[CHƯƠNG 1. GIỚI THIỆU LUẬN VĂN 4](#_Toc56925514)

[1.1. GIỚI THIỆU ĐỀ TÀI 4](#_Toc56925515)

[1.1.1. Phục vụ nhu cầu trong cuộc sống hằng ngày 5](#_Toc56925516)

[1.1.2. Áp dụng vào nhiều lĩnh vực trong cuộc sống 5](#_Toc56925517)

[1.1.3. Tiềm năng kinh tế cao 6](#_Toc56925518)

[1.1.4. Phục vụ nhu cầu sử dụng công nghệ cho đa dạng người trong xã hội 7](#_Toc56925519)

[1.2. LÝ DO LỰA CHỌN ĐỀ TÀI 7](#_Toc56925520)

[1.3. HƯỚNG PHÁT TRIỂN CỦA LUẬN VĂN 8](#_Toc56925521)

[1.4. MỤC TIÊU CỦA LUẬN VĂN 9](#_Toc56925522)

[1.5. PHẠM VI ĐỀ TÀI 9](#_Toc56925523)

[CHƯƠNG 2. LÝ THUYẾT NỀN TẢNG 10](#_Toc56925524)

[2.1. LÝ THUYẾT NỀN TẢNG CỦA NHẬN DẠNG ÂM THANH 10](#_Toc56925525)

[2.1.1. Tổng quan 10](#_Toc56925526)

[2.1.1.1. Tổng quan nhận dạng âm thanh 10](#_Toc56925527)

[2.1.1.2. Các hướng tiếp cận 10](#_Toc56925528)

[2.1.2. Đăc trưng âm học 12](#_Toc56925529)

[2.1.2.1. Bản chất của âm thanh 12](#_Toc56925530)

[2.1.2.2. Ngữ âm Tiếng Việt 12](#_Toc56925531)

[2.1.3. Lý thuyết nền tảng của mạng nơ ron 12](#_Toc56925532)

[2.1.3.1. Cấu trúc tế bào mạng nơ-ron 13](#_Toc56925533)

[2.1.3.2. Kết hợp các tế bào mạng nơ-ron thành một mạng lưới mạng nơ-ron 14](#_Toc56925534)

[2.1.3.3. Hàm kích hoạt (Activation function) 16](#_Toc56925535)

[2.1.3.4. Hàm mất mát (Loss function) 20](#_Toc56925536)

[2.1.3.5. Lan truyền ngược (Back propagation) 20](#_Toc56925537)

[2.1.3.6. Học với lan truyền ngược 20](#_Toc56925538)

[2.1.3.7. Độ giảm dốc 20](#_Toc56925539)

[2.1.4. Các phương pháp huấn luyện mạng nơ ron 20](#_Toc56925540)

[2.1.4.1. Chuẩn hóa theo lô 20](#_Toc56925541)

[2.1.4.2. Dropout 20](#_Toc56925542)

[2.1.5. Mạng nơ ron hồi quy 20](#_Toc56925543)

[2.1.5.1. Tổng quan 20](#_Toc56925544)

[2.1.5.2. BiRNN 20](#_Toc56925545)

[2.1.5.3. GRU 20](#_Toc56925546)

[2.1.5.4. LSTM 20](#_Toc56925547)

[2.1.6. Kỹ thuật trong nhận dạng âm thanh 20](#_Toc56925548)

[2.1.6.1. Phân tích đặc trưng 20](#_Toc56925549)

[2.1.6.2. Phân loại thời gian kết nối CTC 20](#_Toc56925550)

[2.1.6.3. Thuật toán tham lam, beam search 20](#_Toc56925551)

[2.2. HỆ THỐNG NHẬN DẠNG ÂM THANH 20](#_Toc56925552)

[2.2.1. Hệ thống nhận dạng âm thanh truyền thống 20](#_Toc56925553)

[2.2.2. Hệ thống nhận dạng âm thanh theo hướng đầu cuối (end to end) 20](#_Toc56925554)

[CHƯƠNG 3. GIẢI PHÁP XÂY DỰNG MÔ HÌNH ÂM THANH 20](#_Toc56925555)

[3.1. Tổng quan kiến trúc hệ thống 20](#_Toc56925556)

[3.2. Giải pháp xây dựng mô hình âm thanh 23](#_Toc56925557)

[3.2.1. Tổng quan giải pháp 23](#_Toc56925558)

[3.2.2. Mô hình mạng nơ-ron hồi quy RNN 23](#_Toc56925559)

[3.3. Chuẩn hóa hàng loạt (Batch normalization) 27](#_Toc56925560)

[3.4. SortaGrad 28](#_Toc56925561)

[3.5. Biến đổi tần số (Freequency Convolution) 29](#_Toc56925562)

[3.6. Bước sóng (Striding) 29](#_Toc56925563)

[3.7. Mô hình ngôn ngữ (Language model) 29](#_Toc56925564)

[3.8. Giải pháp xây dựng máy chủ 31](#_Toc56925565)

[3.9. Giải pháp xây dựng ứng dụng mẫu 32](#_Toc56925566)

# GIỚI THIỆU LUẬN VĂN

## GIỚI THIỆU ĐỀ TÀI

Cuộc sống ngày càng hiện đại, nhu cầu áp dụng những công nghệ tiên tiến để phục vụ các công việc trong cuộc sống của mỗi người càng cao. Hơn một thập kỷ gần đây, với sự phát triển mạnh mẽ của công nghệ thông tin, công nghệ xử lý ngôn ngữ tự nhiên như mã hóa, nhận dạng tiếng nói, giả lập giọng nói…đã trở thành xu hướng nghiên cứu mới được nhiều nhà khoa học quan tâm ở các lĩnh vực khác nhau như tin học, toán học, điều khiển, điện tử,…

Với xu hướng mới đó, vấn đề giao tiếp giữa con người và máy tính đang đòi hỏi những sự cải tiến vượt bậc. Công nghệ đã làm tăng lên tốc độ xử lý dữ liệu của các thiết bị hiện đại trong cuộc sống hàng ngày của chúng ta. Cùng với sự tăng lên của tốc độ xử lý dữ liệu, thời gian mà máy móc đợi con người nhập liệu cũng tăng lên. Nhìn chung, hầu hết các phương pháp nhập liệu truyền thống, có thể kể đến như nhập liệu bằng tay, nhập liệu thông qua các thiết bị điều khiển, ... có tốc độ và thời gian thao tác nhập dữ liệu tương đối chậm. Để giải quyết vấn đề này, giọng nói là giải pháp hợp lý nhất tính đến thời điểm hiện tại. Với sự trợ giúp của công nghệ, người dùng có thể dễ dàng điều khiển thiết bị và nhập tài liệu bằng giọng nói. Đó là chưa kể đến lợi ích to lớn mà những phần mềm dựa trên nhận dạng giọng nói có thể mang lại cho những người khiếm thị. Họ có thể tận hưởng những tiến bộ công nghệ tương tự như những gì mà một người bình thường có thể làm, không còn khoảng cách xuất hiện do những khiếm khuyết về giác quan. Nhận dạng giọng nói cho phép tạo tài liệu nhanh hơn vì phần mềm nói chung tạo ra các từ thông qua giọng nói thường nhanh hơn nhiều so với một người có thể nhập. Có thể nói, giọng nói đang ngày càng trở nên "quyền lực" hơn khi có thể điều khiển mọi thiết bị công nghệ hiện đại.

Và thực tế, nhận dạng giọng nói tự động đang dần bùng nổ, đem lại nhiều lợi ích và thuận lợi hơn cho con người. Nhận thấy được tiềm năng phát triển to lớn như vậy, hàng loạt các dịch vụ nhận dạng âm thanh tiếng nói cũng như là ứng dụng sử dụng dịch vụ này đã ra đời. Từ các tổ chức đến cá nhân, công nghệ này được sử dụng rộng rãi vì những lợi thế khác nhau mà nó mang lại.

Sau đây nhóm sinh viên sẽ nêu qua một số lĩnh vực mà công nghệ này mang lại.

### Phục vụ nhu cầu trong cuộc sống hằng ngày

Trong cuộc sống hằng ngày, việc sử dụng công nghệ nhận dạng giọng nói giúp giảm tiết kiệm nhiều thời gian, sức lực cho con người. Có thể kể đến như : điều khiển các thiết bị trong nhà thông qua giọng nói; tìm kiếm thông tin, tìm kiếm kênh trên tivi, máy tinh,... mà không cần nhập tay; nhắn tin với bạn bè khi đang bận một công việc khác; ghi chú lại các thông tin trong cuộc hội thảo, cuộc nói chuyện mà không cần tốn thời gian bằng việc ghi lại bằng tay,... Các trợ lý ảo trên điện thoại hoặc các mạch nhận dạng giọng nói trên các thiết bị đang thực hiện tốt các nhiệm vụ trên giúp con người có trải nghiệm tốt nhất trong cuộc sống thường nhật.

### Áp dụng vào nhiều lĩnh vực trong cuộc sống

Công nghệ nhận dạng giọng nói và việc sử dụng trợ lý ảo đã nhanh chóng được áp dụng vào nhiều lĩnh vực khác nhau trong cuộc sống, chẳng hạn như văn phòng, kinh doanh, y tế, vạn vật kết nối, giáo dục, ...

