ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

Final Project

WAKE WORD DETECTION

Trình bày: Chung Kim Khánh (19127644)
Giáo viên hướng dẫn: Nguyễn Đức Hoàng Hạ
Môn: AIoT

Bài toán Trigger word hay Wakeup word

Trigger word là một *tín hiệu kích hoạt* thiết bị bằng giọng nói giống như "Ok Google" - "Hey Google" trong trợ lý ảo của Google, "Alexa" trợ lý ảo Alexa của Amazon và “Hey Siri" của Apple.

Động lực nghiên cứu

Ý nghĩa khoa học

- Giao tiếp là một hoạt động đóng vai trò rất quan trọng trong cuộc sống hàng ngày của chúng ta (đặc biệt là giao tiếp bằng âm thanh).
- Phát triển và mở rộng bài toán nhận diện/phân tích/xử lý giọng nói âm thanh -> Thúc đẩy phát triển công nghệ hoá.
- Ứng dụng và tối ưu hoá các mô hình Deep Learning vào và cải tiến các bài toán về học sâu.

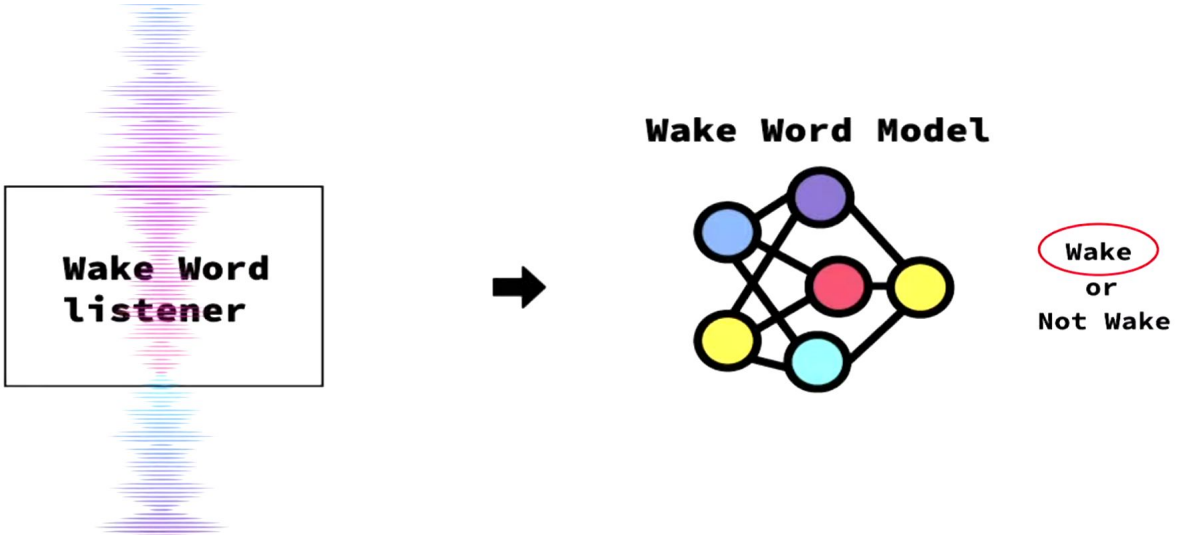
Động lực nghiên cứu

Ứng dụng thực tiễn

- Trigger word như một lời gọi tên, để thiết bị biết là người dùng đang gọi đến thiết bị đó chứ không phải người khác hay thiết bị khác.
- Hạn chế request liên tục các mẫu giọng nói lên trên server.
- Phần nào đó đảm bảo được quyền riêng tư của người dùng.
- Hỗ trợ người lớn tuổi, người khuyết tật tiếp cận với công nghệ.
- Xử lý đa nhiệm

Phát biểu bài toán

- **Input:** Thu âm thanh liên tục
- **Output:** Tín hiệu nhận được từ khoá (có - 1 hoặc không - 0)



Thách thức

_ Gọn nhẹ, chạy được trên các phần cứng cấu hình thấp.

_ Tín hiệu gọi cần gắn ngọn, dễ nói -> đặc trưng, tránh bắt gặp trong câu nói hằng ngày.

_ Độ chính xác cao, detect được giọng nam nữ, các độ tuổi khác nhau, vùng miền khác nhau.

_ Truy tìm từ khoá liên tục

Ví dụ: Giao tiếp liên tục không ngừng

_ Ý nghĩa từ khoá

Ví dụ: Lúc đang nói chuyện thảo luận về một vấn đề gì đó vô tình có cụm từ “Hey Siri". Siri sẽ tự bật lên.

