TRƯỜNG ĐẠI HỌC SƯ PHẠM TPHCM

KHOA CÔNG NGHỆ THÔNG TIN

PHAN LƯƠNG THÙY DƯƠNG

XÂY DỰNG ĐỒ THỊ TRI THỨC CHO PHÂN TÍCH CẢM XÚC TỪ DỮ LIỆU HÌNH ẢNH

KHÓA LUẬN TỐT NGHIỆP

TP. HỒ CHÍ MINH - NĂM 2025

TRƯỜNG ĐẠI HỌC SƯ PHẠM TPHCM

KHOA CÔNG NGHỆ THÔNG TIN

PHAN LƯƠNG THÙY DƯƠNG

XÂY DỰNG ĐỒ THỊ TRI THỨC CHO PHÂN TÍCH CẢM XÚC TỪ DỮ LIỆU HÌNH ẢNH

CHUYÊN NGÀNH: KHOA HỌC MÁY TÍNH

KHÓA LUẬN TỐT NGHIỆP

NGƯỜI HƯỚNG DẪN KHOA HỌC: TS. NGUYỄN VIẾT HƯNG

THS. VÕ LÊ PHÚC HẬU

TP.HCM – NĂM 2025

LỜI CÁM ƠN

Trước hết, em xin chân thành gửi lời cảm ơn sâu sắc đến thầy TS. Nguyễn Viết Hưng và cô Ths. Võ Lê Phúc Hậu là những người thầy, cô đã định hướng, chỉ bảo, giúp đỡ tận tình trong suốt cả thời gian nghiên cứu khóa luận. Với 2 thầy cô có kinh nghiệm và sự tâm huyết với việc nghiên cứu khoa học đã giúp đưa ra nhiều lời khuyên, góp ý hữu ích. Từ đó bài nghiên cứu của em mới hoàn thành một cách trọn vẹn.

Em cũng xin bày tỏ lòng biết ơn đến quý thầy, cô giáo của khoa Công Nghệ Thông Tin đã giảng dạy và truyền đạt kiến thức, kinh nghiệm cho em trong suốt quá trình thực hiện khóa luận tại trường Đại học Sư phạm Thành phố Hồ Chí Minh.

Cuối cùng, em muốn gửi lời cảm ơn đến gia đình và bạn bè cũng như các anh chị khóa trên của em. Những người luôn động viên và ủng hộ để có đủ niềm tin, động lực để hoàn thành khóa luận tốt nghiệp.

Em xin chân thành cảm ơn!

Thành phố Hồ Chí Minh, ngày 24 tháng 03 năm 202

MỤC LỤC

BẢNG CÁC KÝ HIỆU, CHỮ VIẾT TẮT

Từ viết tắt	Từ đầy đủ
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BLIP	Bootstrapping Language-Image Pre-training
CAER	Context Aware Emotion Recognition
CAER-Net	Context-Aware Emotion Recognition Network
CC	Creative Commons
CLIP	Contrastive Language-Image Pre-training
CNN	Convolutional Neural Network
CSL	Circular Skip Link
CV	Computer Vision
EMOTIC	Emotions in Context
GAT	Graph Attention Network
GATs	Graph Attention Networks
GCN	Graph Convolutional Network
GCNs	Graph Convolutional Networks
GIN	Graph Isomorphism Network
GPT-4V	Generative Pre-trained Transformer 4 Vision
GQA	Grounded Question Answering

GraphSAGE	Graph Sample and Aggregate
HECO	Emotion Recognition for Multiple Context Awareness
KG	Knowledge Graph
KTFE	Knowledge-augmented Temporal Feature Extraction
LAION	Large-scale Artificial Intelligence Open Network
LLaVa	Large Language and Vision Assistant
LLM	Large Language Model
LoRA	Low-Rank Adaptation
LSTM	Long Short-Term Memory
mAP	mean Average Precision
MLP	Multi-layer Perceptron
MSE	Mean Squared Error
MUTAG	Mutagenicity
NCI1	National Cancer Institute 1
NLP	Natural Language Processing
OKVQA	Outside Knowledge Visual Question Answering
OpenAL	Open Audio Library
ORC	Optical Character Recognition
PNA	Principal Neighbourhood Aggregation

ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
RoBERTa	Robustly optimized BERT approach
RPN	Region Proposal Network
SBU	Stony Brook University
SK-GCN	Skeleton-based Graph Convolutional Network
VAD	Valence-Arousal-Dominance
ViT	Vision Transformer
VL	Vision Language
VQA	Visual Question Answering
YAGO	Yet Another Great Ontology

BẢNG DANH MỤC CÁC BẢNG BIỂU

BẢNG DANH MỤC CÁC HÌNH VẼ

CHƯƠNG MỞ ĐẦU

1.1. Lý do chọn đề tài.

Trong thời đại hiện nay, phân tích cảm xúc là một trong những lĩnh vực quan trọng trong công nghệ trí tuệ nhân tạo, đặc biệt là trong thời đại mà máy móc và các hệ thống tự động ngày càng tham gia sâu vào đời sống của con người. Trong tương tác người-máy, khả năng hiểu và đáp ứng cảm xúc giúp tạo ra các hệ thống phản ứng linh hoạt và thông minh hơn.

Thực tế, các công ty lớn như Facebook, Google đang đẩy mạnh việc phát triển các hệ thống phân tích cảm xúc trên mạng xã hội để cung cấp nội dung phù hợp, giúp tối ưu hóa trải nghiệm người dùng. Bên cạnh đó, trong lĩnh vực chăm sóc sức khỏe, việc nhận diện cảm xúc tự động còn giúp bác sĩ theo dõi tâm lý bệnh nhân và phát hiện sớm các dấu hiệu rối loạn tâm thần [1]. Trong môi trường giáo dục, phân tích cảm xúc có thể được sử dụng để đánh giá phản ứng của học sinh đối với các phương pháp giảng dạy khác nhau hoặc các bài học cụ thể. Điều này có thể giúp giáo viên điều chỉnh cách dạy để phù hợp hơn với nhu cầu và sự hiểu biết của học sinh [2]. Theo nhiều nghiên cứu lý thuyết về tâm lý học, cảm xúc con người được truyền tải không chỉ qua biểu cảm khuôn mặt mà còn thông qua cử chỉ, tư thế và ngữ cảnh xung quanh [3].

Với các bài nghiên cứu trước đây, các nhà khoa học chủ yếu dựa trên phân tích biểu cảm khuôn mặt [4] hoặc tư thế cơ thể [5]. Tuy nhiên, khi ứng dụng trong các môi trường tự nhiên, các phương pháp này thường gặp hạn chế do không tích hợp đầy đủ yếu tố ngữ cảnh và các thông tin liên quan. Trong khi đó, các nghiên cứu tâm lý học cho thấy ngữ cảnh đóng vai trò quan trọng trong việc hiểu chính xác cảm xúc của con người [6]. Lúc này, Ontology cũng là một phương pháp biểu diễn tri thức có cấu trúc, giúp xác định và mô tả rõ ràng các đối tượng và mối quan hệ giữa chúng. Nó được sử dụng để liên kết các khái niệm trong các hệ thống phức tạp và giúp tạo ra một nền tảng tri thức chung, từ đó cải thiện khả năng phân tích và suy luận. Tuy nhiên, ontology có xu hướng phân cấp và cứng nhắc, dẫn đến khó

khăn khi phải xử lý các mối quan hệ phức tạp và không đồng nhất trong các bối cảnh đa dạng như hình ảnh thực tế.

Với mong muốn phân tích cảm xúc dựa trên ngữ cảnh một cách có hiệu quả, đồ thị tri thức là một phương pháp mới để khắc phục những hạn chế của các phương pháp truyền thống khi phân tích cảm xúc từ hình ảnh. Thay vì chỉ dựa vào biểu cảm khuôn mặt hoặc tư thế cơ thể, đồ thị tri thức cho phép mô hình hóa các mối quan hệ phức tạp giữa các đối tượng và yếu tố ngữ cảnh trong hình ảnh. Các đối tượng này, bao gồm cả nhân vật chính và các đối tượng xung quanh, có thể ảnh hưởng trực tiếp đến trạng thái cảm xúc tổng thể của nhân vật.

Việc sử dụng đồ thị tri thức giúp mã hóa không chỉ các đặc điểm của từng đối tượng mà còn các mối liên hệ cảm xúc giữa chúng, từ đó tạo ra một cấu trúc liên kết giúp mô hình học tập và suy luận về cảm xúc trong các ngữ cảnh phức tạp. Như vậy, việc kết hợp đồ thị tri thức để biểu diễn các yếu tố cảm xúc, ngữ cảnh và mối quan hệ giữa chúng là cần thiết để cải thiện khả năng phân tích cảm xúc tự động trong các hệ thống tương tác.

Từ những cơ sở trên, đề tài "**Xây dựng đồ thị tri thức cho phân tích cảm xúc từ dữ liệu hình ảnh"** được thực hiện dựa trên đồ thị tri thức được xây dựng để nhận dạng cảm xúc của người nhờ vào ứng dụng của thị giác máy tính và xử lý ngôn ngữ tự nhiên. Đồ thị tri thức không chỉ giúp biểu diễn mối quan hệ giữa các yếu tố trong một bức ảnh mà còn hỗ trợ quá trình phân tích cảm xúc một cách hiệu quả và chính xác hơn, mà còn mở ra nhiều cơ hội ứng dụng trong các lĩnh vực khác trong tương lai.

1.2. Mục tiêu

Trong đề tài này, mục tiêu là xây dựng đồ thị tri thức biểu diễn các mối quan hệ giữa các yếu tố cảm xúc từ dữ liệu hình ảnh. Cụ thể, nghiên cứu tập trung vào các đích sau đây:

Xây dựng đồ thị tri thức để phân tích cảm xúc từ hình ảnh nhằm tăng cường khả năng phân tích cảm xúc dựa trên ngữ cảnh.

Thực nghiệm, đánh giá và so sánh kết quả với các phương pháp trước đây.

1.3. Đóng góp của nghiên cứu

Đóng góp của nghiên cứu là xây dựng đồ thị tri thức cho phân tích cảm xúc từ dữ liệu hình ảnh.

1.4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu: Xây dựng đồ thị tri thức cho dữ liệu hình ảnh có chứa cảm xúc và các ngữ cảnh trong tự nhiên, nhằm kết hợp thông tin từ con người và bối cảnh xung quanh để cải thiện khả năng phân tích cảm xúc.

Phạm vi nghiên cứu: Xây dựng đồ thị tri thức dựa trên hình ảnh trong bộ dữ liệu EMOTIC, tập trung vào các yếu tố thị giác và cảm xúc như biểu cảm khuôn mặt, ngữ cảnh xung quanh, và các mối quan hệ giữa cảm xúc và các đối tượng trong hình ảnh.

1.5. Phương pháp nghiên cứu:

Phương pháp nghiên cứu lý thuyết:

- Tìm hiểu tổng quan các công trình nghiên cứu về phân tích cảm xúc từ hình ảnh (khuôn mặt, ngữ cảnh, cử chỉ).
- Nghiên cứu cơ sở lý thuyết liên quan đến đề tài.
- Nghiên cứu kỹ thuật trích xuất đặc trưng trên dữ liệu hình ảnh sử dụng mô hình học sâu.
- Nghiên cứu cách xây dựng đồ thị tri thức.
- Đề xuất hướng phát triển trong tương lai.

Phương pháp nghiên cứu thực nghiệm:

- Thu thập dữ liệu đáp ứng yêu cầu bài toán.
- Tiến hành xây dựng đồ thị tri thức.
- Đánh giá và so sánh kết quả đạt được.

1.6. Ý nghĩa khoa học và thực tiễn:

Ý nghĩa khoa học:

Đề tài cung cấp một khung lý thuyết mới về việc sử dụng đồ thị tri thức trong phân tích cảm xúc từ hình ảnh, đồng thời đóng góp vào sự phát triển của các phương pháp phân tích cảm xúc đa phương thức.

Ý nghĩa thực tiễn:

Ứng dụng của đồ thị tri thức có thể giúp cải thiện khả năng phân tích và dự đoán cảm xúc từ dữ liệu hình ảnh. Điều này góp phần vào việc phát triển các hệ thống tương tác người-máy thông minh hơn, hỗ trợ ra quyết định trong các tình huống thực tế ở nhiều lĩnh vực khác như giáo dục, y tế, marketing, …

1.7. Nội dung khóa luận tốt nghiệp

Khóa luận bao gồm 6 chương:

Chương Mở đầu

Chương này giới thiệu tổng quan về đề tài gồm các nội dung như: lý do chọn đề tài, mục đích, đối tượng và phạm vi nghiên cứu của đề tài.

Chương 1. Tổng quan

Chương này giới thiệu tổng quan về tình hình nghiên cứu trong nước và ngoài nước, các thành tựu và thách thức về xây dựng đồ thị tri thức cho phân tích cảm xúc. Giới thiệu về các tập dữ liệu chuẩn được dùng để nghiên cứu về đồ thị tri thức và phân tích cảm xúc.

Chương 2. Cơ sở lý thuyết

Chương này giới thiệu lý thuyết về đồ thị tri thức, sơ lược về bộ dữ liệu, mô hình sinh caption và mô hình phân tích cảm xúc sử dụng trong bài. Những kiến thức này là tiền đề để áp dụng vào việc xây dựng đồ thị tri thức cho phân tích cảm xúc từ dữ liệu hình ảnh.

Chương 3. Xây dựng đồ thị tri thức cho phân tích cảm xúc từ dữ liệu hình ảnh

Chương này trình bày chi tiết phương pháp mà bài nghiên cứu sử dụng để phân tích cảm xúc của con người nhờ vào đồ thị tri thức. tình trạng của bộ dữ liệu, quá trình xử lý dữ liệu cũng như phương pháp xây dựng nên đồ thị tri thức.

Chương 4. Thực nghiệm và đánh giá

Chương này trình bày quá trình thực nghiệm và phân tích về những ưu điểm, nhược điểm, so sánh và đánh giá kết quả đạt được khi thực hiện chương trình.

Chương 5. Kết luận và hướng phát triển

Chương này tổng kết lại những gì đã đạt được và chưa đạt được sau khi nghiên cứu và thực nghiệm. Sau đó nêu lên những hướng nghiên cứu và phát triển tiếp theo trong tương lai.

CHƯƠNG 1. TỔNG QUAN TÌNH HÌNH NGHIÊN CỨU

1.1. Tổng quan

Trong những năm gần đây, nhận dạng cảm xúc bằng kỹ thuật học máy đã và đang thu hút được sự chú ý đáng kể trong nhiều lĩnh vực khác trong đời sống. Cảm xúc là một yếu tố trung tâm trong đời sống tinh thần của con người, ảnh hưởng sâu sắc đến hành vi, tư duy và khả năng tương tác xã hội. Việc phân tích cảm xúc, tức là khả năng tự động xác định và phân tích trạng thái cảm xúc thông qua dữ liệu như hình ảnh, video, âm thanh hay văn bản đã trở thành một chủ đề nghiên cứu quan trọng trong lĩnh vực trí tuệ nhân tạo.

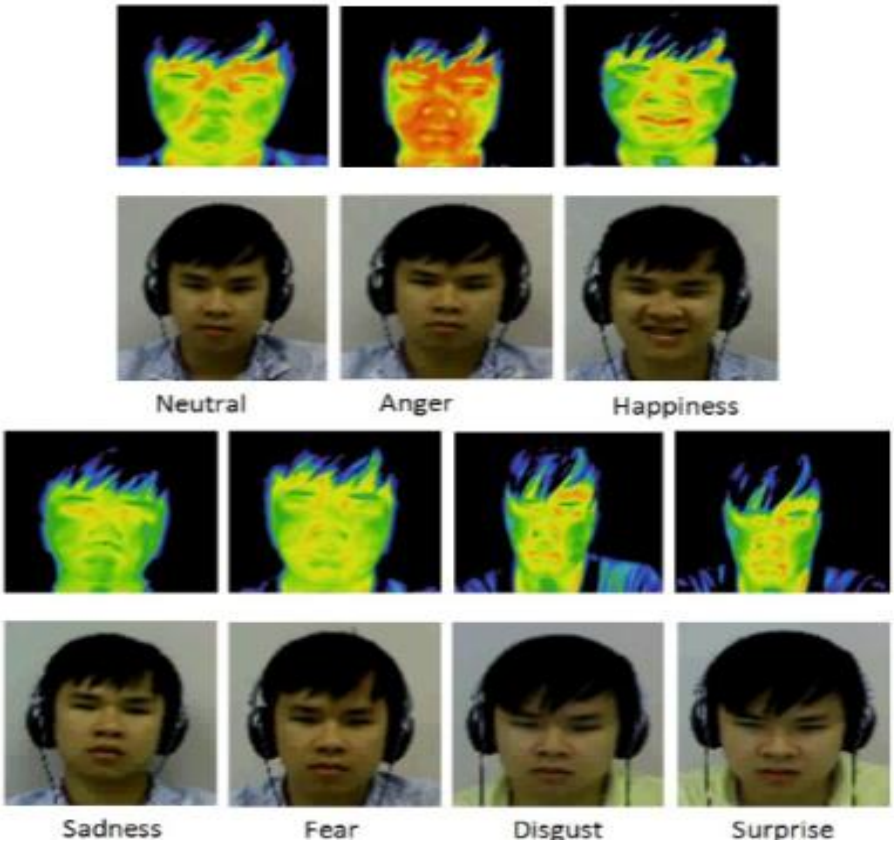
Trên thế giới, nhận diện cảm xúc đang được ứng dụng mạnh mẽ trong nhiều lĩnh vực thiết yếu như giáo dục, y tế, chăm sóc sức khỏe tinh thần và dịch vụ khách hàng. Trong giáo dục, các hệ thống học tập thông minh có khả năng theo dõi biểu cảm khuôn mặt, mức độ tập trung và cảm xúc của học sinh để điều chỉnh cách giảng dạy cho phù hợp, từ đó nâng cao hiệu quả học tập. D'Mello và các cộng sự [7] đã chứng minh rằng việc phát hiện kịp thời các trạng thái cảm xúc như "buồn chán" hay "bối rối" có thể giúp hệ thống học tập đưa ra phản hồi cá nhân hóa, làm tăng sự tương tác giữa người học và nội dung bài giảng.

Trong lĩnh vực y tế, đặc biệt là chăm sóc sức khỏe tâm thần, công nghệ nhận diện cảm xúc hỗ trợ phát hiện sớm các dấu hiệu trầm cảm, lo âu hay rối loạn cảm xúc thông qua biểu hiện khuôn mặt hoặc hành vi phi ngôn ngữ. Các hệ thống này đang được triển khai tại nhiều quốc gia phát triển để hỗ trợ chẩn đoán từ xa hoặc theo dõi tiến trình điều trị. Ngoài ra, trong các ứng dụng chăm sóc người cao tuổi, cảm biến cảm xúc giúp cảnh báo sớm các tình trạng bất ổn về tinh thần mà người bệnh khó tự nhận biết hoặc diễn đạt bằng lời nói. Đặc biệt, sau đại dịch COVID-19, nhu cầu theo dõi sức khỏe tâm thần trong cộng đồng, học sinh, sinh viên và nhân viên văn phòng tăng cao, kéo theo sự quan tâm đến các giải pháp AI có thể "hiểu" được cảm xúc người dùng. Vậy nên có nhiều trường học đã nghiên cứu các ứng dụng hệ thống nhận diện cảm xúc trong lớp học trực tuyến, các video học trục tiếp nhằm theo dõi mức độ chú ý, sự căng thẳng và tương tác của sinh viên.

Theo nhiều nhà tâm lý học, cảm xúc không chỉ ảnh hưởng đến hành vi và nhận thức mà còn đóng vai trò điều hướng các quyết định xã hội, phản ứng thích nghi và học tập. Các lý thuyết kinh điển như thuyết cảm xúc cơ bản của Ekman [8] cho rằng có một số lượng nhỏ cảm xúc phổ quát mà mọi người đều thể hiện và nhận biết được, bất kể văn hóa, bao gồm sáu cảm xúc cơ bản: vui, buồn, tức giận, ghê tởm, sợ hãi và ngạc nhiên. Những lý thuyết này tạo nền tảng cho các nghiên cứu về nhận diện cảm xúc tự động trong khoa học máy tính và trí tuệ nhân tạo.

