Multi-modal Emotion Recognition Utilizing Korean-English Vision and Language Information Alignment Pre-trained Model

Jeeyoung Yun[1,2], San Kim[1], Minyoung Jung[1], Saim Shin[1], and Jin Yea Jang[1]
[1]*Artificial Intelligence Research Center, Korea Electronics Technology Institute*
Seongnam, Republic of Korea
[2]*Department of Computer and Information Engineering, Kwangwoon University*
Seoul, Republic of Korea
{qwer010910, kimsan0622, minyoung.jung, sishin, jinyea.jang}@keti.re.kr

Abstract—**Emotions in humans find expression through a variety of modalities. In this paper, we build a new dataset that combines existing vision and language datasets and use this newly formed multi-modal dataset for the task of multi-modal emotion recognition in empathetic conversation. We utilize a vision-language pre-trained model, VL-KE-T5, to build a model that can process image and text information simultaneously. Comparative experiments show that the proposed model outperforms models that handle image and text separately in emotion recognition.**

Index Terms—**emotion recognition, multi-modal information, VL-KE-T5**

I. INTRODUCTION

Developing an artificial intelligence conversational agent for empathetic interactions [1] places an emphasis on understanding user emotions. To enhance the performance of emotion recognition, leveraging insights from various communication channels – encompassing spoken words, vocal cues, facial expressions, and gestures – proves advantageous. [2]. Consequently, this study performs emotion recognition by utilizing two distinct forms of information: utterance text and facial expression images within empathetic conversational contexts.

To recognize user emotions in empathetic conversations, a multi-modal dataset specialized for empathetic conversations is necessary. The publicly available empathetic conversation dataset [3] contains only textual utterance information, which requires modality augmentation to train a multi-modal emotion recognition model. Therefore, we combined the emotion image dataset [4] with the empathetic conversation data to construct a new multi-modal emotion recognition dataset. The two uni-modal datasets are Korean conversation and Korean facial expression image data, which satisfy cultural congruence with each other.

To combine textual and visual modality information, we propose a model structured with a bi-encoder architecture, where a transformer encoder for each modality exists. The backbone is the multi-modal pre-trained model, denoted as VL-KE-T5[1], designed to align text and image information at the pre-training stage. The representation vectors yielded by each encoder are combined and utilized in the computation of emotion classification. Through comparative experiments involving text-based and image-based models, we substantiate that our proposed multi-modal bi-encoder model surpasses uni-modality models in emotion recognition performance.

II. RELATED WORK

Several studies have introduced diverse approaches in emotion recognition utilizing two or more modalities, such as audio-text [5], audio-video [6], and text-audio-video [7]. However, these studies have limitations, such as the small number of classified emotion labels or the small size of the datasets used to train the models. In terms of the structure of the model, these studies designed the model in that the hidden representation information for each modality input is obtained from a corresponding module that has no information connection to each other and then combined.

We enhance the granularity of emotion labels in multi-modal emotion recognition. Moreover, to strengthen the interconnection of multi-modal information, we adopt a model architecture where vision and language information are aligned during pre-training.

III. MULTI-MODAL EMOTION RECOGNITION MODEL

The backbone model, VL-KE-T5 aligns the embedding vectors of two pre-trained models: KE-T5[2], a language-based model from the Korea Electronics Technology Institute (KETI), and Vision Transformer (ViT)[3], a vision-based model from Google, using a vision-language parallel corpus. KE-T5 is a pre-trained model of T5 [8] structure in Korean and English. It solves natural language processing tasks by reconstructing them in a text-to-text format. KE-T5 is pre-trained on both Korean and English corpora, enabling it to process both languages simultaneously. ViT is a transformer-based model that splits images into 16X16 patches and uses them as tokens. It has demonstrated superior performance over convolutional architectures.

Our bi-encoder emotion recognition model is designed by adding a classification head to VL-KE-T5 that can process

[1] https://github.com/AIRC-KETI/VL-KE-T5

[2] https://github.com/AIRC-KETI/ke-t5

[3] https://github.com/google-research/visiontransformer

979-8-3503-1327-7/23/$31.00 ©2023 IEEE

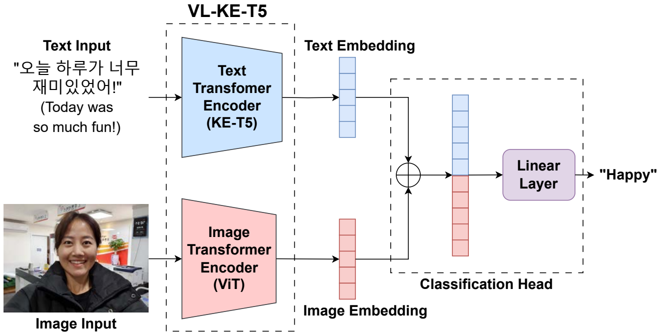

Fig. 1. Multi-modal emotion recognition model

language and vision information together. The structure of the model can be found in Fig. 1. The model takes as input a multi-modal dataset consisting of text and image pairs, which are the users' utterances and facial expressions. Text inputs are converted to text embeddings via a text encoder, and images are converted to image embeddings via an image encoder. At the classification head, the two embeddings are concatenated and passed through a linear layer to compute the probability values of the emotion labels.

