ĐẠI HỌC QUỐC GIA TP.HCM
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN - TIN HỌC

ĐỀ TÀI:

PHÂN LOẠI ÂM THANH MÔI TRƯỜNG BẰNG MẠNG TÍCH CHẬP NƠ RON TRÊN ĐẶC TRƯNG THỜI GIAN - TẦN SỐ LOG-MEL

Thành viên thực hiện:
Đinh Vĩnh Bình Nghi - 23110193

Phân loại âm thanh môi trường (ESC) đóng vai trò then chốt trong các hệ thống giám sát thông minh nhưng thường gặp khó khăn do sự chồng lấn đặc trưng giữa các lớp âm thanh. Đồ án này nghiên cứu khả năng học biểu diễn của mô hình CNN dựa trên hai dạng đầu vào chính: MFCC và Log-Mel Spectrogram. Qua quá trình thử nghiệm trên bộ dữ liệu ESC-50, nhóm nhận thấy Log-Mel Spectrogram mang lại hiệu năng ổn định hơn hẳn. Bên cạnh đó, báo cáo cũng đi sâu vào phân tích định tính thông qua cơ chế Grad-CAM để giải thích các trường hợp nhầm lẫn điển hình, tạo cơ sở cho việc tích hợp các cơ chế chú ý (attention) trong các hướng phát triển tiếp theo.

Mục lục

1 Giới thiệu

1.1 Bối cảnh và động lực của bài toán phân loại âm thanh môi trường

Trong những năm gần đây, cùng với sự phát triển của trí tuệ nhân tạo và các hệ thống cảm biến, âm thanh môi trường ngày càng được sử dụng như một nguồn thông tin bổ sung cho các hệ thống nhận thức ngữ cảnh. Các ứng dụng tiêu biểu bao gồm giám sát an ninh, phát hiện sự cố đô thị, phân tích hành vi trong không gian công cộng và các hệ thống nhà thông minh.

So với các bài toán xử lý âm thanh truyền thống như nhận dạng giọng nói hay âm nhạc, âm thanh môi trường có tính đa dạng cao và thiếu cấu trúc cố định. Các sự kiện âm thanh như tiếng còi xe, tiếng mưa, tiếng chó sủa hay tiếng va chạm thường có thời lượng ngắn, xuất hiện không đều theo thời gian và chịu ảnh hưởng mạnh của nhiễu nền cũng như sự chồng lấp giữa nhiều nguồn âm. Những đặc điểm này khiến việc trích xuất đặc trưng và phân loại trở nên khó khăn hơn.

Sự phát triển của các mô hình học sâu, đặc biệt là mạng nơ-ron tích chập, cho phép khai thác trực tiếp các biểu diễn phổ hai chiều của tín hiệu âm thanh. Thay vì phụ thuộc hoàn toàn vào các đặc trưng thủ công, CNN có khả năng học các mẫu cục bộ trên miền thời gian–tần số, từ đó cải thiện hiệu năng phân loại. Đây là lý do nhóm lựa chọn tiếp cận bài toán ESC theo hướng kết hợp biểu diễn phổ và mô hình CNN.

1.2 Đặc thù và thách thức của âm thanh môi trường

Âm thanh môi trường có nhiều đặc điểm gây khó khăn cho quá trình xử lý và phân loại. Trước hết, các sự kiện âm thanh không tuân theo cấu trúc thời gian cố định và có thể thay đổi mạnh về cường độ cũng như phổ tần số trong cùng một lớp. Một sự kiện có thể diễn ra rất ngắn hoặc kéo dài, với đặc tính phổ không ổn định.

Ngoài ra, sự tương đồng về phổ giữa một số lớp âm thanh khác nhau dẫn đến hiện tượng chồng lấp đặc trưng, làm giảm khả năng phân tách ranh giới lớp. Nhiễu môi trường và sự khác biệt về điều kiện thu âm cũng ảnh hưởng trực tiếp đến chất lượng tín hiệu, khiến các đặc trưng trích xuất từ dữ liệu thô kém ổn định.

Một thách thức quan trọng khác là lựa chọn biểu diễn đặc trưng phù hợp. Các đặc trưng truyền thống như MFCC tuy hiệu quả trong nhận dạng giọng nói nhưng có xu hướng làm mất thông tin cấu trúc thời gian–tần số, vốn quan trọng đối với âm thanh môi trường. Do đó, các biểu diễn phổ hai chiều như Log-Mel Spectrogram được xem là phù hợp hơn khi kết hợp với các mô hình học sâu.

1.3 Mục tiêu và phạm vi đồ án

Mục tiêu chính của đồ án là xây dựng và đánh giá một hệ thống phân loại âm thanh môi trường dựa trên các biểu diễn phổ và mô hình học sâu. Cụ thể, đồ án tập trung vào các mục tiêu sau:

- Khảo sát và so sánh hai phương pháp trích xuất đặc trưng phổ phổ biến là **MFCC** và **Log-Mel Spectrogram** trong bối cảnh bài toán ESC.

- Xây dựng mô hình **Convolutional Neural Network (CNN)** để thực hiện phân loại âm thanh dựa trên các biểu diễn thời gian–tần số.
- Đánh giá hiệu năng mô hình thông qua các thí nghiệm thực nghiệm và phân tích kết quả theo từng lớp âm thanh.
- Phân tích vai trò của biểu diễn thời gian–tần số đối với khả năng học và tổng quát hóa của mô hình.

Trong phạm vi đồ án, nhóm chỉ xem xét các kiến trúc CNN cơ bản và các đặc trưng phổ tiêu chuẩn, chưa triển khai các mô hình học sâu phức tạp hoặc cơ chế chú ý nâng cao. Các hướng mở rộng này sẽ được đề cập như định hướng nghiên cứu trong tương lai.

1.4 Cấu trúc báo cáo

Báo cáo này được tổ chức thành 7 chương, nhằm trình bày một cách hệ thống quá trình xây dựng và đánh giá mô hình phân loại âm thanh môi trường:

- **Chương 1 – Giới thiệu:** Trình bày bối cảnh nghiên cứu, các thách thức đặc thù của bài toán phân loại âm thanh môi trường, động cơ lựa chọn phương pháp học sâu, cũng như mục tiêu và đóng góp chính của đề tài.
- **Chương 2 – Dữ liệu và thiết lập thực nghiệm:** Mô tả tập dữ liệu ESC-50, chiến lược chia tập huấn luyện–kiểm tra nhằm đảm bảo tính khách quan, cùng các tiêu chí đánh giá được sử dụng trong nghiên cứu.
- **Chương 3 – Phương pháp:** Trình bày chi tiết pipeline xử lý tín hiệu âm thanh từ miền thời gian sang miền thời gian–tần số, quá trình trích xuất đặc trưng Log-Mel, Delta và Delta-Delta, thiết kế biểu diễn đầu vào 3 kênh, cũng như kiến trúc mô hình CNN và các kỹ thuật regularization được áp dụng.
- **Chương 4 – Giải thích mô hình:** Giới thiệu phương pháp Grad-CAM nhằm trực quan hóa vùng đặc trưng mà mô hình CNN tập trung khi đưa ra quyết định, từ đó giúp tăng tính minh bạch và khả năng diễn giải của mô hình.
- **Chương 5 – Kết quả và thảo luận:** Phân tích kết quả thực nghiệm, so sánh hiệu năng giữa phương pháp truyền thống và mô hình đề xuất, đồng thời thảo luận nguyên nhân của các trường hợp dự đoán đúng và sai.
- **Chương 6 – Hạn chế và hướng phát triển:** Đánh giá những giới hạn của mô hình hiện tại và đề xuất các hướng cải tiến tiềm năng nhằm nâng cao hiệu năng trong tương lai.
- **Chương 7 – Kết luận:** Tóm tắt những đóng góp chính của đề tài và khẳng định ý nghĩa của phương pháp được đề xuất đối với bài toán phân loại âm thanh môi trường.

2 Dữ liệu và thiết lập thực nghiệm

2.1 Tập dữ liệu ESC-50

Trong nghiên cứu này, tập dữ liệu **ESC-50** (Environmental Sound Classification – 50 classes) được sử dụng làm cơ sở thực nghiệm. Đây là một tập dữ liệu chuẩn, được sử dụng rộng rãi trong các nghiên cứu về phân loại âm thanh môi trường.

ESC-50 bao gồm **2000 đoạn âm thanh**, mỗi đoạn có độ dài **5 giây**, được thu với tần số lấy mẫu **44.1 kHz** và được phân chia đều thành **50 lớp**, tương ứng với các loại âm thanh phổ biến trong môi trường thực tế.

Các lớp âm thanh được chia thành **5 nhóm lớn**:

- Âm thanh động vật (Animal sounds);
- Âm thanh tự nhiên (Natural soundscapes & water sounds);
- Âm thanh con người (Human, non-speech sounds);
- Âm thanh nội thất (Interior/domestic sounds);
- Âm thanh ngoại thất (Exterior/urban noises).

Đặc điểm nổi bật của ESC-50 là số lượng mẫu trên mỗi lớp tương đối nhỏ (40 mẫu/lớp), trong khi các lớp âm thanh có thể có đặc trưng phổ và năng lượng chồng lấn mạnh, khiến bài toán phân loại trở nên đặc biệt thách thức.

2.2 Chiến lược chia tập dữ liệu

ESC-50 được cung cấp sẵn cơ chế chia dữ liệu theo **5 fold**, nhằm hỗ trợ đánh giá mô hình một cách khách quan.

Trong nghiên cứu này, nhóm sử dụng:

- Fold 5 làm tập kiểm tra (test set);
- Các fold 1–4 làm tập huấn luyện và xác thực.

Tập huấn luyện–xác thực được tiếp tục chia theo tỷ lệ **80% huấn luyện – 20% xác thực**, nhằm theo dõi quá trình học của mô hình và áp dụng các kỹ thuật dừng sớm (early stopping).

