2024 IEEE 15th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON) | 979-8-3315-4090-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/UEMCON62879.2024.10754673

Fully Auto-Regressive Multi-modal Large Language Model for Contextual Emotion Recognition

Hiroshi Nonaka
Soka University of America
Aliso Viejo, USA
hnonaka@soka.edu

Damian Valles
Ingram School of Engineering
Texas State University
San Marcos, USA
dvalles@txstate.edu

***Abstract*—Large Language Models (LLMs) have gained popularity due to their high performance in natural language processing. This capability is underpinned by their ability to contextualize relations within text data. The application of LLMs also extends to multi-modal tasks, such as image captioning and video analysis. Research endeavors have been undertaken to leverage LLMs for emotion recognition tasks by utilizing visual, audio, and text data. Inspired by natural human approaches to emotion recognition, we propose a simple yet powerful, multi-modal LLM architecture for emotion recognition in conversations (ERC). Our proposed framework aims to contextualize the change of emotions better. We developed the Fully Auto-Regressive multi-modal LLM for Contextual Emotion Recognition (FARCER) based on this idea. This model consists of the instruction-tuned LLaMA3, a vision encoder, and linear layers that map visual embeddings to the LLM input space. FARCER significantly improved ERC benchmarks over the uni-modal LLaMA3-Instruct, although the LLM and vision encoder's parameters remained frozen during training. Fine-tuned FARCER demonstrated high performance comparable to other state-of-the-art (SOTA) models, highlighting the potential of our context-focused design combined with conversational LLMs for ERC.**

***Index Terms*—Artificial Intelligence, Large Language Models, Vision-Language Models, Multi-modal LLMs, Emotion Recognition**

I. Introduction

Inspired by the emergence of novel architectures such as Transformers [1] and the expansion of data and computational resources, researchers and companies developed Large Language Models (LLMs) for various natural language processing tasks. Recent examples include OpenAI's ChatGPT [2], a widely used LLM for chat-related purposes, Gemini from Google DeepMind [3], and Claude from Anthropic [4]. Meta AI has developed the LLaMA series [5][6], which achieved high scores on several tasks while significantly reducing the parameter size compared to other major LLMs.

Researchers have also attempted to ground multi-modal reasoning in LLMs. Vision Language Models (VLMs) can tackle tasks that involve image and language processing. Some popular approaches include *cross-attention architecture* and *fully auto-regressive architecture*. The cross-attention architecture applies cross-attention layers to interweave visual and text features, modeling the semantic interconnections between different modalities. The fully auto-regressive architecture incorporates a more straightforward strategy of joining visual and text feature vectors as one input. This approach requires a *modality projection* layer to map the visual representation to the same input space of the internal LLM [7]. VLMs with these mechanisms show high capability in image-text reasoning [8][9][10][11][12].

ERC is a meaningful real-life task to which researchers have discussed applying computer programs [13][14][15][16]. Various kinds of information, such as facial expressions, gestures, and dialogue settings deliver emotional clues. Furthermore, contextualizing emotions associated with utterances is key to predicting emotions from continuous dialogues.

This paper introduces a unique, fully auto-regressive architecture (FARCER architecture) to leverage LLMs for contextualizing vision-language dialogues. This framework concatenates visual and text feature tokens alternately and captures inter-modal context. Based on the proposed architecture, we also developed a Fully Auto-Regressive multi-modal LLM for Contextual Emotion Recognition (FARCER). FARCER employs an instruction-tuned LLaMA3 of 8 billion parameters as the core LLM. Vision Transformers (ViT) model pre-trained for facial emotion recognition serves as a visual encoder [17]. Linear layers map image feature vectors from ViT into the LLM input space. We will discuss further design details in Section III.

Our test suggested that employing the FARCER architecture significantly improves the performance of the core LLM. Moreover, one FARCER model with a specific configuration achieved the SOTA-line weighted-F1 and accuracy scores. These results highlight the effectiveness of our context-focused model design and the potential applicability of FARCER architecture to other ERC tasks. We will also provide a comparative analysis of several different implementations and discuss the effectiveness and limitations of our framework.

Our research will provide the following contributions:

- **Novelty**: To the best of our knowledge, FARCER architecture is the first design that leverages in-context concatenation of image and text features for ERC, and this approach demonstrated high accuracy on real-life ERC benchmarks.
- **Simplicity**: FARCER architecture can ground visual ERC reasoning to LLMs solely by training the modality projec-

This research is supported by the National Science Foundation under award number 2150135.

tion layers. This framework does not necessarily require fine-tuning LLMs or visual encoders.
- **Applicability**: FARCER utilizes an instruction-tuned causal LM (LLaMA3) as an internal LLM. FARCER architecture suggests the potential applicability of such LMs to multi-modal and interactive tasks.

II. Related Work

A. Vision-Language Models

VLMs process both text and image/video data. *Flamingo* uses pre-trained vision encoders to extract visual representations and applies cross-attention layers to capture the semantic relationships between text and images [8]. *FROMAGe* [9] is an example of a fully auto-regressive model. The researchers showed that *FROMAGe* presented a high performance on image captioning and retrieval through training only the modality projection layer. *MiniGPT4-Video* [18] incorporates a similar approach to ours for general video understanding; the model converts a video into 45 frame features and concatenates subtitles to each frame. In *Macaw-LLM* [12], text, audio, and visual encoders are trained on many multi-modal instruction data. As a result, *Macaw-LLM* achieved high performance on reasoning, understanding, and question answering. *Video-ChatGPT* [11] is a fine-tuned model on image-text-paired instructions and embeds videos into both spatial and temporal features to pass them to an LLM.

