CMMA: Benchmarking Multi-Affection Detection in Chinese Multi-Modal Conversations

Main Authors： Yazhou Zhang, Yang Yu

Corresponding Author: Qiuchi Li*

Other Authors: Qing Guo, Benyou Wang, Dongming Zhao, Sagar Uprety, Dawei Song, Jing Qin

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

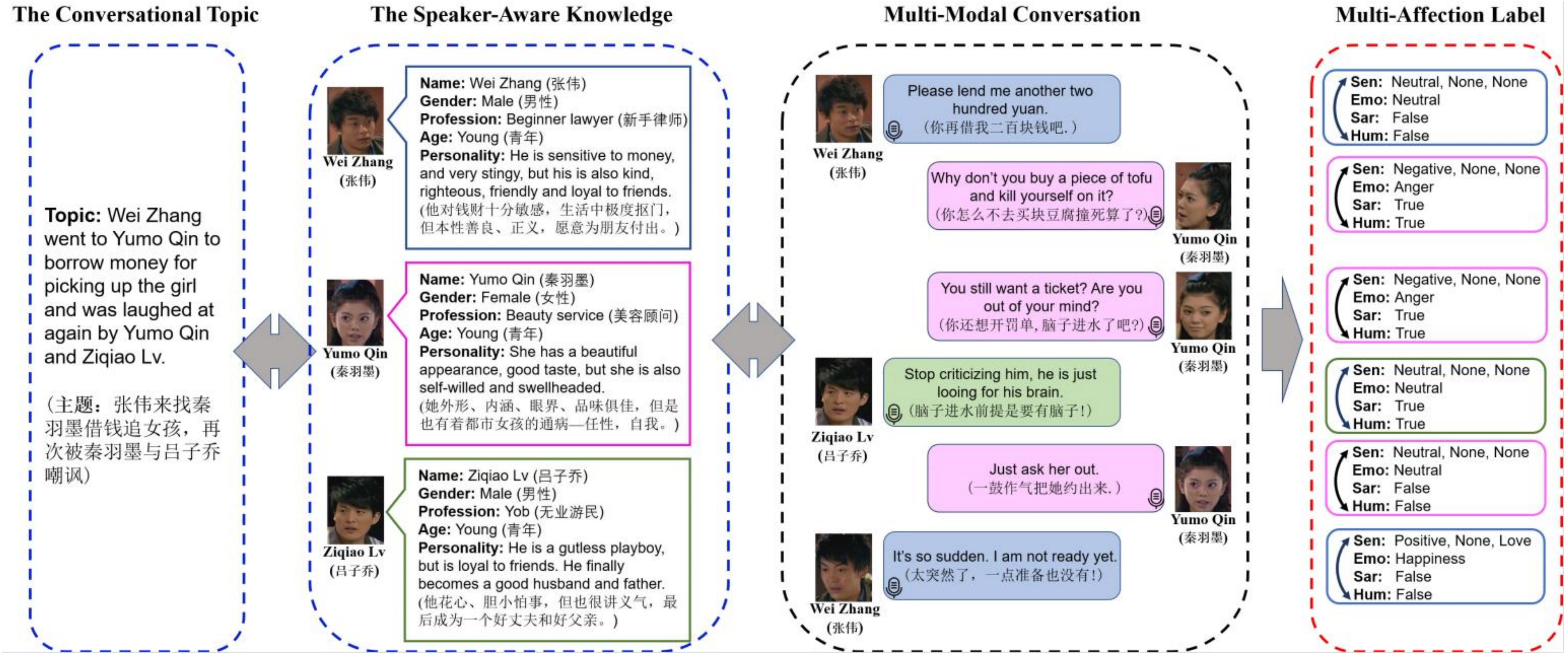
Human communication is multi-modal with textual , visual and audio channels, and also multi-affective in that different types of affects. The interactions of different modalities and inter-correlations between different affect types bring opportunities as well as challenges for multi-modal affect detection, especially in a conversational context.