Phát hiện từ đánh thức (Wake Word Detection)

Các bài toán liên quan (Related work)

Tác giả	Phương pháp/Mô hình	Mô tả
Fengpei Ge và Yonghong Yan (2017)	Phân loại hai giai đoạn (two-stage classification) để tích hợp kiến thức ngữ âm và phân loại dựa trên mô hình vào việc phát hiện các từ đánh thức.	Tối ưu rủi ro Bayes ở cấp thấp nhất để đào tạo, tùy chỉnh mạng giải mã để hấp thụ tiếng ồn xung quanh và giọng nói nền.
Gautam Tiwari và Arindam Mandal (2017)	Mô hình kết nối trực tiếp âm thanh thô với DNN để phát hiện từ đánh thức.	Hệ thống nhận dạng giọng nói thông thường thường trích xuất một biểu diễn tính năng nhỏ gọn dựa trên kiến thức trước đó như năng lượng ngân hàng bộ lọc Log-Mel (LFBE).

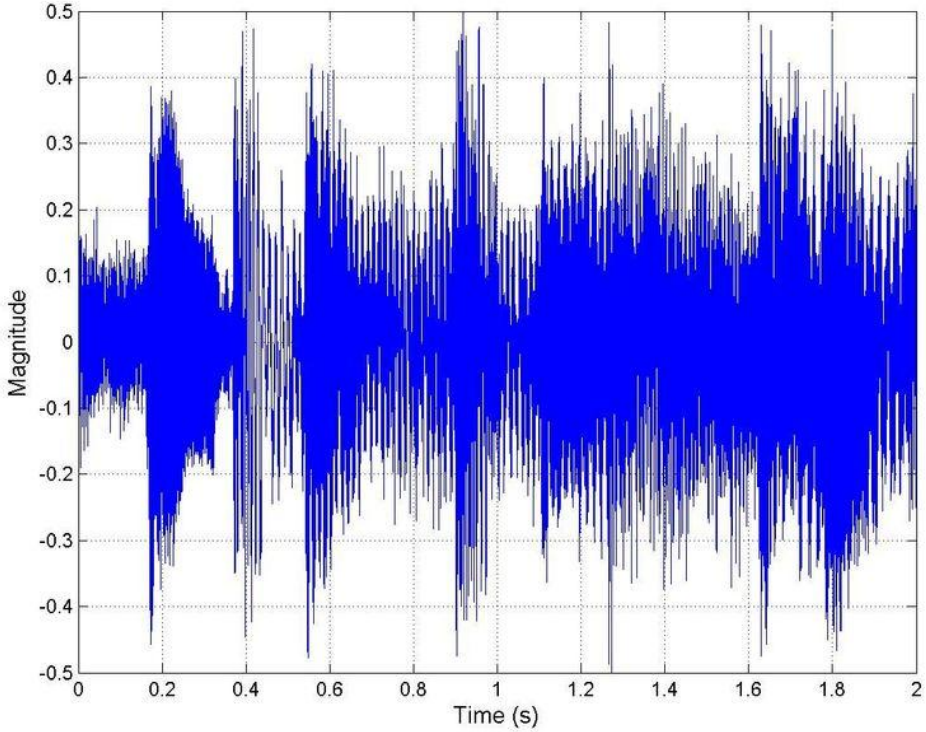
Các bài toán liên quan (Related work)

Tác giả	Phương pháp/Mô hình	Mô tả
Ralf Schluter và Hermann Ney (2016)	Tập trung vào việc đánh giá từng kiến trúc cổ điển để lập mô hình ngôn ngữ nhằm nhận dạng giọng nói từ vựng lớn. Cụ thể là đánh giá mạng lưới đường cao tốc, mạng lưới bên, LSTM và GRU.	Các kết nối thẳng trực tiếp cho phép cả mô hình Recurrent Neural Language và Standalone Feedforward hoàn thiện tốt cấu trúc và cung cấp cải thiện về độ chính xác nhận dạng sau khi nội suy với các mô hình đếm.
Todor Ganchev và Ilyas Potamitis (2005)	Tích hợp sử dụng giọng nói như một phương thức đầu vào tự nhiên để cung cấp cho người dùng khả năng truy cập dễ dàng vào các thiết bị giải trí và thông tin đã được cài đặt trong gia đình.	Dựa trên việc triển khai khối tiền xử lý tín hiệu front-end bao gồm một dãy 8 micrô được kết nối với một soundcard đa kênh và các máy trạm thực hiện tất cả các tác vụ tiền xử lý tín hiệu.

