

A COMMONSENSE KNOWLEDGE ENHANCED NETWORK WITH RETROSPECTIVE LOSS FOR EMOTION RECOGNITION IN SPOKEN DIALOG

Yunhe Xie Chengjie Sun Zhenzhou Ji

Faculty of Computing, Harbin Institute of Technology, China
{xieyh, sunchengjie, jizhenzhou}@hit.edu.cn

ABSTRACT

The recent surges in the open conversational data caused Emotion Recognition in Spoken Dialog (ERSD) to gain much attention. However, the existing ERSD datasets' scale limits the model's complete reasoning. Moreover, the artificial dialogue agent is ideally able to reference past dialogue experiences. This paper proposes a Commonsense Knowledge Enhanced Network with a retrospective loss, namely CKE-Net, to hierarchically perform dialog modeling, external knowledge integration, and historical state retrospect. Specifically, we first adopt a transformer-based encoder to model context in multi-view by elaborating different mask matrices. Then, the graph attention network is used to introduce commonsense knowledge, which benefits the complex emotional reasoning. Finally, a retrospective loss is added to utilize the model's prior experience during training. Experiments on IEMOCAP and MELD datasets demonstrate that every designed module is consistently beneficial to the performance. Extensive experimental results show that our model outperforms the state-of-the-art models across the two benchmark datasets.

Index Terms— emotion recognition in spoken dialog, commonsense knowledge, retrospective loss

1. INTRODUCTION

Researchers have been trying to give machines the ability to recognize, interpret, and express emotions. Early research on emotion recognition focused on comprehending emotions in monologues [1]. The explosion of the open conversational data recently caused researchers to attach great importance to Emotion Recognition in Spoken Dialog (ERSD) [2, 3, 4]. ERSD aims to detect emotions from utterances in spoken dialog [5], which helps create empathy dialogue systems and improve the human-computer interaction experience [6].

Unlike the vanilla emotion recognition of sentences, the context of the dialog is crucial for ERSD. Therefore, recent works focused on the specific factors of the spoken dialog, such as the transactional dynamics [7], the temporality of the

This work was supported by the National Key R&D Program of China via grant 2020YFB1406902.


Case #1	
Utterance	Label
He's had <i>cancer</i> for a while.	Frustrated
<i>cancer</i> 	
illness animal death kill smoking disease breast lung	
Case #2	
Utterance	Label
#8 Simply step to line two A, ma'am, I'm sorry.	Frustrated
.....	
#16 I can't. That's why you need to go to line two A.	Frustrated

Fig. 1. Two cases from the IEMOCAP dataset.

speakers' turns [8], or the speaker-specific information [9]. In addition, some works also proved that utilizing multi-modal information (e.g., text & audio) could often achieve better performance [10, 11]. Gu et al. [12] designed a learnable mutual correlation factor in computing associations across different modalities. Wang et al. [13] adopted end-to-end translation models to mine the subtle correlation between modalities.

However, the existing ERSD datasets are small in scale [14, 15], limiting the model to form a complete chain of reasoning. Intuitively, *commonsense knowledge* and the *model's historical judgment* can be used as auxiliary information. As shown in Case #1 (Figure 1), if the relevant conceptual information of the keyword "cancer" like "death" and "kill" can be obtained as an aid, the model will be more inclined to make negative judgments like "Frustrated". As shown in Case #2, if the model can gain experience from the previous judgment of a similar utterance, it can also avoid misjudgments.

