HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

BÙI THÁI DUY

PHÁT HIỆN TIẾNG NGÁY DỰA TRÊN HỌC SÂU

Chuyên ngành : **HỆ THỐNG THÔNG TIN**

Mã số : **60.48.01.04**

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI - 2020

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: PGS.TS. PHẠM VĂN CƯỜNG

Phản biện 1:

……

Phản biện 2:

……..

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ..... giờ ..... ngày ..... tháng .... năm 2020

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông.

MỞ ĐẦU

Sự tiến bộ của công nghệ đã thúc đẩy cộng đồng nghiên cứu chuyển từ truyền, thu nhận và xử lý dữ liệu mức thấp sang nghiên cứu tích hợp thông tin mức cao, xử lý ngữ cảnh, nhận dạng và suy diễn các hoạt động.

Bên cạnh tác động tới chất lượng giấc ngủ của con người thì ngáy cũng có dấu hiệu của chứng ngưng thở (OSA) sau khi mất ngủ, tỷ lệ mắc bệnh rối loạn giấc ngủ cao nhất, ảnh hưởng đến khoảng 3 - 7% đàn ông trung niên và 2-5% phụ nữ trung niên trong dân số nói chung. OSA được đặc trưng bởi các đợt lặp đi lặp lại của sự khó khăn một phần hoặc hoàn toàn của đường hô hấp trên trong khi ngủ, gây ra sự trao đổi khí bị suy yếu và rối loạn giấc ngủ.

Tổng quan về vấn đề nghiên cứu

Những nghiên cứu trong học sâu từ trước tới nay đã và đang được sử dụng để giải quyết nhiều bài toán về nhận dạng, phát hiện đặc biệt trong lĩnh vực thị giác máy tính. Vì đòi hỏi cần một lượng dữ liệu, thời gian, sức mạnh tính toán đáng kể, các nỗ lực nghiên nghiên cứu cách để tận dụng các mạng CNN được đào tạo trước cho các nhiệm vụ khác như mạng CNN được sử dụng trong các hệ thống nhận dạng. Cho đến nay, rất ít các nghiên cứu thực hiện để khám phá biểu diễn đặc trưng của âm thanh với mạng CNN.

Mục đích, đối tượng, phạm vi và phương pháp nghiên cứu

Đề tài *"Phát hiện tiếng ngáy dựa trên học sâu"* được thực hiện trong khuôn khổ luận văn thạc sĩ chuyên ngành hệ thống thông tin nhằm góp phần đánh giá một số như việc xử lý, lưu trữ âm thanh được thực hiện qua việc xử lý ảnh phổ, kết hợp được việc so sánh, đánh giá các kiến trúc học sâu trong việc phát hiện tiếng ngáy..

Nghiên cứu các kỹ thuật học sâu phù hợp cho bài toán Phát hiện tiếng ngáy dựa trên học sâu.

Nghiên cứu phương pháp phân lớp tiếng ngáy dựa trên phân lớp ảnh dựa trên mạng neural tích chập (CNN) hoặc mô hình hồi quy RNN của quang phổ âm thanh. Để phân lớp các ảnh trên thì sử dụng làm vector đặc trưng.

Nghiên cứu các phương pháp phân lớp phù hợp cho bài toán Phát hiện tiếng ngáy dựa trên học sâu

Cấu trúc của luận văn

Ngoài phần mở đầu và kết luận, luận văn được chia thành ba chương:

Chương 1: Tổng quan về phát hiện tiếng ngáy

Nội dung chương này sẽ bao gồm giới thiệu chung về bài toán phát hiện tiếng ngáy, những khó khăn và ý nghĩa của bài toán này. Chương này cũng trình bày về các nghiên cứu liên quan với các vấn đề về phát hiện âm thanh, nghiên cứu về học máy cũng như học sâu .Từ những cơ sở nghiên cứu này sẽ xác định rõ hướng nghiên cứu của luận văn.

Chương 2: Phương pháp phát hiện và theo dõi tiếng ngáy

Trình bày một số phương pháp học sâu có có tốc độ tính toán nhanh phù hợp với bài toán phát hiện và theo dõi tiếng ngáy. Các âm thanh được trích rút đặc trưng và đi qua các mô hình học sâu như CNN hoặc mô hình hồi quy RNN.

Chương 3: Thử nghiệm và đánh giá

Trong chương này sẽ trình bày các vấn đề: thu thập dữ liệu tiếng ngày; thử nghiệm mô hình CNN hoặc mô hình hồi quy RNN phân tích các âm thanh qua đó có thể đánh giá được các kiến trúc học sâu trong việc phát hiện tiếng ngáy.

CHƯƠNG 1: TỔNG QUAN VỀ PHÁT HIỆN TIẾNG NGÁY

Bài toán phát hiện tiếng ngáy

1.1.1 Tầm quan trọng của tiếng ngáy

Ngáy ngày càng được công nhận là mối quan tâm về sức khỏe cộng đồng. Đây là một
1.1 vấn đề phổ biến ở người lớn và là dấu hiệu của hội chứng ngưng thở khi ngủ do tắc nghẽn (OSA). Một số nghiên cứu về y tế đã chỉ ra các yếu tố liên quan chính đến ngáy dựa trên nghiên cứu đó là lão hóa, giới tính nam, tăng huyết áp, buồn ngủ ban ngày, hút thuốc và huyết thống.

