Ensemble learning for Vietnamese lyrics-and-audio-based music genre classification

Nguyễn Thị Thu Hà[1], Bùi Anh Khôi[2],
Đỗ Trọng Hợp[3], Lưu Thanh Sơn[4], Nguyễn Thành Luân[5]

Trường Đại học Công Nghệ Thông Tin
Đại học Quốc gia Thành phố Hồ Chí Minh
{19521456[1],19520649[2]}@gm.uit.edu.vn
{hoptd[3],sonlt[4],luannt[5]}@uit.edu.vn

Tóm tắt nội dung Trong cuộc sống hiện nay, âm nhạc đang chiếm vị thế lớn trong đời sống tinh thần của mỗi người. Nhờ đó mà âm nhạc ngày càng phát triển về mặt số lượng và chất lượng. Bên cạnh đó, sự xuất hiện các nền tảng nghe nhạc trực tuyến kéo theo những nhu cầu như quản lý thông minh, tìm kiếm và hệ khuyến nghị,.. bài hát là rất cần thiết. Phân loại bài hát theo thể loại sẽ là nền tảng để phát triển những nhu cầu trên. Do đó, chúng tôi đã tiến hành xây dựng mô hình phân loại thể loại bài hát tiếng Việt trên bộ dữ liệu VMusic do chúng tôi tự thu thập. Chúng tôi tiến hành xử lý, thực nghiệm và đánh giá trên các mô hình kết hợp giữa lời bài hát và giai điệu. Kết quả thu được cao nhất là mô hình Bi-LSTM + Resnet18 với F1-Score là 0.552. Sau đó, chúng tôi tiến hành thống kê, so sánh kết quả trên các mô hình để đưa ra kết luận và chọn ra mô hình có hiệu quả tốt nhất và đề ra phương hướng phát triển mô hình trong tương lai.

Keywords: Thể loại bài hát · Phân loại · Phương pháp học sâu · Lời bài hát · Giai điệu

1 Giới thiệu

Từ cổ chí kim, con người đã có nhiều cách thức để thoả mãn những nhu cầu tri giác của bản thân, trong đó âm nhạc là một phần không thể thiếu. Với sự bùng nổ của mạng internet kéo theo đó là sự nở rộ của các nền tảng nghe nhạc trực tuyến. Số lượng bài hát khổng lồ khiến cho việc phân loại và quản lý chúng là vô cùng khó khăn. Việc xây dựng hệ thống phân loại lời bài hát tự động giúp cho người dùng dễ dàng trải nghiệm và tìm kiếm hơn. Bên cạnh đó, việc phân loại bài hát theo thể loại là cơ sở cho việc lưu trữ thông minh và xây dựng hệ khuyến nghị bài hát cho các nền tảng đó. Từ đó, sự xuất hiện của một hệ thống phân loại bài hát tự động dựa trên lời bài hát là rất hợp lý và thiết thực.

Trong bài báo này, chúng tôi trình bày phương pháp xây dựng mô hình tự động dự đoán thể loại bài hát dựa trên hai thành phần tạo nên bài hát là: lời bài hát và giai điệu. Chúng tôi sử dụng bộ dữ liệu VMusic do chúng tôi tự thu thập gồm 6 thể loại đã được gán nhãn: nhạc trẻ, trữ tình, thiếu nhi, cách mạng, rock

việt, rap việt. Chúng tôi sử dụng thuật toán LSTM và BiLSTM dành cho lời bài hát và Resnet18, Resnet34 cho phần giai điệu. Sau đó, tiến hành thực nghiệm và so sánh kết quả giữa các mô hình kết hợp giữa lời bài hát và giai điệu.

Trong mục 2, chúng tôi sẽ trình bày một số công trình nghiên cứu liên quan. Tiếp theo ở mục 3, chúng tôi trình bày các thông tin cơ bản của bộ dữ liệu. Trong mục 4, các giải pháp, mô hình được chúng tôi trình bày và đồng thời, kết quả thử nghiệm sẽ được đánh giá, phân tích ở mục 5. Cuối cùng, mục 6 sẽ là kết luận và hướng phát triển trong tương lai cho các bài toán phân loại nói chung và các bài toán phân loại bài hát nói riêng.

2 Công trình liên quan

2.1 Lyrics-based music genre classification using a hierarchical attention network

Alexandros Tsaptsinos đã nghiên cứu sử dụng các thuật toán cơ sở và HAN để phân loại thể loại bài hát trên hai bộ dữ liệu: 117 thể loại và 20 thể loại. Dữ liệu lấy được thông qua một thỏa thuận nghiên cứu đã ký với LyricFind [8]. Trong nghiên cứu này, tác giả đã thực nghiệm và so sánh mô hình HAN với các mô hình cơ sở: Majority classifier, Logistic regression, Long Short-Term Memory, Hierarchical network trên cả hai tập dữ liệu trên. Mô hình HAN được thực hiện ở hai cấp độ là dòng và đoạn. HAN-L chạy dữ liệu theo dòng và HAN-S chạy dữ liệu theo đoạn. Kết quả thu được các mô hình dựa trên mạng thần kinh thì có kết quả tốt hơn các mô hình đơn giản. Ở bộ dữ liệu 20 thể loại, mô hình LSTM hoạt động tốt hơn HAN và cho ra kết quả 49.77%. Đối với bộ dữ liệu 117 thể loại, cả hai mô hình HAN đều có kết quả tốt, nhất là HAN-L.

2.2 Music Genre Classification using Transfer Learning on log-based MEL Spectrogram

Trong bài báo này, nhóm tác giả đã sử dụng bộ dữ liệu GTZAN gồm 100 bài hát với 10 thể loại bài hát: classical, blues, country, disco, hip-hop, jazz, metal, pop, reggae and rock [6]. Để đưa vào mô hình, giai điệu bài hát sẽ được chuyển thành dạng ảnh quang phổ với sampling rate là 44100Hz và ảnh có kích thước đồng nhất là 224x224. Vì số lượng bài hát trong bộ dữ liệu ít nên nhóm tác giả đã tăng cường dữ liệu bằng cách cắt từ bài hát 30s thành các đoạn nhỏ 3s. Sau đi thực nghiệm bằng các thuật toán Resnet34, Resnet50, VGG16 và AlexNet trên dữ liệu đã tăng cường thì kết quả cho thấy mô hình Resnet34 có Accuracy cao nhất là 79%.