Thus, this paper proposes a hierarchical model from bottom to top to perform context modeling, knowledge enhancement, and historical state retrospect. Our contributions are summarized as follows: 1) We endow the model with the ability of secondary reasoning by utilizing the graph network to introduce the structured knowledge. 2) As far as we know, our work introduces the retrospective loss into the training of the actual task for the first time so that the model can gain

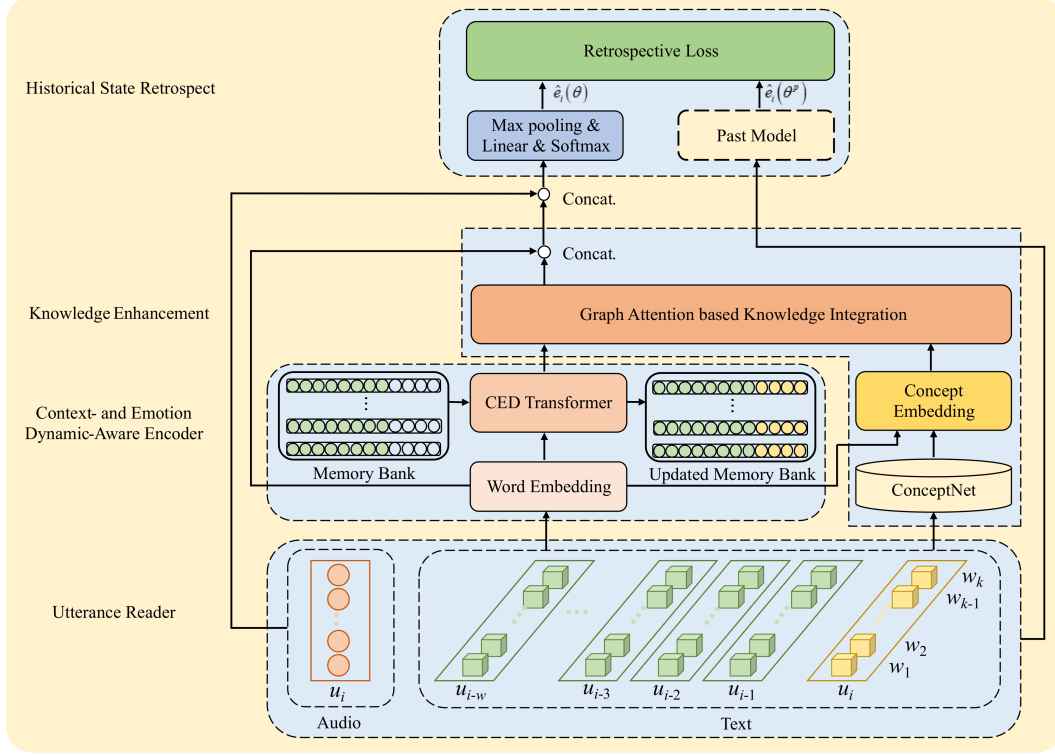


Fig. 2. The architecture of our proposed CKE-Net.

experience from itself. 3) Experiments on two ERSD datasets show that our model outperforms the state-of-the-art models.

2. TASK DEFINITION

Assume that there are a set of spoken dialogs $\mathcal{D} = \{\mathbf{D}_j\}_{j=1}^U$, where U is the number of the dialogs. In each dialog, $\mathbf{D}_j = \{(\mathbf{u}_i, s_i, e_i)\}_{i=1}^{N_i}$ is a sequence of N_i utterances, where the utterance u_i is spoken by the speaker $s_i \in \mathcal{S}$ with a predefined emotion $e_i \in \mathcal{C}$. All speakers compose the set \mathcal{S} and the set \mathcal{C} consists of all emotions, such as anger, happiness. ERSD’s goal is to develop a model to detect each new utterance with an emotion label from \mathcal{C} as accurately as possible. Our proposed framework of CKE-Net is provided in Figure 2.

3. METHODOLOGY

3.1. Utterance Reader Module

In this paper, we focus on a multi-modal scenario where the linguistic content and the acoustic characteristics are employed. For the i^{th} utterance $\mathbf{u}_i = \{w_k\}_{k=1}^{N_k}$ in \mathbf{D}_j , where N_k is the number of words, the acoustic feature of w_d is denoted as \mathbf{A}^d . For the textual feature of w_d , prepend the target utterance \mathbf{u}_i with a special token “[CLS]” and feed it to the embedding layer, then we get $\mathbf{h}_i^0 \in \mathbb{R}^{N_k \times D_h}$, where D_h denotes the input dimension of the encoder mentioned

in Subsection 3.2. \mathbf{h}_i^0 is treated as the input hidden state of the encoder’s first layer and also used in concept embedding layer in Subsection 3.3.