B. Emotion Recognition in Conversations

Understanding how conversations evolve is key to identifying the emotions of speakers. Several researchers have leveraged pre-trained LLMs to this end. In the *InstructERC* architecture [15], LLMs engage in *speaker identification* and *emotion impact prediction* before actually recognizing emotions to reinforce their performance on ERC. *InstructERC* achieved SOTA with several different base LLMs in combination with LoRA fine-tuning [19]. The *CKERC* framework fine-tunes LLMs to uncover the implicit clues of the subject speaker based on past utterances and utilizes the information to predict the emotion [16].

ERC also often necessitates analyzing multiple factors in dialogues. Therefore, it is natural to incorporate multi-modal data for more precise and effective prediction. Researchers have applied multi-modal LLMs to contextualize dialogues better and process multiple modalities for inference. *Emotion-Guided LLM-Based Multimodal Dialogue Method* [14] is a set of three phases that grounds emotional response generation ability in LLMs. In the second *Response Emotion Prediction* stage, models predict the speaker's emotion based on visual and text vectors from the first module, which is trained on a contrastive loss of emotion labels. *DialogueLLM* [13] processes visual data by retrieving video text descriptions. Then, the model merges the captions with utterances based on a template and predicts emotion labels.

III. Methodology

The LM-based models and frameworks discussed in Section II-B achieve high scores on some ERC benchmarks; nevertheless, few of these approaches focus on preserving the inter-modal context of dialogues. Uni-modal LLMs easily lose visual features by solely relying on text information. For multi-modal approaches, VLMs thus far tend to create input as if stacking blocks of different modalities, which is a common strategy observed in several fully auto-regressive architectures [12][14][20]. In real-life situations, however, humans judge emotions based on visual information and speech content in parallel. Regarding this point, the *block-stacking* approach cannot align concurrent visual information and utterance, discarding meaningful inter-modal connectivity. This section will discuss our architecture design that aims to overcome this gap by better contextualizing visual-text totality in conversations.

A. Task Definition

Here, we provide the formal description of ERC tasks FARCER aims to tackle. For a given dialogue, we assume there are n utterances whose set is denoted as $\mathcal{U} = \{u_1, u_2, \ldots, u_n\}$ and n corresponding images $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$. The goal of our training process is to update the model parameters θ and approximate the model to:

$$P(l_n \mid \mathcal{U}, \mathcal{V}; \theta) \tag{1}$$

where l_n is the emotional label of the nth utterance speaker.

B. Model Description

FARCER architecture constructs input in the following way. First, we apply the ViT layers F_{vit} to extract the last hidden states of the image pixels h_i:

$$\mathcal{H} = \{h_i \mid v_i \in \mathcal{V} : h_i = F_{vit}(v_i)\} \tag{2}$$

where $h_i = \{t_{i1}, t_{i2}, \ldots, t_{i197}\}$ is a set of 197 visual feature token embeddings.

Then, we extract the ViT CLS token at index 1 of each last hidden state:

$$\begin{aligned} \mathcal{CLS} &= \{t_{11}, t_{21}, \ldots, t_{n1}\} \\ &= \{cls_1, cls_2, \ldots, cls_n\} \end{aligned} \tag{3}$$

where $cls_i \in \mathbb{R}^{1\times p}$ and p is the hidden size of the ViT.

Next, the modality projection layer $F_{mp} : \mathbb{R}^{n\times p} \rightarrow \mathbb{R}^{n\times q}$ transforms the CLS vectors into a proper LLM input embedding space:

$$\begin{aligned} \mathcal{E}_v &= F_{mp}(\mathcal{CLS}) \\ &= \{e_{v1}, e_{v2}, \cdots, e_{vn}\} \end{aligned} \tag{4}$$

Note that $\mathcal{E}_v \in \mathbb{R}^{n\times q}$ where q is the hidden size of the LLM. Separately, we prepare utterance token embeddings:

$$\mathcal{E}_u = \{e_{u1}, e_{u2}, \ldots, e_{un}\}$$

and prompt embedding e_p by using the LLM embedding layer.

Finally, the model joins $\mathcal{E}_v$, $\mathcal{E}_u$, and e_p in the following manner, aiming to capture the natural visual-text context of the dialogue:

$$\begin{aligned}\mathcal{E} &= [e_p, e_{vi}, e_{ui} | e_{vi} \in E_v, e_{ui} \in E_u] \\ &= [e_p, e_{v1}, e_{u1}, e_{v2}, e_{u2}, \ldots, e_{vn}, e_{un}]\end{aligned} \tag{5}$$

, where $\mathcal{E} \in \mathbb{R}^{(s+t+n)\times q}$ and s is the number of prompt tokens, and t is the total number of text tokens for the dialogue.