3 Bộ dữ liệu

3.1 Thu thập dữ liệu

Bộ dữ liệu VMusic được chúng tôi thu thập theo từng thể loại có sẵn từ trang Nhac.vn. Bộ dữ liệu có 12293 lời bài hát kèm theo audio được gán nhãn theo thể loại của bài hát đó, mô tả sơ lược trong Bảng 1

lyric	genre	ID
Chiều biên giới em ơi Có nơi nào xanh hơn...	cach-mang	99061858
Lặng thầm ngồi đếm bước chân em về qua căn phòng...	nhac-tre	20606639
Có con chim vành khuyên nhỏ Dáng trong thật ngoan ngoãn quá...	thieu-nhi	17533146

Bảng 1: Một số ví dụ cho bộ dữ diệu VMusic.

Trong bài toán này chúng tôi sử dụng thuộc tính đầu vào là lời bài hát và giai điệu để xác định thể loại của chúng, bao gồm 6 loại sau: nhạc trẻ, trữ tình, rap việt , cách mạng, thiếu nhi, rock việt. Tất cả dữ liệu đã được gán nhãn sẵn trên các trang web mà chúng tôi thu thập.

Bộ dữ liệu sẽ được chia thành trainning set, validation set, testing set với tỉ lệ 72:8:20. Chi tiết số lượng nhãn mỗi tập cũng như là nhãn sau khi chuyển thành dạng số được mô tả cụ thể trong Bảng 2.

Nhãn	Nhãn số	Train	Val	Test	Tổng
nhạc trẻ	0	5,479	609	1,522	7,610
trữ tình	1	1,940	216	539	2,695
rap việt	2	658	73	183	914
cách mạng	3	452	50	126	628
thiếu nhi	4	261	29	73	363
rock việt	5	60	7	16	83

Bảng 2: Thống kê số lượng nhãn trên các tập.

3.2 Thách thức bộ dữ liệu

3.2.1 Lời bài hát Mỗi lời bài hát đều có cấu trúc khác nhau. Việc xác định cấu trúc bài hát để loại bỏ những cụm từ đánh dấu thành phần của lời bài hát như '[ĐK:]', '[Hook:]', '[Điệp khúc]', '[Verse 2]', '[amj7]', 'x4',...

Hiện nay, nhạc Việt đang vay mượn quá nhiều từ ngữ nước ngoài trong lời bài hát. Không chỉ sử dụng một số từ thông dụng mà một số bài hát có tiếng nước ngoài chiếm hơn 50% lời bài hát. Điều này gây khó khăn khi sử dụng các phương pháp và tài nguyên có sẵn đặc thù dành cho tiếng Việt.

Lời bài hát thu thập được xuất hiện nhiều từ đặc biệt như: sai chính tả, teencode, sai font chữ,... Hiện tại, chúng tôi đã tìm ra được hơn 4000 nghìn từ bị viết sai chính tả, teencode,... Và không biết là con số đó đã bao quát hết các trường hợp của các từ này hay chưa. Việc tìm ra tất cả các trường hợp và thay thế các từ đặc biệt trên là một thách thức rất lớn mà chúng tôi gặp phải trong bộ dữ liệu này.

Bên cạnh đó, các thể loại có sự chênh lệch lớn về số lượng dữ liệu. Thể loại nhạc trẻ có 7,610 lời bài hát trong khi đó thể loại thiếu nhi và rock việt lần lượt

có 363, 83 lời bài hát. Điều này khiến cho việc xây dựng và dự đoán của mô hình trở nên khó khăn đối với cái thể loại có ít dữ liệu.

Trên đây là một vài những thách thức, khó khăn mà chúng tôi nhận thấy ở thời điểm hiện tại mà đề tài sẽ phải đối mặt. Trong quá trình thực hiện đề tài, chúng tôi sẽ cố gắng giải quyết tốt các vấn đề đã được nêu như trên.

3.2.2 Âm thanh Với hơn 12,000 bài hát trong bộ dữ liệu, dung lượng lưu trữ là khá lớn với 49.7GB gây khó khăn về độ phức tạp cũng như thời gian trong quá trình xử lý, thu thập và lưu trữ. Đồng thời, thời lượng của mỗi bài hát là không giống nhau gây nên sự không đồng đều, tăng thêm các bước xử lý.https://www.overleaf.com/project/61ecc1a60080593b9aeffc1f

4 Phương pháp tiếp cận

4.1 Tiền xử lý dữ liệu

Việc tiền xử lý bộ dữ liệu là một bước rất cần thiết, có ảnh hưởng rất lớn quá trình và kết quả của mô hình. Nên cần đưa dữ liệu về một kiểu nhất quán để giảm chiều dữ liệu và có được kết quả tốt nhất cho mô hình.

4.1.1 Lời bài hát Đầu tiên, chúng tôi sẽ loại bỏ những bài hát có phần lớn các từ không phải ngôn ngữ tiếng Việt hoặc có nội dung không phải là lời bài hát. Sau đó, lời bài hát sẽ đưa về dạng "LowerCase" và loại bỏ các kí tự đặc biệt. Tiếp theo, chúng tôi sẽ xác định các từ đánh dấu cấu trúc của một bài hát ('[ĐK:]', '[Bray:]',...) và loại bỏ chúng.

Trong một số lời bài hát, nhất là trong thể loại rap-viet thường có những từ tượng thanh như: 'eyy', 'zozo', 'skrrrttt',... chúng tôi quyết định giữ lại để giữ được nét đặc trưng của thể loại và đưa các từ về chung một dạng. Ví dụ: 'yeh', 'yeahh', 'yehh', 'ye',... chuyển thành 'yeah'.

