TẠP CHÍ KHOA HỌC
TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH
Tập 20, Số 1 (2023): 153-164
Website: https://journal.hcmue.edu.vn
ISSN: 2734-9918

IO CHI MINH CITY UNIVERSITY OF EDUCATION
JOURNAL OF SCIENCE
Vol. 20, No. 1 (2023): 153-164
https://doi.org/10.54607/hcmue.js.20.1.3627(2023)

Bài báo nghiên cứu

TÍCH HỢP YẾU TỐ CẢM XÚC VÀO NGỮ CẢNH TRONG HỆ THỐNG HỘI THOẠI ĐA PHƯƠNG THỨC

Lê Nguyễn Thuỳ Dương*, Lê Ngọc Tuấn*, Nguyễn Hồng Bửu Long

Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Thành phố Hồ Chí Minh, Việt Nam

**Tác giả liên hệ: Lê Nguyễn Thùy Dương – Email: lethuyduong2000@gmail.com*

Ngày nhận bài: 11-10-2022; ngày nhận bài sửa: 09-11-2022; ngày duyệt đăng: 04-01-2023

TÓM TẮT

Hệ thống hội thoại thuần văn bản sử dụng hướng tiếp cận seq2seq đã xuất hiện nhiều trong các công trình nghiên cứu những năm qua. Tuy nhiên, ngoài việc hội thoại hoàn toàn bằng văn bản thì hình ảnh và cảm xúc cũng là những yếu tố quan trọng. Năm 2021, Zheng và các cộng sự (Zheng et al., 2021) đã đưa ra mô hình cơ sở MOD, mô hình có thể đối thoại bằng văn bản, hình ảnh và có thể phân loại cảm xúc. MOD đã tận dụng thành công sức mạnh của mô hình ngôn ngữ lớn, tuy nhiên ngữ cảnh đầu vào không tận dụng được yếu tố cảm xúc. Vì vậy, chúng tôi thực hiện cải tiến mô hình MOD bằng cách bổ sung sự ràng buộc của yếu tố cảm xúc đối với hai yếu tố còn lại (văn bản, hình ảnh) nhằm mục đích tăng chất lượng thông tin trong ngữ cảnh. Ngoài ra, chúng tôi cũng thực hiện khảo sát hiệu quả khi thêm đặc trưng ảnh được trích xuất từ mạng CNN, nhằm tăng chất lượng đặc trưng ảnh cho ngữ cảnh đầu vào. Thực nghiệm thu được kết quả là tăng 0,19 điểm BLEU-4 và giảm 4,6 ở độ đo Perplexity so với MOD, kết quả cho thấy mô hình cải tiến hoạt động hiệu quả hơn khi có thêm sự ràng buộc của yếu tố cảm xúc trong ngữ cảnh.

Từ khóa: hệ thống hội thoại đa phương thức; học đa tác vụ; mô hình ngôn ngữ lớn; ràng buộc ngữ cảnh; yếu tố cảm xúc

1. Giới thiệu

Hệ thống hội thoại đã được triển khai rộng rãi ở nhiều lĩnh vực trong cuộc sống hiện nay và có rất nhiều tiềm năng trong tương lai. Nhờ vào sự phát triển nhanh của lĩnh vực học sâu trong những năm qua, hệ thống hội thoại đã dần trở thành một trong những hướng nghiên cứu quan trọng trong lĩnh vực xử lí ngôn ngữ tự nhiên.

Trong 5 năm gần đây, hệ thống hội thoại thuần văn bản đã xuất hiện rất nhiều trong các nghiên cứu. Nhưng ngoài việc đối thoại hoàn toàn bằng văn bản thì hình ảnh và cảm xúc cũng là một yếu tố quan trọng không thể thiếu trong các cuộc hội thoại trên các nền tảng nhắn tin trực tuyến ngày nay. Khả năng tương tác kết hợp giữa các phương thức này cho phép con người có thể diễn đạt cảm xúc và ý nghĩ một cách sinh động, hiệu quả hơn.

Cite this article as: Le Nguyen Thuy Duong, Le Ngoc Tuan, & Nguyen Hong Buu Long (2023). Integrating sentiment factors into the context of a multimodal dialogue system. *Ho Chi Minh City University of Education Journal of Science, 20*(1), 153-164.

Các mô hình hệ thống hội thoại đa phương thức trước đây đều tuân theo kiến trúc phân tầng. Các lượt nói trong phần ngữ cảnh được xử lí cục bộ, sau đó được nối lại với nhau bằng phép nối. Do đó, sự ràng buộc xuyên suốt toàn bộ các lượt nói là không tốt. Trong nghiên cứu này, chúng tôi mong muốn xây dựng một hệ thống hội thoại đa phương thức cho ngôn ngữ tiếng Anh, tận dụng khả năng từ các mô hình ngôn ngữ lớn đã được tiền huấn luyện, với đầu vào là ngữ cảnh chứa các yếu tố đa phương thức – kết hợp văn bản, hình ảnh và cảm xúc.

Phạm vi nghiên cứu sẽ chỉ giới hạn ở bộ dữ liệu chứa các cuộc hội thoại miền mở cho tiếng Anh, do đó sẽ không có các hướng phân tích và xử lí liên quan đến kiến thức nền chuyên biệt cho một lĩnh vực cụ thể trong doanh nghiệp (bán lẻ, thương mại điện tử...). Ngoài ra, phần hình ảnh trong tập dữ liệu thực nghiệm sẽ chỉ bao gồm những ảnh meme, sticker được sử dụng rộng rãi trong các nền tảng nhắn tin trực tuyến hiện nay. Tập dữ liệu thực nghiệm sẽ rất hạn chế sự có mặt của các hình ảnh liên quan đến vật thể ngoài đời thật như con người, động vật, thiên nhiên...