- We construct the first Chinese multi-modal multi-affect conversation dataset annotated with the sentiment, emotion, sarcasm, and humor labels, along with well-illustrated quality control and agreement analysis.
- We make the first attempt to manually annotate the relevance intensity between sentiment and emotion, and between sarcasm and humor.
- We show a comprehensive statistics of the dataset, covering the distribution of TV sources, characters and affect types.
- We propose a multi-modal multi-affect joint detection model to evaluate CMMA. The results of SOTA baselines using different feature combinations suggest the need for multi-task learning models.

Related Work

Dataset	Type	Size	Modality	Resource	Language	Annotation	Inter-Task Correlation	Speaker Information	Topic
YouTube	Video	47	Text, Image, Speech	YouTube	English	Sentiment	✗	✗	✗
MOUD	Video	498	Text, Image, Speech	YouTube	English	Sentiment	✗	✗	✗
MOSI	Video	2,199	Text, Image, Speech	YouTube	English	Sentiment	✗	✗	✗
CH-SIMS	Video	2,281	Text, Image, Speech	Movie, TV	Chinese	Sentiment	✗	✗	✗
IEMOCAP	Dialogue	10,039	Text, Image, Speech	Performance	English	Emotion	✗	✓✗	✗
MELD	Dialogue	13,708	Text, Image, Speech	TV Show	English	Sentiment, Emotion	✗	✓✗	✗
MEISD	Dialogue	20,000	Text, Image, Speech	TV Show	English	Sentiment, Emotion	✗	✗	✗
ScenarioSA	Dialogue	24,072	Text	Social Media	English	Sentiment	✗	✗	✓
MUStARD	Dialogue	690	Text, Image, Speech	TV Show	English	Sarcasm	✗	✓✗	✗
Twitter	Tweet	24,635	Text, Image	TV Show	English	Sarcasm	✗	✓✗	✗
Silver-Standard	Instagram post	20K	Text, Image, Speech	TV Show	English	Sarcasm	✗	✓✗	✗
MHD	Dialogue	13,633	Text, Image, Speech	TV Show	English	Humor	✗	✓✗	✗
BBT	Dialogue	39,769	Text, Image, Speech	TV Show	English	Humor	✗	✓✗	✗
UR-FUNNY	TED talk	16,514	Text, Image, Speech	TV Show	English	Humor	✗	✓✗	✓
MUMOR	Dialogue	19,103	Text, Image, Speech	TV Show	English, Chinese	Sentiment, Emotion, Humor	✗	✓✗	✗
MaSaC	Dialogue	15,000	Text, Image, Speech	TV Show	English,Hindi	Sarcasm, Humor	✗	✓✗	✗
Memotion	Internet Meme	8,871	Text, Image	Social Media	English	Sentiment, Emotion, Sarcasm, Humor, Offensive, Motivational	✗	✗	✗
CMMA (Ours)	**Dialogue**	**21,795**	**Text, Image, Speech**	**TV Show**	**Chinese**	**Sentiment, Emotion, Sarcasm, Humor, Pride, Love**	✓	✓	✓

Dataset



Each utterance is annotated with sentiment (including pride and romantic love), emotion, sarcasm and humor labels. Considering that the external knowledge implicitly influences the speaker's affective state, the speaker's background (i.e., name, profession, sex, personality) and the topic of each conversation are provided.

Dataset Construction

Rescource → Processing → **Annotation** → Quality Control → **Division and Statistics**

TV Show
“武林外传” (My Own Swordsman)
“爱情公寓” (iPartment)
“地下交通站” (The Safe House)
“炊事班的故事” (The Story of Cooking Class)
“家有儿女” (Home with Kids)
“媳妇的美好时代” (Beautiful Daughter-in-Law)
“欢乐颂” (Ode to Joy)
“都挺好” (All Is Well)
“三国演义” (Romance of Three Kingdoms)
“父母爱情” (Romance of Our Parents)
“人民的名义” (In the Name of People)
“福贵” (Fu Gui)
“我的团长我的团” (My Chief and My Regiment)
“铁齿铜牙纪晓岚” (Ji Xiaolan)
“白夜追凶” (Day and Night)
“心理罪” (Guilty of Mind)
“天道” (Destiny)
“隐秘的角落” (The Bad Kids)