Để giải quyết vấn đề xuất hiện các từ đặc biệt như chúng tôi đã nêu ở phần trên, chúng tôi đã xây dựng một từ điển có 1203 trường hợp từ bộ dữ liệu này. Từ điển sẽ có từ khóa là các từ đặc biệt và giá trị là từ đã được sửa theo đúng định dạng và chính tả. Cuối cùng, văn bản lời bài hát sẽ được tách từ bằng công cụ VnCoreNLP [9]. Một số trích đoạn trước và sau khi xử lý được chúng tôi mô tả trong ví dụ sau:

"*[Rap]:*
Tôi lạc quan giữa đám đông
Nhưng khi 1 mình thì lại không...",
"*Anh sẽ vì em làm thơ tình ái*
Anh sẽ gom meiiii
Nhưng trên đó làm gì có lâu đài?..."

Sau khi xử lý sẽ trở thành như sau:

"*tôi lạc_ quan giữa đám đông nhưng khi 1 mình thì lại không...*",
"*anh sẽ vì em làm thơ tình_ ái anh sẽ gom mây nhưng trên đó làm gì có lâu_ đài...*"

4.1.2 Âm thanh Tệp tin âm thanh không thể được sử dụng trực tiếp để làm đầu vào của các mô hình học sâu mà phải qua một quá trình biến đổi thành dạng thích hợp. Trong các bài toán phân loại âm thanh hiện tại, các tệp âm thanh sẽ được chuyển đổi thành dạng spectrogram – dạng ảnh quang phổ thể hiện tần số của tín hiệu âm thanh theo thời gian. Trong bài nghiên cứu này, chúng tôi sẽ sử dụng Mel, Harmonic và Percusive spectrogram làm ba đặc trưng được trích xuất từ một tệp âm thanh, chi tiết về những đặc trưng này sẽ được mô tả cụ thể ở phần sau.

Trong quá trình biến đổi tệp âm thanh thành dạng spectrogram gray scale, chúng tôi sử dụng sample rate bằng 44100 Hz vì đây là tiêu chuẩn cho các phương pháp thực nghiệm. Để giảm thiểu độ phức tạp tính toán cùng như thời gian xử lý, chúng tôi đã tiến hành huấn luyện một mạng CNN đơn giản trên tập validation (trích 10% làm tập kiểm thử) với hai kích thước đầu vào khác nhau và chỉ sử dụng một đặc trưng là Mel-spectrogram (kết quả trong Bảng 3):

- Mô hình 1: biến đổi toàn bộ tệp âm thanh thành grayscale Mel-spectrogram 128×1024.
- Mô hình 2: chỉ sử dụng 20s đầu tiên và biến đổi thành dạng grayscale Mel-spectrogram 128×128.

Mô hình	**Accuracy**	**Time**
Mô hình 1	0.525	1547.275
Mô hình 2	0.566	233.833

Bảng 3: Kết quả accuracy trên tập kiểm thử và thời gian xử lý của thí nghiệm

Dựa vào Bảng 3 ta có thể thấy kết quả thu được khi biến đổi 20s đầu tiên ở dạng 128x128 là tương đồng so với khi sử dụng toàn bộ tệp âm thanh ở dạng 128x1024 trong khi thời gian xử lý và huấn luyện cùng như là độ phức tạp của Mô hình 2 là hoàn toàn tối ưu hơn so với Mô hình 1. Vì vậy, các đặc trưng sẽ được biến đổi thành dạng gray-scale spectrogram 128x128.

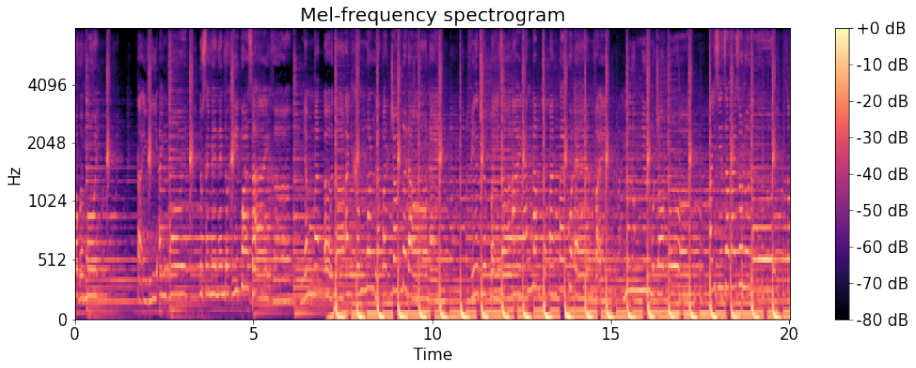
Cuối quá trình, bằng sự trợ giúp của thư viện Librosa[1], với mỗi tệp tin âm thanh chúng tôi sẽ thu được 3 ảnh spectrogram tương ứng với ba đặc trưng kể trên và kết hợp chúng thành dạng ảnh 128x128 với 3 kênh tương ứng.

4.2 Mel Spectrogram

Con người giỏi hơn nhiều trong việc nhận biết những thay đổi nhỏ về cao độ ở tần số thấp hơn là ở tần số cao. Việc kết áp dụng thang đo mel scale để trích

[1] https://librosa.org/

xuất spectrogram này làm cho các thông tin âm thanh trở nên phù hợp hơn với những gì con người nghe thấy [7]. Một ví dụ của Mel spectrogram được mô tả hở Hình 1.



