COIN: Conversational Interactive Networks for Emotion Recognition in Conversation

Haidong Zhang* and **Yekun Chai***
Institute of Automation, Chinese Academy of Sciences
haidong_zhang14@yahoo.com chaiyekun@gmail.com

Abstract

Emotion recognition in conversation has received considerable attention recently because of its practical industrial applications. Existing methods tend to overlook the immediate mutual interaction between different speakers in the speaker-utterance level, or apply single speaker-agnostic RNN for utterances from different speakers. We propose *COIN*, a conversational interactive model to mitigate this problem by applying state mutual interaction within history contexts. In addition, we introduce a stacked global interaction module to capture the contextual and inter-dependency representation in a hierarchical manner. To improve the robustness and generalization during training, we generate adversarial examples by applying the minor perturbations on multimodal feature inputs, unveiling the benefits of adversarial examples for emotion detection. The proposed model empirically achieves the current state-of-the-art results on the IEMOCAP benchmark dataset.

1 Introduction

Emotion recognition in conversation (ERC) has attracted extensive interests due to the prevalence of user-generated contents on social media platforms, such as conversational messages and videos (Poria et al., 2017; Hazarika et al., 2018b; Poria et al., 2019; Hazarika et al., 2021), which aims to detect the speaker's emotions and sentiments within the context of human conversations. Recent works on ERC adopted recurrent neural networks (RNNs) to firstly learn the sequential utterances in conversations and then leveraged high-level context extractor, such as CMN (Hazarika et al., 2018b), DialogueRNN (Majumder et al., 2019), DialogueGCN (Ghosal et al., 2019), to capture the global contextual representation for emotion detection.

This two-step scheme has proven to be effective to achieve success in ERC and can be divided into two categories: one is modeling each speaker with one RNN, such as (Hazarika et al., 2018b; Majumder et al., 2019); the other is speaker-agnostic, *i.e.*, modeling each utterance using one RNN irrespective of its speaker, such as (Poria et al., 2017; Majumder et al., 2019). However, there is no direct dyadic interaction between speaker-specific RNNs in previous work. Different RNNs corresponding to different speakers have been used without mutual interaction (Hazarika et al., 2018b) or interacting through a mediate global RNN (Majumder et al., 2019).

In this paper, the proposed **Co**nversational **I**nteractive **N**etworks (COIN) employs immediate coupling interaction at each state of different speakers and adopts a global extractor to capture the contextual and self-dependency representation for emotion classifier. To enhance the generalization and robustness of our model, we generate adversarial examples by applying minor perturbations on multi-modal embeddings for adversarial training (AT) (Goodfellow et al., 2014).

Our work illustrates that dyadic interaction advances the performance of multimodal emotion recognition in conversation by incorporating mutual interaction and applying adversarial training. Our key contributions are in threefold:

- We introduce state mutual interaction components to allow for the immediate state interaction between different speakers, and global stacked interaction to capture the contextual and inter-dependency representations.
- We unveil the importance of adversarial training in ERC by promoting the model performance with generated adversarial examples on extracted multimodal embeddings.
- We propose a competing model that achieves the state-of-the-art (SOTA) performance on the IEMOCAP dataset, showing that textual

*Equal contribution

Proceedings of the Third Workshop on Multimodal Artificial Intelligence, pages 12–18

and audio features play the most important role in ERC.

2 Methodology

This section is orgnized as follows: Sec. 2.1 describes the definition of ERC task; Sec. 2.2 introduces the approach to extracting multimodal features; Sec. 2.3 gives a detailed description of the proposed model.

2.1 Task Definition

Let there be M parties or speakers $\{p_1, p_2, \cdots, p_M\}$ in a human conversation ($M = 2$ in our experiments). Given the utterances $\{u_1, u_2, \cdots, u_N\}$ from a conversation where the utterance u_t is from the corresponding speaker $p_{s(u_t)}$, the task of ERC is to detect the most likely class from emotion category set $\mathcal{C}$. Here s represents the mapping between the utterances and users.

2.2 Multimodal Feature Extraction

We extract multimodal features using the same setting as (Majumder et al., 2019) for a fair comparison. Multimodal features are simply concatenated along the feature dimension in our systems.

