NGHIÊN CỨU PHƯƠNG PHÁP MIÊU TẢ HÌNH ẢNH TỰ ĐỘNG SỬ DỤNG PHƯƠNG PHÁP HỌC SÂU - ỨNG DỤNG XÂY DỰNG HỖ TRỢ NGƯỜI KHIẾM THỊ

AUTOMATIC IMAGE CAPTIONING WITH DEEP LEARNING – APPLYING TO HELP THE BLIND VISUALIZE THE WORLD

SVTH: Lưu Văn Huy, Dương Sỹ Bình, Mai Hữu Mân, Nguyễn Huy Tưởng

Khoa Công nghệ thông tin, Trường Đại học Bách Khoa, Đại Học Đà Nẵng;

Emai: luuvanhuy2012@gmail.com,huytuong010101@gmail.com,binh9adt@gmail.com,maimanhuu@gmail.com

GVHD: Phạm Minh Tuấn

Khoa Công nghệ thông tin, Trường Đại học Bách Khoa, Đại Học Đà Nẵng;

Email: pmtuan@dut.udn.vn

Tóm tắt - Trong báo cáo này, chúng tôi nghiên cứu phương pháp để mô tả hình ảnh một cách chính xác này. Chúng tôi xây dựng mô hình dựa trên kiến trúc cơ bản CNN-RNN và sử dụng cơ chế chú ý tập trung vào những điểm đáng chú ý của hình ảnh khi mô tả. Hơn nữa, chúng tôi thấy dựa vào kết quả nghiên cứu, chúng tôi thấy rằng việc dự đoán đầu ra của mô hình cơ bản là dựa theo giải thuật tham lam, điều này có thể bỏ qua các kết quả tốt về lâu dài. Chính vì vậy chúng tôi đã sử dụng phương pháp tìm kiếm Beam search để tìm kiếm các kết quả tốt nhất của mô hình.Sau khi đã xây dựng mô hình, chúng tôi áp dụng kết quả để xây dựng hệ thống ứng dụng mô tả môi trường xung quanh cho người khiếm thị.

Từ khóa: Trí tuệ nhân tạo; mạng học sâu; miêu tả tự động; CNN-RNN; cơ chế chú ý; Beam search; người khiếm thị.

Abstract - In this paper, we do a research on how to automaticly caption an image with high accuracy. We build the model based on CNN-RNN architecture and apply Attention machenism in focusing what is highlighted in the image when generating caption. We also apply Beam search instead of Greedy search to maximize the chance of generating the best caption. We apply the result of the research in building a mobile apply to assist the blind in visualizing the world.

Keywords: Artifical intelligence; Image captioning; CNN-RNN; Attention machenism; Beam search; the blind.

1. Đặt vấn đề

Hiện nay, khoa học công nghệ cũng như trí tuệ nhân tạo đang phát triển với tốc độ rất nhanh, nhiều sản phẩm, ứng dụng đã được đưa vào cuộc sống giúp nâng cao năng suất và chất lượng cuộc sống. Một trong những vai trò của trí tuệ nhân tạo là giúp máy tính hiểu được thế giới xung quanh như hình, âm thanh,… Gần đây, vấn đề chú thích cho hình ảnh đã nổi lên như một vấn đề nghiên cứu nổi bật trong lĩnh vực trí tuệ. Nó có thể mô tả được môi trường có cấu trúc phức tạp xung quanh chỉ bằng vài câu đơn giản như con người. Hiện nay đã có rất nhiều nghiên cứu tập trung vào lĩnh vực này, có rất nhiều đột phá tuy nhiên vẫn còn nhiều điểm hạn chế. Trong phạm vi nghiên cứu này, chúng tôi chỉ ra một số điểm hạn chế còn tồn tại ở mô hình mô tả hình ảnh tự động và đề xuất phương pháp giải quyết chúng. Hơn nữa chúng tôi cũng thực hiện một ứng dụng áp dụng mô hình mình xây dụng với mong muốn giúp máy tính có thể hiểu và mô tả lại nội dung của một bức ảnh, giúp đỡ người khiếm thị có thể hình dung được thế giới xung quanh.Những điểm chú ý trong nghiên cứu của chúng tôi :

- Xây dựng mô hình mô tả hình ảnh cơ bản và nhận xét nhược điểm của mô hình.
- Dựa trên mô hình cơ bản, sử dụng cơ chế chú ý để cải tiến mô hình mô tả hình ảnh.
- Chúng tôi thực hiện tìm kiếm những đầu ra tốt nhất của mô hình bằng cách phương pháp beam search.