The LLM computes logits $l_p \in \mathbb{R}^{T\times K}$ (T is the number of label tokens, and K is the vocabulary size of the LLM) by applying the forward function F_{LLM} and slicing it. We can obtain a predicted emotion label with a sampling method F_{sp}:

$$\begin{aligned}l_p &= F_{LLM}(\mathcal{E})[-T:] \\ label &= F_{sp}(l_p)\end{aligned} \tag{6}$$

Fig. 1 visualizes the whole FARCER architecture. The snow icons represent frozen parameters, and the fire marks indicate the model component is trainable. The red and green squares denote the prompt and utterance tokens, respectively. Image data is visualized with the blue squares.

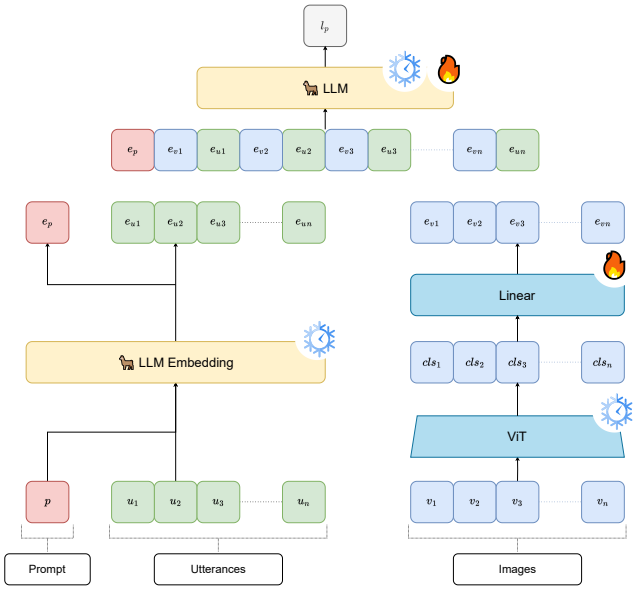

Fig. 1: FARCER Architecture

C. Training Pipeline

We trained our models on **MELD** [21] and **IEMOCAP** [22] datasets.

- **MELD** contains video and text data with 13,708 utterances annotated with seven different emotion labels.
- **IEMOCAP** is a multi-modal dataset with 151 two-party conversations (7,433 utterances) and nine emotion labels.

Table I outlines the dataset size of MELD and IEMOCAP. In IEMOCAP, each dialogue includes 49.2 utterances on average, whereas the average number of utterances in MELD is 9.6. Inputting especially large amounts of data limits the inference of LLMs, blocking effective training. To shorten the dialogues and increase the sample size, we divided dialogues in the training data into sub-dialogues of 5-21 utterances. When we trained our models on IEMOCAP, we initialized the model parameters with the pre-trained weights from MELD.

TABLE I: Dataset Size

	training	validation	test
MELD Dialogues	1039	114	280
MELD Utterances	9989	1109	2610
IEMOCAP Dialogues	120		31
IEMOCAP Utterances	5810		1623
IEMOCAP Sub-dialogues	897	101	224

We trained three types of FARCER models to analyze effective configurations comparatively:

- **Base model**: The modality projection layer is a single linear layer. Only the linear layer is trainable, and the LLM and ViT parameters are fixed.
- **Three-layer model**: The modality projection layer consists of three linear layers. The other components remain frozen.
- **Fine-tuned model**: LLaMA3-Instruct is fine-tuned while training a single modality projection layer. The ViT encoder is kept frozen.

By the default behavior, the models generate a sentence as a response, making it difficult to calculate proper contrastive loss between a response and a target label. To effectively train the models, we provided 7-shot prompting to constrain the output. We created ten prompts with randomly selected 7-shot examples to generalize the model performance.

Fig. 2 shows an input example. The prompt first specifies rules regarding response generation. The few-shot window places special tokens to make LLaMA3-Instruct understand its in-task role. Finally, the formatted input follows the prompt. For a given dialogue, the goal of our training is to minimize the mean cross-entropy loss between the predicted logits $l_p \in \mathbb{R}^{T\times K}$ and the one-hot target label vector $l_t \in \mathbb{R}^{T\times K}$:

$$\mathcal{L}(l_p, l_t) = -\frac{1}{\sum_{i=0}^{T} w_{l_{ti}}} \sum_{i=0}^{T} w_{l_{ti}} \log softmax(l_{pi}) \cdot l_{ti} \tag{7}$$

, where $w_{l_{ti}}$ is the weight assigned to the token l_{ti}.

Since the population of target labels is disproportionate in MELD and IEMOCAP, we calculated the balanced weight for the sub-word tokens of the labels with the *sklearn.compute_class_weight* function. We used the weights to compute the training loss. Table II outlines the label distributions in each dataset.

During the training process, the models performed multiple forward passes until the number of predicted tokens reached that of corresponding target labels. We saved the model parameters with the best-weighted F1 score in the validation data and used them in the testing process.