Textual Feature We employ multi-channel 1-D convolutional neural networks (CNNs) along the sequential dimension to extract n-gram lexical features with kernel sizes of $\{3, 4, 5\}$. Then a global max-pooling layer followed by a linear projection produces the utterance representation. This CNN is trained on emotion classification at the sentence level.

Acoustic Feature We use openSMILE (Eyben et al., 2010) toolkit[§] to extract speech features such as Mel-frequency cepstral coefficients (39 features) and pitch. Z-standardization is applied to normalize the low dimensional feature vectors.

Visual Feature 3D-CNN (Tran et al., 2015) is leveraged to obtain visual features from dialogue videos, followed by a ReLU and max-pooling operation.

2.3 Model Architecture

Fig. 1 illustrates the overview of proposed *COIN* architecture with the history length of $K = 6$. The multimodal inputs of utterances are firstly fed into feature extractor to obtain the multimodal features. Then we adopt Gated Recurrent Units (GRUs) (Chung et al., 2014) to capture the history dialogue of dyadic speakers A/B, followed by the mutual interaction for each state at utterance level. Afterward, the concatenated bidirectional mutual history vectors are fed into a stacked contextual interaction module to capture the inter-dependency between current and history dialogue states.

Speaker Mutual Interaction for Dialogue History Let $\mathbf{u}_i \in \mathbb{R}^d$ represent the extracted d-dimensional multimodal features for i-th speech uttered by speaker $\mathcal{P}$, K be the dialogue history length. We use GRUs in two directions to capture the utterance-level speaker dialogue context. For the forward GRU, we have: $\overrightarrow{\mathbf{h}}^i_{\mathcal{P}} = \overrightarrow{\text{GRU}}^i_{\mathcal{P}}(\mathbf{u}_i), \mathcal{P} \in \{A, B\}, i \in [t-K, t-1]$, where $\mathbf{h}_{\mathcal{P}} \in \mathbb{R}^d$ indicates the hidden state of speaker $\mathcal{P}$ at the step i. The history utterance sequences for speaker $\mathcal{P}$ are denoted as $\mathcal{U}_{\mathcal{P}}$.

We compute the mutual interaction for each history step i by linearly regulating each output of GRU with the previous hidden state of another speaker. In the forward direction, we have:

$$\overrightarrow{\mathbf{m}}_i = \left\{ \begin{array}{ll} \overrightarrow{\mathbf{h}}^i_{\text{A}} \sigma(\overrightarrow{\mathbf{h}}^{i-1}_{\text{B}} \overrightarrow{\mathbf{W}}_{\text{B}} + \overrightarrow{\mathbf{b}}_{\text{B}}) & \text{if } \mathcal{P} = \text{A} \\ \overrightarrow{\mathbf{h}}^i_{\text{B}} \sigma(\overrightarrow{\mathbf{h}}^{i-1}_{\text{A}} \overrightarrow{\mathbf{W}}_{\text{A}} + \overrightarrow{\mathbf{b}}_{\text{A}}) & \text{if } \mathcal{P} = \text{B} \end{array} \right\}, \tag{1}$$

where $\mathbf{h}^0_{\mathcal{P}}$ represents the initial hidden state of speaker $\mathcal{P}$, the sigmoid function $\sigma(x) = 1/(1 + \exp(-x))$, $\{\overrightarrow{\mathbf{W}}_{\text{A}}, \overrightarrow{\mathbf{W}}_{\text{B}}\} \in \mathbb{R}^{d \times d}, \{\overrightarrow{\mathbf{b}}_{\text{A}}, \overrightarrow{\mathbf{b}}_{\text{B}}\} \in \mathbb{R}^d$ represent the trainable parameters. The identical but reversed operation is applied in the backward direction. The output of both forward and backward direction at step i are concatenated along the feature dimension, denoted as $\overleftrightarrow{\mathbf{m}}_i = [\overrightarrow{\mathbf{m}}_i; \overleftarrow{\mathbf{m}}_i] \in \mathbb{R}^{2d}$.

Stacked Contextual Interaction The contextual encoder consists of L identical stacks. In the l-th layer, we feed the history dialogue representations $\mathbf{M}^l$ into a bi-GRU followed by a self-attention (SA) layer to capture the inter-dependency semantics. In the first layer, $\mathbf{M}^l$ is the sequence of encoded context $\overleftrightarrow{\mathbf{m}}_i$, and is bi-GRU's output from previous layer for intermediate stacks, *i.e.*, $\mathbf{M}^{l-1}_g (l > 1)$.

