Mục lục

Danh sách hình ảnh, biểu đồ

Danh sách Bảng

Bảng thuật ngữ

Tên thuật ngữ	Mô tả
Deep Learning	Học sâu
Active Learning	Phương pháp học chủ động
Acoustic Model	Mô hình âm học
Language Model	Mô hình ngôn ngữ
Acoustic Score	Trọng số mô hình âm học
Language Model Score	Trọng số mô hình ngôn ngữ
HMM hoặc Hidden Markov Model	Mô hình Markov ẩn
GMM	Gaussian Mixture Model (Mô hình Gaussian hỗn hợp)
Hybrid	Phương pháp lai
RNN	Mạng nơ ron hồi quy
CTC layer	Connectionist temporal classification
Attention	Cơ chế tập trung, chú ý
LF-MMI	Lattice-free maximum mutual information
End-to-End	Phương pháp học đầu-cuối mà không cần qua nhiều bước trung gian
WER	Word Error Rate - Tỉ lệ lỗi theo từ của câu được nhận dạng để đánh giá độ chính xác của một hệ thống nhận dạng tiếng nói (Tỉ lệ lỗi tốt nhất khi có giá trị bằng 0, tất cả từ đều được nhận dạng đúng).
MFCC	Mel-Frequency Cepstrum Co-efficients (Một phương pháp trích rút đặc trưng biểu diễn tín hiệu âm thanh)
DNN	Deep Neural Network

ASR	Automatic speech recognition - Nhận dạng tiếng nói tự động

Mở đầu

Công nghệ thông tin nói chung và trí tuệ nhân tạo nói riêng đang là một trong những ngành được đầu tư trọng điểm của tất cả các quốc gia trên thế giới. Công nghệ thông tin đã và đang được áp dụng phổ biến vào tất cả ngành nghề. Trong cuộc cách mạng công nghiệp lần thứ 4, máy móc ngày càng thay thế sức lao động của con người nhiều hơn. Hiện nay, việc phát triển máy móc có khả năng xử lý, tư duy như con người đã và đang được rất nhiều nhà khoa học trên thế giới nghiên cứu và phát triển. Đây chính là điều gây nên yêu cầu lớn về nhân lực ngành Trí tuệ nhân tạo. Các hệ thống máy móc như: Nhận dạng hình ảnh, đối tượng, Hệ thống lái xe tự động, Hệ thống nhận dạng Tiếng nói, Dịch máy… đang dần đạt đến độ chính xác của con người.

Để xây dựng nên những tác tử máy thông minh như vậy, tất yếu cần đến sự huấn luyện bởi con người, điều này đòi hỏi con người phải gán nhãn các tập dữ liệu huấn luyện cho mô hình học máy. Chưa bao giờ ngành công nghiệp gán nhãn dữ liệu phát triển như hiện nay. Thay vì làm công việc máy móc làm, giờ đây hàng triệu người đã và đang làm việc với vai trò là các nhân viên gán nhãn dữ liệu: văn bản, ảnh, âm thanh, y tế... Đây là một ví dụ điển hình việc ảnh hưởng của Cuộc cách mạng công nghiệp 4.0 tới sự chuyển dịch của cơ cấu lao động.

Hiện nay, thị trường gán nhãn dữ liệu có giá trị lên tới hàng tỉ đô. Các bài toán khó như xử lý ảnh, nhận dạng âm thanh, dịch máy… yêu cầu hàng chục, hàng trăm nghìn mẫu dữ liệu để có thể đạt độ chính xác tương tự con người. Các nghiên cứu về việc tối ưu lựa chọn những dữ liệu gán nhãn cũng ra đời nhằm đáp ứng việc giảm thiểu chi phí gán nhãn, cũng như hỗ trợ người dùng gán nhãn nhanh nhất, kiểm soát quá trình gán nhãn để đạt được tập dữ liệu tốt nhất cho việc huấn luyện mô hình. Một trong những phương pháp áp dụng hiệu quả cho việc lựa chọn dữ liệu gán nhãn là phương pháp học chủ động. Phương pháp này dựa trên cơ chế bằng cách hỏi một chuyên gia tự động về việc có hay không nên gán nhãn một mẫu dữ liệu.

Trong luận văn này, tôi sẽ trình bày việc áp dụng phương pháp học chủ động trong việc lựa chọn dữ liệu gán nhãn cho bài toán nhận dạng tiếng nói. Bài toán nhận dạng tiếng nói

là một trong những bài toán được đầu tư bởi rất nhiều tập đoàn công nghệ lớn tại Việt Nam trong thời gian gần đây. Việc gán nhãn dữ liệu yêu cầu từ vài trăm giờ dữ liệu đến vài chục nghìn giờ dữ liệu, nó tiêu tốn một lượng lớn ngân quỹ cho việc gán nhãn dữ liệu.

Do đó, luận văn được thực hiện với mục đích chính sau đây:

- Đánh giá mô hình nhận dạng tiếng nói hiện nay, giúp người dùng mới có cái nhìn tổng quan, và dễ tiếp cận bài toán nhận dạng.

- Đề xuất phương pháp lựa chọn dữ liệu "quan trọng" cho việc gán nhãn dữ liệu bài toán nhận dạng tiếng nói sử dụng phương pháp học chủ động. Điều này giúp với cùng số tiền ngân quỹ bỏ ra cho việc gán nhãn, ta thu được những dữ liệu chất lượng nhất cho việc huấn luyện mô hình.

Nội dung của luận văn bao gồm các chương:

- Chương 1 - Giới thiệu bài toán: Luận văn sẽ trình bày tổng quan về thị trường gán nhãn dữ liệu hiện nay. Các vấn đề chính trong một hệ thống gán nhãn dữ liệu nói chung và vấn đề lựa chọn dữ liệu quan trọng cho gán nhãn nói riêng.
- Chương 2 – Mô hình nhận dạng tiếng nói: Luận văn sẽ trình bày về các thành phần chính của một mô hình nhận dạng tiếng nói và một số công cụ nổi bật trong cộng đồng nhận dạng tiếng nói. Đồng thời cũng phân tích và so sánh ưu nhược điểm của một số phương pháp nhận dạng.
- Chương 3 – Phương pháp học chủ động cho bài toán nhận dạng tiếng nói: Luận văn sẽ trình bày tổng quan về phương pháp học chủ động (Active Learning) cho các bài toán học máy. Phương pháp học chủ động được cho là một phương pháp rất phổ biến và hiệu quả đối với các bài toán về xử lý ngôn ngữ tự nhiên, đặc biệt được sử dụng rất nhiều trong các hệ thống gán nhãn dữ liệu. Đồng thời luận văn cũng sẽ khảo sát một số công trình nghiên cứu về cách áp dụng Active Learning trong bài toán nhận dạng tiếng nói.
- Chương 4 – Thí nghiệm: Luận văn sẽ trình bày thí nghiệm trên 2 bộ dữ liệu khác nhau và phân tích sự ảnh hưởng của dữ liệu đối với phương pháp học chủ động.

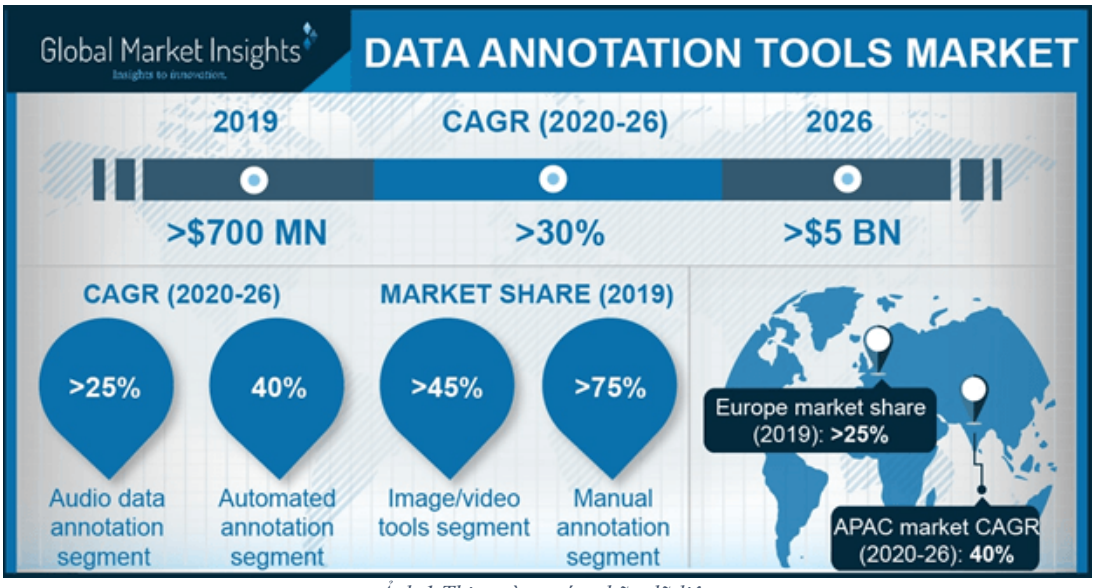
Hiệu quả của phương pháp học chủ động phụ thuộc rất nhiều vào độ dư thừa và trùng lặp của dữ liệu. Đồng thời, luận văn sẽ thí nghiệm việc lựa chọn dữ liệu theo từng tiêu chí về mặt âm học và về mặt ngôn ngữ.

- Chương 5 - Kết luận

Chương 1. Giới thiệu bài toán

1.1. Tổng quan

Sự phát triển của các mô hình học máy và trí tuệ nhân tạo ngày càng trở nên rộng rãi, máy móc ngày càng thay thế cho sức lao động của con người nhiều hơn. Đặc biệt trong những năm gần đây, với sự phát triển của mô hình học sâu đã chứng minh tính hiệu quả trong nhiều bài toán thực tế như: Nhận dạng khuôn mặt, Xử lý tiếng nói, Dịch máy… Đây đều là những bài toán phổ biến, được nhiều tập đoàn công nghệ lớn đầu tư và phát triển.



Ảnh 1 Thị trường gán nhãn dữ liệu

Để phát triển những công cụ học máy với độ chính xác cao, hầu hết các mô hình đều yêu cầu từ hàng trăm ngàn đến hàng triệu mẫu dữ liệu học. Ngành công nghiệp gán nhãn chưa bao giờ phổ biến như hiện nay, điều này phản ánh sự dịch chuyển về cơ cấu lao động. Thay vì làm công việc máy móc đang làm, công việc gán nhãn đã và đang tạo việc làm cho rất nhiều lao động. Hiện nay, rất nhiều công ty đã được mở ra để kinh doanh dịch vụ gán nhãn dữ liệu.

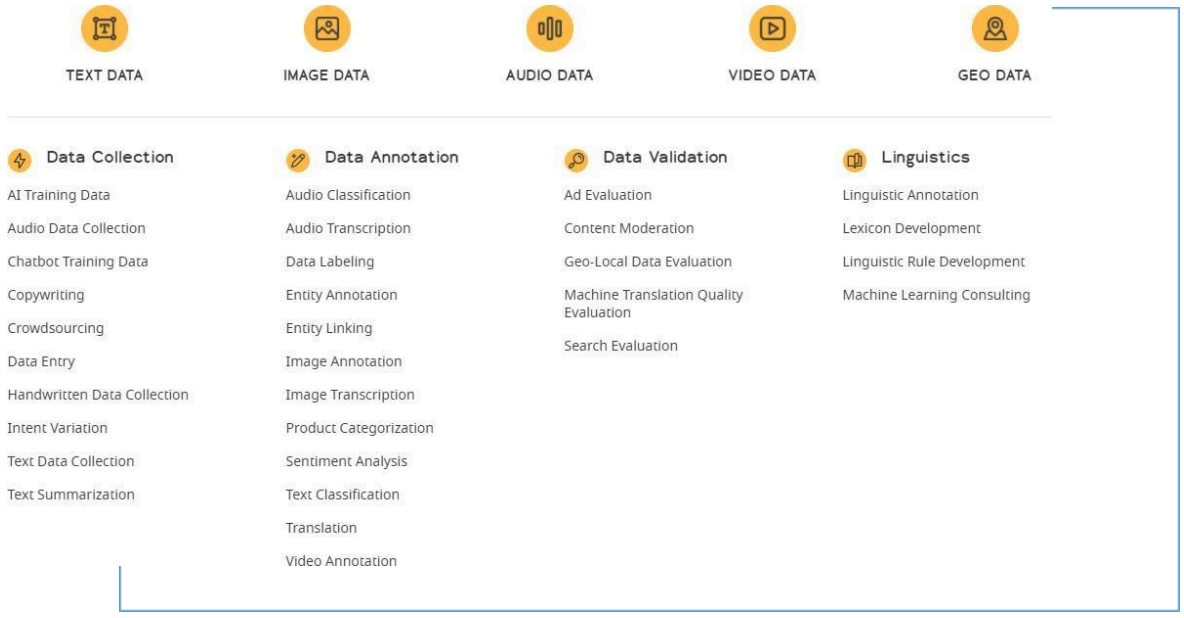
Theo như thống kê, thị trường gán nhãn dữ liệu thủ công năm 2019 là 547 triệu USD, và sẽ tăng gấp hơn 4 lần vào năm 2026. Tương tự với việc gán nhãn tự động, tuy nhiên thị trường gán nhãn tự động thấp hơn thủ công do yêu cầu chủ yếu của việc gán nhãn là độ chính xác, điều này phụ thuộc vào con người.

