Phân loại âm thanh

Sử dụng các mô hình dựa trên CNN

Nhóm 09

Ngày 18 tháng 6 năm 2024

Giới thiệu đề tài

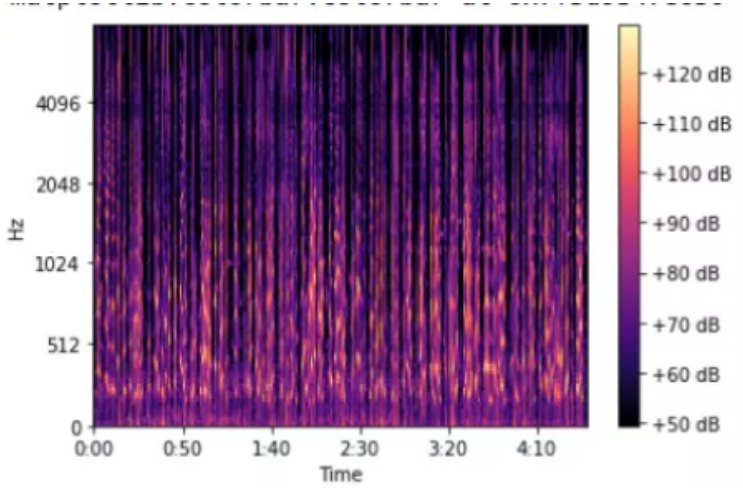
- Phân loại âm thanh là một công việc quan trọng trong lĩnh vực *Học máy* và *Xử lý tín hiệu*, liên quan đến việc phân loại các tín hiệu âm thanh vào các lớp được định nghĩa trước.
- CNN, ban đầu được thiết kế cho xử lý hình ảnh, đã chứng minh sự linh hoạt và tính hiệu quả của nó trong phân tích tín hiệu âm thanh, với khả năng tự động học các hệ thống tính năng không gian thông qua quá trình truyền ngược.
- Việc áp dụng CNN trong phân loại âm thanh đã dẫn đến những cải tiến đáng kể về độ chính xác và hiệu quả so với các phương pháp truyền thống.

Xử lý dữ liệu Audio

- Dữ liệu Audio được lưu trữ dưới dạng tín hiệu âm thanh. Khi được lưu dưới dạng tệp, dữ liệu âm thanh sẽ được nén và khi được load, nó sẽ đại diện bởi một mảng với mỗi phần tử đại diện cho biên độ của âm thanh tại *1/sample_rate* giây. Như vậy, âm thanh đã được biểu diễn bằng một mảng số với một *sample_rate* nhất định.
- Tại một sample, chúng ta cần thêm thông tin về độ cao (tần số) âm thanh để có thể biểu diễn chính xác hơn.
- Vậy, chúng ta đã biểu diễn được âm thanh dưới dạng một bức ảnh mang thông tin về độ cao, biên độ theo thời gian của âm thanh.

Mel-spectrogram

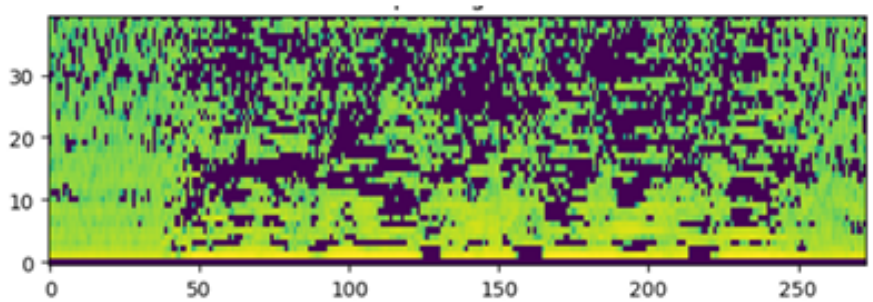
- Spectrogram là biểu đồ mà trong đó, bên cạnh các sample, dữ liệu cũng được chia thành các khoảng có tần số khác nhau và màu sắc của chúng đại diện cho biên độ của âm thanh tại thời điểm và tần số đó.
- Mel-spectrogram là một Spectrogram sau khi đã được chuẩn hoá *mel* và chuẩn hoá *biên độ*.