Denoting the output of bi-GRU as $\mathbf{M}^l_g$, the

[§] https://www.audeering.com/opensmile/

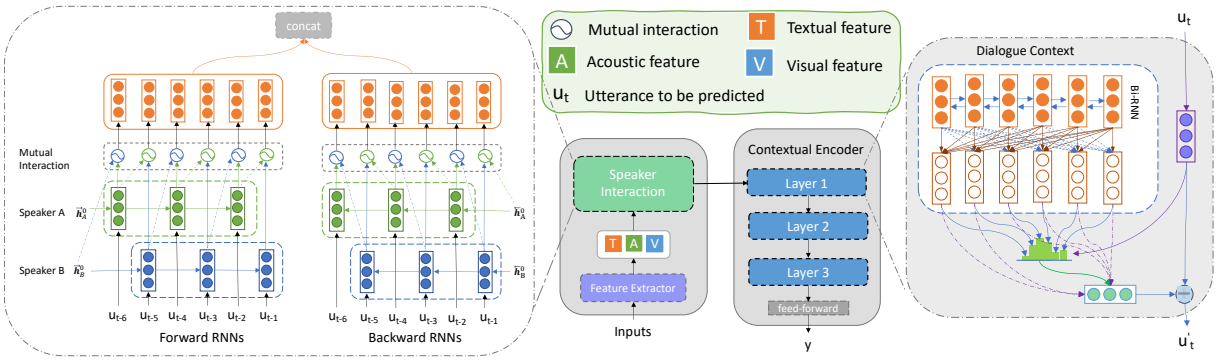


Figure 1: Schematic illustration of the proposed model.

scaled dot product self attention is calculated as:

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \mathbf{M}_g^l \mathbf{W}_Q, \mathbf{M}_g^l \mathbf{W}_K, \mathbf{M}_g^l \mathbf{W}_V, \quad (2)$$

$$\mathbf{M}_{\text{att}}^l = \text{softmax}(d^{-1/2} \mathbf{Q}\mathbf{K}^\top)\mathbf{V}, \quad (3)$$

where $\mathbf{M}_{\text{att}} \in \mathbb{R}^{K \times 2d}$ is passed into the bi-GRU in the next interaction stack as the dialogue context.

Given the encoded utterance $\mathbf{u}_t^l \in \mathbb{R}^{2d}$ at l-th layer (linearly projected multimodal features when $l = 0$), we calculate the context vector for history dialogues:

$$\mathbf{c}^l = \mathbf{M}_{\text{att}}^l \text{softmax}(\mathbf{M}_{\text{att}}^l \mathbf{u}_t^l), \quad (4)$$

$$\mathbf{u}_t' = \tanh(\mathbf{u}_t^l + \mathbf{c}), \quad (5)$$

where the output $\mathbf{u}_t'$ is used as the input of $\{l+1\}$-th layer (*i.e.*, $\mathbf{u}_t^{l+1}$).

Emotion Classifier We use L-th stack's output vector $\mathbf{u}_t'$ to get the final emotion prediction through a linear transformation: $\hat{y} = \arg\max(\mathbf{W}_o \mathbf{u}_t' + \mathbf{b}_o)$, where $\mathbf{W}_o \in \mathbb{R}^{d \times |\mathcal{C}|}$ and $\mathbf{b}_o \in \mathbb{R}^{|\mathcal{C}|}$ are parameters.

2.4 Training

Let $\mathbf{u}$ represent the multimodal features. The cross entropy loss $\mathcal{L}_{\text{xe}}$ between $\hat{y}$ and golden label y is used for training. To improve the generalization, we generate adversarial examples using the model parameterized by $\boldsymbol{\theta}$ as in (Goodfellow et al., 2014)–adding perturbations on extracted multimodal features: $\mathbf{u}_{\text{adv}} = \mathbf{u} + \epsilon \frac{\mathbf{g}}{\|\mathbf{g}\|_2}$, where $\mathbf{g} = \nabla \mathcal{L}_{\text{xe}}(\boldsymbol{\theta}; \mathbf{u})$, $\epsilon \in \mathbb{R}$ is selected on the held out set. The final training objective is defined as:

$$\mathcal{L} = \mathcal{L}(\boldsymbol{\theta}; \mathbf{u}) + \mathcal{L}(\boldsymbol{\theta}; \mathbf{u}_{\text{adv}}). \quad (6)$$

