Thực thi phần cứng một mạng nơ-ron nhân tạo sử dụng thuật toán lan truyền ngược

« Hardware Implementation an Artificial neural network using Backpropagation Algorithm »

Tóm tắt

Mạng nơ-ron nhân tạo (artificial neural network) là công nghệ lõi của học máy (machine learning) với rất nhiều ứng dụng quan trọng hiện nay như phân loại đối tượng, dự báo... Tuy nhiên để mạng nơ-ron nhân tạo xử lý các bài toán lớn thì cần kích thước mạng rất lớn, tiêu tốn nhiều tài nguyên, vì vậy thực thi phần cứng mạng nơ-ron là một hướng đi tối ưu cho tăng tốc thời gian thực thi và giảm tài nguyên tiêu thụ. Trong bài tiểu luận này nghiên cứu về mạng nơ-ron nhân tạo sử dụng thuật toán lan truyền ngược và thực thi mô hình phần cứng mô phỏng mạng nơ-ron đơn giản. Kết quả thực thi đạt được…

# Giới thiệu

Mạng nơ-ron nhân tạo (artifical neural network) là một mô hình toán học hay mô hình tính toán được xây dựng dựa trên các mạng nơ-ron sinh học. Động cơ thúc đẩy sự phát triển của công nghệ mạng nơron bắt nguồn từ mong muốn phát triển một hệ thống nhân tạo có thể thực hiện các nhiệm vụ "thông minh" giống như những gì được thực hiện bởi bộ não con người Mạng nơ-ron là một công cụ mô hình hoá dữ liệu mạnh mẽ có thể nắm bắt và đại diện các mối quan hệ đầu vào/đầu ra phức tạp.

Sức mạnh và tính hữu dụng của mạng nơ-ron nhân tạo đã được chứng minh trong một số ứng dụng bao gồm tổng hợp lời nói, các vấn đề chẩn đoán, y học, kinh doanh và tài chính, kiểm soát robot, xử lý tín hiệu, tầm nhìn máy tính và nhiều vấn đề khác thuộc loại nhận dạng mẫu. Đối với một số lĩnh vực ứng dụng, các mô hình thần kinh hứa hẹn sẽ đạt được kết quả giống như con người thông qua các kỹ thuật giả tạo truyền thống.

Trong mạng nơ-ron nhân tạo, thuật toán Lan truyền ngược (Backpropagation) là một phương pháp cập nhật trọng số phổ biến để huấn luyện mạng dựa trên cơ sở phương pháp tối ưu hóa Gradient descent.

