

SOUND OF SILENCE – THIẾT BỊ HỖ TRỢ GIAO TIẾP THÔNG MINH CHO NGƯỜI KHIẾM THÍNH

Nguyễn Luật Gia Khôi^{1,5}, Võ Ngọc Sang^{1,5}, Ho Trí Kháng^{1,5}, Phạm Bùi Minh Huân^{1,5}, Cao Khánh Gia Hy^{2,5}, Trần Vũ Hà Nhi^{3,5}, Võ Thị Hoàng Quanh^{3,5}, Huỳnh Nhật Huy^{3,5}, Quản Thành Tho^{1,5}, Võ Thanh Hằng^{4,5}

¹Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), 268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Vietnam

²Faculty of Industrial Management, Ho Chi Minh City of Technology (HCMUT), 268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Vietnam

³Faculty of Mechanical Engineering, Ho Chi Minh City of Technology (HCMUT), 268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Vietnam

⁴Faculty of Environment and Natural Resources, Ho Chi Minh City of Technology (HCMUT), 268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Vietnam

⁵Vietnam National University Ho Chi Minh City, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Vietnam

Tóm tắt nội dung—Ở nước ta hiện có khoảng 1,5 đến 2,5 triệu người khiếm thính và người khiếm thính tương đương với cả một tỉnh nằm trong nhóm 1 về quy mô dân số, như Bắc Ninh, Quảng Ninh, tuy nhiên có rất ít phiến dịch viên ngôn ngữ ký hiệu. Riêng tại Hà Nội hiện chỉ có 6 người đạt đến trình độ có thể dịch các lĩnh vực cho người điếc, trong cả nước có khoảng 10 người [1]. Hiện nay, những người khiếm thính, khiếm thính hiện vẫn đang sử dụng thủ ngữ (ngôn ngữ ký hiệu Việt Nam) để giao tiếp trong cuộc sống.)

Tuy nhiên, hệ thống thủ ngữ này có những điểm hạn chế như: người tương tác buộc phải biết thủ ngữ, thời gian tương tác lâu, dễ bị hiểu nhầm. Như cầu sử dụng lớn nhưng tại Việt Nam còn khá ít các trung tâm dạy thủ ngữ, ngành thông dịch viên thủ ngữ cũng ít được giảng dạy tại các trường đại học.

Những hạn chế này được khắc phục và cải thiện phần nào nhờ vào giải pháp sử dụng trí tuệ nhân tạo, công nghệ học máy và phát triển ứng dụng di động. Cụ thể, với công nghệ trí tuệ nhân tạo và thị giác máy tính, các động tác ngôn ngữ ký hiệu có thể được nhận biết nhanh với độ chính xác cao. Bên cạnh đó, việc phát triển ứng dụng di động giúp tăng tính di động của giải pháp. Đặc biệt, với công nghệ lõi là phần mềm tự phát triển và dễ dàng tích hợp với các ứng dụng di động cùng với thiết bị phần cứng có giá thành sản xuất thấp, thiết bị có thể dễ dàng tiếp cận với người tiêu dùng.

Chính vì vậy, dự án “Thiết bị giao tiếp thông minh dành cho người khiếm thính, khiếm thính - Sound of Silence” (thiết bị SOS) được ra đời với mục tiêu là thiết bị giao tiếp chuyển ngôn ngữ ký hiệu sang văn bản và giọng nói để những người khiếm thính, khiếm thính dễ dàng sử dụng và trao đổi thông tin với mọi người. Dự án hiện vẫn đang được phát triển để cải thiện hiệu suất hoạt động cũng như độ chính xác của mô hình học máy.

I. PHẦN GIỚI THIỆU

“Thiết bị giao tiếp thông minh dành cho người khiếm thính, khiếm thính - Sound of Silence” (thiết bị SOS) là thiết bị giao

tiếp chuyển ngôn ngữ ký hiệu sang văn bản và giọng nói để những người khiếm thính, khiếm thính dễ dàng trao đổi thông tin với mọi người. Thiết bị SOS của chúng tôi sử dụng công nghệ AI để chuyển thủ ngữ của những người này sang dạng văn bản và giọng nói phát ra từ các thiết bị điện thoại thông minh trong thời đại công nghiệp 4.0.