3 Experiments

3.1 Experimental Setup

Dataset We evaluate our model on the IEMOCAP dataset (Busso et al., 2008) by reporting the accuracy (acc.) and F1 score on single and overall emotion class. IEMOCAP dataset contains dyadic dialogue videos for ten unique speakers, two of which are used for testing. We maintain the same 80/20 split for training/test set, consisting of 5,810/1,623 utterances respectively. The utterances are annotated as six emotion labels, *i.e.*, happy, sad, neutral, angry, excited, and frustrated.

Implementation Details We experiment using the batch size 512, contextual interaction layer number $L \in \{1, 2, 3, 4, 5, 6\}$, embedding size $d \in \{50, 100, 150, 200\}$, history context size $K \in \{20, 30, 40, 50\}$, the extracted textual/audio/visual feature dimensions of 100/100/512 respectively. We use Adam optimizer (Kingma and Ba, 2015) with initial learning rate of 1e-3. We employ the exponential annealing with base 2 to adjust the learning rate. For adversarial training, we select $\epsilon = 5$ using validation set. To avoid overfitting, we applies dropout keep rate $p \in \{0.2, 0.3, 0.4\}$ and early stopping patience of 10 epoch during training. The optimal hyperparameter settings are: $L = 3, d = 100, K = 40, p = 0.3$. We use an NVIDIA 2080 Ti GPU for experiments.

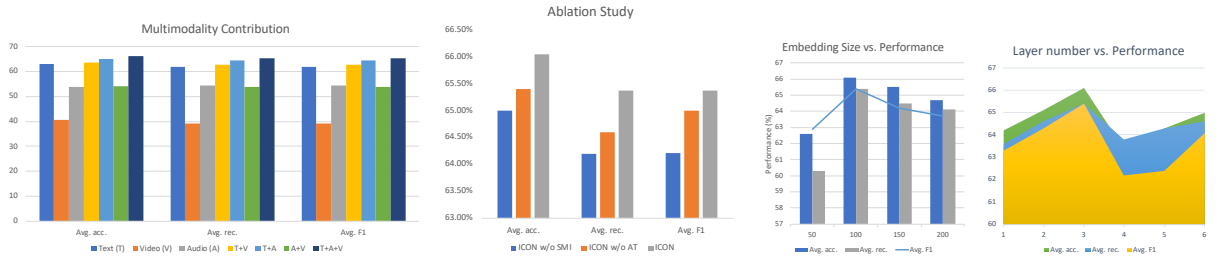
3.2 Results

Table 1 summarizes the performance of the proposed model compared with baseline models, in which our model overshadows previous baselines on both averaged accuracy and F1 metric. We found that the performance of our model ranks first for "angry" and "frustrated" sentiment prediction and achieves similar results on the other emotion classes.

Ablation Study We conduct ablation study on multi-modality (Fig. 2a), adversarial training and Speaker Mutual Interaction (SMI) module (Fig. 2b).

Model	Happy		Sad		Neutral		Angry		Excited		Frustrated		Average	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
CNN (Kim, 2014)	27.77	29.86	57.14	53.83	34.33	40.14	61.17	52.44	46.15	50.09	62.99	55.75	48.92	48.18
MemNet (Sukhbaatar et al., 2015)	25.72	33.53	55.53	61.77	58.12	52.84	59.32	55.39	51.50	58.30	67.20	59.00	55.72	55.10
bc-LSTM (Poria et al., 2017)	29.17	34.43	57.14	60.87	54.17	51.81	57.06	56.73	51.17	57.95	67.19	58.92	55.21	54.95
bc-LSTM+Att (Poria et al., 2017)	30.56	35.63	56.73	62.90	57.55	53.00	59.41	59.24	52.84	58.85	65.88	59.41	56.32	56.91
CMN (Hazarika et al., 2018b)	25.7	32.6	66.5	72.9	53.9	56.2	67.6	64.6	69.9	67.9	71.7	63.1	61.9	61.4
ICON (Hazarika et al., 2018a)	23.6	32.8	70.6	74.4	59.9	60.6	68.2	68.2	72.2	68.4	**71.9**	66.2	64.0	63.5
DialogueRNN (Majumder et al., 2019)	25.69	33.18	75.10	78.80	58.59	59.21	64.71	65.28	**80.27**	**71.86**	61.15	58.91	63.40	62.75
DialogueGCN (Ghosal et al., 2019)	40.62	42.75	89.14	84.54	61.92	63.54	67.53	64.19	65.46	63.08	64.18	66.99	65.25	64.18
IterativeERC (Lu et al., 2020)	-	53.17	-	77.19	-	61.31	-	61.45	-	69.23	-	60.92	-	64.37
COIN	**53.12**	42.50	85.71	73.07	60.05	62.23	66.48	**68.75**	69.13	69.01	61.73	**66.99**	**66.05**	**65.37**