Cách chia này đảm bảo cho các mẫu trong tập kiểm tra không xuất hiện trong quá trình huấn luyện. Việc đánh giá phản ánh đúng khả năng tổng quát hóa của mô hình trên dữ liệu chưa từng thấy.

2.3 Tiêu chí đánh giá

Để đánh giá hiệu năng của các mô hình phân loại âm thanh, nghiên cứu sử dụng **độ chính xác phân loại (Accuracy)** làm chỉ số chính, được định nghĩa như sau:

$$\text{Accuracy} = \frac{\text{Số mẫu dự đoán đúng}}{\text{Tổng số mẫu kiểm tra}} \tag{1}$$

Bên cạnh đó, **ma trận nhầm lẫn (Confusion Matrix)** được sử dụng như một công cụ trực quan nhằm:

- Phân tích chi tiết hiệu năng trên từng lớp;
- Xác định các cặp lớp dễ bị nhầm lẫn;
- Làm cơ sở cho việc giải thích và cải tiến mô hình.

Việc kết hợp chỉ số định lượng (Accuracy) và phân tích định tính (Confusion Matrix, Grad-CAM) cho phép đánh giá mô hình một cách toàn diện hơn.

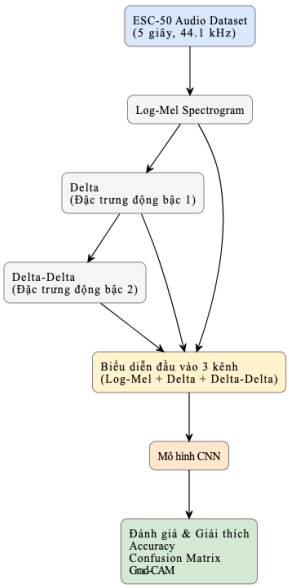
2.4 Tổng quan pipeline thực nghiệm

Toàn bộ quy trình thực nghiệm trong nghiên cứu được mô tả trong Hình 1, bao gồm các bước chính:

- Tiền xử lý tín hiệu âm thanh từ dạng sóng thời gian (waveform);
- Biểu diễn thời gian–tần số thông qua Log-Mel Spectrogram;
- Trích xuất đặc trưng động (Delta và Delta-Delta);
- Xây dựng biểu diễn đầu vào 3 kênh cho mô hình học sâu;
- Huấn luyện mô hình CNN với các kỹ thuật regularization;
- Đánh giá và giải thích kết quả bằng Accuracy, Confusion Matrix và Grad-CAM.

Pipeline này được thiết kế nhằm tận dụng đồng thời:

- Thông tin phổ tĩnh (Log-Mel);
- Sự biến thiên theo thời gian (Delta, Delta-Delta);
- Khả năng học biểu diễn phi tuyến của mạng nơ-ron tích chập.



Hình 1: Pipeline tổng thể của hệ thống phân loại âm thanh môi trường, từ tín hiệu âm thanh thô đến kết quả dự đoán và giải thích mô hình.

3 Biểu diễn đặc trưng âm thanh

3.1 Từ miền thòi gian đến miền thời gian - tần số

3.1.1 Biểu diễn tín hiệu trong miền thời gian (Waveform)

Âm thanh ở dạng nguyên thủy nhất được biểu diễn dưới dạng tín hiệu trong miền thời gian (*waveform*), trong đó biên độ của tín hiệu thay đổi theo thời gian. Đối với tín hiệu âm thanh số, waveform có thể được mô tả như một chuỗi các mẫu rời rạc:

$$x[n], \quad n = 1, 2, \ldots, N \tag{2}$$

trong đó $x[n]$ là biên độ tín hiệu tại thời điểm lấy mẫu thứ n, và N phụ thuộc vào tần số lấy mẫu cũng như độ dài của đoạn âm thanh.

Biểu diễn waveform phản ánh trực tiếp các đặc trưng vật lý của sóng âm, bao gồm:

- **Cường độ âm thanh**, liên quan đến biên độ của tín hiệu
- **Khoảng lặng hoặc các xung âm đột ngột**
- **Thời điểm xuất hiện** của các sự kiện âm thanh

Tuy nhiên, trong bài toán phân loại âm thanh môi trường, biểu diễn waveform tồn tại nhiều hạn chế quan trọng:

- **Thiếu thông tin tần số:** Waveform không cung cấp trực tiếp thông tin về các thành phần tần số của tín hiệu, trong khi các đặc trưng phân biệt giữa các loại âm thanh (ví dụ như tiếng mưa, tiếng động cơ hay tiếng côn trùng) thường nằm trong miền tần số.
- **Độ nhạy cao với nhiễu và dịch chuyển thời gian:** Hai tín hiệu thuộc cùng một loại âm thanh nhưng bị lệch pha hoặc lệch thời điểm có thể có waveform rất khác nhau, gây khó khăn cho các mô hình học máy truyền thống.
- **Khó học đặc trưng hiệu quả với dữ liệu nhỏ:** Việc học trực tiếp từ waveform thường đòi hỏi tập dữ liệu lớn và mô hình có độ phức tạp cao, trong khi bộ dữ liệu ESC-50 có kích thước tương đối hạn chế.

Do đó, mặc dù waveform là điểm khởi đầu tự nhiên trong phân tích tín hiệu âm thanh, biểu diễn này không phù hợp để sử dụng trực tiếp trong các mô hình phân loại âm thanh môi trường. Điều này dẫn đến nhu cầu chuyển đổi tín hiệu từ miền thời gian sang các không gian biểu diễn giàu thông tin hơn, kết hợp đồng thời cả yếu tố thời gian và tần số.

Trong phần tiếp theo, nghiên cứu trình bày phương pháp *Short-Time Fourier Transform* (STFT), cho phép phân tích sự phân bố năng lượng của tín hiệu theo cả thời gian và tần số, từ đó tạo nền tảng cho các biểu diễn phổ được sử dụng trong các mô hình học sâu.

3.1.2 Biến đổi Fourier ngắn hạn (Short-Time Fourier Transform – STFT)

Nghiên cứu sử dụng biến đổi Fourier ngắn hạn (*Short-Time Fourier Transform – STFT*), cho phép phân tích tín hiệu âm thanh trong các cửa sổ thời gian ngắn, từ đó thu được biểu diễn kết hợp giữa miền thời gian và miền tần số.

Về mặt ý tưởng, STFT chia tín hiệu âm thanh thành các đoạn ngắn (*frame*) có độ dài cố định, với giả định rằng tín hiệu là gần như ổn định (*stationary*) trong mỗi đoạn. Sau đó, phép biến đổi Fourier được áp dụng lên từng frame nhằm trích xuất thông tin phổ tần cục bộ theo thời gian.

Về mặt toán học, STFT của tín hiệu rời rạc $x[n]$ được định nghĩa như sau:

$$X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]\, w[n-m]\, e^{-j\omega n} \tag{3}$$

trong đó:

- $w[n]$ là hàm cửa sổ (*window function*), thường sử dụng các cửa sổ như Hamming hoặc Hann
- m biểu thị vị trí của cửa sổ trên trục thời gian
- ω là tần số góc

Kết quả của STFT là một biểu diễn hai chiều, cho thấy cường độ năng lượng của tín hiệu tại từng thời điểm và từng dải tần. Khi biểu diễn biên độ hoặc năng lượng của STFT theo trục thời gian–tần số, ta thu được *spectrogram*.

3.2 Log-Mel Spectrogram

3.2.1 Ánh xạ tần số theo thang Mel

Mặc dù spectrogram thu được từ phép biến đổi Fourier ngắn hạn (STFT) đã cung cấp biểu diễn kết hợp giữa miền thời gian và miền tần số, dạng biểu diễn này vẫn tồn tại một số hạn chế khi được sử dụng trực tiếp trong các mô hình phân loại âm thanh.

Thứ nhất, trục tần số của spectrogram là tuyến tính theo đơn vị Hertz (Hz), trong khi hệ thính giác của con người có độ phân giải cao hơn ở dải tần thấp và giảm dần ở dải tần cao. Thứ hai, biên độ phổ thường có dải động rất lớn, khiến các thành phần năng lượng yếu dễ bị lấn át bởi các thành phần có năng lượng mạnh hơn.

Để khắc phục các hạn chế trên, nghiên cứu sử dụng *Log-Mel Spectrogram*, một biểu diễn phổ được xây dựng thông qua hai bước chính: áp dụng thang Mel và thực hiện phép biến đổi logarit lên năng lượng phổ.

3.2.2 Thang Mel và bộ lọc Mel (Mel Filter Bank)

Thang Mel được đề xuất nhằm mô phỏng mối quan hệ phi tuyến giữa tần số vật lý (Hz) và tần số được cảm nhận bởi tai người. Mối quan hệ này có thể được xấp xỉ bằng công thức:

$$\mathrm{Mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{4}$$

Theo thang Mel, các tần số thấp được phân giải chi tiết hơn so với các tần số cao. Dựa trên nguyên lý này, một bộ lọc Mel (*Mel filter bank*) gồm các bộ lọc tam giác chồng lấn được xây dựng và áp dụng lên phổ công suất thu được từ STFT.

Quá trình áp dụng bộ lọc Mel mang lại các lợi ích sau:

- Giảm số chiều của biểu diễn phổ;
- Nhấn mạnh các dải tần quan trọng về mặt cảm nhận thính giác;
- Tăng tính ổn định của đặc trưng trước nhiễu và các biến thiên nhỏ trong tín hiệu.

Kết quả của bước này là *Mel Spectrogram*, biểu diễn năng lượng của tín hiệu theo các dải tần Mel theo thời gian.

