ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM
KHOA CÔNG NGHỆ THÔNG TIN
BỘ MÔN KHOA HỌC MÁY TÍNH

Nhập môn lập trình điều khiển thiết bị thông minh

Hệ thống nhận diện từ khóa

Giảng viên lý thuyết

TS. Nguyễn Đức Hoàng Hạ
ThS. Đỗ Thị Thanh Hà

Sinh viên thực hiện

Nguyễn Thị Thu Hằng - 18120027

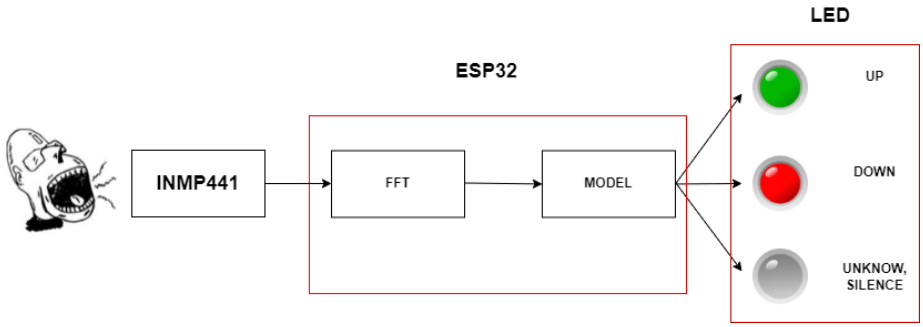
Tháng 1 năm 2022

Mục lục

1 Tổng quan về dự án

Dự án xây dựng một hệ thống có thể nhận biết được các từ khóa một cách cơ bản (ở đây chỉ đơn giản gồm các từ khóa "up", "down") và kết quả sẽ được thể hiện qua hệ thống LED.

Mô tả tổng quát hoạt động của dự án:Đầu tiên hệ thống sẽ nhận tín hiệu âm thanh từ thiết bị INMP441 và tín hiệu sẽ được truyền thời ESP32. Tại đây, ESP32 sẽ biến đổi tín hiệu thu được thành dạng dữ liệu phù hợp thông qua phương thức FFT và cho dữ liệu qua một mô hình đã được thiết lập sẵn để phần loại lớp. Đầu ra của kết quả phân loại sẽ được thể hiện qua đèn LED, cụ thể khi kết quả dự đoán là lớp "up" thì đèn LED màu xanh lá sẽ sáng, và nếu là lớp "down" thì đèn LED mà xanh dương sẽ sáng.



Hình 1: Cách hoạt động của hệ thống

2 Thu thập dữ liệu

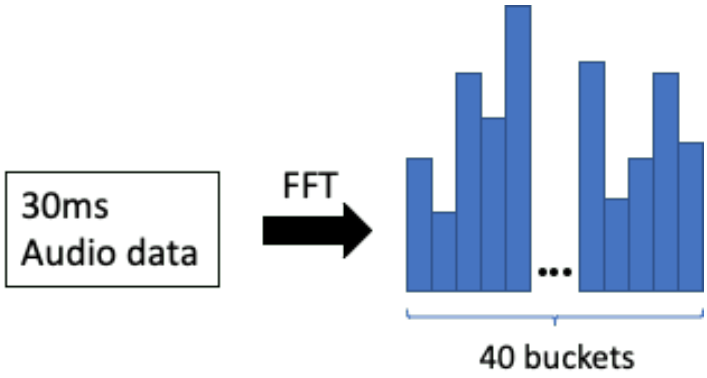
Dữ liệu được em sử dụng để thực hiện đồ án cuối kì này là bộ dữ liệu có sẵn của tensorflow có tên là mini_speech_commands là một phiên bản nhỏ hơn của tập dữ liệu Speech Commands. Tập dữ liệu gốc bao gồm hơn 105.000 tệp âm thanh ở định dạng tệp âm thanh WAV (Dạng sóng) của những người nói 35 từ khác nhau. Dữ liệu này được Google thu thập và phát hành theo giấy phép CC BY. Bộ dữ liệu mini_speech_command gồm 8 trường: no , yes , down , go , left , up , right và stop.