Bên cạnh tác động tới chất lượng giấc ngủ của con người thì ngáy cũng có dấu hiệu của chứng ngưng thở (OSA) sau khi mất ngủ, tỷ lệ mắc bệnh rối loạn giấc ngủ cao nhất, ảnh hưởng đến khoảng 3 - 7% đàn ông trung niên và 2-5% phụ nữ trung niên trong dân số nói chung. OSA được đặc trưng bởi các đợt lặp đi lặp lại của sự khó khăn một phần hoặc hoàn toàn của đường hô hấp trên trong khi ngủ, gây ra sự trao đổi khí bị suy yếu và rối loạn giấc ngủ

1.1.2 Phát biểu bài toán

Đầu vào: Một chuỗi âm thanh

Đầu ra: Phát hiện âm thanh là tiếng ngáy hay không

Với đầu vào là “chuỗi âm thanh” hệ thống sẽ đưa ra được trong chuỗi âm thanh đó có tiếng ngáy hay không không phải tiếng ngáy, hay một nhóm các âm thanh vào thì hệ thống sẽ phát hiện được có bao nhiêu âm thanh trong đó là tiếng ngáy.

1.1.3 Ý nghĩa bài toán

Các nghiên cứu liên quan

1.2.1 Thiết bị phát hiện tiếng ngáy

1.2.2 Mô hình học máy cổ điển trong phát hiện tiếng ngáy

1.2.3 Mô hình học sâu phát hiện tiếng ngáy

1.2.4 Đánh giá các nghiên cứu

Các nghiên cứu gần đây về học máy hay học sâu đã trở thành xu thế nghiên cứu của các nhà khoa học trên thế giới và trong nước. Cùng với đó là một xu thế mới trong việc phát triển các ứng dụng khác nhau mà có sự hỗ trợ của học máy/học sâu để giải quyết những bài toán mà trước đây vô cùng phức tạp hoặc mất nhiều chi phí.

Từ những ngày đầu, các ứng dụng của trí tuệ nhân tạo đã giải quyết các vấn đề đơn nhất, và đến tận ngày nay các ứng dụng này đã phát triển một cách vượt trội qua các ứng dụng phức tạp đòi hỏi việc xử lý thông minh.

1.3 Kết luận chương

Chương 1 đã giới thiệu tổng quan về bài toán phát hiện tiếng ngáy. Tìm hiểu bài toán phát phân loại âm thanh và giới thiệu bài toán phát hiện tiếng ngáy, kèm theo đó là các nghiên cứu liên quan từ các ứng dụng, giải pháp mà được thực hiện từ bài toàn, các mô hình giải quyết bài toán, và các đánh giá về các nghiên cứu qua đó đưa ra những vấn đề cần làm rõ và giải quyết trong luận văn.

Trong chương 2, luận văn sẽ trình bày về hướng giải quyết cho bài toán phát hiện tiếng ngáy, các bước tiến hành khi giải bài toán nhận dạng, phát hiện tiếng ngáy, các đặc trưng của âm thanh, các thành phần xử lý âm thanh và đi sâu hơn trình bày về phương pháp sẽ áp dụng để giải quyết bài toán. Đây cũng là nền tảng cho phương hướng của việc thực nghiệm giải quyết bài toán đã đề ra.

CHƯƠNG 2: PHƯƠNG PHÁP PHÁT HIỆN VÀ THEO DÕI TIẾNG NGÁY

Phương pháp giải quyết bài toán

Để giải quyết bài toán phát hiện và theo dõi tiếng ngáy từ "âm thanh ngáy" đầu vào, mục tiêu cần phải phân lớp và đưa những âm thanh này về lớp "Âm thanh ngáy" và những
2.1 âm thanh còn không phải âm thanh ngáy thì sẽ đưa về lớp "Không phải âm thanh ngáy". Luận văn đã tham khảo và tìm hiểu được các bước thực hiện để xây dựng phương pháp phát hiện và theo dõi tiếng ngáy và được chia làm 2 giai đoạn: huấn luyện và kiểm thử.

Hai giai đoạn huấn luyện và kiểm thử trong phát hiện tiếng ngáy được mô tả như các hình phía trên. Các bước thực hiện trong luận văn sẽ gồm các bước từ trái sang phải sau:

1. Chia dữ liệu thành 2 phần: dữ liệu huấn luyện và dữ liệu kiểm thử
2. Tiền xử lý dữ liệu huấn luyện và kiểm thử trước khi lựa chọn ra các vector đặc trưng, điều này sẽ loại bỏ đi các thông tin có giá trị thấp.
3. Vectơ đặc trưng trích đặc trưng cho tập dữ liệu đã qua tiền xử lý, tại đây sẽ có các đặc trưng riêng của các bài toán được thể hiện ra.
4. Áp dụng các mô hình học sâu (mô hình CNN, mô hình LSTM, mô hình CNN-LSTM) để giải quyết bài toán và so sánh với mô hình học nông
5. Đưa ra mô hình sau khi huấn luyện và kết quả sau khi kiểm thử qua mô hình, từ đó đưa ra được kết quả và đánh giá bài toán

Tại bước 1, luận văn sẽ áp dụng phương pháp cross validation và chia dữ liệu thành 2 phần gồm phần dữ liệu huấn luyện 90%, phần dữ liệu kiểm thử 10% . Cụ thể về phương pháp cross validation sẽ được luận văn trình bày tại mục 3.1 về thu thập dữ liệu.