Cách tiếp cận cho bài toán sinh câu của mô hình sẽ chỉ giới hạn ở hướng tiếp cận seq2seq. Cách tiếp cận chủ yếu của đề tài là tận dụng nguồn tri thức từ mô hình ngôn ngữ lớn tự hồi quy đã được tiền huấn luyện, sau đó xác định bài toán cụ thể và thực hiện thay đổi sao cho có thể tích hợp sự phụ thuộc của các yếu tố đa phương thức. Hệ thống hội thoại cuối cùng phải có khả năng học đa nhiệm (multitask learning) để giải những bài toán sau: sinh câu trả lời (bài toán chính), dự đoán cảm xúc (bài toán phụ), phân lớp ý định sử dụng ảnh cho câu trả lời (bài toán phụ).

2. Đối tượng và phương pháp nghiên cứu

Mô hình cơ sở của nghiên cứu này là mô hình MOD (Zheng et al., 2021), nhóm tác giả đã tiên phong trong việc đưa mô hình ngôn ngữ lớn vào hệ thống hội thoại đa phương thức. Tuy nhiên, chúng tôi nhận thấy ngữ cảnh của mô hình cơ sở thiếu đi yếu tố cảm xúc, mà yếu tố này rất quan trọng trong các cuộc hội thoại miền mở. Vì vậy, chúng tôi đề xuất giải pháp ràng buộc các yếu tố đa phương thức chặt chẽ hơn. Ngoài ra, chúng tôi cũng thực hiện khảo sát hiệu quả của đặc trưng ảnh được trích xuất từ mạng CNN, đã được tiền huấn luyện, nhằm giúp mô hình nhận biết ngữ cảnh tốt hơn.

Mô hình đề xuất mà chúng tôi sử dụng là mô hình hệ thống hội thoại đa phương thức miền mở có thể sinh câu trả lời cùng với cảm xúc (nếu có). Người dùng và hệ thống có thể nói chuyện về bất kì chủ đề nào trong cuộc sống với ngôn ngữ là tiếng Anh, tuy nhiên, chúng tôi cần đưa ra giới hạn cho bài toán như sau:

- Định nghĩa lượt nói: mỗi lượt nói đến từ một người nói khác nhau. Trong một cuộc hội thoại, sẽ chỉ giới hạn 2 người tham gia.
- Nội dung lượt nói: mỗi lượt nói chỉ đề cập đến một chủ đề, một thông tin nào đó xuyên suốt. Cụ thể hơn, tình huống 2 người trao đổi song song cùng lúc nhiều câu chuyện sẽ nằm ngoài phạm vi bài toán mà chúng tôi đặt ra. Do đó, mỗi lượt nói chỉ thể hiện một cảm xúc.

- Các phương thức trong ngữ cảnh: phương thức chính trong bài toán là văn bản, các phương thức khác như hình ảnh và cảm xúc không nhất thiết luôn có mặt trong toàn bộ các lượt nói. Điều này cũng hoàn toàn phù hợp với tình huống thực tế, các lượt nói không phải lúc nào cũng đi kèm với hình ảnh và cảm xúc mà còn tuỳ thuộc vào mong muốn của người nói.
- Hình ảnh: chúng tôi sử dụng tập dữ liệu MOD (Zheng et al., 2021) để huấn luyện mô hình. Hình ảnh được giới hạn là các ảnh meme, nhãn dán, hạn chế sự có mặt của các vật thể ngoài đời thực. Tuy nhiên, hướng tiếp cận của mô hình cơ sở và mô hình cải tiến là hoàn toàn tổng quát, có thể áp dụng cho bất kì loại hình ảnh nào từ đời thực.

Về hướng tiếp cận, mô hình sẽ tận dụng khả năng của mô hình ngôn ngữ lớn đã được tiền huấn luyện, được tích hợp khả năng học đa nhiệm để giải quyết thêm hai bài toán phụ (dự đoán cảm xúc và dự đoán ý định sử dụng ảnh cho câu trả lời) để bổ trợ cho bài toán chính là sinh câu.

Mô hình cơ sở tận dụng mô hình ngôn ngữ lớn đã được tiền huấn luyện CDial-GPT (Wang et al., 2020) để thực hiện học chuyển giao cho bài toán sinh câu. Tuy nhiên, do thực nghiệm trên bộ dữ liệu tiếng Anh mà mô hình CDialGPT chỉ dành cho tiếng Trung Quốc nên chúng tôi quyết định thực hiện học chuyển giao dựa trên một mô hình ngôn ngữ lớn khác là DialoGPT (Zhang et al., 2019). Chúng tôi sử dụng DialoGPT vì đây là mô hình ngôn ngữ lớn sử dụng chung kiến trúc với CDial-GPT. Hơn nữa, DialoGPT tận dụng tri thức từ GPT-2 (Radford et al., 2019) đã được tiền huấn luyện tự giám sát, sau đó được huấn luyện thêm thông qua lượng lớn dữ liệu hội thoại từ Reddit. Do đó, nó sẽ tối ưu quá trình huấn luyện cho bài toán xây dựng hệ thống hội thoại đa phương thức hơn so với các phiên bản mô hình ngôn ngữ lớn khác. Sau đây là 3 bài toán cần giải quyết:

- Bài toán sinh câu – Text response (TR): cho x, đầu vào là ngữ cảnh đa phương thức từ lượt nói đầu tới lượt nói hiện tại, thực hiện dự đoán từ tiếp theo cho câu trả lời. Với N là tổng số từ có trong chuỗi (tổng số token tối đa).

$$L_{TR} = -\sum_{i=1}^{N} \quad logp(x, y_1, \dots, y_{i-1}; \theta) (1)$$

- Bài toán dự đoán cảm xúc - Emotion prediction (EP) có đầu vào là ngữ cảnh từ lượt nói đầu tới lượt nói hiện tại (x) và lượt trả lời (y_{TR}), thực hiện dự đoán cảm xúc của câu trả lời. Đây là bài toán phân lớp truyền thống. Với y_{EP} là nhãn đúng cho cảm xúc.

$$L_{EP} = -logp(x, y_{TR}; \theta) (2)$$

- Bài toán dự đoán ý định sử dụng ảnh cho câu trả lời – Intent classification (IC) có đầu vào là ngữ cảnh từ lượt nói đầu tới lượt nói hiện tại (x) và lượt trả lời (y_{TR}), thực hiện dự đoán câu trả lời hiện tại có cần sử dụng thêm ảnh hay không. Với C là số lượng nhãn. Đây là bài toán phân lớp nhị phân truyền thống. Với y_{IC} là nhãn đúng.

$$L_{IC} = -logp(x, y_{TR}; \theta) (3)$$

Bài toán học đa nhiệm. Cho phần ngữ cảnh của lượt nói hiện tại, mô hình được huấn luyện giải quyết đồng thời cả 3 bài toán trên. Hàm mất mát tổng cho quá trình học đa nhiệm là tổng có trọng số của các hàm mất mát thành phần, với $\lambda_1, \lambda_2, \lambda_3$ là 3 siêu tham số được tinh chỉnh.

$$L_{ML} = \lambda_1 L_{TR} + \lambda_2 L_{EP} + \lambda_3 L_{IC} (4)$$

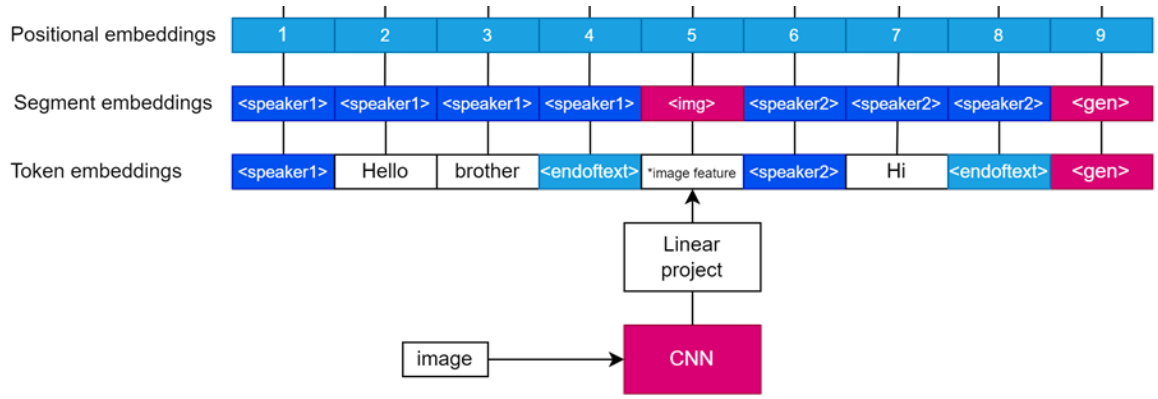
Kiến trúc của mô hình cơ sở được thừa hưởng ý tưởng từ GPT-2, bao gồm n khối Transformer Decoder chồng lên nhau. Đầu vào có 3 dạng vector nhúng:

- Token embeddings: đây là đặc trưng văn bản được trích xuất trực tiếp từ ma trận trọng số Word Embedding có trong mô hình. Ma trận này ánh xạ 1 token (1 từ) sang vector đặc trưng tương ứng.
- Positional embeddings: vector nhúng biểu diễn vị trí của token trong câu. Phép nhúng này tương tự như cơ chế nhúng vị trí của kiến trúc Transformers.
- Segment embeddings: vector nhúng đặc biệt để thể hiện trạng thái của token cùng số chỉ mục trong token embeddings, được trích xuất trực tiếp từ ma trận trọng số Segment embeddings, với số cột là số token đặc biệt được sử dụng trong Segment embeddings.

Mô hình có 5 loại token đặc biệt sau:

- <endoftext>: token đánh dấu kết thúc lượt nói;
- <speaker1>: đánh dấu lượt nói của người thứ nhất;
- <speaker2>: đánh dấu lượt nói của người thứ hai;
- <img>: đánh dấu token hiện tại là đặc trưng ảnh (trích xuất từ mô hình tích chập);
- <gen>: đánh dấu token hiện tại là một token đặc biệt. Lớp ẩn cuối của token sẽ chứa toàn bộ thông tin xuyên suốt câu. Mô hình sẽ dùng thông tin đầu ra của token này đưa qua những lớp tuyến tính để thực hiện giải quyết các bài toán phụ (dự đoán cảm xúc, ý định sinh ảnh).