Các nghiên cứu về phân tích cảm xúc đã được cộng đồng nghiên cứu rộng rãi và phát triển mạnh mẽ, bao gồm các phương pháp truyền thống như phân tích biểu cảm khuôn mặt, và nhận diện cảm xúc từ cử chỉ cơ thể. Hầu hết các công trình hiện có tập trung vào việc phân tích biểu cảm khuôn mặt để dự đoán cảm xúc [9], [10]. Cơ sở của các phương pháp này là Hệ thống mã hóa hành động khuôn mặt [8], mã hóa biểu cảm khuôn mặt bằng một tập hợp các chuyển động cục bộ cụ thể của khuôn mặt, được gọi là Đơn vị hành động. Các phương pháp tiếp cận dựa trên khuôn mặt [8], [11] này thường sử dụng các đặc điểm của hình học khuôn mặt hoặc các đặc điểm ngoại hình để mô tả khuôn mặt. Sau đó, những đặc điểm này được trích xuất đặc trưng và dùng nó để nhận dạng Đơn vị hành động và các cảm xúc cơ bản theo các cảm xúc.

Tuy nhiên, khi phân tích cảm xúc từ hình ảnh có nhiều vấn đề xảy ra như cường độ ánh sáng hay chất lượng ảnh không tốt làm cho việc phân tích không được tốt. Để đảm bảo cho việc không phụ thuộc chất lượng ảnh hay vấn đề môi trường, H. Nguyen và cộng sự [12] đã xây dựng bộ dữ liệu KTFE (Kotani Thermal Facial Emotion), là sự kết hợp ảnh thường và ảnh nhiệt về cảm xúc con người nhằm tăng cường hiệu suất nhận diện cảm xúc của con người.



Hình 1.1 Ảnh thường và nhiệt của bảy loại cảm xúc cơ bản [12]

Bên cạnh ảnh tĩnh, các hướng nghiên cứu khác cũng khai thác các loại dữ liệu đa phương thức như văn bản, âm thanh, video và gần đây là ngữ cảnh tổng thể để phân tích cảm xúc [13], [14], [15]. Trong phân tích văn bản, các mô hình như BERT, RoBERTa [16] hoặc SenticNet đã được áp dụng để trích xuất cảm xúc ẩn trong phát ngôn hoặc bài viết [17]. Trong khi đó, phân tích giọng nói hoặc âm thanh dựa trên các đặc trưng âm học như cao độ, năng lượng, và phổ âm thanh cũng cho thấy hiệu quả trong nhận diện cảm xúc, đặc biệt trong các tình huống giao tiếp thoại [18], [19]. Trong bối cảnh đa phương tiện, video là dạng dữ liệu quan trọng giúp nắm bắt đầy đủ hơn quá trình thay đổi cảm xúc theo thời gian. Các nghiên cứu như [20], [21] đã phát triển các mô hình học sâu kết hợp CNN và LSTM để khai thác mối liên hệ giữa các khung hình và xu hướng cảm xúc theo chuỗi thời gian.

Việc nhận dạng cảm xúc dựa trên bối cảnh nhận được nhiều sự chú ý hơn ở khía cạnh tính toán cảm xúc vì nhiều nghiên cứu [22], [23] trong lĩnh vực tâm lý học đã chứng minh lợi thế của bối cảnh đối với việc dự đoán cảm xúc. Bởi vì, loại thông tin này có thể giải quyết các vấn đề về khuôn mặt hoặc cơ thể bị che khuất. Nghiên cứu của Lee và cộng sự đã đề xuất CAER-Net [24] với hai mạng con để chứng minh tính hiệu quả của bối cảnh. bằng cách chặn khuôn mặt của những người trong ảnh tại một luồng để mạng có thể tìm kiếm các đặc điểm khác nhau nhưng vẫn liên quan đến trạng thái cảm xúc.

Phân tích cảm xúc đa phương thức đã trở thành một hướng nghiên cứu quan trọng trong lĩnh vực trí tuệ nhân tạo, nhằm mục đích khai thác thông tin từ nhiều loại dữ liệu khác nhau như hình ảnh, văn bản và âm thanh để nhận diện chính xác trạng thái cảm xúc của con người. Chìa khóa để xử lý biểu hiện tình cảm chi tiết trong dữ liệu đa phương thức là tìm thông tin quan trọng liên quan đến khía cạnh từ các phương thức khác nhau. Ngoài ra, mối tương quan giữa các phương thức khác nhau có thể được sử dụng để nhận biết thêm thông tin cảm tính. Như trong Hình 5 [25], nội dung đánh giá "Ở đây núi đẹp, nước trong, nhưng ngôi nhà gỗ lại tồi tàn và đáng lẽ phải bị bỏ hoang từ lâu". cùng với hình ảnh đi kèm chứa đựng ba khía cạnh: "núi", "nước" và "nhà gỗ". Nội dung liên quan đến từng khía cạnh có thể được tìm thấy ở vùng tương ứng của hình ảnh. Ví dụ, hình ảnh cho thấy núi đẹp, nước trong nhưng ngôi nhà gỗ lại rất đổ nát. Xem xét các vùng hình ảnh và nội dung văn bản liên quan đến khía cạnh này, chúng ta có thể suy ra rằng sự phân cực tình cảm của các khía cạnh "núi" và "nước" là tích cực, trong khi sự phân cực tình cảm của khía cạnh "ngôi nhà gỗ" là tiêu cực.

Hình 1.2. Hình có chứa các yếu tố "nhà", "nước" và "núi"

Do đó, nhiều mô hình của Yang và cộng sự [26], Zhou và cộng sự [4] đã áp dụng các cơ chế chú ý để khám phá mối tương quan giữa phương thức văn bản và phương thức hình ảnh. Tuy nhiên, vẫn còn nhiều hạn chế hạn chế về dữ liệu tình cảm ngầm. Lian và cộng sự [27] chỉ ra rằng các mẫu dữ liệu trong bộ dữ liệu hiện có chứa thông tin cảm tính ngầm. Nếu không giải quyết được thông tin tình cảm ngầm, các mô hình có thể gặp khó khăn trong việc phân biệt chính xác sự phân cực tình cảm của các bài đánh giá thể hiện sự mỉa mai hoặc ẩn ý. Các mô hình hiện tại có thể không xác định được chính xác cảm xúc trong các bài đánh giá có liên quan đến các yếu tố như sự mỉa mai hoặc cách diễn đạt ngầm. Vậy nên,

sự khó khăn trong việc xử lý mối quan hệ phức tạp giữa các đối tượng dẫn đến hạn chế trong khả năng suy luận toàn diện về cảm xúc.

Để khắc phục vấn đề trên, nhiều nhà nghiên cứu đã nghiên cứu và sử dụng đồ thị tri thức (Knowledge Graph) để tăng cường khả năng diễn giải ngữ cảnh trong phân tích cảm xúc, cải thiện khả năng suy luận về đối tượng chủ thể và ngữ cảnh xung quanh đối tượng. Bởi vì đồ thị tri thức cung cấp khả năng biểu diễn dữ liệu có cấu trúc, tổ chức thông tin dưới dạng các thực thể và mối quan hệ giữa chúng, từ đó cho phép hệ thống phân tích không chỉ các đặc điểm riêng lẻ mà còn hiểu rõ mối tương quan giữa chúng. Đồ thị tri thức giúp kết nối các biểu cảm khuôn mặt, cử chỉ cơ thể và ngữ cảnh xung quanh, tạo ra một mạng lưới thông tin phức tạp nhưng có hệ thống, giúp cải thiện khả năng suy luận cảm xúc trong ngữ cảnh chính xác hơn, diễn giải tình cảm ngầm một cách rõ ràng hơn. Trước đó, đã có nhiều công trình nghiên cứu về đồ thị tri thức. Ví dụ, công ty Google đã giới thiệu cơ sở tri thức bằng đồ thị tri thức có tên là Google Knowledge Graph. Tiếp đó, nhiều đồ thị tri thức được ra đời như WordNet [28], YAGO [29], ... Sau đó đồ thị tri thức được đưa vào trong phân tích cảm xúc trong văn bản, giúp mô hình hóa tri thức về cảm xúc và suy luận một cách hiệu quả. Trong công trình nghiên cứu của Wu và cộng sự [30] đã giới thiệu nhiều nguồn kiến thức, tích hợp kiến thức cấu trúc cú pháp và kiến thức tình cảm bên ngoài vào một mô hình phân tích tình cảm thống nhất, giúp tăng cường các tính năng ngữ nghĩa. Tiếp đó, Zhou và cộng sự [4] đã đề xuất mô hình Cây cú pháp (Parse Tree) và Mạng tích chập đồ thị dựa trên tri thức (SK-GCN), cùng mô hình hóa cây phụ thuộc cú pháp và đồ thị tri thức, kết hợp hiệu quả kiến thức cú pháp và kiến thức bên ngoài. Từ đây có thể thấy, đồ thị tri thức cung cấp cấu trúc tổ chức các khái niệm liên quan đến cảm xúc, cho phép mô hình hiểu sâu hơn về ngữ nghĩa và ngữ cảnh của văn bản, từ đó cải thiện độ chính xác trong nhận diện cảm xúc. Tuy nhiên, việc phân tích cảm xúc chỉ từ văn bản thường gặp hạn chế do thiếu thông tin đa dạng từ các yếu tố phi ngôn ngữ.

Các nghiên cứu gần đây đã kết hợp phân tích cảm xúc từ cả văn bản và hình ảnh, nhằm tăng cường khả năng nhận diện cảm xúc một cách toàn diện hơn. Phân tích hình ảnh không

chỉ cung cấp thêm thông tin về ngôn ngữ cơ thể, biểu cảm khuôn mặt mà còn bổ sung ngữ cảnh quan trọng giúp hệ thống hiểu rõ hơn về tình huống mà cảm xúc diễn ra. Zhang và cộng sự [31] đã sử dụng Graph Convolution Network (GCN) để tìm hiểu mối quan hệ tình cảm giữa tất cả các vùng được đề xuất từ cảnh được trích xuất bởi Graph Convolution Network (RPN). Tuy nhiên, mọi mối quan hệ được đề xuất bằng cách này hay cách khác vẫn chứa đựng một số kết nối dư thừa, dẫn đến những ảnh hưởng xấu đến mạng. Mittal và cộng sự [14] đã trình bày EmotiCon dựa trên Nguyên tắc bối cảnh của Frege từ tâm lý học. Phương pháp của họ bao gồm ba luồng: đa phương thức về khuôn mặt và dáng đi, bối cảnh nền và tương tác giữa các tác nhân năng động xã hội. Từ kết quả thử nghiệm, cho thấy sự tương tác giữa các tác nhân xã hội năng động được trích xuất bằng hình ảnh có chiều sâu, tăng khả năng phân tích cảm xúc. Từ khó khăn [31] và lấy ý tưởng từ [14], Manh-Hung Hoang và cộng sự [32] đã nghiên cứu về chủ đề này và sử dụng cả thông tin bản địa hóa và ngữ nghĩa để giảm được việc nhiều kết nối dư thừa cũng như đánh giá tác động của từng đối tượng lên tác nhân chính một cách tốt hơn. Trong nghiên cứu của Kosti và cộng sự [33]còn kết hợp thêm hành động cử chỉ của cơ thể để phân tích cảm xúc.

Hình 1.3. a) Sử dụng hình ảnh tổng thể để trích xuất thông tin chứa quá nhiều yếu tố không liên quan đến dự đoán cảm xúc. b) Sử dụng các mối quan hệ dày đặc từ các vùng quan tâm, điều này trở thành gánh nặng cho mô hình xử lý các kết nối dự phòng. c) Thiết lập cơ chế quan hệ tập trung với tác nhân chính làm lõi và các đối tượng riêng biệt xung quanh là vệ tinh [32].

Tiếp đó, nghiên cứu của Juan và cộng sự [4] đề xuất một phương pháp nhận diện cảm xúc bằng cách kết hợp nhiều yếu tố đa phương thức từ hình ảnh, bao gồm biểu cảm khuôn mặt, tư thế cơ thể, và ngữ cảnh xung quanh. Sử dụng EmbraceNet+, một mô hình học sâu đa phương thức, phương pháp này tích hợp các loại dữ liệu khác nhau để cải thiện khả năng phân tích cảm xúc. Đáng chú ý, bài báo áp dụng một ontology cảm xúc (EMONTO) [34], nhằm lưu trữ và tổ chức thông tin cảm xúc dựa trên các thực thể và mối quan hệ giữa chúng. Mô hình chỉ hiệu quả đối với những hình ảnh và ngữ nghĩa rõ ràng, còn với anh trừu tường thì vẫn chưa chính xác. Từ những điều trên cho thấy việc sử dụng đồ thị tri thức để phân tích cảm xúc có hiệu quả tuy nhiên vẫn có một số điểm của thực sự được giải quyết như phân tích cảm xúc trên tấm ảnh trừu tượng. Bài nghiên cứu gần đây nhất, Costa và công sự đã đưa ra hướng mới cho đồ thị tri thức và phân tích cảm xúc, điểm nổi bật trong nghiên cứu này là việc sử dụng caption từ hình ảnh để mô tả ngữ cảnh, từ đó hỗ trợ quá trình phân tích cảm xúc chính xác hơn. Bằng cách áp dụng các mạng nơron tích chập đồ thị (Graph Convolutional Networks), bài báo đề xuất mô hình có khả năng kết nối các thực thể trong hình ảnh và các chú thích văn bản, từ đó tạo ra một biểu diễn ngữ cảnh giàu thông tin. Phương pháp này giúp tăng cường khả năng nhận diện cảm xúc bằng cách không chỉ dựa vào các biểu hiện bề ngoài mà còn khai thác các mối quan hệ ngữ nghĩa ẩn giữa các yếu tố trong hình ảnh và chú thích đi kèm. Kết quả cho thấy, việc tích hợp thông tin ngữ cảnh từ caption giúp mô hình đạt được hiệu suất vượt trội so với các phương pháp chỉ sử dụng đặc trưng thị giác thuần túy. Điều này mở ra tiềm năng cho việc ứng dụng đồ thị tri thức trong việc kết nối dữ liệu từ các chú thích và hình ảnh, nhằm xây dựng các mô hình phân tích cảm xúc toàn diện và chính xác hơn.

1.2. Các tập dữ liệu về cảm xúc

Việc xây dựng và huấn luyện các mô hình phân tích cảm xúc dựa trên đồ thị tri thức phụ thuộc rất lớn vào chất lượng và đặc điểm của tập dữ liệu. Trong bối cảnh này, các tập dữ liệu không chỉ cần cung cấp thông tin hình ảnh và cảm xúc, mà còn phải bao gồm các yếu tố ngữ cảnh, chú thích đa dạng và có thể mở rộng thành các thực thể và quan hệ trong

đồ thị tri thức. Nên việc nghiên cứu và phát triển các bộ dữ liệu dùng để xây dựng đồ thị tri thức là một điều cần thiết. Trong những năm gần đây, có vài bộ dữ liệu đã và đang được sử dụng nhằm phục vụ việc nghiên cứu cho xây dựng đồ thị thức và phân tích cảm xúc như EMOTIC [14], CAER [24], HECO [35].

EMOTIC là một trong những bộ dữ liệu đầu tiên kết hợp giữa cảm xúc và thông tin ngữ cảnh trong hình ảnh. Được giới thiệu bởi Kosti và cộng sự, bộ dữ liệu này bao gồm hơn 23.000 hình ảnh, với hơn 34.000 chú thích cảm xúc được gán cho các cá nhân trong bối cảnh thực tế. Điểm nổi bật của EMOTIC là khả năng biểu diễn cảm xúc ở hai cấp độ: cảm xúc rời rạc với 26 lớp (ví dụ như angry, happy, sad, disgusted, ...), và thang đo liên tục Valence-Arousal-Dominance (VAD). Mỗi người trong ảnh được chú thích cảm xúc cùng với hộp giới hạn (bounding box), cho phép mô hình học được mối quan hệ giữa biểu cảm, tư thế cơ thể và môi trường xung quanh. Chính cấu trúc giàu ngữ nghĩa này khiến EMOTIC trở thành một lựa chọn lý tưởng để xây dựng đồ thị tri thức cảm xúc, nơi mỗi thực thể như "người", "địa điểm", "hành động" hay "vật thể" đều có thể trở thành nút trong đồ thị, kết nối với nhau thông qua các mối quan hệ ngữ cảnh.

Trong khi đó, CAER là một bộ dữ liệu được phát triển với mục tiêu mô hình hóa cảm xúc trong ngữ cảnh xã hội động, đặc biệt là từ các đoạn phim truyền hình. Lee và cộng sự đã xây dựng CAER từ các video thực tế trong chương trình truyền hình, với tổng cộng hơn 13.000 clip. Mỗi clip được xử lý thành khung hình và gán nhãn cảm xúc cho từng người, đồng thời giữ lại các yếu tố cảnh trí và tương tác trong khung hình. CAER bao gồm hai phiên bản chính: CAER-S (static) cho ảnh tĩnh và CAER-B (balanced) có phân phối cảm xúc được cân bằng. Không giống như các bộ dữ liệu chỉ tập trung vào khuôn mặt, CAER tận dụng thông tin ngữ cảnh từ toàn bộ khung cảnh – ví dụ như người đang nói chuyện, đối tượng đang nhìn vào ai, có bao nhiêu người trong ảnh – từ đó mở rộng khả năng trích xuất quan hệ và xây dựng mạng lưới tri thức cảm xúc.

Gần đây hơn, bộ dữ liệu HECO (Emotion Recognition for Multiple Context Awareness) được đề xuất bởi Yang và cộng sự đã tiếp tục nâng cao tính ngữ cảnh trong phân tích cảm xúc. Bộ dữ liệu được thiết kế để phản ánh sự phức tạp của các tình huống thực tế, nơi cảm xúc không chỉ được biểu hiện qua khuôn mặt, mà còn thông qua hành vi, mối quan hệ xã hội, và các yếu tố môi trường. Bộ dữ liệu này bao gồm hơn 20.000 hình ảnh được gán nhãn thủ công, mỗi ảnh đều có thông tin chi tiết về biểu cảm khuôn mặt, tư thế cơ thể, tương tác giữa các cá nhân, và ngữ cảnh xung quanh. HECO nổi bật nhờ tính sẵn sàng cho việc xây dựng đồ thị tri thức: các chú thích được tổ chức sao cho có thể trực tiếp ánh xạ thành các thực thể (như người, vật thể, địa điểm) và các mối quan hệ (như "ngồi cạnh", "ôm", "nhìn vào"). Điều này giúp tạo ra mạng tri thức cảm xúc có khả năng suy luận sâu hơn về cảm xúc thực tế trong các tình huống đa chiều.

Hình 1.4. Hình ảnh về các bộ dữ liệu ngữ cảnh a. EMOTIC, b. CAER, c. HECO

Bảng 1.1. Thống kê về các bộ dữ liệu

Tên dữ liệu	**Kích thước**	**Nhãn cảm xúc**	**Loại dữ liệu**	**Đối tượng được chú thích**
EMOTIC [14]	23,571 ảnh	26 loại	ảnh	34,320
CAER [24]	13,201 clips	7 loại	clips	13,201
HECO [35]	9,385 ảnh	8 loại	ảnh	19,781

Cả ba bộ dữ liệu EMOTIC, CAER và HECO đều đóng vai trò quan trọng trong việc phát triển các mô hình phân tích cảm xúc hiện đại, đặc biệt khi kết hợp với đồ thị tri thức. Nếu như EMOTIC cung cấp nền tảng ngữ cảnh phong phú trong hình ảnh tĩnh, thì CAER khai thác tốt các khung cảnh xã hội từ video, còn HECO lại nổi bật với khả năng biểu diễn cảm xúc trong các tình huống đa tác nhân và đa yếu tố. Các bộ dữ liệu này không chỉ cung cấp dữ liệu học chất lượng cao, mà còn mở ra khả năng khai thác các tri thức ngữ nghĩa từ ngữ cảnh – yếu tố cốt lõi giúp cải thiện hiệu quả của hệ thống phân tích cảm xúc thông minh.