3.2. Context- and Emotion Dynamic-Aware Module

We adopt DialogXL [16] as the textual encoder, which introduces dialogue-aware self-attention to capture useful context- and emotion dynamic-aware information. DialogXL can be merged by a function $f(\cdot)$:

$$\mathbf{o}_i^l = f(\mathbf{m}^{l-1}, \mathbf{h}_i^{l-1}, \mathbf{s}). \quad (1)$$

where memory $\mathbf{m}^{l-1} \in \mathbb{R}^{D_m \times D_h}$, $\mathbf{h}_i^{l-1} \in \mathbb{R}^{N_k \times D_h}$, D_m is the memory length. The value of s_{ij} is set to $+\infty$ when the attention is masked, otherwise set to 0. We use an attention mechanism to learn the different impacts across each block and get $\hat{\mathbf{h}}_i^L$, where L is the layers’ number. At the same time, considering that the self-attention mechanism ignores the key information of the utterance sequence information in the conversation, we add a Recurrent Neural Network (RNN) to the last layer to strengthen the utterance position information and get the final output \mathbf{h}_i^L .

3.3. Knowledge Enhancement Module

In this module, we use ConceptNet [17] as the commonsense knowledge source. Each quadruple (concept1, relation, con-

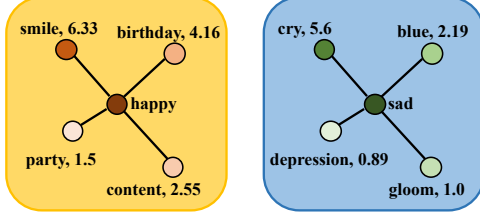


Fig. 3. Illustration of two sub-graphs extracted from the ConceptNet with the keyword “happy” and “sad”.

cept2, weight) denotes an assertion, where the weight is a confidence score assigned to the assertion. We ignore the relation in this paper and get an example assertion of ⟨happy, birthday⟩ with a confidence score of 4.16 as in Figure 3.

For each nonstop token w_d in \mathbf{u}_i , we retrieve a connected knowledge graph $\mathbf{G}(w_d)$ comprising its neighbor node:

$$\mathbf{G}(w_d) = [(n_1, g_1), (n_2, g_2), \dots, (n_{N_n}, g_{N_n})], \quad (2)$$

where n_i denotes the i^{th} connected concept of w_d , g_i denotes its corresponding confidence score and N_n denotes the number of concepts in $\mathbf{G}(w_d)$.

For w_d in \mathbf{u}_i and n_p in $\mathbf{G}(w_d)$, we obtain their token embedding from Subsection 3.1 and denote them as $\mathbf{h}_0^{w_d}$ and $\mathbf{h}_0^{n_p}$ respectively. A graph attention network is leveraged to obtain the knowledge representation \mathbf{K}^d belongs to w_d . We concatenate \mathbf{h}_{id}^L , \mathbf{K}^d and \mathbf{A}^d , finally get $\mathbf{c}_d \in \mathbb{R}^{3D_h}$. \mathbf{c}_d denotes the d^{th} entry of $\mathbf{C}^i \in \mathbb{R}^{3D_h \times N_k}$, which is the target utterance \mathbf{u}_i ’s knowledge-enriched multi-modal representation.