IV. EXPERIMENT

This section explains the process of creating the dataset for our experiments and our experimental approach.

A. Uni-Modality Datasets

As there's no Korean open-source empathetic conversation dataset for multi-modal emotion recognition, we expand the modalities of a text dataset with an image dataset and then create a new multi-modal dataset. We conducted our experiments using Emotional Conversation Corpus [3] (ECC) and Composite Image Data for Korean emotion recognition [4] (CID).

ECC is conversations between human users and chatbot systems. Each conversation consists of a maximum of three turns, which means up to six sentences. The corpus contains information about the user's age, gender, conversation topics, and physical ailments. Emotion labels are classified into six main categories: happy, sad, anger, embarrassment, fear, and hurt along with 60 subcategories.

CID was produced by having an actor portray a given emotion, followed by three annotators who evaluated and reviewed the actor's facial expressions. In addition to location and background information, this dataset contains the user's age, gender, and other relevant details. The emotion labels in the corpus follow Susan David's concept of emotional agility [9], which is categorized into seven labels with an additional neutral label: happy, sad, anger, embarrassment, fear, hurt, and neutral.

B. Building Multi-modal Dataset

The entire process of building the multi-modal emotion dataset is in Fig. 2.

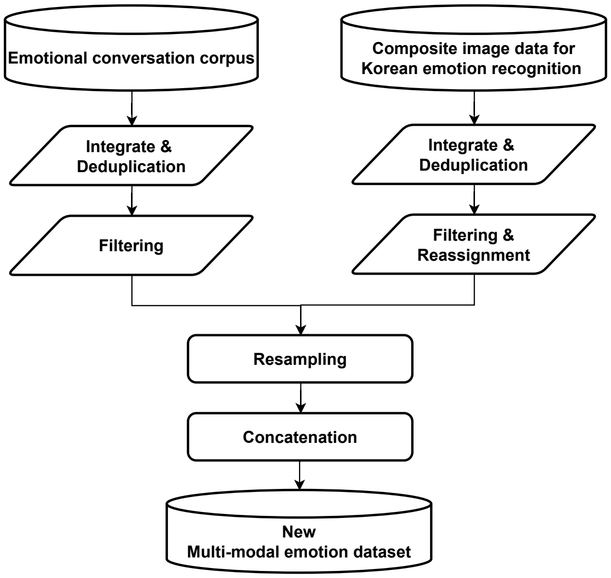
Emotional conversation corpus
Composite image data for Korean emotion recognition
Integrate & Deduplication
Integrate & Deduplication
Filtering
Filtering & Reassignment
Resampling
Concatenation
New Multi-modal emotion dataset

Fig. 2. Building Multi-modal Dataset Process

1) Data Preprocessing: The two aforementioned data were split into two separate sets, train and valid. We integrated the individual sets into a comprehensive set and removed duplicate data.

To ensure consistency, we used only the first and second sentences spoken by human users in ECC. These sentences mainly describe the conversation situation and express significant emotions. As ECC lacks neutral labels, we designated the chatbot system's sentences as neutral labels.

We extracted only data from CID that matched both the emotion expressed by the actor and the emotion perceived by the annotator. While ECC classifies individuals by age as teenagers, young adults, middle-aged, and elderly, CID has individuals in their 10s, 20s, 30s, 40s, 50s, and 60s. To align this data with ECC, we reassigned 10s as teenagers, 20s and 30s as young adults, 40s and 50s as middle-aged, and 60s as elderly.

2) Data Resampling: ECC and CID were not constructed concurrently, resulting in mismatched data counts. Therefore, we analyzed the age and gender distributions of the two datasets and aligned them with the one containing fewer data.

3) Data Concatenation: We concatenated datasets with the same age, gender, and emotion labels in both datasets after adjusting the numbers. Sentence texts are taken from ECC and images are taken from CID. Then we split the newly created datasets into a ratio of 8:1:1 for training, validation, and testing. To avoid cheating in the evaluation, no single person was included in more than one of the three datasets because there were multiple images of the same person with different backgrounds in CID. The final dataset statistics are in Table I.