CMMA:A Chinese Multi-Modal Multi-Affective Dataset (Translated into English)

Video ID : 66

Speaker : **曾小贤** (Xiaoxian Zeng)

The utterance content:

不吓到才怪呢

Click to view the context

What kind of sentiment is in this utterance?
- ○ Positive
- ● Neutral
- ○ Negative

What kind of emotion is in this utterance?
- ○ Joy ○ Sadness
- ● Surprise ○ Anger
- ○ Disgust ○ Fear ○ Neutral

What kind of love in this utterance?
- ○ Immediate love
- ○ Growing love
- ○ Empty love ● Non-love

Is there sarcasm in this utterance?
- ○ Yes ● No

Is there humor in this utterance?
- ● Yes ○ No

Is there pride in this utterance?
- ○ Yes ● No

The correlation between sarcasm and humor
- ○ -2 ○ -1 ● 0 ○ 1 ○ 2

The correlation between emotion and sentiment
- ○ -2 ● -1 ○ 0 ○ 1 ○ 2

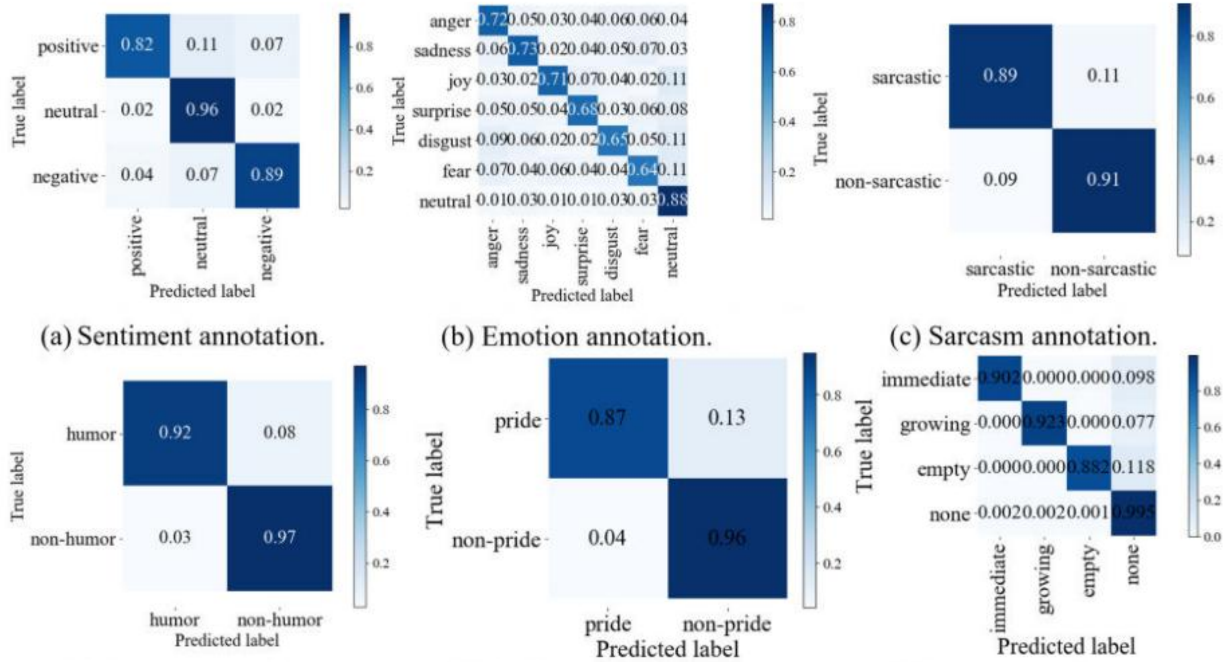
Submit Reset Next

Attention: Please keep quiet, objective, independent and rigorous during the labeling process! Don't pasting and copying! And a 5-level annotation in [−2, −1, 0, 1, 2] is used, where the sign stands for whether an affect contributes to the other or the other way around.