Bài báo này có bố cục như sau. Phần 2 trình bày tổng quan về lý thuyết các kiến trúc sử dụng. Phần 3 trình bày kết quả và so sánh. Phần 4 trình bày bàn luận. Phần 5 trình bày kết luận. Phần 6 trình bày ứng dụng Assisstant.

2. Cơ sở lý thuyết

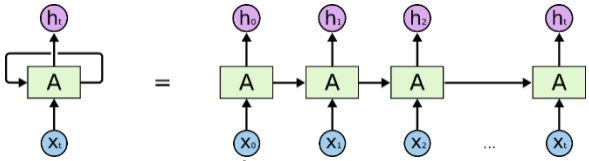
2.1. Mô tả hình ảnh với kiến trúc CNN-RNN

2.1.1 Kiến trúc CNN:

Kiến trúc CNN được sử dụng nhiều trong các bài toán liên quan đến xử lí ảnh và trích xuất đặc trưng của một bức ảnh. Trong dự án này kiến trúc CNN dùng để trích xuất đặc trưng của một bức ảnh[1], rồi từ đặc trưng đó sinh ra một câu bao hàm nội dung của bức ảnh.

2.1.2. Kiến trúc RNN:

Kiến trúc RNN thường được sử dụng trong các kiểu dữ liệu chuỗi, có thứ tự như (Video, Câu, ...), những kiểu dữ liệu này không thể sử dụng kiến trúc CNN được. Bên trong mạng RNN có vòng lặp giúp mạng có thể ghi nhớ những kiến thức đã học trước khi tiếp thu kiến thức mới, giống như việc chúng ta dựa vào những cảnh đã diễn ra để hiểu được cảnh hiện thời của một đoạn video.

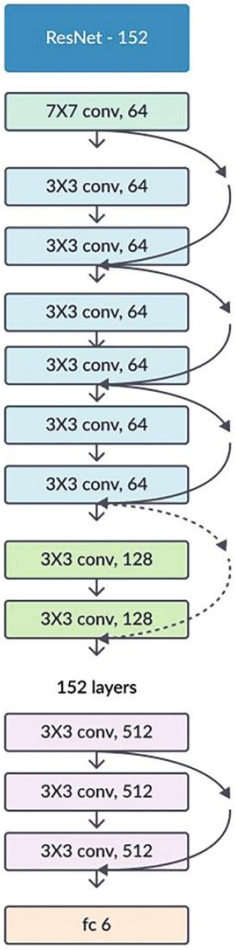


Hình 1. Kiến trúc của mạng RNN.

Kiến trúc RNN được sử dụng ở dự án này với chức năng dự đoán từ tiếp theo của một câu cho trước và lặp lại quá trình đó để được một câu hoàn chỉnh. Một hạn chế của mạng RNN là sau khi ghi nhớ một đoạn kiến thức quá dài, thông tin sẽ bị mất mát, tương tự với việc bạn xem đến cuối đoạn phim thì không thể nhớ đoạn đầu phim có gì. Vì vậy mạng LSTM (Long Short Term Memory networks) ra đời giúp khắc phục tình trạng lãng quên đối với những đoạn dữ liệu dài.

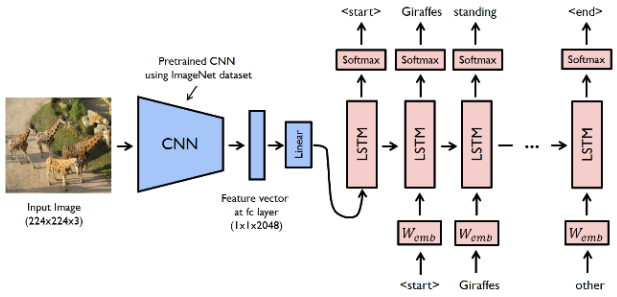
2.1.3. Mô hình kết hợp CNN-RNN:

Trong dự án này chúng tôi sử dụng kiến trúc ResNet152 được huấn luyện trước trên bộ ảnh nổi tiếng ImageNet. Kiến trúc ResNet152(kiến trúc Residual neural network[2]) cùng với bộ ảnh ImageNet đã đạt được tỉ lệ lỗi 3.57% và đạt giải nhất tại ILSVRC 2015 classification task.