1.3. Thách thức trong việc sử dụng đồ thị tri thức cho phân tích cảm xúc từ dữ liệu hình ảnh

Việc áp dụng đồ thị tri thức vào phân tích cảm xúc từ dữ liệu hình ảnh gặp phải nhiều thách thức liên quan đến xây dựng dữ liệu, mô hình hóa quan hệ, tích hợp với mô hình học sâu, tối ưu hóa tính toán và triển khai thực tế. Những vấn đề này đặt ra yêu cầu nghiên cứu chuyên sâu nhằm đảm bảo rằng hệ thống có thể hoạt động hiệu quả, chính xác và có khả năng tổng quát hóa trên nhiều loại dữ liệu khác nhau.

Việc ứng dụng đồ thị tri thức vào phân tích cảm xúc từ dữ liệu hình ảnh mở ra hướng tiếp cận đầy hứa hẹn, khi cho phép hệ thống hiểu sâu hơn về các mối quan hệ ngữ nghĩa giữa các thực thể trong ảnh như con người, đối tượng, hành động, và bối cảnh. Tuy nhiên, quá trình xây dựng và tích hợp đồ thị tri thức vào bài toán nhận diện cảm xúc vẫn đang đối mặt với nhiều thách thức kỹ thuật và lý thuyết.

Thứ nhất, việc trích xuất tri thức từ dữ liệu hình ảnh không hề đơn giản. Không giống như văn bản, nơi các thực thể và quan hệ có thể được xác định rõ ràng bằng các công cụ xử lý ngôn ngữ tự nhiên, hình ảnh đòi hỏi các hệ thống thị giác máy tính phải phát hiện và phân loại chính xác các đối tượng, sau đó suy luận các mối quan hệ giữa chúng. Các mối quan hệ này thường là ngữ cảnh phức tạp, không biểu hiện rõ ràng như "người-ngồi-trên-

ghế" hay "người-cầm-ô". Sai sót trong bước phát hiện thực thể hoặc quan hệ có thể làm suy giảm đáng kể hiệu quả của toàn bộ đồ thị tri thức.

Thứ hai, cảm xúc là một khái niệm trừu tượng, phụ thuộc mạnh vào ngữ cảnh và mang tính chủ quan. Một biểu cảm khuôn mặt hay tư thế cơ thể cụ thể có thể mang ý nghĩa cảm xúc khác nhau trong từng tình huống. Việc ánh xạ các đặc trưng cảm xúc sang các thực thể và quan hệ trong đồ thị tri thức đòi hỏi mô hình phải có khả năng suy luận ngữ nghĩa linh hoạt. Tuy nhiên, hiện nay hầu hết các mô hình đồ thị vẫn hoạt động dựa trên biểu diễn vector cố định, khó thích ứng với các thay đổi phức tạp của cảm xúc trong bối cảnh đa chiều.

Thứ ba, việc thiếu hụt dữ liệu chuẩn hóa là rào cản lớn. Đa số các bộ dữ liệu cảm xúc hiện nay (chẳng hạn như EMOTIC, CAER, …) chưa cung cấp sẵn đồ thị tri thức đi kèm. Việc tự động sinh đồ thị tri thức từ dữ liệu ảnh đòi hỏi xây dựng pipeline gồm nhiều bước phức tạp như phát hiện đối tượng, gán nhãn quan hệ, ánh xạ ngữ nghĩa và cập nhật tri thức – trong khi mỗi bước đều có thể tạo ra lỗi tích lũy. Thêm vào đó, không có một cấu trúc đồ thị tri thức tiêu chuẩn nào được thống nhất để mô tả cảm xúc từ hình ảnh, dẫn đến sự phân mảnh trong hướng tiếp cận giữa các nghiên cứu.

Thứ tư, chi phí tính toán và độ phức tạp của mô hình tăng đáng kể khi tích hợp đồ thị tri thức vào hệ thống học sâu. Các kỹ thuật như Graph Convolutional Networks hoặc Graph Attention Networks yêu cầu tài nguyên tính toán lớn, đặc biệt là khi làm việc với đồ thị có nhiều nút và cạnh. Việc tối ưu hóa các tham số mạng đồ thị đồng thời với mạng CNN cho hình ảnh đòi hỏi thiết kế mô hình cẩn trọng và dữ liệu huấn luyện phong phú, điều mà nhiều nghiên cứu vẫn đang gặp khó khăn.

Cuối cùng, khả năng suy luận và diễn giải của mô hình đồ thị tri thức còn hạn chế. Mặc dù đồ thị tri thức được kỳ vọng mang lại khả năng lý giải tốt hơn cho hệ thống AI, nhưng trong thực tế, quá trình học biểu diễn tri thức vẫn còn mang tính "hộp đen" và thiếu minh

bạch. Việc hiểu rõ mô hình đã sử dụng thông tin gì, từ thực thể nào, quan hệ nào để đưa ra dự đoán cảm xúc vẫn là một bài toán mở.

Như vậy, mặc dù đồ thị tri thức là một công cụ tiềm năng để nâng cao hiệu quả và độ chính xác của hệ thống phân tích cảm xúc từ hình ảnh, nhưng việc tích hợp đồ thị tri thức vào bài toán này vẫn còn đối mặt với nhiều thách thức về mặt kỹ thuật, dữ liệu và lý thuyết. Những khó khăn này đặt ra yêu cầu cho các nghiên cứu tương lai cần phát triển các kỹ thuật trích xuất tri thức mạnh mẽ hơn, xây dựng kho dữ liệu tiêu chuẩn tích hợp tri thức, và tối ưu hóa kiến trúc mạng tích hợp đồ thị tri thức với các mô hình học sâu hiện đại.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1. Lý thuyết về đồ thị tri thức

Đồ thị tri thức (Knowledge Graph - KG) là một cấu trúc dữ liệu bán cấu trúc thể hiện thông tin thông qua các thực thể (entities) và quan hệ (relations) giữa chúng dưới dạng đồ thị. Mỗi nút (node) trong đồ thị đại diện cho một thực thể hoặc khái niệm, còn mỗi cạnh (edge) biểu diễn mối quan hệ giữa các thực thể đó. Đồ thị tri thức thường được biểu diễn dưới dạng bộ ba (triples) có cấu trúc (subject, predicate, object).

Khái niệm đồ thị tri thức được phổ biến rộng rãi sau khi Google công bố hệ thống Google Knowledge Graph vào năm 2012 nhằm nâng cao khả năng hiểu nội dung truy vấn của người dùng dựa trên thực thể và quan hệ. Kể từ đó, KG đã trở thành một trong những công nghệ trọng tâm trong lĩnh vực trí tuệ nhân tạo, đặc biệt trong các ứng dụng như tìm kiếm ngữ nghĩa, hỏi đáp tự động, đề xuất nội dung, và gần đây là phân tích cảm xúc dựa trên hình ảnh và ngữ cảnh.

Đặc điểm nổi bật của đồ thị tri thức bao gồm:

- Tính ngữ nghĩa: Cho phép máy tính hiểu được ý nghĩa sâu hơn của dữ liệu thông qua quan hệ giữa các thực thể.
- Khả năng mở rộng: Có thể dễ dàng cập nhật, bổ sung tri thức mới mà không ảnh hưởng đến cấu trúc tổng thể.
- Tính suy luận: Cho phép hệ thống đưa ra những kết luận mới dựa trên tập hợp tri thức hiện có thông qua các kỹ thuật suy diễn logic [36].

Một đồ thị tri thức điển hình bao gồm bốn thành phần chính:

- Tập thực thể: Đại diện cho các đối tượng trong thế giới thực như người, địa điểm, sự kiện, vật thể, ...

- Tập quan hệ: Là các liên kết ngữ nghĩa giữa các thực thể, ví dụ như "là bạn của", "sống tại", "thuộc loại".
- Thuộc tính: Mô tả các đặc điểm của thực thể, ví dụ như tên, tuổi, giới tính, ...
- Luật suy luận: Hệ thống quy tắc để rút ra các tri thức mới từ các tri thức hiện có.

2.2. Phân tích dữ liệu

Bộ dữ liệu EMOTIC là tập hợp các hình ảnh về con người trong môi trường không bị giới hạn, được chú thích theo trạng thái cảm xúc rõ ràng của họ. Bộ dữ liệu chứa 23.571 hình ảnh và 34.320 người được chú thích. Một số hình ảnh được công cụ tìm kiếm Google thu thập thủ công từ Internet. Bài nghiên cứu đã sử dụng kết hợp các truy vấn chứa nhiều địa điểm, môi trường xã hội, hoạt động khác nhau và nhiều từ khóa khác nhau về trạng thái cảm xúc. Các hình ảnh còn lại thuộc về 2 bộ dữ liệu benchmark công khai: MSCOCO [37]và Ade20k [38]. Bộ dữ liệu bao gồm 26 cảm xúc: Peace, Affection, Esteem, Anticipation, Engagement, Confidence, Happiness, Pleasure, Excitement, Surprise, Sympathy, Doubt/Confusion, Disconnection, Fatigue, Embarrassment, Yearning, Disapproval, Aversion, Annoyance, Anger, Sensitivity, Sadness, Disquietment, Fear, Pain, Suffering. Hình 2.1. đưa ra các hình ảnh mô tả cho từng loại cảm xúc.

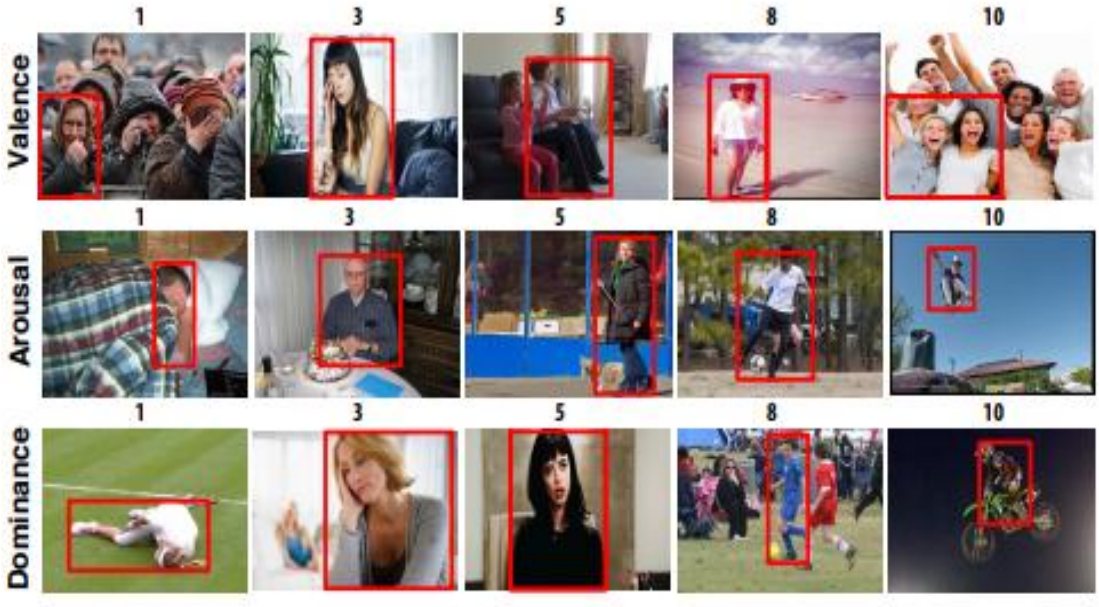
Hình 2.1. Một số hình ảnh trong EMOTIC cho từng loại trong số 26 loại cảm xúc [33]

Bảng 2.1. Ngữ nghĩa của 26 loại cảm xúc

Cảm Xúc	Miêu tả
1: Peace	Trạng thái tốt lành và thư giãn; không lo lắng; có suy nghĩ hoặc cảm giác tích cực; hài lòng
2: Affection	Cảm giác ấm áp; tình yêu; sự dịu dàng
3: Esteem	Cảm giác có ý kiến hoặc đánh giá tích cực; tôn trọng; ngưỡng mộ; biết ơn
4: Anticipation	Trạng thái chờ đợi; hy vọng hoặc chuẩn bị cho các sự kiện tương lai có thể xảy ra
5: Engagement	Chú ý vào điều gì đó; hấp dẫn vào điều gì đó; tò mò; quan tâm
6: Confidence	Cảm giác chắc chắn; niềm tin rằng kết quả sẽ tích cực; động viên; tự hào
7: Happiness	Cảm giác hài lòng; cảm thấy vui vẻ hoặc thích thú
8: Pleasure	Cảm giác hạnh phúc trong các giác quan
9: Excitement	Cảm giác phấn khích; sự kích thích; năng động
10: Surprise	Phát hiện đột ngột của một điều gì đó không mong đợi
11: Sympathy	Trạng thái chia sẻ cảm xúc, mục tiêu hoặc vấn đề của người khác; ủng hộ; thông cảm
12: Doubt/Confusion	Khó khăn trong việc hiểu hoặc quyết định; suy nghĩ về các lựa chọn khác nhau
13: Disconnection	Cảm giác không quan tâm đến sự kiện chính xung quanh; lạnh lùng; buồn chán; mất tập trung
14: Fatigue	Mệt mỏi; mệt mỏi; buồn ngủ
15: Embarrassment	Cảm giác xấu hổ hoặc tội lỗi

Cảm Xúc	Miêu tả
16: Yearning	Mong muốn mạnh mẽ có được điều gì đó; ghen tỵ; ghen tị; ham muốn
17: Disapproval	Cảm giác rằng có điều gì đó sai trái hoặc đáng trách; khinh bỉ; thù địch
18: Aversion	Cảm giác kinh tởm, không thích, ghê sợ; cảm giác ghét bỏ
19: Annoyance	Phiền phức bởi điều gì đó hoặc ai đó; bực tức; không kiên nhẫn; thất vọng
20: Anger	Cảm giác không hài lòng mạnh mẽ hoặc giận dữ; tức giận; phẫn nộ
21: Sensitivity	Cảm giác bị thương tổn về mặt thể chất hoặc tinh thần; cảm giác mong manh hoặc yếu đuối
22: Sadness	Cảm giác không hạnh phúc, buồn bã, thất vọng hoặc tuyệt vọng
23: Disquietment	Lo lắng; lo lắng; bực bội; lo lắng; căng thẳng; áp lực; lo sợ
24: Fear	Cảm giác nghi ngờ hoặc sợ hãi trước nguy hiểm, đe dọa, điều ác hoặc đau đớn; kinh hoàng
25: Pain	Sự đau đớn về mặt thể chất
26: Suffering	Đau đớn tâm lý hoặc cảm xúc; đau khổ; đau đớn

Ngoài ra, trong bộ dữ liệu còn sử dụng mô hình trạng thái cảm xúc VAD (VAD Emotional State Model) bao gồm ba thang đo cảm xúc liên tục: Valence (mức độ tích cực), Arousal (mức độ kích động), và Dominance (mức độ kiểm soát trong tình huống) để giúp mô hình phân tích và dự đoán chính xác trạng thái cảm xúc của con người trong các ngữ cảnh khác nhau đánh giá cảm xúc bằng thang đo từ 1 đến 10.



Hình 2.2. Thang đo đánh giá từ 1 đến 10 của 3 thông số VAD [33]

Ngoài ra bộ EMOTIC còn là dữ liệu đa nhãn. Trong một hình ảnh, một đối tượng có thể được gán đồng thời nhiều nhãn cảm xúc thuộc 26 loại cảm xúc khác nhau, chẳng hạn như "lo lắng", "bối rối", "buồn" và "hy vọng". Bên cạnh yếu tố cảm xúc, dữ liệu này còn tích hợp thông tin ngữ cảnh – một yếu tố đóng vai trò quan trọng trong việc giải mã cảm xúc. Và sự phức tạp cũng gia tăng khi trong một khung hình có thể xuất hiện nhiều người, mỗi người mang trạng thái cảm xúc khác nhau. Điều này đặt ra thách thức lớn trong việc xác định đúng đối tượng cần phân tích.

Để giải quyết vấn đề này, bộ dữ liệu EMOTIC đã bổ sung thông tin về bounding box cho từng cá nhân trong ảnh. Mỗi bounding box xác định vị trí và kích thước của người trong ảnh, giúp mô hình tập trung vào từng đối tượng cụ thể khi trích xuất đặc trưng cảm xúc.

Khi tải về, bộ dữ liệu EMOTIC cung cấp tổng cộng 18,316 hình ảnh với 34,320 đối tượng được gán nhãn, được lưu trữ trong tệp định dạng .mat. Mỗi ảnh được liên kết với một danh sách các thông số bao gồm tọa độ và kích thước của bounding box (tọa độ tâm, chiều dài và chiều rộng). Ngoài ra, bộ dữ liệu còn cung cấp thông tin về 26 nhãn cảm xúc rời rạc, Ba giá trị VAD liên tục (Valence, Arousal, Dominance), Giới tính, Độ tuổi.

Các thông tin này được chuyển đổi và lưu trữ dưới định dạng .csv để thuận tiện cho xử lý và phân tích, như minh họa trong Hình của luận văn. Từ các tọa độ bounding box, các hình ảnh cắt từ ảnh gốc có thể được trích xuất bằng cách tính toán các giá trị:

- x1, y1: tọa độ góc dưới bên trái của bounding box,
- x2, y2: tọa độ góc trên bên phải của bounding box.