Hình 1: Mel-frequency Spectrogram

Tần số ban đầu sẽ được biến đổi theo thang đo mel dựa vào công thức sau với m là Mels và f là tần số Hertz:

$$m = 2595 \log_{10}(1 + \frac{f}{700}) \tag{1}$$

4.3 Harmonic/Percussive Spectrogram

Nói chung, âm thanh có thể được phân thành hai loại [2]. Loại thứ nhất là âm thanh hài hòa (Harmonic) là âm thanh mà chúng ta cảm nhận được cao độ và là thứ khiến chúng ta nghe thấy giai điệu và hợp âm. Loại thứ hai, âm thanh bộ gõ (Percussive) giống như tiếng ồn và thường bắt nguồn từ việc khởi động nhạc cụ như đánh trên trống hoặc từ các phụ âm trong lời nói. Dưới dạng ảnh spectrogram, âm thanh Harmonic thường có dạng những được ngang còn âm thanh Percussive có dạng những đường dọc.

4.4 Tăng cường dữ liệu

4.4.1 Sử dụng pretrained word embbeding(WEMOTE) cho dữ liệu lyric Word embedding bản chất là một không gian vector được sử dụng để biểu diễn dữ liệu có khả năng mô tả được các mối liên hệ, sự tương đồng về mặt ngữ nghĩa, ngữ cảnh của dữ liệu. Trong không gian này, các từ có cùng văn cảnh hoặc ngữ nghĩa sẽ có vị trí gần nhau. Ứng dụng đặc điểm này, chúng tôi đã sử dụng một phương pháp tăng cường dữ liệu sử dụng pretrained embbeding [1].

Đầu tiên, từ không gian vector từ đã được huấn luyện trước, chúng tôi sẽ rút trích được từ điển của các từ tương đồng về mặt văn cảnh và ngữ nghĩa như ví dụ sau:

mỉa_mai : [‘chế_giễu’, ‘chê_bai’, ‘chế_nhạo’, ‘miệt_thị’]
hoan_lạc : [‘khoái_lạc’, ‘lạc_thú’, ‘ân_ái’, ‘giao_hoan’]]

Tiếp theo, với mỗi văn bản thuộc lớp thiểu số, ta gán cho mỗi từ một xác suất với khả năng từ đó sẽ được thay thế bằng một từ đồng nghĩa đã tìm được ở bước trên nếu tồn tại. Cụ thể với xác suất 0.5, sẽ có 50% từ đó sẽ được thay thế bằng từ đồng nghĩa, khi tiến hành trên toàn bộ văn bản sẽ sinh ra như ví dụ sau:

Ta có lời bài hát ban đầu:

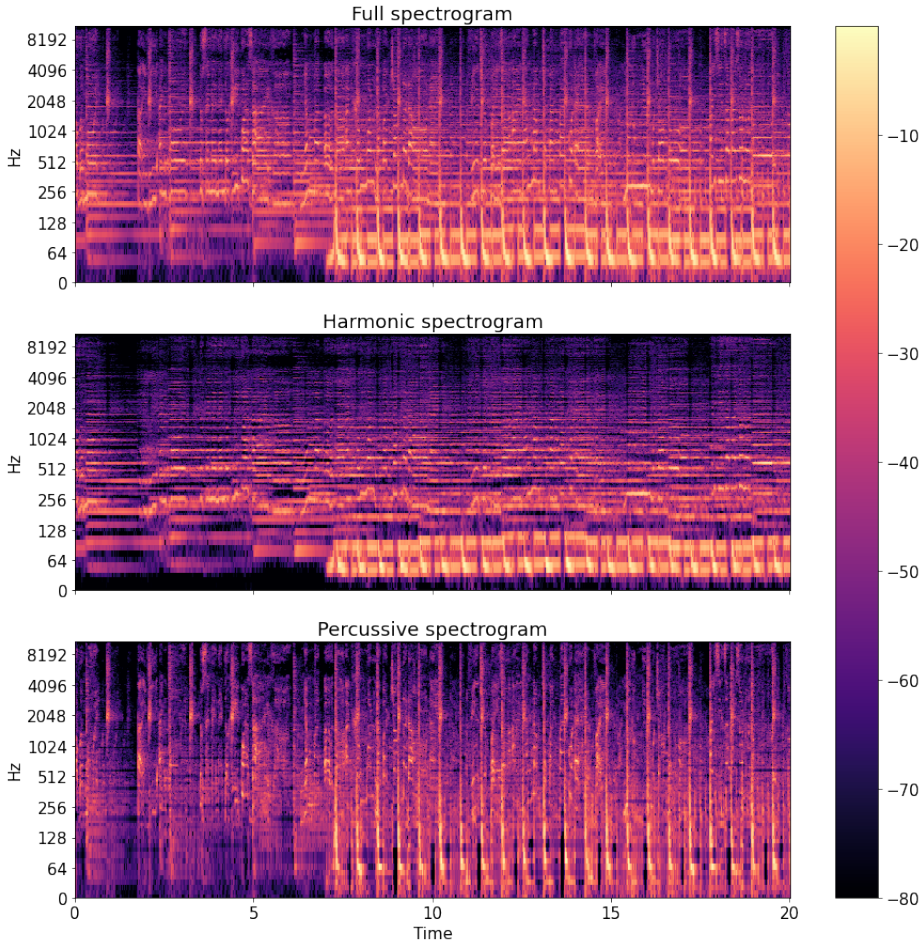
“*bèo dạt mây trôi chốn xa_xôi em ơi anh vẫn chờ bèo dạt . . .*”

Phương pháp trên sẽ sinh ra được các lời bài hát mới như sau:

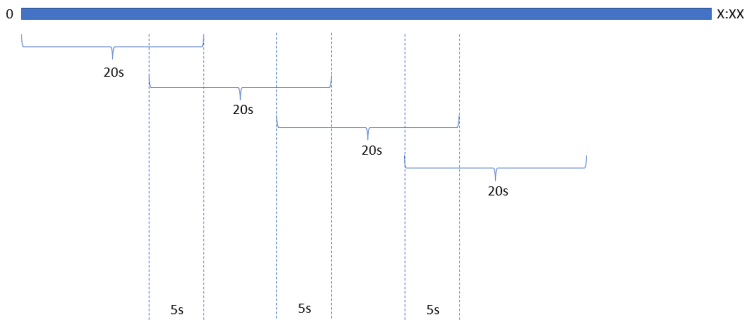
“*bèo mắc_cạn mây cuốn chốn xa_xôi em ơi anh vẫn đợi bèo mắc_cạn...*”