3.2.3 Phép biến đổi logarit

Sau khi áp dụng bộ lọc Mel, năng lượng phổ của tín hiệu vẫn có thể chênh lệch rất lớn giữa các dải tần. Do đó, một phép biến đổi logarit được sử dụng nhằm nén dải động của tín hiệu và làm nổi bật các thành phần năng lượng yếu:

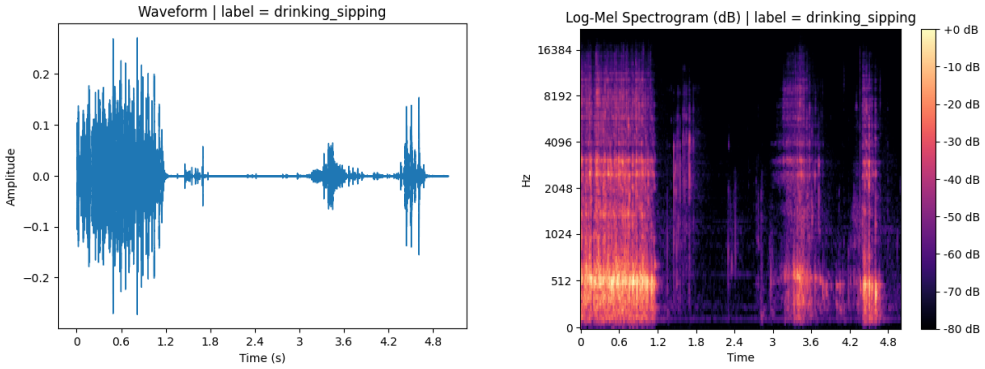
$$S(t, f) = \log\left(M(t, f) + \varepsilon\right) \tag{5}$$

trong đó $M(t, f)$ là năng lượng Mel tại thời điểm t và dải Mel thứ f, còn $\varepsilon > 0$ là một hằng số nhỏ nhằm tránh hiện tượng $\log(0)$.

Phép biến đổi logarit mang lại các lợi ích sau:

- Làm giảm ảnh hưởng của các thành phần năng lượng quá lớn;
- Làm nổi bật các cấu trúc phổ có năng lượng thấp nhưng mang tính phân biệt;
- Đưa biểu diễn phổ tiến gần hơn với cơ chế cảm nhận cường độ âm thanh của hệ thính giác con người.

Sau hai bước áp dụng thang Mel và phép biến đổi logarit, tín hiệu âm thanh được biểu diễn dưới dạng *Log-Mel Spectrogram*, một ma trận hai chiều theo trục thời gian và trục Mel.



(a) Waveform (b) Log-Mel Spectrogram

Hình 2: So sánh biểu diễn tín hiệu âm thanh *drinking_ sipping* trong miền thời gian (waveform) và miền thời gian–tần số (spectrogram). Biểu diễn phổ cho thấy rõ cấu trúc năng lượng theo thời gian và tần số, trong khi waveform chỉ phản ánh biên độ theo thời gian.

3.2.4 Log-Mel Spectrogram như một "ảnh" cho mô hình học sâu

Log-Mel Spectrogram có thể được xem như một ảnh hai chiều, trong đó:

- Trục ngang biểu diễn sự tiến triển theo thời gian
- Trục dọc biểu diễn các dải tần Mel
- Giá trị tại mỗi điểm biểu diễn cường độ năng lượng sau phép biến đổi logarit.

Cách biểu diễn này đặc biệt phù hợp với các mô hình mạng nơ-ron tích chập (CNN), vốn được thiết kế để học các mẫu không gian cục bộ. Trong bối cảnh âm thanh môi trường, các mẫu này có thể tương ứng với các dải năng lượng ổn định theo thời gian, các xung âm ngắn hạn hoặc các cấu trúc phổ lặp lại.

Nhờ đó, Log-Mel Spectrogram đóng vai trò là cầu nối quan trọng giữa xử lý tín hiệu âm thanh truyền thống và các mô hình học sâu hiện đại được sử dụng trong nghiên cứu này.

3.3 Đặc trưng động và biểu diễn ba kênh

Log-Mel Spectrogram cung cấp thông tin phổ tĩnh của tín hiệu âm thanh tại từng thời điểm. Tuy nhiên, trong nhiều loại âm thanh môi trường, sự biến thiên theo thời gian của phổ mang ý nghĩa phân biệt quan trọng không kém so với giá trị phổ tại một thời điểm cụ thể.

Để nắm bắt các đặc trưng động này, nghiên cứu sử dụng các đặc trưng Delta (Δ) và Delta-Delta ($\Delta\Delta$), lần lượt biểu diễn đạo hàm bậc nhất và đạo hàm bậc hai theo thời gian của Log-Mel Spectrogram.

3.3.1 Đặc trưng Delta

Đặc trưng Delta mô tả tốc độ thay đổi của năng lượng phổ theo trục thời gian. Thay vì sử dụng đạo hàm rời rạc đơn giản (dễ nhạy với nhiễu), Delta thường được ước lượng bằng hồi quy tuyến tính trong một cửa sổ $\pm L$ khung:

$$\Delta S(t,f) = \frac{\sum_{n=1}^{L} n\,[S(t+n,f) - S(t-n,f)]}{2\sum_{n=1}^{L} n^2} \tag{6}$$

Đặc trưng Delta cho phép mô hình nhận biết:

- Xu hướng tăng hoặc giảm năng lượng tại các dải tần;
- Thời điểm bắt đầu và kết thúc của các sự kiện âm thanh;
- Các biến thiên ngắn hạn trong cấu trúc phổ.

3.3.2 Đặc trưng Delta-Delta

Delta-Delta thể hiện **gia tốc thay đổi** (độ cong theo thời gian), giúp bắt các biến thiên nhanh như onset/offset của sự kiện âm thanh:

$$\Delta^2 S(t,f) \;=\; \frac{\displaystyle\sum_{n=1}^{L} n\,(\Delta S(t+n,f) - \Delta S(t-n,f))}{\displaystyle 2\sum_{n=1}^{L} n^2}. \tag{7}$$

Ý nghĩa trực quan.

- $S(t,f)$ nắm bắt **hình dạng phổ** tại từng thời điểm.
- $\Delta S(t,f)$ làm nổi bật **xu hướng tăng/giảm năng lượng** theo thời gian (ví dụ lúc âm thanh khởi phát mạnh của tiếng còi).
- $\Delta^2 S(t,f)$ nhấn mạnh **mức độ đột ngột** của biến đổi, hữu ích cho các sự kiện ngắn và sắc nét.

Nhờ vậy, bộ đặc trưng $(S, \Delta S, \Delta^2 S)$ giúp mô hình phân biệt tốt hơn các lớp có phổ tĩnh tương tự nhưng có động học khác nhau.

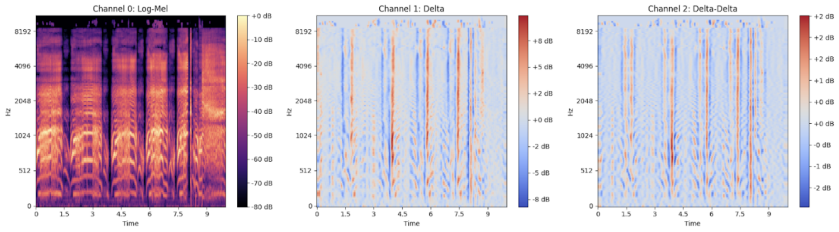
3.3.3 Kết hợp Log-Mel, Delta và Delta-Delta

Trong nghiên cứu này, ba đặc trưng Log-Mel, Delta và Delta-Delta được kết hợp để tạo thành một biểu diễn ba kênh (*3-channel input*), tương tự như cách biểu diễn ảnh màu RGB trong thị giác máy tính.

Cách kết hợp này cho phép các mô hình mạng nơ-ron tích chập (CNN):

- Khai thác thông tin phổ tĩnh (Log-Mel);
- Nắm bắt sự thay đổi tuyến tính theo thời gian (Delta);
- Nhận diện các biến thiên phi tuyến và gia tốc phổ (Delta-Delta).

Việc giữ nguyên cấu trúc hai chiều theo thời gian–tần số của các đặc trưng này giúp CNN học được các mẫu cục bộ theo cả chiều thời gian và chiều tần số một cách hiệu quả.



Hình 3: Biểu diễn đầu vào ba kênh gồm Log-Mel Spectrogram, Delta và Delta-Delta của cùng một mẫu âm thanh.

3.4 Mô hình và Huấn luyện

3.4.1 Baseline: MFCC kết hợp kNN

Trước khi triển khai các mô hình học sâu, nghiên cứu xây dựng một phương pháp baseline dựa trên đặc trưng Mel-Frequency Cepstral Coefficients (MFCC) kết hợp với bộ phân loại k-Nearest Neighbors (kNN), nhằm làm mốc so sánh cho các mô hình đề xuất.

MFCC là một trong những đặc trưng phổ được sử dụng phổ biến trong xử lý tiếng nói và âm thanh, được thiết kế nhằm xấp xỉ cơ chế cảm nhận của hệ thính giác con người. Trong thiết lập này, tín hiệu âm thanh sau khi được chuyển sang miền thời gian–tần số và qua ngân hàng lọc Mel sẽ tiếp tục được biến đổi bằng phép biến đổi cosine rời rạc (*Discrete Cosine Transform – DCT*) để thu được các hệ số MFCC.

Vai trò của DCT và giới hạn của biểu diễn MFCC

Sau bước lọc Mel và lấy log năng lượng, phổ Log-Mel được biến đổi bằng Biến đổi Cosine rời rạc (Discrete Cosine Transform – DCT) nhằm thu được các hệ số MFCC:

$$c_k = \sum_{m=1}^{M} \log(E_m) \cos\left[\frac{\pi k}{M}\left(m - \frac{1}{2}\right)\right], \quad k = 1, \ldots, K$$

trong đó E_m là năng lượng của bộ lọc Mel thứ m, M là số bộ lọc Mel và K là số hệ số MFCC được giữ lại.

Mục tiêu chính của phép biến đổi DCT là khử tương quan giữa các hệ số và nén thông tin phổ vào một số chiều thấp, giúp các mô hình thống kê hoặc dựa trên khoảng cách hoạt động ổn định hơn. Tuy nhiên, quá trình nén này cũng làm **mất đi cấu trúc không gian** và mối quan hệ thời gian–tần số cục bộ của tín hiệu âm thanh.