Các bài toán liên quan (Related work)

Tác giả	Phương pháp/Mô hình	Mô tả
Igor Stefanović và Eleonora Nan (2017)	Hai hoạt động triển khai mô-đun phát hiện từ đánh thức, dựa trên các công cụ nhận dạng giọng nói Pocketsphinx và Snowboy.	Các thí nghiệm đã chỉ ra rằng Pocketsphinx có độ chính xác tốt hơn Snowboy, nhưng hiệu suất của nó bị hạn chế trên RPI2.
Shilpa Srivastava, Ashutosh Kumar Singh và Sanjay Kumar Nayak (2020)	Một mô-đun lệnh bằng giọng nói cho phép điều khiển bằng giọng nói trong phòng thí nghiệm tại nhà.	Xây dựng đơn vị nhẹ có thể hoạt động trên các thiết bị nhúng công suất thấp như Raspberry Pi.

Ý tưởng



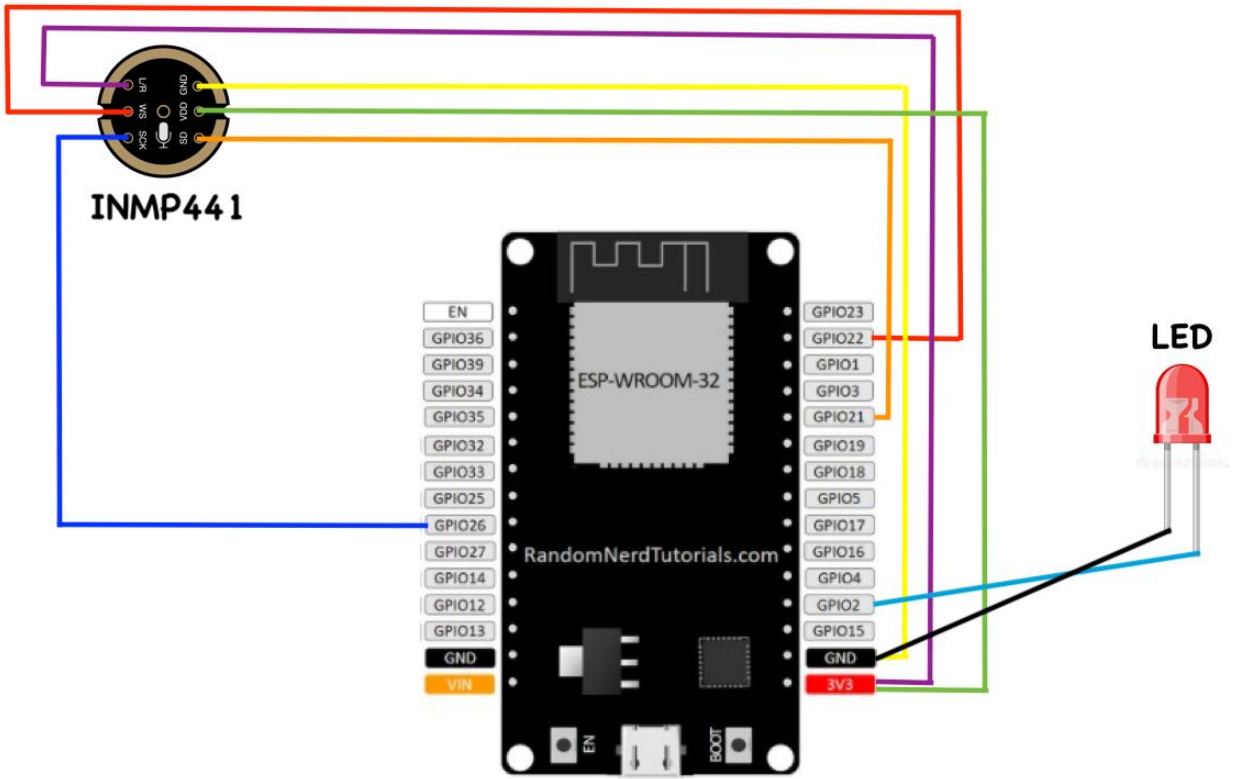
Tách thành các đoạn

Vấn đề quyền riêng tư

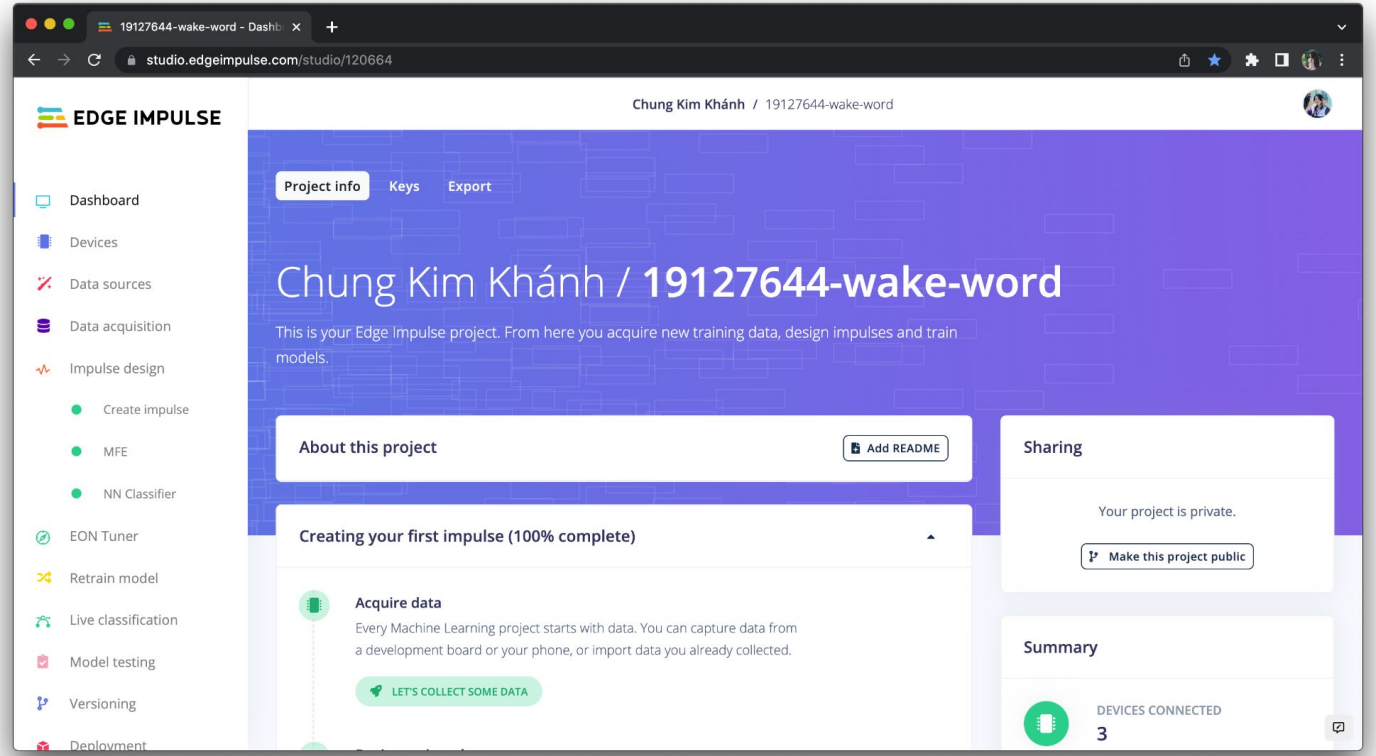
Bài toán khó được đặt ra là liệu thu âm liên tục như thế này có ảnh hưởng gì đến **quyền riêng tư** của cá nhân không?