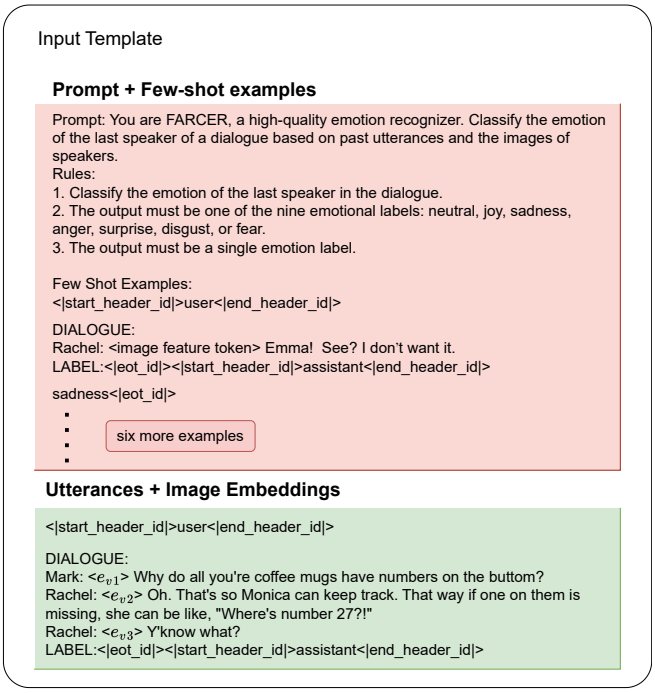

Fig. 2: Input Template

TABLE II: Label Distribution in Training Data

	MELD	IEMOCAP
neutral	45.38%	22.38%
joy	18.98%	-
surprise	11.46%	1.13%
anger	10.69%	13.66%
sadness	7.61%	11.81%
disgust	3.37%	0.18 %
fear	2.51%	0.62 %
other	-	0.97 %
frustration	-	24.49%
excitement	-	13.48%
happiness	-	10.57%

In IEMOCAP, each of the "xxx" labels, on which annotators did not agree, were converted into a randomly selected label among emotion candidates of the corresponding utterance.

D. Implementation Details

We used AdamW [23] as an optimizer and a cosine learning rate scheduler. For LoRA fine-tuning, the target modules were the query, key, and value projection layers of the LLM. We assigned the modality projection layer as a module to save. The batch size was set to 1 because of computational limitations. Table III outlines the hyperparameters in detail.

IV. Experiment

A. Benchmarks

We use the following corpora as benchmarks to evaluate the performance of our models.

TABLE III: Hyperparameters

	Base	Three-layer	Fine-tuned
Epoch	20	20	10
Learning Rate	1e-4	2e-4	1.5e-4 / 2e-4
Activatoin Function	-	GELU	-
Dropout	-	0.05	0.1
LoRA r	-	-	8
LoRA α	-	-	16

- **MELD** test dataset contains 280 dialogue instances and 2610 utterances in total.
- **IEMOCAP** test dataset comprises 31 two-party dialogues. We divided the dialogues into sub-dialogues of 5-26 utterances.

B. Baselines

To evaluate our models comparatively, we introduced five different models as baselines. We selected both multi-modal and text-only Transformer-based models.

- **LLaMA3-8B-Instruct**: We tested the original text-only LLaMA3-Instruct to measure how employing the FARCER architecture improves its performance on the ERC tasks. LLaMA3-Instruct is the LLaMA3 instruction-tuned specifically for dialogic purposes. The LLaMA series has achieved exceptional performance and reduction in the parameter size, although the training corpus only consists of open-source data.
- **MPT-HCL** [24] is a multi-modal ERC model based on Bi-LSTM, Relatoinal Graph Convolutional Networks (RGCN) [27], and the Transformer architecture. MPT-HCL extracts contextual features from Bi-LSTM, speaker dependencies, and contextual information using *Speaker-aware RGCN* and *Context-aware RGCN*, respectively. This model also utilizes *Multimodal Prompt Transformer* to fuse features of different modalities.
- **AccWR** [25] is a uni-modal LM-based architecture that aggregates the contextual representations near the target utterance and feeds it to Ro-BERTa [28] to maximize its inference ability.
- **SDT** [26] uses the cross-attention mechanism to aggregate multi-modal data. The intra-modal and inter-modal steps retrieve interactions within and across different modalities.
- **DialogueLLM** [13] is a LLaMA-powered model that infers emotions based on past utterances and video descriptions. DialogueLLM adopts a different approach from ours by taking visual information as text. Still, this model compares to ours since it utilizes visual data as input.

C. Experimental Setup

In the testing process, the **Base FARCER model**, **Three-layer FARCER model**, **Fine-tuned FARCER model**, and LLaMA3-8B-Instruct predicted the last speaker's emotion in

TABLE IV: Test Results on MELD and IEMOCAP

Table IV shows the F1 scores of each emotion label along with the overall weighted F1 and accuracy scores. We gave the models with "PL" with few-shot prompts where the past utterances are annotated with emotion labels. The "PL" models, except for LLaMA3, are trained on dialogues with the label annotations. The **bold scores** are the highest scores of each section, and the red values represent the best scores across the table.