Index	Folder	Filename	Image Size	BBox	Categorical_Labels	Continuous_Labels	Gender	Age
0	framesdb/images	frame_ghkq7yp0itqz0kn7.jpg	[640, 426]	[279, 18, 623, 425]	['Affection', 'Confidence', 'E	[7, 4, 6]	Female	Adult
1	mscoco/images	COCO_train2014_0000004491	[640, 480]	[379, 210, 520, 479]	['Disapproval', 'Disconnecti	[5, 4, 6]	Female	Teenager
2	mscoco/images	COCO_train2014_0000004491	[640, 480]	[463, 197, 604, 473]	['Anticipation', 'Disconnecti	[4, 5, 6]	Male	Adult
3	mscoco/images	COCO_train2014_000000081	[640, 427]	[409, 120, 600, 422]	['Confidence', 'Happiness',	[7, 5, 7]	Female	Adult
4	mscoco/images	COCO_train2014_000000081	[640, 427]	[247, 91, 365, 285]	['Anticipation', 'Esteem', 'Ex	[7, 4, 6]	Female	Adult
5	mscoco/images	COCO_train2014_000000537	[640, 480]	[229, 106, 392, 363]	['Anticipation', 'Confidence'	[5, 6, 7]	Male	Adult
6	mscoco/images	COCO_val2014_00000014857	[640, 640]	[54, 86, 190, 592]	['Happiness']	[9, 5, 7]	Male	Adult
7	mscoco/images	COCO_val2014_00000014857	[640, 640]	[163, 98, 311, 555]	['Affection', 'Esteem', 'Excit	[8, 8, 8]	Male	Adult
8	mscoco/images	COCO_train2014_000000049	[640, 427]	[183, 147, 324, 421]	['Anticipation', 'Esteem', 'Ha	[6, 7, 7]	Male	Kid
9	mscoco/images	COCO_train2014_000000353	[640, 429]	[234, 78, 348, 373]	['Anticipation', 'Confidence'	[6, 9, 7]	Female	Adult
10	mscoco/images	COCO_train2014_000000336	[640, 427]	[375, 19, 534, 398]	['Affection', 'Engagement', 'I	[6, 6, 5]	Male	Kid
11	mscoco/images	COCO_train2014_000000572	[320, 240]	[17, 35, 307, 236]	['Anticipation', 'Confidence'	[7, 5, 5]	Female	Adult
12	mscoco/images	COCO_train2014_000000418	[640, 427]	[117, 14, 529, 426]	['Pleasure', 'Sympathy']	[7, 4, 7]	Male	Adult
13	mscoco/images	COCO_train2014_000000418	[640, 427]	[98, 42, 330, 419]	['Affection', 'Happiness', 'Pl	[8, 5, 7]	Male	Adult
14	emodb_small/images	gya78toe8acx9f1oqb.jpg	[430, 400]	[71, 6, 341, 400]	['Anger', 'Annoyance', 'Avers	[4, 7, 8]	Male	Adult
15	mscoco/images	COCO_val2014_00000026222	[640, 425]	[442, 33, 548, 358]	['Annoyance', 'Anticipation',	[5, 8, 8]	Female	Teenager
16	mscoco/images	COCO_val2014_00000026222	[640, 425]	[63, 104, 255, 386]	['Anticipation', 'Confidence'	[5, 8, 8]	Female	Teenager

Hình 2.3. Hình ảnh mô tả tập dữ liệu sau khi chạy file .mat

Trong đó:

- Index: Số thứ tự ảnh trong dữ liệu
- Folder: Nơi lưu trữ ảnh
- Filename: Tên của ảnh
- Image Size: Kích thước của ảnh
- Bbox: Tọa độ của bounding box đối tượng
- Categorical_Labels: Nhãn cảm xúc của đối tượng (26 loại cảm xúc)
- Continuous_Labels: Nhãn độ đo VAD
- Gender: Giới tính (nam, nữ)
- Age: Độ tuổi gồm Kid (0-12 tuổi), Teenager (13-20 tuổi), Adult (hơn 20 tuổi)

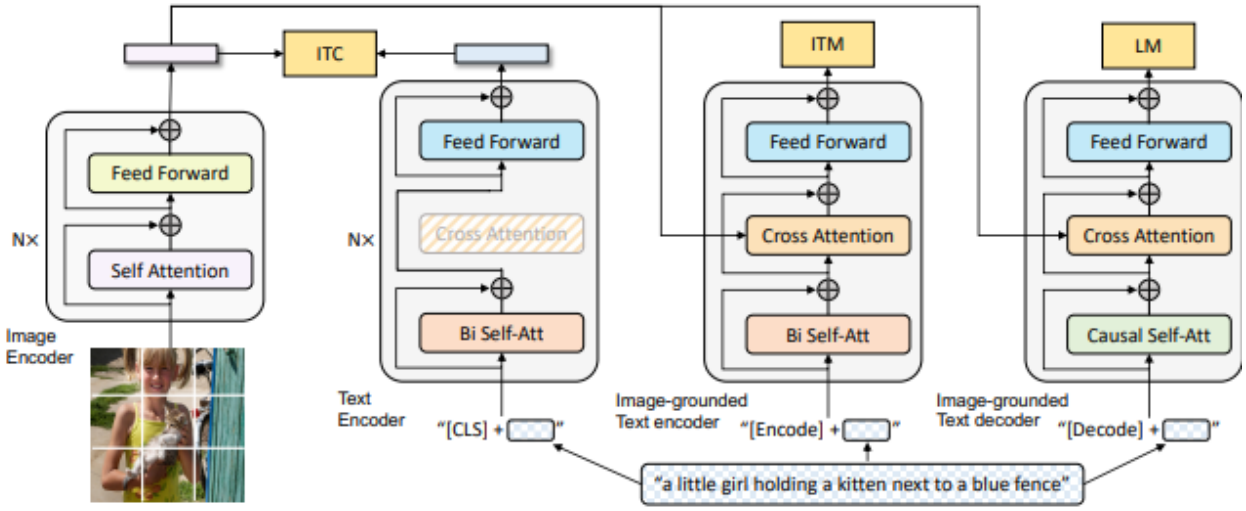
Việc trích xuất các đối tượng riêng biệt từ ảnh gốc là bước quan trọng trong quá trình huấn luyện mô hình, giúp hệ thống tập trung vào đặc trưng biểu cảm của từng người, đồng thời kết hợp với thông tin ngữ cảnh xung quanh để nâng cao độ chính xác của việc nhận diện cảm xúc.

2.2.1. Mô hình sinh Chú Thích

2.2.2. Mô hình BLIP

Trong bối cảnh trí tuệ nhân tạo ngày càng phát triển, sự kết hợp giữa xử lý hình ảnh và xử lý ngôn ngữ tự nhiên đã mở ra một hướng nghiên cứu đầy tiềm năng, được gọi là các bài toán Vision-Language (VL). Các mô hình VL không chỉ đòi hỏi khả năng hiểu nội dung hình ảnh mà còn phải liên kết chúng với ngôn ngữ tự nhiên một cách chính xác và hiệu quả. Trong số các mô hình nổi bật, BLIP [39] (Bootstrapping Language-Image Pre-training), được phát triển bởi Salesforce Research, đã thu hút sự chú ý nhờ cách tiếp cận sáng tạo và hiệu suất vượt trội. Mục tiêu chính của BLIP là xây dựng một hệ thống đa phương thức có khả năng học từ dữ liệu thô, đồng thời tối ưu hóa hiệu quả huấn luyện mà không cần phụ thuộc quá nhiều vào các tập dữ liệu được gắn nhãn thủ công.

BLIP được thiết kế dựa trên ý tưởng tận dụng dữ liệu sẵn có từ internet – một nguồn tài nguyên phong phú nhưng thường chứa nhiều nhiễu. Không giống như các mô hình truyền thống yêu cầu dữ liệu huấn luyện được chuẩn bị kỹ lưỡng, BLIP sử dụng các kỹ thuật tự giám sát để tự học từ các cặp ảnh-văn bản không hoàn hảo. Quá trình này được gọi là bootstrapping, trong đó mô hình tự cải thiện chất lượng dữ liệu thông qua các bước huấn luyện lặp đi lặp lại. Sự đổi mới này không chỉ giúp giảm chi phí chuẩn bị dữ liệu mà còn mở rộng khả năng ứng dụng của mô hình trong thực tế, nơi dữ liệu thường không được kiểm soát chặt chẽ.



Hình 2.4. Cấu trúc mô hình BLIP [39]

Về mặt kiến trúc, BLIP bao gồm ba thành phần chính: Image Encoder, Text Encoder và Text Decoder. Image Encoder, thường dựa trên Vision Transformer, chịu trách nhiệm trích xuất đặc trưng từ hình ảnh, biến chúng thành các biểu diễn số mà mô hình có thể hiểu được. Text Encoder, dựa trên kiến trúc Transformer, xử lý văn bản đầu vào để tạo ra các biểu diễn ngôn ngữ tương ứng. Cuối cùng, Text Decoder đóng vai trò sinh ra văn bản từ các biểu diễn kết hợp giữa hình ảnh và ngôn ngữ, hỗ trợ các tác vụ như tạo chú thích ảnh. Điểm độc đáo của BLIP nằm ở việc tích hợp cả hai khả năng "hiểu" và "sinh" trong một kiến trúc thống nhất, cho phép các thành phần chia sẻ thông tin và học hỏi lẫn nhau trong quá trình huấn luyện.

Một trong những thách thức lớn nhất của các bài toán Vision-Language là chất lượng dữ liệu huấn luyện. Dữ liệu thu thập từ web thường không đồng nhất, với các cặp ảnh-văn bản không luôn khớp nhau về nội dung. Để giải quyết vấn đề này, BLIP sử dụng hai mô hình phụ trợ: Captioner và Filter. Captioner tạo ra các chú thích mới cho hình ảnh dựa trên nội dung thực tế, trong khi Filter loại bỏ các cặp dữ liệu không liên quan hoặc sai lệch. Quá trình này không chỉ cải thiện độ chính xác của mô hình mà còn giúp BLIP học được các mối liên hệ sâu sắc hơn giữa hình ảnh và ngôn ngữ. Nhờ vậy, BLIP có thể hoạt động hiệu

quả ngay cả khi dữ liệu đầu vào ban đầu không hoàn hảo, một ưu điểm lớn so với các phương pháp truyền thống.

BLIP thể hiện tính linh hoạt qua khả năng ứng dụng trong nhiều tác vụ VL khác nhau. Chẳng hạn, trong bài toán Image Captioning, mô hình có thể tự động sinh ra chú thích phù hợp với nội dung hình ảnh, chẳng hạn như mô tả một bức ảnh phong cảnh hoặc sự kiện. Trong Visual Question Answering (VQA), BLIP trả lời các câu hỏi dựa trên thông tin từ hình ảnh, ví dụ như "Màu sắc của chiếc xe trong ảnh là gì?". Ngoài ra, BLIP còn hỗ trợ Image-Text Retrieval, cho phép tìm kiếm ảnh dựa trên truy vấn văn bản hoặc ngược lại. Những ứng dụng này không chỉ có giá trị trong nghiên cứu mà còn mang lại tiềm năng lớn trong các lĩnh vực thực tế như thương mại điện tử, giáo dục và truyền thông.

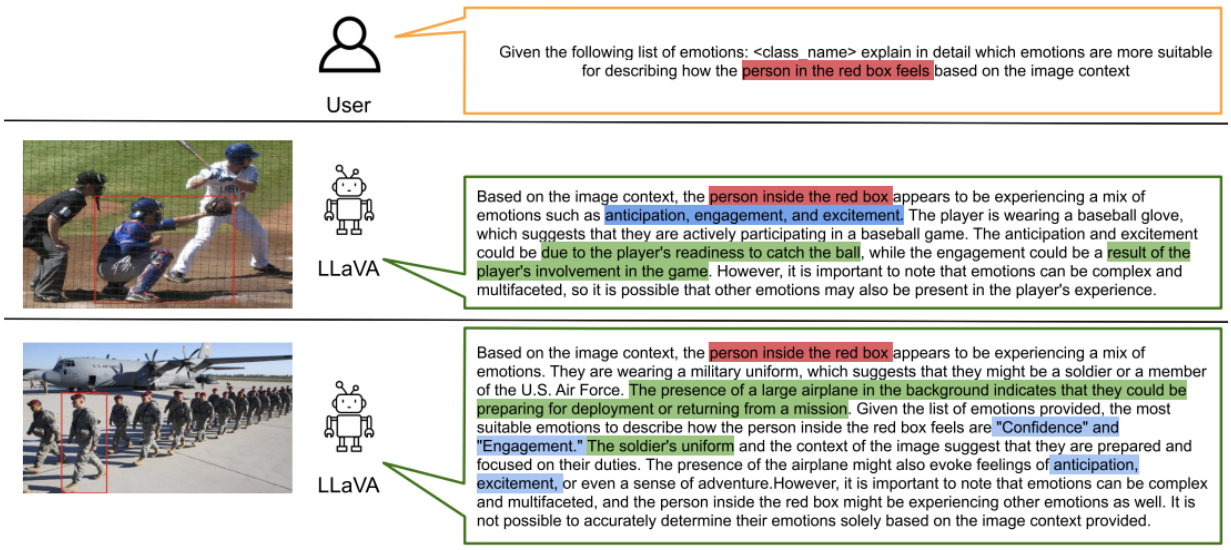
Mặc dù sở hữu nhiều ưu điểm, BLIP không phải không có hạn chế. Để đạt hiệu suất tối ưu, mô hình vẫn cần huấn luyện end-to-end trên các tập dữ liệu lớn, điều này đòi hỏi tài nguyên tính toán đáng kể. Hơn nữa, hiệu quả của BLIP phụ thuộc nhiều vào chất lượng ban đầu của dữ liệu thô; nếu dữ liệu quá nhiễu hoặc không đủ đa dạng, kết quả có thể bị ảnh hưởng. Tuy nhiên, những nhược điểm này không làm giảm giá trị của BLIP. Ngược lại, chúng mở ra cơ hội cho các nghiên cứu tiếp theo nhằm cải thiện khả năng thích nghi và hiệu suất của mô hình trong các điều kiện khác nhau.

BLIP là một bước tiến quan trọng trong lĩnh vực Vision-Language, kết hợp giữa sáng tạo kỹ thuật và tính thực tiễn cao. Với khả năng tận dụng dữ liệu thô, kiến trúc linh hoạt và hiệu suất ấn tượng trên nhiều tác vụ, BLIP không chỉ là một công cụ hữu ích cho các nhà nghiên cứu mà còn đặt nền móng cho các ứng dụng AI đa phương thức trong tương lai.

2.2.3. Mô hình LLaVa-1.5

LLaVa-1.5 [40] (Large Language and Vision Assistant) là một mô hình tiên tiến trong lĩnh vực hiểu và xử lý ngôn ngữ kết hợp với thị giác, được thiết kế để kết hợp thông tin từ hình ảnh và văn bản nhằm thực hiện các tác vụ như trả lời câu hỏi từ hình ảnh sinh caption mô tả hình ảnh và hiểu ngữ cảnh trực quan. Mô hình LLaVa được phát triển bởi nhóm

nghiên cứu từ Đại học Wisconsin-Madison, Microsoft Research và Đại học Columbia. Ra mắt vào ngày 5 tháng 10 năm 2023, LLaVa-1.5 không chỉ gây ấn tượng bởi hiệu suất vượt trội mà còn bởi thiết kế đơn giản, tiết kiệm tài nguyên và tính mã nguồn mở, cho phép cộng đồng nghiên cứu dễ dàng tiếp cận và phát triển thêm. Với khả năng xử lý đồng thời văn bản và hình ảnh, mô hình này đã đạt được kết quả tốt nhất trên 11 bài kiểm tra tiêu chuẩn, cạnh tranh trực tiếp với các hệ thống tiên tiến như GPT-4V của OpenAI, đồng thời đặt nền móng cho các ứng dụng thực tiễn trong tương lai.



Hình 2.5. Mô hình LlaVa [40]

LLaVA-1.5 được xây dựng dựa trên sự kết hợp giữa một bộ mã hóa thị giác đã được huấn luyện sẵn (CLIP-ViT-L-336px) và một mô hình ngôn ngữ lớn là Vicuna, dựa trên LlaMA dùng để liên kết thông qua một lớp ánh xạ MLP cải tiến thay vì lớp tuyến tính đơn giản như phiên bản trước. Mô hình LLaVA được xây dựng trên nền tảng của các mô hình mạnh mẽ trong lĩnh vực thị giác – ngôn ngữ, với ba thành phần chính:

- Vision Encoder (Mã hóa hình ảnh):

- Sử dụng Vision Transformer để trích xuất đặc trưng hình ảnh, giúp mô hình nhận diện các đối tượng, bối cảnh và chi tiết trong ảnh.

- Language Model (Mô hình ngôn ngữ lớn - LLM):
  - LLaVA sử dụng các mô hình LLM như LLaMA, GPT hoặc T5, giúp mô hình có thể hiểu và suy luận dựa trên thông tin hình ảnh đã mã hóa.

- Multi-modal Connector (Bộ nối kết đa phương thức):
  - Thành phần này giúp liên kết đầu ra từ Vision Encoder với LLM, cho phép mô hình có thể hiểu và tạo phản hồi dựa trên cả hai dạng dữ liệu (hình ảnh và văn bản).

Quá trình huấn luyện của LLaVA-1.5 được chia thành hai giai đoạn chính: giai đoạn căn chỉnh đặc trưng sử dụng 558 nghìn cặp hình ảnh-văn bản từ các nguồn dữ liệu công khai như LAION/CC/SBU, và giai đoạn tinh chỉnh toàn diện trên 150 nghìn mẫu hướng dẫn đa phương thức do GPT tạo ra, cùng với hơn 515 nghìn mẫu câu hỏi-thị giác (VQA) học thuật từ các bộ dữ liệu như OKVQA, và GQA. Điểm đặc biệt là mô hình chỉ cần khoảng 1,2 triệu mẫu dữ liệu – một con số khiêm tốn so với các đối thủ như Qwen-VL-Chat [41]– và hoàn tất huấn luyện trong vòng một ngày trên một node với 8 GPU A100. Điều này không chỉ chứng minh tính hiệu quả về mặt tính toán mà còn khẳng định LLaVA-1.5 là một giải pháp tiết kiệm tài nguyên, phù hợp cho các tổ chức hoặc cá nhân với nguồn lực hạn chế.

Về hiệu suất, LLaVa-1.5 thể hiện khả năng vượt trội trong các tác vụ như trả lời câu hỏi dựa trên hình ảnh, mô tả hình ảnh và suy luận đa phương thức. Tuy nhiên, LLaVa-1.5 vẫn tồn tại một số hạn chế nhất định. Khả năng nhận diện ký tự quang học (OCR) của mô hình chưa thực sự hoàn hảo, thường gặp lỗi khi xử lý văn bản trong hình ảnh phức tạp, và hiện tượng "ảo giác"– tức việc tạo ra thông tin không chính xác – vẫn xảy ra trong một số

trường hợp. Ngoài ra, mô hình chưa được tối ưu hóa cho các tác vụ như mã hóa giao diện từ hình ảnh thiết kế, giới hạn phạm vi ứng dụng trong một số lĩnh vực kỹ thuật cụ thể.

Mặc dù vậy, với tính mã nguồn mở và sự hỗ trợ từ cộng đồng, LLaVa-1.5 mang lại tiềm năng ứng dụng rộng lớn. Người dùng có thể thử nghiệm mô hình thông qua các bản demo trên Hugging Face hoặc Gradio, hoặc cài đặt cục bộ với yêu cầu phần cứng tương đối thấp (GPU tương thích CUDA, Python 3.10+). Các nhà nghiên cứu cũng có thể tinh chỉnh mô hình trên dữ liệu riêng bằng kỹ thuật LoRA [42] để giảm yêu cầu tài nguyên mà vẫn duy trì hiệu suất. So với các đối thủ, LLaVa-1.5 không chỉ nổi bật về chi phí thấp và tính linh hoạt, mà còn là minh chứng cho khả năng đạt được hiệu suất cao với dữ liệu và tài nguyên hạn chế.

So với các mô hình truyền thống như BLIP, CLIP hoặc ExpansionNet-v2 [43], LLaVa có khả năng tích hợp ngữ cảnh tốt hơn, tạo ra các mô tả chi tiết và tự nhiên hơn về nội dung hình ảnh. Sự kết hợp này giúp LLaVA có thể suy luận về hình ảnh theo cách tương tự như con người, chẳng hạn như mô tả một bức ảnh không chỉ dựa trên nội dung trực quan mà còn hiểu rõ về bối cảnh và ý nghĩa của nó. Trong tương lai, với sự phát triển không ngừng của công nghệ AI đa phương thức, LLaVa-1.5 hứa hẹn sẽ tiếp tục đóng vai trò quan trọng trong việc thúc đẩy các ứng dụng thực tiễn, từ trợ lý thông minh đến hỗ trợ giáo dục và nghiên cứu khoa học.

2.3. Mô hình mạng nơ-ron tích chập

2.3.1. Mô hình GCN

Trong lĩnh vực trí tuệ nhân tạo và học sâu, dữ liệu dạng đồ thị ngày càng trở nên quan trọng nhờ khả năng biểu diễn các mối quan hệ phức tạp giữa các thực thể, như mạng xã hội, cấu trúc phân tử, hay mạng lưới giao thông. Để xử lý loại dữ liệu này, các mô hình mạng nơ-ron truyền thống như CNN hay RNN không phù hợp do chúng chủ yếu hoạt động trên dữ liệu có cấu trúc lưới hoặc chuỗi Graph Convolutional Network, được giới thiệu trong bài báo “Semi-Supervised Classification with Graph Convolutional Networks” [44] đã mở

ra một hướng tiếp cận mới bằng cách tổng quát hóa khái niệm tích chập từ dữ liệu Euclidean sang dữ liệu phi Euclidean như đồ thị. GCN không chỉ tận dụng thông tin cấu trúc của đồ thị mà còn kết hợp đặc trưng của các nút để thực hiện các tác vụ như phân loại nút, dự đoán liên kết, hay phân loại đồ thị.

Cấu trúc mô hình

Kiến trúc GCN được xây dựng dựa trên ý tưởng truyền thông tin giữa các nút trong đồ thị thông qua các lớp tích chập đồ thị. Một đồ thị được định nghĩa bởi tập hợp các nút V và cạnh E, thường biểu diễn dưới dạng ma trận kề A và ma trận đặc trưng X chứa thông tin của từng nút. Trong GCN, mỗi lớp tích chập thực hiện hai bước chính: một là tổng hợp thông tin từ các nút láng giềng và hai là cập nhật biểu diễn của nút. Công thức cơ bản (1) của một lớp GCN được biểu diễn như sau:

$$H^{(k)} = \sigma\left(\widetilde{D}^{-\frac{1}{2}}\tilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(k-1)}W^{(k-1)}\right) \qquad (1)$$

Trong đó:

$H^{(k)}$ là ma trận biểu diễn nút tại lớp k

$\tilde{A} = A + I$ với A là ma trận kề, I là ma trận đơn vị, nhằm đảm bảo tính ổn định.