Hình 2. Kiến trúc Resnet152.

Qua kiến trúc CNN này chúng ta trích xuất được đặc trưng của bức ảnh, đây là đầu vào của kiến trúc tiếp theo, kiến trúc LSTM (Nâng cấp của RNN). Mạng LSTM nhận đầu vào là vector đặc trưng của bức ảnh và một phần câu đã hoàn thiện, tự hoàn thiện câu bằng cách sinh tiếp tục những từ tiếp theo cho đến khi gặp từ khóa kết thúc. Ban đầu mạng LSTM sẽ nhận vào vector đặc trưng ảnh và từ khóa bắt đầu.



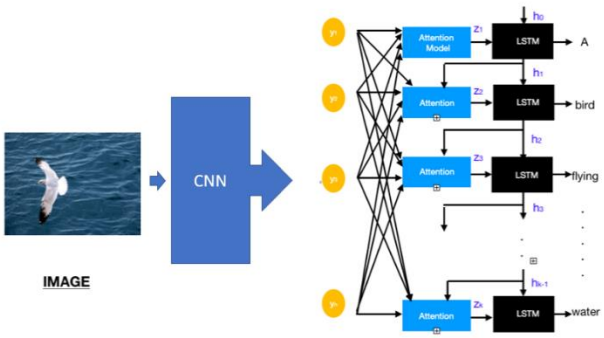
Hình 3. Mô hình kết hợp kiến trúc CNN-RNN

2.2. Sử dụng cơ chế Attention trong mô tả hình ảnh

Vấn đề của kiến trúc mô tả hình ảnh chỉ sử dụng sự kết hợp CNN-RNN đó là việc nắm bắt bố cục của hình ảnh để sinh ra các từ khác nhau trong câu mô tả. Nguyên nhân của việc này đó là mỗi từ trong câu thường chỉ mô tả một phần nội dung của bức ảnh trong

khi với mô hình cũ việc dự đoán mỗi từ của lớp RNN đều dựa trên toàn bộ đặc trưng trích xuất được từ lớp CNN. Ý tưởng của chúng tôi là áp dụng cơ chế Attention đã được nghiên cứu trước đây của mô hình dịch ngôn ngữ do Dzmitry Bahdanau[3] đề xuất vào mô hình mô tả hình ảnh, khi đó mỗi từ sẽ được dự đoán tập trung vào một số đặc trưng của bức ảnh mà thể hiện rõ nhất nội dung của nó.

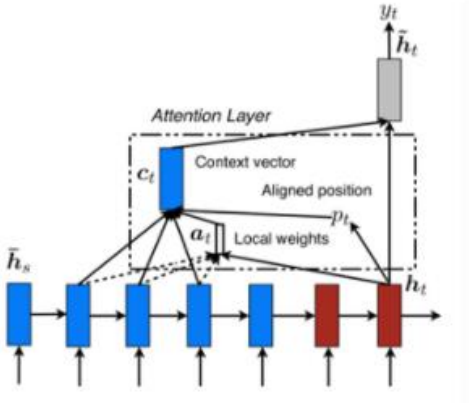
2.1.1 Kiến trúc ứng dụng cơ chế Attention



Hình 4. Kiến trúc ứng dụng cơ chế attention.

Thay vì giải mã bức ảnh thành một vector thuộc tính có 2048 chiều, mạng CNN được tinh chỉnh để giải mã một bức ảnh đầu vào thành 49 vector có 2048 chiều chứa các đặc trưng của bức ảnh. Các vector này sẽ là đầu vào của lớp attention, cùng với đầu ra của lớp LSTM trước đó để tính toán đầu vào cho lớp LSTM tiếp theo.

Cơ chế chú ý - Attention có thể chia làm 2 loại là Global Attention[4] (Luong's Attention) và Local Attention[5] (Bahdanau Attention). Tiến hành thực nghiệm và so sánh kết quả cho thấy Local Attention cho câu mô tả nội dung của bức ảnh chính xác hơn so với sử dụng Global Attention cùng với đó chi phí tính toán của Local Attention cũng thấp hơn.