MELD

Models	Neutral	Joy	Surprise	Anger	Sadness	Disgust	Fear	Accuracy	w-F1
MPT-HCL [24]	77.82	60.18	58.26	59.25	45.15	30.36	21.52	65.86	65.02
AccWR [25]	-	-	-	-	-	-	-	64.99	59.40
SDT [26]	80.19	64.29	59.07	54.33	43.69	28.78	17.88	67.55	66.60
DialogueLLM [13]	-	-	-	-	-	-	-	71.96	71.90
LLaMA3-Instruct - 3 shot	40.68	45.45	17.50	37.04	25.0	15.38	0.0	35.0	36.03
LLaMA3-Instruct - 7 shot	41.99	48.12	17.86	32.76	21.05	11.76	0.0	35.36	36.26
Base FARCER - 3 shot	70.59	54.24	36.67	51.52	34.15	30.77	0.0	56.79	56.41
Base FARCER - 7 shot	70.80	57.14	35.71	**54.84**	34.15	16.67	**25.0**	58.21	57.38
Three-layer FARCER - 3 shot	73.36	60.71	40.82	51.35	19.05	**42.86**	0.0	58.93	58.44
Three-layer FARCER - 7 shot	75.27	58.82	44.0	53.12	19.51	30.77	0.0	60.36	59.19
Fine-tuned FARCER - 3 shot	79.86	**66.04**	43.64	54.24	**34.29**	28.57	**25.0**	**66.07**	**64.37**
Fine-tuned FARCER - 7 shot	**80.88**	64.86	**45.28**	50.79	33.33	25.0	22.22	65.36	64.17
LLaMA3-Instruct PL - 3 shot	60.09	44.66	31.37	55.32	25.53	31.58	0.00	48.57	48.92
LLaMA3-Instruct PL - 7 shot	59.59	43.81	26.67	56.47	33.33	22.22	0.00	48.57	48.55
Base FARCER PL - 3 shot	79.09	63.37	46.43	59.46	34.04	18.18	0.00	64.29	63.63
Base FARCER PL - 7 shot	77.98	59.79	48.00	60.27	36.36	18.18	0.00	64.29	62.89
Three-layer FARCER PL - 3 shot	76.87	60.95	36.00	55.38	35.09	28.57	0.00	62.86	61.05
Three-layer FARCER PL - 7 shot	76.71	52.53	51.72	58.18	42.11	33.33	0.00	63.93	61.98
Fine-tuned FARCER PL - 3 shot	82.63	**69.16**	**56.25**	61.29	41.86	33.33	**28.57**	69.29	69.18
Fine-tuned FARCER PL - 7 shot	**83.02**	69.09	55.74	**64.41**	**50.00**	**40.00**	0.00	**70.00**	**70.01**

IEMOCAP

Models	Frustration	Neutral	Anger	Sadness	Excitement	Happiness	Accuracy	w-F1
MPT-HCL [24]	69.09	66.75	69.96	85.97	74.06	58.13	72.83	72.51
AccWR [25]	-	-	-	-	-	-	-	64.99
SDT [26]	68.68	74.62	69.73	81.84	80.17	66.19	73.95	74.08
DialogueLLM [13]	-	-	-	-	-	-	70.62	69.93
LLaMA3-Instruct - 3 shot	51.91	39.44	52.0	50.0	50.0	34.15	44.0	45.39
LLaMA3-Instruct - 7 shot	46.3	30.3	46.67	58.23	60.32	35.29	43.56	44.52
Base FARCER - 3 shot	49.57	45.98	53.85	61.97	51.52	13.33	47.11	47.35
Base FARCER - 7 shot	53.45	48.89	51.16	66.67	54.79	20.0	50.22	50.51
Three-layer FARCER - 3 shot	57.66	**52.73**	29.41	64.29	59.26	32.26	50.67	50.39
Three-layer FARCER - 7 shot	59.38	38.46	48.78	65.17	**64.41**	**52.63**	53.78	53.05
Fine-tuned FARCER - 3 shot	**60.0**	47.06	58.82	64.62	60.0	35.29	**54.67**	54.44
Fine-tuned FARCER - 7 shot	55.05	49.54	**59.65**	**67.69**	60.0	38.89	**54.67**	**54.45**
LLaMA3-Instruct PL - 3 shot	68.42	49.35	59.65	69.57	66.67	**66.67**	**60.89**	**61.25**
LLaMA3-Instruct PL - 7 shot	61.11	48.65	52.63	**71.79**	**77.78**	63.83	**60.89**	60.18
Base FARCER PL - 3 shot	48.08	52.63	41.03	64.2	53.85	42.86	49.78	49.57
Base FARCER PL - 7 shot	61.54	61.86	50.0	70.89	73.53	52.63	60.0	60.76
Three-layer FARCER PL - 3 shot	42.35	50.42	58.33	64.65	34.15	54.55	50.22	47.93
Three-layer FARCER PL - 7 shot	55.56	54.74	**60.87**	68.04	48.15	50.0	54.67	54.51
Fine-tuned FARCER PL - 3 shot	51.85	42.86	51.06	66.67	64.52	50.0	52.44	53.32
Fine-tuned FARCER PL - 7 shot	56.41	45.83	51.16	67.65	64.52	53.66	54.67	55.49

each dialogue. We prepared 3-shot and 7-shot prompts to investigate the effect of few-shot prompting. We used greedy decoding as a sampling method F_{sp}, namely:

$$\begin{aligned} label &= F_{sp}(l_p) \\ &= \underset{dim=-1}{\arg\max}(l_p) \end{aligned} \tag{8}$$

Table IV outlines the test results.

