VIEMO: ADVANCING EMOTION RECOGNITION IN VIETNAMESE LANGUAGE THROUGH MULTIMODAL LEARNING

Nhóm 9 – Giảng viên hướng dẫn: Nguyễn Thành Luân

Thành viên nhóm:

22520505 – Kiều Quí Hùng

22520752 – Nguyễn Duy Liêm

22521123 – Mạc Nguyên Phúc

TỔNG QUAN

- Multimodal Emotion Recognition - MER - là bài toán xác định và hiểu các trạng thái cảm xúc của con người bằng cách kết hợp các tín hiệu khác nhau.

- Trong báo cáo này, chúng tôi tập trung vào hai nhiệm vụ: (1) xây dựng bộ dữ liệu đa phương thức tiếng Việt, tích hợp văn bản, hình ảnh và âm thanh; (2) đánh giá hiệu quả của các phương pháp late fusion trên bộ dữ liệu đã xây dựng.

ĐẶT VẤN ĐỀ

First Problem

Việc chỉ dựa vào nội dung của văn bản có thể không phản ánh được chính xác trạng thái cảm xúc thực sự của người nói.

Second Problem

Theo khảo sát của chúng tôi thì hiện nay chưa có bài báo nào giải quyết vấn đề này trên tiếng Việt.

Third Problem

Tiếng Việt có cấu trúc đặc thù và ngữ nghĩa riêng biệt so với các ngôn ngữ khác.

CÁC NGHIÊN CỨU LIÊN QUAN

1. "Multimodal Emotion Recognition", Anatoli de Bradké, Maël Fabien, Raphaël Lederman, and Stéphane Reynal (2019)

2. “A systematic survey on multimodal emotion recognition using learning algorithms”, Naveed Ahmed, Zaher Al Aghbari, Shini Girija (2023)

3. “Emotion Recognition for Vietnamese Social Media Text”, Vong Ho, Duong Nguyen, Danh Nguyen, Linh Pham, Kiet Nguyen and Ngan Nguyen (2019)

GIỚI THIỆU BÀI TOÁN

Nhãn dữ liệu cho cảm xúc: 6 cảm xúc cơ bản theo Paul Ekman và I nhãn khác cho các trường hợp còn lại.

- happy
- sad
- neutral
- disgust
- surprise
- angry
- fear

GIỚI THIỆU BÀI TOÁN

Input: 1 điểm dữ liệu gồm 3 nguồn khác nhau, được lấy cùng một đoạn video ngắn chỉ gồm đoạn hội thoại của người nói cùng với nhãn đơn cảm xúc của từng modality. 3 modality gồm:

- **Text**
  Lời của người nói
- **Image**
  1 hình trong đoạn video của người nói
- **Audio**
  Âm điệu trong đoạn hội thoại

Output: 1 nhãn cảm xúc dự đoán cho đoạn hội thoại.