Trong bước 2, tiền xử lý, dữ liệu đầu vào là âm thanh cần phải loại bỏ các yếu tố dư thừa của dữ liệu như các đoạn không có thu nhận được âm thanh...

Các phần tiếp theo của chương sẽ trình bày chi tiết về các phương pháp, mô hình và đề xuất lựa chọn và áp dụng vào việc phát hiện tiếng ngáy trong hệ thống phát hiện, theo dõi tiếng ngáy..

Xử lý âm thanh

2.2.1 Biến đổi Fourier (FT)

Âm thanh là một chuỗi tín hiệu dài biến thiên theo thời gian, nhưng hàm lượng thông tin trong đó không nhiều. Kết quả sẽ nhận được 1 cách biểu diễn giàu thông tin hơn so với cách biểu diễn thông thường.

2.2

Công thức biến đổi Fourier cho hàm f(x) liên tục trong công thức (2.1) :

$$f(x) = \int_{-\infty}^{\infty} F(k) e^{2\pi i k x} dk \quad \textbf{(2.1)}$$

$$F(k) = \int_{-\infty}^{\infty} f(x) e^{-2\pi i k x} dx \quad \textbf{(2.2)}$$

Trong đó, F(k) là công thức biến đổi fourier ngược trong công thức 2.2

Công thức biến đổi Fourier rời rạc (DFT) trong công thức 2.3

$$X(k) = \sum_{n=-\infty}^{\infty} x[n] e^{-jkn} \quad \textbf{(2.3)}$$

Biến đổi Fourier là phép biến đổi đối xứng, tức một thông tin được biến đổi Fourier từ miền thời gian sang miền tần số, có thể biến đổi Fourier ngược để khôi phục thông tin từ miền tần số lại về miền thời gian. Dưới đây là minh hoạ cho sóng vuông được phân giải thành các sóng Sin. Có thể thấy với giá trị n càng cao, độ chính xác càng lớn.

Phép biến đổi Fourier thường dùng cho phân tích các tín hiệu audio. Tuy nhiên, nó có hạn chế là ta không thể biết được tại một thời điểm sẽ xuất hiện những thành phần tần số nào. Để khắc phục nhược điểm này, các nhà khoa học sử dụng biến đổi Fourier thời gian ngắn STFT (Short time Fourier transform). Theo đó, tín hiệu được chia thành các khoảng nhỏ và được biến đổi Fourier trong từng khoảng đó.

2.2.2 Biến đổi Fourier thời gian ngắn (STFT)

Nguyên tắc của phương pháp này là phân chia tín hiệu ra thành từng đoạn đủ nhỏ sao cho có thể xem tín hiệu trong mỗi đoạn là tín hiệu ổn định, sau đó, thực hiện biến đổi Fourier trên từng đoạn tín hiệu này.

2.2.3 Phương pháp hệ số biểu diễn phổ của phổ (MFCC)

MFCC (Mel Frequency Cepstral Coefficients) là các hệ số biểu diễn phổ của phổ (spectrum-of-a-spectrum) của đoạn âm thanh. Kỹ thuật này dựa trên việc thực hiện biến đổi để chuyển dữ liệu âm thanh đầu vào (đã được biến đổi Fourier cho phổ) về thang đo tần số Mel, một thang đo diễn tả tốt hơn sự nhạy cảm của tai người đối với âm thanh.

Mạch tăng cường

Do các âm ở tần số thấp có mức năng lượng cao, các âm ở tần số cao lại có mức năng lượng khá thấp. Trong khi đó, các tần số cao này vẫn chứa nhiều thông tin về âm vị.. Do đó, nhấn mạnh trước được sử dụng để tăng năng lượng từ thấp đến cao, được thể hiện trong công thức 2.4

$$\tilde{x}(n) = x(n) - \alpha x(n-1) \quad \textbf{(2.4)}$$

Trong đó x(n) là tín hiệu và, n là số lượng mẫu lấy, và α là giá trị trong khoảng từ 0.9 tới 1.0.

Khung

Khung được sử dụng để chia $\tilde{x}(n)$ thành N thời gian của khung với các khung liền kề được phân tách bằng dịch chuyển khung P. Giả định rằng tồn tại một thuộc tính tín hiệu không đổi trong mỗi khung, tuy nhiên, việc phân chia tín hiệu đột ngột (ở cả hai đầu) bằng cách tạo khung dẫn đến mất thông tin hoặc mất đặc trưng. Dựa trên thời gian đo N, phạm vi từ 10 đến 30ms và thời gian trùng khớp < 0,5. Mỗi khung được ước lượng giá trị như sau: Công thức (2.5) là tính số lượng của khung trong tín hiệu. công thức (2.6) thể hiện giá trị ước lượng của khung f

$$\eta = \frac{p + [\tau - N]}{p} \quad \textbf{(2.5)}$$

$$\tilde{f}_j(n) = \tilde{x}(p_j + n) \quad \textbf{(2.6)}$$

Trong đó có $0 \leq n \leq N-1, 0 \leq j \leq \eta.\eta$ là số lượng của khung trong tín hiệu, τ là tổng số mẫu của tín hiệu

