GRACE: GRadient-based Active Learning with Curriculum Enhancement for Multimodal Sentiment Analysis Supplementary Materials

Xinyu Li*
MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China
Hefei, China
xinyli@mail.ustc.edu.cn

Wenqing Ye*
University of Science and Technology of China
Hefei, China
wenqy@mail.ustc.edu.cn

Yueyi Zhang
MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China
Hefei, China
zhyuey@ustc.edu.cn

Xiaoyan Sun†
MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China
Hefei, China
sunxiaoyan@ustc.edu.cn

A Dataset Descriptions

SIMSv2 [1] contains 4403 labeled instances from 145 video clips, with an average clip duration of 3.67s. The original videos are collected from 11 different scenarios, simulating real-world situations and including numerous instances with weaker textual dependencies. The sentiment annotations assigned to each instance in the dataset range from Strong Negative (-1, -0.8), Weak Negative (-0.6, -0.4, -0.2), Neutral (0), Weak Positive (0.2, 0.4, 0.6), and Strong Positive (0.8, 1). The data is split into three subsets: a training set (2722), a validation set (647), and a test set (1034). **CHERMA** [2] is a large Chinese multimodal emotion recognition dataset, containing 28,717 utterances mainly acquired from TV series. The samples are divided into training, validation, and test datasets using a 6:2:2 ratio. The annotations adhere to Ekman's system of emotion theory, encompassing seven categories: happiness, sadness, fear, anger, surprise, disgust, and neutrality.

B Data Pre-processing

The inputs of the model from the two public datasets are preprocessed feature sequences, and the detailed process is described as follows.

SIMSv2: The length of the text is fixed to 50 by padding or truncation, and then the features are obtained by a pre-trained Chinese BERT model with dimension 768. OpenSMILE is used to extract audio features, generating a feature sequence with a dimension of 25 and a length of 925. For visual data, the image stream is passed through OpenFace to extract 177-dimensional features, with a fixed length of 232.

CHERMA: The pre-trained Chinese BERT model is also used for the pre-processing of the text. The length of the raw data is padded to 78. By adding CLS and SEP tokens, the 768-dimensional text features of length 80 are obtained. The audio is pre-processed using the pre-trained wav2vec, which generates 768-dimensional feature sequences with the original data length. After cropping facial images using MTCNN, the visual data is passed through a pre-trained Resnet 18 to obtain a feature sequence with a length of 64 and a dimension of 512.

C Effect of Initial Pool and Budget

We conduct experiments on SIMSv2, varying the initial pools and budgets to investigate how these factors influence model performance. Figure 1 illustrates the scenario where both the initial pool and budget are set at 50, while Figure 2 showcases the case where the initial pool and budget are both 200. Across both figures, it is evident that GRACE consistently outperforms other methods, demonstrating its robustness and superiority under different initial pools and budgets. When the number of labeled samples reaches 800, the MAE metrics of GRACE in Figure 1 and Figure 2 are both 0.322, rounded to three decimal places. In addition, it is observed that the performance improvement of all methods in Figure 1 is small at the early stages. This can be attributed to the extremely limited initial data, leading to a cold start failure and subsequently trapping the network in a local optimum. As more data becomes available, the model gradually commences a normal learning process. On the other hand, Figure 2 shows a more significant performance improvement as the active learning cycle increases, due to the larger size of the initial pool. However, the performance of Figure 2 at 800 samples is slightly lower than the experiment in Sec. 4.2 of the main paper at the same data volume. This is because of the limited number of active cycles. Overall, these experiments highlight the importance of the initial pool and budget.

D Emotion-wise Comparison

To further analyze performance on the classification task over SIMSv2, we report the overall and emotion-wise F1 scores for the compared methods in Table 1. Our approach outperforms other methods in terms of the overall F1 score, once again demonstrating

*Both authors contributed equally to this research.

†Corresponding author.

Table 1: Experimental results of F1 score on CHERMA. The results are reported when the number of selected samples reaches 5000. The best results are highlighted in bold.

Methods	anger	disgust	fear	happy	neutral	sad	surprise	overall
Full data	75.08	43.72	66.52	77.94	66.55	79.37	65.08	69.22
BALD	72.22	20.09	44.43	76.01	63.14	72.97	49.08	60.53
BADGE	**72.54**	13.80	40.11	75.89	63.52	74.03	52.85	60.13
BMMAL	70.07	32.71	**56.15**	**76.41**	63.49	73.26	56.47	63.53
Random	70.38	14.11	40.88	75.49	61.77	72.27	45.54	58.37
GRACE	71.24	**35.39**	52.92	75.68	**64.03**	**75.87**	**57.91**	**64.11**

Table 2: Experimental results of multiple metrics on SIMSv2 using LFMIM. The results are reported when the number of selected samples reaches 800. The best results are highlighted in bold.