* **Trong lĩnh vực văn phòng**

Công nghệ nhận dạng giọng nói ở lĩnh vực văn phòng đã phát triển mạnh, có vai trò kết hợp các nhiệm vụ đơn giản để tăng hiệu quả, giảm các công việc truyền thống do con người thực hiện để tăng năng suất.

Các công việc mà trợ lý ảo có thể thực hiện được như: tìm kiếm tài liệu, văn bản, báo cáo,... trên thiết bị; nhận dạng thông tin muốn đưa vào máy tính; in, xuất tài liệu; thực hiện các công việc trong cuộc hội thảo; tạo các biểu đồ, bảng báo cáo bằng dữ liệu,... Tất cả các yêu cầu trên đều có thể giao tiếp với trợ lý ảo bằng giọng nói.

* **Trong lĩnh vực kinh doanh**

Nhận dạng giọng nói giúp tăng cường nhu cầu tìm kiếm bằng giọng nói, từ đó mở ra một khía cạnh mới mà nhà tiếp thị có thể tiếp cận người tiêu dùng. Dữ liệu giọng nói của người dùng có thể giải thích độ tuổi, tầng lớp và các thông tin liên quan đến sinh trắc nhân khẩu như văn hóa của họ. Các nhà tiếp thị có thể tối ưu hóa những thông tin này để đón đầu các xu hướng tiêu dùng.

* **Trong lĩnh vực y tế**

Đặc điểm của công nghệ nhận dạng giọng nói là giảm thiểu thời gian giao tiếp với thiết bị công nghệ. Do đó, việc áp dụng công nghệ này vào các bệnh viện sẽ tiết kiệm thời gian quý giá bởi việc truy cập thông tin rảnh tay, mang lại ích lợi vô cùng tích cực đến sự an toàn của bệnh nhân, nâng cao hiệu quả y tế. Một số thuận lợi cụ thể của nhận dạng giọng nói trong lĩnh vực ý tế như: tìm các hồ sơ y tế nhanh chóng; truy vấn thông tin về quản lý như số giường trống, số bệnh nhân hiện có, ... cải thiện quy trình làm việc; rút gọn thời gian nhập liệu; trong tương lai có thể có sự tương tác giữa bệnh nhân và trợ lý ảo tại nhà để tiết kiệm thời gian đợi ở bệnh viện,...

* **Trong lĩnh vực vạn vật kết nối**

Các trợ lý ảo trên điện thoại thông minh cho phép người sử dụng tương tác, điều khiển các thiết bị trong mạng lưới vạn vật kết nối (Internet of Things - IoT), đặc biệt là công nghệ nhà thông minh (Smart Home). Trong hiện tương lai, công nghệ nhận dạng có thể giúp người dùng trực tiếp ‘ra lệnh’ cho các thiết bị này mà không cần thiết phải chạm trực tiếp vào điện thoại. Điều này tạo cảm giác thoải mái, tiện lợi cho người dùng.

### Tiềm năng kinh tế cao

Nhận dạng giọng nói cho phép chuyển đổi giọng nói thành văn bản, giúp việc tạo và sử dụng thông tin trở nên dễ dàng hơn bởi vì văn bản dễ dàng hơn để lưu trữ, xử lý và sử dụng, cho cả máy tính và con người. Các ứng dụng được xây dựng để phục vụ nhu cầu nhận dạng và tích hợp các dịch vụ nhận dạng đang ngày càng phát triển và là lĩnh vực hấp dẫn đối với thị trường công nghệ. Sẽ rất thuận tiện nếu chúng ta có thể sử dụng phần mềm dựa trên nhận dạng giọng nói để mua hàng, kiểm tra thời tiết, gửi email, tìm kiếm thông tin trên internet theo một cách mới để tương tác với máy móc.

Các ứng dụng chụp ảnh giải trí trên các thiết bị di động không những mang lại giá trị cho bản thân người dùng mà còn đem đến những giá trị về kinh tế cho phía nhà phát triển. Bắt kịp xu thế và nhu cầu của người dùng, các doanh nghiệp, cùng với những cải tiến ấn tượng về khả năng hiểu ngôn ngữ tự nhiên và tỷ lệ chính xác của nhận dạng giọng nói, đã khiến các công ty ngày càng muốn tiếp tục xây dựng tính năng hỗ trợ giọng nói trải nghiệm thậm chí vượt ra ngoài phạm vi riêng tư, các công ty, doanh nghiệp chủ động hơn trong cách tiếp cận với nhận dạng giọng nói để giới thiệu hoặc kết hợp công nghệ nhận dạng giọng nói vào các sản phẩm của họ.

Hơn thế nữa, lĩnh vực nhận dạng giọng nói và tổng quan hơn là trí tuệ nhân tạo sẽ ngày càng trở nên phức tạp hơn trong tương lai. Ngành công nghiệp này đã và đang thu hút với hàng trăm công ty trên thế giới thử nghiệm và tích hợp sản phẩm và dịch vụ của họ với công nghệ nhận dạng giọng nói. Theo lời của Brian Roemmele, người sáng lập và Tổng biên tập của tạp chí Multiplex, "60 năm qua, con người đã thích nghi với máy tính. 60 năm tới, máy tính sẽ thích ứng với chúng ta. Tiếng nói của chúng ta sẽ là nền tảng cho điều đó, nó sẽ là một cuộc cách mạng và nó sẽ thay đổi mọi thứ. ".

Trên cơ sở đó, việc xây dựng các dịch vụ nhận dạng giọng nói cũng như các ứng dụng các dịch vụ này vào các phần mềm hiện có trong nhiều lĩnh vực của cuộc sống đem lại một tiềm năng kinh tế rất cao.

### Phục vụ nhu cầu sử dụng công nghệ cho đa dạng người trong xã hội

Công nghệ ngày càng hiện đại, một số tầng lớp của xã hội gần như bị ‘lạc hậu’ đi vì khó khăn trong việc sử dụng các thiết bị điện tử thông minh. Người lớn tuổi, người cận thị, người tiếp cận với thiết bị điện tử hiện đại muộn khó khăn trong việc nhập thông tin từ bàn phím, màn hình điện thoại di động với phím nhỏ; người khiếm thị khó khăn trong việc sử dụng các thiết bị thông minh,... Công nghệ nhận dạng giọng nói giúp giải quyết các khó khăn trên.

## LÝ DO LỰA CHỌN ĐỀ TÀI

Trong những năm gần đây, các nhà nghiên cứu đang tập trung vào công nghệ nhận dạng giọng nói và đã có một số thành công đối với việc nhận dạng tiếng Anh và một số ngôn ngữ khác. Triển khai, xây dựng mô hình và ứng dụng các mô hình này vào thực tế ứng dụng vấn đề này là một việc làm hết sức có ý nghĩa đặc biệt trong giai đoạn công nghiệp hoá hiện đại hoá hiện nay.

Đối với Tiếng Việt, Tiếng Việt là một trong ngôn ngữ có nhiều thanh điệu nhất thế giới, cùng với những đặc trưng riêng về vùng miền, mặc dù đã có một số công trình nghiên cứu về lĩnh vực nhận dạng giọng nói Tiếng Việt nhưng lĩnh vực này nói chung vẫn còn khá mới ở nước ta.

Áp dụng các kiến thức đã học của bản thân trong quá trình học tập cũng như quá trình tìm hiểu, nghiên cứu vào việc xây dựng mô hình có ý nghĩa lớn, có tiềm năng cao trong tương lai, nhóm sinh viên lựa chọn đề tài “Xây dựng mô hình nhận dạng âm thanh tiếng Việt”. Mục tiêu cơ bản của đề tài này, nhóm sinh viên muốn nghiên cứu, thu thập dữ liệu và xây dựng mô hình nhận dạng âm thanh Tiếng Việt hiệu quả, đạt độ chính xác cao so với các ứng dụng hiện có trên thị trường.

Ngoài ra, việc chọn đề tài này giúp nhóm sinh viên tiếp cận với lĩnh vực học máy, nghiên cứu tìm hiểu thông tin từ các nguồn tài liệu quý giá. Hơn thế nữa, sau khi thực hiện đề tài, nhóm sinh viên sẽ có thêm kinh nghiệm và hiểu biết về quy trình làm ra một dự án thực tế, không những thế chúng em còn được học hỏi các kiến thức chuyên môn liên quan đến học sâu, huấn luyện mô hình, và áp dụng vào xây dựng ứng dụng minh họa. Chính những điều đó sẽ là nền tảng quý báu hỗ trợ đắc lực cho chúng em trên con đường học vấn và việc làm trong tương lai.