Riêng đối với phần đặc trưng ảnh (*image features), được trích xuất từ mô hình tích chập lớn EfficientNet (Tan et al., 2019), chúng tôi sẽ cho mô hình sử dụng lần lượt 2 phiên bản: EfficientNet với đặc trưng ảnh mặc định và EfficientNet được huấn luyện thêm trên ảnh của tập dữ liệu huấn luyện, học phân lớp cảm xúc của lượt nói. Mục đích không phải để mạng tích chập học cách phân lớp mà là để nó học cách biểu diễn đặc trưng ảnh mà trước đây chưa từng được thấy qua, từ đó nó sẽ rút trích được đặc trưng chất lượng hơn. Chúng tôi thực hiện điều này nhằm đánh giá độ hiệu quả của đặc trưng ảnh khi chúng được sinh ra bởi mạng tích chập, và mạng này đã hiểu được cách biểu diễn ảnh từ tập huấn luyện. Đầu ra của mạng tích chập sẽ là một vector biểu diễn đặc trưng cho ảnh đầu vào. Đẩy vector này qua một lớp biến đổi tuyến tính (Fully connected) để đưa về cùng kích thước ẩn của mô hình. Sau đó vector này được đưa vào phần ngữ cảnh ngay vị trí token với chú thích (*image features). Nếu như câu hiện tại không có ảnh, phần token cho ảnh sẽ không có mặt.

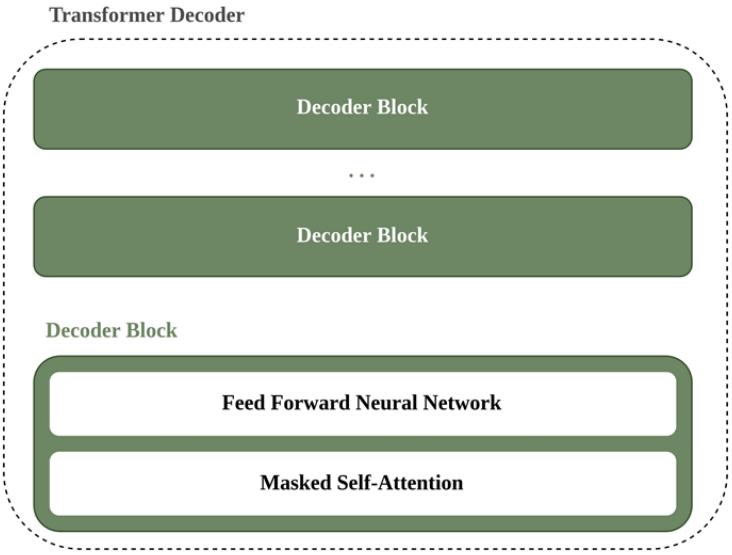


Hình 1. *Mô tả đầu vào của mô hình*

Cả ba dạng biểu diễn đầu vào sẽ được cộng lại có được biểu diễn đa phương thức đầu vào E^t cho lượt nói hiện tại:

$$E^t = TE^t + PE^t + SE^t (5)$$

Biểu diễn đa phương thức đầu vào sẽ truyền qua Masked multihead attention (Vaswani et al, 2017) với Query (Q), Key (K), Value (V) đều là E^t, bên trong các khối Transformer Decoder (DB). Mô hình chứa 12 khối Decoder, mỗi khối có hai lớp: Masked Seft-Attention và Mạng truyền thẳng theo vị trí (Position-wise feed-forward network).

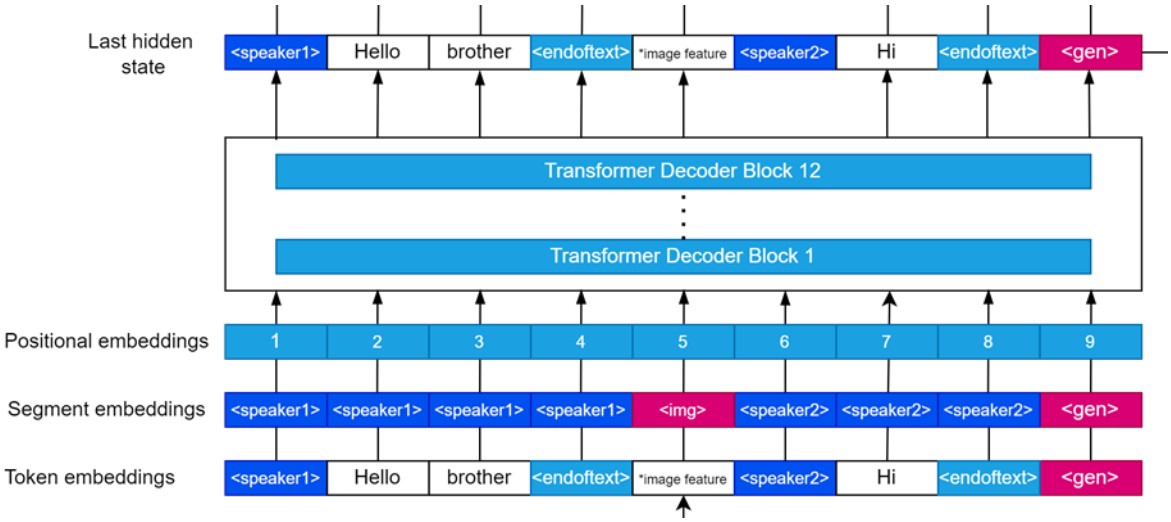


Hình 2. *Kiến trúc của một khối Transformer Decoder*

Đầu ra của mỗi khối Decoder là đầu vào của khối Decoder tiếp theo, cho tới khối cuối cùng. Đầu ra của mô hình là lớp ẩn cuối (last hidden state), bản chất nó vẫn là chuỗi đầu vào ban đầu nhưng đã được tăng cường thêm thông tin xuyên suốt ngữ cảnh đa phương thức. Mô hình sẽ dựa vào thông tin lớp ẩn này để đưa ra giải quyết 3 bài toán đặt ra:

- Sinh câu: mô hình sẽ sử dụng thông tin lớp ẩn cuối của các token, đưa qua một lớp biến đổi tuyến tính có kích thước là tập từ vựng để tạo ra phân phối xác suất của tập từ vựng, từ đó có thể dự đoán được các từ tiếp theo tương ứng;

- Dự đoán cảm xúc: mô hình sử dụng thông tin lớp ẩn cuối của token <gen>. Token này đứng ở cuối câu nên có khả năng nắm trọn toàn bộ thông tin của các token trước nó (theo cơ chế của masked multihead-attention). Thực hiện đẩy lớp ẩn cuối của token <gen> qua một lớp tuyến tính với đầu ra là số nhãn cảm xúc, ta có được đầu ra tương ứng;
- Dự đoán ý định sử dụng ảnh kèm câu trả lời: thực hiện tương tự với dự đoán cảm xúc.

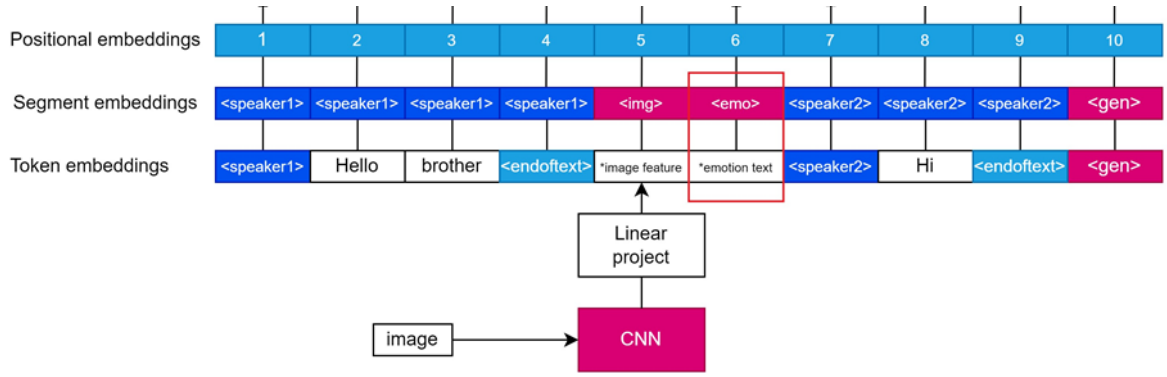


Hình 3. *Kiến trúc của một khối Transformer Decoder*

Quá trình suy diễn: mô hình nhận đầu vào là ngữ cảnh (văn bản, hình ảnh), sau đó thực hiện sinh câu trả lời. Mô hình thực hiện dự đoán từng token tiếp theo cho tới khi gặp token <endoftext> thì dừng lại. Sau đó thực hiện gắn thêm token <gen> ở cuối. Truyền thẳng chuỗi cuối cùng này qua mô hình một lần nữa, lấy ra được lớp ẩn cuối của token <gen>. Từ đó dự đoán được cảm xúc và ý định sử dụng ảnh cho câu trả lời.

Cải tiến cho mô hình cơ sở: Nhận thấy mô hình cơ sở không có sự tham gia của yếu tố cảm xúc trong phần ngữ cảnh, chúng tôi quyết định giải quyết vấn đề này. Giải pháp là đưa yếu tố cảm xúc vào phần ngữ cảnh, tương ứng theo từng lượt nói (lượt nào không có cảm xúc sẽ được bỏ qua). Cảm xúc có dạng là dạng nhãn định danh được chuyển thành văn bản mô tả cảm xúc tương ứng, sau đó đưa vào ngữ cảnh ở giữa những lượt nói. Bằng cách đó, ràng buộc trong ngữ cảnh có thêm sự tham gia của phương thức cảm xúc. Khi cảm xúc đã xuất hiện trong chuỗi đầu vào, mô hình sẽ học được sự tương tác giữa thông tin của cảm xúc với thông tin của văn bản và hình ảnh bằng cơ chế tập trung.

Ngoài ra, để tăng hiệu quả của mô hình trong việc hiểu được thái độ và cảm xúc của câu, mô hình cải tiến còn đưa thêm một token mô tả cảm xúc ở cuối (sau token <gen>), nhằm ép mô hình dự đoán token này – văn bản mô tả cảm xúc của câu trả lời. Mô hình lúc này buộc phải dự đoán yếu tố cảm xúc của câu trả lời dưới 2 dạng: nhãn định danh và token văn bản. Điều này giúp mô hình tăng khả năng hiểu ngữ cảnh hơn nữa, đồng thời cũng tạo ra sự ràng buộc mạnh mẽ hơn cho phương thức cảm xúc với các phương thức còn lại (văn bản, hình ảnh).