V. Results

A. Effect of FARCER Architecture

Employing the FARCER architecture significantly improved the performance of LLaMA3-Instruct in MELD. The weighted F1 score increased from 36.26 to 57.38 for the base FARCER model and 59.19 for the three-layer model. The FARCER architecture successfully improved the ERC performance on most emotion labels. Nonetheless, identifying fear remained a highly challenging task for the models without fine-tuning. This result implies the limitation of our model design. Our framework has difficulty ameliorating the core LLM's critical weakness as it augments the reasoning ability by adding external linear layers without updating the parameters of the core LLM. In contrast, the LoRA fine-tuned FARCER marked the stable scores on the "fear" label, overcoming the LLaMA3-Instruct's fundamental inability by attaching additional trainable parameters to the internal attention layers.

We observed a similar improvement in IEMOCAP. The weighted F1 scores increased by 5.99 points for the base model and 8.53 points for three-layer FARCER with 7-shot prompting. The difference in improvement rates between MELD and IEMOCAP is attributed to the discrepancy between the ViT encoder and the nature of the IEMOCAP dataset. While the ViT encoder is designed to classify images across seven emotions: {neutral, happiness, surprise, anger, sadness, disgust, fear}, IEMOCAP introduces three additional emotions: {frustration, excitement, and other}. Fig. 3 and 4 are the confusion matrices of Base FARCER and LLaMA3-Instruct with 7-shot prompting on IEMOCAP. These figures show that our vision-language models are more likely to misclassify "anger" utterances as "frustration" and "happiness" utterances as "excitement" or "neutral" than the text-only LLaMA3-Instruct.

The greater clarity of this pattern in the vision-language models indicates that the vision encoder mapped facial expressions annotated with the new labels onto one of the seven standard emotions, and the CLS tokens from ViT obscure the differences in those similar emotion features. In contrast, this observation also explains the high improvement rate in MELD. Since the labels classified by the vision encoder match the emotion labels of MELD, the CLS features were properly transformed and reinforced the inference of the LLM. Therefore, we speculate we can maximize the performance of FARCER by applying it to ERC with the seven distinct labels or by fine-tuning the vision encoder for more diverse emotion labels.

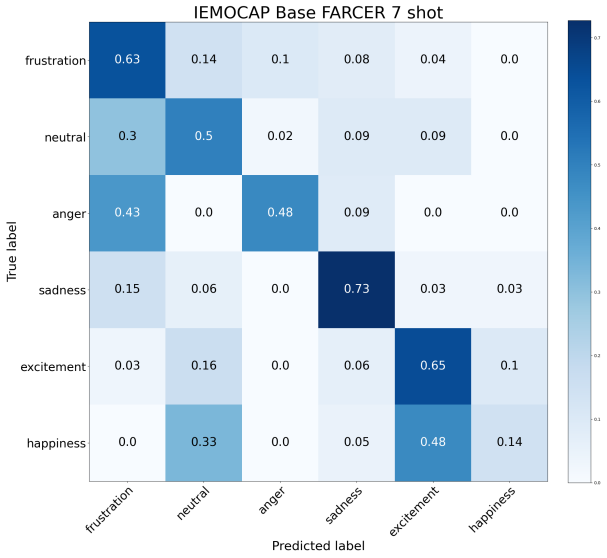

Fig. 3: Confusion Matrix for Base FARCER

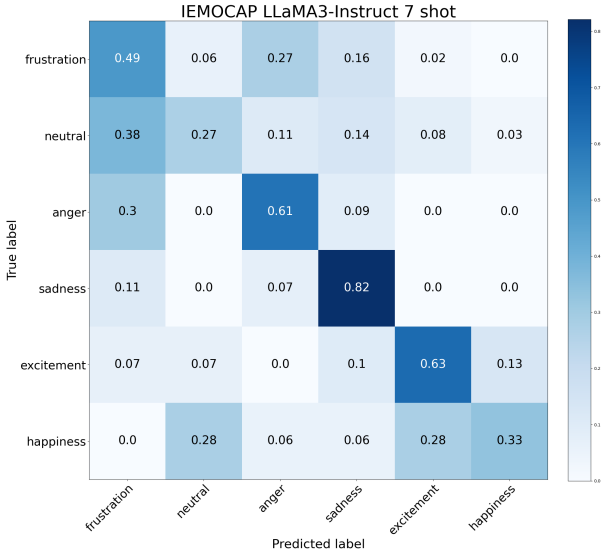

Fig. 4: Confusion Matrix for LLaMA

B. Modality Projection Layer Size

Increasing the number of modality projection layers also improved the overall F1 score in MELD from 57.38 to 59.19 with the 7-shot prompt. This phenomenon is explained by the neural networks' capability to learn more complicated and hierarchical representations from input visual features [29]. The weighted F1 score for the three-layer model with 7-shot prompting increased by 4.47 points in the "neutral" emotion, 8.29 points in "surprise", and 14.1 points in "disgust" from the single-layer counterpart.

In the same principle, the three-layer FARCER achieved

higher scores in IEMOCAP. Fig. 5 displays the confusion matrix of the three-layer model in IEMOCAP. By training the deeper modality projection layer, the model marked higher accuracy in classifying ambiguous labels such as "happiness", "excitement," and "frustration" than the base model.