$\widetilde{D}$ là ma trận bậc (degree matrix) của $\tilde{A}$, với $\widetilde{D}_{ii} = \sum_j \tilde{A}_{ij}$

$W^{(k-1)}$ là ma trận trọng số huấn luyện tại lớp k

σ là hàm kích hoạt phi tuyến

Quá trình chuẩn hóa $\widetilde{D}^{-\frac{1}{2}}\tilde{A}\widetilde{D}^{-\frac{1}{2}}$ giúp đảm bảo rằng thông tin từ các nút láng giềng được tổng hợp một cách cân bằng, tránh hiện tượng gradient biến mất hoặc bùng nổ trong quá trình huấn luyện. Kết quả là sau mỗi lớp, GCN tạo ra một biểu diễn mới cho các nút, kết hợp cả đặc trưng ban đầu và thông tin cấu trúc từ đồ thị.

GCN hoạt động theo cơ chế lan truyền thông tin, trong đó mỗi nút tổng hợp đặc trưng từ các nút láng giềng dựa trên ma trận kề. Không giống CNN sử dụng bộ lọc trượt trên lưới pixel, GCN sử dụng cấu trúc đồ thị để xác định "vùng láng giềng" của mỗi nút. Quá trình này có thể được xem như một dạng trung bình có trọng số của đặc trưng láng giềng, với trọng số được điều chỉnh thông qua huấn luyện. Trong bài toán phân loại nút bán giám sát, GCN tận dụng một phần nhỏ nhãn có sẵn để dự đoán nhãn cho các nút còn lại, nhờ khả năng lan truyền thông tin hiệu quả qua các lớp.

Về mặt huấn luyện, GCN thường sử dụng phương pháp lan truyền ngược với hàm mất mát như cross-entropy. Tuy nhiên, do tính chất toàn cục của đồ thị, GCN ban đầu được thiết kế theo kiểu full-batch gradient descent, tức là tính toán gradient trên toàn bộ đồ thị trong mỗi lần lặp. Điều này dẫn đến hạn chế về bộ nhớ và thời gian khi xử lý đồ thị lớn, một vấn đề đã được các mô hình sau này như GraphSAGE [45] cải thiện bằng cách sử dụng mini-batch.

GCN sở hữu nhiều ưu điểm nổi bật. Thứ nhất, nó đơn giản và hiệu quả, chỉ cần vài lớp tích chập để đạt hiệu suất cao trên các bài toán như phân loại nút trên tập dữ liệu CORA [44] (đạt độ chính xác khoảng 81,5%). Thứ hai, GCN tận dụng tốt cấu trúc đồ thị, cho phép học biểu diễn nút mà không cần phụ thuộc hoàn toàn vào dữ liệu nhãn. Tuy nhiên, mô hình cũng có những hạn chế đáng kể. Việc sử dụng full-batch khiến GCN không khả thi với đồ thị quy mô lớn (hàng triệu nút). Ngoài ra, GCN giả định rằng các nút láng giềng có mức độ quan trọng như nhau, điều này không phù hợp với các đồ thị mà mối quan hệ giữa các nút có sự khác biệt đáng kể. Hơn nữa, khi số lớp tích chập tăng lên, hiện tượng "quá mượt" có thể xảy ra, làm mất đi sự khác biệt giữa các biểu diễn nút.

GCN đã được ứng dụng rộng rãi trong nhiều lĩnh vực, từ phân tích mạng xã hội về dự đoán hành vi người dùng, hóa học, đến gợi ý sản phẩm. Sự thành công của GCN đã truyền cảm hứng cho nhiều biến thể cải tiến, như Graph Attention Network [46]– bổ sung cơ chế attention để phân bổ trọng số khác nhau cho các nút láng giềng, hay GraphSAGE – hỗ trợ

huấn luyện inductive trên đồ thị động. Những phát triển này khắc phục một phần hạn chế của GCN, mở rộng khả năng ứng dụng trong thực tế.

2.3.2. Mô hình GIN

Trong lĩnh vực học sâu trên dữ liệu đồ thị, Graph Isomorphism Network (GIN) là một kiến trúc mạng nơ-ron đồ thị (GNN) nổi bật, được đề xuất bởi Xu và các cộng sự [47]vào năm 2018. GIN được thiết kế để đạt được sức mạnh phân biệt tối đa trong việc biểu diễn cấu trúc đồ thị, tương đương với bài kiểm tra đẳng cấu đồ thị Weisfeiler-Lehman (WL test) – một phương pháp truyền thống dùng để kiểm tra xem hai đồ thị có đồng cấu hay không. Với khả năng nắm bắt tốt các đặc trưng cấu trúc và mối quan hệ giữa các nút trong đồ thị, GIN đã trở thành một công cụ quan trọng trong các tác vụ như phân loại đồ thị, dự đoán tính chất vật liệu, hay phân tích mạng xã hội.

GIN thuộc nhóm GNN không gian (spatial-based GNN), tập trung vào việc tổng hợp thông tin từ các nút láng giềng dựa trên mối quan hệ không gian trong đồ thị, thay vì phân tích phổ như các mô hình GCN. Cấu trúc của GIN bao gồm nhiều lớp, mỗi lớp thực hiện quá trình cập nhật biểu diễn của các nút thông qua hai bước chính: tổng hợp và biến đổi. Công thức toán học (2) cốt lõi của một lớp GIN được định nghĩa như sau:

$$h_v^{(k)} = MLP^{(k)}\left((1 + \epsilon^{(k)}\right) \cdot h_v^{(k-1)} + \sum_{u \in N(v)} h_u^{(k-1)} \qquad (2)$$

Trong đó:

$h_v^{(k)}$ là biểu diễn (embedding) của nút v tại lớp k

$\epsilon^{(k)}$ là tham số có thể huấn luyện

$N(v)$ là tập các nút lân cận trực tiếp của nút v

MLP là mạng neural đa lớp giúp mô hình học được các đặc trưng phi tuyến tính của đồ thị

Công thức này cho thấy GIN tổng hợp thông tin bằng cách cộng tất cả biểu diễn của các nút láng giềng và chính nút v, sau đó áp dụng một MLP để tạo ra biểu diễn mới. Sau nhiều lớp, biểu diễn của toàn bộ đồ thị được tạo ra bằng cách tổng hợp (hoặc gộp) các biểu diễn nút, thường sử dụng phép cộng và kết hợp với một hàm readout (ví dụ: sum, mean) để phục vụ các tác vụ như phân loại đồ thị.

GIN có nhiều ưu điểm nổi bật so với các GNN khác như:

- Sức mạnh phân biệt: GIN đạt được độ phân biệt tương đương với bài kiểm tra WL 1 chiều (1-WL), giúp nó phân biệt hầu hết các đồ thị không đồng cấu mà các mô hình như GCN không làm được.
- Tính linh hoạt: Việc sử dụng MLP cho phép GIN học các hàm phức tạp hơn, phù hợp với nhiều loại đồ thị và tác vụ khác nhau.
- Hiệu quả thực nghiệm: GIN đã chứng minh hiệu suất vượt trội trên nhiều tập dữ liệu chuẩn như PROTEINS [48], NCI1[47]trong các bài toán phân loại đồ thị.

Dù mạnh mẽ, GIN vẫn tồn tại một số hạn chế:

- Phụ thuộc vào độ sâu: Số lớp k phải được chọn sao cho phù hợp với đường kính đồ thị, nếu không GIN có thể không nắm bắt được toàn bộ cấu trúc.
- Không phân biệt một số đồ thị đặc biệt: GIN chỉ tương đương với 1-WL, nên không thể phân biệt một số đồ thị mà các biến thể WL cao hơn (như 2-WL, 3-WL) có thể xử lý.
- Chi phí tính toán: Việc sử dụng MLP trong mỗi lớp làm tăng độ phức tạp tính toán, đặc biệt với đồ thị lớn.

GIN đã được ứng dụng rộng rãi trong các lĩnh vực như hóa học, sinh học và vật lý. Các cải tiến sau này, như GIN với đặc trưng cạnh hoặc kết hợp với cơ chế attention, đã khắc phục một phần hạn chế của phiên bản gốc, mở rộng phạm vi ứng dụng.

GIN là một kiến trúc GNN mạnh mẽ với cấu trúc đơn giản nhưng hiệu quả, tận dụng tốt mối quan hệ không gian trong đồ thị. Dù còn một số thách thức, GIN đã đặt nền móng cho các nghiên cứu tiếp theo trong lĩnh vực học sâu trên đồ thị, hứa hẹn mang lại nhiều đóng góp quan trọng trong tương lai.

2.3.1. Mô hình PNA

Mô hình Principal Neighbourhood Aggregation là một kiến trúc học sâu dựa trên Graph Neural Networks, được thiết kế để xử lý các bài toán liên quan đến dữ liệu đồ thị. PNA cải thiện hiệu suất của các mô hình GNN truyền thống bằng cách sử dụng một tập hợp các hàm tổng hợp và các hàm chuẩn hóa nhằm tăng cường khả năng biểu diễn của mô hình. Phương pháp này được giới thiệu trong bài báo "Principal Neighbourhood Aggregation for Graph Nets" bởi Corso et al [51], được công bố tại hội nghị NeurIPS 2020.

PNA được phát triển để giải quyết vấn đề về khả năng tổng quát hóa và biểu diễn của các GNN khi xử lý các đồ thị có cấu trúc phức tạp và đa dạng. Không giống như các mô hình GNN truyền thống như GCN hay GraphSAGE, vốn chỉ sử dụng một hàm tổng hợp đơn lẻ (ví dụ: trung bình, tổng, hoặc max-pooling), PNA kết hợp nhiều hàm tổng hợp như trung bình, tổng, tối đa, tối thiểu và độ lệch chuẩn. Điều này giúp mô hình nắm bắt được các đặc trưng đa dạng từ láng giềng của một nút, từ đó cải thiện khả năng phân biệt các cấu trúc đồ thị khác nhau.

Ngoài ra, PNA sử dụng các hàm chuẩn hóa để điều chỉnh giá trị tổng hợp dựa trên kích thước của tập láng giềng, giúp giảm thiểu vấn đề mất cân bằng khi các nút có số lượng láng giềng khác nhau. PNA đã đạt được hiệu suất vượt trội trên nhiều tập dữ liệu chuẩn như ZINC, MUTAG và PPI, chứng minh tính hiệu quả của phương pháp này trong các bài toán phân loại nút, phân loại đồ thị và dự đoán thuộc tính.

CHƯƠNG 3. XÂY DỰNG ĐỒ THỊ TRI THỨC CHO PHÂN TÍCH CẢM XÚC TỪ DỮ LIỆU HÌNH ẢNH

Với chương 3, bài nghiên cứu trình bày việc xử lý bộ dữ liệu EMOTIC và quá trình xây dựng nên đồ thị tri thức cho phân tích cảm xúc từ dữ liệu hình ảnh.

3.1. Tạo chú thích cho hình ảnh trong bộ dữ liệu

Khi đã lấy được các hình ảnh trong bộ dữ liệu EMOTIC, dữ liệu được đưa vào mô hình BLIP để sinh câu chú thích. Mô hình BLIP được sử dụng để tạo ra các chú thích mô tả chính xác nội dung hình ảnh đầu vào. Bên cạnh đó, BLIP là mô hình nằm trong "State of the art" về khả năng sinh chú thích bao quát về ngữ cảnh.

Dữ liệu đầu vào là file .csv chứa thông tin về thư mục, tên ảnh và chỉ số index tương ứng. Các đường dẫn ảnh được chuẩn hóa lại theo cấu trúc:

Tên thư mục + Tên ảnh + _ + Index + .jpg

Sau đó, từng ảnh được truyền qua mô hình BLIP để sinh caption. Mỗi caption mô tả nội dung tổng quát của ảnh, từ đó giúp cung cấp thêm ngữ cảnh cho hệ thống nhận diện cảm xúc. Kết quả đầu ra là tập file trước đó với thêm cột Caption được sinh ra, được lưu lại trong file kết quả (captions_train.csv, captions_test.csv hoặc caption_full.csv).

Folder	Filename	Image Size	BBox	Categorical_l	Continuous_	Gender	Age	Caption
mscoco/im	COCO_val2	[640, 640]	[86, 58, 56	['Disconnecti	[5, 3, 9]	Male	Adult	a man wearing a purple shirt and a purple tie
mscoco/im	COCO_trair	[640, 480]	[485, 149,	['Anticipation	[6, 4, 7]	Male	Adult	a bride and groom cut their wedding cake
mscoco/im	COCO_val2	[640, 480]	[305, 92, 4	['Engagemen	[7, 8, 8]	Male	Teenager	a man standing in a field holding a frisbee
mscoco/im	COCO_trair	[480, 640]	[221, 63, 4	['Aversion', 'P	[8, 9, 8]	Male	Kid	a little boy sitting at a table with a dog
mscoco/im	COCO_trair	[500, 333]	[44, 143, 1	['Confidence'	[7, 9, 10]	Male	Adult	a baseball player swinging a bat at a ball
mscoco/im	COCO_trair	[640, 478]	[42, 32, 41	['Anticipation	[3, 6, 8]	Male	Adult	a man taking a picture of himself in a mirror
mscoco/im	COCO_val2	[500, 375]	[257, 39, 4	['Anticipation	[6, 7, 7]	Male	Adult	a man sitting at a desk with headphones on
mscoco/im	COCO_trair	[640, 429]	[336, 80, 4	['Anticipation	[7, 7, 8]	Male	Kid	two people on skis standing on top of a snow covered slope
mscoco/im	COCO_trair	[640, 480]	[188, 109,	['Engagemen	[7, 4, 7]	Female	Kid	a little girl sitting on a couch holding an umbrella
mscoco/im	COCO_trair	[350, 500]	[198, 29, 3	['Fatigue', 'Ha	[7, 7, 6]	Male	Teenager	a young girl paddles a raft in the water

Hình 3.1. Tập captions_train.csv sau khi tạo chú thích

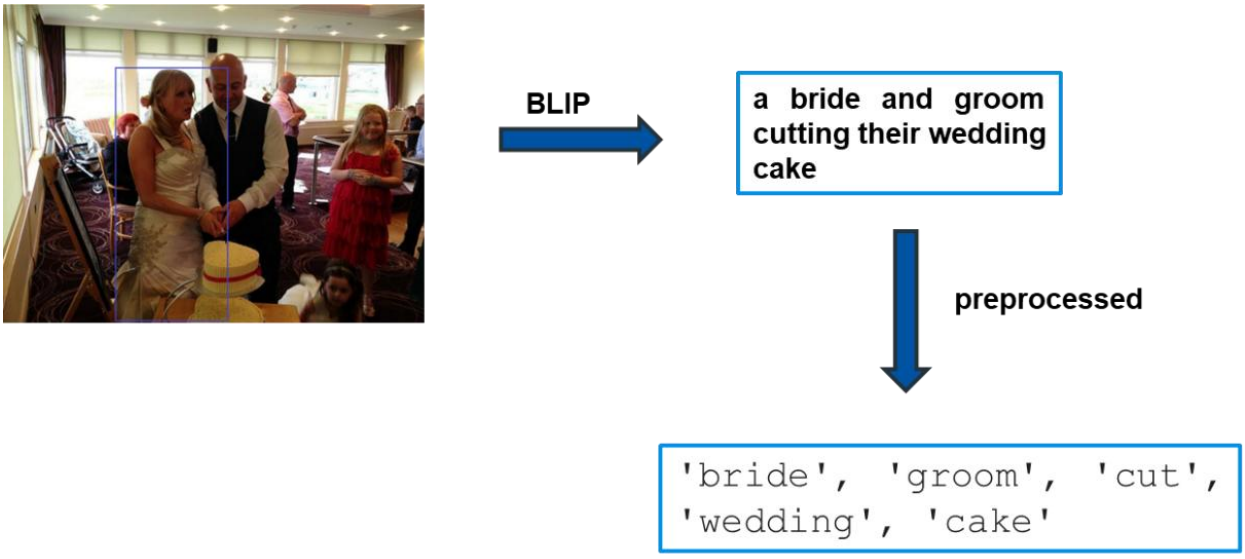
Sau khi thu được các caption mô tả ảnh, bước tiếp theo là tiến hành tiền xử lý ngôn ngữ tự nhiên để trích xuất ra các từ khóa tiềm năng phục vụ xây dựng đồ thị tri thức. Sử dụng thư viện spaCy, các caption được phân tích cú pháp và lọc bỏ những từ stop words (như “is”, “in”, “the”, ...) và danh từ chung đại diện cho con người (như "man", "woman", "children", "people", ...) nhằm tập trung vào những từ mang tính mô tả hành động, trạng thái, đồ vật, hoặc ngữ cảnh.

Tiếp đó, mỗi từ còn lại được chuẩn hóa về dạng gốc (lemma) nhằm giảm độ đa dạng từ vựng và dễ dàng tra cứu trong các nguồn tri thức (như SenticNet). Ví dụ, các từ như “running”, “ran”, “runs” sẽ được chuyển về “run”. Tập chú thích sau khi được xử lý sẽ được lưu lại thành một cột mới có tên “processed_caption” trong file dữ liệu để phục vụ cho các bước xây dựng đồ thị tri thức hoặc khai thác tri thức tiếp theo.

Folder	Filename	Image Size	BBox	Categorical_l	Continuous_	Gender	Age	Caption	processed_caption
mscoco/im	COCO_val2	[640, 640]	[86, 58, 5(	['Disconnecti	[5, 3, 9]	Male	Adult	a man wear	['wear', 'purple', 'shirt', 'tie']
mscoco/im	COCO_trair	[640, 480]	[485, 149,	['Anticipation	[6, 4, 7]	Male	Adult	a bride and	['bride', 'groom', 'cut', 'wedding', 'cake']
mscoco/im	COCO_val2	[640, 480]	[305, 92, 4	['Engagemen	[7, 8, 8]	Male	Teenager	a man stan	['stand', 'field', 'hold', 'frisbee']
mscoco/im	COCO_trair	[480, 640]	[221, 63, 4	['Aversion', 'P	[8, 9, 8]	Male	Kid	a little boy s	['little', 'sit', 'table', 'dog']
mscoco/im	COCO_trair	[500, 333]	[44, 143, 1	['Confidence'	[7, 9, 10]	Male	Adult	a baseball p	['baseball', 'player', 'swinge', 'bat', 'ball']
mscoco/im	COCO_trair	[640, 478]	[42, 32, 41	['Anticipation	[3, 6, 8]	Male	Adult	a man takin	['take', 'picture', 'mirror']
mscoco/im	COCO_val2	[500, 375]	[257, 39, 4	['Anticipation	[6, 7, 7]	Male	Adult	a man sittir	['sit', 'desk', 'headphone']
mscoco/im	COCO_trair	[640, 429]	[336, 80, 4	['Anticipation	[7, 7, 8]	Male	Kid	two people	['ski', 'stand', 'snow', 'cover', 'slope']
mscoco/im	COCO_trair	[640, 480]	[188, 109,	['Engagemen	[7, 4, 7]	Female	Kid	a little girl s	['little', 'sit', 'couch', 'hold', 'umbrella']
mscoco/im	COCO_trair	[350, 500]	[198, 29, 3	['Fatigue', 'Ha	[7, 7, 6]	Male	Teenager	a young girl	['young', 'paddle', 'raft', 'water']

Hình 3.2. Tập captions_train.csv sau khi tiền xử lý

Và các bước sinh chú thích cũng như xử lý dữ liệu được biểu diễn tại hình 3.3.



Hình 3.3: Quy trình xử lý chú thích

Sau đó, tiến hành chỉ lấy những cột cần thiết để tiến hành làm đầu vào cho mô hình BLIP. Tập dữ liệu sau khi lọc bỏ các cột không cần thiết có hình như dưới. Và đây cũng là cấu trúc bộ dữ liệu đầu vào để xây dựng đồ thị tri thức cũng như huấn luyện và test.