Chuyển đổi số được thực hiện cho tất cả các ngành nghề, do đó việc gán nhãn dữ liệu có thể đến từ tất cả lĩnh vực như: Tài chính, Kinh tế, Nông nghiệp, Y tế, Viễn thông, Tự động hóa…

Các dữ liệu gán nhãn cũng rất đa dạng, phong phú và có thể được lấy từ nhiều nguồn:

- Dữ liệu văn bản
- Dữ liệu hình ảnh
- Dữ liệu âm thanh
- Dữ liệu video
- Dữ liệu có cấu trúc (HTML, XML, Excel)

Đối với dữ liệu văn bản, ta có nhiều bài toán cần gán nhãn như: Tóm tắt, trích rút thực thể, phân loại văn bản. Đối với dữ liệu về ảnh, ta có các lớp bài toán như phân loại đối tượng, phát hiện đối tượng, phân vùng ảnh. Đối với dữ liệu tiếng nói, ta có bài toán về nhận dạng tiếng nói, tổng hợp tiếng nói. Ngoài việc cung cấp hệ thống gán nhãn dữ liệu, một số doanh nghiệp còn có thể cung cấp về nhân lực con người.



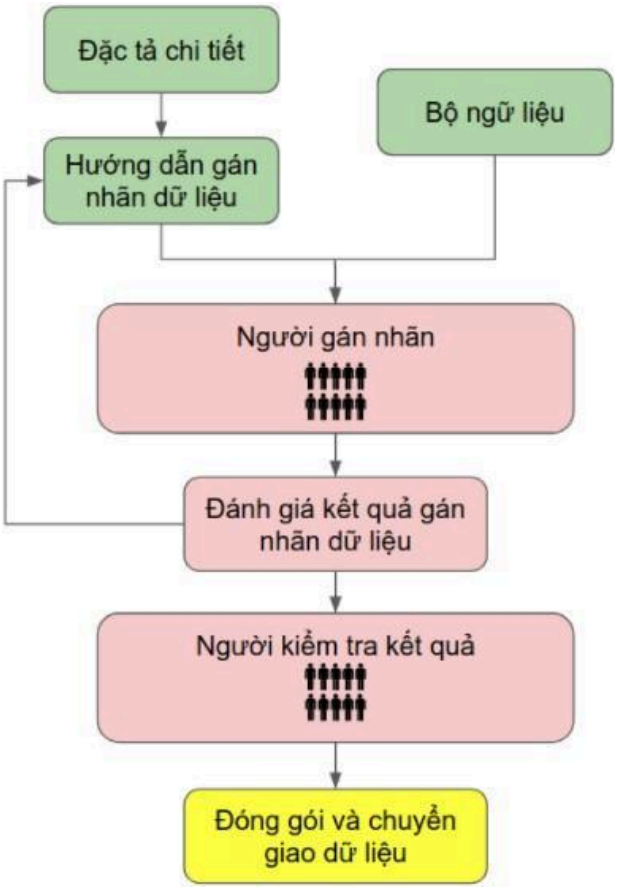
Ảnh 2 Một số loại dữ liệu và các bài toán gán nhãn (Lionbridge AI)

Một số nhà cung cấp các công cụ, dịch vụ gán nhãn phổ biến như:

- Lionbridge AI

- Amazon Mechanical Turk
- Computer Vision Annotation Tool (CVAT)
- SuperAnnotate
- Dataturks

1.2. Quy trình gán nhãn dữ liệu



Ảnh 3 Quy trình gán nhãn dữ liệu

Để có một hệ thống gán nhãn dữ liệu hoàn chỉnh, ta cần các thành phần sau:

- Tài liệu đặc tả sử dụng phần mềm
- Tài liệu hướng dẫn nhân viên gán nhãn và đánh giá dữ liệu. Đối với các loại dữ liệu yêu cầu chuyển môn, cần có tài liệu hướng dẫn cụ thể và chi tiết.
- Bộ ngữ liệu: Tập dữ liệu cần để gán nhãn.
- Người gán nhãn: Nhân viên thực hiện gán nhãn dữ liệu

- Đánh giá kết quả gán nhãn dữ liệu: Bước tự động đánh giá kết quả sử dụng mô hình đã huấn luyện sẵn.
- Người kiểm tra kết quả: Nhân viên đánh giá, xem xét lại kết quả gán nhãn cuối cùng
- Đóng gói và chuyển giao dữ liệu: Đóng gói dữ liệu gán nhãn và chuyển cho khách hàng.

Đây là thành phần thiết yếu cần cho một hệ thống gán nhãn dữ liệu. Tuy nhiên, tùy vào mỗi hệ thống gán nhãn và yêu cầu của bài toán gán nhãn mà ta có thể mở rộng kiến trúc hệ thống gán nhãn này để phù hợp và chi tiết hơn với việc gán nhãn và kiểm soát chất lượng gán nhãn của bài toán.

1.3. Vấn đề chính trong một hệ thống gán nhãn dữ liệu.

Một hệ thống gán nhãn dữ liệu thường gặp 2 vấn đề chính sau đây:

- Lựa chọn dữ liệu gán nhãn: bước quan trọng nhất trong hệ thống gán nhãn. Lựa chọn dữ liệu không những giúp giảm thiểu số lượng mẫu cần gán nhãn, giảm chi phí ngân quỹ gán nhãn mà còn giúp cải thiện chất lượng, thời gian huấn luyện mô hình.
- Kiểm tra, đánh giá các dữ liệu đã gán nhãn: Đây là bước quan trọng để đảm bảo lỗi dữ liệu gán nhãn ở mức thấp nhất, tránh ảnh hưởng đến tỉ lệ lỗi của mô hình.

1.3.1. Các phương pháp lựa chọn dữ liệu gán nhãn

Luận văn tập trung vào việc lựa chọn dữ liệu gán nhãn (cụ thể cho bài toán nhận dạng tiếng nói). Bước lựa chọn dữ liệu gán nhãn là bước quan trọng đối với hầu hết các hệ thống gán nhãn. Trong doanh nghiệp, việc lựa chọn dữ liệu gán nhãn tốt giúp giảm số lượng thời gian, ngân quỹ đáng kể cho việc làm dữ liệu mà vẫn đảm bảo độ chính xác của hệ thống.

Hiện nay, có hai phương pháp chính trong việc lựa chọn dữ liệu:

- Phương pháp học chủ động (Active Learning)

- Phương pháp lựa chọn tập lõi (Core-Set Selection)

Phương pháp học chủ động lựa chọn mẫu dữ liệu để gán nhãn từ một hồ dữ liệu chưa được gán nhãn, và lặp đi lặp lại quá trình lựa chọn dữ liệu và huấn luyện mô hình để được tập dữ liệu cho việc gán nhãn. Khác với phương pháp học chủ động, phương pháp lựa chọn tập lõi có thể thực hiện cho cả tập dữ liệu đã gán nhãn và chưa gán nhãn. Mục đích của phương pháp chọn tập lõi là tìm tập con nhỏ nhất có độ chính xác xấp xỉ toàn bộ tập dữ liệu. Thuật toán thường sử dụng cho phương pháp lựa chọn tập lõi là phương pháp phân cụm k-means hoặc k-median. Sau khi lựa chọn được các tập Core-Set, ta có thể lựa chọn các mẫu theo tỉ lệ nhất định từ mỗi tập Core-Set này.

Phương pháp Core-Set là phương pháp đơn giản, chủ yếu dựa vào phân cụm và khó kết hợp đối với tập dữ liệu đã gán nhãn sẵn hoặc mẫu có đặc trưng phức tạp. Ví dụ trong trường hợp nhận dạng tiếng nói, ta có thể phân cụm các mẫu trong tập dữ liệu chưa gán nhãn bằng đặc trưng âm học (MFCC), tuy nhiên sẽ không hiệu quả vì đây là đặc trưng theo thời gian. Ta có thể thay bằng tìm tập Core-Set cho nhãn các câu được giải mã bằng máy, nhưng phụ thuộc vào độ chính xác của mô hình học và không thể kiểm tra đối với các mẫu đã gán nhãn.

Phương pháp học chủ động là phương pháp tốt nhất để lựa chọn các dữ liệu quan trọng cho một hệ thống gán nhãn (hay mô hình học máy), có thể hoạt động trên nhiều bài toán và kiểu dữ liệu.

Do đó, trong luận văn này, luận văn sẽ tập trung vào bài toán nhận dạng tiếng nói và việc áp dụng phương pháp học chủ động cho bài toán nhận dạng tiếng nói.

1.3.2. Đánh giá chất lượng gán nhãn

Để đánh giá chất lượng gán nhãn, ta có thể sử dụng 2 phương pháp tự động hoặc thủ công.

- Phương pháp thủ công: Cần có các nhóm người với vai trò người đánh giá. Nhóm sẽ xem xét các mẫu dữ liệu nhân viên gán nhãn và thực hiện và thực hiện đánh giá, chỉnh sửa lại.

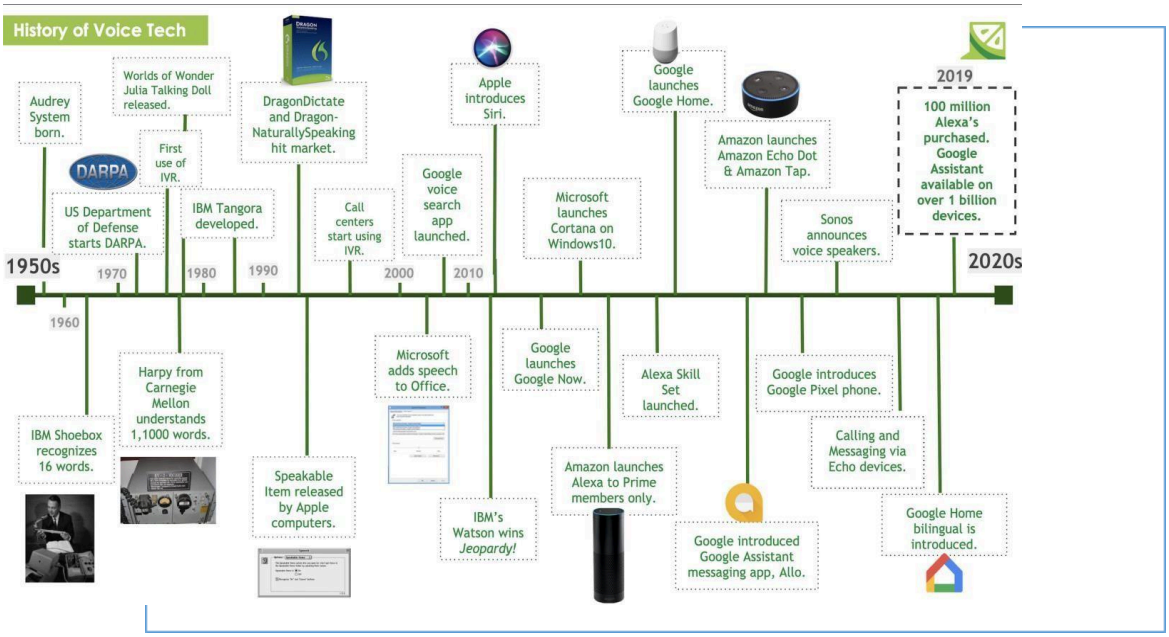
- Phương pháp tự động: Có nhiều phương pháp đánh giá tự động, tuy nhiên việc đánh giá tự động không đảm bảo được tính chính xác tuyệt đối.
    - Phương pháp 1 - Sử dụng tập dữ liệu đã gán nhãn: Đưa các dữ liệu này vào tập dữ liệu cần gán nhãn. Kiểm tra tính chính xác của nhân viên gán nhãn bằng cách đối chiếu với nhãn thực tế.
    - Phương pháp 2 – So sánh chéo: So sánh nhiều mẫu được thực hiện bởi các nhân viên gán nhãn. So sánh và đối chiếu độ chính xác dựa trên các mẫu dữ liệu này.

Chương 2. Mô hình nhận dạng tiếng nói.

2.1. Giới thiệu

Bài toán nhận dạng tiếng nói là bài toán khó và gần đây rất được chú ý và nghiên cứu bởi cộng đồng. Nhưng thực tế bài toán nhận dạng tiếng nói được các nhà khoa học nghiên cứu từ rất sớm, từ đầu những năm 1950. Bài toán nhận dạng tiếng nói đi từ các bài toán đơn giản như nhận dạng từng chữ số, phát triển đến nhận dạng 26 ký tự trong bảng từ điển Tiếng Anh, và hiện nay là có thể nhận dạng được theo cả từ và câu.

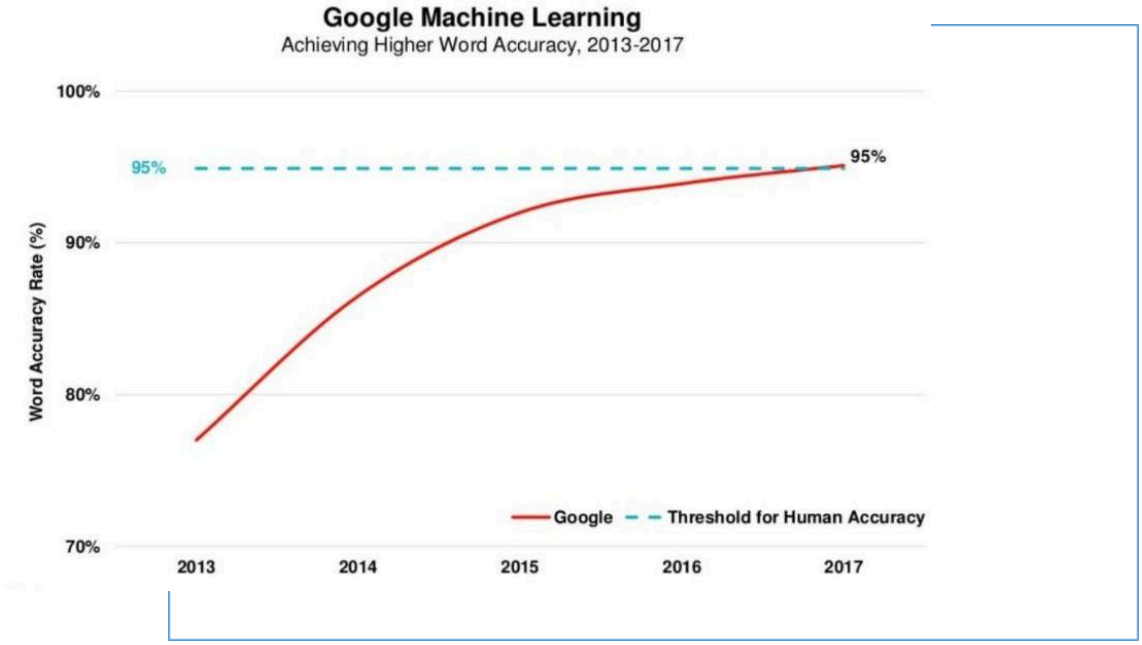
Quá trình phát triển của các mô hình nhận dạng tiếng nói [1].