## HƯỚNG PHÁT TRIỂN CỦA LUẬN VĂN

Tiếng Việt được coi là một ngôn ngữ khó học với người nước ngoài bởi ngữ pháp, thanh điệu và đặc trưng vùng miền. Máy tính cũng giống như người nước ngoài - để nó nghe hiểu và diễn giải được giọng nói tiếng Việt thành dạng văn bản không phải là việc dễ dàng. Nhận dạng tiếng nói đóng vai trò quan trọng trong giao tiếp giữa người và máy. Nó giúp máy móc hiểu và thực hiện các hiệu lệnh của con người. Hiện nay trên thế giới, lĩnh vực nhận dạng tiếng nói đã đạt được nhiều tiến bộ vượt bậc. Đối với ngôn ngữ tiếng Anh, việc nhận dạng có thể đạt độ chính xác tới 99%. Ở Việt Nam, lĩnh vực nhận dạng giọng nói còn khá mới và hiện độ chính xác nhìn chung chưa cao. Luận văn này, nhóm sinh viên sẽ xây dựng, huấn luyện mô hình nhận dạng âm thanh Tiếng Việt và xây dựng ứng dụng demo với mục tiêu cơ bản là:

* Mô hình được xây dựng đạt độ chính xác tối thiểu 75%.
* Ứng dụng mẫu sẽ được phát triển trên nền tảng web với chức năng chính nhận một tập tin âm thanh Tiếng Việt và chuyển sang văn bản Tiếng Việt, bên cạnh đó cho phép ghi âm trực tiếp âm thanh và tải lên tập tin âm thanh, văn bản sau khi chuyển đổi hỗ trợ tải về ở nhiều định dạng.

## MỤC TIÊU CỦA LUẬN VĂN

Để hoàn thành tốt đề tài luận văn, bản luận văn và sản phẩm cuối cùng của nhóm sinh viên sẽ đảm bảo các mục tiêu tối thiểu sau đây:

* Bản luận văn trình bày lý thuyết nền tảng và giải pháp để xử lý việc nhận 1 tập tin âm thanh tiếng Việt và xuất ra nội dung văn bản ở dạng tiếng Việt.
* Xây dựng, thu thập dữ liệu, và đào tạo mô hình để nhận 1 file âm thanh tiếng Việt và xuất ra nội dung ở dạng văn bản
* Cải tiến độ chính xác của mô hình với mục tiêu là 75%.
* Xây dựng ứng dụng web chuyển đổi từ giọng nói sang văn bản Tiếng Việt. Ứng dụng mẫu áp dụng mô hình được xây dựng là ứng dụng web dựa trên nền tảng Flask. Ứng dụng cho phép định dạng, chỉnh sửa văn bản trực tuyến (căn lề, kích thước chữ, phông chữ, định dạng chữ, …), bên cạnh đó ứng dụng cho phép người dùng tải lên tập tin âm thanh ít nhất 3 định dạng (.mp3, .acc, .wav). và cho phép người dùng tải xuống tập tin văn bản ít nhất 3 định dạng (.txt, .pdf, .doc). Ứng dụng cũng hỗ trợ ghi âm, nghe lại đoạn âm thanh đã ghi âm trước khi thực hiện chuyển đổi.

## PHẠM VI ĐỀ TÀI

Sản phẩm của đề tài “Xây dựng mô hình nhận dạng âm thanh tiếng Việt” là một mô hình cho phép nhận vào một tập tin âm thanh Tiếng Việt và trả về một đoạn văn bản Tiếng Việt. Nhóm sinh viên sẽ tập trung vào việc phát triển một cách đầy mô hình này dựa trên engine đã có sẵn trên thị trường. Khả năng nhận dạng ngôn ngữ Tiếng Việt của mô hình được đào tạo từ bộ dữ liệu âm thanh Tiếng Việt do nhóm tìm hiểu, thu thập từ các nguồn dữ liệu mở. Sản phẩm luận văn được thực hiện một cách toàn diện nên sẽ được áp dụng, mở rộng vào các dịch vụ web (API). Nhóm sẽ xây dựng một API sử dụng mô hình trên và một trang web mẫu sử dụng API này để biểu diễn mô hình.

# LÝ THUYẾT NỀN TẢNG

Ở chương 1, nhóm sinh viên đã đề xuất sử dụng phương pháp nhận dạng mẫu để nhận dạng giọng nói. Trong luận văn, nhóm sinh viên thực hiện xây dựng mô hình nhận dạng đầu cuối (end- to -end) sử dụng RNN (mạng nơ-ron hồi quy) và kết hợp với mô hình ngôn ngữ để phát triển mô hình. Sau đây nhóm sinh viên sẽ trình bày các kiến thức liên quan được ghi lại trong quá trình thực hiện luận văn để xây dựng một mô hình nhận dạng giọng nói.

## LÝ THUYẾT NỀN TẢNG CỦA NHẬN DẠNG ÂM THANH

### Tổng quan

#### Tổng quan nhận dạng âm thanh

#### Các hướng tiếp cận

**Tiếp cận âm học**

Phương pháp này dựa trên lý thuyết về Âm học-Ngữ âm học. Lý thuyết đó cho biết có sự tồn tại của các đơn vị ngữ âm trong ngôn ngữ tiếng nói, các đơn vị ngữ âm này được biểu diễn đặc trưng bởi một tập hợp những thuộc tính thể hiện trong tín hiệu âm thanh hay biểu diễn phổ theo thời gian. Đặc điểm của phương pháp nhận dạng tiếng nói theo hướng tiếp cận Âm học-Ngữ âm học:

* Người thiết kế phải có kiến thức khá sâu rộng về Âm học-Ngữ âm học.
* Phân tích các khối ngữ âm mang tính trực giác, thiếu chính xác.
* Phân loại tiếng nói theo các khối ngữ âm thường không tối ưu do khó sử dụng các công cụ toán học để phân tích.

**Tiếp cận nhận dạng mẫu thống kê**

Đây là một phương pháp sử dụng trực tiếp các mẫu tiếng nói (chính là đoạn tiếng nói cần nhận dạng) mà không cần xác định thật rõ các đặc trưng và cũng không cần phân đoạn tín hiệu. Phương pháp này cũng có 2 bước:

* Bước 1: thu thập các mẫu tiếng nói: sử dụng tập mẫu tiếng nói (cơ sở dữ liệu mẫu tiếng nói) để đào tạo các mẫu tiếng nói đặc trưng (mẫu tham chiếu) hoặc các tham số hệ thống.
* Bước 2: nhận dạng mẫu: đối sánh mẫu tiếng nói từ ngoài với các mẫu đặc trưng để ra quyết định. Cơ sở dữ liệu tiếng nói cho đào tạo có đủ các mẫu cần nhận dạng thì quá trình đào tạo có thể xác định chính xác các đặc tính âm học của mẫu, từ đó tăng độ chính xác cho mô hình.

Tiếp cận nhận dạng mẫu thường được lựa chọn cho các ứng dụng nhận dạng tiếng nói bởi các lý do sau:

* Tính dễ sử dụng và dễ hiểu trong thuật toán.
* Tính bất biến và khả năng thích nghi đối với những từ vững, người sử dụng, các tập hợp đặc trưng, các thuật toán so sánh mẫu và các quy tắc quyết định khác nhau.
* Khẳng định tính năng cao trong thực tế.
* Hiện nay, một số kỹ thuật nhận dạng mẫu được áp dụng thành công trong nhận dạng tiếng nói là lượng tử hóa vector, so sánh thời gian động (DTW), mô hình Markov ẩn (HMM), mạng nơron nhân tạo (ANN), sử dụng cơ sở tri thức,...

**Cách tiếp cận học máy**

Tiếp cận học máy là phương pháp cố gắng “máy móc hóa” chức năng nhận dạng theo cách mà con người áp dụng trí thông minh của mình trong việc quan sát, phân tích và thực hiện những quyết định trên các đặc trưng âm học của tín hiệu.

Cách tiếp cận này kết hợp các cách tiếp cận trên nhằm tận dụng tối đa các ưu điểm của chúng.

Đặc điểm của các hệ thống nhận dạng theo cách tiếp cận học máy:

• Sử dụng hệ chuyên gia để phân đoạn, gán nhãn ngữ âm. Điều này làm đơn giản hóa hệ thống so với phương pháp nhận dạng ngữ âm.

• Sử dụng mạng nơron nhân tạo để học mối quan hệ giữa các ngữ âm, sau đó dùng nó để nhận dạng tiếng nói.

Đây sẽ là hướng tiếp cận tương lai của nhận dạng tiếng nói.

### Đăc trưng âm học

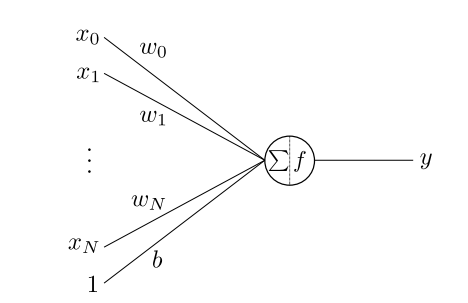
#### Bản chất của âm thanh

#### Ngữ âm Tiếng Việt

### Lý thuyết nền tảng của mạng nơ ron

Mạng nơ ron là một hệ thống học tập tính toán sử dụng một mạng các chức năng để hiểu và dịch đầu vào dữ liệu của một đầu vào này thành đầu ra mong muốn. Sử dụng một chuỗi các thuật toán cố gắng nhận ra các mối quan hệ trong một tập hợp dữ liệu thông qua một quy trình. Khái niệm về mạng lưới thần kinh nhân tạo được lấy cảm hứng từ sinh học của con người và cách các tế bào thần kinh của não người hoạt động để nhận biết đầu vào.