Hình 5. Cơ chế Local Attention.

Dưới đây là công thức tính độ quan trọng của một vector thuộc tính khi tiến hành sinh ra một từ trong câu mô tả:

$$e_{jt} = f_{ATT}(s_{t-1}, h_j) \quad (1)$$

Trong đó s_{t-1} là lớp ẩn khi sinh ra từ trước đó, h_j là vector thuộc tính đang xem xét. f_{ATT} là một tế bào neuron đơn giản:

$$f_{ATT} = V_{attn}^T * \tanh(U_{attn} * h_j + W_{attn} * s_t) \quad (2)$$

Để đưa về phân phối xác suất chúng tôi sử dụng hàm Softmax:

$$\alpha_{it} = \text{Softmax}(e_{jt}) \quad (3)$$

Vector bối cảnh Ct được tính như sau:

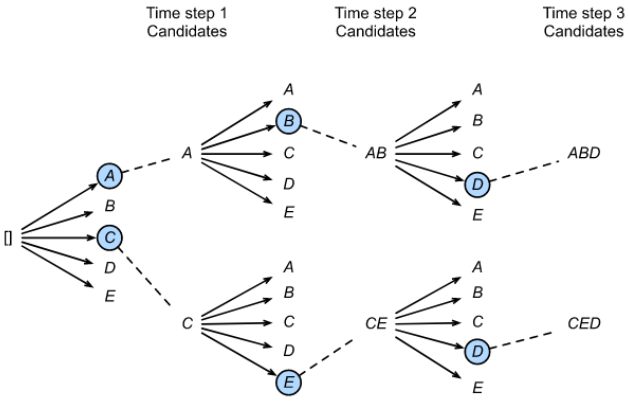
$$C_t = \sum_{j=1}^{T} \alpha_{jt} h_j \quad (4)$$

Sau đó lớp LSTM sẽ nhận đầu vào là C_t và từ trước đó để sinh ra từ tiếp theo trong câu tương tự như kiến trúc RNN đã được đề cập ở trên.

2.3. Phương pháp dựa đoán đầu ra sử dụng Beam Search

Tại mỗi bước tìm kiếm kết quả, chúng ta cung cấp cho mô hình đặc trưng của bức ảnh và câu mô tả trước đó. Kết quả dự đoán là xác suất xuất hiện của từ tiếp theo được thể hiện qua một vector với số chiều là số lượng từ vựng, giá trị là xác suất xuất hiện tiếp theo của từ đó. Vì vậy với cách tìm kiếm truyền thống (tìm kiếm tham lam), tại mỗi bước chúng ta chỉ việc lấy từ có xác suất xuất hiện cao nhất, thêm từ đó vào mô tả và tiếp tục tìm kiếm từ tiếp theo cho đến khi gặp từ khóa kết thúc.

Tuy nhiên qua kết quả thực nghiệm cho thấy việc chỉ chọn từ có xác suất xuất hiện cao nhất đã bỏ lỡ đi nhiều câu có triển vọng hơn là câu được sinh ra. Để khắc phục tình trạng đó, Beam Search ra đời giúp chúng ta mở rộng phạm vi tìm kiếm, tránh việc bỏ sót những câu có triển vọng nhưng không được chọn.



Hình 6. Tìm kiếm Beam Search với beam width bằng 2.

Ý tưởng của Beam Search là thay vì chọn duy nhất một từ triển vọng để tạo thành câu (kết thúc quá trình tìm kiếm chỉ chọn được một câu) thì chúng ta sẽ chọn beam width từ tiếp theo có xác suất cao nhất để tạo thành câu. Vậy tại mỗi bước tìm kiếm chúng ta chỉ giữ lại beam width câu tốt nhất. Cho đến khi tìm kiếm kết thúc, chúng ta có được một cây xác suất, từ đó có thể nhân xác suất trên mỗi câu và chọn câu có xác suất cao nhất. Vì tích xác suất thường nhỏ nên hàm Logarit thường được sử dụng. Beam width thường được chọn từ 5 đến 10 để đảm bảo thời gian tìm kiếm không quá lâu.