Ảnh 4 Lịch sử phát triển của hệ thống nhận dạng tiếng nói

Các mô hình nhận dạng tiếng nói cũng đi từ phương pháp thô sơ đến các phương pháp phức tạp như phương pháp học sâu đầu cuối. Các phương pháp ban đầu của nhận dạng tiếng nói chủ yếu dựa vào phương pháp phân loại đặc trưng âm thanh của các ký tự chữ hoặc số tương ứng. Đến những năm 1980, với sự phát triển của mô hình Markov ẩn (Hidden Markov Model, viết tắt HMM) là mô hình học máy dựa vào thống kê có thể xử lý dữ liệu theo chuỗi thời gian, các hệ thống nhận dạng tiếng nói trở nên phổ biến, được nghiên cứu nhiều hơn và độ chính xác được cải tiến đáng kể. Sau này, với sự phát triển

của mạng học sâu và phần cứng GPU, mô hình nhận dạng tiếng nói chuyển dịch dần sang mô hình lai (kết hợp HMM và mạng học sâu) từ đầu những năm 2010. Từ năm 2013 đến nay, các mô hình học sâu đầu cuối đã bước đầu thay thế các phương pháp lai vì sự tiện lợi và dễ dàng trong việc chuẩn bị dữ liệu, huấn luyện mô hình cũng như khi triển khai thực tế.



Ảnh 5 Độ chính xác của Google Voice qua các thời kỳ [2]

Độ chính xác của mô hình nhận dạng học máy tăng nhanh từ khoảng từ năm 2013 trở lại đây. Hình trên cho thấy từ năm 2013, google chỉ đạt độ chính xác gần 78%, nhưng đến nay đã đạt độ chính xác tương tự con người với mức 95%.

Công nghệ nhận dạng tiếng nói cũng được tìm hiểu và nghiên cứu từ đầu những năm 2014, 2015 bởi các tập đoàn lớn như Viettel, FPT, Zalo, Vingroup, … Việc triển khai hệ thống nhận dạng tiếng nói cho Tiếng Việt gặp nhiều khó khăn hơn tiếng Anh do một số nguyên nhân sau:

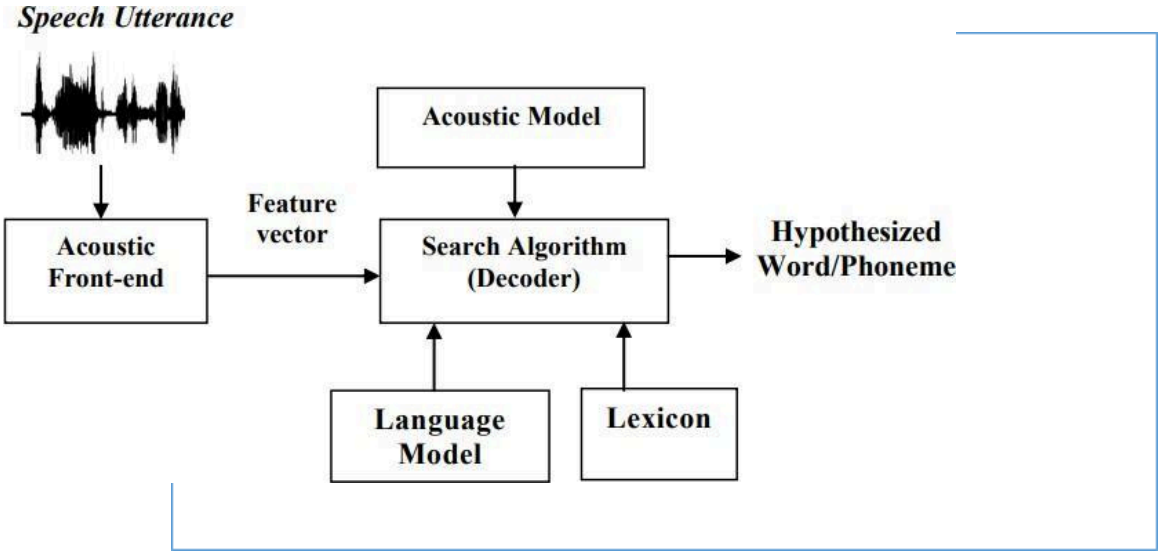
- Tiếng Việt có ngữ pháp đa dạng phong phú
- Tiếng Việt là ngôn ngữ từ ghép.
- Tiếng Việt có nhiều phát âm vùng miền…

Để phát triển một hệ thống nhận dạng tiếng nói tốt thì yêu cầu từ vài trăm giờ đến vài chục nghìn giờ dữ liệu huấn luyện. Với số ngân quỹ cố định cho việc gán nhãn, bài toán đặt ra là lựa chọn những dữ liệu tốt nhất cho mô hình học. Đây là vấn đề gặp phải với hầu hết các doanh nghiệp khi làm bài toán nhận dạng tiếng nói. Trong luận văn này, tôi sẽ trình bày về các nội dung nhằm giải quyết vấn đề lựa chọn dữ liệu quan trọng cho quá trình gán nhãn để huấn luyện mô hình nhận dạng tiếng nói như phân tích độ dư thừa dữ liệu và sử dụng phương pháp học chủ động (Active Learning) để lựa chọn dữ liệu quan trọng cho quá trình gán nhãn dữ liệu.

2.2. Kiến trúc mô hình nhận dạng tiếng nói

Kiến trúc của một mô hình nhận dạng tiếng nói cơ bản bao gồm 4 thành phần chính:

- Acoustic Front-end
- Acoustic Model
- Search Algorithm (Decoder)
- Language Model



Ảnh 6 Kiến trúc mô hình nhận dạng tiếng nói [16]

Acoustic Front-end có vai trò chuyển tín hiệu tiếng nói thành đặc trưng đầu vào để huấn luyện mô hình học máy. Tín hiệu âm thành từ mic (microphone - thiết bị ghi âm tiếng nói) sẽ được chuyển thành các vector âm học có số chiều cố định. Các tham số của mô

hình được ước lượng từ các acoustic vector của bộ dữ liệu huấn luyện. Sau đó, bộ giải mã (decoder) sẽ tìm kiếm tất cả các chuỗi từ để từ đó tìm ra chuỗi từ có xác suất cao nhất khớp với tín hiệu tiếng nói đầu vào.

Chức năng của hệ thống nhận dạng tiếng nói tự động có thể được mô tả như việc trích xuất các tham số tiếng nói từ tín hiệu tiếng nói âm thanh cho mỗi từ. Các tham số của lời nói mô tả một từ thay đổi theo thời gian và chúng cùng nhau tạo nên một mẫu đặc trưng cho từ. Trong giai đoạn huấn luyện mô hình, các mẫu đặc trưng của từ được học và lưu trữ. Khi muốn nhận dạng một từ, mẫu đặc trưng của nó sẽ được so sánh với các mẫu đã lưu trữ và trả về kết quả phù hợp nhất với mẫu được chọn. Phương pháp này được gọi là nhận dạng mẫu.

2.2.1. Đặc trưng âm học (Acoustic Front-end)

Acoustic front-end liên quan đến việc xử lý tín hiệu và trích xuất đặc trưng. Trong nhận dạng tiếng nói, mục tiêu chính của bước trích xuất đặc trưng là tính toán một chuỗi các vector đặc trưng cho một biểu diễn dạng số của tín hiệu đầu vào đã cho. Việc trích rút đặc trưng thường bao gồm 3 giai đoạn.

Giai đoạn đầu tiên được gọi là phân tích tiếng nói. Nó thực hiện phân tích phổ của tín hiệu âm thanh và tạo ra các đặc trưng thô mô tả phổ của các khoảng tiếng nói trong một thời gian ngắn.

Giai đoạn thứ hai tổng hợp mở rộng đặc trưng của vector bao gồm kết hợp các đặc trưng hoặc đưa thêm các thông tin tĩnh và động.

Giai đoạn cuối cùng là biến đổi những vector đặc trưng thành các vector nhỏ gọn như nén, phân tích thành phần chính, sau đó được đưa vào huấn luyện mô hình nhận dạng.

Trích rút đặc trưng âm thanh có rất nhiều loại, và cho nhiều biểu diễn khác nhau. Để tìm được phương pháp trích rút đặc trưng tốt thì chúng phải cho phép hệ thống tự động phân biệt giữa các âm thanh khác nhau thông qua âm thanh tiếng nói tương tự, chúng phải cho phép tạo tự động các mô hình âm thanh cho các âm thanh mà không cần quá nhiều dữ liệu

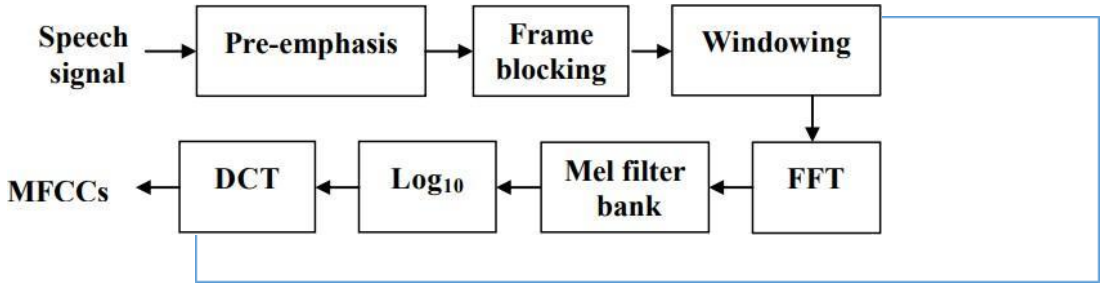
huấn luyện và chúng phải thể hiện số liệu thống kê phần lớn bất biến đối với người nói và môi trường nói.

Có rất nhiều phương pháp để mô tả tín hiệu tiếng nói dưới dạng số. Một số phương pháp trích xuất đặc trưng như: Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA), Independent Component Analysis (ICA), Linear Predictive Coding (LPC), Cepstral Analysis, Mel-Frequency Scale Analysis, Filter-Bank Analysis, Mel-Frequency Cepstrum Co-efficients (MFCC), Kernal Based Feature Extraction, Dynamic Feature Extraction, Wavelet based features, Spectral Subtraction and Cepstral Mean Subtraction (CMS). Đối với nhận dạng tiếng nói trong môi trường có tiếng ồn, nhiều phương pháp trích xuất đặc trưng như: biên độ đỉnh giao nhau bằng không (ZCPA), phát hiện đồng bộ cục bộ trung bình (ALSD), đáp ứng không méo phương sai tối thiểu theo cảm nhận (PMVDR), hệ số song song chuẩn hóa công suất (PNCC), Các tính năng tích hợp bất biến (IIF), hạt nhân tái tạo thính giác thưa thớt (SPARK), và các đặc trưng Filter-Bank Gabor được áp dụng hiệu quả.

Có nhiều biểu diễn đặc trưng được sử dụng, nhưng phổ biến nhất trong nhận dạng tiếng nói là phương pháp trích rút đặc trưng MFCC. Phương pháp MFCC bao gồm các bước sau:

- Pre-emphasis: Tăng mức năng lượng cho các âm có tần số cao.
- Frame blocking và Windowing: Chia tín hiệu đầu vào thành các đoạn có khoảng thời gian đủ nhỏ. Điều này được thực hiện bằng cách tạo ra cửa sổ với độ rộng N miliseconds và bước dịch chuyển là M miliseconds. Cửa sổ sẽ trượt theo bước dịch chuyển để lấy ra các đoạn tín hiệu âm thanh liên tục.
- Discrete Fourier Transform (FFT - Fast Fourier Transform): Sử dụng phép biến đổi Fourier nhanh (FFT) cho mỗi đoạn tín hiệu để biến đổi âm thanh từ miền thời gian, biên độ sang miền tần số.
- Mel Filter Bank: Tai người cảm nhận âm có tần số thấp tốt, kém nhạy cảm với các âm có tần số cao. Do đó, ta cần chuẩn hóa lại các vector tần số này sao cho thích hợp nhất với đặc trưng của tai người.

- Log: Lấy logarit thập phân của các tín hiệu phổ Mel để giảm độ chênh lệch tần số.
- DCT: sử dụng phép biến đổi cosine rời rạc dựa trên đặc trưng vừa thu được. Đầu ra của DCT là vector 13 chiều
- MFCCs: Bổ sung các chiều đặc trưng thể hiện sự biến đổi của tín hiệu bằng các đạo hàm cấp 1 và đạo hàm cấp 2 trên đặc trưng vừa thu được.



Ảnh 7 Các bước trích rút đặc trưng MFCC

2.2.2. Mô hình âm học (Acoustic Model)

Mô hình âm học (Acoustic Model) là một trong những thành phần quan trọng nhất trong một hệ thống nhận dạng tiếng nói tự động, hệ thống này đại diện cho các đặc điểm âm thanh để hình thành các đơn vị ngữ âm được nhận dạng.