3.4. Retrospective Historical State Module

For the target utterance \mathbf{u}_i , we compute its utterance-level representation through max pooling:

$$\hat{\mathbf{C}}^i = \text{maxpooling}(\mathbf{W}_c \mathbf{C}^i + \mathbf{b}_c), \quad (3)$$

where $\mathbf{W}_c \in \mathbb{R}^{h_c \times 3D_h}$, $\mathbf{b}_c \in \mathbb{R}^{h_c \times N_k}$ are model parameters. We compute the final classification probabilities as follows:

$$\hat{\mathbf{y}}^i = \text{softmax}(\mathbf{W}_e \hat{\mathbf{C}}^i + \mathbf{b}_e), \quad (4)$$

where $\mathbf{W}_e \in \mathbb{R}^{h_e \times h_c}$, $\mathbf{b}_e \in \mathbb{R}^{h_e}$ are model parameters and h_e denotes the number of predefined emotions. We compute the loss of ERSD task in standard cross-entropy loss:

$$\text{loss}_{ersd} = - \sum_{j=1}^U \sum_{i=1}^{N_i} \sum_{e=1}^{h_e} y_e^i \log \hat{y}_e^i + (1 - y_e^i) (1 - \log \hat{y}_e^i), \quad (5)$$

where y_e^i denotes the ground truth emotion value of \mathbf{u}_i .

Equation (1) to (4) can be regarded as a mapping $\theta(\cdot)$ where $\theta(\mathbf{u}_i) = \hat{\mathbf{y}}^i$ and if we save the past model parameter,

Table 1. Split of Experimental Datasets.

Dataset	Dialog (Train/Val/Test)	Utter. (Train/Val/Test)
IEMOCAP	100/20/31	4810/1000/1523
MELD	1038/114/280	9989/1109/2610

we can get $\theta_p(\cdot)$. Inspired by [18], the proposed retrospective loss is derived by the above two mappings:

$$\text{loss}_{retro} = (\beta + 1) \|\theta(\mathbf{u}_i) - \mathbf{y}^i\| - \beta \|\theta(\mathbf{u}_i) - \theta_p(\mathbf{u}_i)\|, \quad (6)$$

where β is a hyperparameter and “ $\|\cdot\|$ ” denotes L_1 -norm.

The retrospective loss aims to reference the model’s historical judgment during training. Two additional hyperparameters warm-up period P and update frequency F are added where P refers to the start epoch for introducing such retrospective updates and F indicates the introduction frequency. The final loss function is computed as:

$$L = \text{loss}_{ersd} + \text{loss}_{retro}. \quad (7)$$

4. EXPERIMENT

4.1. Datasets

We use two benchmark datasets to evaluate our CKE-Net. IEMOCAP [19] is a dyadic conversation dataset with emotion labels neutral, happiness, sadness, anger, frustration, and excitement. MELD [20] is a multi-modal dataset collected from the TV show Friends. The labels are neutral, happiness, surprise, sadness, anger, disgust, and fear. The details about the training/validation/testing split are provided in Table 1.

4.2. Experimental Details

We tokenize and pre-process all datasets and ConceptNet using the XLNet tokenizer. For hyper-parameter setting, $D_h = 768$, $L=12$, $P=60$, $F=1$, h_c and D_m depends on the dataset. We employ the AdamW optimizer during training. For a fair comparison, we use the same initial modal features as in [9, 21]. For IEMOCAP and MELD, we use the weighted-F1 score as the metric. The results reported in our experiments are all based on an average of 5 random runs on the test set.

4.3. Baselines

We first make a comparison with the following baselines: AGHMN [3], bc-LSTM [5], DialogueRNN [6] and DDIN [21] use different variants of RNN to capture the information transfer of the dialog flow. ICON [8] and CMN [10] utilize the memory network to summarize the context. QMNN [4] and ConGCN [9] models dialog from the perspective of quantum theory and graph neural networks, respectively.

Table 2. Performance comparison of ours, baselines, and the state-of-the-art method on IEMOCAP and MELD. The baseline performance is obtained from [9, 21]. “–” means the original paper does not give the corresponding result. The numbers with * indicate that the improvement of our model over all baselines is statistically significant with $p < 0.05$ under the t-test.