Sản phẩm là một thiết bị tích hợp phần mềm thông minh nhằm hỗ trợ người khiếm thính có thể tương tác với mọi người một cách dễ dàng và thuận tiện hơn bằng cách chuyển đổi thủ ngữ thành văn bản và giọng nói. Thiết bị bao gồm bao gồm hai thành phần: mô-đun camera được gắn trên nón và ứng dụng trên điện thoại Android

II. TỔNG QUAN LÝ THUYẾT

Nói về các giải pháp hiện tại (giấy, người phiên dịch, bla bla) Trong thực tế người khiếm thính đã và đang áp dụng một số cách thức giao tiếp, có thể kể đến như: sử dụng trực tiếp ngôn ngữ ký hiệu để giao tiếp, sử dụng giấy bút, gõ ký tự trên điện thoại thông minh, hay gần đây nhất là sử dụng dịch vụ thông dịch viên trực tuyến. Cả 4 phương pháp trên đều gặp những vấn đề về sự bất tiện [Bảng I].

Cụ thể, đối với phương pháp sử dụng trực tiếp ngôn ngữ ký hiệu, đối tượng giao tiếp của người khiếm thính khiếm thính buộc phải có kiến thức về ngôn ngữ ký hiệu để hiểu được thông điệp giao tiếp. Trong khi đó, số lượng người có kiến thức về ngôn ngữ ký hiệu rất ít, do đó, phương pháp giao tiếp này không hiệu quả ở hiện tại.

Đối với phương pháp sử dụng giấy bút, vấn đề truyền tải và thấu hiểu thông điệp đã được giải quyết do khi đó cả hai đối tượng tham gia giao tiếp đều sử dụng chung ngôn ngữ tiếng Việt. Tuy nhiên, vấn đề của phương pháp này đó là yêu cầu một trong hai đối tượng giao tiếp phải có giấy, bút bên mình. Bên cạnh đó, quá trình ghi thông điệp ra giấy cũng tiêu tốn nhiều thời gian cho cả hai.

Cùng với sự phát triển của thiết bị di động, đặc biệt là điện thoại thông minh, phương pháp gõ ký tự trên điện thoại thông minh ra đời mở ra cơ hội giao tiếp tốt hơn cho người khiếm thính khiếm thính. Giờ đây, người sử dụng có thể sử dụng điện thoại thông minh và gõ thông điệp giao tiếp của mình để người đối diện có thể hiểu. Tuy nhiên, điểm yếu duy nhất của phương pháp này đó là thời gian giao tiếp vẫn còn chưa đủ nhanh mặc dù thời gian gõ phím trên điện thoại thông minh nhanh hơn viết trên giấy.

Một giải pháp cho cả 3 phương pháp trên đã được đề xuất, đó là sử dụng thông dịch viên trực tuyến thông qua nền tảng giao tiếp video. Khi đó, người khiếm thính khiếm thính sẽ

Giải pháp	Hạn chế
	<ul style="list-style-type: none"> Người giao tiếp phải hiểu ngôn ngữ ký hiệu
	<ul style="list-style-type: none"> Yêu cầu về giấy bút Tốc độ giao tiếp thấp
	<ul style="list-style-type: none"> Tốc độ giao tiếp chưa cao
	<ul style="list-style-type: none"> Yêu cầu kết nối internet Số lượng thông dịch viên không nhiều

Bảng I

CÁC GIẢI PHÁP GIAO TIẾP NGÔN NGỮ KÝ HIỆU HIỆN NAY

diễn đạt ngôn ngữ ký hiệu cho thông dịch viên, sau đó, người thông dịch viên sẽ phiên dịch thông điệp này sang tiếng Việt cho người đối diện. Mặc dù giải pháp đã đưa vào thực tế và đem lại hiệu quả nhờ chi phí dịch vụ thấp và tốc độ giao tiếp nhanh. Tuy vậy, giải pháp này vẫn còn tồn đọng nhiều mặt hạn chế, có thể kể đến như yêu cầu người dùng phải có kết nối internet di động ổn định, điều mà chưa nhiều người dùng ở Việt Nam có thể đáp ứng, cũng như số lượng thông dịch viên phải đủ lớn để đáp ứng nhu cầu ngày càng tăng trong khi nghịch lý là số lượng người theo nghề thông dịch viên ngôn ngữ ký hiệu ngày càng ít.