“*bèo_bọt lênh_đênh mây trôi chốn xa_xôi cháu ơi vợ vẫn chờ lục_bình dạt...*”

4.4.2 Tăng cường dữ liệu âm thanh Việc chỉ lấy 20 giây giúp cho dữ liệu âm thanh có thể được tăng cường một cách đáng kể. Cụ thể, với mỗi tệp âm thanh cần tăng cường, chúng tôi sẽ tiến hành cắt thành nhiều đoạn nhỏ với thời lượng 20 giây và trùng nhau 5 giây. Toàn bộ quá trình trên được mô tả trong sơ đồ Hình 3.



Hình 2: Harmonic/Percussive Spectrogram

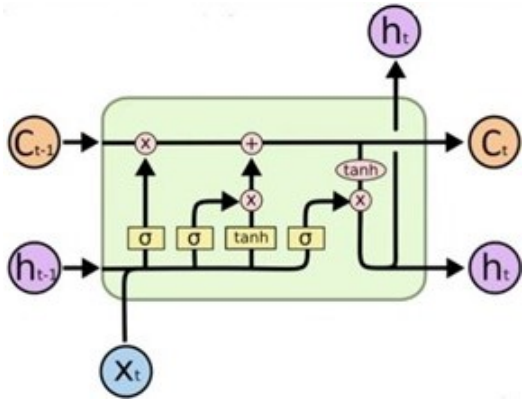


Hình 3: Sơ đồ mô tả quá trình tăng cường dữ liệu âm thanh

4.5 LSTM

Long Short-Term Memory (LSTM) là một phiên bản mở rộng của RNN, được đề xuất vào năm 1997 bởi Sepp Hochreiter và Jurgen Schmidhuber [5]. LSTM được thiết kế để giải quyết các bài toán về phụ thuộc xa trong mạng RNN do bị ảnh hưởng bởi vấn đề gradient vanishing.

Một đơn vị LSTM (được mô tả trong Hình 4) thông thường bao gồm một tế bào (cell), một cổng vào (input gate), một cổng ra (output gate) và một cổng quên (forget gate). Tế bào ghi nhớ các giá trị trong các khoảng thời gian bất kỳ và ba cổng sẽ điều chỉnh luồng thông tin ra/vào tế bào.



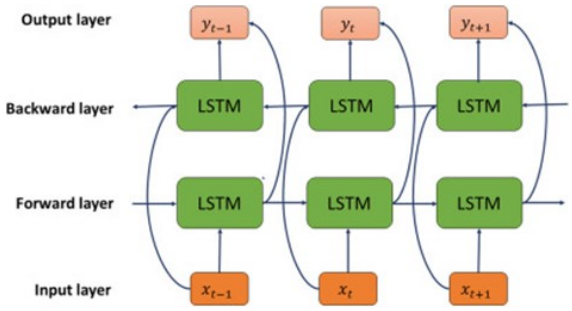
Hình 4: Mô hình LSTM

4.6 BiLSTM

Bi-LSTM (LSTM hai chiều), là một mô hình xử lý trình tự bao gồm hai LSTM đơn được sử dụng đồng thời và độc lập để mô hình hóa chuỗi đầu vào theo hai hướng: một lấy đầu vào theo hướng thuận (forward LSTM) và một theo hướng ngược lại (backward LSTM). Bi-LSTM làm tăng hiệu quả lượng thông tin có sẵn cho mạng, cải thiện ngữ cảnh có sẵn cho thuật toán. Cấu trúc mô hình được chúng tôi mô tả trong Hình 5.

4.7 Transfer learning với Resnet

Một mạng nơ-ron được huấn luyện tốt cũng tương tự với việc một bộ não được dạy dỗ tốt, có khả năng đưa ra quyết định trên nhiều dạng dữ liệu khác nhau. Áp dụng phương pháp Transfer learning, chúng ta có thể xây dựng mô hình từ mô hình đã được chuyển tiếp "kiến thức" đã được huấn luyện trước đó. Transfer Learning đã được chứng minh là một phương pháp rất hiệu quả về thời gian và



Hình 5: Mô hình Bi-LSTM

sức mạnh tính toán. Có nhiều kiến trúc mô hình học chuyển tiếp khác nhau, được đào tạo trên các bộ dữ liệu khác nhau trên khắp thế giới. Một số kiến trúc mô hình như là AlexNet, SqueezeNet, VGG, DenseNet, InceptionV3, GoogleNet, ShuffleNet, MobileNet, Resnet, ...

Dựa vào các công trình liên quan ở Phần 2 và dựa vào [4], chúng tôi sẽ sử dụng các kiến trúc mô hình thuộc họ Resnet để áp dụng Transfer Learning nhờ những hiệu quả và khả năng phân lớp mà chúng đã thể hiện trong những nghiên cứu trên [3].

4.8 Weighted Average Stacking Ensemble

Stacking là một kỹ thuật kết hợp kết quả dự đoán từ nhiều mô hình khác nhau để cải thiện hiệu suất dự đoán sau cùng. Quá trình này chia làm hai giai đoạn chính:

- Giai đoạn 1: Huấn luyện các mô hình cơ sở
- Giai đoạn 2: Huấn luyện một mô hình cuối cùng học từ kết quả dự đoán của những mô hình cơ sở

Tuy nhiên trong một số trường hợp, chúng ta sẽ muốn cho một mô hình tốt nhất trong số các mô hình cơ sở đóng góp nhiều hơn vào kết quả dự đoán cuối cùng và các mô hình yếu hơn sẽ có vai trò thấp hơn theo một trọng số nào đó. Weighted Average Ensemble là một hướng tiếp cận cho phép các mô hình cơ sở có một mức đóng góp theo tỷ lệ nhất định vào kết quả dự đoán cuối cùng.