Cửa sổ Hamming

Cửa sổ Hamming được sử dụng để tránh quá trình mất thông tin có thể xảy ra trong quá trình đóng khung. Hơn nữa, nó được sử dụng để ngăn chặn sự cắt giảm liên tục khung hình ở cả hai đầu của tín hiệu (âm thanh ngáy). Để thực hiện cửa sổ trên tín hiệu, các khung được thực hiện bởi cửa sổ hamming theo công thức (2.7) như sau

$$f_j = \omega(n) \times f_j(n), 0 \le n \le N-1 \quad \textbf{(2.7)}$$

$$\omega(n) = \left[-\beta\cos\left(\frac{2\pi n}{N-1}\right) - (\beta - 1\right] \quad \textbf{(2.8)}$$

$$0 \le n \le N-1$$

Giá trị của β được đặt là 0.46

Biến đổi Fourier nhanh

Biến đổi Fourier nhanh sử dụng tín hiệu liên tục và định kỳ trong một khung và chuyển đổi từng tín hiệu trong miền thời gian sang miền tần số

Biến đổi Fourier nhanh (FFT) được thực hiện để chuyển đổi mỗi khung với N mẫu từ miền thời gian sang miền tần số. Tín hiệu gốc cần được thực hiện biến đổi Fourier qua bộ lọc thông dải để xử lý độ lệch tần số Mel. Biến đổi Fourier chuẩn không được sử dụng do tín hiệu âm thanh không xác định trên toàn miền thời gian. Thông thường hay sử dụng biến đổi DFT.

Mel filter DCT

Thang tần số Mel được định nghĩa như sau với giá trị của f được lấy từ công thức (2.7) ta được giá trị của tần số Mel trong công thức (2.9)

$$Mel(f) = 2595\log_{10}\left(\frac{f}{700} + 1\right) \quad \textbf{(2.9)}$$

2.3 Sau khi tín hiệu âm thanh được biểu diễn phổ phổ của phổ âm thanh thông qua MFCC thì được biểu diễn như trong hình 2.6. Trên hình 2.6 có thể thấy được những thời gian ngáy thì có một đường thẳng kéo dài từ dưới đi lên trên.

Mô hình học nông

2.3.1 Trích đặc trưng của âm thanh

Trích chọn đặc trưng bao gồm hai phần: tách/trích xuất đặc trưng (feature extraction) và lựa chọn đặc trưng (feature selection). Trích chọn đặc trưng nhằm rút gọn các tín hiệu

thành các đặc trưng để phân biệt các hoạt động đang có và sau đó được sử dụng làm dữ liệu đầu vào cho bước phân lớp. Tùy thuộc vào từng hệ thống cụ thể mà lựa chọn đặc trưng có thể được thực hiện hoặc không.

Các đặc trưng có thể được trích xuất tự động hoặc dựa trên tri thức chuyên gia. Tập các đặc trưng có được từ dữ liệu được gọi là không gian đặc trưng. Nói chung, khi các hoạt động được phân tách càng rõ ràng trong không gian đặc trưng thì hiệu suất nhận dạng của hệ thống càng cao.

2.3.2 Mô hình học máy SVM

Mô hình học máy SVM là mô hình kinh điển trong bài toán phân loại. Tư tưởng của SVM là định nghĩa ra một siêu mặt phẳng có thể phân tách các tập dữ liệu cần phân loại sao cho khoảng cách (margin) từ siêu mặt phẳng đến các tập cần phân loại là tương đương nhau và lớn nhất

Trong không gian Euclid có cách tính khoảng cách từ một điểm có tọa độ (x_0, y_0) tới đường thẳng có phương trình $w_1x + w_2y + b = 0$ được tính bằng:

$$h = \frac{|w_1x_0 + w_2y_0 + b|}{\sqrt{w_1^2 + w_2^2}} \quad \textbf{(2.10)}$$

Trong không gian ba chiều khoảng cách từ một điểm có tọa độ (x_0, y_0, z_0) tới một mặt phẳng có phương trình $w_1x + w_2y + w_3z + b = 0$ được tính bằng

$$h = \frac{|w_1x_0 + w_2y_0 + w_3\,z_0 + b|}{\sqrt{w_1^2 + w_2^2 + w_3^2}} \quad \textbf{(2.11)}$$

Nhận thấy nếu bỏ dấu giá trị tuyệt đối thì có thể xác định được điểm đang xét nằm phía nào của đường thẳng hay mặt phẳng. Từ đó, có thể tổng quát cho rằng nếu biểu thức bỏ dấu giá trị tuyệt đối thì những điểm nào cùng mang dấu với nhau thì nằm cùng phía với nhau và có được công thức tính khoảng cách trong không gian có n số chiều mà trong đó có khoảng cách được tính bằng:

$$h = \frac{|w^Tx_0 + b|}{\sqrt{\sum_{i=1}^{n} w_i^2}} \quad \textbf{(2.12)}$$

Giả sử với xét các cặp dữ liệu đào tạo là $(x_1, y_1), (x_2, y_2)$ ..... (x_n, y_n) tượng trưng cho dữ liệu đầu vào của một điểm dữ liệu

Bài toán SVM trở thành đi tìm w và b sao cho khoảng cách này đạt giá trị lớn nhất.