Table 2: Statistics of CMMA. (t,v,a) = (text, video, audio).

Item	Train	Dev	Test
#Modalities	(t,v,a)	(t,v,a)	(t,v,a)
#Conversations	1800	600	600
#Utterances	13788	4046	3961
#Speakers	299	78	119
#Words	115,434	35,487	34,521
#Unique words	2,677	1,842	1,988
#Video duration	9.2h	3.0h	3.0h
#Average utterances per conversation	7.7	6.8	6.6
#Average words per conversation	64.1	59.1	57.5
#Average words per utterance	8.4	8.8	8.7
#Average duration of a conversation	18.5s	18.4s	17.8s
#Average duration of an utterance	2.4s	2.7s	2.8s
#Average turns per conversation	3.7	3.3	3.2

Annotation agreement



(a) Sentiment annotation. (b) Emotion annotation. (c) Sarcasm annotation.

(d) Humor annotation. (e) Pride annotation. (f) Love annotation.

Table 2: Inter-agreement comparison between CMMA and other datasets.

Affection	CMMA (ours)	MOSI	MELD	IEMOCAP	ScenarioSA	MUStARD	EmotionLines	MUMOR	MaSaC	MEISD
Sentiment	**0.85**	0.77	0.48	0.57	0.57	-	-	0.84	-	0.75
Emotion	**0.69**	-	0.43	0.40	-	-	0.33	0.45	-	0.67
Sarcasm	**0.68**	-	-	-	-	0.58	-	-	0.65	-
Humor	**0.85**	-	-	-	-	-	-	0.81	0.68	-
Pride	**0.71**	-	-	-	-	-	-	-	-	-
Love	**0.83**	-	-	-	-	-	-	-	-	-
Num. of Annotators	**9**	5	3	6	5	3	5	3	5	4

(1) **Percent agreement calculation approach:** 88.8%, 71.5%, 86.8%, 94.5%, 82.5%, 94.9%.

(2) **Fleiss' kappa score**: 0.85, 0.69, 0.68, 0.85, 0.71, 0.83.

Compared with other related datasets, we have attained the highest inter-agreement scores on all tasks

Benchmark

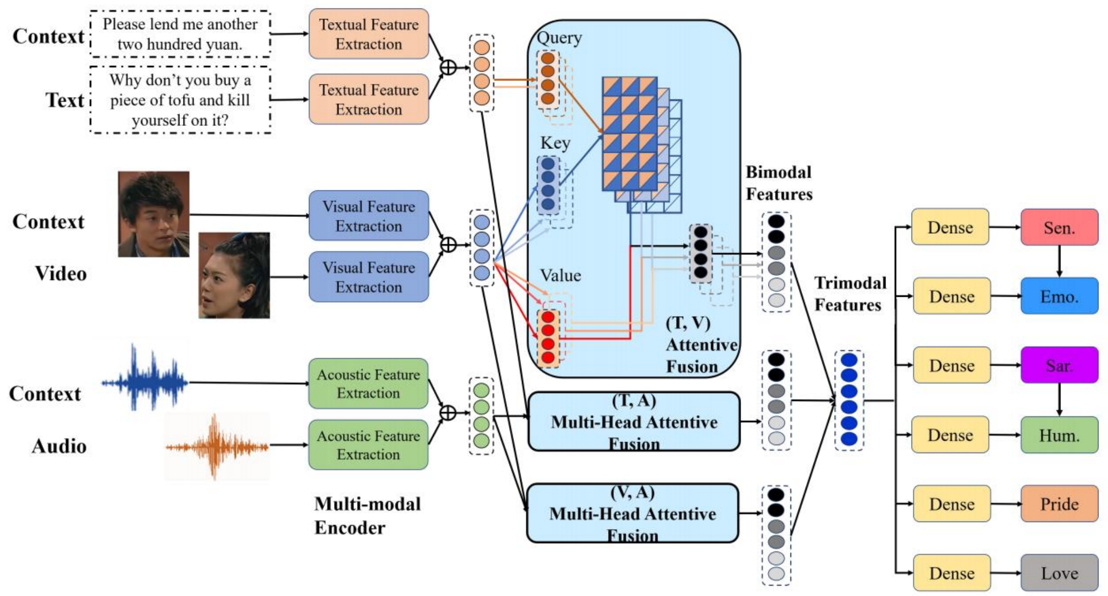


Figure 5: Multi-modal multi-affect joint detection model.