Mạng nơ-ron là một trong nhiều công cụ và cách tiếp cận được sử dụng trong học máy. Mạng nơ-ron đang được áp dụng cho nhiều vấn đề trong cuộc sống ngày nay, bao gồm nhận dạng giọng nói và hình ảnh, lọc email spam, chẩn đoán y tế, thị giác máy tính, …



Hình 2.1. Hình minh họa một nơ-ron mạng thần kinh điển hình. Nơron bao gồm đầu ra y, một tập hợp các đầu vào , trọng số đầu vào , độ sai lệch (bias) b và hàm kích hoạt f. Ở đây, độ sai lệch b được coi là trọng số của đầu vào của đơn vị cố định.

#### Cấu trúc tế bào mạng nơ-ron

Trong mục này, ta sẽ trình bày lí thuyết cơ sở về nơ-ron, đơn vị cơ bản của mạng nơ-ron.

Đối với một tế bào nơ-ron k đã cho có N đầu vào với tín hiệu và trọng số , và một độ sai lệch (bias) b ta có tế nào (nút) của mạng nơ-ron là một hàm của tập các trọng số tương ứng với các giá trị đầu vào.

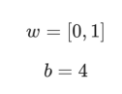
Trong đó:

* : trọng số của đầu vào
* : hàm kích hoạt (activation function)
* : độ sai lệch (bias)

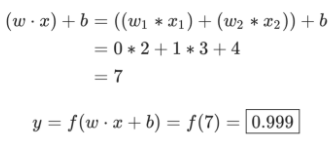
Ta sẽ sử dụng kí hiệu ma trận để làm đơn giản cách thể hiện khi đó đầu ra là:

Hàm kích hoạt được sử dụng để biến các đầu vào không bị ràng buộc thành đầu ra có hình thức và dễ dự đoán. Một hàm kích hoạt thường được sử dụng là sigmod sẽ được trình bày trong phần sau.

Lấy một ví dụ đơn giản, giả sử ta có một nơ-ron với hai đầu vào sử dụng hàm kích hoạt sigmoid và có các tham số sau:



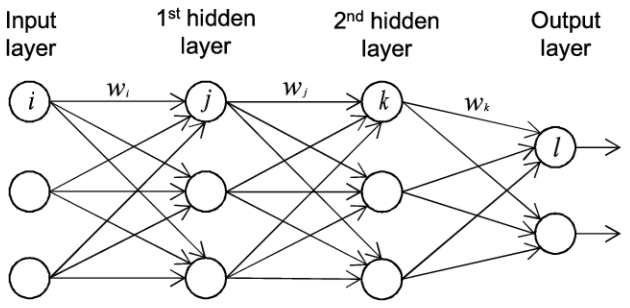
Và ta có một nơ-ron với đầu vào là x = [2, 3], ta sẽ tính đầu ra như sau:



Nơ-ron cho kết quả đầu ra 0,999 với các đầu vào x = [2,3]. Quá trình chuyển đầu vào về phía trước để đạt được đầu ra như ví dụ trên được gọi là chuyển tiếp (feedforward). Hình 1 đã mô tả kiến trúc của một nơ-ron mạng thần kinh điển hình.

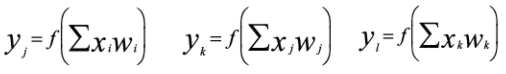
#### Kết hợp các tế bào mạng nơ-ron thành một mạng lưới mạng nơ-ron

Phần trên vừa trình bày chi tiết về một tế bào mạng nơ-ron, mà thực chất một mạng lưới thần kinh là một nhóm các tế bào thần kinh được kết nối với nhau.



Hình 2.2. Khối xây dựng của mạng nơ-ron sâu

Với các đầu ra được tính như sau, các đầu ra ở các lớp ẩn có thể cho kết quả như nhau.



Hình 2 mô tả cấu trúc mạng nơ- ron tổng quan, mạng lưới mạng nơ- ron thường được tổ chức theo từng lớp, có 3 lớp cơ bản: lớp đầu vào (Input layer) và lớp đầu ra (output layer) chỉ gồm 1 layer , hidden layer có thể có 1 hay nhiều layer tùy vào bài toán cụ thể. Các lớp được tạo thành từ nhiều lớp được kết nối với nhau chứa một hàm kích hoạt.

Kiến trúc như trên được gọi là kiến trúc của mạng nơ- ron nhân tạo Artificial Neural Network (ANN). Trong ANN, trừ lớp đầu vào thì tất cả các nút thuộc các lớp khác đều kết nối đầy đủ (full-connected) với các nút thuộc lớp trước nó. Mỗi nút thuộc lớp ẩn nhận vào ma trận đầu vào từ lớp trước và kết hợp với trọng số để cho ra được kết quả.

**Lớp đầu vào (Input layer)**

Mục đích của lớp đầu vào là nhận giá trị đầu vào của các thuộc tính cho mỗi quan sát. Lớp đầu vào trình bày các mẫu cho mạng, mạng này giao tiếp với một hoặc nhiều lớp ẩn. Các nút của lớp đầu vào là thụ động, nghĩa là chúng không thay đổi dữ liệu trong mạng. Chúng nhận một giá trị duy nhất trên đầu vào của họ và nhân đôi giá trị đó và tạo ra nhiều đầu ra và gửi đến tất cả các nút ẩn.

**Lớp ẩn (Hidden layer)**

Các lớp ẩn áp dụng các phép biến đổi đã cho cho các giá trị đầu vào bên trong mạng. Trong đó, nó nhận giá trị từ đầu vào hoặc từ nút ẩn trước đó, tính toán giá trị đầu ra tại nút đó và sử dụng giá trị đầu ra đó, kết nối làm đầu vào cho nút ẩn tiếp theo hoặc đầu ra của mạng. Trong lớp ẩn, quá trình xử lý thực tế được thực hiện thông qua một hệ thống các kết nối có trọng số, có thể có một hoặc nhiều lớp ẩn. Các giá trị nhập vào một nút ẩn nhân với trọng số, trọng số là một tập hợp các số xác định trước. Các đầu vào có trọng số sau đó được thêm vào để tạo ra một số duy nhất.

**Lớp đầu ra (Output layer)**

Các lớp ẩn cuối cùng trong hệ thống các lớp ẩn sẽ liên kết với một lớp đầu ra. Lớp đầu ra nhận các kết nối từ các lớp ẩn hoặc từ lớp đầu vào (nếu mạng nơ- ron chỉ có một lớp ẩn). Nó trả về một giá trị đầu ra tương ứng với dự đoán của biến phản hồi. Trong các bài toán phân loại, thường chỉ có một nút đầu ra. Các nút hoạt động của lớp đầu ra kết hợp và thay đổi dữ liệu để tạo ra các giá trị đầu ra. Khả năng của mạng nơ-ron để cung cấp thao tác dữ liệu có ích nằm ở việc lựa chọn trọng số thích hợp. Điều này khác với xử lý thông tin thông thường.

#### Hàm kích hoạt (Activation function)

Các hàm kích hoạt mạng thần kinh là một thành phần quan trọng của học sâu. Các hàm kích hoạt xác định đầu ra của mô hình học sâu, độ chính xác của nó và cả hiệu quả tính toán của việc đào tạo một mô hình. Các hàm kích hoạt cũng có ảnh hưởng lớn đến khả năng hội tụ và tốc độ hội tụ của mạng nơ-ron, hoặc trong một số trường hợp, các hàm kích hoạt có thể ngăn các mạng nơ-ron hội tụ ngay từ đầu. Các hàm kích hoạt này có thể được sử dụng trong các lớp ẩn (để tính giá trị output cho nút) hoặc được sử dụng trên lớp đầu ra.

Hàm kích hoạt là phần rất quan trọng trong mạng nơ-ron nhiều lớp ẩn. Nếu không có hàm kích hoạt phi tuyến tính, cho dù mạng nơ-ron có nhiều lớp ẩn đến cỡ nào thì cũng chỉ có sức mạnh đại diện cho phân loại tuyến tính, điều này tương đương với một mạng mà không có lớp ẩn nào. Vì bản chất tổng hợp các hàm tuyến tính là một hàm tuyến tính. Do đó, hàm kích hoạt là một hàm phi tuyến tính được áp dụng cho đầu ra tại mỗi nút, cho phép mạng nơ-ron nhiều lớp ẩn học các hàm phi tuyến phức tạp.

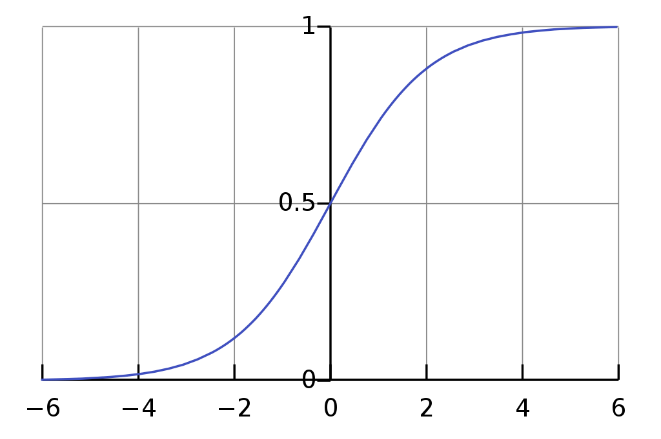
Tiếp theo, nhóm sẽ trình bày một số hàm kích hoạt thường được sử dụng trong mạng nơ- ron.