C. Implementation Details

The experimental environment ran on Ubuntu 20.04, utilizing an Nvidia TITAN RTX. The model hyper-parameters were the batch size of 8, the learning rate of 1e-4, and the AdamW

TABLE I
MULTI-MODAL EMOTION DATA STATISTICS

Label	Train	Valid	Test	Total
Happy	6,349	762	778	7,889
Sad	7,835	1,048	992	9,875
Anger	7,737	1,099	877	9,713
Embarrassment	6,992	1,002	862	8,856
Fear	5,320	795	587	6,702
Hurt	4,292	561	417	5,270
Neutral	8,056	980	989	10,025
All	46,582	6,247	5,502	58,330

TABLE II
EXPERIMENTAL RESULTS (F1-SCORE)

Label	KE-T5 based	ViT based	VL-KE-T5 based
Happy	0.80	0.90	**0.92** (+0.12, +0.02)
Sad	0.54	0.83	**0.84** (+0.30, +0.01)
Anger	0.43	0.79	**0.83** (+0.40, +0.04)
Embarrassment	0.46	**0.87**	**0.87** (+0.41, +0.00)
Fear	0.39	0.56	**0.65** (+0.26, +0.09)
Hurt	0.15	0.39	**0.44** (+0.29, +0.05)
Neutral	**1.00**	0.85	**1.00** (+0.00, +0.15)
All	0.59	0.78	**0.84** (+0.25, +0.06)

TABLE III
EXAMPLES OF A MULTI-MODAL MODEL BEING CORRECT
AND A UNI-MODAL MODEL BEING INCORRECT

Utterance	True Label	KE-T5 based prediction	Vit based prediction
My daughter wants a laptop, so I’m buying her one.	Happy	Embarrassment	-
I’m glad I don’t have to see those selfish people anymore.	Sad	Happy	-
Now I feel like a burden to my parents. I’m such a piece of shit.	Anger	-	Neutral
How about now?	Neutral	-	Fear

optimizer. The model was trained for 30 epochs, with the test conducted on the model with the best validation performance.

V. RESULTS

For performance comparison, we trained the KE-T5 and ViT-based uni-modality models also. As the KE-T5 is a language-based model, it exclusively utilizes text from the newly constructed dataset as its input. In contrast, ViT is a vision-based model, therefore it takes only images from the dataset as input. All three models were trained by 30 epochs, then the test dataset results from the models that showed the highest accuracy on the validation dataset.

The results of the experiment can be found in Table II, where the F1 scores on each model and emotion label are reported. The numbers in parentheses in the table are the score difference between the multi-modal model and each uni-modal model. The experimental results show that the overall score of the VL-KE-T5 based model, followed by the ViT based model, and then the KE-T5 based model had the highest emotion recognition performance. Examples where uni-modal models failed but multi-modal models succeeded in predicting are shown in Table III. The English utterances in the table are translations of the original Korean data.

For individual emotion labels, the VL-KE-T5 based model achieved the highest scores across all emotions except ’neutral’. The KE-T5 based model accurately predicted all instances of the ’neutral’ label. This is presumed to be due to the construction of data where ’neutral’ user utterances, absent from the ECC dataset, were taken using chatbot system utterances. Hence, it can be inferred that during the training of KE-T5, the model learned the differences in tone between user and chatbot system utterances, when handling the ’neutral’ label.

In terms of information modalities, the combination of vision information was more helpful in recognizing negative emotions than the use of language information alone. Comparing the use of vision alone with the use of additional language information, it was relatively more helpful in recognizing ‘fear’ and ‘hurt’ among the negative emotions.

VI. CONCLUSIONS

This research introduced a new multi-modal emotion recognition model, leveraging a pre-trained model that aligns both language and visual information. The conducted experiments demonstrated that employing multiple modalities leads to more precise and comprehensive emotion recognition in contrast to relying on a single modality. In future work, we plan to evaluate the effectiveness of information alignment between modalities within a pre-trained model and compare it to a model without such alignment. We are also considering exploring multi-modal emotion recognition models with additional modalities such as audio information.

ACKNOWLEDGMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) (No. 2022-0-00320) grant funded by the Korea government (MSIT)

REFERENCES

[1] J. Gao, Y. Liu, H. Deng, W. Wang, Y. Cao, J. Du, and R/ Xu, “Improving Empathetic Response Generation by Recognizing Emotion Cause in Conversations,” In Findings of the Association for Computational Linguistics EMNLP 2021, pp. 807–819, 2021

[2] G. Subramanian, N. Cholendiran, K. Prathyusha, N. Balasubramanain and J. Aravinth, ”Multimodal Emotion Recognition Using Different Fusion Techniques,” 2021 Seventh International Conference on Bio Signals, Images, and Instrumentation (ICBSII), Chennai, India, 2021, pp. 1-6

[3] Emotional Conversation Corpus, https://aihub.or.kr/aidata/7978

[4] Composite Image Data for Korean Emotion Recognition, https://aihub.or.kr/aidata/27716

[5] S. Yoon, S. Byun and K. Jung, “Multimodal speech emotion recognition using audio and text”, 2018 IEEE Spoken Language Technology Workshop (SLT), pp. 112-118, 2018.

[6] M. Singh and Y. Fang, ”Emotion recognition in audio and video using deep neural networks,” 2020.

[7] A. Joshi, A. Bhat, A. Jain, A. Singh, and A. Modi, “COGMEN: COntextualized GNN based Multimodal Emotion recognitioN,” 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4148–4164, 2022.

[8] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. “Exploring the limits of transfer learning with a unified text-to-text transformer.” The Journal of Machine Learning Research 21.1 (2020): 5485-5551.

[9] S. David, “Emotional agility: Get unstuck, embrace change, and thrive in work and life.” Penguin, 2016.