3. Triển khai và đánh giá kết quả

3.1. Bộ dữ liệu

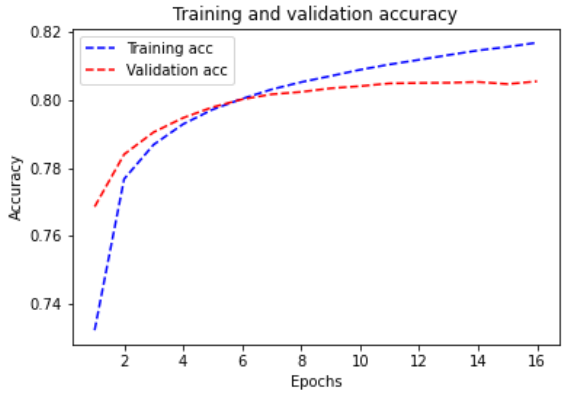
Trong nghiên cứu này, chúng tôi dùng tập dữ liệu Flickr 8k. Flirkr 8k là bộ dữ liệu gồm hình ảnh và chú thích dùng trong bài toán mô tả ảnh. Nó gồm 8000 ảnh trong đó có 6000 ảnh thuộc tập huấn luyện, 1000 ảnh thuộc tập kiểm thử và 1000 ảnh thuộc tập đánh giá. Với mỗi ảnh sẽ có 5 chú thích khác nhau.

Tất cả các chú thích được xử lý để đảm bảo độ dài của chú thích là 39 từ. Chúng tôi cũng xử lý phần chú thích như đưa các từ về dạng viết thường, loại bỏ các từ không xuất hiện trên 5 lần, và thu được tập tập từ điển gồm 2632 từ.

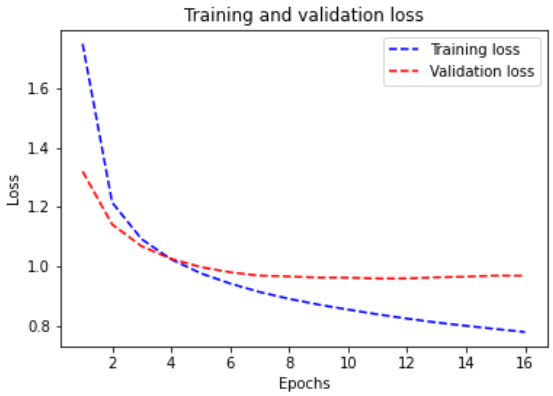
Các ảnh được định dạng về kích thước 224x224 và được chuẩn hóa về miền giá trị (-1,1) trước khi đưa vào các mô hình.

3.2. Kết quả thu được

3.2.1 Kết quả thu được từ mô hình CNN-RNN



Hình 7. Giá trị hàm mất mát của tập huấn luyện và tập kiểm tra theo vòng lặp.



Hình 8. Giá trị độ chính xác của tập huấn luyện và tập kiểm tra theo vòng lặp.

Hình 9. Kết quả thu được từ 1 ảnh trong tập thử nghiệm.

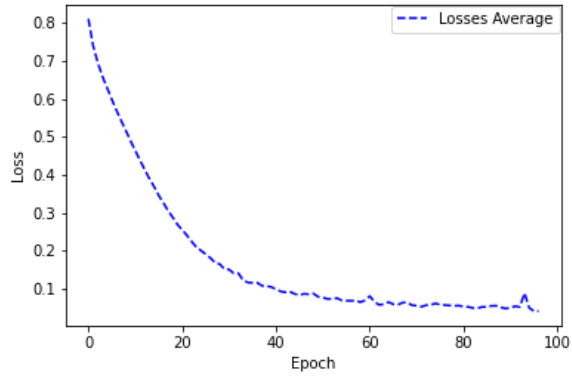
Hình 10. Kết quả thu được từ 1 ảnh trong tập thử nghiệm.

Kết quả sau khi huấn luyện cho độ chính xác cao, một số ảnh cho ra mô tả tương đối chính xác(Hình 9. Kết quả thu được từ 1 ảnh trong tập thử nghiệm.Hình 9), nhưng còn nhiều ảnh mô tả không chính xác(Hình 10) nên cần cải tiến cách trích xuất đặc trưng từ ảnh.