5 Thực nghiệm

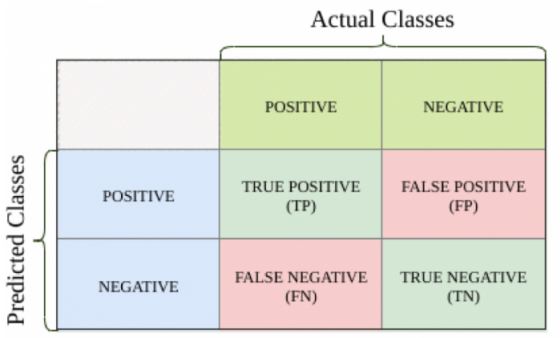
Trong phần này, chúng tôi sẽ xây dựng các mô hình kết hợp dựa trên hai giai đoạn trên cả hai tập dữ liệu train trước và sau khi áp dụng các phương pháp tăng cường. Cụ thể phân bô nhãn được mô tả trong bảng sau:

Nhãn	Nhãn số	Trước	Sau
nhạc trẻ	0	5,479	5,479
trữ tình	1	1,940	1,940
rap việt	2	658	1,974
cách mạng	3	452	1,808
thiếu nhi	4	261	1,566
rock việt	5	60	1,380

Bảng 4: Thống kê số lượng nhãn train trước và sau khi tăng cường.

5.1 Độ đo

Các độ đo phổ biến thường được dùng để đánh giá một bài toán phân loại là: Precision, Recall, F1-score. Trước khi đi sâu vào các độ đo, chúng ta cần làm rõ một số khái niệm trong Hình 6.



Hình 6: Mô tả ma trận nhầm lẫn

- True Positive – TP: giá trị thực tế và dự đoán đều là positive.
- False Positive – FP: giá trị thực tế là negative nhưng dự đoán là positive.
- True Negative – TN: giá trị thực tế và dự đoán đều là negative.
- True Positive – TP: giá trị dự đoán là negative nhưng thực tế là positive.

Trong bài toán phân loại nhiều lớp, ta lần lượt xem một lớp là positive, các lớp còn lại là negative.

5.1.1 Precision Precision được định nghĩa là tỉ lệ số điểm Positive mô hình dự đoán đúng trên tổng số điểm mô hình dự đoán là Positive.

$$Precision = \frac{TP}{TP + FP}$$

Precision càng cao, tức là số điểm mô hình dự đoán là positive đều là positive càng nhiều. Precision = 1, tức là tất cả số điểm mô hình dự doán là Positive đều đúng, hay không có điểm nào có nhãn là Negative mà mô hình dự đoán nhầm là Positive.
Trong bài toán phân loại nhiều lớp:

- Macro-precision: Là trung bình cộng của precision.
- Micro-precision: Là tỉ lệ tổng số điểm của toàn bộ các điểm positive thuộc các lớp trên tổng số điểm mô hình dự đoán là positive thuộc các lớp.

5.1.2 Recall Recall được định nghĩa là tỉ lệ số điểm Positive mô hình dự đoán đúng trên tổng số điểm thật sự là Positive (hay tổng số điểm được gán nhãn là Positive ban đầu).

$$Recall = \frac{TP}{TP + FN}$$

Recall càng cao, tức là số điểm là positive bị bỏ sót càng ít. Recall = 1, tức là tất cả số điểm có nhãn là Positive đều được mô hình nhận ra. Trong bài toán phân loại nhiều lớp:

- Macro-recall: Là trung bình cộng của recall.
- Micro-recall: Là tỉ lệ tổng số điểm của toàn bộ các điểm positive thuộc các lớp trên tổng số điểm thực sự là Positive thuộc các lớp.

5.1.3 F1-score Tuy nhiên, chỉ có Precision hay chỉ có Recall thì không đánh giá được chất lượng mô hình.

- Chỉ dùng Precision, mô hình chỉ đưa ra dự đoán cho một điểm mà nó chắc chắn nhất. Khi đó Precision = 1, tuy nhiên ta không thể nói là mô hình này tốt.
- Chỉ dùng Recall, nếu mô hình dự đoán tất cả các điểm đều là positive. Khi đó Recall = 1, tuy nhiên ta cũng không thể nói đây là mô hình tốt.

Khi đó F1-score được sử dụng. F1-score là trung bình điều hòa (harmonic mean) của precision và recall (giả sử hai đại lượng này khác 0). F1-score được tính theo công thức:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Trong bài toán phân loại nhiều lớp:

- Macro F1-score được tính tương tự theo macro-precision và macro-recall.
- Micro F1-score được tính tương tự theo micro-precision và micro-recall.

5.2 Giai đoạn 1

Ở bước này, chúng tôi sẽ tiến hành xây dựng các mô hình cơ sở cho hai nhánh trên tập train (trước và sau khi tăng cường dữ liệu) cụ thể như sau:

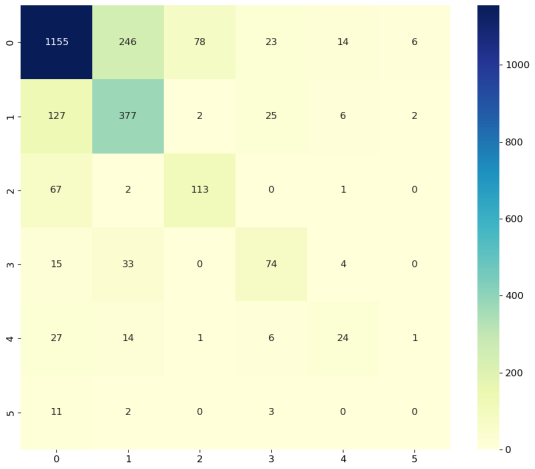
- Nhánh Lyric: LSTM và Bi-LSTM với độ dài câu tối đa 500 từ đồng thời sử dụng pretrained word embedding PhoW2V với số chiều là 300.
- Nhánh Audio: Xây dựng hai mô hình thuộc họ Resnet là Resnet18 và Resnet34 sử dụng pretrained weights từ image-net.