Trong việc xây dựng một mô hình âm học, một vấn đề cơ bản và quan trọng là lựa chọn các đơn vị cơ bản cho mô hình học. Tùy vào ngôn ngữ khác nhau mà một số loại đơn vị từ phụ có thể được sử dụng để mô hình hóa âm thanh. Các đơn vị cơ bản này có thể là từ, ký tự hay mức độ nhỏ hơn là âm vị (phone)... Đơn vị cơ bản khác nhau được lựa chọn để huấn luyện mô hình có thể tạo ra sự khác biệt đáng kể về độ chính xác của hệ thống nhận dạng tiếng nói. Mô hình âm thanh của tiếng nói thường được học dựa trên các biểu diễn thống kê của các chuỗi vector đặc trưng được tính toán từ dạng sóng của tiếng nói.

Mô hình Markov ẩn (HMM) là một trong những mô hình thống kê được sử dụng phổ biến nhất để xây dựng các mô hình âm học trong bài toán nhận dạng tiếng nói. Các mô hình âm thanh khác bao gồm mô hình phân đoạn, mô hình siêu phân đoạn (bao gồm cả mô hình động ẩn), mạng nơron, mô hình entropy cực đại và trường ngẫu nhiên có điều kiện (ẩn), v.v. Mô hình âm học là một mô hình học các biểu diễn thống kê cho từng tín hiệu âm thanh riêng biệt tạo nên một từ. Mỗi biểu diễn thống kê này được gán một nhãn gọi là

mô hình âm vị. Các âm vị (phonemes) được tạo ra bằng cách lấy một cơ sở dữ liệu lớn của một ngôn ngữ, sau đó sử dụng các thuật toán huấn luyện đặc biệt để tạo ra các biểu diễn thống kê cho mỗi âm vị trong một ngôn ngữ. Mỗi âm vị tương đương với một trạng thái ẩn khác nhau trong mô hình HMM. Bộ giải mã tiếng nói lắng nghe các âm thanh khác nhau được nói bởi người dùng và sau đó tìm kiếm trạng thái ẩn HMM phù hợp trong mô hình âm học để thu được các âm vị của câu nói. Mỗi từ được nói sẽ được phân tách thành một chuỗi âm thanh cơ bản được gọi là âm vị cơ bản. Mô hình âm học mô tả xác suất của một quan sát cụ thể đối với một âm vị cơ bản.

Hiện nay, các mô hình học ở mức âm vị thường cho kết quả tốt nhất. Tuy nhiên, các mô hình học ở mức ký tự có thể được huấn luyện và giải mã đơn giản hơn. HMM là một trong nhưng mô hình cơ bản của nhận dạng tiếng nói, được sử dụng trong thời gian dài. Tuy nhiên, hiện nay, các mô hình lai và các mô hình mạng học sâu cho kết quả tốt hơn rất nhiều so với sử dụng mô hình HMM đơn thuần.

2.2.3. Mô hình ngôn ngữ (Language Model)

Mô hình ngôn ngữ là một tập hợp các ràng buộc về chuỗi các từ được chấp nhận trong một ngôn ngữ nhất định. Những ràng buộc này có thể được biểu diễn, ví dụ, bằng các quy tắc của ngữ pháp chung hoặc đơn giản bằng số liệu thống kê về mỗi cặp từ được ước tính trên một tập ngữ liệu huấn luyện. Mặc dù có những từ có âm thanh tương tự điện thoại, nhưng con người nhìn chung không khó nhận ra từ đó. Điều này chủ yếu là do họ biết ngữ cảnh và cũng có ý tưởng về những từ hoặc cụm từ có thể xảy ra trong ngữ cảnh. Cung cấp ngữ cảnh này cho hệ thống nhận dạng tiếng nói là mục đích của mô hình ngôn ngữ. Mô hình ngôn ngữ chỉ định những từ hợp lệ trong ngôn ngữ là gì và chúng có thể xảy ra theo trình tự nào.

Các mô hình ngôn ngữ thường được huấn luyện dựa trên xác suất các n-gram (chuỗi n từ liên tiếp nhau trong một câu được gọi là các n-gram của câu) được tính bằng cách thống kê các chuỗi từ liên tiếp trong một kho văn bản. Các mô hình ngôn ngữ phổ biến là mô hình bigram và trigram. Mô hình ngôn ngữ giúp ta có thể xác định được xác suất của các từ tiếp theo mà người nói có thể nói, dựa trên lịch sử của các từ đã nói trước đó. Do đó,

việc sử dụng mô hình ngôn ngữ đối với việc chuẩn hóa lại đầu ra của mô hình nhận dạng tiếng nói đặc biệt hiệu quả. Tuy nhiên, người ta đã quan sát thấy rằng việc giảm sự hỗn loạn (Perplexity) của mô hình ngôn ngữ không nhất thiết dẫn đến kết quả nhận dạng tiếng nói tốt hơn.

Hiện nay, ngoài mô hình ngôn ngữ n-gram, các mô hình phức tạp hơn sử dụng học sâu cho kết quả tốt hơn do có thể học được quan hệ ngữ nghĩa giữa các từ như RNNLM, TransformerXL…

2.2.4. Bộ giải mã (Decoder)

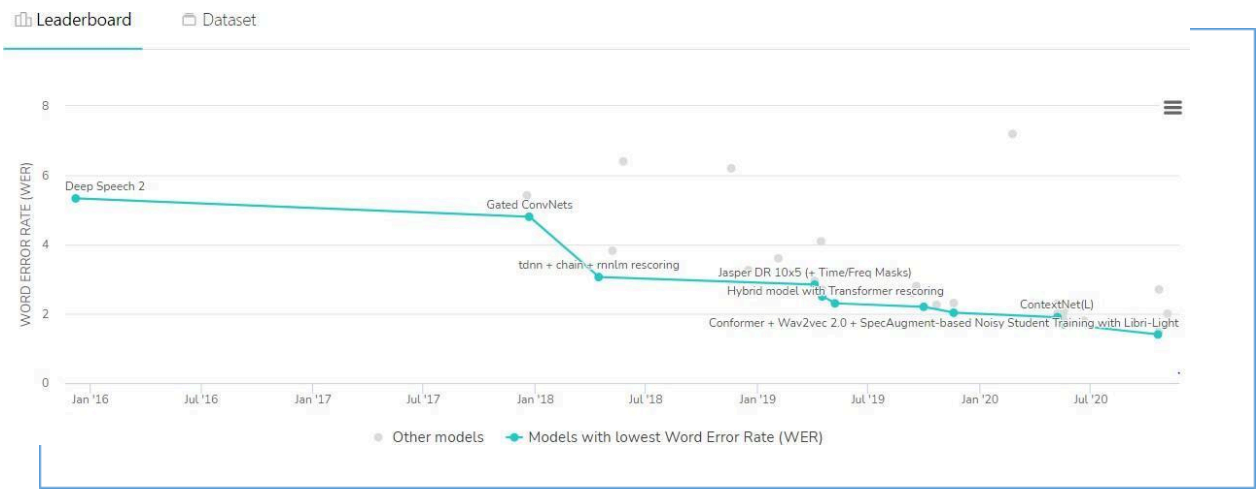
Nhiệm vụ của giai đoạn giải mã là tìm chuỗi từ W có khả năng xảy ra nhất với chuỗi quan sát O, và mô hình âm học-ngôn ngữ. Vấn đề giải mã có thể được giải quyết bằng cách sử dụng các thuật toán quy hoạch động. Thay vì đánh giá khả năng của tất cả các đường dẫn mô hình có thể tạo ra chuỗi quan sát O, trọng tâm là tìm một đường dẫn duy nhất qua mạng mang lại kết quả phù hợp nhất với O. Để ước tính trình tự trạng thái tốt nhất cho chuỗi quan sát đã cho, ta thường sử dụng thuật toán Viterbi. Trong trường hợp các nhiệm vụ nhận dạng có số lượng từ vựng lớn, sẽ rất khó để xem xét tất cả các từ có thể có bằng phương pháp đệ quy của thuật toán Viterbi. Để giải quyết vấn đề này, tìm kiếm chùm (Beam search) có thể được sử dụng cho phép lặp Viterbi với chỉ những từ có xác suất đường dẫn trên một ngưỡng mới được xem xét khi mở rộng đường dẫn đến bước thời gian tiếp theo. Cách tiếp cận này đẩy nhanh quá trình tìm kiếm với chi phí đánh đổi là độ chính xác lời giải tốt nhất của việc giải mã. Thuật toán Viterbi giả định rằng mỗi đường dẫn tốt nhất tại thời điểm t phải là phần mở rộng của mỗi đường dẫn tốt nhất kết thúc tại thời điểm t - 1. Tuy nhiên, đường dẫn có xác suất thấp hơn những đường khác ngay từ đầu có thể trở thành đường dẫn tốt nhất cho toàn bộ chuỗi.

2.3. Khảo sát mô hình nhận dạng tiếng nói hiện nay.

Hiện nay, có hai phương pháp nhận dạng tiếng nói phổ biến là phương pháp lai giữa mô hình Markov ẩn với mạng học sâu và phương pháp sử dụng mạng học sâu đầu cuối. Phương pháp lai giữa mô hình markov ẩn và mạng học sâu trở nên phổ biến và được sử dụng rộng rãi đến năm 2020. Nhưng hiện nay ngày càng nhiều mô hình sử dụng phương pháp học sâu end-to-end cho kết quả tốt hơn phương pháp lai trên các tập dữ liệu.

Ưu điểm của phương pháp học sâu đầu cuối:

- Chuẩn bị dữ liệu huấn luyện dễ dàng, có thể sử dụng cấp phonemes hoặc cấp ký tự, cấp từ.
- Có thể huấn luyện song song hoàn toàn bằng GPU.
- Không cần sử dụng máy chuyển đổi trạng thái hữu hạn trong quá trình giải mã.



Ảnh 8 Các mô hình nhận dạng mới nhất trên bộ dữ liệu librispeech-test-clean [3]

Một số công cụ, mô hình nhận dạng tiếng nói phổ biến hiện nay như:

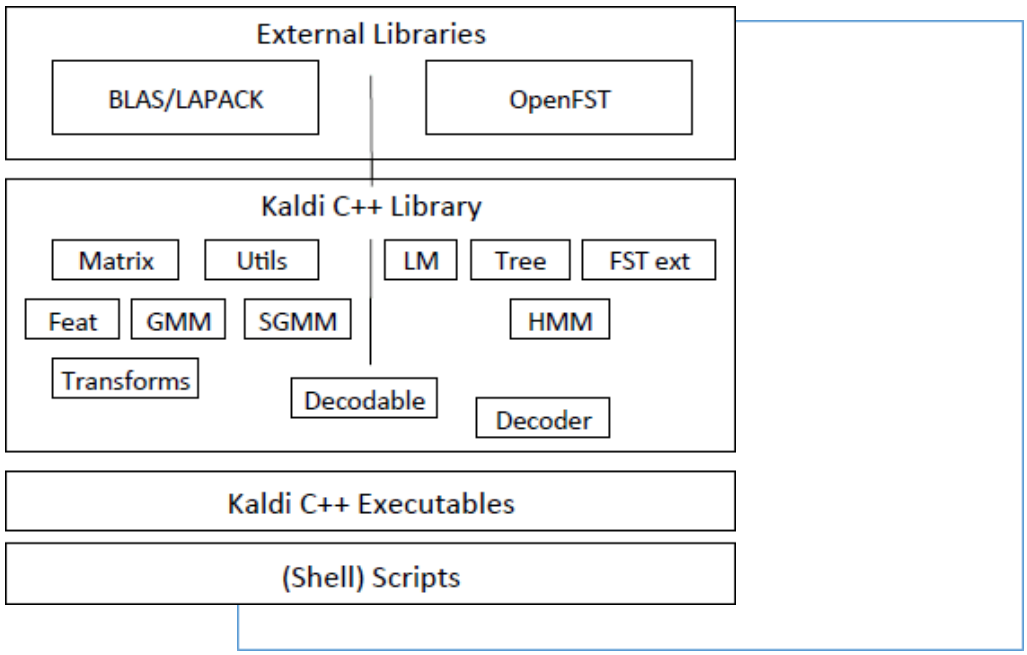
- Mô hình lai Kaldi [4]
- Mô hình lai Pytorch Kaldi [5].

- Mô hình end-to-end Deep Speech [6].
- Mô hình Wav2letter [7].
- Mô hình Pychain [8].
- Mô hình mạng Nvidia QuartzNet [9].
- Mô hình mạng Conformer [10].

2.3.1. Công cụ Kaldi

Kaldi là công cụ được phát triển bởi đại học Johns Hopkins vào năm 2009 dành cho nhận dạng tiếng nói.

Mặc dù rất nhiều mô hình đầu cuối ra đời, nhưng mô hình lai của khung công việc Kaldi vẫn có vị trí nhất định trong công đồng nghiên cứu nhận dạng tiếng nói. Mô hình kaldi là mô hình tiếng nói cho kết quả tốt nhất cho Tiếng Việt cho tới tận năm 2019 và mới được thay thế bởi mô hình học sâu đầu cuối - Conformer nào năm 2020.