Do đặc trưng MFCC được biểu diễn dưới dạng các vector một chiều theo từng khung thời gian, các vector này được tổng hợp theo thời gian thông qua các thống kê

đơn giản như trung bình và phương sai, nhằm tạo ra một biểu diễn đặc trưng cố định cho mỗi mẫu âm thanh. Sau đó, bộ phân loại kNN được sử dụng để thực hiện phân loại dựa trên khoảng cách trong không gian đặc trưng.

Phương pháp MFCC kết hợp kNN có một số ưu điểm như:

- Cấu trúc đơn giản, dễ triển khai;
- Chi phí tính toán thấp;
- Phù hợp làm chuẩn tham chiếu trong các bài toán phân loại âm thanh.

Tuy nhiên, phương pháp này cũng tồn tại những hạn chế cơ bản:

- Không thực hiện học biểu diễn đặc trưng (*representation learning*) mà phụ thuộc hoàn toàn vào đặc trưng thủ công;
- Làm mất cấu trúc hai chiều thời gian–tần số của tín hiệu khi rút gọn thành vector đặc trưng;
- Khả năng phân biệt hạn chế đối với các lớp âm thanh có đặc tính phổ tương đồng.

Do đó, phương pháp MFCC + kNN được sử dụng trong nghiên cứu này với vai trò baseline, nhằm định lượng mức cải thiện hiệu năng khi chuyển sang các mô hình học sâu dựa trên biểu diễn phổ hai chiều.

3.4.2 Mô hình đề xuất: CNN trên biểu diễn Log-Mel ba kênh

Dựa trên các phân tích ở các mục trước, nghiên cứu đề xuất sử dụng mô hình Convolutional Neural Network (CNN) cho bài toán phân loại âm thanh môi trường, với đầu vào là biểu diễn Log-Mel Spectrogram ba kênh gồm phổ tĩnh (Log-Mel), Delta và Delta-Delta.

Không giống như các phương pháp truyền thống dựa trên đặc trưng rút gọn một chiều, mô hình CNN làm việc trực tiếp trên biểu diễn hai chiều theo trục thời gian–tần số. Điều này cho phép mô hình khai thác đồng thời cấu trúc phổ và sự biến thiên theo thời gian của tín hiệu âm thanh, từ đó học được các đặc trưng mang tính phân biệt cao hơn giữa các lớp âm thanh.

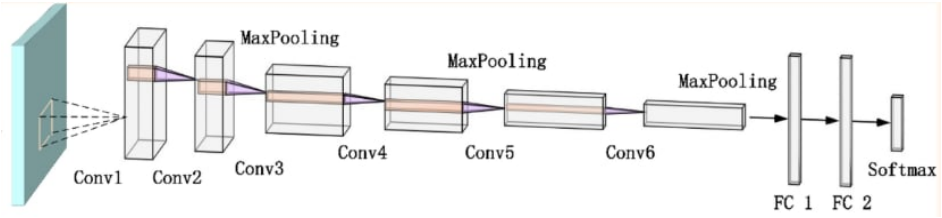
Kiến trúc mô hình được sử dụng trong đồ án

Trong đồ án này, nhóm sử dụng một kiến trúc CNN cơ bản nhưng hiệu quả để thực hiện bài toán phân loại âm thanh môi trường. Mô hình được thiết kế nhằm cân bằng giữa khả năng học biểu diễn và độ phức tạp tính toán, phù hợp với quy mô dữ liệu và mục tiêu của đồ án.

Kiến trúc tổng quát của mô hình bao gồm các thành phần chính sau:

- **Lớp đầu vào (Input layer):**
  Nhận Log-Mel Spectrogram của mỗi mẫu âm thanh, được xem như một ảnh hai chiều với một kênh (single-channel).
- **Các khối tích chập (Convolutional blocks):**
  Mỗi khối bao gồm một hoặc nhiều lớp tích chập 2D, theo sau là hàm kích hoạt phi tuyến (ReLU) và lớp pooling. Các lớp này có nhiệm vụ trích xuất các đặc trưng cục bộ trong miền thời gian–tần số và giảm dần kích thước không gian của đặc trưng.

- **Lớp gộp (Pooling layer):**
  Giúp giảm độ nhạy của mô hình đối với các dịch chuyển nhỏ theo thời gian hoặc tần số, đồng thời giảm số lượng tham số cần học.
- **Các lớp fully connected:**
  Sau khi trích xuất đặc trưng, các lớp fully connected kết hợp các đặc trưng cấp cao và thực hiện phân loại.
- **Lớp đầu ra (Output layer):**
  Sử dụng hàm kích hoạt Softmax để dự đoán xác suất thuộc về từng lớp âm thanh môi trường.



Hình 4: Kiến trúc tổng quát của mô hình CNN sử dụng trong đồ án.

Kiến trúc này cho phép mô hình học được các đặc trưng quan trọng từ dữ liệu phổ Log-Mel, đồng thời đảm bảo tính ổn định và khả năng tổng quát hóa trong quá trình huấn luyện và kiểm thử.

Đầu vào ba kênh cho CNN

Đối với mỗi mẫu âm thanh, đầu vào của mô hình CNN được biểu diễn dưới dạng một tensor ba chiều:

$$\mathbf{X} \in \mathbb{R}^{T \times F \times 3}, \tag{8}$$

trong đó T là số khung thời gian, F là số băng tần Mel, và ba kênh lần lượt tương ứng với:

- Kênh 1: Log-Mel Spectrogram (đặc trưng phổ tĩnh);
- Kênh 2: Delta (tốc độ thay đổi theo thời gian);
- Kênh 3: Delta-Delta (gia tốc thay đổi theo thời gian).

Việc sử dụng biểu diễn ba kênh cho phép mô hình CNN đồng thời tiếp cận thông tin về cấu trúc phổ và động học của tín hiệu, thay vì chỉ dựa trên một dạng đặc trưng đơn lẻ.

Vì sao xếp theo kênh lại "thông minh hơn"?

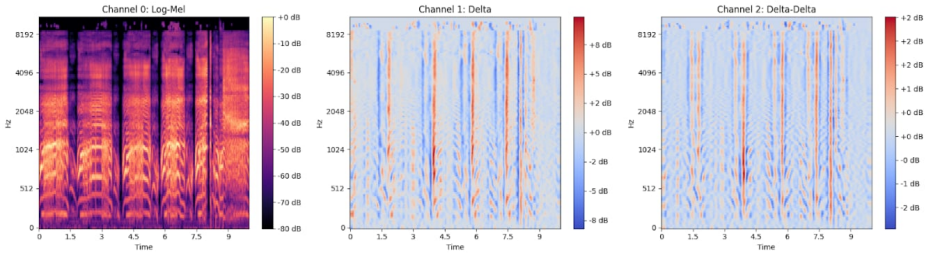
- **Giữ nguyên cấu trúc 2D thời gian–tần số:** CNN học các mẫu cục bộ (local patterns) trên mặt phẳng (t, f) như cạnh (edges), dải năng lượng, formant-like bands, onset.

- **Cho phép tích chập khai thác tương tác liên kênh:** Với kernel $W \in \mathbb{R}^{k_t \times k_f \times 3}$, một feature map đầu ra được tính:

$$Y = \sigma\left(\sum_{c=1}^{3} W_c * X_c + b\right), \tag{9}$$

  nghĩa là mô hình có thể học cách kết hợp tối ưu giữa "phổ tĩnh" và "động học" ngay ở tầng đầu.
- **Ổn định hơn so với nối chiều (concatenate):** Nếu nối đặc trưng theo chiều tần số hoặc chiều đặc trưng, mô hình dễ bị mất tính tương ứng giữa các vùng thời gian–tần số và làm tăng kích thước đầu vào không cần thiết.

Cách biểu diễn 3 kênh có thể xem tương tự như một ảnh RGB, trong đó mỗi "màu" mang một loại thông tin khác nhau: tĩnh (structure), tốc độ thay đổi (trend) và gia tốc (abruptness). Đây là lý do nhóm lựa chọn biểu diễn 3 kênh để cải thiện hiệu năng phân loại âm thanh môi trường.



Hình 5: Biểu diễn nhiều kênh.

Huấn luyện mô hình

Trong quá trình huấn luyện, mô hình CNN được học trực tiếp từ dữ liệu huấn luyện thông qua việc tối ưu hàm mất mát phân loại đa lớp. Các tham số của mạng được cập nhật dựa trên tập xác thực nhằm giảm thiểu hiện tượng học quá khớp.

Biểu diễn Log-Mel Spectrogram ba kênh kết hợp với kiến trúc CNN cho phép mô hình tự động học các đặc trưng phù hợp cho bài toán phân loại âm thanh môi trường mà không cần thiết kế thủ công các đặc trưng phức tạp.

Để tăng khả năng tổng quát hóa của mô hình trong bối cảnh dữ liệu huấn luyện hạn chế, các kỹ thuật regularization và chiến lược tối ưu hóa sẽ được trình bày trong các mục tiếp theo. Đặc biệt, nghiên cứu áp dụng kỹ thuật tăng cường dữ liệu *SpecAugment* nhằm cải thiện hiệu năng trên tập kiểm tra.

3.5 Kỹ thuật tăng cường dữ liệu SpecAugment

Trong các bài toán phân loại âm thanh môi trường, dữ liệu huấn luyện thường có quy mô hạn chế và chịu ảnh hưởng mạnh từ nhiễu, điều kiện thu âm và sự đa dạng của

ngữ cảnh thực tế. Điều này dễ dẫn đến hiện tượng *overfitting*, đặc biệt khi sử dụng các mô hình học sâu có số lượng tham số lớn như CNN.

Để cải thiện khả năng tổng quát hóa của mô hình, nhóm áp dụng kỹ thuật tăng cường dữ liệu SpecAugment, một phương pháp tăng cường trực tiếp trên biểu diễn phổ thời gian–tần số thay vì trên tín hiệu âm thanh gốc. SpecAugment đã được chứng minh hiệu quả trong nhiều bài toán xử lý âm thanh và đặc biệt phù hợp với các mô hình dựa trên spectrogram.