Đối với bài toán phân lớp mà có số lớp n > 2 thì có thể sử dụng bằng cách chuyển bài toán phân lớp nhị phân giữa 1 lớp và (n-1) lớp còn lại. Tức là sẽ phải thực hiện n lần giữa phân lớp giữa lớp thứ i và (n-i) lớp còn lại.

Khoảng cách từ chiều tới mặt

2.3.3 Đánh giá mô hình học máy SVM

Mô hình CNN cho phát hiện tiếng ngáy

Mạng neural được lấy cảm hứng từ cấu tạo về não của con người, khi mà từ thông tin
2.4 tiếp nhận và được xử lý khi đi qua các neural và khi đến cuối của neural thì thông tin đã được xử lý xong hoàn toàn. Mô hình mạng neural được mô tả thông qua hình sau:

Lớp đầu tiên là lớp input, các layer ở giữa được gọi là các lớp ẩn, lớp cuối cùng là lớp đầu ra. Các hình tròn được gọi là node

2.4.1 Giới thiệu về kiến trúc mạng CNN

Mạng neural tích chập (CNN) là một trong những mô hình mạng neural Deep Learning tiên tiến giúp cho việc xây dựng được những hệ thống thông minh với độ chính xác cao. Thường được sử dụng trong tín hiệu số (Signal Processing), phân lớp ảnh (Image Classification).

2.4.2 Tích chập trong mạng neural

Tích chập được sử dụng đầu tiên trong xử lý tín hiệu số (Signal processing). Nhờ vào nguyên lý biến đổi thông tin, các nhà khoa học đã áp dụng kĩ thuật này vào xử lý ảnh và video số.

CNNs gồm một vài layer của convolution kết hợp với các hàm kích hoạt phi tuyến (nonlinear activation function) như ReLU hay tanh để tạo ra thông tin trừu tượng hơn (abstract/higher-level) cho các layer tiếp theo.

Có ba tầng chính để xây dựng kiến trúc cho một mạng nơron tích chập:

1. Tầng tích chập

2. Tầng gộp (pooling layer)

3. Tầng được kết nối đầy đủ (fully-connected).

Tầng kết nối đầy đủ giống như các mạng nơron thông thường, và tầng chập thực hiện tích chập nhiều lần trên tầng trước. Tầng gộp có thể làm giảm kích thước mẫu trên từng khối 2x2 của tầng trước đó. Ở các mạng nơron tích chập, kiến trúc mạng thường chồng ba tầng này để xây dựng kiến trúc đầy đủ

2.4.3 Mô hình mạng CNN trong phát hiện tiếng ngáy

CNNs có tính bất biến và tính kết hợp cục bộ (Location Invariance and Compositionality). Với cùng một đối tượng, nếu đối tượng này được chiếu theo các gốc độ khác nhau (translation, rotation, scaling) thì độ chính xác của thuật toán sẽ bị ảnh hưởng đáng kể. Pooling layer biểu hiện được tính bất biến đối với phép dịch chuyển (translation), phép quay (rotation) và phép co giãn (scaling).

2.5 Mô hình LSTM cho phát hiện tiếng ngáy

Sau quá trình tìm hiểu và tham khảo, với điều kiện thực nghiệm còn hạn chế với kiến trúc CNN, luận văn quyết định áp dụng 4 convolutional layer với các thông số sau:

	Feature maps	**Patch size**	**Pool size**
Conv layer 1	64	193x1	191x64
Conv layer 2	64	191x64	189x64
Conv layer 3	128	189x64	63x64
Conv layer 4	128	61x128	59x128

2.5.1 Giới thiệu về mạng neural hồi quy

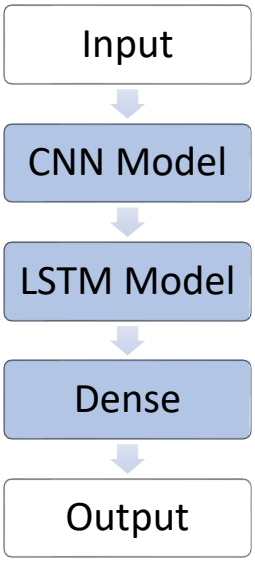
2.5.2 Hồi quy trong mạng neural và mô hình LSTM.

2.5.3 Mô hình mạng LSTM trong phát hiện tiếng ngáy

Như đã giới thiệu ở phần trên về mạng RNN, RNN có thể xử lý thông tin dạng chuỗi, như dự đoán hành dộng trong chuỗi ảnh, hay số tăng giảm giá nhà từ trong các dữ liệu trong lịch sử. RNN mang thông tin của các trạng thái trước tới các trạng thái sau, rồi ở trạng thái cuối là sự kết hợp của các trạng thái đã diễn ra để dự đoán kết quả.