**Sigmoid**

Hàm kích hoạt được chọn một cách cổ điển là hàm sigmoid, ánh xạ các giá trị đầu vào vào khoảng biên từ 0 đến 1, do đó, nó đặc biệt được sử dụng cho các mô hình mà phải dự đoán xác suất như một đầu xác suất đầu ra (vì xác suất tồn tại trong khoảng từ 0 đến 1). Có công thức theo phương trình:

Trong đó:

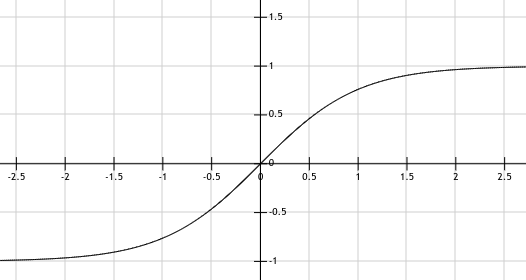
* : trọng số tại lớp thứ i
* : tham số sai lệch tại lớp thứ i
* : hàm kích hoạt



Hình 2.3. Hình minh họa đồ thị hàm kích hoạt sigmoid (Nguồn Wikipedia)

**Tanh**

Bên cạnh đó, có một hàm kích hoạt luôn hoạt động tốt hơn hàm sigmoid là hàm tiếp tuyến hyperbolic (hyperbolic tangent function. Hàm nhận đầu vào là một số thực và chuyển thành một giá trị trong khoảng (-1; 1). Cũng như Sigmoid, hàm Tanh bị bão hoà ở 2 đầu (gradient thay đổi rất ít ở 2 đầu). Tuy nhiên hàm Tanh lại đối xứng qua 0 nên khắc phục được một nhược điểm của Sigmoid. Có công thức là:

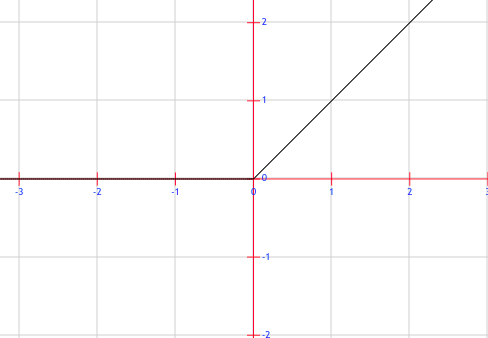


Hình 2.4. Minh họa hàm kích hoạt tanh. (Nguồn Wikipedia)

**ReLU (Rectified Linear Unit)**

Với nhược điểm tốc độ hội tụ chậm khi z quá lớn của hàm sigmoid và hàm tanh, hàm ReLU đang được sử dụng khá nhiều trong những năm gần đây khi huấn luyện các mạng neuron. ReLU đơn giản lọc các giá trị < 0. Có công thức như sau:

ReLU cung cấp một số ưu điểm vượt trội. Thứ nhất, ReLU có tốc độ hội tụ nhanh hơn hẳn, điều này có thể do ReLU không bị bão hoà ở hai đầu như hàm Sigmoid và Tanh. Hơn thế nữa, ReLU cung cấp khả năng tính toán nhanh hơn. Hàm Tanh và Sigmoid sử dụng hàm mũ và công thức phức tạp hơn ReLU rất nhiều do vậy sẽ tốn nhiều chi phí hơn để tính toán.



Hình 2.8 Minh họa hàm kích hoạt ReLu. (Nguồn: Wikipedia)

**Softmax**

Ngoài ra, có một hàm kích hoạt rất thú vị là hàm kích hoạt Softmax, hàm kích hoạt này thường được sử dụng trong lớp đầu ra của mạng nơ- ron. Nguyên nhân là vì, hàm kích hoạt Softmax không chỉ ánh xạ đầu ra của chúng ta trong phạm vi [0,1] mà còn ánh xạ từng đầu ra theo cách sao cho tổng các đầu ra là 1. Do đó, đầu ra của Softmax là một phân phối xác suất.

Hàm softmax lấy các kích hoạt tuyến tính của lớp cuối cùng làm đầu vào. Đối với đơn vị thứ i của lớp ẩn, kích hoạt softmax được cho bởi:

Với và

Trong đó:

* là đầu ra thứ i tại lớp ẩn trước lớp đầu ra, cũng là đầu vào thứ i tại lớp đầu ra.
* Đầu vào z của lớp đầu ra là vector với K số thực.

Có thể nhận thấy rằng, Softmax được sử dụng để đa phân loại trong mô hình hồi quy logistic trong khi Sigmoid được sử dụng để phân loại nhị phân trong mô hình hồi quy logistic.

#### Hàm mất mát (Loss function)

#### Lan truyền ngược (Back propagation)

#### Học với lan truyền ngược

#### Độ giảm dốc

### Các phương pháp huấn luyện mạng nơ ron

#### Chuẩn hóa theo lô

#### Dropout

### Mạng nơ ron hồi quy

#### Tổng quan

#### BiRNN

#### GRU

#### LSTM

### Kỹ thuật trong nhận dạng âm thanh

#### Phân tích đặc trưng

#### Phân loại thời gian kết nối CTC

#### Thuật toán tham lam, beam search

## HỆ THỐNG NHẬN DẠNG ÂM THANH

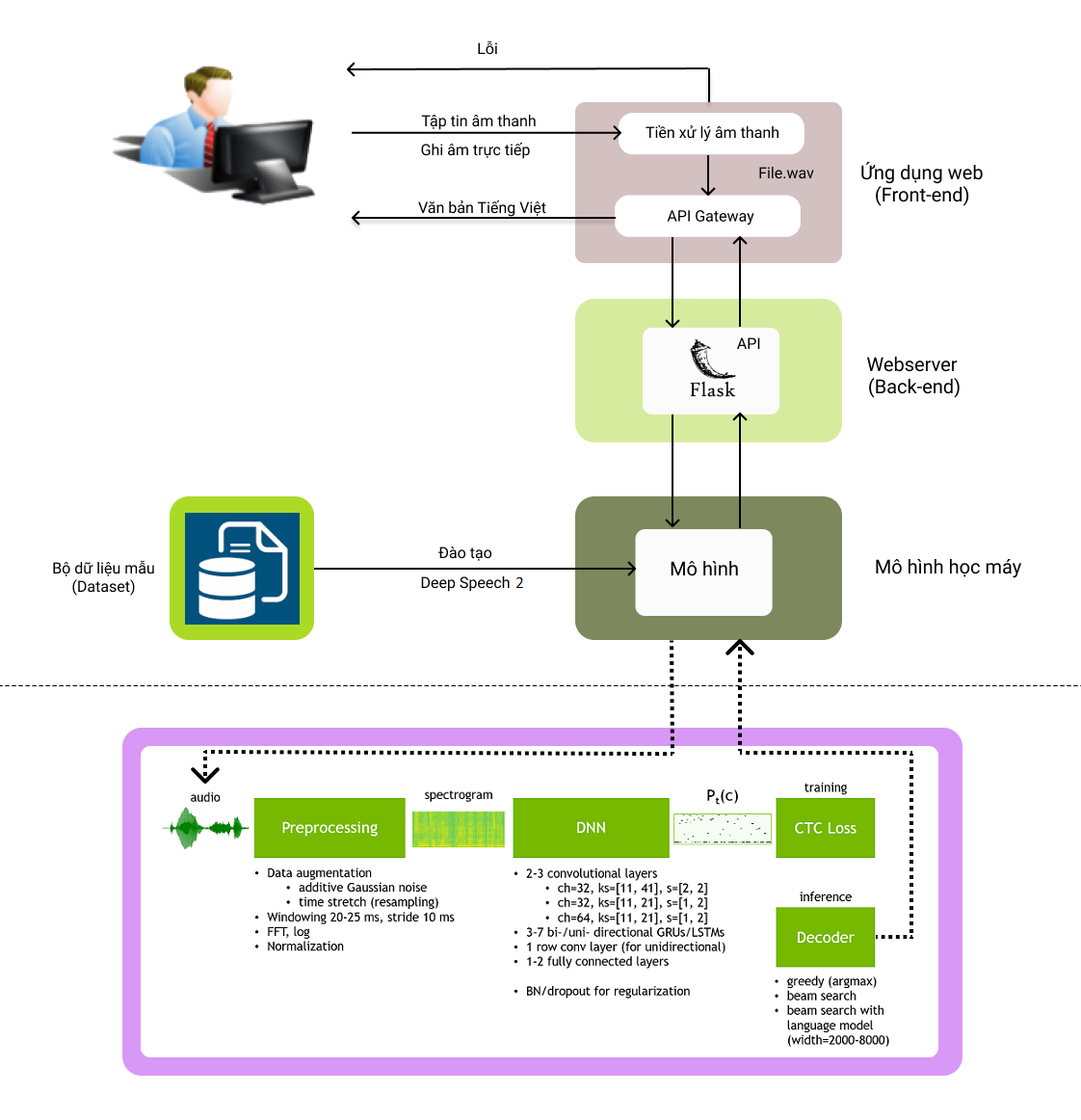
### Hệ thống nhận dạng âm thanh truyền thống