THỰC NGHIỆM

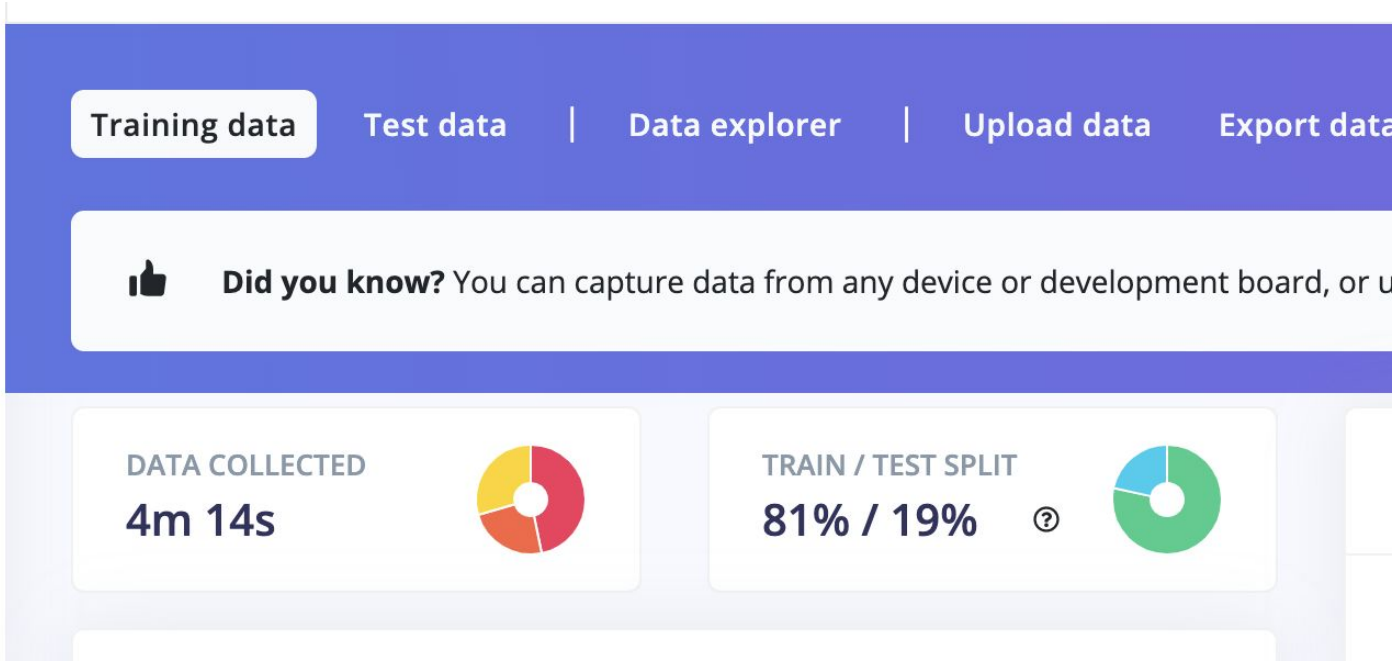
Sơ đồ mạch điện



Huấn luyện mô hình

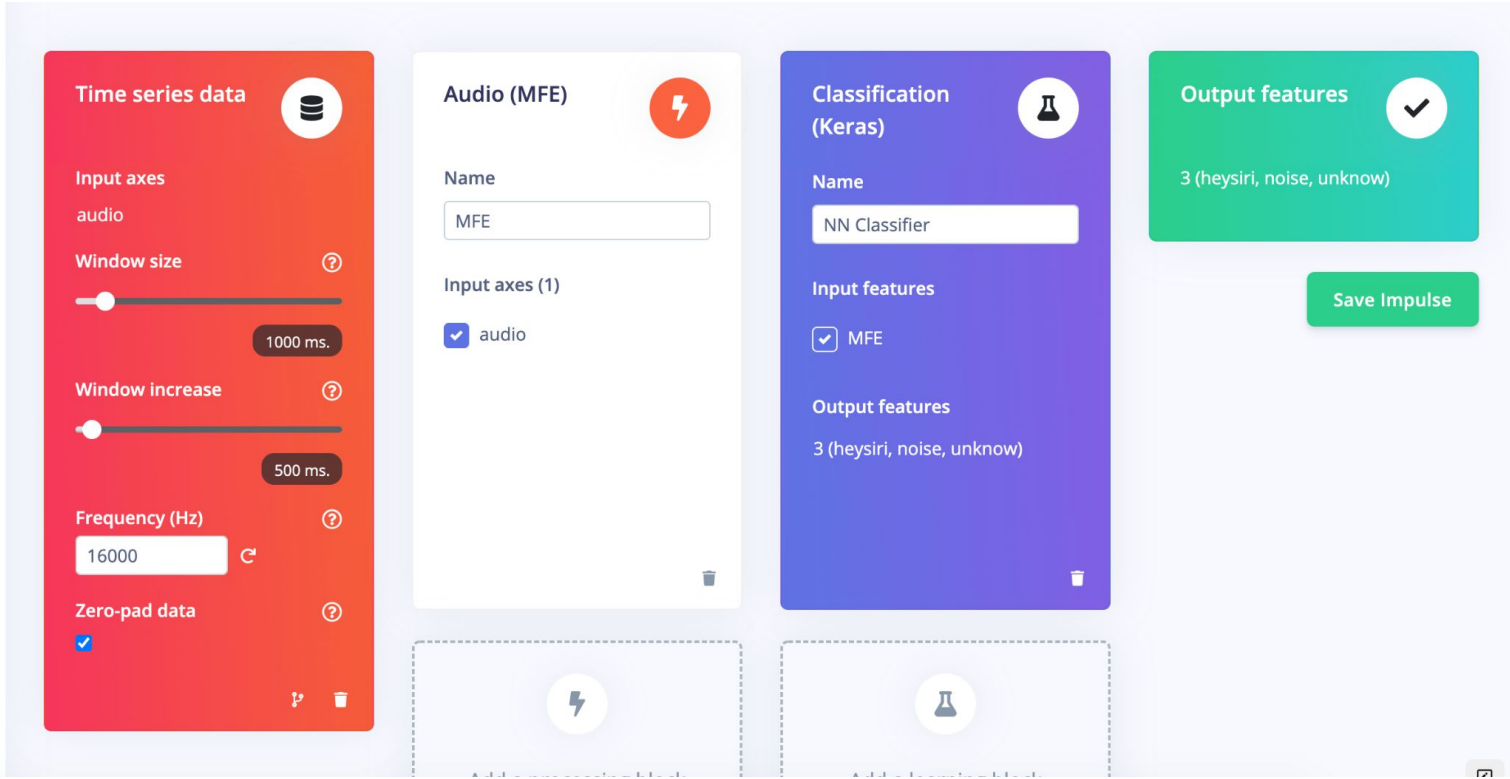


Tập dữ liệu