Folder	Filename	Categorical Labels	Continuous Labels	Caption	Processed caption
mscoco/images	COCO_val2014_000000562243.jpg	['Disconnection', 'Doubt/Confusion']	[5, 3, 9]	a man wearing a purple shirt and a purple tie	['wear', 'purple', 'shirt', 'purple', 'tie']
mscoco/images	COCO_train2014_000000288841.jpg	['Anticipation']	[6, 4, 7]	a bride and groom cut their wedding cake	['bride', 'groom', 'cut', 'wedding', 'cake']
mscoco/images	COCO_val2014_000000558171.jpg	['Engagement', 'Excitement', 'Happiness']	[7, 8, 8]	a man standing in a field holding a frisbee	['stand', 'field', 'hold', 'frisbee']
mscoco/images	COCO_train2014_000000369575.jpg	['Aversion', 'Pleasure']	[8, 9, 8]	a little boy sitting at a table with a dog	['little', 'sit', 'table', 'dog']
mscoco/images	COCO_train2014_000000213009.jpg	['Confidence', 'Excitement']	[7, 9, 10]	a baseball player swinging a bat at a ball	['baseball', 'player', 'swinge', 'bat', 'ball']

Hình 3.4: Bảng dữ liệu sau khi chọn lọc

3.2. Xây dựng đồ thị tri thức

Đồ thị tri thức được xây dựng từ dữ liệu caption sau khi thực hiện qua một số bước cụ thể như sau. Đầu tiên, các từ hợp lệ được đưa vào. Tiếp theo, xây dựng đồ thị tri thức dựa trên những từ hợp lệ này bằng cách thêm từng từ làm nút trong đồ thị. Hai sáu cảm xúc cũng lần lượt được được thêm vào đồ thị tương tự. Giữa từ và cảm xúc ($e = (V_w, V_{C_i})$) được liên kết với nhau bằng trọng số vào dựa trên xác suất điều kiện, tính từ ma trận đồng xuất hiện giữa từ và cảm xúc (Mc). Cụ thể, trọng số của cạnh từ-cảm xúc được tính bằng cách lấy số lần từ xuất hiện cùng cảm xúc đó chia cho tổng số lần từ xuất hiện với tất cả các cảm xúc khác dựa theo công thức 3:

$$w_e = P(C_i|W) = \frac{M_{C_{W,i}}}{\sum M_{C_W}} \qquad (3)$$

Trong đó:

W: Một từ hợp lệ (valid word) được trích xuất từ caption của ảnh.

C_i : Một loại cảm xúc thứ i (trong 26 loại cảm xúc của EMOTIC).

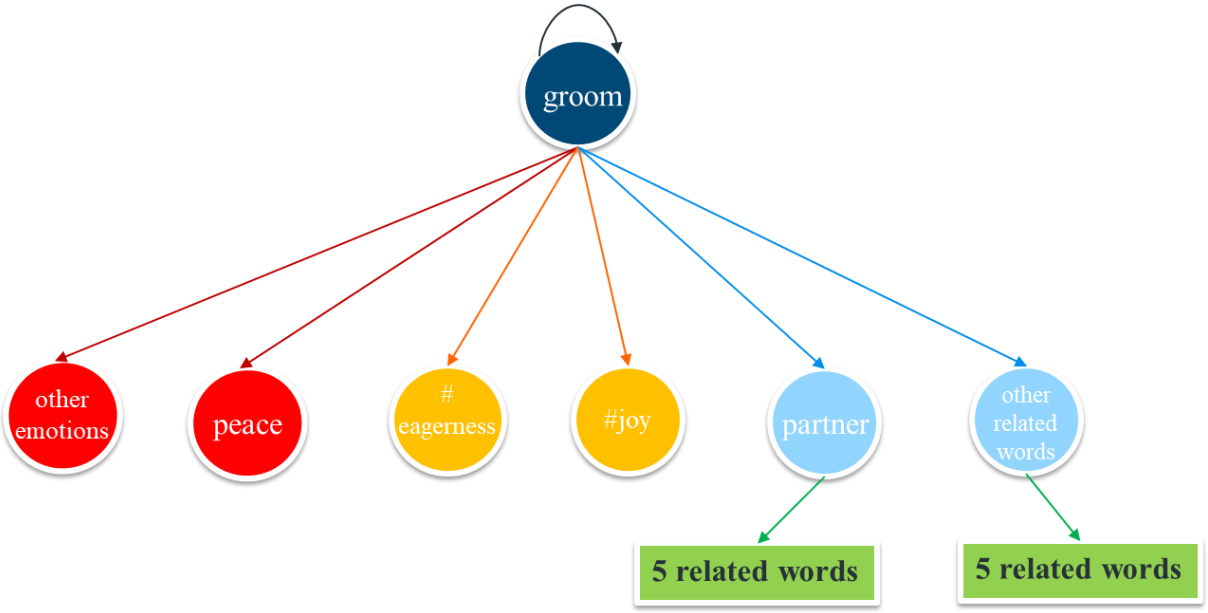
$M_{C_{W,i}}$: Số lần cảm xúc C_i xuất hiện cùng từ W trong tập dữ liệu (lấy từ ma trận đồng xuất hiện *co-occurrence matrix*).

$\sum M_{C_W}$: Tổng số lần mọi cảm xúc xuất hiện cùng từ W.

Cuối cùng, việc xây dựng các cạnh giữa các nút từ-từ, dựa vào ma trận đồng xuất hiện (Ms). Trọng số các cạnh này được tính bằng cách lấy tần suất đồng xuất hiện giữa hai từ chia cho tổng số lần xuất hiện của từ đầu tiên. Bước tiếp theo là bổ sung thêm thông tin ngữ nghĩa cho từng từ từ cơ sở tri thức SenticNet 5 [49]. SenticNet là một công cụ phân tích ngữ nghĩa và cảm xúc dựa trên đồ thị tri thức cảm xúc. Nó bao gồm các từ và khái niệm liên quan tới cảm xúc, cung cấp cơ sở dữ liệu phong phú để phục vụ phân tích cảm xúc từ

hình ảnh. SenticNet là bộ dữ liệu tri thức cảm xúc phong phú, chứa thông tin về các khái niệm và mối quan hệ ngữ nghĩa liên quan đến cảm xúc, thường được sử dụng trong xây dựng các đồ thị tri thức cảm xúc để phân tích và nhận diện cảm xúc chính xác.

Với mỗi từ hợp lệ, ta truy xuất cơ sở SenticNet để lấy các thông tin quan trọng như giá trị mức độ dễ chịu (pleasantness value), giá trị phân cực cảm xúc (polarity value), các nhãn tâm trạng (mood tags) và các từ liên quan về mặt ngữ nghĩa (related words). Với nhãn tâm trạng thì được truy xuất ra 2 từ và các từ liên quan được lấy ra 5 từ. Các mood tags được thêm vào đồ thị như các nút loại mood, và trọng số cạnh nối từ và mood được xác định bằng giá trị pleasantness value. Các từ liên quan về mặt ngữ nghĩa cũng được thêm vào đồ thị, với cạnh nối giữa từ hợp lệ và từ liên quan này có trọng số là giá trị phân cực (polarity value). Với các từ liên quan trên lại lấy thêm 5 từ liên quan nữa để tăng cường ngữ nghĩa cho mô hình. Trong quá trình lấy các từ liên quan mà bộ Senticnet không có đủ 5 từ thì sẽ các từ trong WordNet [50] để bổ sung. WordNet là một cơ sở dữ liệu từ vựng lớn được thiết kế dưới dạng mạng lưới ngữ nghĩa, trong đó mỗi từ được nhóm vào các nhóm đồng nghĩa, kèm theo mô tả ngữ nghĩa và quan hệ với các synsets khác như quan hệ đồng nghĩa, trái nghĩa, siêu loại. Do đó WordNet cũng là bộ dữ liệu hữu ích cho việc xây dựng đồ thị tri thức.



Hình 3.5. Hình minh họa cho các từ hợp lệ trong mỗi đồ thị tri thức

Từ các bước trên, kết quả cuối cùng là các đồ thị tri thức biểu diễn mối liên hệ giữa các từ và cảm xúc, phục vụ cho quá trình huấn luyện và đánh giá mô hình nhận diện cảm xúc sau này.

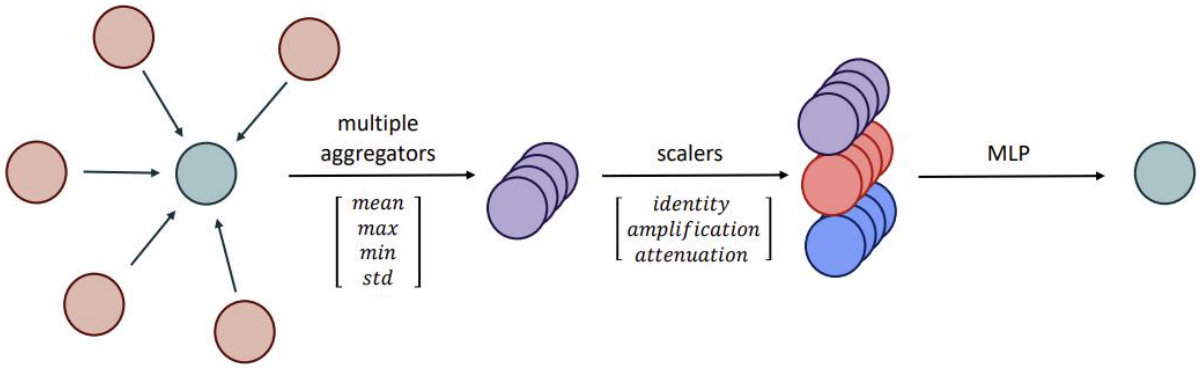
3.3. Sử dụng mô hình PNA

Khi đã có đồ thi tri thức cho mỗi tấm ảnh, đồ thị được tiến hành trích xuất đặc trưng. Trong đồ thị tri thức, mỗi nút được gán một vector đặc trưng, được lấy từ GloVe để nhúng. Ở đây sử dụng GloVe 200, chứa các vector nhúng 200 chiều được huấn luyện trước trên tập dữ liệu lớn, mỗi từ khóa tương ứng với một vector 200 chiều. Việc sử dụng GloVe để nhúng giúp cung cấp các biểu diễn ngữ nghĩa phong phú, được huấn luyện trên dữ liệu lớn, giúp mô hình hiểu được ý nghĩa của các từ khóa trong ngữ cảnh.

Các đồ thị tri thức sau đó sẽ được tổ chức thành các batch để đưa vào mô hình PNA. Mỗi đồ thị tương ứng với một mẫu dữ liệu (một hình ảnh và chú thích) được lấy từ tập dữ liệu dựa trên chỉ số index. Một batch đồ thị được tạo bằng cách sử dụng hàm dgl.batch(graph_batch) của thư viện DGL, trong đó graph_batch là danh sách các đồ thị riêng lẻ. Đặc trưng nút được trích xuất từ graphs.ndata.pop('x'), là các vector GloVe 200

chiều, và được chuyển thành kiểu dữ liệu thực (float32) để tính toán. Trọng số cạnh được lấy từ graphs.edata['weight'] và cũng được chuyển thành kiểu dữ liệu thực. Việc chuẩn bị batch đồ thị đảm bảo rằng dữ liệu được định dạng phù hợp với yêu cầu của PNA, đồng thời tối ưu hóa hiệu suất tính toán.

Mô hình PNA được lấy từ framework PyTorch. Mô hình PNAPredictor được cấu hình với các tham số đầu vào chính như sau: kích thước đặc trưng nút đầu vào là 200 (node_in_feats=200), kích thước đặc trưng cạnh đầu vào là 1 (edge_in_feats=1), và số lớp ẩn là 256 (node_out_feats=256). Mô hình bao gồm hai lớp PNA (num_layers=2), sử dụng tập hợp các hàm tổng hợp gồm trung bình (mean), tối đa (max), tối thiểu (min), độ lệch chuẩn (std), và moment bậc bốn (moment4). Ngoài ra, các hàm chuẩn hóa (scalers) bao gồm identity, amplification, và attenuation được áp dụng để điều chỉnh giá trị tổng hợp dựa trên kích thước tập láng giềng, với tham số delta được đặt là 2.5 nhằm kiểm soát độ nhạy của các hàm chuẩn hóa. Đầu ra của mô hình bao gồm 26 lớp phân loại (n_tasks=26) cho nhiệm vụ phân loại.



Hình 3.6. Cấu trúc mô hình PNA [51]

Kiến trúc của mô hình bao gồm ba thành phần chính: mô-đun nhúng lớp PNA, và mô-đun đọc dữ liệu. Trước tiên, các đặc trưng nút và cạnh được nhúng thông qua các mô-đun tuyến tính, tiếp theo là chuẩn hóa theo batch và hàm kích hoạt Mô-đun PNA, được triển khai thông qua lớp PNAConv của DGL, thực hiện quá trình truyền tin để cập nhật biểu diễn nút dựa trên thông tin từ các nút láng giềng và đặc trưng cạnh. Quá trình này được lặp lại

qua hai lớp PNA, với sự tích hợp của một mô-đun GRU để duy trì và cập nhật trạng thái ẩn của các biểu diễn nút, từ đó tăng cường khả năng học các đặc trưng phức tạp.

Sau khi các biểu diễn nút được cập nhật, mô-đun đọc dữ liệu sử dụng cơ chế Set2Set để tổng hợp thông tin từ toàn bộ đồ thị thành một vector đặc trưng cố định. Cơ chế này thực hiện hai bước lặp với một lớp ẩn, đảm bảo rằng các đặc trưng đồ thị được biểu diễn một cách toàn diện. Cuối cùng, vector đặc trưng đồ thị được đưa qua một mạng nơ-ron tuyến tính với hai lớp ẩn: lớp đầu tiên giảm kích thước từ 2 * node_out_feats xuống node_out_feats, áp dụng hàm ReLU, và lớp thứ hai ánh xạ xuống số chiều đầu ra tương ứng với nhiệm vụ. Với lớp cuối dùng sẽ sử dụng dropout bằng 0.3.

Về quá trình huấn luyện, mô hình được tối ưu hóa bằng thuật toán AdamW với tốc độ học ban đầu là 0.001 và suy giảm trọng số 0.0005. Một lập lịch tốc độ học dạng ReduceLROnPlateau được áp dụng để điều chỉnh tốc độ học dựa trên giá trị mất mát trên tập validation, với hệ số giảm 0.1 và độ kiên nhẫn là 5 epoch. Hàm mất mát tổng hợp được tính bằng cách kết hợp hai thành phần: mất mát phân loại với trọng số 0.9 và mất mát liên tục với trọng số 0.1, nhằm cân bằng giữa hai nhiệm vụ.

Kiến trúc tận dụng các ưu điểm của PNA, đặc biệt là khả năng kết hợp đa dạng các hàm tổng hợp và chuẩn hóa, để học được các biểu diễn đồ thị phong phú và hiệu quả. Việc tích hợp GRU và Set2Set giúp mô hình không chỉ nắm bắt được các đặc trưng cục bộ từ láng giềng mà còn tổng hợp thông tin toàn cục của đồ thị một cách chính xác, phù hợp với các bài toán đa nhiệm phức tạp trên dữ liệu đồ thị.

3.4. Hàm loss

Trong học máy hàm loss đóng một vai trò quan trọng trong việc đánh giá hiệu suất của mô hình. Hàm loss có chức năng đánh giá mức độ sai lệch của giá trị dự đoán của mô hình và giá trị thức tế của dữ liệu huấn luyện. Trong nghiên cứu việc sử dụng bộ dữ liệu có dự đoán đa nhãn một đối tượng được gán nhiều nhãn cảm xúc đã được phân tích trên phần mô

hình. Trong bài nghiên cứu đã sử dụng lại hàm loss theo bài [48]. Trong bài nghiên cứu đã sử dụng hai hàm mất mát riêng biệt tương ứng với hai loại nhãn trên.

Đối với nhãn cảm xúc rời rạc đa nhãn, tác giả sử dụng một hàm mất mát tự thiết kế mang tên DiscreteLoss, được xây dựng dựa trên mean squared error (MSE) có trọng số. Hàm này đo lường độ lệch giữa dự đoán và nhãn cảm xúc thực tế theo từng lớp cảm xúc:

$$L_{discrete} = \sum_{i=1}^{26} w_1 . (p_i - y_i)^2 \qquad (5)$$

Trong đó:

p_i: xác suất dự đoán cho cảm xúc thứ ii,

y_i: nhãn thực tế cho cảm xúc thứ ii,

w_i: trọng số tương ứng với cảm xúc thứ ii, giúp điều chỉnh độ quan trọng giữa các lớp.

Công thức tính trọng số động:

$$w_i = \begin{cases} \frac{1}{\log(s_i+1.2)}, & Nếu\ s_i > 0 \\ 0.0001, & Nếu\ s_i = 0 \end{cases} \qquad (6)$$

với s_i là tổng số lần cảm xúc i xuất hiện trong batch hiện tại.

Chiến lược trọng số này giúp mô hình học cân bằng hơn, giúp giải quyết vấn đề không cân bằng nhãn, thường gặp trong các bài toán đồ thị hoặc giúp tránh mất cảm với nhãn hiếm trong bài toán cảm xúc rời rạc. Trong bài nghiên cứu đã sử dụng dynamic để tiến hành huấn luyện cho tập dữ liệu.

CHƯƠNG 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ

4.1. Môi trường thực nghiệm

Về thông tin máy tính chạy thực nghiệm:

- Hệ điều hành: Window 10 – 64 bit.
- Bộ vi xử lý: Intel(R) Core (TM) i5-1035G4 CPU @ 1.10GHz.
- Bộ nhớ RAM: 8.0 GB.

Về ngôn ngữ lập trình:

- Sử dụng ngôn ngữ l ập trình Python 3.10 cùng với các gói thư viện SpaCy, Tranformers, PyTorch, …

4.2. Triển khai mô hình

Dữ liệu đầu vào của mô hình là đồ thị tri thức. Trong đó, dữ liệu được chia 70% Training, 10% Validation và Testing 20% để huấn luyện mô hình. Sau đó những tập này được đưa và hàm tao đồ thị tri thức rồi được nhúng vào GloVe200 để trích xuất đặc trưng. Cuối cùng được đưa vào mô hình PNA có thiết lập như đã trình bày ở phần 3.3.

4.3. Thang đo mAP

mAP (Mean Average Precision) là một chỉ số phổ biến trong các bài toán đánh giá mô hình machine learning, đặc biệt được sử dụng rộng rãi trong các bài toán nhận diện, phân loại đa nhãn (multi-label classification), và truy xuất thông tin (information retrieval).

Trong bối cảnh đồ thị tri thức, đặc biệt là phân tích cảm xúc hoặc nhận diện các đặc trưng phức tạp từ hình ảnh, mAP (mean Average Precision) thường được áp dụng để đánh giá khả năng nhận diện đúng các nút hoặc các mối liên kết trong đồ thị.

mAP (mean Average Precision) là thang đo trung bình của giá trị Average Precision (AP) qua tất cả các lớp hoặc tất cả các truy vấn.

Công thức tính mAP:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \qquad (9)$$

Trong đó:

N là số lượng lớp

AP_i là giá trị precision trung bình của lớp thứ i

AP là trung bình trọng số precision qua các mức recall khác nhau, được sử dụng khi xét khả năng xếp hạng các kết quả dự đoán. Nó đo lường khả năng mô hình xếp hạng kết quả phù hợp lên trước (đánh giá cả độ chính xác và khả năng gọi ra kết quả đúng).

AP thường được tính dựa trên đường cong Precision-Recall.

Công thức tính AP:

$$AP = \frac{\sum_{k=1}^{k} P(k)\times rel(k)}{Số\ lượng\ kết\ quả\ đúng} \qquad (10)$$

Hoặc là

$$AP = \sum_n (R_n - R_{n-1}) \times P_n \qquad (11)$$

P_n: Độ chính xác (Precision) tại điểm n

R_n: Độ thu hồi (Recall) tại điểm n

Công thức độ chính xác:

Là tỉ lệ mẫu dự đoán chính xác trong tổng số các dự đoán mô hình đưa ra.

$$Precision = \frac{TP}{TP+FP} \qquad (12)$$

TP: Dự đoán đúng (true positive).

FP: Dự đoán sai nhưng được phân loại vào lớp này (false positive).

4.4. Đánh giá kết quả thực nghiệm

Khi đã huấn luyện mô hình, kết quả thực nghiệm được trình bày ở Bảng 4.1. là kết quả Average Precision (AP) của từng lớp cảm xúc khi sử dụng hai mô hình học sâu dựa trên đồ thị là PNA, GIN và GraphSAGE.