### Hệ thống nhận dạng âm thanh theo hướng đầu cuối (end to end)

# GIẢI PHÁP XÂY DỰNG MÔ HÌNH ÂM THANH

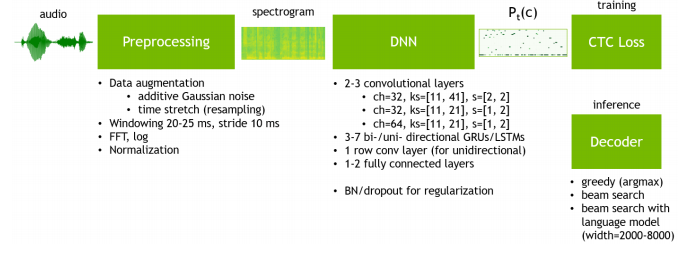
## Tổng quan kiến trúc hệ thống

Nhóm sinh viên dự kiến xây dựng sản phẩm là trang web với kiến trúc và mô hình được minh họa trong hình 3.1



Hình 3.1

Đề tài dự kiến sử dụng kiến trúc mô hình nhận dạng đầu cuối (End-to-End), được xây dựng kết hợp cùng ý tưởng DeepSpeech 2, một nghiên cứu của Baidu được công bố vào ngày 08/12/2015 tại Silicon Valley AI Lab. Nội dung của bài báo trình bày về nhận dạng giọng nói được thực hiện trên ngôn ngữ Tiếng Anh (English) và tiếng Quan Thoại (Mandarin).

Deep Speech 2 sử dụng mô hình Mạng nơ-ron hồi quy (RNN – Recurrent Neural Network) và sử dụng Connectionist Temporal Classification (CTC) để dự đoán đầu ra:

Đầu vào của hệ thống là đoạn âm thanh thô chưa được xử lý. Phần tiền xử lý (Preprocessing) lấy một tín hiệu dạng sóng âm thanh thô và chuyển nó thành một biểu đồ phổ có kích thước (N\_timesteps, N\_frequency\_features). N\_timesteps phụ thuộc vào thời lượng của tệp âm thanh gốc, N\_frequency\_features có thể được chỉ định trong tệp cấu hình của mô hình dưới dạng thông số “num\_audio\_features”.

Phần Deep Neural Network (DNN) tạo ra phân phối xác suất Pt(c) trên các ký tự từ vựng c cho mỗi bước thời gian *t*. Deep Speech 2 được đào tạo với nhiều thử nghiệm với mạng nơ-ron được huấn luyện với chức năng suy giảm phân loại theo thời gian kết nối (CTC) để dự đoán phiên âm giọng nói từ âm thanh. Tỷ lệ lỗi từ (WER) là số liệu đánh giá độ chính xác của mô hình.

Để đưa các từ ra khỏi một mô hình được đào tạo, cần sử dụng một bộ giải mã (Decoder). Bộ giải mã chuyển đổi phân phối xác suất trên các ký tự thành văn bản. Có hai loại bộ giải mã thường được sử dụng với các mô hình dựa trên CTC: bộ giải mã tham lam (Greedy decoder) và bộ giải mã tìm kiếm chùm (Beam search decoder) với mô hình ngôn ngữ. Một bộ giải mã tham lam xuất ra ký tự có thể xảy ra nhất ở mỗi bước thời gian. Nó có tốc độ xử lý nhanh và có thể tạo ra các câu rất chính xác, nhưng có thể mắc nhiều lỗi chính tả nhỏ. Tuy nhiên, do bản chất của chỉ số WER, một lỗi ký tự cũng làm cho một từ không chính xác. Một bộ giải mã tìm kiếm chùm có chức năng ghi lại mô hình ngôn ngữ cho phép kiểm tra nhiều giải mã bằng cách chỉ định điểm cao hơn cho nhiều N-grams tùy vào mô hình ngôn ngữ nhất định. Mô hình ngôn ngữ cũng giúp sửa lỗi chính tả. Nhược điểm là nó chậm hơn đáng kể so với một bộ giải mã tham lam.

Đầu ra của hệ thống là một đoạn văn bản Tiếng Việt hoàn chỉnh.

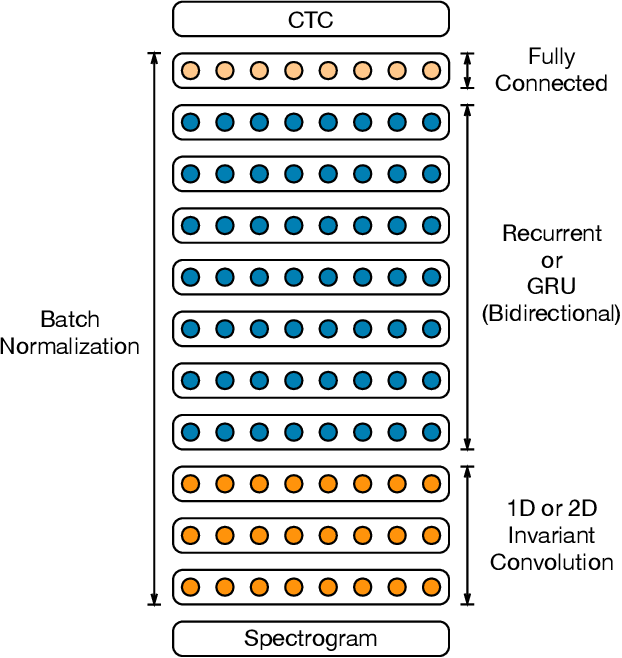
## Giải pháp xây dựng mô hình âm thanh

### Tổng quan giải pháp

Một mô hình nhiều lớp đơn giản với một lớp hồi quy không thể khai thác hàng nghìn giờ dữ liệu lời nói đã được gắn nhãn. Để mô hình học tập được tập dữ liệu lớn, cần phải tăng dung lượng mô hình thông qua chiều sâu. Nhóm sinh viên dự kiến sử dụng kiến trúc mô hình có 11 lớp bao gồm nhiều lớp hồi quy hai chiều (bidirectional recurrent) và lớp tích tụ (convolutional). Đồng thời, để tối ưu hóa mô hình, nhóm sử dụng chuẩn hóa hàng loạt cho RNN và một chương trình tối ưu hóa mới gọi là SortaGrad.

### Mô hình mạng nơ-ron hồi quy RNN

Một mạng nơ tron hồi quy được đào tạo để nhận các âm phổ đầu vào và cho ra các đoạn văn bản minh họa trong hình 3.2



Hình 3.2

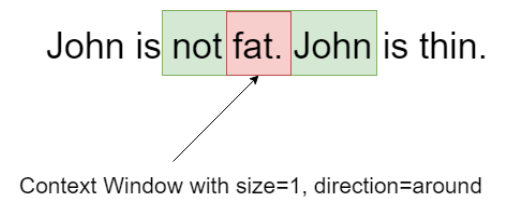
Cụ thể, trong một tập huấn luyện với x là phát âm đầu vào và y là nhãn văn bản Tiếng Việt tương ứng với đầu ra. Mỗi phát âm là một chuỗi thời gian có độ dài là . Trong đó, mỗi lát cắt thời gian nhất định của là một vector của các đặc trưng âm thanh , t = 1 ,2 , ... .

Nhóm sinh viên tiếp tục sử dụng âm phổ của các đoạn âm thanh được chuẩn hóa làm các đặc trưng cho hệ thống, do đó biểu thị cho độ lớn của khoảng tần số thứ *p* trong khung âm thanh tại thời điểm *t.*

Mục tiêu của RNN là chuyển đổi đầu vào thành chỗi ký tự cho nhãn tương ứng. Với mỗi khoảng thời gian *t,* RNN cho ra một xác suất dự đoán trên tập hợp các ký tự: với là một ký tự trong bảng chữ cái hoặc khoảng trắng, nháy đơn,…

Mô hình RNN mà nhóm dự kiến xây dựng bao gồm một số lớp đơn vị ẩn (hidden unit), một hoặc nhiều lớp phức hợp (convolution layer), tiếp đó là một hoặc nhiều lớp hồi quy (recurrent layer), cuối cùng là một hoặc nhiều lớp kết nối đầy đủ (connected layer)

**Lớp tích chập:** đơn vị biểu diễn ẩn tại lớp được đặt là với quy ước biểu diễn cho đầu vào . Phần dưới cùng của mạng là một hoặc nhiều vòng chập theo thời gian của đầu vào. Cửa sổ ngữ cảnh (Context window) là số lượng từ được dự đoán có thể xuất hiện trong phạm vi của từ đã cho (Hình 3.2)



Đối với cửa sổ ngữ cảnh (context window) có kích thước *c*, nút kích hoạt thứ *i* ở bước thời gian *t* của lớp tích chập (the i-th activation at time-step t of the convolutional layer) được đưa ra bởi công thức :

).

Trong đó: biểu thị yếu tố khôn ngoan giữa bộ lọc thứ *i* và của sổ ngữ cảnh của các lớp kích hoạt trước đó và hàm *f* là một hàm phi tuyến tính bậc nhất.

Hàm chỉnh lưu tuyến tính (ReLU) được rút gọn thành làm độ phi tuyến tính cho mô hình.