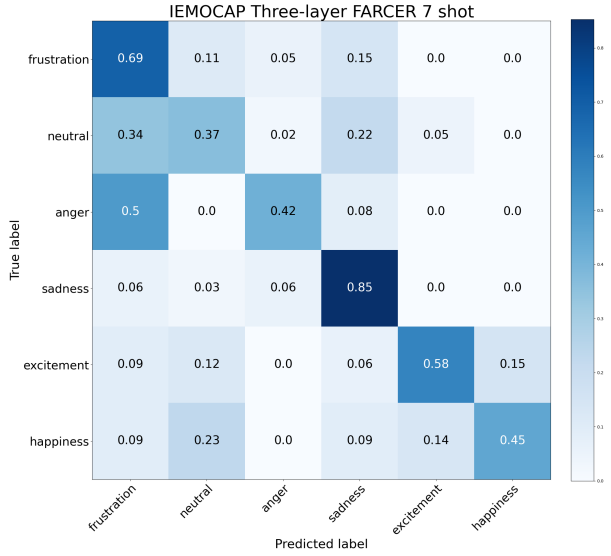

Fig. 5: Confusion Matrix for Three-layer FARCER

C. Ablation Study on Past Label Annotations

We also tested our model trained on dialogues where utterances were annotated with emotion labels. Our motivation was to analyze the effect of integrating the contextual flow of emotions as input features. The results are shown in the bottom sections with "PL" (past labels).

In MELD, the overall F1 scores improved from the counterpart models; on average, the scores increased by 4.8 points. The fine-tuned FARCER model with PLs achieved the weighted F1 score of 70.01, a SOTA-line performance. Notably, the base FARCER and three-layer models reached the same weighted F1 score and accuracy level when the past emotion labels were provided. The base single-layer model improved by 6.37 points by introducing PLs, whereas the three-layer model improved by only 2.7 on average. Given the observation above that more modality projection layers underpin the models' visual abstraction ability, the result suggests that providing emotional context for dialogues supports more accurate reasoning of the core LLM, filling the gap between single-layer models and those with more layers.

In IEMOCAP, LLaMA3-Instruct with PLs improved the F1 scores for all the emotion labels. This result supports our argument that giving contextual emotion clues improves the inference of the core LLM. In FARCER, the overall improvement was less conspicuous. The recall scores for "anger" did not increase, especially for the models with a single modality projection layer. On average, the recall scores changed from 50.67 to 46.67, suggesting that the false prediction of "anger" as "frustration" was unimproved. The smaller size of the modality projection layer had difficulty distinguishing a CLS token for the two similar features, especially due to the unbalanced label availability in the training data, as Table II shows. On the other hand, all the FARCER models successfully increased the recall scores for "happiness", which was often misclassified as "excitement", by 22.46 on average with the more balanced label size. The fundamental incongruency between the vision encoder and the dataset elicited this test result.

D. Few-shot Prompting

Few-shot prompting is reported to improve the reasoning of language models [30]. At the same time, Zhang et al. [13] show that providing excessive examples deteriorates the LLM's performance on ERC because of information redundancy and long input sequences. Our research observed that the weighted F1 scores of the FARCER models increased by 0.44 in MELD and 4.30 in IEMOCAP on average using 7-shot prompting.

VI. Conclusion

In this work, we delved into the effectiveness of our context-focused, fully auto-regressive mechanism with causal LM for ERC. The test results showed that introducing our simple FARCER architecture significantly increased F1 scores under suitable settings, improving the LLM's reasoning in ERC. Moreover, our research indicated that annotating emotion labels to past utterances elevates the LLM's contextual understanding of emotions. We hope this paper provided overarching analysis and critical insight into applying conversational LLMs to ERC tasks.

VII. Future Work

Further research should be done to expand the community's knowledge on effectively implementing conversational LLMs for ERC. One possible research topic is testing the performance of FARCER architecture on more complicated and interactive tasks, such as integrating ERC and emotional response generation. This type of research will clarify the real-life applicability of our proposed framework. Another fundamental interest is in data accessibility to negative emotions. In contemporary social settings, people are less often exposed to so-called negative emotions compared to neutral states or joy. Furthermore, it is socially difficult to express those negative emotions even amid uncomfortable settings. This social-behavioral tendency is also reflected in the lower ERC dataset availability of negative emotions as TableII show. The consequence is ERC models' relatively poor performance in recognizing a human's disgust, anger, or fear. Therefore, collecting more data on those emotions will help ERC models improve their emotion recognition ability.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and

I. Polosukhin, "Attention Is All You Need," Aug. 2023, arXiv:1706.03762 [cs]. [Online]. Available: http://arxiv.org/abs/1706.03762

[2] e. a. Long Ouyang, Jeff Wu, "Training language models to follow instructions with human feedback," 2022. [Online]. Available: https://arxiv.org/abs/2203.02155

[3] e. a. Gemini Team, "Gemini: A family of highly capable multimodal models," 2024. [Online]. Available: https://arxiv.org/abs/2312.11805

[4] Anthropic, "Claude," accessed on July 22nd, 2024. [Online]. Available: https://claude.ai

