho kết quả tốt hơn về mọi số liệu khác, ngoại trừ AUPR. Cuối cùng, đối với tập dữ liệu *new*, mô hình của chúng tôi hoạt động tốt hơn trong mọi chỉ số ngoại trừ AUROC. Kết quả khi training trên tập dữ liệu *cross-spieces* và test trên tập dữ liệu *new* (Bảng 4) chứng minh rằng cách tiếp cận của chúng tôi có thể xác định tốt các pre-miRNA trong một loài mới mặc dù nó được huấn luyện trên một tập hợp các loài hỗn hợp.

Cũng quan trọng để so sánh hai cách tiếp cận của chúng tôi với nhau. Nếu không có ma trận ghép nối (pairing matrix), kết quả sẽ tồi tệ hơn nhiều trong mỗi tập dữ liệu và mỗi số liệu. Do đó, ma trận ghép nối là một phần quan trọng trong phương pháp của chúng tôi. Các đầu vào có kích thước cố định và có kích thước biến thiên cho kết quả tương tự cho hầu hết dữ liệu. Với đầu vào có kích thước cố định, chúng tôi cần nhiều bộ nhớ để huấn luyện hơn so với đầu vào có kích thước thay đổi. Do đó, chúng tôi khuyến khích sử dụng đầu vào có kích thước biến thiên cho nhiệm vụ này.

Xem xét tất cả các kết quả và so sánh với các phương pháp khác, kết quả cho thấy phương pháp đề xuất của chúng tôi mang lại hiệu suất tốt nhất trong việc xác định các pre-miRNA.



Hình 11. Biểu đồ các chỉ số đánh giá trên bộ dữ liệu 5-fold và training

# Kết luận và công việc tương lai

Kết quả khi sử dụng *pairing* matrix tốt hơn nhiều so với khi không sử dụng nó. Điều đó có nghĩa rằng *pairing matrix* và cấu trúc bậc hai đóng vai trò quan trọng trong nhận biết miRNA. Bằng cách nhúng cấu trúc bậc hai vào đầu vào của mạng ConvNet, không gian đặc trưng sẽ lớn hơn. Vì thế, chúng ta cần một bộ dữ liệu lớn hơn để tổng quát hoá được dữ liệu. Bên cạnh đó, chúng ta phải đối mặt với sự mất cân bằng giữa các nhãn (positive và negative) trong dữ liệu. Một hướng hứa hẹn cho công việc tương lai để cải thiện kết quả nhận biết là sưu tập thêm nhiều dữ liệu, sau đó sử dụng mạng nơron lớn hơn, chẳng hạn AlexNet [43] hoặc ResNet [38].

Trong đồ án này, em đã đề xuất cách thể hiện thông tin cấu trúc bậc hai của RNA, hay còn gọi là *pairing matrix,* làm đầu vào cho một mạng ConvNet để trích xuất các đặc trưng một cách tự động. Phương pháp này đã giành được kết quả tốt nhất trên ba bộ dữ liệu chuẩn *human*, *cross-species* and new, đặc biệt cho độ đo F-score. Trong tương lai, em sẽ mở rộng phương pháp để giải quyết các bài toán khác liên quan đến miRNA, chẳng hạn: dự đoán chuỗi mRNA đích của miRNA, dự đoán chức năng của miRNA [49].