Hình 4. *Kiến trúc của một khối Transformer Decoder*

Các thiết lập thực nghiệm: Cả 2 mô hình cơ sở và mô hình cải tiến đều sử dụng chung một thiết lập để đảm bảo tính công bằng. Các thiết lập này đã được chúng tôi tinh chỉnh để rút ra được kết quả thực nghiệm tối ưu. Chúng tôi sử dụng thuật toán tối ưu AdamW (Loshchilov et al., 2017) (biến thể amsgrad) để học các trọng số của mô hình với hệ số học gốc là 3e-5, cùng với thuật toán điều khiển hệ số học Cyclical (Smith et al., 2017), với hệ số học tối đa là 5e-3. Chúng tôi huấn luyện hệ thống trên 8 epoch với kích thước batch là 64. Độ dài chuỗi tối đa cho phép là 256 (phần ngữ cảnh tối đa 233, phần trả lời tối đa 23). Thuật toán tách từ được sử dụng là Byte Pair Encoding (Sennrich et al., 2015). Chúng tôi sử dụng mô hình ngôn ngữ lớn được tiền huấn luyện DialoGPT (Zhang et al., 2019), phiên bản "small" với 125 triệu tham số, kiến trúc có kích thước ẩn là 768 và 12 khối Transformer Decoder. Mô hình có thêm 2 lớp neural truyền thẳng (không lớp ẩn), dành cho 2 bài toán phụ. Trọng số tiêu biến của mô hình (weight decay) được đặt mặc định là 0.001. Thuật toán sinh câu được lựa chọn trong quá trình đánh giá là Lấy mẫu Nucleus, với giá trị ngưỡng p bằng 0.75 và giá trị nhiệt độ temperature là 1 theo mặc định. Quá trình tinh chỉnh siêu tham số được thực hiện trên tập đánh giá, sau đó mới được đánh giá và lấy kết quả cuối cùng trên tập kiểm tra.

Ngôn ngữ lập trình cho quá trình thực nghiệm được sử dụng là Python, cùng với thư viện Pytorch và HuggingFace để lập trình học sâu cho mô hình ngôn ngữ. Phần thực nghiệm được tiến hành chạy trên 1 GPU NVIDIA A100 40GB PCIe, và một lần huấn luyện sẽ tốn khoảng 1,5 ngày.

3. Kết quả và thảo luận

Đối với **bài toán sinh câu**, chúng tôi sử dụng độ đo Perplexity và BLEU với 2-gram, 3-gram và 4-gram. Ngoài ra, độ đo Distinct n-gram (Dist-n) (Li et al., 2015) được sử dụng để đo mức độ đa dạng trong cách sử dụng ngôn ngữ với n là 1 và 2. Các dấu sao (*) ở trước tên mô hình đại diện cho việc mô hình này sử dụng đặc trưng ảnh được tạo ra từ mạng tích chập đã được tiền huấn luyện trên ảnh của tập dữ liệu huấn luyện. Để thuận tiện cho việc gọi tên, chúng tôi sẽ gọi "mô hình*" để phân biệt với mô hình dùng đặc trưng gốc. Kết quả chi tiết ở Bảng 1.

Bảng 1. *Kết quả thực nghiệm bài toán sinh câu*

	Perplexity	**BLEU-2**	**BLEU-3**	**BLEU-4**	**Dist-1**	**Dist-2**
Mô hình cơ sở	26,65	14,68	4,35	1,01	2,81	17,58
Mô hình cải tiến	**22,04**	**14,88**	**4,44**	**1,13**	2,75	16,09
*Mô hình cơ sở	26,54	14,73	4,40	1,09	2,95	17,93
*Mô hình cải tiến	**21,94**	**14,91**	**4,57**	**1,28**	2,71	15,64

Kết quả thực nghiệm cho thấy mô hình cải tiến đã có những cải thiện ở độ đo Perplexity (giảm 4,6 ở mô hình thường và 4,6 ở mô hình*), BLEU-2 (tăng 0,2 ở mô hình thường và 0,18 ở mô hình*), BLEU-3 (tăng 0,09 ở mô hình thường và 0,17 ở mô hình*), BLEU-4 (tăng 0,12 ở mô hình thường và 0,19 ở mô hình*). Tuy nhiên, có thể thấy mô hình cải tiến bị giảm chỉ số Distinct-n gram đi 1 lượng nhỏ. Mô hình* cũng hoàn toàn tốt hơn so với mô hình gốc ở cả phiên bản cơ sở lẫn cải tiến.

Đối với **bài toán dự đoán cảm xúc**, chúng tôi sử dụng độ chính xác (accuracy) top 1, top 3 và top 5. Sử dụng thêm độ đo weighted F1 để đánh giá chính xác hơn vì nhãn cảm xúc có sự mất cân bằng nên cần có sự tham gia của tỉ lệ nhãn trong kết quả đánh giá. Kết quả chi tiết ở Bảng 2.

Bảng 2. *Kết quả thực nghiệm bài toán dự đoán cảm xúc*

	Accuracy top1	**Accuracy top3**	**Accuracy top5**	**Weighted F1**
Mô hình cơ sở	67,77	73,09	76,52	56,65
Mô hình cải tiến	**69,01**	**75,24**	**78,95**	**57,42**
*Mô hình cơ sở	68,30	75,38	79,45	56,79
*Mô hình cải tiến	**69,49**	**78,17**	**82,94**	**57,54**

Từ kết quả ở Bảng 2, ta thấy rõ mô hình cải tiến tăng mạnh đối với cả mô hình bình thường và mô hình* trên tất cả các độ đo: Accuracy top 1 (tăng 1,3%, 1,19%), Accuracy top 3 (tăng 2,2%, 2,79%), Accuracy top 5 (tăng 2,43%, 3,49%), và điểm Weighted F1 (tăng 0,7%, 0,75%).