Mô hình CNN-LSTM cho phát hiện tiếng ngáy

Kiến trúc này khi xác định hai mô hình con: Mô hình CNN để trích xuất đặc trưng và Mô hình LSTM để diễn giải các tính năng theo các bước thời gian. Điều này, sẽ mang lại cho mô hình tận dụng được ưu điểm của từng mô hình con[22] đối với kết quả sau khi học tập.
2.6 Sau quá trình tham khảo và nghiên cứu luận văn nhận thấy kiến trúc phát hiện tiếng ngáy sử dụng mô hình học sâu CNN-LSTM sẽ được mô tả như sau:



Hình 2.1. Kiến trúc mô hình học sâu với CNN LSTM cho nhận dạng tiếng ngáy

Trong mô hình CNN-LSTM mà luận văn sử dụng có tham khảo từ các mô hình CNN
2.7 và LSTM mà luận văn đã lựa chọn ở trong hai phần đã được trình bày ở trên.

Kết luận chương

Trong chương này đã trình bày về quá trình tìm hiểu và áp dụng mô hình học nông SVM các mô hình học sâu CNN, LSTM, CNN-LSTM. Bênh cạnh đó chương này cũng trình bày giới thiệu về thuật toán SVM, mạng neural tích chập, mạng neural hồi quy và mạng neural tích chập và hồi quy đẻ phân lớp dữ liệu.

Với những kiến thức đã tìm hiểu và trình bày tại chương, luận văn sẽ áp dụng kiến trúc mạng neural tích chập, kiến trúc mạng neural hồi quy – LSTM và so sánh với SVM.

Chương 3 sẽ tiến hành thực nghiệm dữ liệu với phương pháp đã đề xuất dựa trên các kịch bản khác nhau, sau đó sẽ đánh giá độ chính xác và đưa ra đề xuất định hướng tiếp theo

CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ

Trong chương này sẽ trình bày các vấn đề: thu thập dữ liệu tiếng ngày; thử nghiệm mô hình CNN hoặc mô hình hồi quy RNN phân tích các âm thanh qua đó có thể đánh giá được các kiến trúc học sâu trong việc phát hiện tiếng ngáy.

3.1 Thu thập dữ liệu

Sau khi thực hiện gán nhãn, các tập dữ liệu về các lớp âm thanh ngáy, và không ngáy số lượng cụ thể thu được sau quá trình gán nhãn âm thanh ngáy được mô tả tại bảng sau.

Bảng 3.1. Thống kê dữ liệu thực nghiệm

Dữ liệu âm thanh ngáy			
	Thời gian ngáy	Tổng thời gian	Tỉ lệ tiếng ngáy/ tổng thời gian
Ngáy 1	36 phút	40 phút	0.9
Ngáy 2	25 phút	30 phút	0.83
Dữ liệu Kaggle	8 phút	16 phút	1
Tổng cộng	69 phút	86 phút	

Với dữ liệu thực nghiệm như trên thì có đủ các âm thanh ngáy/ không ngáy từ những người xuất hiện tình trạng ngáy khi ngủ và thêm vào đó có thêm các dữ liệu của Kaggle về các lớp ngáy/ không ngáy được thu thập trên trang mạng chia sẻ âm thanh.

3.2 Kết quả thử nghiệm

Môi trường thử nghiệm các mô hình học sâu được tìm hiểu thông qua Google Colab hay Colaboratory notebooks. Google Colab cung cấp cho chúng ta khả năng tính toán mạnh hơn với Tesla K80 GPU, thay vì phải code và train model với máy tính, laptop cá nhân. Google Colab cũng hỗ trợ khá toàn diện các thư viện trong python, phiên bản mới nhất của tensoflow, keras, PyTorch,. Cv2 .. trong việc cài đặt các mô hình.

Để đánh giá các mô hình thì luận văn sử dụng 2 độ đo là Precision và Reall trong đó:

TP: là số âm thanh tiếng ngáy mà mô hình đoán là tiếng ngáy.

FP: là số âm thanh tiếng ngáy mà mô hình đoán là không phải tiếng ngáy.

FN: là số âm thanh không phải là tiếng ngáy mà mô hình dựa đoán là tiếng ngáy.

Precision được định nghĩa là tỉ lệ số điểm TP trong số những điểm được phân loại là chủ động của mô hình (TP+FP) với công thức (3.1) được tính như sau:

$$Precision = \frac{TP}{TP + FP} \quad \textbf{(3.1)}$$

Recall được định nghĩa là tỉ lệ số điểm TP trong số những điểm thực sự là do mô hình dự đoán ra (TP+FN) với công thức (3.2) được tính như sau:

$$Recall = \frac{TP}{TP + FN} \quad \textbf{(3.2)}$$

Ngoài ra, hai độ đo trên không phải lúc nào cũng tăng giảm tương ứng với nhau, có trường hợp Recall cao còn Precision thấp và ngược lại, để cho đánh giá tổng quát hơn thì F-measure là trung bình điều hòa của 2 độ đó trên với hệ số 0.5 (tầm quan trọng của 2 hệ số ngang nhau) được tính với công thức (3.3) như sau:

$$F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = 2\frac{precision\,.recall}{precision + recall} \quad \textbf{(3.3)}$$

3.2.1 Kết quả học nông SVM

Với mô hình học nông SVM với các tham số được xác định khi chạy thực nghiệm là tham số C và gamma là hai tham số rất quan trọng trong việc huấn luyện SVM

Kết quả thực nghiệm của SVM thu được:

Bảng 3.2. Kết quả của phương pháp học nông SVM

SVM			
Acc (%)	0.724637681		
	Presion	Recall	F1
Tiếng ngáy	0.71559633	0.75	0.732394366
Không ngáy	0.734693878	0.699029126	0.71641791

Dựa trên bảng kết qua của mô hình SVM ta có thể nhận thấy trong SVM thì tỉ lệ phát hiện tiếng ngáy/ không ngáy gần như bằng nhau. Tỉ lệ chính xác khoảng gần 0.72.