METHODOLOGY

Image Emotion Recognition

Áp dụng mạng CNN để trích xuất đặc trưng từ hình ảnh

Text Emotion Recognition

Chuẩn hóa bộ dữ liệu UIT-VSMEC để huấn luyện mô hình

Speech Emotion Recogniton

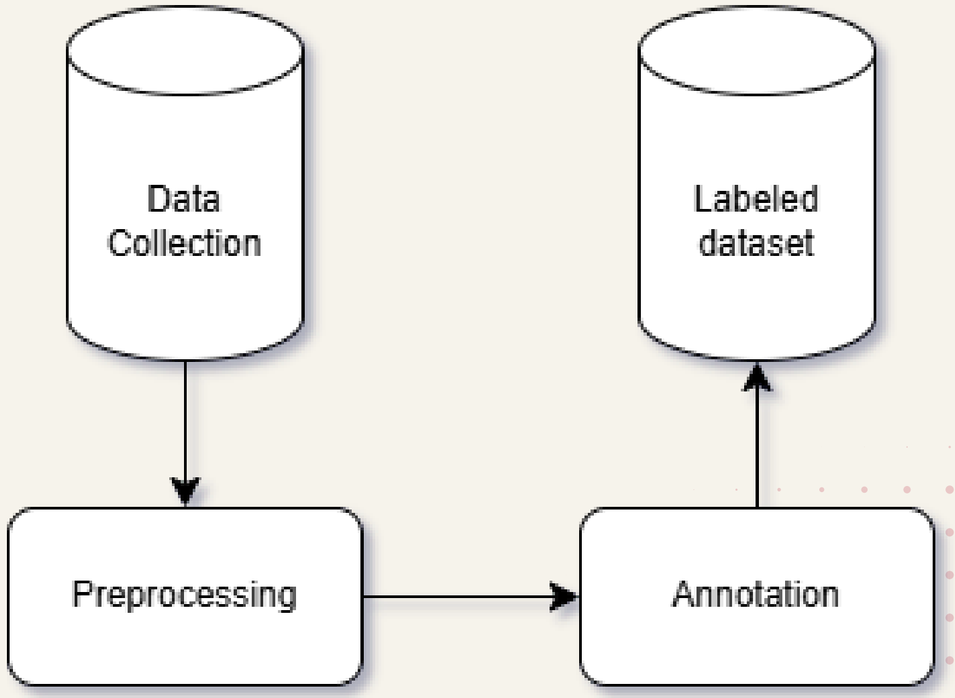
Trích xuất thông tin bằng việc sử dụng MFCC

Fusion Method

Sử dụng PROD fusion và weight average để kết hợp các dự đoán từ các modality khác nhau

THU THẬP DỮ LIỆU

- Dữ liệu
  - tự thu thập cho từng modality ứng với từng đoạn video
  - thu thập dữ liệu có transcript
- Preprocessing: xử lí, thống nhất text về I dạng.
- Annotation: huấn luyện annotator đến khi đạt độ đồng thuận Kappa nhất định.



BỘ DỮ LIỆU

Độ đồng thuận: 0.770

Bộ dữ liệu gồm 1093 dữ liệu gồm 8 thuộc tính cho từng điểm dữ liệu.

Tên thuộc tính	Nội dung
start_time	Thời điểm bắt đầu đoạn video
end_time	Thời điểm kết thúc đoạn video
Emotion_Text	Nhãn đơn cảm xúc cho lời nói
Text	Lời nói trong đoạn video
Emotion_Audio	Nhãn đơn cảm xúc cho audio
Audio	Đường dẫn đến audio của đoạn video
Emotion_Image	Nhãn đơn cảm xúc cho hình ảnh
Image	Đường dẫn đến hình ảnh của đoạn video

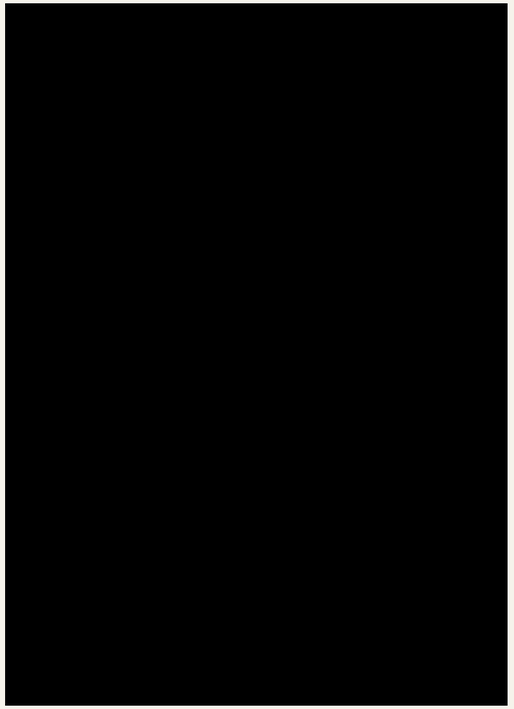
BỘ DỮ LIỆU

Số lượng nhãn mỗi cảm xúc của từng modal

	angry	disgust	fear	happy	neutral	surprise	sad
Emotion_Text	262	94	52	173	216	136	150
Emtion_Audio	233	172	73	77	233	107	188
Emotion_Image	219	82	60	203	159	102	257

BỘ DỮ LIỆU

Text:
Thì là lần đầu của em nên là em muốn có kỷ niệm.