3.2.2 Kết quả thu được từ mô hình sử dụng cơ chế Attention

Với mô hình sử dụng cơ chế Attention, chúng tôi thực hiện huấn luyện trên tập huấn luyện và thu được

giá trị hàm loss theo từng vòng lặp. Giá trị hàm mất mát thấp hơn khi so sánh với mô hình CNN-RNN.



Hình 11. Giá trị hàm mất mát theo vòng lặp.

Hình 12. Kết quả thu được từ 1 ảnh trong tập thử nghiệm.

3.3 So sánh các mô hình

Chúng tôi thực hiện thử nghiệm trên 4 ảnh thuộc tập kiểm định để so sánh độ chính xác và độ tự nhiên của các mô tả được tạo ra cho mỗi ảnh.(Bảng 1)

Ảnh	Kết quả từ CNN-RNN	Kết quả từ attention	Caption thật
	a football player is holding a football	a football player in a red and white uniform	guy in red and white football uniform
	a man is wearing a blue shirt is wearing a blue shirt	a man smiling in his uniform with a closeup of broadway	this man is smiling very big at the camera
	a group of people are sitting on the street	the view of a dog in a shopping cart down a hand	a man with a pug dog bends over to pick something off the sidewalk
	A black dog runs through the grass	a dog sniffing catching a frisbee	two dogs are catching blue frisbee in grass

Bảng 1. So sánh một số kết quả mô tả hình ảnh giữa các mô hình.

3.4 Đánh giá kết quả

3.4.1 Sự khác nhau giữa các kết quả thu được từ 2 mô hình

Cấu trúc CNN-RNN là cấu trúc nguyên sơ, sử dụng CNN để trích xuất đặc trưng của cả bức ảnh vào một vector đặc trưng duy nhất dẫn đến tình trạng không có sự chú ý vào một chi tiết nào đó tại một ngữ cảnh nhất định. Qua mô hình sử dụng cơ chế Attention đã khắc phục được tình trạng này, tại mỗi thời điểm, mô hình sẽ học được cách để chú ý vào những phần liên quan đến việc sinh ra từ kế tiếp hơn là chú ý đến đặc trưng của cả bức ảnh. Tuy nhiên hầu hết những mô tả sinh ra có phần gượng gạo, chưa tự nhiên và gần với ngôn ngữ con người nhất.

3.4.2 Tại sao có những kết quả của attention có điểm số BLEU[6] cao hơn nhưng vẫn không có độ tự nhiên?

Kiến trúc attention có vẻ cho độ chính xác cao tuy nhiên vì chú tâm vào độ chính xác mà lời mô tả có phần gượng gạo, một phần là do trong quá trình xử lí dữ liệu đã cắt bỏ đi những từ ít xuất hiện.

3.4.3 Liệu mô hình có thể mô tả được tất cả sự vật, sự việc?

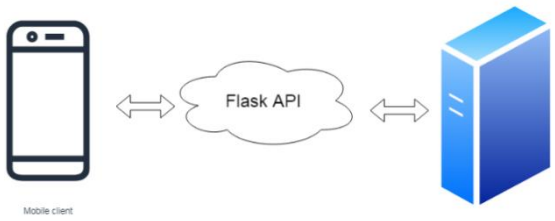
Mô hình được huấn luyện trên bộ dữ liệu Flirck8k gồm 6000 ảnh huấn luyện, tuy nhiên vẫn còn hạn chế về số lượng sự vật, sự việc mà mô hình đã học được. Để sản phẩm có thể thương mại hóa cần có thời gian huấn luyện mô hình trên những tập dữ liệu lớn hơn, đồng thời kết hợp hệ thống feedback để cải thiện bộ dữ liệu đã học.

4. Kết luận

Qua từng bước cải thiện, mô hình tạo mô tả cho ảnh đã tương đối hoàn thiện. Những mô tả sinh ra đã có độ chính xác tương đối cao và tự nhiên hơn trong cách diễn đạt. Tuy nhiên, bộ dữ liệu huấn luyện hiện tại là tập Flickr8k, tuy số lượng ảnh lớn nhưng chưa bao quát được nhiều sự vật, sự việc của thế giới xung quanh. Vì vậy còn gặp những vật thể lạ mà mô hình không thể nhận diện mà mô tả cho đúng được. Việc này có thể khắc phục ở quy mô sản phẩm bằng cách huấn luyện trên các bộ dữ liệu lớn hơn (COCO dataset, Flickr30k).