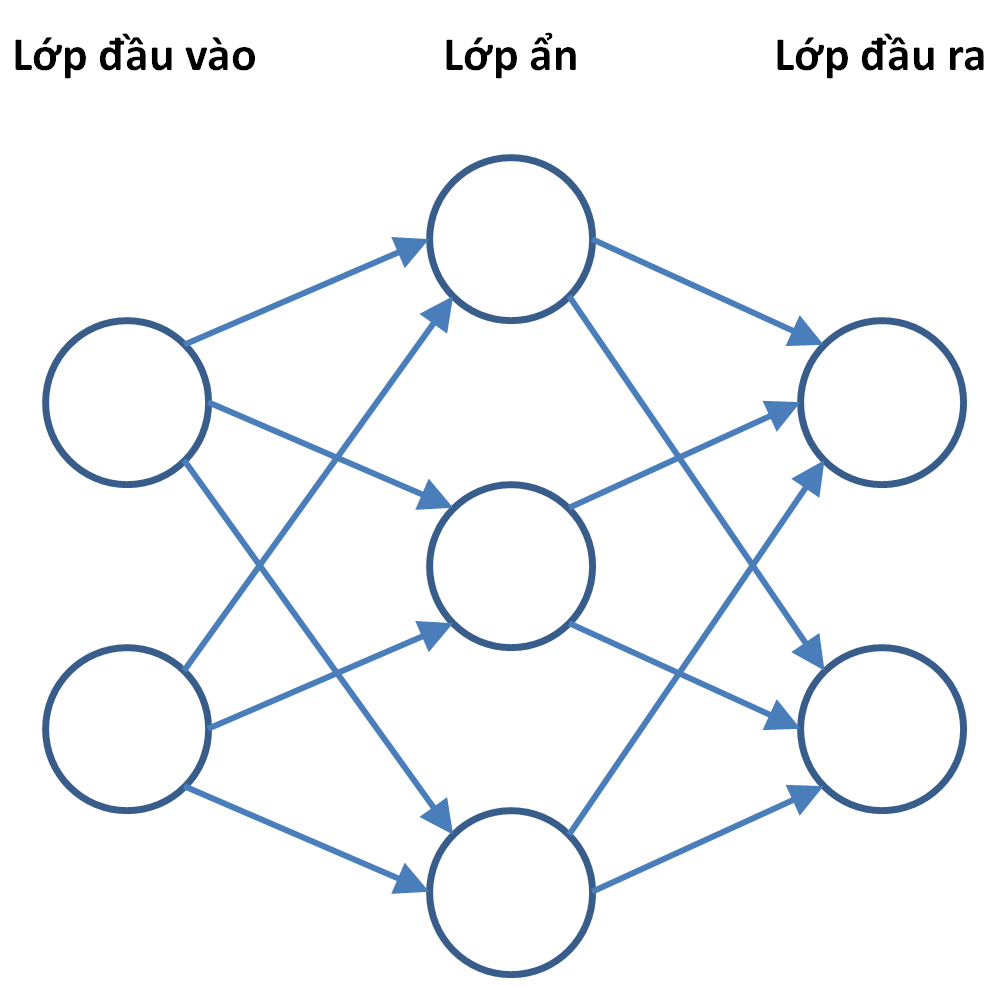
Mạng nơ-ron nhân tạo hữu ích trong việc xấp xỉ các hàm của tập dữ liệu đầu vào có sẵn với đầu ra kỳ vọng. Tuy nhiên khi tập dữ liệu vào rất lớn độ phức tạp tăng lên tuyến tính, các bài toán thực tế với đầu vào hình ảnh, âm thanh, video… tạo một mạng nơ-ron độ phức tạp, khó thực thi trên các phần mềm và máy tính nhúng thông thường. Vì vậy thực thi phần cứng mạng nơ-ron là một cách tiếp cận tối ưu giúp tăng tốc độ tính toán với khả năng thực hiện song song.

Trong bài tiểu luận này sẽ đề cập đến thuật toán đào tạo có giám sát cho mạng nơ-ron nhân tạo và thuật toán Lan truyền ngược cho mạng. Sau đó bài tiểu luận tiếp cận các phương pháp tối ưu phổ biến cho mạng nơ-ron. Sau đó, một kiến trúc phần cứng mạng nơ-ron sẽ được biểu diễn. Cuối cùng sẽ là sự phân tích kết quả mô phỏng của thuật toán Lan truyền ngược trên phần cứng.

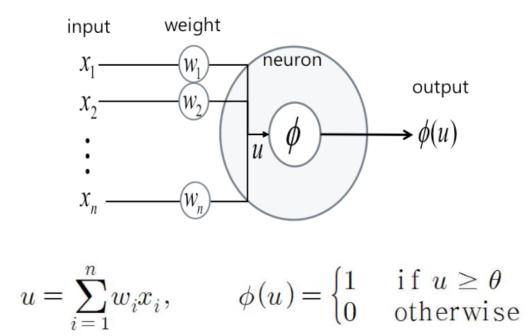
# Mạng nơ-ron và thuật toán lan truyền ngược

Từ cấu tạo của bộ não trong sinh học, các nhà nghiên cứu đã khám phá ra mạng nơ-ron – một cách tiếp cận mới mạng tính thuật toán để xử lý thông tin. Mạng nơ-ron là một công cụ mô hình hóa dữ liệu mạnh mẽ có thể nắm bắt và đại diện các mối quan hệ đầu vào/đầu ra phức tạp. Động cơ thúc đẩy sự phát triển của công nghệ mạng nơron bắt nguồn từ mong muốn phát triển một hệ thống nhân tạo có thể thực hiện các nhiệm vụ "thông minh" giống như những gì được thực hiện bởi bộ não con người. Mạng thần kinh giống với não người theo hai cách sau:

* Một mạng thần kinh thu thập kiến ​​thức thông qua học tập.
* Một kiến ​​thức về mạng nơ ron được lưu trữ trong các thế mạnh liên kết neuron gọi là trọng lượng synap.



Hình : Một mạng nơ-ron cơ bản gồm 3 thành phần: lớp đầu vào, lớp ẩn và lớp đầu ra.



Hình : Kiến trúc của một nơ-ron,

# Kiến trúc phần cứng

# Kết quả

# Kết luận