Emotion_Text: happy

Emotion_Audio: happy

Emotion_Image: surprise

BỘ DỮ LIỆU

Speech TESS Toronto

Gồm các dữ liệu về audio với định dạng wav và labels của 2 diễn viên ở hai độ tuổi khác nhau

Text UIT-VSMEC

Gồm 6927 điểm dữ liệu gồm các kiểu dữ liệu văn bản truyền thông xã hội với 7 nhãn

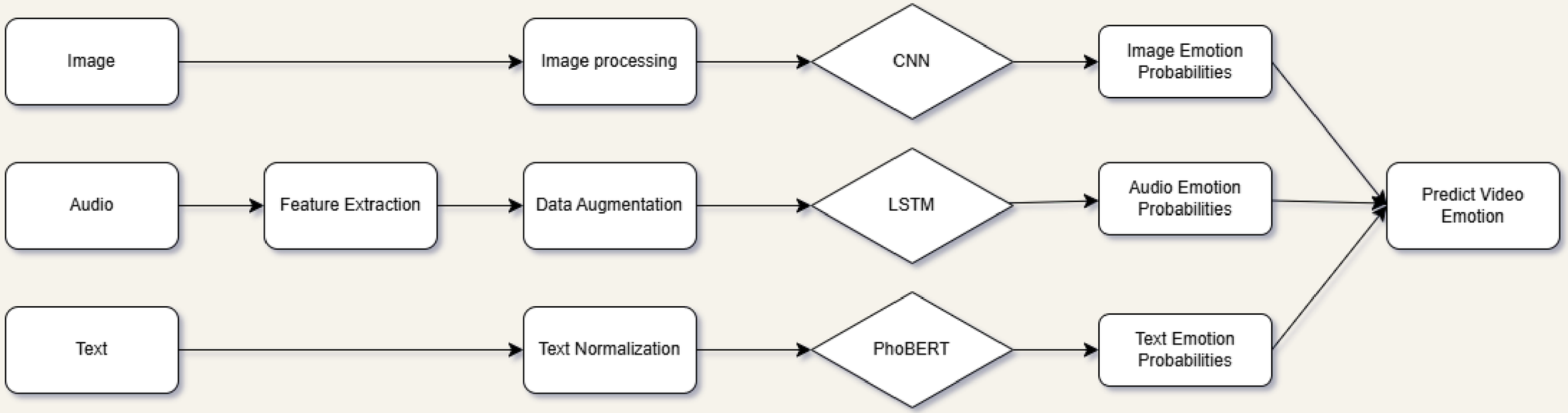
Image trên github

Gồm 35887 điểm dữ liệu gồm các hình ảnh và 7 loại label

Dataset tự thu thập

Crawl bằng code và tay với các thuộc tính gồm Audio, Text, Image và labels của từng loại modalities và label chung.

HUẤN LUYỆN MÔ HÌNH



HUẤN LUYỆN MÔ HÌNH

Thông số trong CNN

Layer Type	Output Shape	Parameters	Activation	Dropout
Conv2D	(46, 46, 32)	320	ReLU	-
MaxPooling2D	(23, 23, 32)	-	-	-
Dropout	(23, 23, 32)	-	-	0.2
Conv2D	(21, 21, 64)	18496	ReLU	-
MaxPooling2D	(10, 10, 64)	-	-	-
Dropout	(10, 10, 64)	-	-	0.2
Conv2D	(8, 8, 128)	73728	ReLU	-
MaxPooling2D	(4, 4, 128)	-	-	-
Flatten	2048	-	-	-
Dense	128	262272	ReLU	-
Dropout	128	-	-	0.5
Dense	7	903	Softmax	-

Thông số huấn luyện CNN

- validation_split = 0.1
- epochs = 15
- batch_size = 2
- optimizer = 'adam'
- loss = 'categorical_crossentropy'
- metrics = ['accuracy']
- learning_rate = 0.001