Sau quá trình xây dựng và kiểm thử các mô hình, chúng tôi thu được kết quả thể hiện hiệu suất phân loại của các mô hình cơ sở được thể hiện trong Bảng 5.

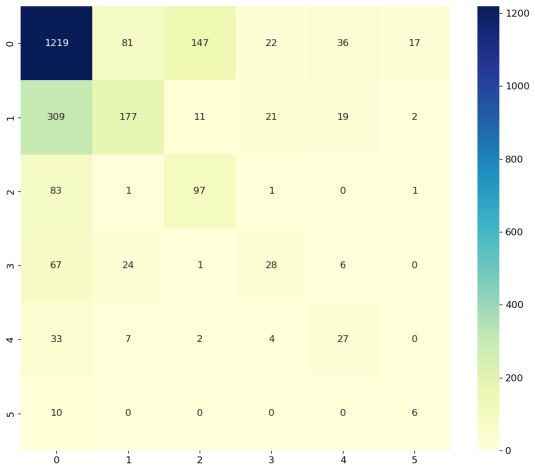
Mô hình	**Ban đầu**	**Tăng cường**
(lyric)LSTM	0.223	0.479
(lyric)Bi-LSTM	0.492	0.497
(audio)Resnet18	0.130	0.411
(audio)Resnet34	0.128	0.420

Bảng 5: F1-Score (macro) trên tập test của các mô hình cơ sở

Dựa vào bảng kết quả, mô hình cho kết quả tốt nhất là Bi-LSTM (nhánh Lyric) trên tập dữ liệu tăng cường với F1-score (macro) bằng 0.497. Đối với nhánh Audio, mô hình Resnet34 trên tập dữ liệu tăng cường cho kết quả tốt hơn Resnet18 với F1-score (macro) bằng 0.420 so với 0.411. Các mô hình xây dựng trên tập dữ liệu tăng cường đều cho kết quả cải thiện hoàn toàn so với mô hình tương ứng trên tập dữ liệu huấn luyện gốc, chứng tỏ sự hiệu quả của việc áp dụng các phương pháp tăng cường dữ liệu. Đặc biệt ở nhánh Lyric, mô hình Bi-LSTM cho kết quả tương đồng ở cả hai tập dữ liệu huấn luyện điều đó cho thấy phần nào lợi thế nhìn chuỗi theo hai chiều của kiến trúc mô hình. Hầu hết các mô hình thuộc nhánh Lyric đều có F1-score tốt hơn Audio tuy nhiên xét hai ma trận nhầm lẫn ở Hình 7 và Hình 8, mô hình Resnet34 lại dự đoán được nhãn số 5 (nhãn có số lượng ít nhất) trong khi mô hình Bi-LSTM lại không thể làm được điều tương tự.



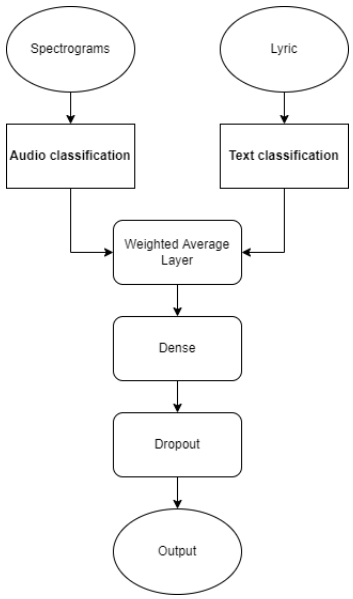
Hình 7: Ma trận nhầm lẫn của mô hình Bi-LSTM (trên tập tăng cường)



Hình 8: Ma trận nhầm lẫn của mô hình Resnet34 (trên tập tăng cường)

5.3 Giai đoạn 2

Ở giai đoạn này, chúng tôi xây dựng các mô hình kết hợp tự các mô hình cơ sở theo từng đôi Lyric-Audio, chi tiết kiến trúc được mô tả trong Hình 9. Kết quả của quá trình trên được thể hiện trong Bảng 6



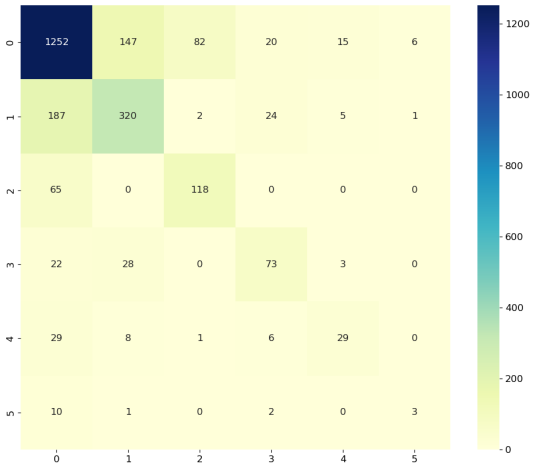
Hình 9: Mô tả kiến trúc Weighted Average Ensemble

Bảng 6 cho thấy các mô hình kết hợp đều cải thiện hiệu suất phân loại đáng kể so với các mô hình cơ sở. Trong đó, mô hình cho kết hợp giữa Bi-LSTM và Resnet18 trên tập dữ liệu tăng cường cho kết quả cao nhất với F1-score (macro) bằng 0.552. Và một lần nữa chúng ta có thể thấy được sự hiệu quả của việc tăng cường dữ liệu khi các mô hình kết hợp xây dựng trên tập dữ liệu tăng cường đều cho kết quả vượt trội (F1-score vượt ngưỡng 0.5). Nhìn vào ma trận nhầm lẫn ở Hình 10 ta có thể thấy được mô hình kết hợp đã kết thừa được điểm mạnh của các mô hình cơ sở khi đã dự đoán được nhãn 5.