Bảng 4.1. Kết quả AP của từng lớp của mô hình PNA, GIN và GraphSAGE

Emotion	PNA	GIN	GraphSAGE
Affection	37.13	31.82	34.38
Anger	13.55	9.77	11.89
Annoyance	12.80	11.88	12.34
Anticipation	57.59	59.04	57.01
Aversion	7.06	9.87	6.42
Confidence	76.69	81.31	75.78
Disapproval	11.68	11.56	11.99
Disconnection	26.18	26.55	25.68
Disquietment	17.44	19.19	17.43
Doubt/Confusion	19.40	18.70	18.47
Embarrassment	2.87	2.90	2.76
Engagement	85.99	84.57	85.05
Esteem	16.92	13.45	16.08

Emotion	PNA	GIN	GraphSAGE
Excitement	70.07	73.83	69.03
Fatigue	13.73	14.51	12.51
Fear	7.98	8.74	8.14
Happiness	66.15	61.80	64.20
Pain	11.43	7.03	10.20
Peace	24.31	25.14	23.31
Pleasure	43.94	42.16	42.23
Sadness	29.14	23.74	26.91
Sensitivity	7.35	7.96	6.81
Suffering	21.79	23.18	19.91
Surprise	8.00	8.47	7.68
Sympathy	15.30	13.18	15.05
Yearning	10.47	9.30	9.16
mAP	**27.50**	**26.91**	**26.55**

Trong quá trình thực nghiệm nhận diện cảm xúc dựa trên bộ dữ liệu EMOTIC, ba mô hình được triển khai là PNA (Principal Neighbourhood Aggregation), GIN (Graph Isomorphism Network) và GraphSAGE, với mAP trung bình lần lượt là 27.50, 26.91 và 26.55. Kết quả này cho thấy PNA có hiệu suất tổng thể nhỉnh hơn cả GIN và GraphSAGE, mặc dù sự chênh lệch giữa các mô hình không quá lớn. Phân tích chi tiết bảng kết quả cho thấy mỗi mô hình có ưu thế riêng ở các cảm xúc khác nhau, đồng thời cũng bộc lộ những hạn chế nhất định trong việc nhận diện một số cảm xúc phức tạp.

So với GIN và GraphSAGE, PNA đạt hiệu suất cao hơn ở một số cảm xúc quan trọng. Cụ thể, PNA vượt trội ở các cảm xúc như Affection (37.13 so với 31.82 và 34.38), Anger (13.55 so với 9.77 và 11.89), Happiness (66.15 so với 61.80 và 64.20), Sadness (29.14 so với 23.74 và 26.91), Pain (11.43 so với 7.03 và 10.20), và Pleasure (43.94 so với 42.16 và

42.23). Những cảm xúc này bao gồm cả các trạng thái tích cực (như Happiness, Pleasure) và tiêu cực (như Anger, Sadness, Pain), cho thấy PNA có khả năng học được các đặc trưng đa dạng hơn từ dữ liệu hình ảnh và ngữ cảnh. Đặc biệt, sự cải thiện rõ rệt ở Happiness và Sadness, vốn là hai cảm xúc phổ biến và có biểu hiện trực quan rõ ràng, chứng minh rằng PNA tận dụng tốt các đặc điểm thị giác như biểu cảm khuôn mặt và tư thế cơ thể. Ngoài ra, PNA cũng vượt trội ở Sympathy (15.30 so với 13.18 và 15.05) và Yearning (10.47 so với 9.30 và 9.16), vốn là các cảm xúc mang tính xã hội và phụ thuộc nhiều vào ngữ cảnh.

Trong khi đó, GIN đạt hiệu suất cao hơn PNA và GraphSAGE ở một số cảm xúc như Confidence (81.31 so với 76.69 và 75.78), Anticipation (59.04 so với 57.59 và 57.01), Excitement (73.83 so với 70.07 và 69.03), Disquietment (19.19 so với 17.44 và 17.43), và Aversion (9.87 so với 7.06 và 6.42). Những cảm xúc này thường liên quan đến các trạng thái tinh thần phức tạp hoặc biểu hiện tinh tế, cho thấy GIN có khả năng xử lý tốt hơn các đặc trưng cấp cao trong dữ liệu đồ thị, đặc biệt khi tận dụng thông tin từ các mối quan hệ ngữ cảnh. Ví dụ, Confidence và Excitement có thể được nhận diện tốt hơn thông qua các đặc trưng tổng hợp từ tư thế, ánh mắt, và môi trường xung quanh, điều mà GIN tối ưu hóa hiệu quả hơn nhờ cơ chế đồng hình đồ thị.

Về phía GraphSAGE, hiệu suất của mô hình này nhìn chung thấp hơn một chút so với PNA và GIN trên hầu hết các cảm xúc. Tuy nhiên, GraphSAGE vẫn duy trì kết quả khá cạnh tranh ở các cảm xúc như Engagement (85.05), Happiness (64.20) và Pleasure (42.23). Điều này cho thấy GraphSAGE vẫn có khả năng khai thác hiệu quả các đặc trưng khu vực cục bộ trong đồ thị, nhờ cơ chế trích xuất đặc trưng dựa trên quá trình lấy mẫu và tổng hợp thông tin từ hàng xóm. Tuy nhiên, sự chênh lệch nhỏ về mAP trung bình (26.55) cũng phản ánh hạn chế của GraphSAGE trong việc xử lý các mối quan hệ ngữ cảnh phức tạp hoặc những cảm xúc đòi hỏi sự tích hợp thông tin đa chiều hơn.

Trong cả ba mô hình, cảm xúc có mAP cao nhất là Engagement, với PNA đạt 85.99, GIN đạt 84.57 và GraphSAGE đạt 85.05. Đây là một kết quả đáng chú ý, vì Engagement

phản ánh sự tham gia tích cực của đối tượng trong các hoạt động hoặc tương tác xã hội, thường được thể hiện qua các đặc trưng rõ ràng như cử chỉ năng động, ánh mắt tập trung hoặc ngữ cảnh nhóm. Sự tương đồng về hiệu suất cao ở cả ba mô hình cho thấy Engagement là một cảm xúc dễ nhận diện, có thể do bộ dữ liệu EMOTIC chứa nhiều mẫu hình ảnh với các đặc điểm thị giác nổi bật liên quan đến cảm xúc này.

Ngược lại, cảm xúc có mAP thấp nhất ở cả ba mô hình là Embarrassment, với PNA đạt 2.87, GIN đạt 2.90 và GraphSAGE đạt 2.76. Các cảm xúc khác thuộc nhóm thấp (1-10%) bao gồm Disapproval, Fear, Sensitivity, Surprise và Aversion. Những cảm xúc này thường mang tính chủ quan cao, phụ thuộc vào các yếu tố ngữ cảnh phức tạp hoặc biểu hiện khuôn mặt tinh tế, khiến việc nhận diện trở nên khó khăn cho cả ba mô hình.

Tổng thể, PNA đạt mAP trung bình cao nhất nhờ vào cơ chế tổng hợp hàng xóm chính (Principal Neighbourhood Aggregation), cho phép mô hình cân bằng tốt giữa các đặc trưng cục bộ và toàn cục trong dữ liệu đồ thị. Điều này đặc biệt hữu ích trong việc nhận diện các cảm xúc như Happiness, Sadness và Sympathy. Trong khi đó, GIN nổi bật ở các cảm xúc đòi hỏi xử lý đặc trưng cấp cao như Confidence và Excitement, còn GraphSAGE cho thấy khả năng khai thác ổn định nhưng chưa vượt trội khi so sánh với hai mô hình còn lại. Từ đó, có thể khẳng định rằng việc sử dụng mô hình PNA trong bài toán xây dựng đồ thị tri thức cho phân tích cảm xúc là hoàn toàn phù hợp, đặc biệt khi mô hình thể hiện ưu thế rõ rệt trong việc khai thác các mối quan hệ ngữ cảnh giữa các đặc trưng cảm xúc. Bên cạnh đó, khi so sánh với các phương pháp của các nhóm khác, cụ thể là so sánh độ chính xác trung bình (AP) theo từng lớp cảm xúc giữa mô hình đề xuất và ba phương pháp đã được công bố trước đó gồm Kosti [33], Lee [14] và Chen [52]. Kết quả thu được là Bảng 4.2. Bảng so sánh kết quả của bài nghiên cứu so với các bài khác mà có sử dụng bộ dữ liệu là EMOTIC

Bảng 4.2. So sánh phương pháp đề xuất với các phương pháp khác

Method / Emotion	Kosti [33]	Lee [14]	Chen [52]	Our
Anticipation	9.49	12.88	11.67	**13.55**
Aversion	14.06	14.42	**16.56**	12.80
Confidence	58.64	52.85	**67.26**	57.59
Disapproval	7.48	3.26	**9.5**	7.06
Disconnection	78.35	72.68	**84.68**	76.69
Disquietment	14.97	**15.37**	15.32	11.68
Affection	27.85	22.36	**41.89**	37.13
Anger	9.49	12.88	11.67	**13.55**
Annoyance	14.06	14.42	**16.56**	12.80
Anticipation	58.64	52.85	**67.26**	57.59
Aversion	7.48	3.26	**9.5**	7.06
Confidence	78.35	72.68	**84.68**	76.69
Disapproval	14.97	**15.37**	15.32	11.68
Disconnection	21.32	22.01	**38.53**	26.18
Disquietment	16.89	10.84	**22.14**	17.44
Doubt/Confusion	**29.63**	26.07	25.26	19.40
Embarrassment	3.18	1.88	**4.60**	2.87
Engagement	87.53	73.71	**90.12**	85.99
Esteem	17.73	15.38	**24.79**	16.92
Excitement	77.16	70.42	**78.95**	70.07
Fatigue	9.7	6.29	**15.74**	13.73
Fear	**14.14**	7.47	8.76	7.98

Method / Emotion	Kosti [33]	Lee [14]	Chen [52]	Our
Happiness	58.26	53.73	**74.13**	66.15
Pain	8.94	8.16	8.58	**11.43**
Peace	21.56	19.55	**32.98**	24.31
Pleasure	45.46	34.12	**52.5**	43.94
Sadness	19.66	17.75	18.3	**29.14**
Sensitivity	**9.28**	6.94	9.9	7.35
Suffering	18.84	14.85	17.91	**21.79**
Surprise	**18.81**	17.46	12.54	8.00
Sympathy	14.71	14.89	**20.3**	15.30
Yearing	8.34	4.84	**12.52**	10.47
Mean	**27.38**	**23.85**	**31.36**	**27.5**

Có thể thấy rằng mô hình đề xuất đạt mAP là 27.50, cao hơn so với Kosti (27.38) và Lee (23.85), tuy vẫn còn thấp hơn so với Chen (34.27). Cụ thể, so với Kosti phương pháp của nghiên cứu cải thiện đáng kể ở một số cảm xúc như Anger (18.00 so với 9.49), Annoyance (37.13 so với 27.85), Happiness (66.15 so với 58.26), Sadness (29.14 so với 19.66), và Suffering (21.79 so với 18.84). Những cải tiến này cho thấy khả năng nhận diện tốt hơn các cảm xúc tiêu cực và phức tạp, vốn thường khó phân biệt do sự tinh tế trong biểu hiện khuôn mặt hoặc ngữ cảnh. So với Lee, phương pháp của nghiên cứu vượt trội ở các cảm xúc như Affection (35.69 so với 19.9), Excitement (70.07 so với 70.42), Pleasure (43.94 so với 34.12), và Happiness (66.15 so với 53.73), chứng minh khả năng xử lý tốt hơn các cảm xúc tích cực và liên quan đến tương tác xã hội. Tuy nhiên, khi so sánh với Chen, phương pháp đề xuất vẫn chưa đạt hiệu suất cao ở một số cảm xúc như Disconnection (76.69 so với 84.68) và Happiness (66.15 so với 74.13), có thể do sự khác biệt trong cách xây dựng mô hình hoặc dữ liệu huấn luyện.

Trong số 26 cảm xúc được đánh giá trên phương pháp được đề xuất, Excitement đạt mAP cao nhất với 70.07, tiếp theo là Happiness (66.15) và Confidence (57.59). Những cảm xúc này có hiệu suất cao có thể do đặc trưng rõ ràng trong dữ liệu hình ảnh, chẳng hạn như nụ cười, ánh mắt tích cực, hoặc ngữ cảnh xã hội vui vẻ, giúp mô hình dễ dàng nhận diện. Ngược lại, nhóm cảm xúc có mAP thấp nhất bao gồm Esteem (2.87), Disapproval (7.06), Fear (7.98), Sensitivity (7.35), và Surprise (8.00), đều nằm trong khoảng 1-10%. Các cảm xúc này thường phức tạp, mang tính chủ quan cao, hoặc có ít dữ liệu huấn luyện chất lượng, dẫn đến khó khăn trong việc phân loại chính xác. Khi phân loại theo mức độ hiệu suất, không có cảm xúc nào tcủa bài nghiên cứu đạt mAP trên 80%, nhưng các cảm xúc như Excitement, Happiness, Confidence, và Disconnection (76.69) thuộc nhóm hiệu suất cao (>50%). Trong khi đó, nhóm thấp (1-10%) đã được liệt kê ở trên, và nhóm trung bình (10-50%) bao gồm các cảm xúc như Affection (35.69), Annoyance (37.13), Pleasure (43.94), và Sadness (29.14).

Lý do các cảm xúc như Excitement và Happiness đạt mAP cao xuất phát từ việc bộ dữ liệu EMOTIC chứa nhiều mẫu hình ảnh với các biểu hiện rõ ràng chẳng hạn như nụ cười, cử chỉ tích cực, hoặc ngữ cảnh xã hội cũng như tiệc tùng, lễ hội. Những đặc trưng này dễ được các mô hình học sâu nhận diện thông qua các đặc điểm thị giác như đường nét khuôn mặt hoặc màu sắc hình ảnh. Ngược lại các cảm xúc như Esteem, Disapproval, và Sensitivity thường mang tính trừu tượng, phụ thuộc vào ngữ cảnh phức tạp hoặc biểu hiện tinh tế, khó phân biệt chỉ dựa trên hình ảnh. Ngoài ra, sự thiếu hụt dữ liệu huấn luyện cho các cảm xúc này có thể làm giảm hiệu suất của mô hình. Ví dụ, Esteem có thể yêu cầu phân tích các yếu tố văn hoá hoặc xã hội cụ thể. Trong khi Suprise có thể bị nhầm lẫn với các cảm xúc khác như là Fear do sự tương đồng trong biểu hiện khuôn mặt.

Phân tích cụ thể theo từng lớp cảm xúc, mô hình thực nghiệm ghi nhận các kết quả như sau: Với Anticipation, AP của bài là (13.55) cao nhất, vượt Lee (12.88). Kosti (9.49) thấp nhất. Đối với Aversion, hiệu suất là 12.80, thấp hơn Kosti (14.06), Lee (14.42), và Chen (16.56). Với Confidence, Chen (67.26) cao nhất, Lee (52.85) thấp nhất. Và bài nghiên cứu

có kết quả (57.59) gần Kosti nhưng thấp hơn Chen. Ở Disapproval, kết quả Chen (9.5) cao nhất, Lee (3.26) thấp nhất. Kết quả của bài (7.06) gần Kosti.

Với Disconnection, mô hình có thứ hạng giống với Disapproval. Đối với Disquietment, hiệu suất của Lee (15.37) cao nhất, còn bài nghiên cứu (11.68) là thấp nhất. Với Affection, mô hình đạt 37.13, cao hơn Kosti (27.85) và Lee (22.36) nhưng thấp hơn Chen (41.89). Ở Anger, hiệu suất là 13.55, cao hơn Kosti (9.49), Lee (12.88), và Chen (11.67). Với Annoyance, kết quả là 12.80, thấp hơn Kosti (14.06), Lee (14.42), và Chen (16.56). Đối với Doubt/Confusion, mô hình đạt 19.40, thấp hơn Kosti (29.63), Lee (26.07), và Chen (25.26). Với Embarrassment, hiệu suất là 2.87, thấp hơn Kosti (3.18), Lee (1.88), và Chen (4.60). Ở Engagement, mô hình đạt 85.99, cao hơn Kosti (87.53) và Lee (73.71) nhưng thấp hơn Chen (90.12). Với Esteem, kết quả là 16.9, thấp hơn Kosti (17.73), Lee (15.38), và Chen (24.79). Đối với Excitement, mô hình đạt 70.07, thấp hơn Kosti (77.16), Lee (70.42), và Chen (78.95). Với Fatigue, hiệu suất là 13.73, cao hơn Kosti (9.7) và Lee (6.29) nhưng thấp hơn Chen (15.74). Ở Fear, mô hình đạt 7.98, thấp hơn Kosti (14.14), Lee (7.47), và Chen (8.76). Với Happiness, kết quả là 66.15, cao hơn Kosti (58.26) và Lee (53.73) nhưng thấp hơn Chen (74.13). Đối với Pain, mô hình đạt 11.43, cao hơn Kosti (8.94), Lee (8.16), và Chen (8.58). Với Peace, hiệu suất là 24.31, cao hơn Kosti (21.56ss) và Lee (19.55) nhưng thấp hơn Chen (32.98). Ở Pleasure, mô hình đạt 43.94, thấp hơn Kosti (45.46) và Chen (52.5) nhưng cao hơn Lee (34.12). Với Sadness, kết quả là 29.14, cao hơn Kosti (19.66), Lee (17.75), và Chen (18.3). Đối với Sensitivity, mô hình đạt 7.35, thấp hơn Kosti (9.28), Lee (6.94), và Chen (9.9). Với Suffering, hiệu suất là 21.79, cao hơn Kosti (18.84), Lee (14.85), và Chen (17.91). Ở Surprise, mô hình đạt 8.00, thấp hơn Kosti (18.81), Lee (17.46), và Chen (12.54). Với Sympathy, kết quả là 15.30, cao hơn Kosti (14.71) và Lee (14.89) nhưng thấp hơn Chen (20.3). Cuối cùng, với Yearning, mô hình đạt 10.47, cao hơn Kosti (8.34) và Lee (4.84) nhưng thấp hơn Chen (12.52). Tổng thể, mô hình của bài vượt trội ở các cảm xúc như Anger, Pain, Sadness, và Suffering, nhưng cần cải thiện ở các cảm xúc

như Aversion, Doubt/Confusion, Fear, Surprise, và Sensitivity để đạt hiệu suất cạnh tranh hơn so với Chen.

Phương pháp đề xuất đạt được sự cải thiện trong nhận diện cảm xúc nhờ vào việc sử dụng mô hình dựa trên mạng nơ-ron tích chập. Ngoài ra, việc sử dụng nhúng GloVe để mã hóa thông tin ngữ nghĩa và xây dựng đồ thị tri thức ngữ cảnh đã hỗ trợ mô hình hiểu rõ hơn về mối quan hệ giữa các cảm xúc và ngữ cảnh. Những cải tiến này cho phép mô hình học được các đặc trưng phức tạp hơn, đặc biệt là với các cảm xúc có sự khác biệt tinh tế, từ đó nâng cao độ chính xác so với các phương pháp của Kosti và Lee. Tuy nhiên, để vượt qua phương pháp của Chen, bài nghiên cứu cần tiếp tục tối ưu hóa mô hình, chẳng hạn như tăng số lớp mạng, sử dụng các hàm tổng hợp tiên tiến hơn (PNA).

Tóm lại, phương pháp đề xuất của không chỉ cao hơn một số các nghiên cứu cơ sở mà còn mang lại những cải tiến đáng kể ở nhiều cảm xúc. Những kết quả này khẳng định tiềm năng của phương pháp trong việc ứng dụng thực tế, chẳng hạn như phân tích cảm xúc trong tương tác xã hội hoặc hỗ trợ trí tuệ nhân tạo trong các hệ thống giao tiếp thông minh. Điều này cho thấy sự cải tiến đáng kể trong cách xây dựng và khai thác thông tin từ đồ thị tri thức, đồng thời mở ra hướng nghiên cứu tiềm năng trong việc kết hợp giữa biểu cảm hình ảnh và đặc trưng ngữ cảnh để nâng cao chất lượng phân loại cảm xúc.