Ảnh 9 Kiến trúc công cụ Kaldi

Kaldi là công cụ được viết bằng ngôn ngữ C++. Ban đầu, công cụ Kaldi được xây dựng bằng mô hình Gaussian Mixture Model cho bài toán nhận dạng tiếng nói và các mô hình liên quan đến ngôn ngữ như máy chuyển đổi trạng thái hữu hạn. Sau này, Kaldi tích hợp thêm các mô hình học sâu, để kết hợp tạo thành mô hình lai giữa HMM và DNN. Mô hình này cho kết quả nhận dạng tốt nhất trong nhiều năm. Do đó, Kaldi được rất nhiều công đồng nghiên cứu và phát triển. Định dạng dữ liệu của Kaldi là một trong các chuẩn mà nhiều công cụ nhận dạng hiện tại hỗ trợ và tuân theo. Hiện nay, Kaldi là công cụ hỗ trợ đầy đủ cho tác vụ nhận dạng tiếng nói bao gồm nhiều bài toán như:

- Trích rút đặc trưng tín hiệu,
- Bài toán xử lý nhiễu.
- Mô hình nhận dạng âm học.
- Mô hình ngôn ngữ n-gram, RNNLM.
- Nhận dạng người nói.
- Phát hiện hoạt động nói (VAD)

Huấn luyện mô hình lai cho công cụ nhận dạng tiếng nói của Kaldi bao gồm 2 pha chính. Pha thứ nhất là việc sử huấn luyện mô hình Markov ẩn để huấn luyện mô hình tiếng nói, đầu ra được sử dụng để căn lề giữa tín hiệu tiếng nói và các âm vị (phone). Pha thứ hai là sử dụng nhãn dữ liệu đã được align (căn lề - với mỗi đặc trưng âm thanh tại một thời điểm sẽ được gán với nhãn của mỗi âm vị tương ứng) để huấn luyện mô hình học sâu. Do việc chia làm hai pha này mà khi lượng dữ liệu lên đến hàng nghìn giờ thì việc huấn luyện của Kaldi trở nên chậm chạp hơn. Pha đầu tiên của Kaldi - huấn luyện mô hình HMM chỉ có thể được thực hiện trên bộ xử lý trung tâm của máy tính (CPU), nên hiệu năng chạy song song kém hơn bộ xử lý đồ họa của máy (GPU) vì cần rất nhiều CPU. Pha thứ hai huấn luyện dựa trên mạng học sâu có thể được thực hiện trên GPU và có thời gian tính toán nhanh hơn rất nhiều so với mô hình học sâu đầu cuối vì không cần tự động căn lề các nhãn đầu ra.

Do phụ thuộc cả vào CPU và GPU nên việc đánh giá tốc độ tính toán của mô hình lai so với các mô hình học sâu là rất khó. Với số lượng dữ liệu vài trăm giờ và đủ số lượng CPU core thì thời gian huấn luyện mô hình của Kaldi nhanh hơn rất nhiều so với các phương pháp End-to-End.

Mô hình lai giữa HMM và DNN tận dụng ưu điểm của việc sử dụng HMM-GMM để căn lề giữa tín hiệu tiếng nói và âm vị. Tuy nhiên việc căn lề này khá phức tạp, bao gồm phải kết hợp nhiều pha như căn lề cấp mono phone, bi phone, tri phone… để được một căn lề tốt. Trong khi đó, sự xuất hiện kiến trúc End-To-End kết hợp với Connectionist Temporal Classification layer (CTC layer - tự động căn lề dữ liệu đầu vào và đầu ra của mạng nơron) đã tạo thành một trào lưu mới trong việc áp dụng End-To-End vào các bài toán sequence to sequence, đặc biệt trong các bài toán dịch máy, nhận dạng tiếng nói, chữ viết… Việc áp dụng kiến trúc End-To-End hiện nay đang là một hướng đi mới trong cộng đồng nhận dạng tiếng nói. Nó cũng tồn tại ưu điểm và nhược điểm so với hệ thống cũ.

Ưu điểm:

Việc áp dụng CTC layer hay encoder-decoder giúp là giảm bước alignment phức tạp và tốn thời gian của HMM-GMM. Giúp toàn bộ mạng chỉ phụ thuộc vào một lần học. So với phương pháp cũ, độ chính xác của HMM-DNN sẽ phụ thuộc vào độ chính xác của việc alignment sử dụng HMM-GMM. Do vậy sẽ dễ dàng điều chỉnh và thay đổi mô hình.

Việc áp dụng End-To-End có thể thực hiện trên cấp kí tự hoặc từ, điều này giúp giảm đáng kể công sức chuẩn bị dữ liệu so với phương pháp cũ, cần từ điển phoneme và phiên âm các từ thành các phonemes.

Giảm ảnh hưởng của mô hình ngôn ngữ đối với mô hình học. Trong kiến trúc cũ, việc mô hình ngôn ngữ là thành phần không thể thiếu để có thể chuyển các chuỗi phone thành các word trong câu. Trong khi việc sử dụng End-To-End có thể tiếp cận học độc lập với mô hình ngôn ngữ, không phụ thuộc vào mô hình ngôn ngữ. Tuy vậy, vẫn cần có mô hình ngôn ngữ trong việc tinh chỉnh lại kết quả theo đúng cú pháp, ngữ nghĩa.

Mô hình mạng End-To-End dễ triển khai, dễ thực thi, dễ thay đổi mô hình hơn.

Nhiều mô hình mạng học sâu khác nhau ra đời giúp cho việc tiếp cận End-To-End trở nên dễ dàng và cho kết quả đáng mong đợi hơn.

Nhược điểm:

Mô hình cũ sử dụng phần căn lề phức tạp nhưng phần HMM-DNN lại trở nên đơn giản hơn. Do đó kiến trúc pha DNN không phức tạp bằng kiến trúc DNN trong các mạng End-To-End.

Hiện nay, End-To-End trong nhận dạng tiếng nói có rất nhiều tiếp cận, tuy nhiên vẫn chưa thể thay thế hoàn toàn mô hình cũ.

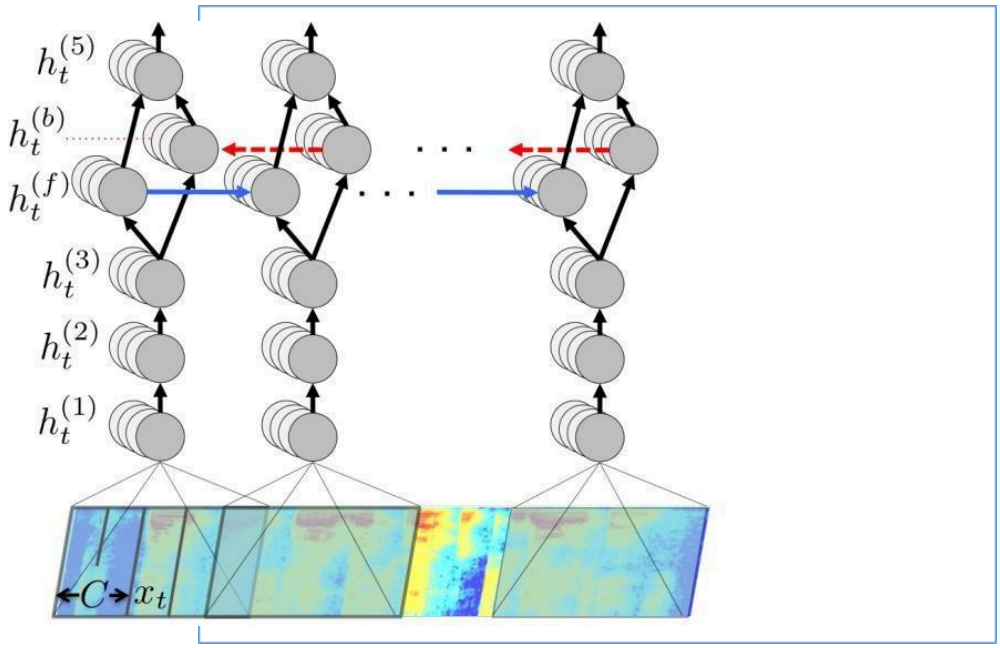
Deep Speech là một trong những mạng End-To-End đầu tiên, đo đó có thời gian huấn luyện khá lớn. Hiện nay, ngày càng nhiều mạng End-To-End có thời gian huấn luyện nhanh hơn và nhiều báo cáo trên các bài báo khoa học cho độ chính xác cao hơn so với phương pháp truyền thống ở một số tập dữ liệu [7] [8] [9]. Do đó hướng tiếp cận End-To-End hiện nay đang rất phổ biến và dần thay thế cho công cụ Kaldi.

2.3.2. Deep Speech: Scaling up end-to-end speech recognition

Deep Speech được phát triển bởi phòng nghiên cứu Baidu Research. Deep Speech nổi lên từ năm 2014, là một trong những mô hình đầu tiên áp dụng end-to-end cho bài toán nhận dạng tiếng nói. Mô hình Deep Speech bao gồm các thành phần như:

- Mạng RNN đơn giản thay bởi Long short term memory (LSTM) phức tạp.
- Sử dụng CTC layer để tự động căn lề giữa đầu vào và đầu ra của mạng.
- Kết hợp mô hình ngôn ngữ.

Do sử dụng CTC layer, nên tốc độ hội tụ của việc huấn luyện khá chậm. Đồng thời, kết quả nhận dạng cho độ chính xác thấp hơn so với phương pháp dạng lai như Kaldi.



Ảnh 10 End-to-End Deep Speech

Mạng end-to-end được thực hiện theo kiến trúc deep-speech 2 của google, là mạng dựa trên kiến trúc thuần không kết nối lặp và mạng nơron kết nối lặp mà không sử dụng kiến trúc có bộ nhớ phức tạp như LSTM. Phương pháp này sẽ làm giảm độ phức tạp và cải thiện hiệu năng của hệ thống nhận dạng. Bằng việc kết hợp với hàm kích hoạt là ReLu, kiến trúc này giải quyết được vấn đề vanishing gradient trong kiến trúc RNN truyền thống và học được các phụ thuộc dài của LSTM.

Kiến trúc mạng bao gồm 5 tầng ẩn. 3 tầng đầu là mạng không lặp, là các tầng truyền thẳng cơ bản. Với hàm kích hoạt ReLu

$$g(z) = \min\{\max\{0, z\}, 20\}$$

$$h_t^{(l)} = g(W^{(l)} h_t^{(l-1)} + b^{(l)})$$

Tầng thứ 4 là tầng lặp 2 chiều (bi-directional recurrent layer):

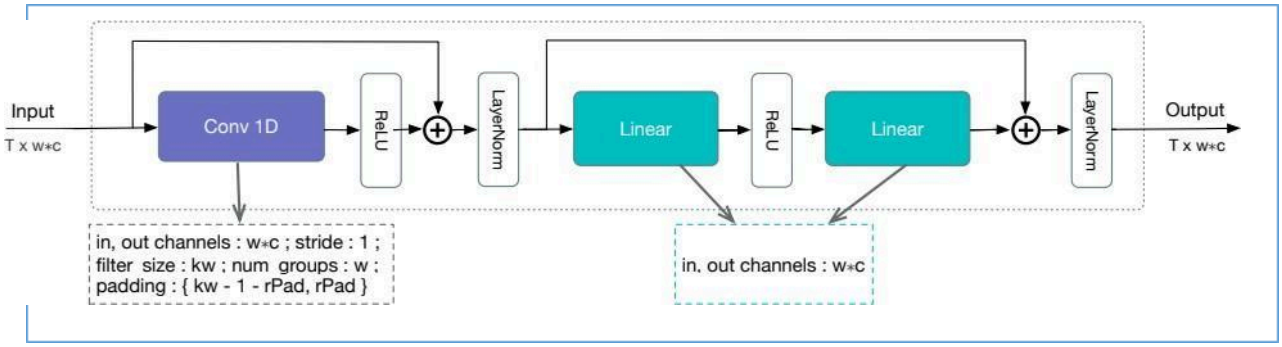
$$h_t^{(f)} = g(W^{(4)} h_t^{(3)} + W_r^{(f)} h_{t-1}^{(f)} + b^{(4)})$$
$$h_t^{(b)} = g(W^{(4)} h_t^{(3)} + W_r^{(b)} h_{t+1}^{(b)} + b^{(4)})$$

Tầng cuối cùng trong kiến trúc mạng là tầng CTC. Đây là tầng quan trọng nhất trong các kiến trúc mạng End-to-End ban đầu. Tầng CTC có chức năng thực hiện tự động alignment các nhãn dữ liệu đầu vào và nhãn dữ liệu đầu ra. Điều này giúp các tín hiệu tiếng nói và kí tự chữ có thể tự động được học mà không cần sự gán nhãn của con người.

2.3.3. Wav2letter++ Scaling Up Online Speech Recognition Using ConvNets.

Được phát triển bởi Facebook và nổi lên với tốc độ huấn luyện nhanh hơn so với các phương pháp khác. Wav2letter++ sử dụng các khối tích chập 1 chiều (Convolution 1D). Điểm nhấn chính của bài báo:

- Sử dụng các khối Time-Depth Separable dựa trên mạng tích chập 1 chiều.
- Sử dụng CTC layer để căn lề dữ liệu đầu vào và đầu ra.
- Pha giải mã sử dụng Beam Search.