Methods	MAE (↓)	Acc-2 (↑)	Acc-3 (↑)	Acc-5 (↑)	F1 score (↑)	Corr (↑)
Full data	0.258	83.46	77.47	61.99	83.55	76.78
CoreSet	0.311	81.33	75.53	55.13	81.42	70.86
GCNAL	0.297	81.53	76.21	56.38	81.64	72.29
DBAL	0.312	81.82	75.15	52.90	81.93	72.17
Random	0.318	80.95	72.44	51.84	81.05	69.44
GRACE	**0.289**	**82.11**	**76.60**	**58.80**	**82.16**	**73.56**

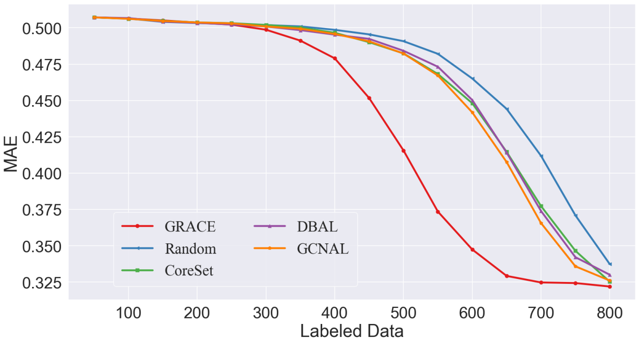

Figure 1: Model performance comparison on SIMSv2. The initial pool contains 50 samples and increases to 800 by fifteen AL cycles. A smaller MAE score indicates better performance.

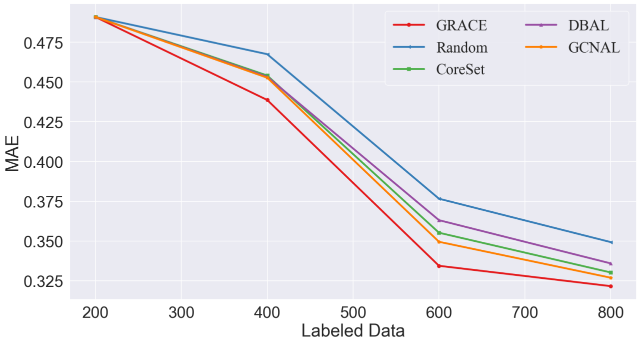

Figure 2: Model performance comparison on SIMSv2. The initial pool contains 200 samples and increases to 800 by three AL cycles. A smaller MAE score indicates better performance.

its superiority. We also achieve the best performance across multiple emotions, particularly with a significant margin surpassing other methods in terms of emotion disgust, one of the most difficult categories to recognize. In emotion anger, fear, and happy, our performance does not reach the optimal level, but the gap is quite small. This might be due to the fact that GRACE prematurely selects difficult samples with emotions like disgust, while the learning of simpler samples is not sufficient. Therefore, we can introduce adaptive adjustment strategies of the curriculum factor in future studies. It is worth noting that the active learning method BMMAL, which also considers multimodal properties, performs well across multiple metrics, far exceeding other active learning methods designed for unimodal tasks.

E Robustness to Task Model

In order to verify the robustness of GRACE to the task model, we choose LFMIM [2], a transformer-based architecture, as the task model for comparison on SIMSv2. The LFMIM comprises three unimodal transformers and a multimodal transformer. The outputs of each layer of the unimodal transformers are concatenated and used as the input for each layer of the multimodal transformer. As can be seen from Table 2, all methods show superior performance compared to previous experiments shown in Sec. 4.2 due to the strong task model. Compared with the full data, the difference in MAE of GRACE is 0.031, and the difference in other metrics is also less than 3.3%. Overall, our method still outperforms other active learning methods on the LFMIM model across all metrics, demonstrating the insensitivity of GRACE to the task model.

F Algorithm Procedure

The full procedure of our GRACE is depicted in Algorithm 1. Initially, a small number of randomly selected labeled samples forms the labeled dataset L, while the remaining unlabeled data constitutes the unlabeled dataset U. A curriculum factor α and its corresponding curriculum decay α_d are introduced for curriculum learning. Before the phase of active learning, the network $F(\cdot;\theta)$ is initially trained using L. For each active learning cycle, we begin by computing both the unimodal and multimodal gradients $\mathbf{G}$ for each sample in U. Subsequently, we iterate over U and calculate the informativeness and easiness scores for each unlabeled data x^p. Due to the representativeness criterion considering the overall distribution of data, we compute the sum of distances between x^p and all other x^q in the unlabeled dataset, excluding x^p itself. After normalization, the final score is computed. Next, we select the top K samples from U based on their final scores and obtain labels for them through Oracle. The datasets L, U, and the curriculum factor α are then updated. Finally, we train the task model $F(\cdot;\theta)$ using the current labeled dataset L. This process is repeated until the termination condition is met.