4.5. Thực nghiệm mô hình

Sau khi xây dựng thành công đồ thị tri thức để biểu diễn các mối quan hệ giữa các trạng thái cảm xúc và các đặc điểm hình ảnh. Tiếp đó, tiến hành thực nghiệm mô hình phân loại cảm xúc trên hình ảnh của một vận động viên bóng chày trong tư thế chuẩn bị ném bóng. Nhãn thực tế được gán cho hình ảnh bao gồm: Anticipation, Confidence, Engagement và Excitement.

Labels: Anticipation, Confidence, Engagement, Excitement

Prediction: Anticipation, Confidence, Engagement, Excitement]

Hình 4.1: Kết quả dự đoán (1)

Kết quả dự đoán từ mô hình cũng trả về các nhãn: Anticipation, Confidence, Engagement và Excitement, trùng khớp hoàn toàn với nhãn thực tế. Độ chính xác cho thấy mô hình đã nhận diện đúng các trạng thái cảm xúc dựa trên các đặc điểm hình ảnh như tư thế, bối cảnh và môi trường. Tuy nhiên, để đánh giá khả năng tổng quát hóa của mô hình, cần tiến hành thêm các thử nghiệm trên tập dữ liệu đa dạng hơn với các điều kiện và bối cảnh khác nhau. Tiếp theo, hình ảnh thứ hai được đưa vào thực nghiệm là một vận động viên bóng chày trẻ trong tư thế chuẩn bị đánh bóng. Nhãn thực tế được gán cho hình ảnh này bao gồm: Anticipation, Confidence, Disquietment, Engagement, Excitement và Happiness. Kết quả dự đoán từ mô hình trả về các nhãn: Anticipation, Confidence, Engagement và Excitement. So sánh với nhãn thực tế, mô hình nhận diện chính xác bốn trạng thái cảm xúc, nhưng không phát hiện được Disquietment và Happiness.

Labels: Anticipation, Confidence, Disquietment, Engagement, Excitement, Happiness

Prediction: Anticipation, Confidence, Engagement, Excitement]

Hình 4.2: Kết quả dự đoán (2)

Kết quả này cho thấy mô hình có khả năng nhận diện một số trạng thái cảm xúc dựa trên tư thế và bối cản. Tuy nhiên với những với tấm hình có nhiều đối tượng như tấm hình này, thì mô hình vẫn dự đoán chưa được tốt lắm. Tiếp tục, hình ảnh thứ ba được đưa vào thực nghiệm mô tả một người đang cho hươu cao cổ ăn trong môi trường tự nhiên, với nhãn thực tế được gán bao gồm: Affection, Anticipation, Confidence, Engagement, Sensitivity và Sympathy. Kết quả dự đoán từ mô hình trả về các nhãn: Affection, Peace và Sensitivity.

Labels: Affection, Sympathy Anticipation, Confidence, Engagement, Sensitivity,

Prediction: Affection, Peace, Sensitivity

Hình 4.3: Kết quả dự đoán (3)

So sánh với nhãn thực tế, mô hình chỉ nhận diện chính xác hai trạng thái cảm xúc (Affection và Sensitivity) và dự đoán thêm Peace, một trạng thái không có trong nhãn thực tế, đồng thời bỏ sót Anticipation, Confidence, Engagement và Sympathy.

Labels: Anticipation, Confidence, Disquietment Engagement, Excitement

Prediction: Pain, Fear, Doubt/confusion, Confidence, Anticipation, Excitement, Sadness, Yearning, Fatigue

Hình 4.4: Kết quả dự đoán (4)

Với tấm ảnh một đám đông tụ tập trước một tòa nhà, với nhãn thực tế được gán bao gồm: Anticipation, Confidence, Disquietment, Engagement và Excitement. Kết quả dự đoán từ mô hình trả về các nhãn: Pain, Fear, Doubt/Confusion, Confidence, Anticipation, Excitement, Sadness, Yearning và Fatigue. So sánh với nhãn thực tế, mô hình nhận diện chính xác ba trạng thái cảm xúc (Anticipation, Confidence, Excitement), nhưng bỏ sót Disquietment và Engagement, đồng thời dự đoán thêm nhiều trạng thái không có trong nhãn thực tế như Pain, Fear, Doubt/Confusion, Sadness, Yearning và Fatigue. Kết quả này cho thấy mô hình có xu hướng nhận diện quá mức các trạng thái cảm xúc tiêu cực trong bối cảnh đám đông, có thể do sự phức tạp của cảnh và số lượng người lớn.

Như vậy, quá trình thực nghiệm mô hình phân loại cảm xúc dựa trên đồ thị tri thức cho thấy mô hình có ưu điểm trong việc nhận diện chính xác một số trạng thái cảm xúc cơ bản, đặc biệt trong các bối cảnh đơn giản. Tuy nhiên, mô hình cũng bộc lộ nhược điểm khi gặp khó khăn trong việc nhận biết đầy đủ các trạng thái cảm xúc phức tạp, thường bỏ sót một số nhãn quan trọng và có xu hướng dự đoán thêm nhiều trạng thái không có trong nhãn thực

tế, đặc biệt trong các tình huống đông người hoặc đa dạng về ngữ cảnh. Để cải thiện, cần tập trung mở rộng tập dữ liệu huấn luyện và tinh chỉnh mô hình nhằm tăng khả năng tổng quát hóa và giảm tỷ lệ dự đoán sai lệch.

CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1. Kết luận

Nghiên cứu này của em đã tập trung vào phát triển, cải thiện một đồ thị tri thức cho mô hình nhận diện cảm xúc dựa trên ngữ cảnh. Kết quả của thử nghiệm này đã rõ ràng chỉ ra rằng việc mà sử dụng đồ thị tri thức để phân tích tình cảm là một phương pháp có thể dùng để đánh giá cảm xúc dựa trên ngữ cảnh.

Tuy nhiên, trong một bức hình lại có nhiều đặc trưng thể hiện nhiều khía cạnh có thể nhận diện được cảm xúc ví dụ như có quá nhiều đối tượng trong hình, các mối tương tác của các đối tượng với nhau, … Ngoài ra, việc các đối tượng che khuất lẫn nhau các góc chụp, ánh sáng cũng làm cho sai lệch qua cảm xúc khác. Qua đó, việc sử dụng đồ thị tri thức với một nhánh khác như khuôn mặt, dáng đứng, … có thể cải thiện được việc phân tích cảm xúc từ hình ảnh. Do đó, Việc thêm nhánh đặc trưng cho mô hình là điều cần thiết.

5.2. Hướng phát triển

Tuy mô hình đã đạt được kết quả khả quan so với mô hình gốc của tác giả tập dữ liệu EMOTIC, nhưng đồ thị vẫn còn nhiều hạn chế. Vì thế việc phát triển và tìm kiếm những phương pháp nhằm ứng dụng cho việc giúp nâng cao tính ngữ nghĩa và liên kết cần được đầu tư. Trong tương lai, việc phát triển của mô hình trí tuệ nhân tạo ứng dụng vào thực tế là điều cần thiết đáp ứng nhu cầu của con người. Bên cạnh đó, việc thêm các nhánh khác (khuôn mặt, hình dáng, …) để tăng thêm các đặc trưng vào mô hình. Ngoài ra, sử dụng các mô hình ngôn ngữ lớn kết hợp mô hình Vision-Language Models (VLMs) nhằm phân tích các ngữ nghĩa trong ảnh và so khớp với đặc trưng ảnh.

DANH MỤC TÀI LIỆU THAM KHẢO

[1] Azher Uddin, A. Uddin, Joolekha Bibi Joolee, J. B. Joolee, Young Koo Lee, and Y.-K. Lee, "Depression Level Prediction Using Deep Spatiotemporal Features and Multilayer Bi-LTSM," *IEEE Trans. Affect. Comput.*, no. 1, pp. 1–1, Jan. 2020, doi: 10.1109/taffc.2020.2970418.

[2] Sai Manvitha Enadula, Sai Manvitha Enadula, Akshith Sriram Enadula, Akshith Sriram Enadula, Rama Devi Burri, and Rama Devi Burri, "Recognition of Student Emotions in an Online Education System," *IEEE Int. Conf. Electr. Comput. Commun. Technol.*, Sep. 2021, doi: 10.1109/icecct52121.2021.9616788.

[3] Alan Cowen, A. S. Cowen, Dacher Keltner, and D. Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 114, no. 38, p. 201702247, Sep. 2017, doi: 10.1073/pnas.1702247114.

[4] Jie Zhou *et al.*, "SK-GCN: Modeling Syntax and Knowledge via Graph Convolutional Network for aspect-level sentiment classification," *Knowl. Based Syst.*, vol. 205, p. 106292, Oct. 2020, doi: 10.1016/j.knosys.2020.106292.

[5] Zhengyuan Yang *et al.*, "Pose-based Body Language Recognition for Emotion and Psychiatric Symptom Interpretation," *ArXiv Comput. Vis. Pattern Recognit.*, 2020, doi: 10.1109/icpr48806.2021.9412591.

[6] Lisa Feldman Barrett, L. F. Barrett, Batja Mesquita, B. Mesquita, Maria Gendron, and M. Gendron, "Context in Emotion Perception," *Curr. Dir. Psychol. Sci.*, vol. 20, no. 5, pp. 286–290, Oct. 2011, doi: 10.1177/0963721411422522.

[7] Sidney K. D'Mello, S. K. D'Mello, Art Graesser, and A. C. Graesser, "Dynamics of affective states during complex learning," *Learn. Instr.*, vol. 22, no. 2, pp. 145–157, Apr. 2012, doi: 10.1016/j.learninstruc.2011.10.001.

[8] E. Prince, Katherine B. Martin, and D. Messinger, "Facial Action Coding System," 2015, doi: 10.4135/9781483381411.n178.

[9] Maja Pantić, M. Pantic, Léon J. M. Rothkrantz, and L. J. M. Rothkrantz, "Expert system for automatic analysis of facial expressions," *Image Vis. Comput.*, vol. 18, no. 11, pp. 881–905, Aug. 2000, doi: 10.1016/s0262-8856(00)00034-2.

[10] Zisheng Li, Z. Li, Jun-ichi Imai, J. Imai, Masahide Kaneko, and M. Kaneko, "Facial-component-based bag of words and PHOG descriptor for facial expression recognition," *IEEE Int. Conf. Syst. Man Cybern.*, pp. 1353–1358, Oct. 2009, doi: 10.1109/icsmc.2009.5346254.

[11] Paul Ekman, P. Ekman, Wallace V. Friesen, and W. V. Friesen, "Constants across cultures in the face and emotion.," *J. Pers. Soc. Psychol.*, vol. 17, no. 2, pp. 124–129, Feb. 1971, doi: 10.1037/h0030377.

[12] Hung Son Nguyen *et al.*, "A Thermal Facial Emotion Database and Its Analysis," *Pac.-Rim Symp. Image Video Technol.*, pp. 397–408, Oct. 2013, doi: 10.1007/978-3-642-53842-1_34.

[13] Ronak Kosti *et al.*, "Emotion Recognition in Context," *Comput. Vis. Pattern Recognit.*, pp. 1960–1968, Jul. 2017, doi: 10.1109/cvpr.2017.212.

[14] Trisha Mittal *et al.*, "EmotiCon: Context-Aware Multimodal Emotion Recognition Using Frege's Principle," *Comput. Vis. Pattern Recognit.*, pp. 14234–14243, Jun. 2020, doi: 10.1109/cvpr42600.2020.01424.

[15] Trisha Mittal, Puneet Mathur, Aniket Bera, and Dinesh Manocha, "Affect2MM: Affective Analysis of Multimedia Content Using Emotion Causality," *Comput. Vis. Pattern Recognit.*, 2021, doi: 10.1109/cvpr46437.2021.00561.

[16] Armin Seyeditabari *et al.*, "Emotion Detection in Text: Focusing on Latent Representation.," *ArXiv Comput. Lang.*, Jul. 2019.
[17] Zhaoxia Wang *et al.*, "SenticNet 6: Ensemble Application of Symbolic and Subsymbolic AI for Sentiment Analysis," *Int. Conf. Inf. Knowl. Manag.*, pp. 105–114, 2020, doi: 10.1145/3340531.3412003.
[18] Sadil Chamishka *et al.*, "A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling," *Multimed. Tools Appl.*, Jun. 2022, doi: 10.1007/s11042-022-13363-4.
[19] M. Shamim Hossain, M. S. Hossain, Ghulam Muhammad, and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," *Inf. Fusion*, vol. 49, pp. 69–78, Sep. 2019, doi: 10.1016/j.inffus.2018.09.008.
[20] Harshit Varma, Harshit Varma, Nagarajan Ganapathy, Nagarajan Ganapathy, Thomas M. Deserno, and Thomas M. Deserno, "Video-based driver emotion recognition using hybrid deep spatio-temporal feature learning," *Med. Imaging*, Apr. 2022, doi: 10.1117/12.2613118.
[21] Bingdian Yang, Qian Zhang, and Zhichao Liu, "ICANet: A Method of Short Video Emotion Recognition Driven by Multimodal Data," *2022 2nd Int. Conf. Netw. Syst. AI INSAI*, Oct. 2022, doi: 10.1109/insai56792.2022.00014.
[22] Hillel Aviezer, H. Aviezer, Yaacov Trope, Y. Trope, Alexander Todorov, and A. Todorov, "Body cues, not facial expressions, discriminate between intense positive and negative emotions," *Science*, vol. 338, no. 6111, pp. 1225–1229, Nov. 2012, doi: 10.1126/science.1224313.
[23] James K. McNulty, J. K. McNulty, Frank D. Fincham, and F. D. Fincham, "Beyond positive psychology? Toward a contextual view of psychological processes and well-being.," *Am. Psychol.*, vol. 67, no. 2, pp. 101–110, Feb. 2012, doi: 10.1037/a0024572.
[24] Jiyoung Lee *et al.*, "Context-Aware Emotion Recognition Networks," *IEEE Int. Conf. Comput. Vis.*, pp. 10142–10151, Jan. 2019, doi: 10.1109/iccv.2019.01024.
[25] Yujie Wan, Yuzhong Chen, Jiali Lin, Jiayuan Zhong, and Chen Dong, "A knowledge-augmented heterogeneous graph convolutional network for aspect-level multimodal sentiment analysis," *Comput. Speech Lang.*, 2023, doi: 10.1016/j.csl.2023.101587.
[26] Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang, "Multimodal Sentiment Detection Based on Multi-channel Graph Neural Networks," *Annu. Meet. Assoc. Comput. Linguist.*, 2021, doi: 10.18653/v1/2021.acl-long.28.
[27] Jian Liao *et al.*, "Dynamic commonsense knowledge fused method for Chinese implicit sentiment analysis," *Inf. Process. Manag.*, vol. 59, no. 3, pp. 102934–102934, May 2022, doi: 10.1016/j.ipm.2022.102934.
[28] Mohannad AlMousa, M. AlMousa, Rachid Benlamri, R. Benlamri, Richard Khoury, and R. Khoury, "Exploiting non-taxonomic relations for measuring semantic similarity and relatedness in WordNet," *Knowl.-Based Syst.*, vol. 212, p. 106565, Jun. 2020, doi: 10.1016/j.knosys.2020.106565.
[29] Fabian M. Suchanek, F. M. Suchanek, Gjergji Kasneci, G. Kasneci, Gerhard Weikum, and G. Weikum, "YAGO: A Large Ontology from Wikipedia and WordNet," *J. Web Semant.*, vol. 6, no. 3, pp. 203–217, Sep. 2008, doi: 10.1016/j.websem.2008.06.001.
[30] Sixing Wu *et al.*, "Aspect-based sentiment analysis via fusing multiple sources of textual knowledge," *Knowl. Based Syst.*, vol. 183, p. 104868, Jul. 2019, doi: 10.1016/j.knosys.2019.104868.

[31] Minghui Zhang *et al.*, "Context-Aware Affective Graph Reasoning for Emotion Recognition," *IEEE Int. Conf. Multimed. Expo*, pp. 151–156, Jul. 2019, doi: 10.1109/icme.2019.00034.
[32] Manh-Hung Hoang *et al.*, "Context-Aware Emotion Recognition Based on Visual Relationship Detection," *IEEE Access*, vol. 9, pp. 90465–90474, Jan. 2021, doi: 10.1109/access.2021.3091169.
[33] Ronak Kosti *et al.*, "Context Based Emotion Recognition Using EMOTIC Dataset," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 11, pp. 2755–2766, Nov. 2020, doi: 10.1109/tpami.2019.2916866.
[34] Wilfredo Graterol *et al.*, "Emotion Detection for Social Robots Based on NLP Transformers and an Emotion Ontology.," *Sensors*, vol. 21, no. 4, p. 1322, 2021, doi: 10.3390/s21041322.
[35] Dingkang Yang *et al.*, "Emotion Recognition for Multiple Context Awareness," *Eur. Conf. Comput. Vis.*, pp. 144–162, Jan. 2022, doi: 10.1007/978-3-031-19836-6_9.
[36] M. Nickel *et al.*, "A Review of Relational Machine Learning for Knowledge Graphs," *ArXiv Mach. Learn.*, 2015, doi: 10.1109/jproc.2015.2483592.
[37] Tsung-Yi Lin *et al.*, "Microsoft COCO: Common Objects in Context," *ArXiv Comput. Vis. Pattern Recognit.*, May 2014, doi: 10.1007/978-3-319-10602-1_48.
[38] Bolei Zhou *et al.*, "Semantic Understanding of Scenes Through the ADE20K Dataset," *Int. J. Comput. Vis.*, vol. 127, no. 3, pp. 302–321, Mar. 2019, doi: 10.1007/s11263-018-1140-0.
[39] Junnan Li, Dongxu Li, Caiming Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," *Int. Conf. Mach. Learn.*, 2022.
[40] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee, "Improved Baselines with Visual Instruction Tuning," *Comput. Vis. Pattern Recognit.*, 2023, doi: 10.1109/cvpr52733.2024.02484.
[41] Jinze Bai *et al.*, "Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond," *arXiv.org*, Aug. 2023, doi: 10.48550/arxiv.2308.12966.
[42] J. E. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," *Int. Conf. Learn. Represent.*, 2021.
[43] J. Hu, R. Cavicchioli, and Alessandro Capotondi, "Exploiting Multiple Sequence Lengths in Fast End to End Training for Image Captioning," *BigData Congr. Serv. Soc.*, 2022, doi: 10.1109/bigdata59044.2023.10386812.
[44] Thomas Kipf, T. Kipf, Max Welling, and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," *ArXiv Learn.*, Sep. 2016.
[45] William L. Hamilton, Z. Ying, and J. Leskovec, "Inductive Representation Learning on Large Graphs," *Neural Inf. Process. Syst.*, 2017.
[46] Petar Veličković *et al.*, "Graph Attention Networks," *ArXiv Mach. Learn.*, Oct. 2017, doi: 10.17863/cam.48429.
[47] Keyulu Xu, Weihua Hu, J. Leskovec, and S. Jegelka, "How Powerful are Graph Neural Networks?," *Int. Conf. Learn. Represent.*, 2018.
[48] Karsten Borgwardt *et al.*, "Protein function prediction via graph kernels," *Bioinformatics*, vol. 21, no. 1, pp. 47–56, Jan. 2005, doi: 10.1093/bioinformatics/bti1007.

[49] Erik Cambria *et al.*, “SenticNet 5: Discovering Conceptual Primitives for Sentiment Analysis by Means of Context Embeddings.,” *AAAI Conf. Artif. Intell.*, pp. 1795–1802, Jan. 2018, doi: 10.1609/aaai.v32i1.11559.

[50] G. A. Miller, “WordNet : a lexical database for English : New horizons in commercial and industrial AI,” *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Jan. 1995.

[51] Gabriele Corso *et al.*, “Principal Neighbourhood Aggregation for Graph Nets,” *ArXiv Learn.*, Apr. 2020.

[52] Jing Chen *et al.*, “Incorporating structured emotion commonsense knowledge and interpersonal relation into context-aware emotion recognition,” *Appl. Intell.*, Jun. 2022, doi: 10.1007/s10489-022-03729-4.