Lấy mẫu phụ là một kỹ thuật đã được phát minh ra để giảm sự phụ thuộc vào định vị chính xác trong các danh sách các ánh xạ đặc trưng được tạo ra bởi các lớp phức hợp trong CNN. Mẫu phụ là một trong bộ nhỏ () được chọn bằng cách lấy ngẫu nhiên một cách đơn giản từ mẫu được sử dụng để bắt đầu lặp lại trên mẫu đó (. Nhóm sinh viên dự kiến lấy mẫu phụ bằng cách tách một số khung tính chập trong một số lớp, thường là lớp đầu tiên. Mục đích của việc này là rút ngắn số bước thời gian cho các lớp hồi quy ở trên.

**Lớp hồi quy hai chiều:** các lớp hồi quy kích hoạt tiến theo thời gian ( và lùi theo thời gian () được tính bằng công thức :

Hai tập hợp các nút kích hoạt được cộng lại để tạo thành các đầu ra cho lớp .

Hàm có thể rút gọn bằng phép hồi quy:

(1)

Trong đó, là trọng số của các đơn vị ẩn trong ma trận, là trọng số của ma trận hồi quy và là số hạng chệch.

**Lớp kết nối đầy đủ**: sau các lớp hồi quy hai chiều, nhóm áp dụng một hoặc nhiều lớp kết nối đầy đủ với .

Đầu ra của lớp là một phép tính softmax phân bố xác suất (computing a probability distribution) trên các ký tự được cho bởi công thức:

Mô hình được huấn luyện bằng cách sử dụng hàm mất mát CTC (CTC loss function), với (x,y) là cặp giá trị đầu vào và đầu ra, là các tham số hiện tại của mạng.

Đạo hàm L theo các tham số của mạng được sử dụng đề cập nhật thông số của mạng thông qua sự lan truyền ngược (backpropagation) bằng thuật toán thời gian.

## Chuẩn hóa hàng loạt (Batch normalization)

Để mở rộng mô hình khi tăng thêm tập dữ liệu huấn luyện, nhóm sinh viên dự kiến tăng độ sâu của mạng bằng cách thêm nhiều lớp ẩn hơn thay vì làm cho các lớp lớn hơn. Chuẩn hóa hàng loạt là một kỹ thuật đào tạo các mạng nơ-ron theo chiều sâu, chuẩn hóa các đầu vào thành một lớp cho từng lô nhỏ. Mục đích chính của chuẩn hóa hàng loạt là giúp ổn định quá trình đào tạo và giảm số lượng epoch cần thiết để đào tạo mô hình.

Một phép biến đổi BatchNorm có công thức :

Trong đó:

* E và Var là giá trị trung bình và phương sai thực nghiệm trên một lô nhỏ.
* Các tham số cho phép lớp chia tỉ lệ và dịch chuyển từng đơn vị ẩn như mong muốn, hằng số rất nhỏ dùng để ổn định số.

Trong các lớp tích chập, giá trị trung bình và phương sai được ước tính trên các đơn vị đầu ra tạm thơi cho một bộ lọc tích chập nhất định trên lô nhỏ.

Có hai phương pháp mở rộng BatchNorm cho RNN hai chiều:

Thêm một phép biến đổi BatchNorm ngay trước mọi trường hợp phi tuyến tính, phương trình g(.) (1) trở thành :

.

Trong phương pháp này, trung bình và phương sai được tích lũy qua một bước thời gian duy nhất của một lô nhỏ. Kỹ thuật này không dẫn để cải tiến trong việc tối ưu hóa.

Tính toán hồi quy theo công thức :

Qua mỗi đơn vị ẩn, trung bình và phương sai được tính toán trên tất cả các lô nhỏ theo chiều dài của chuỗi. Sử dụng kỹ thuật này, nhóm có thể đánh giá một lời nói duy nhất tại một thời điểm với kết quả tốt hơn so với đánh giá một loạt lớn đầu vào.

## SortaGrad

Khi huấn luyện các âm thanh mẫu có độ dài khác nhau đặt ra một số thách thức về thuật toán. Một giải pháp khả thi là cắt bớt sự lan truyền ngược theo thơi gian để tất cả mẫu có độ dài và trình tự giống nhau trong quá trình đào tạo. Tuy nhiên, điều này hạn chế khả năng học tập dài hạn cho mô hình.

Hàm chi phí CTC (CTC Cost Function) mà nhóm sẽ sử dụng ngầm định phụ thuộc vào độ dài câu nói.

(2)

Trong đó, Align (x, y) là tập hợp tất cả các căn chỉnh có thể có của các ký tự của phiên mã y với các âm thanh đầu vào x dưới toán tử CTC. Ví dụ: từ SUN có các Align S-UN, SU-N, -SUN,…

Trong phương trình (2), số hạng bên trong là tích theo các bước thời gian của dãy, số hạng này thu hẹp lại theo độ dài của dãy vì . là tích các nên < 1, do đó có thể nhỏ hơn 1, -log của tổng này là đầu ra của phương trình.

thúc đẩy chương trình đào tạo được gọi là SortaGrad, sử dụng độ dài của câu làm kinh nghiệm cho độ khó, vì các câu nói dài có chi phí cao hơn những câu nói ngắn.

## Biến đổi tần số (Freequency Convolution)

Tích chập tạm thời (Temporal convolution) thường được sử dụng trong nhận dạng giọng nói để mô hình hóa hiệu quả theo thời gian cho các phát âm có độ dài thay đổi. Nhiều mô hình giọng nói có lớp đầu tiên xử lý các khung đầu vào với một cửa sổ ngữ cảnh. Ngoài ra, lấy mẫu phụ là điều cần thiết để mạng nơ-ron hồi quy có thể kiếm soát tính toán với âm thanh có tốc độ mẫu cao (hign sample-rate).

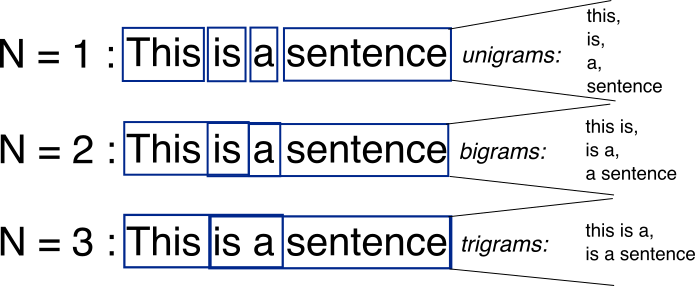
Các biến đổi trong miền tần số và thời gian khi áp dụng cho các đặc trưng đầu vào trước bất kỳ quá trình xử lý nào khác có thể cải thiện hiệu suất ASR. Sự biến đổi tần số cố gắng mô hình hóa phương sai do dự biến đổi của loa ngắn hơn so với các mạng kết nối đầy đủ lớn. Vì thứ tự âm phổ của các đặc trưng bị loại bỏ bởi các lớp kết nối đầy đủ và lớp hồi quy, nên biến đổi tần số hoạt động tốt hơn trên các lớp đầu tiên của mạng.

## Bước sóng (Striding)

Trong các lớp tích chập, áp dụng bước sóng dài hơn và bối cảnh (context) rộng hơn để tăng tốc quá trình đào tạo vì chỉ cần ít bước thời gian hơn để mô hình hóa một đầu vào. Việc lấy âm thanh mẫu đầu vào thông qua FFT và bước sóng chập làm giảm số lượng bước thời gian và tính toán cần thiết trong các lớp, nhưng làm giảm hiệu suất.

## Mô hình ngôn ngữ (Language model)

Mô hình RNN được đào tạo qua bộ dữ liệu lớn. Để mạng có thể thành thạo chính ram không có ràng buộc về ngôn ngữ bên ngoài, cải thiện WER, nhóm dự kiến sử dụng mô hình ngôn ngữ n-gram vì chúng mở rộng quy mô phù hợp với lượng lớn văn bản không được đánh nhãn.



Trong quá trình tìm hiểu, phiên âm y tối đa hóa Q(y):

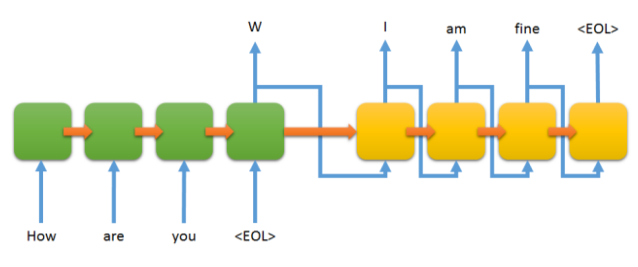
(3)

Trọng số kiểm soát các đóng góp tương đối của mô hình ngôn ngữ và mạng CTC.

Trọng số khuyến khích nhiều từ hơn trong phiên âm.