3.2.2 Kết quả của phương pháp CNN

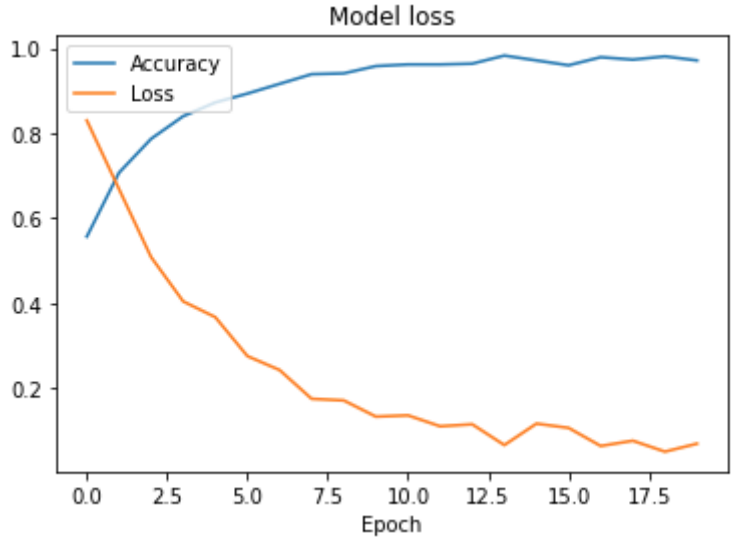
Mô hình học sâu với mạng mô hình CNN đã được lựa chọn trong phần 2.4 mô hình mạng CNN trong phát hiện tiếng ngáy.

Kết quả thực nghiệm của CNN được thể hiện như sau.

Bảng 3.3. Kết quả của mô hình CNN

CNN			
Acc	0.768115942		
	Presion	Recall	F1
Tiếng ngáy	0.689189189	0.980769231	0.80952381
Không ngáy	0.966101695	0.80952381	0.703703704

Mô hình học CNN đánh giá mô hình



Hình 3.1. Thực nghiệm độ chính xác của mô hình CNN qua số lần epoch

Thời gian mà mô hình đào tạo hết tổng cộng 17 giây, kiểm tra độ chính xác đạt, 0.968 và đạt điểm 0.12452.

Dựa trên bảng kết qua của mô hình học sâu CNN, kết quả thực nghiệm, kết quả đo đánh giá mô hình, kết quả huấn luyện mô hình ta có thể nhận thấy mô hình mạng CNN có tỉ lệ chính xác vượt trội hơn so với phương pháp học sâu với độ chính xác lên tới 0.76. Các độ đo về độ chính xác khi phát hiện âm thanh ngáy là 0.689 nhỏ hơn nhiều so với việc phát hiện ra âm thanh đó không phải tiếng ngáy là 0.9661.

3.2.3 Kết quả của phương pháp LSTM

Mô hình học sâu với mạng mô hình LSTM đã được lựa chọn trong phần 2.5 mô hình mạng LSTM trong phát hiện tiếng ngáy.

Kết quả thực nghiệm của LSTM được thể hiện như sau:

Bảng 3.4. Kết quả của mô hình LSTM

LSTM			
Acc (%)	0.753623188		
	Presion	**Recall**	**F1**
Tiếng ngáy	0.702290076	0.884615385	0.782978723
Không ngáy	0.842105263	0.621359223	0.715083799

Mô hình học LSTM đánh giá mô hình

Hình 3.2. Thực nghiệm độ chính xác mô hình LSTM qua số lần epoch

Thời gian mà mô hình đào tạo hết tổng cộng 205 giây, kiểm tra độ chính xác đạt, 0.7635 và đạt điểm : 0.466

Dựa trên bảng kết quả của mô hình học sâu LSTM, kết quả thực nghiệm, kết quả đo đánh giá mô hình, kết quả huấn luyện mô hình ta có thể nhận thấy mô hình mạng LSTM có

tỉ lệ chính xác vượt trội hơn so với phương pháp học sâu với độ chính xác lên tới 0.6328. Các độ đo về độ chính xác khi phát hiện âm thanh ngáy là 0.7022 nhỏ hơn nhiều so với việc phát hiện ra âm thanh đó không phải tiếng ngáy là 0.8421.