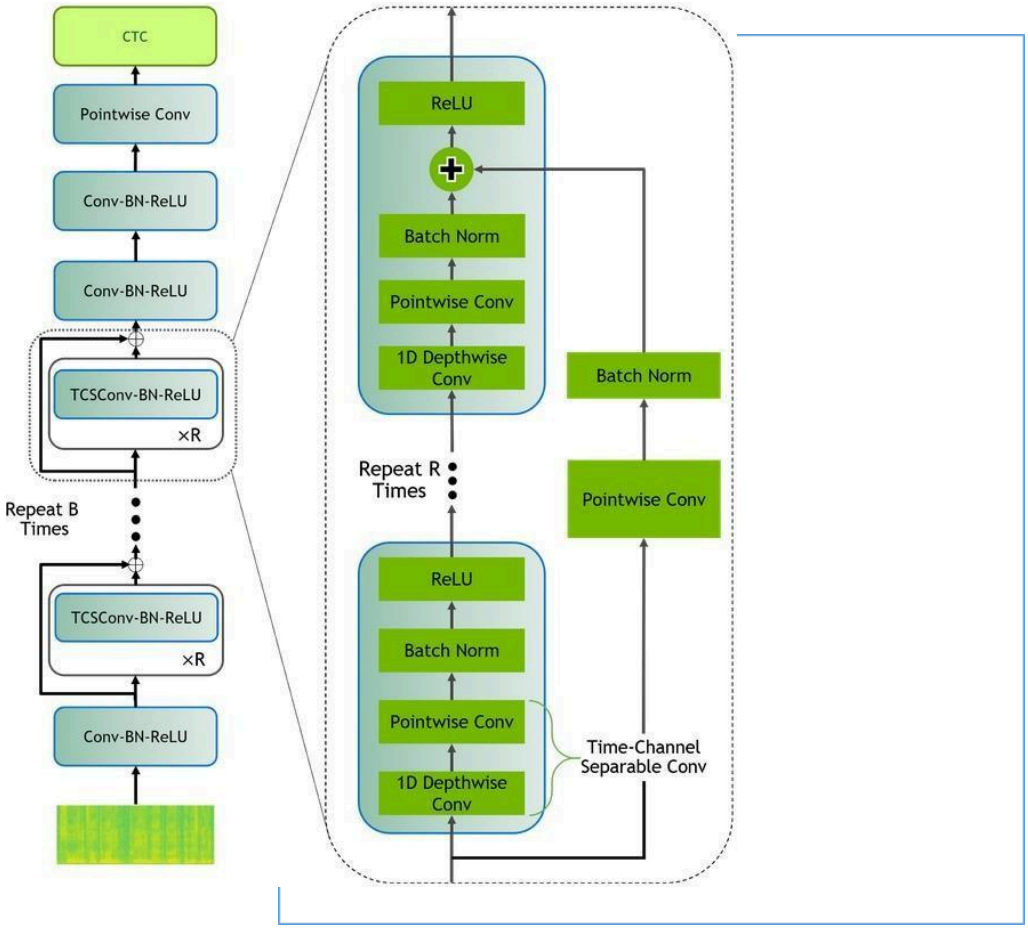
Ảnh 11 Khối Time-Depth Separable

Wav2letter++ cho kết quả tốt hơn mô hình lai Kaldi trên một số tập test, tuy nhiên trên một số tập dữ liệu được ghi nhận vẫn kém hơn mô hình lai.

2.3.4. Mô hình QuartzNet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions

Được phát triển bởi Nvidia, là một mô hình được phát triển dựa trên Jasper nhưng có số lượng tham số nhỏ hơn. Mô hình QuartzNet cũng gần tương tự với Wav2letter++, đều dựa trên mạng tích chập 1 chiều. Mạng tích chập 1D có thể thay thế các mạng truyền

thẳng truyền thống và cho hiệu quả cao hơn, hiện nay rất được ưa chuộng và thường xuyên được sử dụng bởi Facebook.



Ảnh 12 Kiến trúc mạng QuartzNet

Kết quả của QuartzNet tốt hơn mạng Wav2letter++, tuy không thực sự khác biệt quá nhiều về tỉ lệ lỗi nhưng có số lượng tham số nhỏ hơn đáng kể.

Model	Language Model	93-test WER (%)	92-eval WER (%)	# Params, M
RNN-CTC	3-gram	–	8.7	26.5
ResCNN-LAS	3-gram	9.7	6.7	6.6
Wav2Letter++	4-gram	9.5	5.6	17
	convLM	7.5	4.1	
QuartzNet-5×3	**4-gram**	**8.1**	**5.8**	**6.4**
	Transformer-XL	**7.0**	**4.5**	

Table 1 Kết quả so sánh QuartzNet với một số mô hình

2.3.5. PyChain: A Fully Parallelized PyTorch Implementation of LF-MMI for End-to-End ASR

Đây là mô hình được phát triển bởi YiwenShaoStephen. Được cho là mô hình end to end cho kết quả tương đương với mô hình lai Kaldi, đã được tích hợp vào công cụ Espresso. Bằng việc sử dụng LF-MMI, là phương pháp tính hàm mất mát hiệu quả và được cho là có kết quả tốt hơn việc sử dụng CTC layer.

System	# Params (M)	WER (%)
Zeghidour et al. [32]	17	5.6
Baskar et al. [33]	∼100	3.8
Likhomanenko et al. [34]	17	3.6
Zeghidour et al. [35]	17	3.5
Wang et al. [9]	18	**3.4**
Hadian et al. [14]	9.1	4.3
PyChain	6.3	3.5

Table 2 Hiệu năng so sánh của Pychain

Kết quả sử dụng LF-MMI làm hàm mất mát cho thấy kết quả tốt hơn so với phương pháp end-to-end khác và đồng thời đạt được hiệu năng tương đương với khung công việc Kaldi với số lượng tham số nhỏ hơn đáng kể (6 triệu tham số).

2.3.6. Conformer: Convolution-augmented Transformer for Speech Recognition

Conformer được phát triển bởi Google và được phát hành vào tháng 6/2020. Đây được xem là mô hình nhận dạng tiếng nói cho kết quả tốt nhất hiện nay. Mô hình được xây dựng bằng các khối conformer dựa trên kết hợp từ mạng tích chập, mạng truyền tiến, mạng multi-head self attention cụ thể như sau:

Ảnh 13 Khối Conformer

Mạng conformer cho kết quả khá khác biệt so với các phương pháp end-to-end sử dụng CTC layer, và cho kết quả tốt hơn các phương pháp sử dụng transducer.

Method	#Params (M)	WER Without LM		WER With LM	
		testclean	testother	testclean	testother
Hybrid					
Transformer [33]	-	-	-	2.26	4.85
CTC					
QuartzNet [9]	19	3.90	11.28	2.69	7.25
LAS					
Transformer [34]	270	2.89	6.98	2.33	5.17
Transformer [19]	-	2.2	5.6	2.6	5.7
LSTM	360	2.6	6.0	2.2	5.2
Transducer					
Transformer [7]	139	2.4	5.6	2.0	4.6
ContextNet(S) [10]	10.8	2.9	7.0	2.3	5.5
ContextNet(M) [10]	31.4	2.4	5.4	**2.0**	4.5
ContextNet(L) [10]	112.7	**2.1**	4.6	**1.9**	4.1
Conformer (Ours)					
Conformer(S)	10.3	**2.7**	**6.3**	**2.1**	**5.0**
Conformer(M)	30.7	**2.3**	**5.0**	**2.0**	**4.3**
Conformer(L)	118.8	**2.1**	**4.3**	**1.9**	**3.9**

Table 3 Bảng so sánh độ chính xác Conformer

Hiện nay, các mô hình ra đời được xây dựng dựa trên các khối Conformer đã cho kết quả tốt hơn trên một số tập dữ liệu kiểm thử. Đối với nhận dạng tiếng nói Tiếng Việt, Conformer hiện đang là mô hình cho kết quả tốt nhất tại VLSP 2020.

Chương 3. Phương pháp học chủ động cho bài toán nhận dạng tiếng nói.

3.1. Cơ sở lý thuyết [11]

Học chủ động là phương pháp học có tương tác với truy vấn người dùng (hoặc một số nguồn thông tin khác) để gán nhãn các điểm dữ liệu mới với kết quả đầu ra mong muốn. Trong thống kê, nó còn được gọi là thiết kế thực nghiệm tối ưu (Optimal experimental design).

Dữ liệu chưa được gán nhãn thường rất dồi dào và việc gán nhãn thủ công tất cả dữ liệu thường rất tốn kém, có thể gây ra dư thừa cả về thời gian và tiền bạc. Trong trường hợp này, các thuật toán học tập có thể chủ động hỏi người dùng hoặc giáo viên về các nhãn của dữ liệu. Nếu giáo viên trả lời nhãn dữ liệu này đã chính xác hoặc có độ tin cậy cao, thì ta có thể loại bỏ mẫu này trong việc gán nhãn nó. Nếu giáo viên trả lời dữ liệu này không chính xác, cần phải gán nhãn nó thì ta sẽ đưa mẫu này vào cho các nhân viên gán nhãn dữ liệu thực hiện. Lặp đi lặp lại thủ tục này để có thể tìm được tập dữ liệu gán nhãn phù hợp nhất với mô hình và tiết kiệm chi phí này được gọi là học chủ động. Vì người học có thể chọn các mẫu học tốt nhất từ gợi ý của giáo viên, nên số lượng mẫu để học có thể thấp hơn nhiều so với số lượng cần thiết trong cách học có giám sát thông thường.

3.1.1. Định nghĩa cụ thể của phương pháp học chủ động như sau

Giả sử T là tập chứa tất cả các dữ liệu cần xem xét để gán nhãn dữ liệu. Tại mỗi vòng lặp, tập T bao gồm 3 thành phần sau:

- T(K, i): Tập dữ liệu đã được gán nhãn tại vòng lặp i
- T(U, i): Tập dữ liệu chưa biết nhãn tại vòng lặp i.
- T(C, i): là tập con của T(U,i), là tập dữ liệu được chọn để gán nhãn.

Mục tiêu của phương pháp học chủ động là chọn ra tập dữ liệu T(C, i) tốt nhất để đưa vào gán nhãn dữ liệu, sau đó thực hiện huấn luyện mô hình.

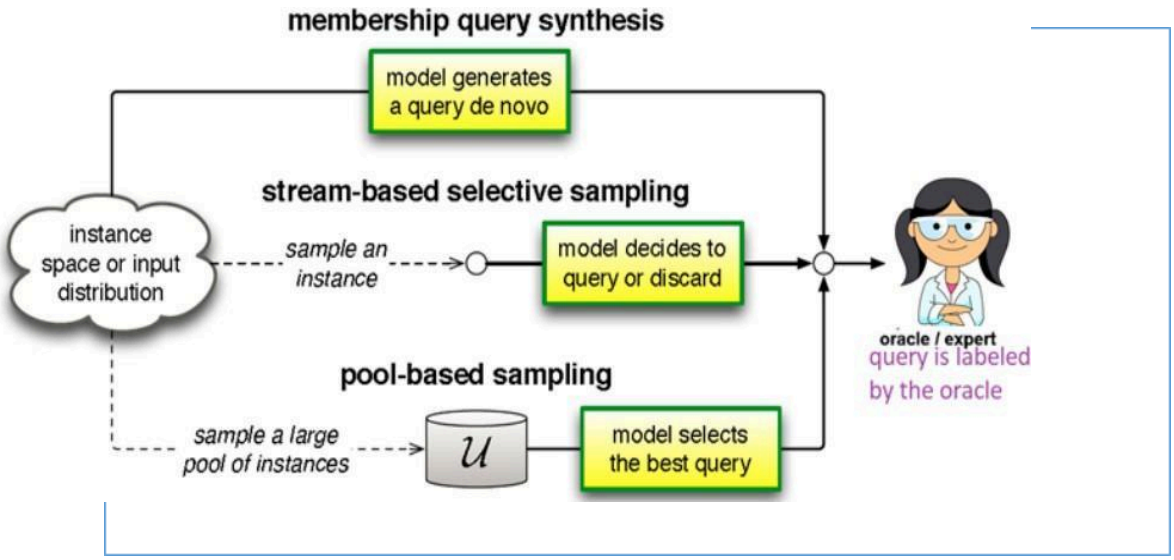
3.1.2. Ngữ cảnh chính của phương pháp học chủ động

Phương pháp học chủ động có 3 ngữ cảnh áp dụng chính:

- Tổng hợp Truy vấn Thành viên (Membership Query Synthesis): Thường chỉ áp dụng với các bài toán theo mô hình sinh, bộ học sẽ tự tạo ra các mẫu truy vấn mới và yêu cầu chuyên gia gán nhãn dữ liệu.

- Lấy mẫu chọn lọc dựa trên luồng (Stream-Based Selective Sampling): Với phương pháp này, từng điểm dữ liệu không được gán nhãn sẽ được kiểm tra lần lượt bởi người học. Người học tự quyết định xem có gán nhãn cho mỗi điểm dữ liệu hay không, nếu không thì điểm dữ liệu này sẽ được bỏ qua.

- Lấy mẫu dựa trên nhóm (Pool-Based Sampling): Trong trường hợp này, các mẫu học được lấy từ toàn bộ dữ liệu và được đánh giá tính thông tin theo từng mẫu. Sau đó, hệ thống sẽ chọn các mẫu chứa nhiều thông tin nhất và truy vấn giáo viên về các nhãn.



Ảnh 14 Các ngữ cảnh chính trong phương pháp học chủ động [12]

3.1.3. Chiến lược truy vấn của phương pháp học chủ động

Có nhiều chiến lược truy vấn, lựa chọn mẫu cho gán nhãn, sau đây là một số chiến lược phổ biến được áp dụng:

- Lấy mẫu không chắc chắn: gắn nhãn những điểm mà mô hình hiện tại ít chắc chắn nhất về kết quả đầu ra chính xác.
- Truy vấn theo hội đồng: xây dựng nhiều mô hình, và huấn luyện các mô hình này trên tập dữ liệu đã gán nhãn. Các mô hình sẽ bỏ phiếu cho dữ liệu không được gán nhãn; gán nhãn những điểm mà "hội đồng" không đồng ý nhất.
- Thay đổi mô hình dự kiến: gán nhãn những điểm có thể thay đổi mô hình hiện tại nhiều nhất.
- Giảm lỗi mong đợi: gán nhãn những điểm có thể làm giảm nhiều nhất lỗi tổng quát của mô hình.