Kết quả thực nghiệm của **bài toán phân lớp ý định dùng ảnh** thể hiện chi tiết ở Bảng 3. Kết quả cho thấy mô hình cải tiến tăng nhẹ trên Accuracy (tăng 0,6%, 1,13%) và Weighted F1 (tăng 0,5%, 0,1%).

Bảng 3. *Kết quả thực nghiệm bài toán phân lớp ý định dùng ảnh*

	Accuracy	**Weighted F1**
Mô hình cơ sở	58,22	53,08
Mô hình cải tiến	**58,84**	**54,53**
*Mô hình cơ sở	59,06	54,47
*Mô hình cải tiến	**60,19**	**55,57**

Dựa vào những kết quả khả quan trên, có thể thấy việc tích hợp yếu tố cảm xúc vào sự phụ thuộc ngữ cảnh, và việc tinh chỉnh mô hình dự đoán phần token mô tả cảm xúc đã đem lại những cải thiện rõ rệt. Mô hình lúc này phải dự đoán yếu tố cảm xúc dưới 2 dạng là nhãn định danh và token văn bản. Vì vậy, mô hình đã học được thêm thông tin từ yếu tố cảm xúc và sự ràng buộc của cảm xúc với văn bản và hình ảnh, từ đó, giúp nó hiểu sâu được ngữ cảnh hơn nữa và củng cố thêm chất lượng dự đoán của mô hình. Đầu ra của mô hình được cải tiến khi có thêm sự đóng góp của yếu tố cảm xúc. Tuy thay đổi trên có phần làm cho mô hình cải tiến sinh câu kém đa dạng hơn so với mô hình cơ sở, nhưng ngược lại nó lại hiệu quả hơn trong việc xử lí các bài toán đặt ra. Mô hình* hoàn toàn tốt hơn so với mô hình sử dụng đặc trưng gốc. Mạng tích chập học cách biểu diễn ảnh trong tập huấn luyện đã cung cấp đặc trưng ảnh tốt hơn. Chất lượng đặc trưng ảnh tốt hơn giúp mô hình nhận biết được ngữ cảnh sâu hơn, và đạt được kết quả đánh giá khả quan hơn.

Về phần suy diễn, chúng tôi đưa ra hai kết quả thể hiện Hình 4 và Hình 5, mô hình cải tiến có phần vượt trội hơn khi câu sinh ra có chất lượng tốt hơn so với mô hình cơ sở. Câu trả lời không phải lúc nào cũng ổn định vì có thêm thông tin cảm xúc mà hệ thống dự đoán. Tuy nhiên, cả hai mô hình đều có thể sinh ra các câu trả lời hoàn chỉnh, hợp với văn phong nói và ngữ cảnh của cuộc hội thoại.

Hình 3. *Kết quả suy diễn thứ nhất*

Hình 4. *Kết quả suy diễn thứ hai*

4. Kết luận

Trong nghiên cứu này, chúng tôi đã tiến hành khảo sát, nghiên cứu các hướng tiếp cận cho bài toán xây dựng hệ thống hội thoại đa phương thức. Chúng tôi đã sử dụng hướng tiếp cận là tận dụng khả năng của mô hình ngôn ngữ lớn để trích xuất thông tin văn bản, hình ảnh, cảm xúc từ các cuộc hội thoại, đồng thời tạo sự ràng buộc ngữ cảnh cho các phương thức khác nhau. Mô hình có khả năng học đa nhiệm, xử lí cùng lúc 3 bài toán. Chúng tôi cũng thực hiện tiền huấn luyện phân lớp cho mô hình CNN nhằm tăng chất lượng đặc trưng ảnh đầu vào cho mô hình hội thoại.

Từ đó, chúng tôi tiến hành thực nghiệm và đạt được kết quả tốt hơn so với mô hình cơ sở. Việc đưa yếu tố cảm xúc vào ngữ cảnh và thực hiện tinh chỉnh mô hình dự đoán cảm xúc dưới dạng nhãn định danh và token văn bản miêu tả cảm xúc đã tạo được sự phụ thuộc đầy đủ đa phương thức cho ngữ cảnh. Ngoài ra, việc cho mô hình tích chập học trước cách biểu diễn ảnh từ tập huấn luyện cũng giúp cải thiện đáng kể kết quả thực nghiệm.

Trong tương lai, chúng tôi đề xuất 4 hướng phát triển cho đề tài này. Thứ nhất, nghiên cứu và thực hiện tích hợp thêm hướng giải quyết cho bài toán truy hồi ảnh (image retrieval). Mô hình sẽ có khả năng học đa nhiệm giải quyết cùng lúc 4 bài toán: sinh câu, dự đoán cảm xúc, phân lớp ý định dùng ảnh cho câu trả lời, và truy hồi ảnh. Thứ hai, trích xuất hình ảnh thành 1 chuỗi các token trong ngữ cảnh như hướng tiếp cận của Vision Transformers (Dosovitskiy et al., 2020). Thứ ba, thực hiện kết hợp mô hình ngôn ngữ lớn với kiến trúc phân tầng, theo như ý tưởng của mô hình Hierarchical Transformer (Santra et al., 2020). Việc kết hợp mô hình ngôn ngữ lớn với kiến trúc phân tầng, cùng với sự tham gia của các phương thức khác (hình ảnh, cảm xúc) sẽ tạo ra một hướng tiếp cận khá mới mẻ và đầy tiềm năng cho bài toán xây dựng hệ thống hội thoại đa phương thức. Thứ tư, sử dụng kết hợp hướng tiếp cận truy hồi và sinh câu cho hệ thống hội thoại đa phương thức (Zhang et al., 2021) trên hệ thống hội thoại truyền thống. Đây cũng là một hướng tiếp cận hoàn toàn mới mẻ cho bài toán xây dựng hệ thống hội thoại đa phương thức, khắc phục được điểm yếu của hướng tiếp cận sinh câu. Tổng kết lại, để đạt được những kết quả khả quan hơn nữa trên bài toán xây dựng hệ thống hội thoại đa phương thức, và áp dụng kết quả nghiên cứu vào thực tiễn, chúng ta cần những bước tiến nghiên cứu vượt trội hơn nữa trong tương lai.