# Tài liệu Tham Khảo

|  |  |
| --- | --- |
| [1] | B. Isaac, “Identification of hundreds of conserved and nonconserved human microRNAs,” *Nature Genetics,* tập 37, p. 766, 2005. |
| [2] | M. Jones-Rhoades, “MicroRNAs and their regulatory roles in plants,” *Annu. Rev. Plant Biol.,* tập 57, pp. 19-53, 2006. |
| [3] | S. Pfeffer, “Identification of virus-encoded microRNAs,” *Science,* tập 304, pp. 734-736, 2004. |
| [4] | D. P. Bartel, “MicroRNAs: genomics, biogenesis, mechanism, and function,” *Cell,* tập 116, pp. 281-297, 2004. |
| [5] | M. D. Jansson, “MicroRNA and cancer,” *Molecular oncology,* tập 6, pp. 590-610, 2012. |
| [6] | D. Wang, “Human microRNA oncogenes and tumor suppressors show significantly different biological patterns: from functions to targets,” *PloS one,* tập 5, 2010. |
| [7] | Tüfekci, “The role of microRNAs in human diseases,” *miRNomics: MicroRNA Biology and Computational Analysis,* pp. 33-50, 2014. |
| [8] | A. Mathelier, “MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data,” *Bioinformatics,* tập 26, pp. 2226-2234, 2010. |
| [9] | A. Gudyś, “HuntMi: an efficient and taxon-specific approach in pre-miRNA identification,” *BMC bioinformatics,* tập 14, p. 83, 2013. |
| [10] | S. Tempel, “miRBoost: boosting support vector machines for microRNA precursor classification,” *RNA,* 2015. |
| [11] | S. Agarwal, “Prediction of novel precursor miRNAs using a context-sensitive hidden Markov model (CSHMM),” *BMC bioinformatics,* tập 11, p. S29, 2010. |
| [12] | R. Batuwita, “microPred: effective classification of pre-miRNAs for human miRNA gene prediction,” *Bioinformatics,* tập 25, pp. 989-995, 2009. |
| [13] | P. Jiang, “MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features,” *Nucleic acids research,* tập 35, số W339-W344, p. Oxford University Press, 2007. |
| [14] | C. Xue, “Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine,” *BMC bioinformatics,* tập 6, số BioMed Central, p. 310, 2005. |
| [15] | M. E. Rahman, “MiRANN: A reliable approach for improved classification of precursor microRNA using Artificial Neural Network model,” *Genomics,* tập 99, số Elsevier, pp. 189-194, 2012. |
| [16] | J. Thomas, “DP-miRNA: An improved prediction of precursor microRNA using deep learning model,” *Big Data and Smart Computing (BigComp), 2017 IEEE International Conference,* số IEEE, pp. 96-99, 2017. |
| [17] | S. Park, “Deep recurrent neural network-based identification of precursor micrornas,” *Advances in Neural Information Processing Systems,* pp. 2891-2900, 2017. |
| [18] | I. L. Hofacker, “Vienna RNA secondary structure server,” *Nucleic acids research,* tập 31, số Oxford University Press, pp. 3429-3431, 2003. |
| [19] | A. Mathuriya, “GTfold: a scalable multicore code for RNA secondary structure prediction,” *Proceedings of the 2009 ACM symposium on Applied Computing,* số ACM, pp. 981-988, 2009. |
| [20] | E. Bindewald, “Cylofold: secondary structure prediction including pseudoknots,” *Nucleic acids research,* tập 38, số Oxford University Press, pp. W368-W372, 2010. |
| [21] | L. Jiang, “BP neural network could help improve pre-miRNA identification in various species,” *BioMed research international,* số Hindawi, 2016. |
| [22] | K. L. S. Ng, “De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures,” *Bioinformatics,* tập 23, số Oxford University Press, pp. 1321-1330, 2007. |
| [23] | V. D. T. Tran, “miRBoost: boosting support vector machines for microRNA precursor classification,” *RNA,* tập 21, p. 775. |
| [24] | Demirci, “ On the performance of pre-microRNA detection algorithms,” *Nature Communications,* tập 8, p. 330. |
| [25] | J. Kuka{\v{c}}ka, “Regularization for Deep Learning: A Taxonomy,” *arXiv,* 2017. |
| [26] | L. Cun, “Deep learning,” *Nature,* tập 521, p. 436. |
| [27] | J. Mendenhall, “Improving quantitative structure--activity relationship models using Artificial Neural Networks trained with dropout,” *Journal of computer-aided molecular design,* tập 30, số Springer, pp. 177-189, 2016. |
| [28] | V. Golkov, “3D deep learning for biological function prediction from physical fields,” *arXiv,* 2017. |
| [29] | M. Spencer, “A deep learning network approach to ab initio protein secondary structure prediction,” *IEEE/ACM transactions on computational biology and bioinformatics (TCBB),* tập 12, pp. 103-112, 2015. |
| [30] | M. AlQuraishi, “End-to-end differentiable learning of protein structure,” *bioRxiv,* số Cold Spring Harbor Laboratory, 2018. |
| [31] | V. Golkov, “Protein contact prediction from amino acid co-evolution using convolutional networks for graph-valued images,” *Advances in Neural Information Processing Systems,* pp. 4222-4230, 2016. |
| [32] | S. Hochreiter, “Long short-term memory,” *Neural computation,* tập 9, pp. 1735-1780, 1997. |
| [33] | A. Churkin, “RNA dot plots: an image representation for RNA secondary structure analysis and manipulations,” *Wiley Interdisciplinary Reviews: RNA,* tập 4, số Wiley Online Library, pp. 205-216, 2013. |
| [34] | W. M. Fitch, “Locating gaps in amino acid sequences to optimize the homology between two proteins,” *Biochemical genetics,* tập 3, pp. 99-108, 1969. |
| [35] | J. V. Maizel, “Enhanced graphic matrix analysis of nucleic acid and protein sequences,” tập 78, pp. 7665-7669, 1981. |
| [36] | R. Staden, “An interactive graphics program for comparing and aligning nucleic acid and amino acid sequences,” *Nucleic Acids Research,* tập 10, số Oxford University Press, pp. 2951-2961, 1982. |
| [37] | X. Wang, “miRDB: a microRNA target prediction and functional annotation database with a wiki interface.,” *RNA,* tập 14, số Cold Spring Harbor Lab, pp. 1012-1017, 2008. |
| [38] | K. He, "Deep residual learning for image recognition," in *the IEEE conference on computer vision and pattern recognition*, 2016. |
| [39] | A. Singh, “Sentiment Transfer using Seq2Seq Adversarial Autoencoders,” *Arxiv,* 2018. |
| [40] | G. H. Weinberger, “Densely Connected Convolutional Networks,” trong *Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017. |
| [41] | J. Duchi, “Adaptive subgradient methods for online learning and stochastic optimization,” *Journal of Machine Learning Research,* tập 12, pp. 2121-2159, 2011. |
| [42] | D. Kingma, “A method for stochastic optimization,” *arXiv,* 2014. |
| [43] | A. Krizhevsky, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012. |
| [44] | N. Srivastava, “Dropout: a simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research,* tập 15, pp. 1929-1958, 2014. |
| [45] | A. Krogh, “A simple weight decay can improve generalization,” *Advances in neural information processing systems,* pp. 950-957, 1992. |
| [46] | Griffiths-Jones, “miRBase: microRNA sequences, targets and gene nomenclature,” *Nucleic acids research,* tập 34, số Oxford University Press, pp. D140-D144, 2006. |
| [47] | D. Bu, “NONCODE v3. 0: integrative annotation of long noncoding RNAs,” *Nucleic acids research,* tập 40, số Oxford University Press, pp. D210-D215, 2011. |
| [48] | L. Lestrade, “snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs,” *Nucleic acids research,* tập 34, số Oxford University Press, pp. D158-D162, 2006. |
| [49] | S. Singh, "Machine Learning Techniques in Exploring MicroRNA Gene Discovery, Targets, and Functions," *Bioinformatics in MicroRNA Research,* no. Springer, pp. 211-224, 2017. |