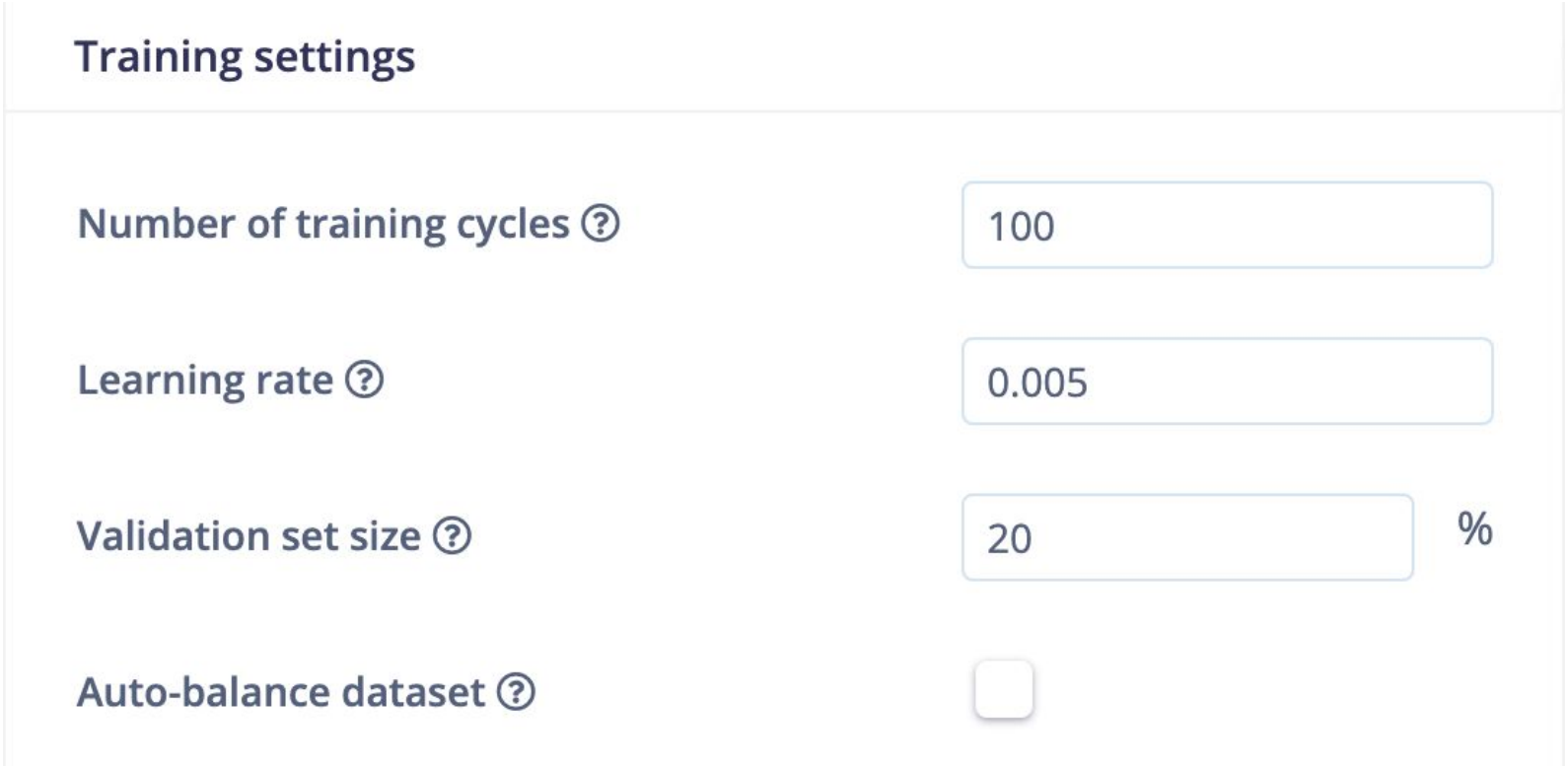


Tập dữ liệu tự thu thập

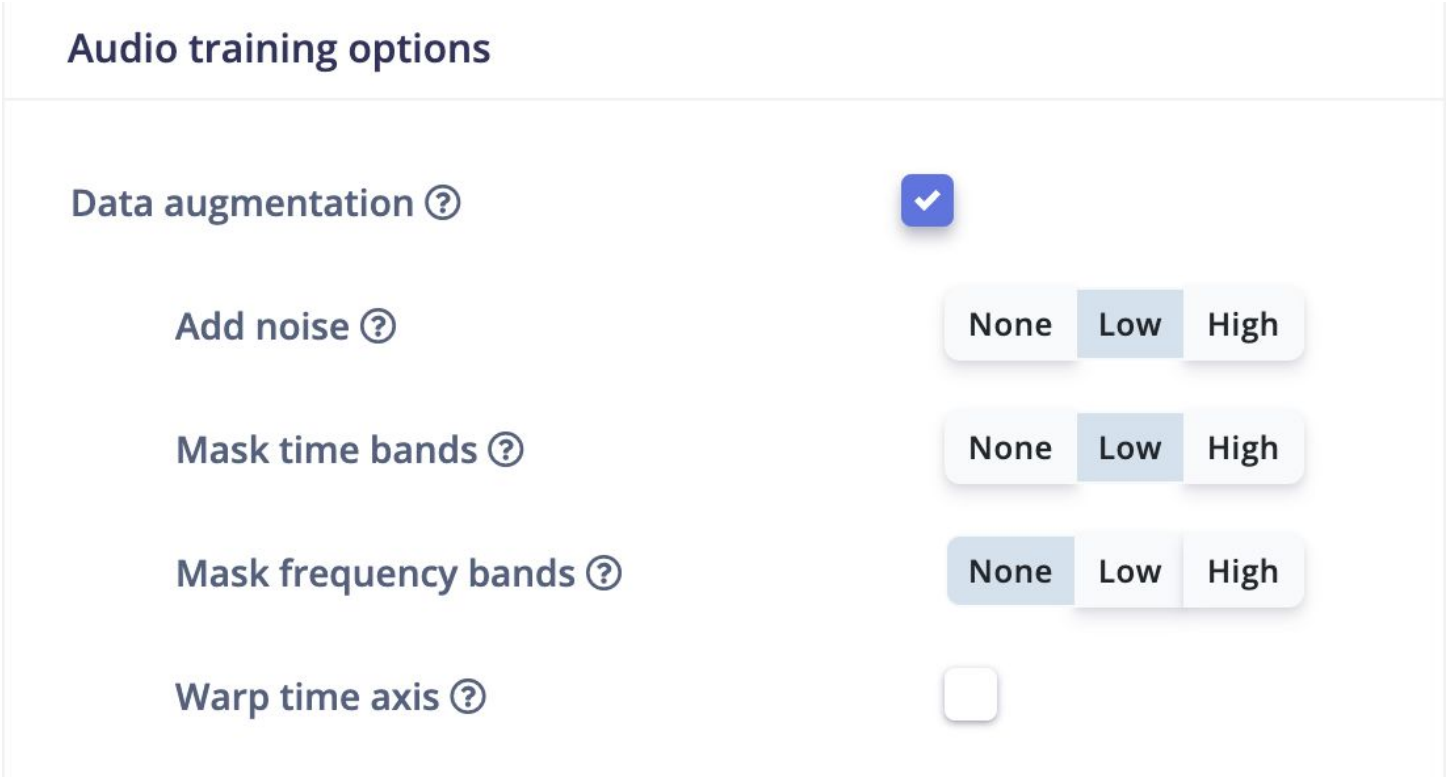
Thiết lập mô hình huấn luyện