Qua các phân tích trên ta có thể thấy nhu cầu rất cao về một thiết bị có thể dễ dàng mang theo, thời gian thông dịch nhanh và không cần kết nối internet. Nắm bắt được nhu cầu của người dùng, cùng với sự phát triển nhanh chóng của công nghệ trí tuệ nhân tạo, thiết bị giao tiếp thông minh dành cho người khiếm thính - Sound of Silence được nghiên cứu và phát triển với mục tiêu tạo ra sản phẩm là thiết bị có khả năng thông dịch ngôn ngữ ký hiệu một cách nhanh chóng thông qua kết nối thiết bị với điện thoại thông minh.

III. CHI TIẾT SẢN PHẨM

A. Mô tả sản phẩm

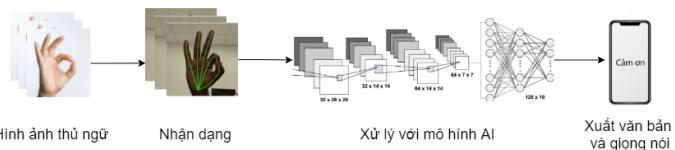
Thiết bị gồm 2 phần chính, camera module gắn trên nón và mô hình phiên dựng ở dạng ứng dụng được cài trên thiết bị di động.

Camera nhỏ gọn và các linh kiện điện tử được lắp đặt trong hộp nhựa gắn vào nón của người dùng. Việc tích hợp camera vào nón cố định thiết bị giúp tín hiệu hình ảnh đầu vào được đảm bảo và tầm nhìn của camera sẽ không bị ảnh hưởng bởi yếu tố ngoại tác nhiều. Ngoài ra vị trí trên nón còn giúp tăng khả năng cơ động của thiết bị khi có thể được tháo lắp dễ dàng, không cản trở quá trình sinh hoạt của người dùng. Vật liệu sử dụng cho hộp chứa linh kiện điện tử là nhựa sinh học (Polylactic Acid - PLA) không độc hại, không gây ảnh hưởng đến sức khỏe người tiêu dùng và có khả năng dễ phân hủy, thân thiện với môi trường. Khi người dùng nhấn nút khởi động của thiết bị, camera này sẽ ghi nhận và chuyển dòng hình ảnh vào ứng dụng đã cài đặt sẵn trên điện thoại thông minh của người dùng.

Ứng dụng sẽ sử dụng các công nghệ ở trên để chuyển dòng hình ảnh thủ ngữ thành văn bản và giọng nói tương ứng hiện lên màn hình và phát ra loa. Bằng cách này, người đối diện có thể hiểu được điều mà người khiếm thính cần truyền đạt mà không cần phải học qua thủ ngữ. Ứng dụng được thiết kế với dung lượng nhỏ gọn, không đòi hỏi phần cứng quá mạnh mẽ để đạt được tốc độ và độ chính xác cao, có khả năng chạy trên nhiều mẫu thiết bị mới và cũ. Việc không yêu cầu phần cứng mạnh mẽ cũng phù hợp đối tượng người dùng khiếm thính, khiếm thính là những người có khả năng tài chính không mạnh. Ngoài ra, ứng dụng còn có chức năng lưu lại thủ ngữ cùng với văn bản và phát âm tương ứng. Chức năng này giúp cho những người muốn học thủ ngữ có thể lưu trữ bài học một cách dễ dàng. Nói cách khác, hệ thống vừa là thiết bị dịch thủ ngữ sang văn bản và lời nói, vừa là ứng dụng hỗ trợ học thủ ngữ hiệu quả.