3.5.1 Nguyên lý của SpecAugment

SpecAugment hoạt động bằng cách che (mask) có chủ đích một số vùng trong biểu diễn Log-Mel Spectrogram, buộc mô hình phải học các đặc trưng mang tính tổng quát hơn thay vì phụ thuộc quá mức vào một số vùng phổ cụ thể. Không giống như nhiễu ngẫu nhiên trong miền thời gian, SpecAugment khai thác trực tiếp cấu trúc hai chiều của phổ thời gian–tần số.

Ba phép biến đổi chính của SpecAugment bao gồm:

Che theo trục tần số (Frequency Masking).
Một dải tần liên tiếp $[f_0, f_0 + \Delta f)$ được chọn ngẫu nhiên và đặt bằng 0:

$$S(t, f) = 0, \quad \forall f \in [f_0, f_0 + \Delta f). \tag{10}$$

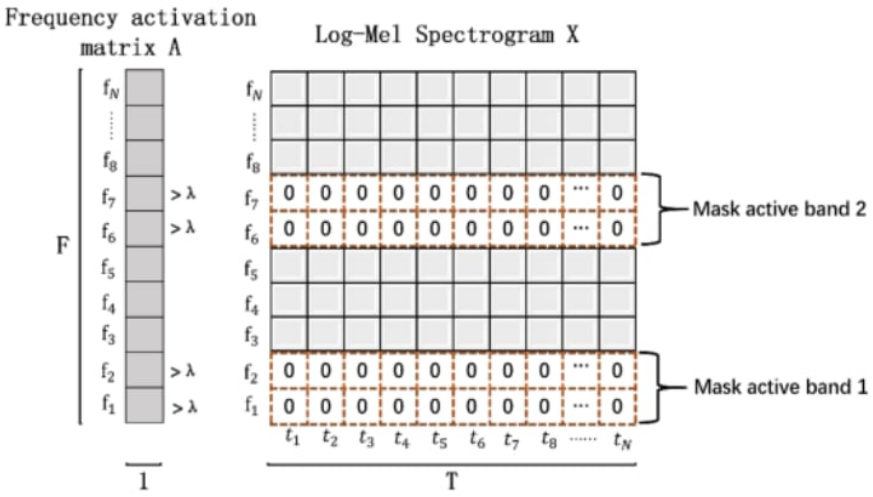
Phép che này mô phỏng sự thiếu hụt thông tin ở một số băng tần, giúp mô hình không phụ thuộc vào một dải tần cố định.

Che theo trục thời gian (Time Masking).
Một khoảng thời gian $[t_0, t_0 + \Delta t)$ được che:

$$S(t, f) = 0, \quad \forall t \in [t_0, t_0 + \Delta t). \tag{11}$$

Điều này giúp mô hình học cách nhận dạng sự kiện âm thanh ngay cả khi một phần tín hiệu bị mất hoặc bị che khuất.

Trong đồ án này, nhóm tập trung sử dụng hai dạng che theo thời gian và tần số, do chúng phù hợp trực tiếp với bài toán ESC và không làm biến dạng nhãn của dữ liệu.



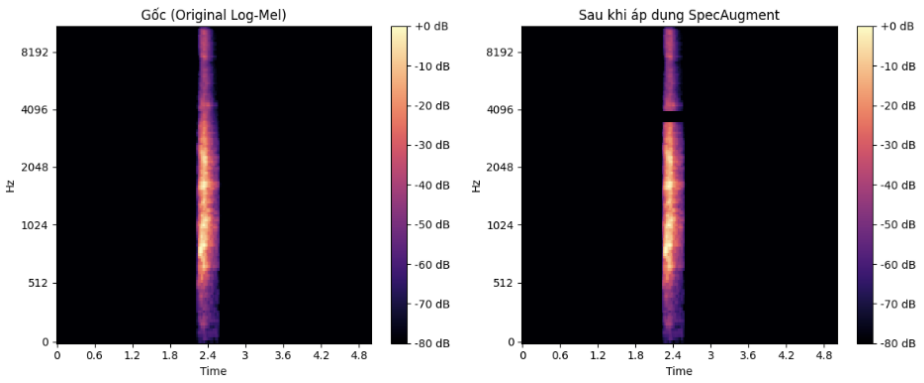
Hình 6: Minh họa nguyên lý che băng tần trong SpecAugment.

3.5.2 Áp dụng SpecAugment cho biểu diễn 3 kênh

Khi kết hợp với biểu diễn 3 kênh gồm Log-Mel Spectrogram, Delta và Delta-Delta, SpecAugment được áp dụng **đồng thời trên cả ba kênh** nhằm đảm bảo tính nhất quán về mặt thời gian–tần số. Cụ thể, cùng một mặt nạ (mask) theo thời gian hoặc tần số được sử dụng cho cả ba kênh:

$$(S,\ \Delta S,\ \Delta^2 S).$$

Cách áp dụng này giúp mô hình học được mối quan hệ tương đối giữa đặc trưng tĩnh và đặc trưng động ngay cả khi một phần thông tin bị che, thay vì làm nhiễu ngẫu nhiên từng kênh riêng lẻ.



Hình 7: Minh họa Log-Mel Spectrogram trước và sau khi áp dụng SpecAugment. Các vùng bị che theo trục thời gian và tần số buộc mô hình học các đặc trưng tổng quát hơn.

3.5.3 Lợi ích của SpecAugment trong bài toán ESC

Việc sử dụng SpecAugment mang lại nhiều lợi ích cho bài toán phân loại âm thanh môi trường:

- **Giảm overfitting:** Mô hình không thể ghi nhớ chi tiết cục bộ của phổ, mà phải học các đặc trưng ổn định hơn.
- **Tăng tính bất biến theo thời gian và tần số:** Phù hợp với bản chất không ổn định của âm thanh môi trường.
- **Không làm thay đổi nhãn dữ liệu:** Khác với một số kỹ thuật augmentation trong miền thời gian có thể làm méo nghĩa của sự kiện âm thanh.
- **Tương thích tự nhiên với CNN:** SpecAugment hoạt động trực tiếp trên "ảnh phổ", phù hợp với kiến trúc tích chập.

Nhờ đó, SpecAugment đóng vai trò như một thành phần quan trọng giúp mô hình CNN kết hợp Log-Mel Spectrogram và các đặc trưng động đạt được hiệu năng ổn định hơn trên tập kiểm tra.

3.6 Tối ưu hóa và chiến lược huấn luyện

Việc lựa chọn chiến lược tối ưu hóa và huấn luyện phù hợp đóng vai trò quan trọng trong bối cảnh tập dữ liệu ESC-50 có quy mô tương đối nhỏ, trong khi mô hình CNN có số lượng tham số lớn. Do đó, nghiên cứu tập trung vào các thiết lập nhằm đảm bảo quá trình huấn luyện hội tụ ổn định, đồng thời hạn chế hiện tượng quá khớp (*overfitting*).

3.6.1 Bộ tối ưu và lịch học

Mô hình CNN được huấn luyện bằng bộ tối ưu AdamW, một biến thể của Adam có bổ sung cơ chế *weight decay*. Cơ chế này giúp kiểm soát độ lớn của trọng số trong quá

trình cập nhật tham số, từ đó cải thiện khả năng tổng quát hóa của mô hình so với Adam tiêu chuẩn.

Tốc độ học được điều chỉnh thông qua chiến lược *Cosine Learning Rate Decay*, cho phép tốc độ học giảm dần theo dạng hàm cosine trong suốt quá trình huấn luyện. Chiến lược này mang lại các lợi ích sau:

- Giúp mô hình học nhanh trong các epoch đầu, khi tham số còn ở xa nghiệm tối ưu;
- Giảm dần tốc độ học ở các giai đoạn sau để tinh chỉnh tham số một cách ổn định;
- Tránh dao động mạnh quanh nghiệm tối ưu so với việc sử dụng tốc độ học cố định.

3.6.2 Chiến lược giám sát huấn luyện

Để hạn chế hiện tượng học quá khớp, nghiên cứu áp dụng chiến lược *Early Stopping* dựa trên hiệu năng của tập xác thực. Quá trình huấn luyện được dừng lại khi độ chính xác trên tập xác thực không còn cải thiện sau một số epoch liên tiếp.

Bên cạnh đó, mô hình tốt nhất được lưu lại thông qua cơ chế *model checkpointing*, dựa trên tiêu chí độ chính xác của tập xác thực. Cách tiếp cận này đảm bảo rằng mô hình được sử dụng cho quá trình đánh giá là mô hình có khả năng tổng quát hóa tốt nhất trên dữ liệu chưa từng thấy.

3.6.3 Tổng kết chiến lược huấn luyện

Sự kết hợp giữa bộ tối ưu AdamW, chiến lược Cosine Learning Rate Decay, cơ chế Early Stopping, model checkpointing và kỹ thuật tăng cường dữ liệu SpecAugment tạo thành một chiến lược huấn luyện đồng bộ. Chiến lược này giúp mô hình CNN đạt được hiệu năng ổn định và khả năng tổng quát hóa tốt trong điều kiện dữ liệu huấn luyện hạn chế của bài toán phân loại âm thanh môi trường.

4 Giải thích mô hình bằng Grad-CAM

Mặc dù mô hình CNN đạt được kết quả phân loại khả quan trên bài toán phân loại âm thanh môi trường, các mô hình học sâu thường bị xem là "hộp đen" do khó diễn giải quá trình ra quyết định. Trong các ứng dụng thực tế, việc hiểu được mô hình đang dựa vào *vùng đặc trưng nào* trên biểu diễn thời gian–tần số là yếu tố quan trọng nhằm tăng độ tin cậy, hỗ trợ kiểm chứng và giúp phân tích nguyên nhân của các trường hợp dự đoán sai.

Do đó, nghiên cứu áp dụng phương pháp **Gradient-weighted Class Activation Mapping (Grad-CAM)** để trực quan hóa những vùng thời gian–tần số có ảnh hưởng lớn nhất đến quyết định của mô hình CNN trên Log-Mel Spectrogram ba kênh.