Table 4: Comparison of different models.

Model	Text	Video	Audio	Sentiment			Emotion			Sarcasm			Humor		
				P	R	M_a-F1	P	R	M_a-F1	P	R	M_a-F1	P	R	M_a-F1
Text	BiLSTM	-	-	50.36	51.22	50.74	41.52	62.12	44.74	56.43	50.64	52.29	43.69	55.9	49.05
	BERT	-	-	56.77	55.51	54.89	51.85	70.87	56.18	54.56	53.88	53.61	51.5	56.94	54.08
	GPT-2	-	-	53.88	58.01	54.35	45.33	44.37	45.21	51.41	53.48	52.42	44.81	64.39	52.85
	GPT-3	-	-	54.66	54.21	54.43	48.72	47.65	48.18	53.27	54.87	54.06	49.84	47.21	48.49
Video	-	EfficientNet	-	42.86	45.12	42.84	38.08	61.58	42.18	46.77	61.66	53.19	38.06	52.8	44.23
	-	ResNet	-	48.92	51.53	49.40	47.65	47.89	47.66	57.66	57.84	57.75	41.84	55.69	47.78
Audio	-	-	VGGish	41.15	62.12	44.89	33.24	26.70	30.64	42.19	43.54	42.85	34.98	44.81	46.84
Text+Video	BiLSTM	EfficientNet	-	49.68	52.33	50.20	40.51	39.69	40.10	45.70	57.17	50.80	44.67	58.18	50.53
	BiLSTM	ResNet	-	48.77	51.27	49.30	36.68	48.86	37.49	50.69	57.17	53.74	42.51	61.7	50.34
	BERT	EfficientNet	-	65.47	69.88	66.75	41.16	61.68	44.29	55.74	58.35	57.02	**53.59**	61.9	**58.44**
	BERT	ResNet	-	**67.32**	**73.36**	**68.89**	**56.24**	**68.54**	**57.82**	**67.84**	65.69	**66.75**	52.03	**66.25**	58.29
	GPT-2	EfficientNet	-	58.13	64.24	59.17	38.08	61.58	42.18	45.45	56.05	50.20	46.02	63.35	53.31
	GPT-2	ResNet	-	59.09	66.32	60.03	42.17	61.80	45.91	50.55	61.65	55.56	45.75	64.6	53.56
Video+Audio	-	EfficientNet	VGGish	49.22	50.21	48.27	41.15	62.12	44.89	38.59	59.19	46.73	40.89	64.18	49.96
	-	ResNet	VGGish	52.47	53.52	51.62	52.12	51.04	51.44	42.12	58.74	49.06	42.63	65.84	51.75
Text+Audio	BiLSTM	-	VGGish	46.97	49.55	46.84	43.13	64.83	46.82	40.85	66.59	50.64	42.23	64.18	50.94
	BERT	-	VGGish	54.41	55.25	55.74	46.93	63.31	50.36	43.57	68.39	53.22	48.99	65.22	55.95
	GPT-2	-	VGGish	51.41	53.48	52.42	45.23	66.98	49.44	41.52	**69.73**	52.05	45.29	63.77	52.97
Text+Video+Audio	BERT	EfficientNet	VGGish	69.59	73.98	71.12	53.03	74.37	57.36	69.38	65.02	67.13	63.76	69.57	66.53
	BERT	ResNet	VGGish	71.64	**76.31**	**73.29**	**56.71**	**76.32**	**61.76**	**76.28**	**74.22**	**75.23**	**76.47**	**75.36**	**75.91**
	GPT-2	EfficientNet	VGGish	65.66	69.47	66.86	47.06	73.71	51.49	58.95	62.78	60.8	58.16	62.73	60.36
	GPT-2	ResNet	VGGish	**71.76**	74.87	72.88	52.09	73.82	56.17	74.44	67.26	70.67	65.80	73.29	69.34
Trimodal *vs* Bimodal (%)	-	-	-	+6.6	+4.0	+6.4	+0.8	+11.3	+6.8	+12.4	+6.4	+12.6	+42.6	+13.7	+29.6