Model	IEMOCAP	MELD
bc-LSTM	–	56.8
CMN	56.7	55.5
ICON	61.4	56.3
DialogueRNN	58	57
AGHMN	60.1	–
DDIN	61.7	–
QMNN	–	58
ConGCN	–	59.4
CKE-Net	66.5*	62.9*

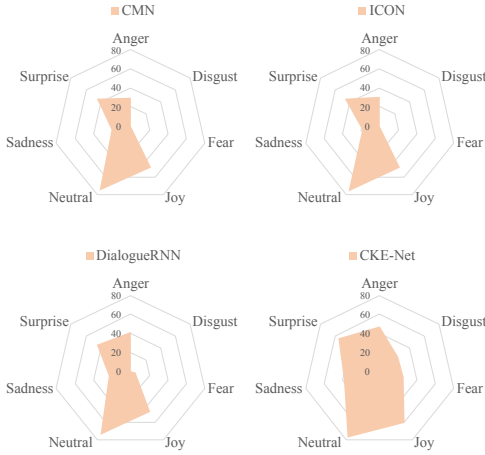


Fig. 4. Radar chart of emotion-specific result on MELD.

4.4. Overall Results

Overall results are listed in Table 2, and we can find early works such as CMN and ICON perform the worst, with a performance difference of 3.9%-5% compared with the current state-of-the-art model. The subsequent models design corresponding modules and thus achieve substantial performance improvements. Based on the full consideration of the context and emotional dynamics, our model significantly outperforms the state-of-the-art model by 4.8% on IEMOCAP and 3.5% on MELD by introducing external auxiliary knowledge and reference model historical judgment.

4.5. Emotion-Specific Results

We further draw an emotional radar chart based on the baseline and CKE-Net specific emotion results on MELD. From Figure 4, we can find that CKE-Net achieves a more balanced

Table 3. Ablation study results on IEMOCAP and MELD.

Method	IEMOCAP	MELD
CKE-Net	66.5	62.9
- retrospective loss	66.1(↓0.4)	62.7(↓0.2)
- knowledge	66.3(↓0.2)	62.4(↓0.5)

effect on all emotions. Considering that the MELD is a multi-party short conversation dataset, this once again proves the powerful reasoning ability of CKE-Net. Furthermore, the improvement of specific emotional performance does not come at the cost of the decline in other emotions’ performance.

4.6. Ablation Study

We perform an ablation study for our designed modules. For “- retrospective loss”, we only use the standard cross-entropy loss. “-commonsense” means knowledge enhancement is discarded. The results on IEMOCAP and MELD are shown in Table 3. We observe that the performance of our model drops on IEMOCAP and MELD with any of the components removed. For IEMOCAP, there is no considerable degradation in performance with the removal of knowledge enhancement and retrospective loss. IEMOCAP contains more utterances (around 50) for each dialog. Obviously, for the emotional recognition of long conversations, capturing the clues in the historical content of the conversation is far more critical than using external knowledge to assist reasoning, which has been achieved in our context- and emotion dynamic-aware module. For MELD, commonsense knowledge benefits more on the task. Since MELD has shorter utterances (around 9) for each dialog, commonsense knowledge can enrich semantics.

5. CONCLUSION

This paper proposes a Commonsense Knowledge Enhanced Network with a retrospective loss, namely CKE-Net, to solve emotion recognition in spoken dialog. Three modules were designed for conversation modeling, external knowledge integration, and model historical state retrospect. Extensive experiments were conducted on two ERSD benchmarks, and the results show that the proposed model outperforms all the baselines on the datasets. Experiment results indicate that the modules in CKE-Net are practical for an ERSD system. Furthermore, for the emotional recognition of long dialogs, capturing the clues in the historical content of the conversation is far more critical. In contrast, knowledge introduction is very precious in short conversations. In future work, we will focus on studying how to rationally combine each module’s inference results to make judgments closer to the actual situation.