HUẤN LUYỆN MÔ HÌNH

Thông số trong LSTM

Layer Type	Output Shape	Param #
LSTM	(None, 500, 256)	304,128
Dropout	(None, 500, 256)	0
BatchNormalization	(None, 500, 256)	1,024
LSTM	(None, 500, 256)	525,312
Dropout	(None, 500, 256)	0
BatchNormalization	(None, 500, 256)	1,024
LSTM	(None, 128)	197,120
Dropout	(None, 128)	0
Dense	(None, 7)	903

Thông số huấn luyện LSTM

- validation_split = 0.2
- epochs = 20
- batch_size = 64
- optimizer = 'adam'
- loss = 'categorical_crossentropy'
- metrics = ['accuracy']

HUẤN LUYỆN MÔ HÌNH

Thông số huấn luyện BARTpho-syllable

- optimizer: AdamW
- learning_rate = 5e-5
- tokenizer: AutoTokenizer, BARTpho-syllable
- truncation=True
- padding = True
- epochs = 3
- loss = default Masked Language Model loss

Thông số huấn luyện PhoBERT

- optimizer: AdamW
- learning_rate = 5e-5
- tokenizer: AutoTokenizer, PhoBERT-base
- truncation=True
- padding = True
- batch_size = 16
- epochs = 3
- loss = cross_entropy

HUẤN LUYỆN MÔ HÌNH

Đánh giá mô hình trên từng modality riêng lẻ

	Precision	Recall	F1-score	support
angry	0.42	0.16	0.24	49
disgust	0.47	0.71	0.56	135
fear	0.47	0.55	0.51	31
happy	0.7	0.7	0.7	214
neutral	0.49	0.29	0.36	141
surprise	0.35	0.37	0.36	30
sad	0.49	0.56	0.52	86
accuracy			0.54	686
macro avg	0.48	0.48	0.46	686
weighted avg	0.54	0.54	0.52	686

Text

	Precision	Recall	F1-score	support
angry	1	0.98	0.99	333
disgust	1	0.98	0.99	306
fear	1	1	1	349
happy	0.99	0.98	0.98	325
neutral	1	1	1	307
surprise	0.95	1	0.97	312
sad	1	1	1	308
accuracy			0.99	2240
macro avg	0.99	0.99	0.99	2240
weighted avg	0.99	0.99	0.99	2240

Audio

HUẤN LUYỆN MÔ HÌNH

Đánh giá mô hình trên từng modality riêng lẻ

	Precision	Recall	F1-score	support
angry	0.46	0.44	0.45	960
disgust	0.9	0.16	0.27	111
fear	0.3	0.3	0.3	1018
happy	0.75	0.83	0.79	1825
neutral	0.44	0.66	0.53	1216
surprise	0.44	0.48	0.46	1139
sad	0.93	0.05	0.09	797
accuracy			0.52	7066
macro avg	0.6	0.42	0.41	7066
weighted avg	0.57	0.52	0.49	7066

Image

FUSION METHOD

- **Product Fusion**

  Hiệu suất cao
  Ứng dụng đa dạng

- **Weighted Average Fusion**

  Có thể dùng để cải thiện hiệu suất của dữ liệu với các độ tin cậy khác nhau bằng cách sử dụng các trọng số

RESULT

Phương pháp fusion đưa ra dự đoán bị sai lệch khá nhiều với bộ dữ liệu tự tạo.

Phương pháp	Tỷ lệ
PROD Fusion	20%
Audio WA Fusion	8%
Text WA Fusion	8%

Table 1: Bảng tỷ lệ dự đoán của fusion method

HẠN CHẾ

- **Dataset tự thu thập**
  Khuôn mặt bị đơ và chỉ có một cảm xúc, audio bị nhiều và bị mất thông tin trong lúc thu thập.

- **Bộ dữ liệu không đồng nhất**
  Bộ audio TESS chỉ gồm 2 người nữ sử dụng ngôn ngữ khác, dữ liệu bị lệch nhãn.

CONCLUSION

- Hiệu suất các modality riêng lẻ chỉ trên trung bình (F1-score, accuracy).
- Prod fusion tận dụng không được tốt thông tin từ các modalities.

THANK YOU