- **Textual Feature Extraction:** BiLSTM, BERT, GPT-2
- **Visual Feature Extraction:** EffcientNet, ResNet
- **Acoustic Feature Extraction:** VGGish

- For each modality, the encoded utterance is concatenated with its encoded context, and the unimodal contextual features are combined by multi-modal fusion. The obtained multi-modal representation is then passed through task-specific dense layers for each affect detection task. The labels of all tasks are produced in the forward pass, where we set different weights for different tasks.

Results

NEURAL INFORMATION PROCESSING SYSTEMS

Table 5: Comparison of different multi-modal fusion strategies.

Trimodal Accuracy	Sentiment		Emotion		Sarcasm		Humor	
	Validation	Test	Validation	Test	Validation	Test	Validation	Test
Multi-head Attention	74.81	**78.48**	72.24	**77.09**	**82.44**	**85.64**	84.31	**86.15**
Concatenate	**76.76**	76.31	73.14	76.32	82.22	84.28	**85.06**	85.88
Add	71.62	77.39	73.33	76.36	82.37	84.86	85.06	82.93
Multiply	69.85	72.22	70.39	73.05	78.77	78.54	80.91	81.31
Maximum	75.95	76.38	**74.11**	72.47	81.25	83.13	81.66	79.42

Table 7: Effect of the relevance between sentiment-emotion / sarcasm-humor.

Setup	Sentiment		Emotion		Sarcasm		Humor	
	$\mathrm{M_a}$-F1	Acc	$\mathrm{M_a}$-F1	Acc	$\mathrm{M_a}$-F1	Acc	$\mathrm{M_a}$-F1	Acc
STL	71.17	72.22	59.75	72.47	71.97	83.13	73.21	79.41
$S-MTL:Emo.$	71.61	72.85	**61.76**	76.32	-	-	-	-
$S-MTL:Sen.$	73.14	75.52	60.12	73.39	-	-	-	-
RaM	**74.31**	**79.55**	**61.76**	**77.09**	-	-	-	-
$S-MTL:Sar.$	-	-	-	-	74.22	**85.64**	73.77	80.46
$S-MTL:Hum.$	-	-	-	-	72.27	83.84	74.51	85.42
RaM	-	-	-	-	**75.23**	**85.64**	**75.91**	**86.15**

Conclusions and Future work

- Few works (including the recent large language models) have set foot in multi-affect joint detection in conversations, largely due to the lack of multi-modal conversation datasets with multi-affect annotations. We have filled this gap by proposing CMMA, the first multi-modal multi-affect conversation dataset. CMMA consists of 21,795 multi-modal utterances from 3,000 multi-party conversations. Apart from rich affect labels including sentiment, emotion, sarcasm and humor, the dataset contains annotation relevance between affect types.
- We have performed comprehensive qualitative and quantitative studies for analyzing the dataset, and presented a range of baselines to evaluate the potential of CMMA. The results demonstrate the quality of the dataset and indicate the need of novel investigations in models in multi-modal multi-affect joint detection in conversations.

NEURAL INFORMATION PROCESSING SYSTEMS

Thanks for listening

This work is partly supported by a Project of Strategic Importance grant of The Hong Kong Polytechnic University (project no. 1-ZE2Q). This paper is also partly supported by a grant under Hong Kong RGC Theme-based Research Scheme (project no. T45-401/22-N). This work is also supported by National Science Foundation of China under grant No. 62006212, Fellowship from the China Postdoctoral Science Foundation (2023M733907).

中国移动通信
CHINA MOBILE