4.1 Nguyên lý của Grad-CAM

Grad-CAM là phương pháp giải thích mô hình dựa trên gradient, thường được áp dụng cho các kiến trúc mạng nơ-ron tích chập (CNN). Ý tưởng cốt lõi là sử dụng gradient của điểm số lớp mục tiêu đối với các bản đồ đặc trưng (*feature maps*) tại tầng tích chập cuối cùng, từ đó suy ra mức độ "quan trọng" của từng feature map đối với dự đoán của lớp đó.

Gọi A^k là feature map thứ k tại tầng tích chập cuối cùng và y^c là điểm số (logit) hoặc xác suất tương ứng với lớp c. Trọng số tầm quan trọng α_k^c được tính bằng trung bình toàn cục của gradient:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \tag{12}$$

trong đó i, j lần lượt là chỉ số theo hai chiều (thời gian–tần số) của feature map, và Z là số phần tử của feature map (hệ số chuẩn hóa).

Bản đồ kích hoạt lớp (class activation map) của Grad-CAM cho lớp c được tính như sau:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k \right), \tag{13}$$

trong đó hàm ReLU chỉ giữ các đóng góp dương. Cuối cùng, $L_{\text{Grad-CAM}}^c$ được nội suy (upsample) về cùng kích thước với biểu diễn đầu vào để có thể chồng (overlay) lên Log-Mel Spectrogram.

4.2 Áp dụng Grad-CAM cho phân loại âm thanh môi trường

Trong nghiên cứu này, mô hình CNN nhận đầu vào là biểu diễn phổ thời gian–tần số dưới dạng tensor ba kênh $(S, \Delta S, \Delta^2 S)$. Để trực quan hóa quá trình ra quyết định, Grad-CAM được tính trên tầng tích chập cuối cùng và được chồng lên biểu diễn spectrogram nhằm xác định:

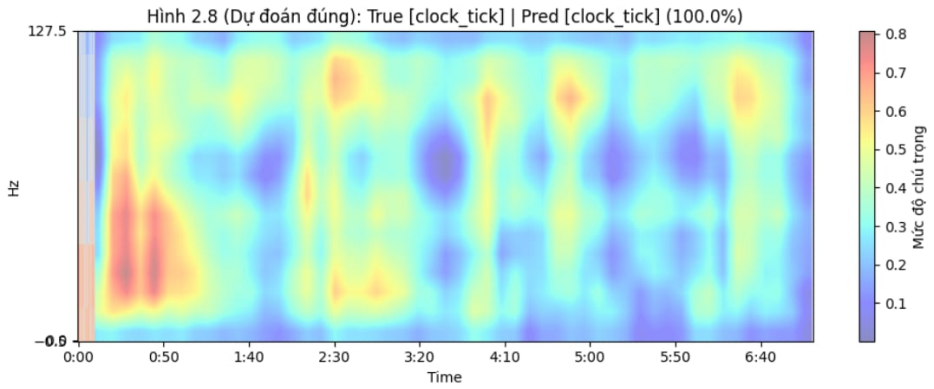
- Các vùng thời gian–tần số mà mô hình tập trung khi dự đoán đúng;
- Các vùng gây nhầm lẫn trong các trường hợp dự đoán sai;
- Mức độ mô hình có dựa vào nhiễu nền hay không.

Việc sử dụng Grad-CAM không nhằm thay thế đánh giá định lượng (accuracy), mà bổ sung một tầng kiểm chứng định tính giúp giải thích tại sao mô hình đạt (hoặc chưa đạt) hiệu năng kỳ vọng trên các lớp âm thanh có tính chất vật lý tương đồng.

4.3 Phân tích định tính bằng Grad-CAM

4.3.1 Ví dụ dự đoán đúng

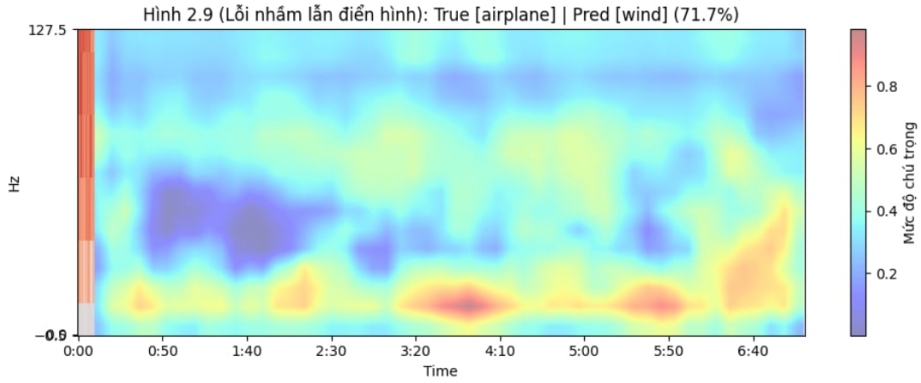
Hình 8 minh họa một trường hợp dự đoán đúng với lớp *clock_tick*. Bản đồ Grad-CAM cho thấy mô hình tập trung vào các vùng năng lượng ngắn và lặp lại theo thời gian, phù hợp với bản chất "tích tắc" đặc trưng của âm thanh đồng hồ. Điều này cho thấy mô hình đang khai thác các dấu hiệu mang tính phân biệt, thay vì học theo nhiễu nền ngẫu nhiên.



Hình 8: Ví dụ dự đoán đúng với Grad-CAM.

4.3.2 Ví dụ lỗi nhầm lẫn điển hình

Hình 9 minh họa một lỗi nhầm lẫn điển hình, trong đó mẫu âm thanh thuộc lớp *airplane* nhưng bị dự đoán nhầm thành *wind*. Quan sát bản đồ Grad-CAM cho thấy mô hình tập trung vào các dải năng lượng thấp–trung kéo dài theo thời gian—đây là đặc trưng phổ có thể xuất hiện ở cả hai lớp *airplane* và *wind*. Do đó, lỗi dự đoán không mang tính ngẫu nhiên mà phản ánh sự chồng lấn đặc trưng vật lý giữa các lớp âm thanh trong môi trường thực.



Hình 9: Ví dụ lỗi nhầm lẫn điển hình được phân tích bằng Grad-CAM.

4.3.3 Kết luận từ phân tích Grad-CAM

Từ các ví dụ trên, Grad-CAM cung cấp bằng chứng rằng mô hình CNN thường tập trung vào các vùng thời gian–tần số hợp lý về mặt vật lý khi đưa ra quyết định phân loại. Các lỗi dự đoán chủ yếu xuất phát từ sự tương đồng về phổ năng lượng và động

học giữa các lớp âm thanh, đặc biệt trong các tình huống nhiễu nền mạnh hoặc khi đặc trưng phân biệt không đủ nổi bật.

Nhờ đó, Grad-CAM đóng vai trò như một công cụ kiểm chứng định tính quan trọng, giúp củng cố tính đúng đắn của phương pháp đề xuất và hỗ trợ định hướng các cải tiến trong tương lai (ví dụ: tăng cường dữ liệu, kiến trúc mạnh hơn, hoặc kỹ thuật chú ý theo thời gian–tần số).

5 Kết quả và thảo luận

Chương này trình bày kết quả thực nghiệm của hai hướng tiếp cận: (i) baseline truyền thống dựa trên MFCC kết hợp kNN, và (ii) mô hình CNN trên biểu diễn Log-Mel ba kênh kết hợp các chiến lược huấn luyện/regularization. Bên cạnh các chỉ số định lượng, nhóm sử dụng confusion matrix và Grad-CAM để phân tích định tính, nhằm làm rõ nguyên nhân của các trường hợp dự đoán đúng và các lỗi nhầm lẫn điển hình.

5.1 So sánh hiệu năng

Bảng 1 tổng hợp độ chính xác (accuracy) của các mô hình trên tập kiểm tra. Kết quả cho thấy mô hình CNN trên biểu diễn Log-Mel ba kênh đạt độ chính xác cao hơn baseline MFCC+kNN, phản ánh lợi thế của việc học biểu diễn đặc trưng trực tiếp từ dữ liệu trong không gian thời gian–tần số. Mức cải thiện khoảng 7 điểm phần trăm

Bảng 1: So sánh độ chính xác (Accuracy) trên tập kiểm tra.

Mô hình	**Accuracy (Test)**
Baseline: MFCC + kNN	≈ 0.65
Đề xuất: CNN + Log-Mel 3-channel	≈ 0.72

cho thấy mô hình CNN có khả năng khai thác tốt hơn cấu trúc thời gian–tần số và động học của tín hiệu (thông qua bộ đặc trưng $S, \Delta S, \Delta^2 S$), trong khi kNN chỉ dựa trên khoảng cách trong một không gian đặc trưng cố định.

5.2 So sánh với các nghiên cứu liên quan

Để có cái nhìn khách quan về hiệu năng của mô hình đề xuất, nhóm tiến hành so sánh kết quả đạt được với các nghiên cứu tiêu biểu đã công bố trên cùng bộ dữ liệu ESC-50.

Kết quả thực nghiệm cho thấy mô hình CNN ba kênh (3-channel CNN) sử dụng các đặc trưng Log-Mel, Delta và Delta-Delta đạt độ chính xác 72% trên tập kiểm thử Fold-5. Mặc dù kết quả này vẫn còn một khoảng cách so với mức 84.4% được báo cáo trong nghiên cứu về mô hình TFCNN đăng trên tạp chí *Scientific Reports*, nhóm đánh giá đây là một kết quả khích lệ đối với một hệ thống được xây dựng thủ công từ đầu (hand-made), không sử dụng các kiến trúc phức tạp hoặc kỹ thuật tiền huấn luyện.