B. Nguyên lý hoạt động

Ứng dụng AI để chuyển các động tác thủ ngữ thành văn bản và giọng nói giúp cho việc giao tiếp của người khiếm thính, khiếm thính với cộng đồng dễ dàng hơn. Các bước thực hiện như sau (Hình 1):



Hình 1. Luồng xử lý

Môđun camera ghi lại chuỗi hình ảnh thủ ngữ của người dùng, truyền làm đầu vào cho ứng dụng.

Bộ công cụ Mediapipe được phát triển bởi Google sẽ tiếp nhận hình ảnh và trích xuất các đặc trưng về các điểm chốt của bàn tay.

Mô hình AI được huấn luyện trước bằng công nghệ học sâu

sẽ tiếp nhận các đặc trưng và dự đoán kết quả tương ứng. Đầu ra của bước này sẽ là văn bản tương ứng với chuỗi hình ảnh thủ ngữ được tiếp nhận ban đầu.

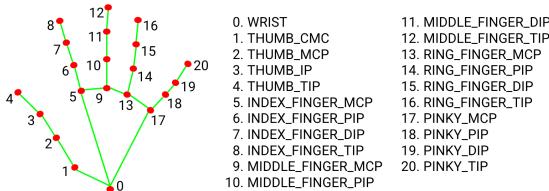
Công cụ Text2Speech được phát triển bởi Google sẽ tiếp nhận và chuyển văn bản thành giọng nói tương ứng.

Kết quả sau cùng ở cả hai dạng văn bản và giọng nói sẽ được hiển thị trên thiết bị của người dùng.

IV. XÂY DỰNG MÔ HÌNH VÀ HOÀN THIỆN SẢN PHẨM

A. Xây dựng mô hình học máy

Ở bước tiền xử lý, mô hình Mediapipe[9] sẽ nhận vào chuỗi hình ảnh thủ ngữ từ camera và trích xuất ra những vec tơ đặc trưng tương ứng. Mỗi một động tác tay được ghi nhận sẽ tương ứng với một vec tơ đầu ra 2 chiều \vec{v} . Vec tơ \vec{v} , có kích thước (21,3), bao gồm 21 phần tử chứa thông tin tọa độ của 21 điểm nốt trên bàn tay (Hình 2). Mỗi một điểm nốt sẽ được ghi nhận 3 cặp số tọa độ (x, y, z) tương ứng. Trong đó, x và y là tọa độ ngang và dọc của điểm nốt so với gốc tọa độ là điểm chính giữa của bàn tay. x và y được chuẩn hóa về khoảng [0,1] so với kích thước ảnh. z thể hiện độ xa của điểm nốt so với cổ tay và cũng được chuẩn hóa về khoảng [0, 1].



Hình 2. 21 điểm nốt trên bàn tay được nhận dạng bởi mô hình Mediapipe

Sau khi đã trích xuất ra được vec tơ đặc trưng v , vec tơ này sẽ được chuyển hóa thành ma trận khoảng cách tương ứng M . M là một ma trận hai chiều có kích thước (21 x 21). Trong đó M_{ij} sẽ tương ứng với khoảng cách Euclid (Euclidean distance) giữa điểm nốt i và điểm nốt j, kí hiệu d_{ij} . Như đã đề cập trước đó, mỗi điểm nốt sẽ bao gồm 3 cặp số x, y, z tương ứng. Như vậy, khoảng cách Euclid giữa hai điểm nốt sẽ là khoảng cách giữa 2 điểm trong không gian 3 chiều:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