6. REFERENCES

- [1] Nourah Alswaidan and Mohamed El Bachir Menai, “A survey of state-of-the-art approaches for emotion recognition in text,” *Knowledge & Information Systems*, vol. 62, no. 8, 2020.
- [2] Sung-Lin Yeh, Yun-Shao Lin, and Chi-Chun Lee, “A dialogical emotion decoder for speech emotion recognition in spoken dialog,” in *ICASSP 2020*. IEEE, 2020, pp. 6479–6483.
- [3] Wenxiang Jiao, Michael Lyu, and Irwin King, “Real-time emotion recognition via attention gated hierarchical memory network,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 8002–8009.
- [4] Qiuchi Li, Dimitris Gkoumas, Alessandro Sordoni, Jian-Yun Nie, and Massimo Melucci, “Quantum-inspired neural network for conversational emotion recognition,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 13270–13278.
- [5] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency, “Context-dependent sentiment analysis in user-generated videos,” in *Proceedings of the 55th annual meeting of the association for computational linguistics*, 2017, pp. 873–883.
- [6] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria, “Dialoguerrnn: An attentive rnn for emotion detection in conversations,” in *Proceedings AAAI Conference on Artificial Intelligence*, 2019, pp. 6818–6825.
- [7] Sung-Lin Yeh, Yun-Shao Lin, and Chi-Chun Lee, “An interaction-aware attention network for speech emotion recognition in spoken dialogs,” in *ICASSP 2019*, 2019, pp. 6685–6689.
- [8] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann, “Icon: Interactive conversational memory network for multimodal emotion detection,” in *EMNLP*, 2018.
- [9] Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou, “Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations,” in *IJCAI*, 2019, pp. 5415–5421.
- [10] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann, “Conversational memory network for emotion recognition in dyadic dialogue videos,” in *NAACL-HLT*, 2018.
- [11] Licai Sun, Bin Liu, Jianhua Tao, and Zheng Lian, “Multimodal cross-and self-attention network for speech emotion recognition,” in *ICASSP 2021*. IEEE, 2021, pp. 4275–4279.
- [12] Yue Gu, Xinyu Lyu, Xinyu Sun, and Ivan Marsic, “Mutual correlation attentive factors in dyadic fusion networks for speech emotion recognition,” in *ACM MM 2019*, 2019, pp. 157–166.
- [13] Zilong Wang, Zhaohong Wan, and Xiaojun Wan, “Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis,” in *Proceedings of The Web Conference 2020*, 2020, pp. 2514–2520.
- [14] Wenxiang Jiao, Michael Lyu, and Irwin King, “Exploiting unsupervised data for emotion recognition in conversations,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020, pp. 4839–4846.
- [15] Devamanyu Hazarika, Soujanya Poria, Roger Zimmermann, and Rada Mihalcea, “Conversational transfer learning for emotion recognition,” *Information Fusion*, vol. 65, pp. 1–12, 2021.
- [16] Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie, “Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition,” in *AAAI Conference on Artificial Intelligence*, 2021, pp. 13789–13797.
- [17] Robyn Speer, Joshua Chin, and Catherine Havasi, “Conceptnet 5.5: An open multilingual graph of general knowledge,” in *Thirty-first AAAI conference on artificial intelligence*, 2017.
- [18] Surgan Jandial, Ayush Chopra, Mausoom Sarkar, Piyush Gupta, Balaji Krishnamurthy, and Vineeth Balasubramanian, “Retrospective loss: Looking back to improve training of deep neural networks,” in *Proceedings of the 26th ACM SIGKDD International Conference*, 2020, pp. 1123–1131.
- [19] Carlos Busso, Murtaza Bulut, Sungbok Lee, and Shrikanth S Narayanan, “Iemocap: Interactive emotional dyadic motion capture database,” *Language resources and evaluation*, pp. 335–359, 2008.
- [20] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea, “Meld: A multimodal multi-party dataset for emotion recognition in conversations,” in *ACL 2019*, 2019, pp. 527–536.
- [21] Dong Zhang, Weisheng Zhang, Shoushan Li, Qiaoming Zhu, and Guodong Zhou, “Modeling both intra-and inter-modal influence for real-time emotion detection in conversations,” in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 503–511.