Sự chênh lệch hiệu năng khoảng 12% có thể được lý giải thông qua hai nguyên nhân chính, được nhóm rút ra sau khi phân tích các tài liệu tham khảo liên quan:

- **Cơ chế chú ý (Attention Mechanism):** Mô hình TFCNN tích hợp các khối chú ý theo cả miền thời gian và miền tần số, cho phép mạng tập trung vào các vùng thông tin quan trọng trong phổ âm thanh. Trong khi đó, mô hình của nhóm chưa sử dụng cơ chế chú ý, mà tập trung vào việc làm giàu đặc trưng đầu vào thông qua các kênh động học (Delta và Delta-Delta).
- **Kỹ thuật tiền xử lý chuyên sâu:** Nghiên cứu tham chiếu áp dụng thuật toán HPSS (Harmonic–Percussive Source Separation) nhằm tách và làm nổi bật các thành phần hài âm và gõ trong tín hiệu âm thanh trước khi đưa vào huấn luyện. Đây là một kỹ thuật tiền xử lý nâng cao mà nhóm chưa triển khai trong nghiên cứu hiện tại và dự kiến sẽ tích hợp trong các hướng phát triển tiếp theo.

Mặc dù vậy, việc đạt được độ chính xác 72%, tương đương mức cải thiện hơn 7% so với baseline sử dụng đặc trưng MFCC truyền thống, đã khẳng định tính hợp lý của hướng tiếp cận "thị giác hóa âm thanh" mà nhóm theo đuổi. Thay vì chỉ tập trung vào việc tối ưu hóa chỉ số độ chính xác, nhóm chú trọng đến khả năng diễn giải mô hình thông qua kỹ thuật Grad-CAM, nhằm đảm bảo tính minh bạch và giúp làm rõ mối liên hệ giữa các đặc trưng học được và bản chất vật lý của tín hiệu âm thanh.

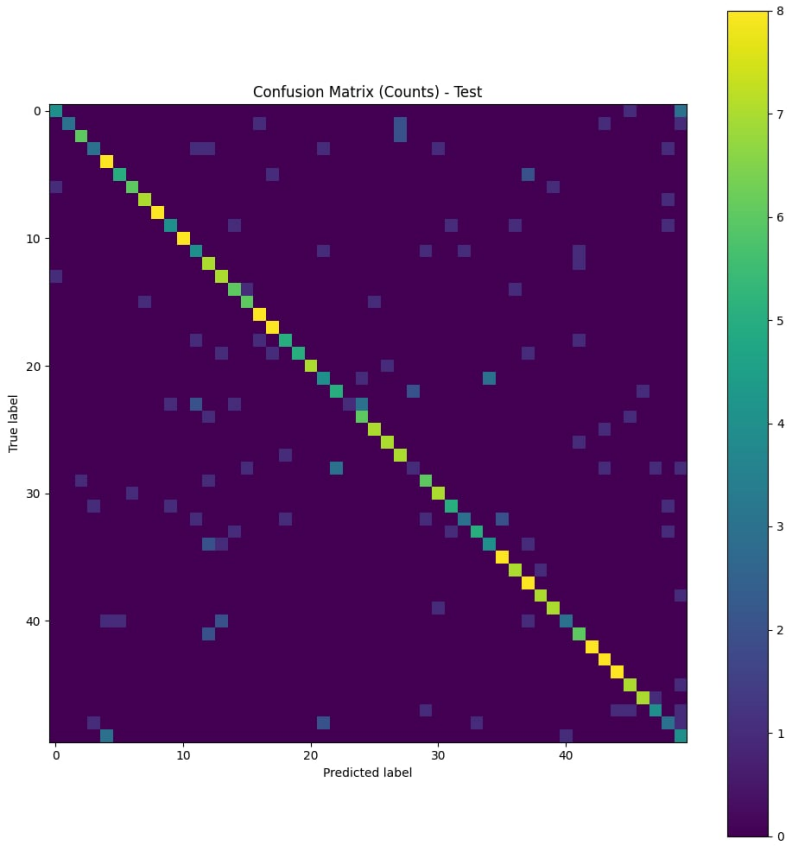
5.3 Trực quan hoá kết quả

Để đánh giá toàn diện hơn, nhóm sử dụng trực quan hoá nhằm quan sát hành vi mô hình và phân bố lỗi:

- **Learning curves:** thể hiện quá trình hội tụ và dấu hiệu overfitting/underfitting.
- **Confusion matrix:** cho biết các lớp nào thường bị nhầm lẫn với nhau.
- **Ví dụ dự đoán đúng/sai:** kết hợp Grad-CAM để kiểm chứng vùng đặc trưng mô hình tập trung.

5.3.1 Confusion matrix trên tập kiểm tra

Hình 10 trình bày confusion matrix (dạng số lượng mẫu) trên tập kiểm tra. Ma trận có xu hướng tập trung dọc đường chéo chính, cho thấy mô hình dự đoán đúng phần lớn mẫu. Tuy nhiên, các điểm sáng ngoài đường chéo thể hiện các cặp lớp bị nhầm lẫn, thường xuất phát từ sự tương đồng về đặc trưng phổ và nền âm (background texture).



Hình 10: Confusion matrix (counts) của mô hình CNN trên tập kiểm tra.

5.3.2 Chỉ số đánh giá theo lớp

Từ confusion matrix, với mỗi lớp c ta có:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \tag{14}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c}, \tag{15}$$

$$\text{F1}_c = \frac{2\,\text{Precision}_c\,\text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}. \tag{16}$$

Để đánh giá mức độ thiên vị giữa các lớp, báo cáo thêm:

$$\text{Macro-F1} = \frac{1}{C}\sum_{c=1}^{C}\text{F1}_c, \quad \text{Weighted-F1} = \sum_{c=1}^{C}\frac{n_c}{N}\text{F1}_c. \tag{17}$$

Macro-F1 nhạy với các lớp khó/ít mẫu, do đó phản ánh công bằng hơn khi so sánh hiệu năng giữa các lớp.

Bảng 2: Tổng hợp các chỉ số đánh giá trên tập kiểm thử (Fold 5)

Chỉ số	Precision	Recall	F1-score	Support
Accuracy		0.7200		400
Macro average	0.7385	0.7200	0.7073	400
Weighted average	0.7385	0.7200	0.7073	400

5.3.3 Ví dụ dự đoán đúng/sai kết hợp Grad-CAM

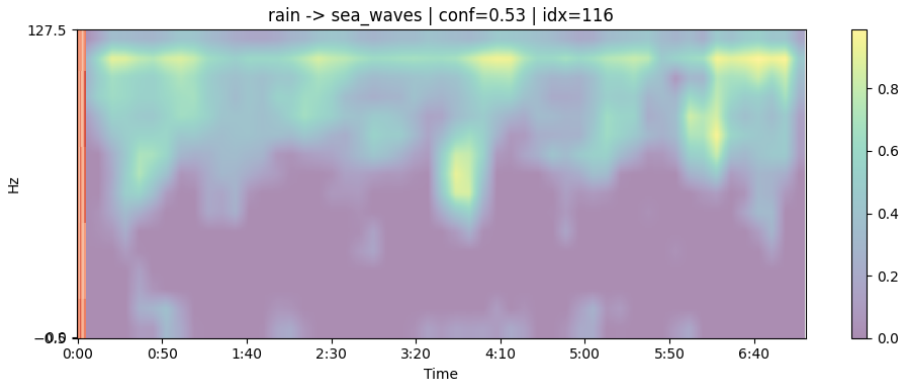
Bên cạnh confusion matrix, nhóm chọn một số trường hợp tiêu biểu để phân tích định tính. Ở Chương 4, Grad-CAM đã được sử dụng để trực quan hóa vùng thời gian–tần số mà mô hình tập trung. Trong chương này, các ví dụ nhầm lẫn sẽ được dùng để giải thích nguyên nhân lỗi theo góc nhìn đặc trưng vật lý.

5.4 Phân tích nhầm lẫn tiêu biểu

Trong bài toán ESC-50, nhiều lớp âm thanh có thể chia sẻ các đặc trưng phổ tương tự nhau (ví dụ: nền nhiễu broadband, năng lượng dải thấp kéo dài, hoặc cấu trúc lặp). Dưới đây là một số lỗi nhầm lẫn điển hình và phân tích tương ứng.

5.4.1 Rain → Sea_waves

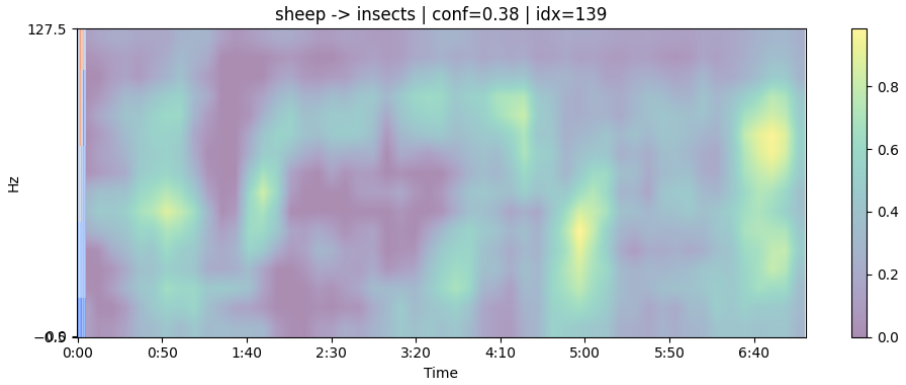
Hình 11 cho thấy một trường hợp mẫu thuộc lớp *rain* nhưng bị mô hình dự đoán nhầm thành *sea_waves* (độ tin cậy ≈ 0.53). Quan sát vùng kích hoạt cho thấy mô hình tập trung vào các vùng năng lượng nền trải dài theo thời gian, đặc biệt ở dải tần thấp–trung. Đây là dạng "texture" phổ liên tục vốn có thể xuất hiện ở cả tiếng mưa và tiếng sóng biển, khiến ranh giới giữa hai lớp trở nên mờ, nhất là trong điều kiện dữ liệu hạn chế.