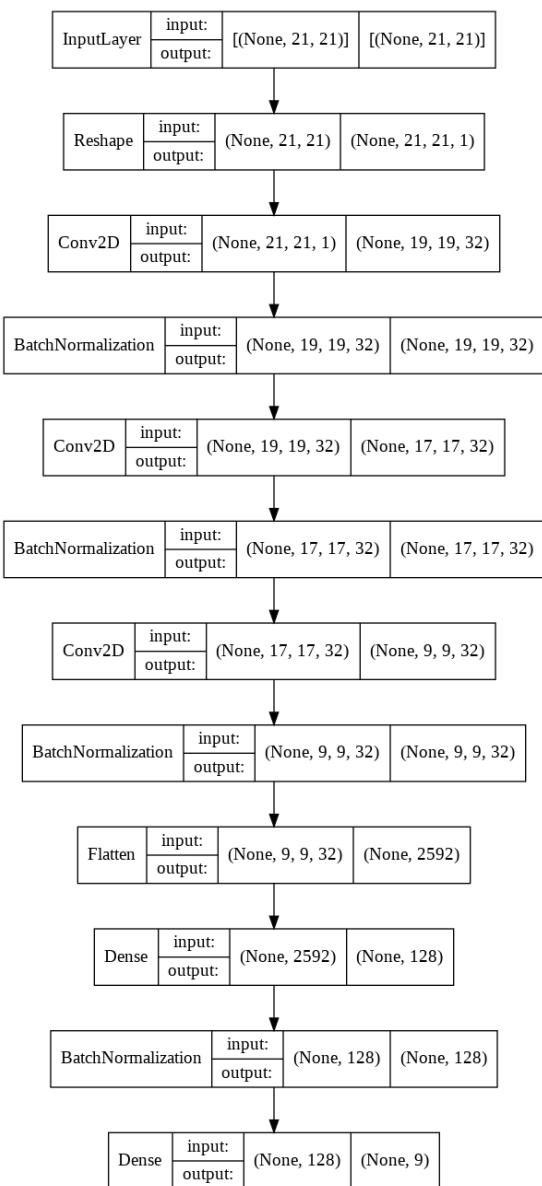
Cấu trúc mô hình:

Mô hình mà chúng tôi xây dựng lấy cảm hứng từ mạng Nơron tích chập (Convolutional Neural Network [10]) sẽ nhận đầu vào là ma trận khoảng cách M và xuất ra một vec tơ dự đoán từ ngữ tương ứng với ma trận M . Cấu trúc của mô hình bao gồm: 3 lớp Convolution chồng lên nhau, ngay sau mỗi lớp Convolution là một lớp Batch Normalization, cuối cùng là 2 lớp Dense với hàm kích hoạt softmax ở lớp Dense sau cùng để dự đoán đầu ra của bài toán. Hai lớp Convolution đầu tiên được cấu hình với kích thước filter = 3, padding = 0, stride = 1 và theo sau bởi hàm kích hoạt tanh. Riêng lớp Convolution thứ ba có kích thước filter = 5, padding 'same'

và stride = 2. Chi tiết về cấu trúc mô hình được thể hiện ở Hình 3. Số lượng lớp đầu ra của mô hình sẽ phụ thuộc hoàn toàn vào tập dữ liệu dùng để huấn luyện mô hình. Mô hình gốc mà chúng tôi tạo ra được huấn luyện dựa trên 9 nhóm đầu ra, tương ứng với 9 từ hoặc cụm từ thông dụng trong tiếng Việt.

Huấn luyện:

Mô hình được huấn luyện trên công cụ Google Colab với tập dữ liệu bao gồm 193 hình ảnh động tác thủ ngữ. Mỗi hình ảnh sau khi thông qua bước tiền xử lý sẽ được chuyển hóa thành ma trận khoảng cách M tương ứng và đưa vào mô hình huấn luyện. Quá trình huấn luyện sử dụng thuật toán tối ưu Adam[11] với tốc độ học là 0.001.



Hình 3. Cấu trúc mô hình. Từ trái sang phải: Tên lớp, Kích thước đầu vào, kích thước đầu ra

B. Hoàn thiện sản phẩm

Do yêu cầu về tính tiện dụng của sản phẩm, đó là chạy được trên các thiết bị di động với dung lượng nhỏ và có thể sử dụng trong môi trường không có kết nối Internet, mô hình xây dựng bên cạnh việc tối ưu để đạt độ chính xác cao, còn cần thiết phải chú trọng vào yếu tố khả dụng và độ hiệu quả trên các thiết bị.