3.2. Một số áp dụng phương pháp học chủ động cho bài toán nhận dạng tiếng nói.

3.2.1. Active Learning For Automatic Speech Recognition [13]

Đây là công trình nghiên cứu từ rất sớm (năm 2002) về học chủ động cho bài toán nhận dạng tiếng nói của AT&T Lab. Do công trình nghiên cứu này ra đời từ sớm nên việc học đơn giản là sử dụng phương pháp lọc dựa trên tiêu chí độ tin cậy của đầu ra của mạng huấn luyện.

1. Train acoustic and language models, AM_i and LM_i, for recognition, using S_t (i is the iteration number)
2. Recognize the utterances in set S_u using AM_i and LM_i, and compute the confidence scores for all the words
3. Compute confidence scores of utterances
4. Select k utterances which have the smallest confidence scores from S_u, and transcribe them. Call the new transcribed set as S_i
5. $S_t = S_t \bigcup S_i$; $S_u = S_u - S_i$
6. Stop if word accuracy has converged, otherwise go to Step 1

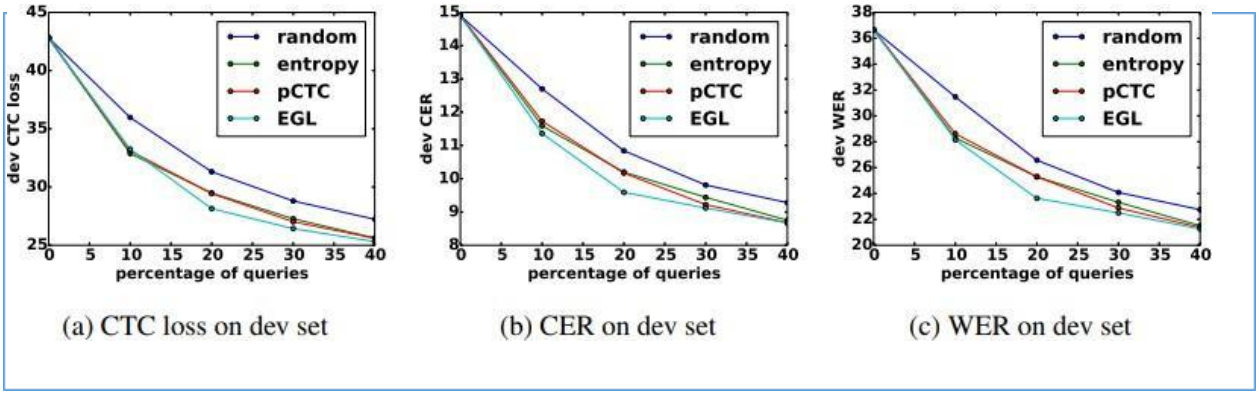
Ảnh 15 Các bước chính được thực hiện bằng phương pháp học chủ động

Các bước chính của thuật toán bao gồm việc huấn luyện mô hình ngữ âm và mô hình ngôn ngữ. Sau đó nhận dạng danh sách các câu cần gán nhãn dựa vào mô hình cần học. Sau đó đánh giá độ chính xác kết quả nhận dạng của từng câu. Tại mỗi vòng lặp sẽ chọn ra k câu có độ tin cậy nhỏ nhất và đưa nó vào gán nhãn.

Kết quả thí nghiệm được các tác giả báo cáo là giảm 27% lượng dữ liệu và cho độ chính xác tương đương.

3.2.2. Active Learning for Speech Recognition: the Power of Gradients [14]

Mô hình nhận dạng được áp dụng cho bài toán nhận dạng tiếng nói được nhóm tác giả sử dụng là mạng RNN và CTC layer. Trong bài báo của các tác giả, tác giả đánh giá phương pháp học chủ động theo nhiều tiêu chí lựa chọn như: Chọn mẫu ngẫu nhiên, chọn mẫu dựa trên độ tin cậy, chọn mẫu dựa trên sự thay đổi của mô hình (Expected Gradient Length)



Ảnh 16 Đánh giá độ chính xác theo các tiêu chí lựa chọn

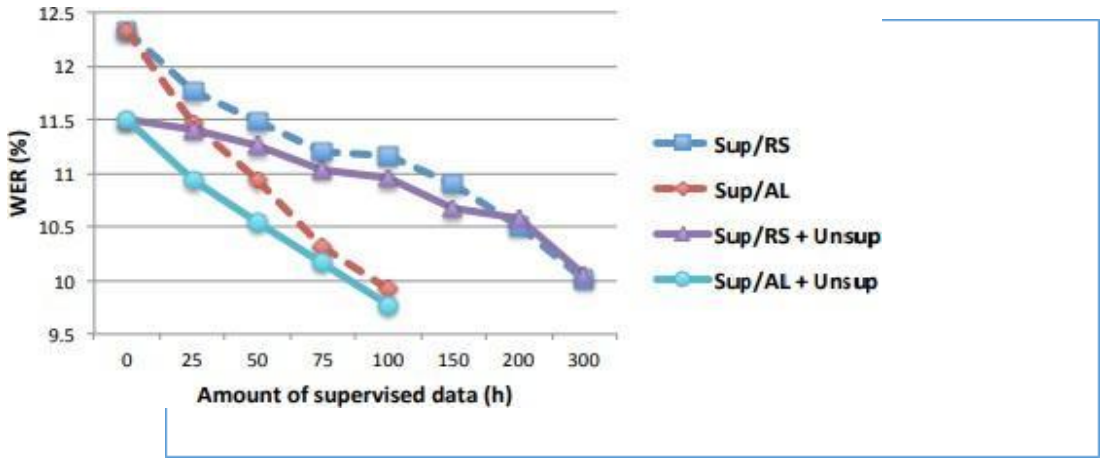
Kết quả thí nghiệm cho thấy rằng việc sử dụng phương pháp học chủ động cho kết quả tốt hơn so với phương pháp chọn ngẫu nhiên. Trong đó, phương pháp dựa trên sự thay đổi của mô hình cho kết quả tốt nhất, tuy nhiên sau khi lựa chọn được dữ liệu nhất định thì sẽ cho kết quả tương đương.

3.2.3. Active and Semi-Supervised Learning in ASR: Benefits on the Acoustic and Language Models [15]

Được công bố vào năm 2019 bởi các nhà nghiên cứu ở Amazon. Thí nghiệm của các tác giả được thực hiện trên mô hình lai HMM-DNN. Đóng góp chính của các tác giả là việc

áp dụng phương pháp học chủ động trong việc lựa chọn dữ liệu quan trọng cho gán nhãn và đánh giá tính hiệu quả của việc kết hợp với việc học bán giám sát.

Đối với việc sử dụng Active Learning, bài báo dựa trên tiêu chí confidence score. Các tác giả thực nghiệm rất nhiều phạm vi lựa chọn và tìm ra được phương pháp lựa chọn tốt nhất là việc lựa chọn ngẫu nhiên với các mẫu dữ liệu có độ tin cậy thấp từ 0 tới 0.7. Với việc thêm 100h dữ liệu được chọn bởi Active Learning thì kết quả giảm 2% tỷ lệ lỗi tương quan so với phương pháp chọn ngẫu nhiên.



Ảnh 17 Kết quả áp dụng phương pháp học chủ động và học bán giám sát

Ngoài ra, các tác giả còn kết hợp huấn luyện với các mẫu dữ liệu bán giám sát. Tuy nhiên, khi lượng dữ liệu đủ nhiều, việc thêm các dữ liệu bán giám sát sẽ mất dần đi tính hiệu quả của nó.

Chương 4. Cài đặt thực nghiệm.

Luận văn thí nghiệm trên 2 bộ dữ liệu. Cả hai bộ dữ liệu đều có khối lượng 100 giờ dữ liệu. Bộ dữ liệu thứ nhất (Set 1) là bộ dữ liệu có số lượng trùng lặp ít, ngữ cảnh đa dạng là các hội thoại sinh hoạt thường ngày. Bộ dữ liệu thứ hai (Set 2) là bộ dữ liệu có âm thanh rõ ràng, có số lượng trùng lặp rất nhiều do chỉ tập trung vào một vài ngữ cảnh.

Bộ dữ liệu	Số lượng (giờ)	Số câu	Độ dư thừa
Set 1	100	140543	4% mức câu
Set 2	100	124870	20% mức câu

Bảng 4 Tập dữ liệu kiểm thử

Mô hình ASR được luận văn lựa chọn để làm thực nghiệm là mô hình Kaldi. Độ tin cậy nhận dạng (Confidence Score) của dữ liệu kiểm thử được tổng hợp từ 2 điểm: Một điểm (acoustic score) là độ chính xác của mô hình nhận dạng âm, một điểm (language model score) là độ chính xác về mặt ngữ nghĩa theo mô hình ngôn ngữ. Tại mỗi vòng lặp trong học chủ động, các bài báo đã công bố thường chỉ sử độ tin cậy mặc định từ đầu ra của bộ giải mã (tỉ lệ trọng số giữa mô hình âm học và mô hình ngôn ngữ là 1:1). Trong luận văn này sẽ thực nghiệm việc lựa chọn dữ liệu bằng cách sử dụng kết hợp dựa trên cả 2 tiêu chí.

Thí nghiệm 1: Thí nghiệm ảnh hưởng của dư thừa dữ liệu đến hiệu quả phương pháp học chủ động

Dataset	Test 1	Test 2
Random	31.34	22.48
AL	30.87	21.35

Tỉ lệ lỗi tương quan	1%	5%

Bảng 5 Bảng thí nghiệm so sánh AL và phương pháp ngẫu nhiên (đơn vị WER)

Word Error Rate (WER) là tỉ lệ lỗi từ của một hệ thống nhận dạng tiếng nói khi nhận dạng một tín hiệu âm thanh đầu vào. Tỉ lệ WER càng tiến gần đến 0 thì hệ thống nhận dạng tiếng nói càng tốt. Tỉ lệ lỗi tương quan là tỉ lệ phần trăm cải tiến của phương pháp sử dụng học chủ động so với phương pháp lựa chọn ngẫu nhiên (lựa chọn mẫu dữ liệu ngẫu nhiên để gán nhãn)

Từ bảng thực nghiệm trên, ta thấy rằng hiệu quả của phương pháp học chủ động phụ thuộc nhiều vào độ dư thừa của dữ liệu. Trên tập dữ liệu kiểm thử thứ nhất, đây là bộ dữ liệu có độ dư thừa thông tin ít (dữ liệu trùng lặp ít, ngữ cảnh đa dạng) nên việc áp dụng phương pháp học chủ động chỉ cho kết quả cải tiến 1% so với phương pháp lựa chọn ngẫu nhiên. Ngược lại đối với bộ dữ liệu thứ hai, bộ dữ liệu này có độ trùng lặp và phân bố với mật độ dày nên việc áp dụng Active Learning cho kết quả tốt hơn, sai lệch 5% so với phương pháp chọn ngẫu nhiên. Đây là điều dễ hiểu đối với tính chất của phương pháp học chủ động. Phương pháp học chủ động là phương pháp lựa chọn ra những mẫu quan trọng nhất cho việc huấn luyện, tức là các mẫu cần phải chứa nhiều thông tin nhất, ít bị dư thừa nhất. Đối với dữ liệu có lượng dư thừa thông tin lớn, việc sử dụng phương pháp học chủ động có thể loại bỏ phần lớn thông tin dư thừa không có ích và dễ dàng chọn lựa thông tin ý nghĩa hơn.

Để phân tích độ dư thừa của một tập dữ liệu, người đọc có thể dựa vào các thuật toán phân cụm và lý thuyết dư thừa thông tin, chi tiết tham khảo tại đồ án tốt nghiệp đại học [17] của Nguyễn Văn Phong, anh Đỗ Văn Hải và tôi thực hiện.

Thí nghiệm 2: Thí nghiệm lựa chọn dữ liệu theo 2 tiêu chí điểm âm học (acoustic score) và điểm ngôn ngữ (language score).

Các bài báo áp dụng phương pháp active learning thường chỉ sử dụng một độ đo tin cậy được tổng hợp từ kết quả đầu ra của bộ giải mã. Tuy nhiên, ảnh hưởng của hai độ đo là hoàn toàn khác nhau, dẫn tới sự sai lệch đối với thang điểm đánh giá của tiêu chí lựa chọn.

Với hệ thống có mô hình âm học tốt, nhưng mô hình ngôn ngữ kém, ta cần ưu tiên gán nhãn những mẫu ví dụ học tốt cho mô hình ngôn ngữ. Tương tự ngược lại với mô hình ngôn ngữ tốt và mô hình âm học có độ chính xác kém, ta cần ưu tiên chọn mẫu có độ chính xác thấp đối với mô hình âm học. Do vậy, luận văn đề xuất ý tưởng lựa chọn dữ liệu gán nhãn đồng thời dựa trên cả 2 tiêu chí lựa chọn: mô hình âm học, mô hình ngôn ngữ. Với mỗi tiêu chí lựa chọn một nửa dữ liệu so với phương pháp thông thường.

Để đánh giá hiệu quả của phương pháp lựa chọn dữ liệu này, luận văn thực hiện đánh giá trên bộ dữ liệu Test 2, là bộ dữ liệu có lượng dư thừa lớn. Dữ liệu được chia thành các tập dữ liệu sau:

- 50 giờ (Gọi là tập dữ liệu D) được đưa vào huấn luyện.
- 50 giờ còn lại (Gọi là tập dữ liệu P) sẽ sử dụng phương pháp học chủ động để lựa chọn ra một số mẫu dữ liệu để gán nhãn.
- Tập dữ liệu 15 giờ kiểm thử (Gọi là tập dữ liệu

T) Sơ đồ thuật toán cụ thể như sau:

- Bước 1: Huấn luyện mô hình trên tập dữ liệu D.
- Bước 2: Đưa tập dữ liệu P vào mô hình đã huấn luyện để giải mã.
- Bước 3: Đưa tập dữ liệu T vào mô hình huấn luyện để giải mã và lấy độ chính xác của mô hình, và xuất ra độ chính xác của tập dữ liệu T tại mỗi vòng lặp để đánh giá hiệu quả của phương pháp học chủ động.
- Bước 4: Tính acoustic weight và language model weight cho mỗi mẫu dữ liệu đã giải mã.
- Bước 5: Cập nhật trọng số của lattice (Đồ thị biểu diễn máy chuyển đổi trạng thái hữu hạn) các câu theo 2 tỉ lệ **1:alpha** (acoustic trội với alpha < 1) và **alpha:1** (language model trội với alpha < 1). Sau đó tính độ tin cậy của mỗi giả thuyết trong lattice (mỗi giả thuyết trong đồ thị lattice là một câu tương ứng với tín hiệu âm thanh đầu vào được mô hình tiếng nói nhận dạng) để được độ tin cậy của chúng.