Tuy nhiên xuyên suốt hai giai đoạn, ở tất cả mô hình vẫn tồn tại những dự đoán nhầm lẫn ở ba nhãn lớn là 0, 1 và 2 tương ứng với nhạc trẻ, trữ tình và rap việt. Về sự nhầm lẫn giữa nhạc trẻ và rap việt, đây là hai thể loại tương đối mới đối với giới trẻ Việt và có thể pha trộn kết hợp với nhau cho nên tồn tại sự tương đồng về mặt giai điệu. Đồng thời hiện tượng vay mượn từ ngữ tiếng nước ngoài cũng là một đặc điểm chung gây nên sự nhầm lẫn giữa hai thể loại trên. Tương tự đối với sự nhầm lẫn giữa nhạc trẻ và trữ tình, điều này có thể là do sự tương đồng về mặt nội dung, từ vựng, đều viết về tình yêu đôi lứa, cùng với đó

Mô hình	Ban đầu	Tăng cường
LSTM+resnet18	0.223	0.526
LSTM+resnet34	0.159	0.505
bi-LSTM+resnet18	0.495	0.552
bi-LSTM+resnet34	0.497	0.526

Bảng 6: F1-Score (macro) trên tập test của các mô hình kết hợp



Hình 10: Ma trận nhầm lẫn của Bi-LSTM+Resnet18 (trên tập tăng cường)

là giai điệu du dương, tình cảm. Dưới đây là một ví dụ về nhạc trẻ bị dự đoán nhầm thành rap việt bởi mô hình kết hợp có kết quả cao nhất, ta có thể thấy được hiện tượng vay mượn từ vựng tiếng Anh trong lời nhạc:

> "*hãy nhấc cánh_tay em đang có gì hãy nhắm_mắt xem em nhìn thấy gì họ muốn gì từ em_em vẫn ngây_ngô chốn đông người vì sao mất nhau chắc em buồn_buồn là vì em yêu yêu anh yêu lắm xong đêm mai đâu mất_tăm anh tìm rồi mấy trăm ngày qua em đi về đây em nói luôn cần 1 bến đỗ nơi anh này cho em là trò_chơi mà chơi chơi thôi mà uhm thắng hay thua tuỳ em tuỳ vào em chọn ai hoà nhau thì cũng đau anh đau nhất em nhì cùng 1 ván nhưng_mà đố em tìm ai có được cách chơi sâu_sắc như anh i like the way you dance clap your hands baby em sẽ nhớ đêm nhạc này hơn_bao_giờ_hết đối_với anh ngày nào cũng là tiệc một_khi nhạc vào việc all black everything từ trên nhìn xuống đầu anh lại bổ luống nói cho anh nghe em là ai và em từ đâu lỡ mai_kia nếu_như ta không gặp lại anh sẽ nhớ trong đầu you are my playgirl you play my game girl*

and i play your game girl i see you là trò_ chơi mà chơi chơi thôi mà uhm thắng hay thua tuỳ em tuỳ vào em chọn ai hoà nhau thì cũng đau anh đau nhất em nhì cùng 1 ván nhưng_ mà đố em tìm ai có được cách chơi sâu_ sắc như anh you are my playgirl you are my playgirl you play my game girl and i play your game girl ”

6 Kết luận

Trong bài báo này, chúng tôi đã áp dụng phương pháp Weighted Average Ensemble để xây dựng các mô hình dự đoán thể loại nhạc Việt kết hợp giữa nhạc và lời trên bộ dữ liệu VMusic tự thu thập với hơn 12,000 điểm dữ liệu. Sau quá trình xử lý, tăng cường dữ liệu cũng như thực nghiệm qua nhiều loại mô hình, chúng tôi thu được mô hình tốt nhất là mô hình kết hợp giữa Bi-LSTM và Resnet18 trên tập dữ liệu đã được tăng cường với F1-score (macro) bằng 0.552. Đây là một kết quả tương đối ổn tuy nhiên vẫn tồn tại một vài dự đoán nhầm lẫn giữa các nhãn có số lượng lớn. Qua đó, chúng tôi cũng đã thấy được sự hiệu quả của các phương pháp tăng cường dữ liệu đối với việc cải thiện kết quả của các bài toán phân loại trên tập dữ liệu mất cân bằng. Trong tương lai, chúng tôi sẽ mở rộng thêm bộ dữ liệu cũng như là tạo ra bộ word embedding cho miền dữ liệu lời bài hát và cùng với đó chúng tôi sẽ xây dựng nhiều dạng mô hình kết hợp phức tạp hơn cũng như nhiều trích xuất thêm nhiều dạng đặc trưng nhằm cải thiện kết quả mô hình.

Tài liệu tham khảo

[1] Tao Chen, Ruifeng Xu, Bin Liu, Qin Lu, and Jun Xu. Wemote-word embedding based minority oversampling technique for imbalanced emotion and sentiment classification. In *Workshop on Issues of Sentiment Discovery and Opinion Mining*. Citeseer, 2014.

[2] Aggelos Gkiokas, Vassilis Papavassiliou, Vassilis Katsouros, and George Carayannis. Deploying nonlinear image filters to spectrogram for harmonic/percussive separation. In *Proceedings of the International Conference on Digital Audio Effects (DAFx), York, UK*, pages 17–21, 2012.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[6] Jash Mehta, Deep Gandhi, Govind Thakur, and Pratik Kanani. Music genre classification using transfer learning on log-based mel spectrogram. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 1101–1107. IEEE, 2021.

[7] Stanley Smith Stevens, John Volkmann, and Edwin Broomell Newman. A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*, 8(3):185–190, 1937.

[8] Alexandros Tsaptsinos. Lyrics-based music genre classification using a hierarchical attention network. *arXiv preprint arXiv:1707.04678*, 2017.

[9] Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. VnCoreNLP: A Vietnamese natural language processing toolkit. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 56–60, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.