Sau khi hoàn thành quá trình huấn luyện, chúng tôi đã chuyển hóa mô hình thành một phiên bản gọn nhẹ hơn sử dụng Tensorflow Lite. Tensorflow Lite về cơ bản là một bộ công cụ giúp tinh giản và chạy mô hình trên các thiết bị di động và thiết bị nhúng, đảm bảo được tính cục bộ của mô hình đồng thời tối ưu hóa tài nguyên của thiết bị.

V. PHÂN TÍCH ĐỘ HIỆU QUẢ

Trước tiên, cần phải đề cập đến tính di động của thiết bị. Những mô hình trí tuệ nhân tạo được xây dựng với mục đích tương tự, thông thường sẽ tiêu tốn rất nhiều tài nguyên và bộ nhớ bởi cách tiếp cận vẫn đề chưa được tối ưu hóa. Tuy nhiên đối với sản phẩm này, nhờ có hương tiếp cận mới, kết hợp với việc phân loại cử chỉ tay, nhận diện vị trí bàn tay và phép ánh xạ mỗi một tổ hợp đầu ra với một từ ngữ, những công đoạn về dịch thô ngữ trực tiếp sang dạng văn bản, tưởng chừng như rất phức tạp, lại có thể được thực hiện ngay trên điện thoại di động.

Nói về độ chính xác, việc nhận diện cử chỉ tay được thực hiện rất tốt. Cụ thể, nhiệm vụ nhận diện cử chỉ tay của mô hình đạt tỉ lệ chính xác trung bình khoảng 89%.

Tuy nhiên, mô hình cho nhiệm vụ nhận diện vị trí và sự chuyển động của bàn tay vẫn còn nhiều mặt hạn chế. Ở phương diện này, mô hình chỉ đạt tỉ lệ chính xác vào khoản 60/

VI. NHỮNG THÀNH TÍCH VÀ GIẢI THƯỞNG

Sản phẩm của nhóm cũng đã gây được sự chú ý của cộng đồng. Đặc biệt là vào tháng 11 năm 2021, nhận thấy tính nhân văn cần thiết của sản phẩm, Đài Phát Thanh và Truyền Hình Vĩnh Long đã phối hợp với trường Đại học Bách Khoa TPHCM cũng như các thành viên trong nhóm để đưa tin, truyền bá về sản phẩm. Dưới đây là những hình ảnh của bản tin về sản phẩm của nhóm được chiếu tại kênh Truyền Hình Vĩnh Long ngày 30/11/2021.

VII. PHẦN TỔNG KẾT

“Thiết bị giao tiếp thông minh dành cho người khiếm thính, khiếm thính - Sound of Silence” (thiết bị SOS), nói tóm lại, là thiết bị giao tiếp chuyển ngôn ngữ ký hiệu sang văn bản và giọng nói để những người khiếm thính, khiếm thính dễ dàng trao đổi thông tin với mọi người. Thiết bị được áp dụng công nghệ AI để chuyển thủ ngữ của những người này sang dạng văn bản và giọng nói phát ra từ các thiết bị điện thoại thông minh, thứ được sử dụng phổ biến trong thời đại công nghiệp 4.0.

Ngoài những điểm mạnh như tính di động, tốc độ và độ chính xác, sản phẩm bên cạnh đó vẫn còn đang có một vài mặt hạn chế như việc chưa thể dịch hết tất cả thô ngữ sang văn bản (nguyên nhân chủ yếu là do nhóm đang thiếu dữ liệu),



Hình 4. Đài Phát Thanh và Truyền Hình Vĩnh Long đưa tin về sản phẩm của nhóm

cũng như việc nhận diện trị trí và sự chuyển động của bàn tay đạt còn vận hành chưa được như kì vọng.