Hình: Mel-spectrogram

LFCC

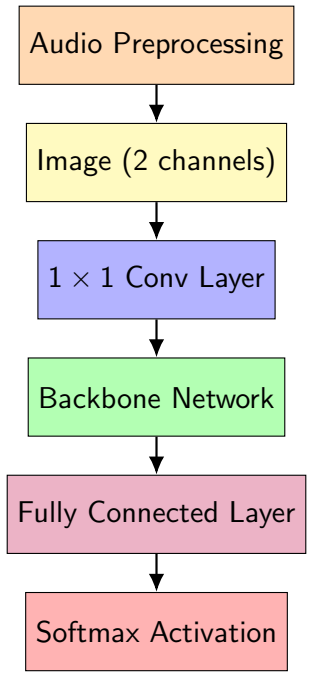
LFCC (linear-frequency cepstrum coefficients) là phép biển đổi âm thanh về các miền tần số với khoảng chia nhỏ, áp dụng mel-frequency filterbanks để mô phỏng cách tai người nghe âm thanh, sau đó tính toán logarit lên các giá trị biên độ, áp dụng biến đổi Fourier ngược lên giá trị đó và cuối cùng là chuẩn hoá.



Hình: LFCC

Mô hình

Hầu hết các mô hình CNN đều được xây dựng dựa trên kiến trúc gồm *backbone (xương sống)* và *lớp kết nối đầy đủ*. Lớp kết nối đầy đủ này sẽ mang các thuộc tính đã được trích xuất từ backbone để phục vụ cho các tác vụ khác nhau như phân loại, hồi quy, xác định, ...



Các độ đo

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

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

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

$$\text{WeightedAverageMetric} = \sum_{i=1}^{C} w_i \times \text{Metric}_i$$

Thực nghiệm

Sử dụng tập dữ liệu **UrbanSound8K**.

- Tập dữ liệu này chứa 8732 đoạn trích âm thanh (không quá 4 giây) định dạng **WAV**;
- Các âm thanh được gán nhãn từ 10 lớp (*air_conditioner*, *car_horn*, *children_playing*, *dog_bark*, *drilling*, *enginge_idling*, *gun_shot*, *jackhammer*, *siren* và *street_music*) và các lớp này có phân bố không đồng đều.

Cài đặt

Các cài đặt cho quá trình đào tạo ở các model giống nhau.

Mục	**Giá trị**
Criterion	Cross Entropy Loss
Optimizer	Adam
Learning rate	0.001
Scheduler	OneCircleLR
Epochs	30
Batch_size	16

Kết quả

Backbone	Mel	LFCC	Precision	Recall	Accuracy	Params	GFlops
Efficientnet V2 Small	x		0.9527	0.9525	0.9525	21.5M	8.37
		x	0.9036	0.9026	0.9026		
ConvNext Tiny	x		0.9445	0.9433	0.9433	28.6M	4.46
		x	0.9278	0.9280	0.9280		
Resnet-18	x		0.9637	0.9633	0.9633	11.7M	1.81
		x	***0.9491***	***0.9490***	***0.9490***		
MobileNet V3 Small	x		0.9200	0.9198	0.9198	2.06M	0.06
		x	0.8986	0.8981	0.8981		
MobileNet V2	x		0.9315	0.9313	0.9313	3.5M	0.3
		x	0.9025	0.9026	0.9026		
Alexnet	x		0.8485	0.8465	0.8465	61.1M	0.71
		x	0.8044	0.8036	0.8036		
GoogleNet	x		**0.9674**	**0.9674**	**0.9674**	6.6M	1.5
		x	0.9388	0.9381	0.9381		
DenseNet-121	x		0.9530	0.9525	0.9525	8.05M	2.83
		x	0.9449	0.9444	0.9444		
ShuffleNet V2 x1.0	x		0.9320	0.9307	0.9307	2.3M	0.14
		x	0.8973	0.8969	0.8969		

(Kết quả tốt nhất được **tô đậm**, kết quả tốt thứ hai được **gạch chân** và kết quả tốt nhất với phép biến đổi LFCC được **tô đậm và in nghiêng**)

Kết luận

- Lớp bài toán phân loại âm thanh là một hướng đi tiềm năng giúp nhận biết và tận dụng được dữ liệu âm thanh trong thực tế.
- Thống kê đã chỉ ra rằng phương pháp này sẽ có nhiều ứng dụng trong thực tế, nhất là với các mô hình CNN đang ngày càng được tối ưu về tốc độ, kích thước và độ chính xác.
- Trong tương lai gần, với sự phát triển của các họ mô hình khác như các mô hình chuỗi thời gian, vision-transformer, ... cũng có thể được áp dụng để bài toán đạt hiệu suất tốt hơn.