Trong đồ án cuối kì, em chỉ trích chọn ra 2 trường là "up" và "down" để thực hiện

3 Cách nhúng dữ liệu

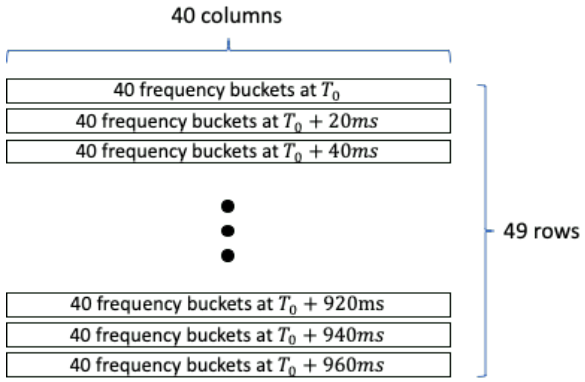
Dữ liệu thuộc tập dữ liệu và dữ liệu được thu từ INMP441 đều được chuyển đổi thành dạng ma trận thông qua kĩ thuật nhúng dữ liệu âm thành bằng các biến đổi FFT (Fast Fourier Transformation). "Biến đổi Fourier nhanh" chuyển đổi một tín hiệu thành các thành phần phổ riêng lẻ và do đó cung cấp thông tin tần số về tín hiệu.

Thay vì sử dụng trực tiếp từng dữ liệu âm thanh dài một giây để đào tạo mô hình ML, chúng tôi đã chuyển đổi từng dữ liệu thành một biểu đồ quang phổ. Để tạo mỗi biểu đồ quang phổ, chúng tôi đã tạo 49 phần của đoạn âm thanh 30 ms từ dữ liệu âm thanh dài một giây, trượt cửa sổ 30 ms đi 20 ms. Sau đó, chúng tôi chuyển đổi mỗi đoạn âm thanh 30 ms thành 40 nhóm tần số. [**?**]



Hình 2: Chuyển từng đoạn âm thanh có độ dài 30s thành 40 nhóm tần số

Bằng cách đó ta sẽ thu được một ma trận đại diện cho đoạn âm thanh đó với kích thước là 40x49



Hình 3: Mỗi đoạn âm thành sau khi được nhúng thành ma trận 40x49

4 Kiến trúc mạng

Trong đồ án này, kiến trúc mạng được em sử đụng là Convolutional Neural Network(CNN). Với một lớp CNN, 2 lớp MLP.Mô hình này thực tế không tạo ra một mô hình có độ chính xác cao (theo nghiên cứu mô hình học sâu thì có độ sâu càng sâu càng tốt), nhưng chúng ta vẫn phải chấp nhận việc đó vì khi đưa mô hình lên các thiết bị điện thử thông minh có dung lượng hạn chế, việc đưa một mô hình quá đồ sộ sẽ gây phản tác dụng. Ngoài ra, mô hình sử dụng đầu vào bằng giọng nói được xử lý trước do đó việc tận dụng một mô hình đơn giản hơn để có kết quả chính xác là phù hợp.

Model thực hiện nhiệm vụ chính là phân loại âm thành thành các lớp phù hợp. Cụ thể ở đây là các lớp "yes", "no", "unknown", "silence".

5 Thiết kết phần cứng

Kết nối ESP32 với INMP441 theo các thông số sau:

ESP32 Pins	INMP411
VCC	VDD
GND	GND
33	SD
26	SCK
25	WS

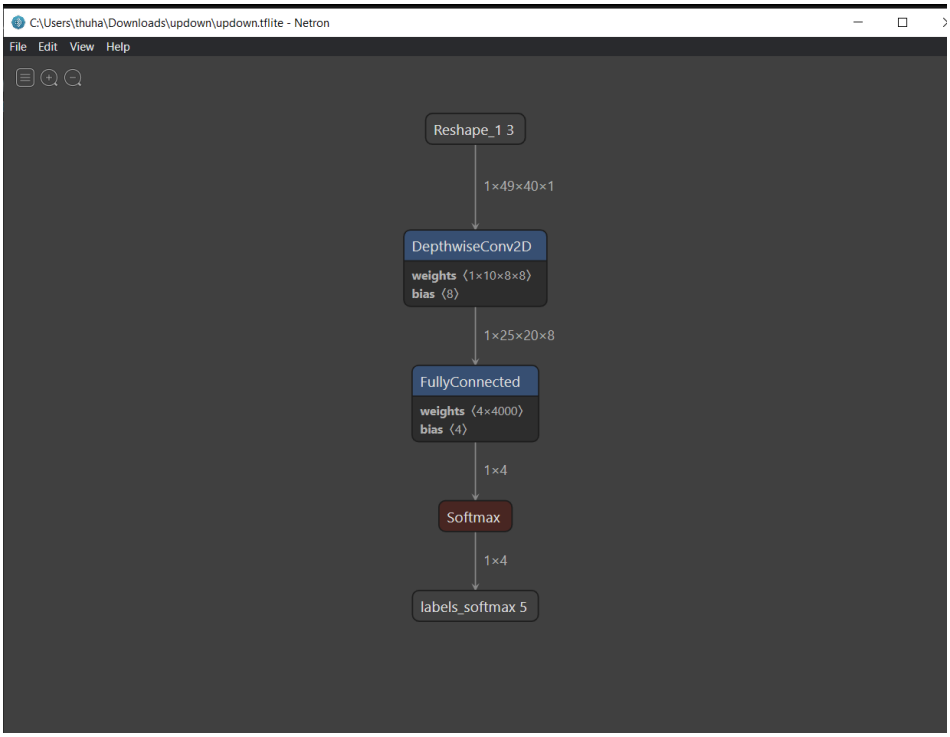
Kết nối ESP32 với LED

ESP32 Pins	LED
23	LED1(+)
23	LED2(+)
GND	GND 2 LED

6 Lập trình thực hiện dự án

Đầu tiên, em dùng ví dụ về nhận biết từ khóa đã có sẵn trong TensorFlowLite_ESP32 để làm sườn và sau đó thực hiện custom lại code để có thể đoán được từ khóa khác.

Sau đó, em sử dụng mã code của sẵn trên nền tảng colab được cung cấp bởi tensorflow để thực hiện training mô hình [**?**] và thu được một mô hình dưới dạng file .tflite như sau.



Hình 4

Bước tiếp, theo em sẽ chuyển updown.tflite sang dạng mã nguồn c bằng cứ pháp.

$$xxd - iupdown.tflite > updown.cc$$

Và cuối cùng là sẽ chỉnh sửa lại code mẫu của thư viện TensorFlowLite_ESP32, tại các file sau:

- micro_model_setting.cc chỉnh sửa tên class cho phù hợp

```
#include "micro_model_settings.h"

const char* kCategoryLabels[kCategoryCount] = {
    "silence",
    "unknown",
    "up",
    "down",
};
```

Hình 5

- tiny_conv_micro_features_model_data.cpp được thay thế bằng mô hình trong tệp updown.cc
- micro_speech_ESP-EYE.cpp thêm 2 thiết lập led trong phần void setup()

```
void setup() {
  pinMode(23, OUTPUT);
  pinMode(22, OUTPUT);
  xQueueAudioWave = xQueueCreate(QueueAudioWaveSize, sizeof(int16_t));
```

Hình 6

- command_responder.cpp để thêm khối lệnh hiện thị output thông qua LED

```
if (strcmp(found_command,"up")==0)
{
  digitalWrite(23, HIGH);
  delay(100);
  digitalWrite(23, LOW);
}
else
{
  if (strcmp(found_command,"down")==0)
  {
    digitalWrite(22, HIGH);
    delay(100);
    digitalWrite(22, LOW);
  }
}
```

Hình 7

Phần code full tại đây: https://github.com/NT-ThuHang/AIoT_Project_Up_Down

7 Quá trình thử nghiệm

Quá trình thử nghiệm đơn giản,với INMP441 đã được hàn chì chân và chưa được hàn chì. Khi chưa hàn chì chân INMP441 vẫn thực hiện khá tốt nhiệm vụ thu tín hiệu âm thanh của mình nhưng đa số các lớp phân loại bị sai và nhiễu cao(tức lớp silence và unknown xuất hiện nhiều khi mình vẫn phát âm down up đúng). Sau khi đã được hàn chì, việc thu tín hiệu âm thanh của INMP441 rất tốt, mô hình hoạt động ổn, sai lệch không nhiều.

8 Tổng kết

Đồ án hoàn thiện và đáp ứng đủ yêu cầu của giáo viên đặt ra. Tuy nhiên, hướng phát triển của hệ thống còn dài và theo thực tế thì hiện chưa áp dụng được. Tương lai để có thể sử dụng được trong thực tế, mô hình cần phải linh hoạt nhận biết thêm nhiều từ khóa hơn, có thể thay thế thiết bị thu thập tín hiệu âm thanh để có hiệu suất tốt hơn. Không chỉ vậy về phần lập trình có thể thay thế việc nhúng dữ liệu và mô hình phân lớp bằng các mô hình SOTA dành cho thiết bị nhúng. Đa số các bài báo khoa học thướng hướng tới việc triển khai các thuật toán SOTA ứng dụng trong nhiều lĩnh vực tuy nhiên với các thiết bị nhúng như thể này vẫn còn hạn chế, đây sẽ là hướng phát triển có thể theo đuổi trong tương lai.