Tuy nhiên, về mặt lâu dài, nhóm vẫn sẽ tiếp tục theo đuổi dự án, trước mắt có thể là tìm kiếm, cung cấp thêm dữ liệu về thô ngữ nhằm tạo sự khả thi cho việc dịch một số lượng lớn ngôn ngữ khí hiệu, sau đó là đến việc cải thiện tốc độ và độ chính xác của mô hình,... Tất cả những sự cố gắng này, mục đích là để, bằng một cách nào đó, có thể giúp cho cộng đồng người khiếm thanh khiếm thính có thể giao tiếp được với mọi người, từ đó xóa bỏ đi rào cản ngôn ngữ cũng như tạo cho họ thêm niềm tin, sự đón nhận trong cuộc sống.

TÀI LIỆU

- [1] Số lượng người khiếm thính ở Việt Nam, *Trung Tâm Nghiên cứu Giáo dục Người Khiếm Thính*, <http://www.philoinhuan.org/to-chuc/trung-tam-nghiencuu-giao-duc-nguoikhiem-thinh-ced>
- [2] Công Ty TNHH Sản Xuất Thương Mại Mỹ Toàn. “Nhựa PLA”. Available: <https://nhuadinhthinh.vn/nhua-pla/>
- [3] Thành Hòa, “Hơn 2.000 học sinh khiếm thính được học bằng ngôn ngữ ký hiệu”. *Báo Dân Sinh*, chuyên mục Giáo dục nghề nghiệp, 2019. Available: <https://baodansinh.vn/hon-2000-hoc-sinh-khiem-thinh-duoc-hoc-bang-ngon-Ngu-ky-hieu-20190927155542301.htm>
- [4] Phương Nhhung. “Dịch vụ nghe, nói “hộ” người khiếm thính”. *Người lao động*, chuyên mục Trong nước, 2018. Available: <https://nld.com.vn/thoi-su/dich-vu-nghe-noi-ho-nguoikhiem-thinh-20180215011145489.htm>
- [5] Kiều Tuyết, Nguyễn Yên. “Thiếu phiến dịch viên, bệnh nhân đặc biệt thiệt thòi về cơ hội chăm sóc y tế”. *VOV Giao thông*, chuyên mục Góc nhìn, 2020. Available: <https://vovgiaothong.vn/thieu-phiendich-vien-benh-nhan-dac-biet-thiet-thoi-ve-co-hoi-cham-soc-y-te>
- [6] Nguyễn Nhất Duy, “iPrice Group: Tổng kết Thương mại điện tử Việt Nam năm 2020 hướng đến 2021”. *iPrice*, chuyên mục Xu hướng, 2020. Available: <https://iprice.vn/xu-huong/insights/iprice-group-tong-ket-thuong-mai-dien-tu-viet-nam-nam-2020-huong-den-2021/>
- [7] Viện nghiên cứu Xã hội, Kinh tế và Môi trường. “Xóa bỏ kỳ thị - Quan điểm và đánh giá của người khuyết tật”. Nhà xuất bản Tri thức, 2017. Available: <http://isee.org.vn/wp-content/uploads/2020/02/XOA-BO-KY-THI-VN.pdf>
- [8] Hải Anh. “Phiến dịch viên Ngôn ngữ ký hiệu, “nghề” kén người học bậc nhất”. *Tạp chí điện tử Giáo dục Việt Nam*, chuyên mục Giáo dục 24h, 2021. Available: <https://giaoduc.net.vn/giao-duc-24h/phiendich-vien-ngo-ky-hieu-ken-nguo-hoc-bac-nhat-post217305.gd>
- [9] Fan Zhang et al., MediaPipe Hands: On-device Real-time Hand Tracking, <https://arxiv.org/abs/2006.10214>
- [10] Wikipedia, Mạng thần kinh tích chập, https://en.wikipedia.org/wiki/Convolutional_neural_network
- [11] Diederik P. Kingma et al., Adam: A Method for Stochastic Optimization, <https://arxiv.org/abs/1412.6980>