Algorithm 1: The Sketch of GRACE Method

```
Input: initial labeled dataset $L$ and unlabeled dataset $U$,
       curriculum factor $\alpha$ and decay $\alpha_d$.
Initialize the task model $F(\cdot;\theta)$ using the labeled dataset $L$.
for $i$ to $query\_rounds$ do
    Get the gradients $\mathbf{G} = \{\mathbf{g}_t, \mathbf{g}_a, \mathbf{g}_v, \mathbf{g}_{m_t}, \mathbf{g}_{m_a}, \mathbf{g}_{m_v}\}$ of the
      unlabeled dataset $U$.
    for $x^p \in U$ do
        Calculate $\hat{s}_i(x^p)$, $\hat{s}_e(x^p)$ using $\mathbf{G}(x^p)$.        Eq.(4,6)
        Initialize $\hat{s}_r(x^p) = 0$.
        for $x^q \in U \backslash x^p$ do
            Accumulate $\hat{s}_r(x^p)$ by the distance between $x^p$
              and $x^q$ using $\mathbf{G}(x^p)$ and $\mathbf{G}(x^q)$.        Eq.(5)
        end
        Normalize to obtain $s_i(x^p), s_r(x^p), s_e(x^p)$.
        $s(x^p) = s_e(x^p) \cdot \alpha + s_i(x^p) \cdot s_r(x^p)$.        Eq.(7)
    end
    $Q \longleftarrow$ Query top $K$ samples from $U$ based on $s$.
    Oracle labeling $Y^{new}$ for samples in $Q$.
    Update dataset: $L \longleftarrow L \bigcup (Q, Y^{new}), U \longleftarrow U \backslash Q$.
    Update curriculum factor: $\alpha \longleftarrow \alpha - \alpha_d$.
    Train the task model $F(\cdot;\theta)$ using the labeled dataset $L$.
end
```

G Training Efficiency

We compare training efficiency across methods in two aspects. 1) Querying time. The time for one AL cycle and the overall training process using an RTX 3090 GPU are shown in Table 3. For small datasets (like SIMSv2), the querying time for AL is negligible compared to overall training. For large datasets (like CHERMA), GRACE demonstrates advantages in efficiency. 2) Learning speed. Taking Figure 4 in the paper as an example, GRACE reaches MAE 0.35 requiring 200 fewer labeled data than other methods, and thus incurs 2X fewer model training and querying rounds. This indicates GRACE achieves the same performance faster. In practical applications, we can collect more unlabeled data, which is less costly than labeling, to enhance model capacity.

Methods	Time		Methods	Time	
	querying	overall		querying	overall
CoreSet	5.09s	15m 46s	BALD	211.92s	35m 48s
GCNAL	16.04s	16m 33s	BADGE	256.62s	42m 52s
DBAL	5.01s	15m 31s	BMMAL	301.96s	53m 13s
GRACE	25.73s	18m 13s	GRACE	161.24s	28m 13s

Table 3: Time taken on SIMSv2 and CHERMA datasets.

H Further Study on Representativeness

We replace the calculation method with the sum of distances between representations of the four modalities (GRACE-rp), and the distance between the fusion modality representations (GRACE-rp(m)). The results are shown in Row 1 & 2 of Table 4. The original GRACE remains the best in all metrics. This is because representations are the mappings of data in feature space, while gradient embeddings reflect the underlying parameter update directions and magnitudes of the network. Therefore, gradient embeddings can effectively guide the AL process.

Methods	MAE	Acc-2	Acc-3	Acc-5	F1 score	Corr
GRACE-rp	0.326	79.37	70.76	48.87	79.49	69.31
GRACE-rp(m)	0.329	79.05	70.73	48.55	79.17	68.32
$\beta = 1, \gamma = 1$	0.322	80.46	71.56	49.03	80.57	69.85
$\beta = 10, \gamma = 1$	0.325	80.08	70.56	48.42	80.30	70.12
$\beta = 1, \gamma = 10$	0.329	79.66	70.31	48.48	79.78	69.66
GRACE	**0.319**	**81.17**	**72.86**	**50.52**	**81.26**	**70.75**

Table 4: Additional experimental results on SIMSv2.

I Balance between Criteria

We have proposed a curriculum factor to balance sample difficulty and active value. The active value is the product of informativeness and representativeness since we aim to select samples with both scores being high. However, the easiness doesn't always need to be high and should decrease with training, so the additive relationship is adopted. To further study the balance, we propose a new additive relation as a comparison: $s(x^p) = \alpha \cdot s_e(x^p) + \beta \cdot s_i(x^p) + \gamma \cdot s_r(x^p)$. The setting of α is consistent with the paper, and the experiments of β and γ are shown in Row 3-5 of Table 4. The original GRACE remains the best in all metrics. Using the sum leads to reduced performance by selecting unexpected samples, such as those with s_i being extremely high yet s_r extremely low.

References

[1] Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. 2022. Make acoustic and visual cues matter: CH-SIMS v2. 0 dataset and AV-Mixup consistent module. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 247–258.

[2] Jun Sun, Shoukang Han, Yu-Ping Ruan, Xiaoning Zhang, Shu-Kai Zheng, Yulong Liu, Yuxin Huang, and Taihao Li. 2023. Layer-wise fusion with modality independence modeling for multi-modal emotion recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. 658–670.