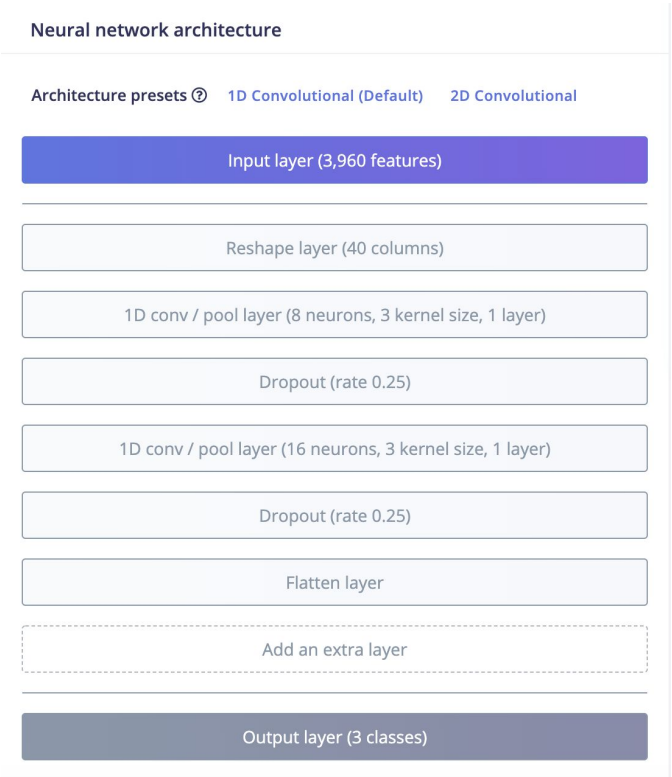
Cài đặt mạng Nơ-ron



Cài đặt mạng Nơ-ron



Cài đặt mạng Nơ-ron



Kết quả

Last training performance (validation set)