Hình 11: Lỗi nhầm lẫn tiêu biểu: *rain* bị dự đoán nhầm thành *sea_ waves* (conf ≈ 0.53). Vùng kích hoạt trải dài theo thời gian và tập trung ở dải tần thấp–trung, phản ánh sự tương đồng "texture" phổ giữa hai lớp.

5.4.2 Sheep → Insects

Hình 12 minh họa một trường hợp mẫu thuộc lớp *sheep* nhưng bị dự đoán nhầm thành *insects* (độ tin cậy ≈ 0.38). Ở các mẫu thực tế, tiếng động vật và tiếng côn trùng có thể cùng xuất hiện trong bối cảnh ngoài trời, kèm theo nền âm tự nhiên (gió, lá cây, tiếng xa). Vì vậy, mô hình có xu hướng dựa vào các vùng năng lượng nền và các dải phổ kéo dài, dẫn đến nhầm lẫn giữa các lớp thuộc nhóm "animal/nature" có đặc trưng giao nhau.



Hình 12: Lỗi nhầm lẫn tiêu biểu: *sheep* bị dự đoán nhầm thành *insects* (conf ≈ 0.38). Sự giao thoa nền âm tự nhiên và cấu trúc phổ kéo dài có thể làm giảm khả năng phân biệt giữa các lớp.

5.4.3 Các nhầm lẫn thường gặp khác

Ngoài các ví dụ trên, một số nhầm lẫn thường gặp trong bài toán ESC có thể bao gồm:

- **Insects ↔ Birds:** đều có cấu trúc dải năng lượng ở tần số trung–cao và có tính lặp.
- **Dog ↔ Footsteps:** dễ nhầm khi tiếng bước chân bị nhiễu mạnh hoặc tiếng chó sủa ở xa có biên độ nhỏ.
- **Airplane ↔ Wind:** đều có nền năng lượng kéo dài ở dải thấp–trung (đã phân tích ở Chương 4).

Các nhóm nhầm lẫn này gợi ý rằng hạn chế chính không chỉ đến từ mô hình, mà còn từ **tính chồng lấn đặc trưng vật lý** giữa các lớp âm thanh trong điều kiện môi trường thực.

6 Hướng phát triển

Trong đồ án này, mô hình CNN kết hợp với Log-Mel Spectrogram, các đặc trưng động và SpecAugment đã cho thấy hiệu năng phân loại ổn định đối với bài toán âm thanh môi trường. Tuy nhiên, các kết quả phân tích ở Chương 4 và Chương 5 cũng cho thấy mô hình vẫn gặp khó khăn trong việc phân biệt một số lớp âm thanh có đặc trưng phổ chồng lấp hoặc có cấu trúc thời gian phức tạp.

Một hạn chế chung của CNN tiêu chuẩn là mô hình xử lý mọi vùng thời gian–tần số như nhau, trong khi trên thực tế, không phải mọi phần của phổ đều đóng vai trò quan trọng như nhau cho quá trình phân loại. Điều này đặt ra nhu cầu phát triển các cơ chế giúp mô hình có khả năng tập trung chọn lọc vào các vùng thông tin quan trọng hơn của tín hiệu âm thanh.

6.1 Cơ chế chú ý theo thời gian (Temporal Attention)

Trong âm thanh môi trường, nhiều sự kiện mang tính ngắn hạn và chỉ xuất hiện trong một khoảng thời gian rất nhỏ của toàn bộ tín hiệu. Các cơ chế chú ý theo thời gian (Temporal Attention) được đề xuất nhằm giúp mô hình tự động xác định và nhấn mạnh các khoảng thời gian có chứa thông tin phân biệt quan trọng.

Thay vì xử lý đồng đều toàn bộ trục thời gian, Temporal Attention gán các trọng số khác nhau cho từng khung thời gian, qua đó tăng ảnh hưởng của các đoạn tín hiệu quan trọng và giảm tác động của các đoạn ít thông tin hoặc chứa nhiễu. Điều này đặc biệt hữu ích đối với các âm thanh có dạng bùng phát nhanh như tiếng còi xe, tiếng va chạm hoặc tiếng động đột ngột.

Việc tích hợp cơ chế chú ý theo thời gian vào CNN cho phép mô hình không chỉ học các mẫu phổ cục bộ mà còn khai thác tốt hơn cấu trúc động học của tín hiệu, góp phần cải thiện khả năng phân loại các sự kiện ngắn và phức tạp.

6.2 Cơ chế chú ý theo tần số (Frequency Attention)

Bên cạnh trục thời gian, trục tần số cũng đóng vai trò quan trọng trong việc phân biệt các lớp âm thanh môi trường. Một số lớp âm thanh đặc trưng bởi các dải tần nhất định, trong khi các dải tần khác có thể ít thông tin hoặc chủ yếu chứa nhiễu.

Cơ chế chú ý theo tần số (Frequency Attention) cho phép mô hình tự động học các trọng số cho từng dải tần, từ đó tập trung vào các vùng phổ mang tính phân biệt cao hơn. Khác với các bộ lọc cố định trong quá trình trích xuất đặc trưng, Frequency Attention mang tính thích nghi và có thể thay đổi theo dữ liệu và ngữ cảnh cụ thể.

Việc áp dụng chú ý theo tần số giúp mô hình giảm sự phụ thuộc vào các dải phổ không quan trọng, đồng thời tăng cường khả năng phân biệt giữa các lớp âm thanh có phổ chồng lấp trong một số vùng tần số.

6.3 Kết hợp chú ý thời gian – tần số trong mô hình CNN

Một hướng phát triển tiềm năng là kết hợp đồng thời cơ chế chú ý theo thời gian và theo tần số nhằm khai thác đầy đủ cấu trúc hai chiều của biểu diễn Log-Mel Spectrogram. Trong cách tiếp cận này, mô hình không chỉ học cách tập trung vào các khoảng thời gian quan trọng mà còn xác định được các dải tần mang nhiều thông tin nhất cho từng lớp âm thanh.

Sự kết hợp này đặc biệt phù hợp với bản chất của âm thanh môi trường, nơi thông tin phân biệt thường xuất hiện cục bộ trong cả hai chiều thời gian và tần số. Các nghiên cứu gần đây cho thấy việc tích hợp chú ý thời gian–tần số vào CNN có thể giúp cải thiện đáng kể hiệu năng phân loại, đặc biệt trong các kịch bản dữ liệu phức tạp và nhiễu cao.

6.4 Hướng mở rộng trong tương lai

Ngoài các cơ chế chú ý theo thời gian và tần số, nhiều hướng mở rộng khác cũng có thể được xem xét trong các nghiên cứu tiếp theo. Chẳng hạn, việc sử dụng các kiến trúc học sâu sâu hơn hoặc kết hợp CNN với các mô hình tuần tự như RNN hoặc Transformer có thể giúp mô hình nắm bắt tốt hơn các phụ thuộc dài hạn theo thời gian.

Bên cạnh đó, các chiến lược tăng cường dữ liệu nâng cao, tối ưu siêu tham số hoặc huấn luyện trên tập dữ liệu lớn hơn cũng là những hướng tiềm năng để cải thiện hiệu năng và khả năng tổng quát hóa của hệ thống phân loại âm thanh môi trường.

7 Kết luận

Trong đồ án này, nhóm đã xây dựng một hệ thống phân loại âm thanh môi trường trên tập dữ liệu ESC-50, từ hướng tiếp cận truyền thống (baseline) đến mô hình học sâu dựa trên biểu diễn thời gian–tần số.

Cụ thể, nhóm đã:

- Mô tả và phân tích bài toán phân loại âm thanh môi trường cùng các thách thức đặc thù như nhiễu nền, tính không dừng theo thời gian và sự tương đồng vật lý giữa các lớp.
- Xây dựng pipeline đặc trưng dựa trên Log-Mel Spectrogram, kết hợp Delta và Delta-Delta, tạo thành biểu diễn ba kênh phù hợp với CNN.
- Thiết lập mô hình baseline MFCC + kNN nhằm làm mốc so sánh, từ đó định lượng mức cải thiện khi chuyển sang học sâu.

- Huấn luyện mô hình CNN với các chiến lược tối ưu hóa và regularization nhằm tăng khả năng tổng quát hóa.
- Áp dụng Grad-CAM để giải thích mô hình, cung cấp kiểm chứng định tính về vùng thời gian–tần số mà mô hình tập trung, đồng thời hỗ trợ phân tích nguyên nhân của các lỗi nhầm lẫn điển hình.

Trong tương lai, để tiếp tục cải thiện hiệu năng, nhóm định hướng mở rộng theo các hướng: (i) tăng cường dữ liệu mạnh hơn, (ii) sử dụng mô hình/kiến trúc mạnh hơn hoặc pretrained (transfer learning), và (iii) khai thác cơ chế chú ý theo thời gian–tần số nhằm tăng khả năng phân biệt giữa các lớp có phổ tương đồng.

A Phụ lục

A.1 Tài liệu tham khảo bổ sung

Các tài liệu trong mục này được sử dụng làm cơ sở lý thuyết và tham khảo kỹ thuật trong quá trình xây dựng baseline và triển khai thực nghiệm, nhưng không được xem là các công trình so sánh trực tiếp trong phần kết quả chính của báo cáo.

Tài liệu

[1] K. J. Piczak, *ESC-50: Dataset for Environmental Sound Classification*. Available: https://github.com/karolpiczak/ESC-50. Accessed: Jan. 2026.

[2] W. Mu, B. Yin, X. Huang, J. Xu, Z. Du, *Environmental sound classification using temporal-frequency attention based convolutional neural network*, Scientific Reports, vol. 11, 2021.

[3] S. Chu, *Environmental sound recognition with time-frequency audio features*, IEEE Transactions on Audio, Speech, and Language Processing, 2009.

[4] N. T. Thành, *Feature Extraction – MFCC cho xử lý tiếng nói*, Viblo Asia. Available: https://viblo.asia/p/feature-extraction-mfcc-cho-xu-ly-tieng-noi-4dbZN2xmZYM. Accessed: Jan. 2026.

[5] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer, 2nd ed., 2009.