- Bước 6: Chọn ngẫu nhiên 2500 câu có độ tin cậy acoustic trội từ 0=>0.8 và 2500 câu có độ tin cậy language model trội từ 0 => 0.8. Nếu không có câu nào có độ tin cậy từ 0=>0.8 thì sẽ nhảy tới bước 8 (Kết thúc).
- Bước 7: Lấy 5000 câu này và nhãn văn bản đã được gán của nó đưa vào tập dữ liệu D để huấn luyện. Đồng thời loại bỏ 5000 câu này khỏi tập dữ liệu P. Sau đó lặp lại bước 1.
- Bước 8: Kết thúc

Đối với phương pháp lựa chọn ngẫu nhiên, thay bước 5 và bước 6 bằng lựa chọn ngẫu nhiên 5000 câu có độ tin cậy (confidence score) từ 0=>0.8.

Tại sao lựa chọn 5000 câu tại mỗi vòng lặp?

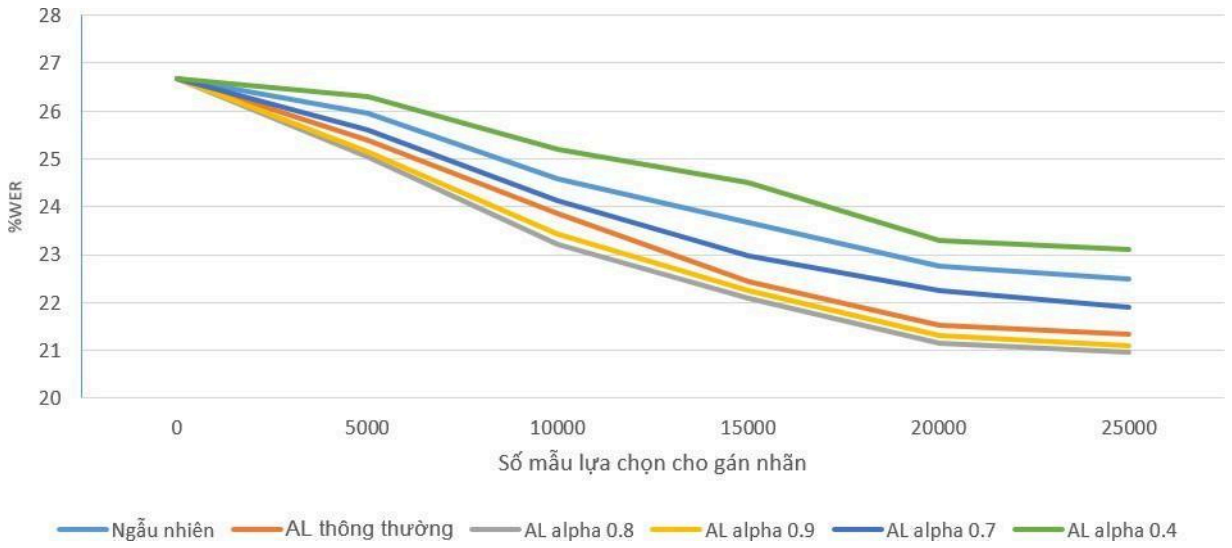
Thông thường tại mỗi vòng lặp, số câu lựa chọn càng nhỏ càng tốt, vì điều này có thể giảm lượng dư thừa trong tập mẫu dữ liệu gán nhãn vừa chọn (Dữ liệu vừa chọn có dư thừa đối với tập dữ liệu huấn luyện thấp, nhưng trong bản thân dữ liệu vừa chọn lại có nhiều dữ liệu trùng lặp giống nhau). Tuy nhiên do thời gian huấn luyện một mô hình nhận dạng tiếng nói cần rất nhiều thời gian, lên đến vài ngày hoặc vài tuần với mỗi một vòng lặp thí nghiệm selection, cộng thêm việc lựa chọn ngẫu nhiên những mẫu có độ tin cậy thấp, ta có thể ước lượng một con số vừa đủ để sau một số vòng lặp, ta thu được một lượng dữ liệu để gán nhãn tương ứng. Ở đây, chúng tôi chọn 5000 câu, tương ứng với 4 vòng lặp huấn luyện để thu được khoảng 20h cho việc gán nhãn.

Tại sao lựa chọn ngưỡng độ tin cậy từ 0=>0.8?

Đây là ngưỡng có độ tin cậy cho các mẫu đủ thấp để cần gán nhãn lại dữ liệu. Việc lựa chọn ngẫu nhiên mà không lựa chọn theo top các mẫu có độ tin cậy thấp nhất vì tránh việc nhiều mẫu giống nhau và có độ tin cậy thấp sẽ được chọn cùng nhau. Thí nghiệm các khoảng lựa chọn này cũng được chỉ ra tại mục 3.2.1 của tác giả bài báo 15 với giá trị từ 0 => 0.7. Tuy nhiên, trong thí nghiệm của luận văn, số lượng mẫu có độ tin cậy từ 0=> 0.7 có số lượng khá thấp nên luận văn đã điều chỉnh lên 0.8 để có thể thu được nhiều mẫu hơn.

Số câu lựa chọn	Random	AL thông thường alpha 1.0	AL alpha 0.8	AL alpha 0.9	AL alpha 0.7	AL alpha 0.4
0	26.67	26.67	26.67	26.67	26.67	26.67
5000	25.94	25.38	25.04	25.14	25.6	26.3
10000	24.59	23.85	23.22	23.42	24.12	25.2
15000	23.67	22.44	22.08	22.25	22.96	24.5
20000	22.75	21.53	21.15	21.3	22.24	23.3
25000	22.48	21.35	20.95	21.1	21.9	23.1

Bảng 6 Thí nghiệm với tham số alpha (đơn vị %WER - Phần tram tỉ lệ lỗi từ khi giải mã)



Ảnh 18 Đồ thị bảng 7

Thí nghiệm so sánh phương pháp đề xuất với phương pháp lựa chọn ngẫu nhiên các mẫu để đưa vào gán nhãn. Kết quả cho thấy rằng các phương pháp sử dụng chọn mẫu dựa trên Confidence Score và phương pháp tách riêng theo 2 tiêu chí trọng số về âm học và trọng số về ngôn ngữ cho kết quả tốt hơn so với phương pháp ngẫu nhiên và đạt kết quả tốt nhất là 20.95% tỉ lệ lỗi từ (WER) trong tập dữ liệu kiểm thử.

Tỉ lệ trội giữa mô hình âm học và mô hình ngôn ngữ cho kết quả tốt nhất ở ngưỡng alpha=0.8 và giảm dần khi alpha có giá trị nhỏ hơn. Điều này có thể giải thích do lý do khi tỉ lệ trội quá cao, thì độ đo tổng hợp quá thiên về một độ đo và mất đi tính chính xác của độ tin cậy (Khi alpha giảm tới 0.4 cho hiệu quả kém phương pháp ngẫu nhiên). Lúc này việc lựa chọn ngẫu nhiên các câu có tỉ lệ tin cậy thấp không còn chính xác nữa.

Việc tách việc lựa chọn dữ liệu theo 2 tiêu chí cho kết quả tốt hơn từ 1% tới 3% tỉ lệ lỗi tương quan so với phương pháp thông thường và 2% tới 5% so với phương pháp ngẫu nhiên tùy vào mỗi vòng lặp.

Chương 5: Kết luận.

Những vấn đề đã giải quyết trong luận văn

- Luận văn đã tiến hành khảo sát bài toán nhận dạng tiếng nói. Đây là bài toán có ứng dụng rất nhiều trong thực tế và được phát triển bởi rất nhiều trường đại học và tập đoàn công nghệ lớn. Luận văn đã trình bày sơ lược quá trình phát triển của bài toán nhận dạng tiếng nói. Đồng thời, khảo sát các mô hình nhận dạng tiếng nói mới nhất để người mới tiếp cận có cái nhìn tổng quan và dễ dàng xây dựng một hệ thống nhận dạng tiếng nói.
- Độ hiệu quả của phương pháp học chủ động phụ thuộc vào tính chất và độ dư thừa thông tin của dữ liệu, dữ liệu có độ dư thừa càng lớn thì hiệu quả của phương pháp học chủ động càng cao. Do đó, khi bạn bắt đầu sử dụng phương pháp học chủ động trên một tập dữ liệu, hay bài toán nào đó, trước tiên bạn cần phân tích độ dư thừa của dữ liệu, dữ liệu có phân bố đều, hay tập trung thành các cụm với mật độ cao hay không. Nếu độ dư thừa cao thì việc áp dụng phương pháp học chủ động sẽ rất hiệu quả. Nếu độ dư thừa thấp thì có thể việc áp dụng phương pháp học chủ động sẽ cho kết quả tốt hơn không đáng kể so với phương pháp chọn ngẫu nhiên.
- Bài toán nhận dạng tiếng nói dựa trên 2 độ đo là độ đo về mặt âm học và về mặt ngôn ngữ. Phương pháp học chủ động với việc tách biệt 2 độ đo có thể giúp bổ sung các dữ liệu tốt về mặt âm học nếu mô hình âm học không tốt, hoặc nếu mô hình âm học đã tốt mà mô hình ngôn ngữ không tốt thì sẽ bổ sung được đúng mẫu tốt cho mô hình ngôn ngữ. Điều này sẽ giúp cải tiến hơn so với phương pháp thông thường chỉ sử dụng trên một độ đo kết hợp.

Công việc nghiên cứu trong tương lai

- Tỉ lệ lỗi thực tế của đầu ra có độ tương quan không lớn với tiêu chí độ tin cậy của đầu ra mô hình học. Điều này dẫn đến việc sử dụng tiêu chí này để lựa chọn dữ liệu quan trọng không thực sự quá hiệu quả với lỗi thực tế. Do đó, ta cần cải tiến độ tương quan của Confidence Score với Word Error Rate.

- Việc sử dụng phương pháp học chủ động cho bài toán nhận dạng tiếng nói tốn khá nhiều thời gian do mỗi vòng lặp lựa chọn được dữ liệu quan trọng thì cần phải huấn luyện lại mô hình. Điều này gây khó khăn cho việc triển khai thực tế khi không kịp tiến độ gán nhãn. Do đó, ta cần tìm hiểu các phương pháp Transfer Learning hiệu quả để giảm thời gian huấn luyện lại mô hình giữa các vòng lặp.
- Thí nghiệm, đánh giá phương pháp lựa chọn dữ liệu chỉ dựa vào tiêu chí của mô hình ngôn ngữ. Điều này có thể bỏ qua bước huấn luyện mô hình âm học mất rất nhiều thời gian.

TÀI LIỆU THAM KHẢO

Tiếng Anh

[1] https://info.keylimeinteractive.com/history-of-voice-technology

[2] https://www.vox.com/2017/5/31/15720118/google-understand-language-speech-equivalent- humans-code-conference-mary-meeker

[3] https://paperswithcode.com/sota/speech-recognition-on-librispeech-test-clean

[4] **Daniel Povey và cộng sự** (2011). *The Kaldi Speech Recognition Toolkit IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*

[5] **M. Ravanelli, T. Parcollet, Y. Bengio** (2018). *The PyTorch-Kaldi Speech Recognition Toolkit*

[6] **A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng.** (2014) *Deep speech: Scaling up end-to-end speech recognition*

[7] **Pratap và cộng sự (2020).** *Scaling Online Speech Recognition Using ConvNets*

[8] **Yiwen Shao, Yiming Wang, Daniel Povey, Sanjeev Khudanpur** (2020). *PyChain: A Fully Parallelized PyTorch Implementation of LF-MMI for End-to-End ASR*

[9] **Samuel Kriman và cộng sự** (2019). *QuartzNet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions*

[10] **Anmol Gulati và cộng sự** (2020). *Conformer: Convolution-augmented Transformer for Speech Recognition*

[11] **Philip Bachman, Alessandro Sordoni, Adam Trischler** (2017*) Learning Algorithms for Active Learning*

[12] **Settles, Burr** (2010). *Active learning literature survey.*

[13] **Dilek Hakkani-Tür và cộng sự** (2002). *Active learning for automatic speech recognition*

[14] **Jiaji Huang và cộng sự** (2016). *Active Learning for Speech Recognition: the Power of Gradients*

[15] **Thomas Drugman, Janne Pylkkonen, Reinhard Kneser** (2019). *Active and Semi-Supervised Learning in ASR: Benefits on the Acoustic and Language Models*

[16] **Karpagavalli and Chandra và cộng sự** (2016). *A Review on Automatic Speech Recognition Architecture and Approaches.*

Tiếng Việt

[17] **Nguyễn Văn Phong, Đỗ Văn Hải, Nguyễn Minh Sơn.** *Đồ án tốt nghiệp Đại học Thủy Lợi - Phương pháp lựa chọn dữ liệu quan trọng cho quá trình gán nhãn và huấn luyện mô hình nhận dạng tiếng nói.*