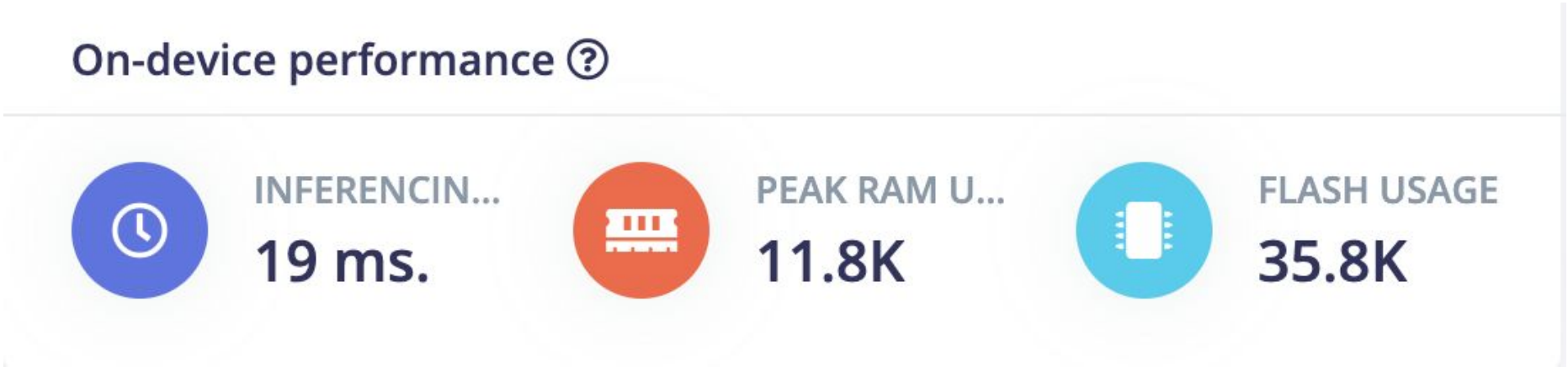
ACCURACY
95.8%

LOSS
0.17

Confusion matrix (validation set)

	HEYSIRI	NOISE	UNKNOW
HEYSIRI	100%	0%	0%
NOISE	0%	93.8%	6.3%
UNKNOW	4.5%	4.5%	90.9%
F1 SCORE	0.99	0.94	0.93

Hiệu suất hoạt động trên thiết bị



Link Video

https://youtu.be/U3EoiOzjzxI

Cảm ơn thầy và các bạn đã lắng nghe!

Nguồn tham khảo

Hey Siri: An On-device DNN-powered Voice Trigger for Apple's Personal Assistant, Apple Machine Learning Research, August 2017

OK google - Hey Siri - Sunnie ơi, Nguyen Van Duc, 2020

Long short term memory (LSTM), Tuan Nguyen, 2019

Recurrent Neural Network: Từ RNN đến LSTM, Nguyen Thanh Huyen, 2021

Supriya, Kalyanam. (2020). Trigger Word Recognition using LSTM. International Journal of Engineering Research and. V9. 10.17577/IJERTV9IS060092.

F. Ge and Y. Yan, "Deep neural network based wake-up-word speech recognition with two-stage detection," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 2761-2765, doi: 10.1109/ICASSP.2017.7952659.

K. Kumatani et al., "Direct modeling of raw audio with DNNS for wake word detection," 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2017, pp. 252-257, doi: 10.1109/ASRU.2017.8268943.

Giannakopoulos, Theodoros & Tatlas, Nicolas-Alexander & Ganchev, Todor & Potamitis, Ilyas. (2005). Practical, real-time speech-driven home automation front-end. Consumer Electronics, IEEE Transactions on. 51. 514 - 523. 10.1109/TCE.2005.1467995.

I. Stefanović, E. Nan and B. Radin, "Implementation of the wake word for smart home automation system," 2017 IEEE 7th International Conference on Consumer Electronics - Berlin (ICCE-Berlin), 2017, pp. 271-272, doi: 10.1109/ICCE-Berlin.2017.8210649.

Supriya, Kalyanam. (2020). Trigger Word Recognition using LSTM. International Journal of Engineering Research and. V9. 10.17577/IJERTV9IS060092.