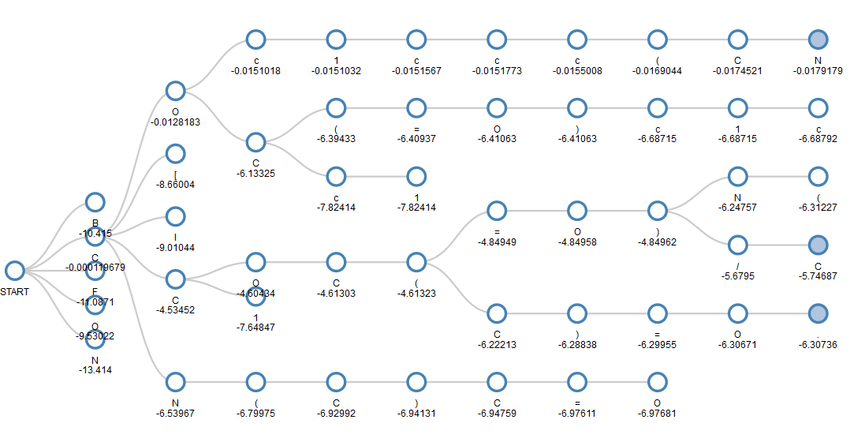
Các tham số này được điều chỉnh nhiều hơn trong một tập phát triển, nhóm dự kiến sử dụng thuật toán tìm kiếm chùm (BeamSearch) để tìm phiên âm tối ưu.

Ý tưởng của BeamSearch là tìm kiếm từ có xác suất xuất hiện cao nhất, và từ này sẽ là vector đầu vào để dự đoán từ tiếp theo.



Hình 3.4

Thuật toán BeamSearch có tham số gọi là beam width, tại mỗi bước dự đoán, thay vì chọn từ xác suất lớn nhất, nhóm chọn beam witdh có xác suát cao nhất. Tiếp tục tính toán các bộ xác suất cho đến khi kết thúc, ta sẽ thu được một cây xác xuất như hình 3.5

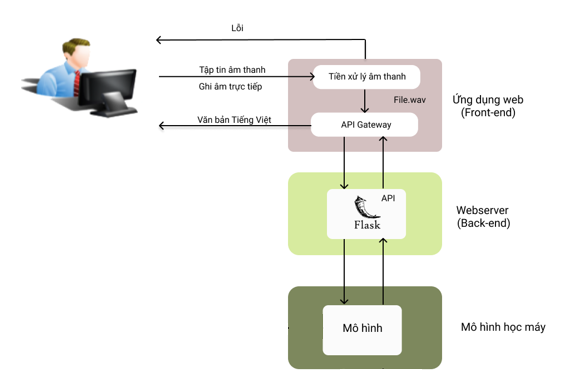


Hình 3.5

Nhân các xác suất của cả chuỗi có điều kiệ, câu kết quả là câu có xác xuất lớn nhất.

## Giải pháp xây dựng máy chủ

Nhóm sinh viên dự kiến xây dựng máy chủ với mục đích tạo ra giao diện lập trình ứng dụng (API), làm trung gian giữa ứng dụng mẫu và mô hình âm thanh đã được huấn luyện.



Máy chủ cho phép ứng dụng mẫu gửi đến một đoạn âm thanh dạng wav của tiếng Việt, sử dụng mô hình học máy xử lý đoạn âm thanh và trả về đoạn văn bản tiếng Việt cho ứng dụng mẫu.

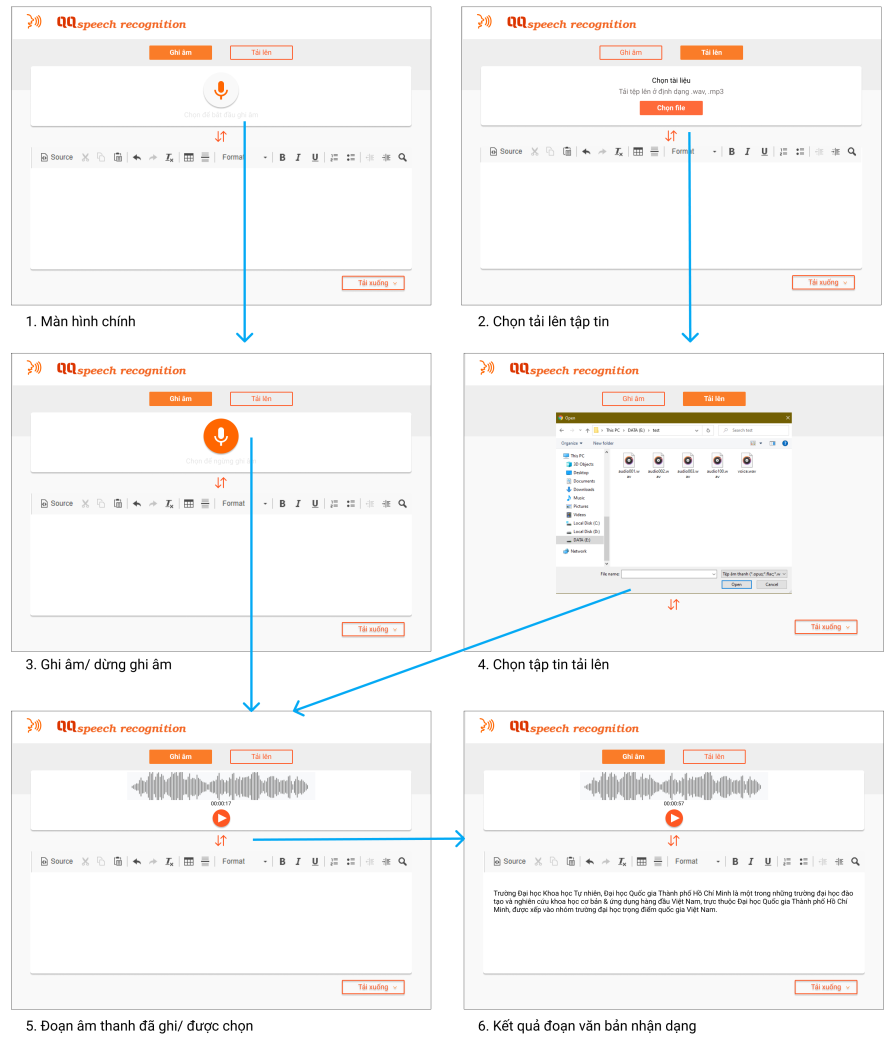
Nhóm lựa chọn sử dụng Framework Flask của python để xây dựng máy chủ. Vì Flask dễ cài đặt và triển khai, phù hợp với ứng dụng web quy mô vừa và nhỏ, giúp tập trung phát triển mục tiêu của nhóm là xây dựng giao diện lập trình ứng dụng, dễ dàng triển khai trên các dịch vụ máy chủ đám mây.

Sau khi xây dựng máy chủ, nhóm dự kiến triển khai máy chủ lên nền tảng đám mây Heroku.

## Giải pháp xây dựng ứng dụng mẫu

Theo giải pháp về hệ thống, nhóm sinh viên xây dựng một ứng dụng mẫu với nhiệm vụ cho phép người dùng sử dụng dịch vụ chuyển đổi âm thanh tiếng Việt thành văn bản tiếng Việt. Ứng dụng mẫu sử dụng giao diện lập trình ứng dụng (API) của máy chủ vừa được tạo ở trên.

Bản mẫu giao diện ứng dụng:



|  |  |  |  |
| --- | --- | --- | --- |
| Prior State | New State | Valid Transition | Comment |
| 1 | 1 | N |  |
| 1 | 2 | Y | Xe vào bãi và được nhận dạng |
| 1 | 3 | N |  |
| 1 | 4 | N |  |
| 1 | 5 | N |  |
| 1 | 6 | N |  |
| 2 | 1 | N |  |
| 2 | 2 | N |  |
| 2 | 3 | Y | Xe được xử lí nhân dạng khi vào |
| 2 | 4 | Y | Xe vừa vào và nhận diện ra |
| 2 | 5 | N | Xe chưa vào bãi xe |
| 2 | 6 | Y | Hủy gửi xe |
| 3 | 1 | N |  |
| 3 | 2 | N | Xe chưa ra khỏi bãi xe |
| 3 | 3 | N |  |
| 3 | 4 | Y | Lấy xe ra |
| 3 | 5 | Y | Nhận dạng ra thất bại, |
| 3 | 6 | Y | Xe chưa nhận dạng ra |
| 4 | 1 | N |  |
| 4 | 2 | N | Xe chưa ra khỏi bãi xe |
| 4 | 3 | N | Xe chưa nhận dạng ra |
| 4 | 4 | N |  |
| 4 | 5 | Y | Xe nhận dạng ra thất bại |
| 4 | 6 | Y | Xe nhận dạng ra thành công, ra khỏi bãi xe |
| 5 | 1 | N |  |
| 5 | 2 | N | Xe nhận dạng ra sai, chưa ra khỏi bãi xe |
| 5 | 3 | Y | Xe nhận dạng ra sai, bị giữ lại |
| 5 | 4 | Y | Nhận dạng ra lại cho xe |
| 5 | 5 | N |  |
| 5 | 6 | N | Xe bị không hợp lệ không được ra khỏi bãi xe |
| 6 | 1 | N |  |
| 6 | 2 | Y | Xe ra khỏi bãi và gửi lại |
| 6 | 3 | N | Xe chưa được kiểm tra nhận dạng vào |
| 6 | 4 | N | Xe chưa được kiểm tra nhận dạng vào |
| 6 | 5 | N | Xe chưa được kiểm tra nhận dạng vào |
| 6 | 6 | N |  |