❖ ***Tuyên bố về quyền lợi:*** *Các tác giả xác nhận hoàn toàn không có xung đột về quyền lợi.*

❖ ***Lời cảm ơn:*** *Chúng tôi xin chân thành cảm ơn Trung tâm Ngôn ngữ học Tính toán, Đại học Khoa học Tự nhiên, ĐHQG-HCM hỗ trợ trong suốt quá trình nghiên cứu.*

TÀI LIỆU THAM KHẢO

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones,… Illia Polosukhin. (2017). Attention is all you need. In Proceedings of the 31st *International Conference on Neural Information Processing Systems* (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000-6010.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P.,... & Amodei, D. (2020). *Language models are few-shot learners*. Advances in neural information processing systems, (33), 1877-1901.

Fei, Z., Li, Z., Zhang, J., Feng, Y., & Zhou, J. (2021). *Towards Expressive Communication with Internet Memes: A New Multimodal Conversation Dataset and Benchmark*. ArXiv, abs/2109.01839.

Weidong He, Zhi Li, Dongcai Lu, Enhong Chen, Tong Xu, Baoxing Huai, & Jing Yuan. (2020). Multimodal Dialogue Systems via Capturing Context-aware Dependencies of Semantic Elements. In Proceedings of the 28th ACM International Conference on Multimedia (MM '20). *Association for Computing Machinery*, New York, NY, USA, 2755-2764. https://doi.org/10.1145/3394171.3413679

Li, J., Galley, M., Brockett, C., Gao, J., & Dolan, B. (2015). *A diversity-promoting objective function for neural conversation models*. arXiv preprint arXiv:1510.03055.

Loshchilov, I., & Hutter, F. (2017). *Decoupled weight decay regularization*. arXiv preprint arXiv:1711.05101.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*.

Saha, A., Khapra, M., & Sankaranarayanan, K. (2018, April). Towards building large scale multimodal domain-aware conversation systems. In Proceedings of the AAAI Conference on *Artificial Intelligence* (Vol. 32, No. 1).

Sennrich, R., Haddow, B., & Birch, A. (2015). *Neural machine translation of rare words with subword units*. arXiv preprint arXiv:1508.07909.

Smith, L. N. (2017, March). Cyclical learning rates for training neural networks. In 2017 IEEE winter conference on *Applications of computer vision* (WACV) (pp. 464-472). IEEE.

Tan, M., & Le, Q. V. (2019). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. ArXiv, abs/1905.11946.

Wang, Y., Ke, P., Zheng, Y., Huang, K., Jiang, Y., Zhu, X., & Huang, M. (2020). *A Large-Scale Chinese Short-Text Conversation Dataset*. NLPCC.

Weidong He, Zhi Li, Dongcai Lu, Enhong Chen, Tong Xu, Baoxing Huai, & Jing Yuan. (2020). Multimodal Dialogue Systems via Capturing Context-aware Dependencies of Semantic Elements. In Proceedings of the 28th ACM International Conference on Multimedia (MM '20). *Association for Computing Machinery*, New York, NY, USA, 2755-2764. https://doi.org/10.1145/3394171.3413679

Zhang, Y., Sun, S., Galley, M., Chen, Y. C., Brockett, C., Gao, X.,... Dolan, B. (2019). *Dialogpt: Large-scale generative pre-training for conversational response generation*. arXiv preprint arXiv:1911.00536.

Zhang, Y., Sun, S., Gao, X., Fang, Y., Brockett, C., Galley, M.,... Dolan, B. (2021). *Joint retrieval and generation training for grounded text generation*. arXiv preprint arXiv:2105.06597.

INTEGRATING SENTIMENT FACTORS INTO THE CONTEXT OF A MULTIMODAL DIALOGUE SYSTEM

Le Nguyen Thuy Duong*, Le Ngoc Tuan, Nguyen Hong Buu Long

University of Science, Ho Chi Minh City, Vietnam National University Ho Chi Minh City, Vietnam

**Corresponding author: Le Nguyen Thuy Duong – Email: lethuyduong2000@gmail.com*

Received: October 11, 2022; Revised: November 08, 2022; Accepted: January 12, 2023

ABSTRACT

The text-based dialogue systems using the seq2seq model have been extensively used in recent research. However, besides purely text conversations, images and emotions are also important factors. In 2021, Zheng et al. presented MOD which can dialogue with text, visuals, and classify emotions. In spite of the promising performance of MOD, the input context does not use the emotional element. In this article, we improve MOD by adding the sentiment factor binding to the other two factors (text, image) to enhance the quality of the information in the context and help the model capture the context more deeply. Finally, we incorporate image features retrieved from the CNN network for the input context to improve the quality of visual features. Finally, our model improved the BLUE-4 score by 0.19 and the Perplexity by 4.6 compared to MOD. The results show that our model (integrating the sentiment factor into the context) performs better.

Keywords: context-aware dependency; large language model; multimodal dialogue system; multitask learning; sentiment factor.