3.2.4 Kết quả của phương pháp CNN-LSTM

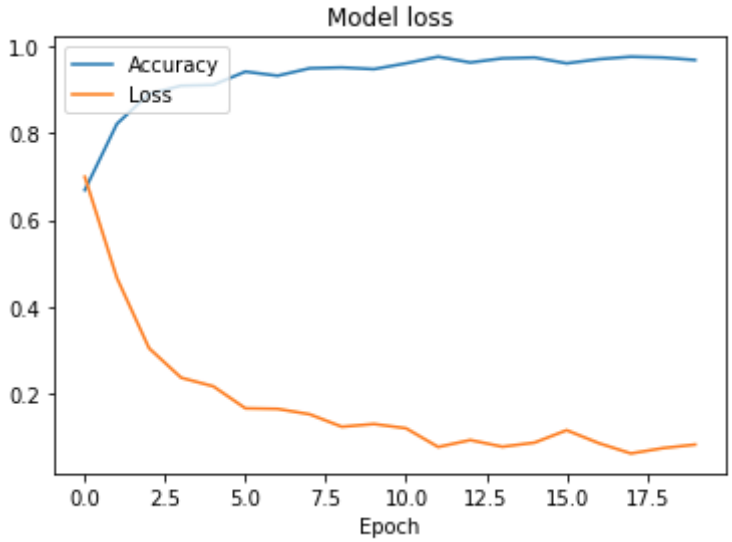
Mô hình học sâu với mạng mô hình CNN-LSTM đã được lựa chọn trong phần 2.6 mô hình mạng CNN-LSTM trong phát hiện tiếng ngáy.

Kết quả thực nghiệm của CNN-LSTM được thể hiện như sau:

Bảng 3.5. Kết quả của mô hình CNN-LSTM

CNN-LSTM			
Acc (%)	0.917874396		
	Presion	Recall	F1
Tiếng ngáy	0.871794872	0.980769231	0.923076923
Không ngáy	0.977777778	0.854368932	0.911917098

Mô hình học CNN-LSTM đánh giá mô hình



Hình 3.3. Thực nghiệm độ chính xác mô hình CNN-LSTM qua số lần epoch

Dựa trên bảng kết quả của mô hình CNN-LSTM ta có thể nhận thấy thời gian mà mô hình đào tạo hết tổng cộng 52 giây, kiểm tra độ chính xác đạt, 0.9772 và đạt điểm: 0.0489

Phân tích và đánh giá

Dựa vào kết quả của các đánh giá trên thì nhận thấy được các mạng học sâu đều cho kết quả phát hiện âm thanh ngáy tốt hơn nhiều so với mạng học nông mà cụ thể ở đây là SVM.

3.3 Độ chính xác, đánh giá qua các độ đo được nên ra trong phần kết quả thử nghiệm gồm Pression, Recall, F1-score thì có thể thấy được các phương pháp có kết quả được xếp từ thấp lên cao như sau:

Bảng 3.6. Độ chính xác của các mô hình

Mô hình	Độ chính xác
Mô hình học nông SVM	0.724637681
Mô hình mạng CNN	0.768115942
Mô hình mạng LSTM	0.753623188
Mô hình mạng CNN-LSTM	**0.917874396**

Kết quả của các mô hình được thực nghiệm trong luận văn có thể nhận thấy rằng, mô hình mạng học sâu có kết quả tốt hơn hẳn so với mô hình mạng học nông như SVM, kết quả của mô hình mạng học sâu CNN-LSTM cho ra là kết quả tốt nhất, nhờ có sự kết hợp giữa ưu điểm của mô hình CNN và LSTM điều này có sự tương đồng với các nghiên cứu về phân lớp âm thanh có liên quan.

3.4 Kết luận chương

Trong chương này sẽ trình bày các vấn đề: thu thập dữ liệu tiếng ngày; thử nghiệm mô hình CNN hoặc mô hình hồi quy RNN phân tích các âm thanh qua đó có thể đánh giá được các kiến trúc học sâu trong việc phát hiện tiếng ngáy. Sau quá trình thử nghiệm với tập dữ liệu và cài đặt với các mô hình, phương pháp học máy khác nhau thì thu được kết quả tốt nhất thuộc về mô hình mạng học sâu kết hợp CNN-LSTM với kết quả tốt hơn nhiều so với các phương pháp còn lại.

KẾT LUẬN

Nghiên cứu về phát hiện âm thanh nói chung, về bài toán phát hiện tiếng ngáy dựa trên học sâu nói riêng với tôi là công nghệ mới, thời gian nghiên cứu còn ngắn nên vẫn còn nhiều vấn đề chưa thực sự nắm bắt tốt. Tuy nhiên, qua quá trình nghiên cứu, luận văn đã tìm hiểu sâu về các giai đoạn từ tiền xử lý dữ liệu đến các phướng pháp xử lý âm thanh, các phương pháp học máy mà đặc biệt là các mô hình học sâu với mạng neural, phương pháp học sâu để xây dựng mô hình phân lớp dữ liệu (mô hình hình CNN, LSTM, CNN-LSTM) và so sánh với mô hình học nông SVM.

Sử dụng các mạng neural nói chung hay CNN, LSTM và CNN-LSTM nói riêng trong học sâu là một hướng đi có kỹ thuật và hiệu quả trong các bài toán xử lý chuỗi và hiện đang trở thành xu thế của các nhà nghiên cứu.

Trong tương lai, luận văn có thể được phát triển nghiên cứu các mô hình khác, giải quyết các bài toán khác về theo dõi, nhân diện âm thanh, hoặc phát triển thành những ứng dụng y tế mà có thể hỗ trợ cho nhiều người trong cộng đồng..