[5] e. a. Hugo Touvron, Thibaut Lavril, "LLaMA: Open and Efficient Foundation Language Models," Feb. 2023, arXiv:2302.13971 [cs]. [Online]. Available: http://arxiv.org/abs/2302.13971

[6] e. a. Hugo Touvron, Louis Martin, "Llama 2: Open Foundation and Fine-Tuned Chat Models," Jul. 2023, arXiv:2307.09288 [cs]. [Online]. Available: http://arxiv.org/abs/2307.09288

[7] H. Laurençon, L. Tronchon, M. Cord, and V. Sanh, "What matters when building vision-language models?" May 2024, arXiv:2405.02246 [cs]. [Online]. Available: http://arxiv.org/abs/2405.02246

[8] e. a. Jean-Baptiste Alayrac, Jeff Donahue, "Flamingo: a visual language model for few-shot learning," 2022. [Online]. Available: https://arxiv.org/abs/2204.14198

[9] J. Y. Koh, R. Salakhutdinov, and D. Fried, "Grounding Language Models to Images for Multimodal Inputs and Outputs," Jun. 2023, arXiv:2301.13823 [cs]. [Online]. Available: http://arxiv.org/abs/2301.13823

[10] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, "VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text," Dec. 2021, arXiv:2104.11178 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2104.11178

[11] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models," Jun. 2024, arXiv:2306.05424 [cs]. [Online]. Available: http://arxiv.org/abs/2306.05424

[12] C. Lyu, M. Wu, L. Wang, X. Huang, B. Liu, Z. Du, S. Shi, and Z. Tu, "Macaw-LLM: Multi-Modal Language Modeling with Image, Audio, Video, and Text Integration," Jun. 2023, arXiv:2306.09093 [cs]. [Online]. Available: http://arxiv.org/abs/2306.09093

[13] Y. Zhang, M. Wang, Y. Wu, P. Tiwari, Q. Li, B. Wang, and J. Qin, "DialogueLLM: Context and Emotion Knowledge-Tuned Large Language Models for Emotion Recognition in Conversations," Jan. 2024, arXiv:2310.11374 [cs]. [Online]. Available: http://arxiv.org/abs/2310.11374

[14] C. Liu, Z. Xie, S. Zhao, J. Zhou, T. Xu, M. Li, and E. Chen, "Speak From Heart: An Emotion-Guided LLM-Based Multimodal Method for Emotional Dialogue Generation," in *Proceedings of the 2024 International Conference on Multimedia Retrieval*, ser. ICMR '24. New York, NY, USA: Association for Computing Machinery, Jun. 2024, pp. 533–542. [Online]. Available: https://doi.org/10.1145/3652583.3658104

[15] S. Lei, G. Dong, X. Wang, K. Wang, and S. Wang, "InstructERC: Reforming Emotion Recognition in Conversation with a Retrieval Multi-task LLMs Framework," Mar. 2024, arXiv:2309.11911 [cs]. [Online]. Available: http://arxiv.org/abs/2309.11911

[16] Y. Fu, "Ckerc : Joint large language models with commonsense knowledge for emotion recognition in conversation," 2024. [Online]. Available: https://arxiv.org/abs/2403.07260

[17] Todor Pakov, "vit-face-expression (revision 78ed8d3)," 2024. [Online]. Available: https://huggingface.co/trpakov/vit-face-expression

[18] K. Ataallah, X. Shen, E. Abdelrahman, E. Sleiman, D. Zhu, J. Ding, and M. Elhoseiny, "MiniGPT4-Video: Advancing Multimodal LLMs for Video Understanding with Interleaved Visual-Textual Tokens," Apr. 2024, arXiv:2404.03413 [cs]. [Online]. Available: http://arxiv.org/abs/2404.03413

[19] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[20] Z. Cheng, Z.-Q. Cheng, J.-Y. He, J. Sun, K. Wang, Y. Lin, Z. Lian, X. Peng, and A. Hauptmann, "Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning," 2024. [Online]. Available: https://arxiv.org/abs/2406.11161

[21] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," 2019. [Online]. Available: https://arxiv.org/abs/1810.02508

[22] C. Busso, M. Bulut, C.-C. Lee, E. A. Kazemzadeh, E. M. Provost, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:11820063

[23] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019. [Online]. Available: https://arxiv.org/abs/1711.05101

[24] H. Ma, J. Wang, H. Lin, B. Zhang, Y. Zhang, and B. Xu, "A transformer-based model with self-distillation for multimodal emotion recognition in conversations," *IEEE Transactions on Multimedia*, pp. 1–13, 2023.

[25] X. Jieying, N. P. Minh, M. Blake, and N. Minh Le, "Accumulating word representations in multi-level context integration for erc task," in *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*, 2023, pp. 1–6.

[26] H. Ma, J. Wang, H. Lin, B. Zhang, Y. Zhang, and B. Xu, "A transformer-based model with self-distillation for multimodal emotion recognition in conversations," *IEEE Transactions on Multimedia*, pp. 1–13, 2023.

[27] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," 2017. [Online]. Available: https://arxiv.org/abs/1703.06103

[28] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[29] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–44, 05 2015.

[30] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020. [Online]. Available: https://arxiv.org/abs/2005.14165