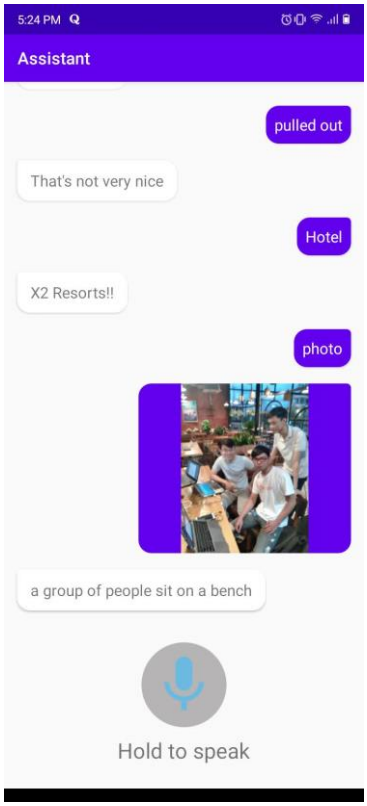
5. Giới thiệu ứng dụng trợ lí ảo giúp đỡ người khiếm thị

Để tiếp cận tới người dùng, mô hình sinh mô tả ảnh được triển khai dưới dạng API và ứng dụng vào phần mềm trợ lí ảo trên điện thoại.



Hình 13. Mô hình ứng dụng.

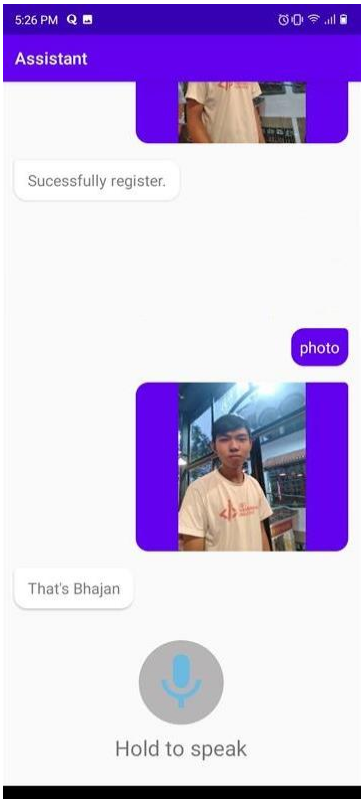
Ứng dụng Assistance được xây dựng trên kết quả thu được sau nghiên cứu. Ứng dụng được triển khai dưới dạng ứng dụng Android với API sử dụng Flask. Ứng dụng có khả năng giao tiếp với người dùng bằng giọng nói, miêu tả cảnh vật xung quanh thông qua ảnh được chụp bằng camera, nhận diện người quen đã được đăng kí khuôn mặt.



Hình 14. Giao diện chức năng mô tả hình ảnh.

Ứng dụng giao tiếp với người dùng qua micro. Khi câu người dùng nói có từ "photo" thì ứng dụng sẽ nhận mở camera để chụp ảnh. Ảnh sau khi chụp sẽ được gửi lên server thông qua Flask API và kết quả nhận được từ server sẽ là nội dung của tấm ảnh. Chức năng này giống như việc camera thay thế cho mắt người để quan sát cảnh vật xung quanh và sau đó sẽ miêu tả lại cho người dùng thông qua loa.

Ngoài việc miêu tả tấm ảnh, ứng dụng còn có thể nhận diện người quen thông qua việc đăng kí khuôn mặt (Hình 15).



Hình 15. Giao diện chức năng đăng kí và nhận diện khuôn mặt.

Tài liệu tham khảo

[1] Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks

[2] Deep Residual Learning for Image Recognition.Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun.

[3] Neural Machine Translation by Jointly Learning to Align and Translate. Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. arXiv: 1409.0473, Sep 2014.

[4] Effective Approaches to Attention-based Neural Machine Translation. Minh-Thang Luong, Hieu Pham, Christopher D. Manning. arXiv:1508.04025, Sep 2015.

[5] Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun sCho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio. arXiv: 1502.03044, Apr 2016.

[6] BLEU: a Method for Automatic Evaluation of Machine Translation. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing ZhuIBM T. J. Watson Research CenterYorktown Heights, NY 10598, USA.