Table 1: Overall performance of emotion recognition models on IEMOCAP dataset.



(a) Modality contribution (b) Ablation study (c) Embedding size (d) Layer number

Figure 2: Visualization of experimental analysis.

Fig. 2c and Fig. 2d show the influence of embedding size and stack number of context interactions. It can be seen model performance reaches its peak by taking the embedding size of 100 (Fig. 2c) and layer number $L = 3$ (Fig. 2d). Fig. 2b witnesses the influence of adversarial training and SMI on emotion detection. We further conduct experiments by applying adversarial training on various baselines, finding that our model achieves the best results among them. See Appendix B for discussion.

Fig. 2a show that among uni-modality, textual features contribute most followed by the acoustic setting whereas video features perform worst in our system. We guess high-level visual features extracted from CNN-3D lack of fine-grain facial representations, which requires further improvement. In dual modality settings, textual and acoustic features make the most contribution to predict emotion categories in comparison with tri-modal fusion settings.

Case Study Fig. 3 shows an instance of dialogue snippet, where our model captures the emotion dynamics of the male speaker during the conversation process. Using different RNNs to modeling various speaker utterances may circumvent the fluctuation of emotion transitions and effectively capture the emotion transition of disparate speakers. It is also observed in more examples in Appendix C.

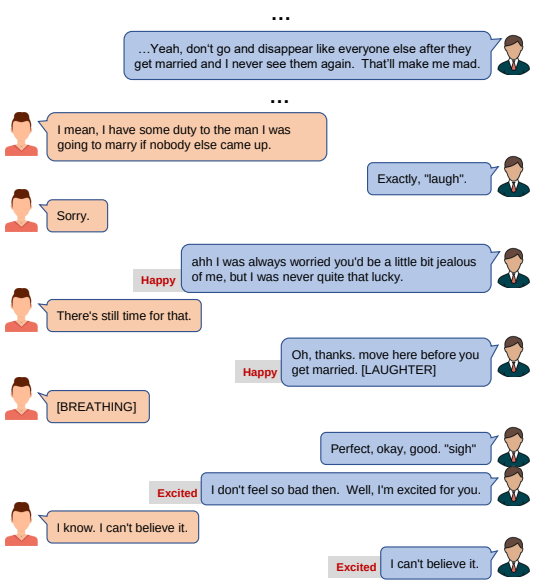


Figure 3: Case study.

4 Conclusion

We propose a new dialogue contextual interaction architecture to focus on the compact interaction for both speaker-level dialogue history and current utterance. By adopting adversarial training, our model achieves the SOTA performance on the IEMOCAP dataset for emotion recognition in conversation. In the future, multimodal fusion methods could be investigated to capture richer modelily-interactive representations at modality level.

References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Junyoung Chung, Çaglar Gülçehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *ArXiv*, abs/1412.3555.

Florian Eyben, Martin Wöllmer, and Björn W. Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *ACM Multimedia*.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *EMNLP-IJCNLP*.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. In *ICLR*.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. Icon: Interactive conversational memory network for multimodal emotion detection. In *EMNLP*.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. Conversational memory network for emotion recognition in dyadic dialogue videos. *NAACL*.

Devamanyu Hazarika, Soujanya Poria, R. Zimmermann, and R. Mihalcea. 2021. Conversational transfer learning for emotion recognition. *Inf. Fusion*, 65:1–12.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

X. Lu, Yanyan Zhao, Yang Wu, Yijian Tian, H. Chen, and Bing Qin. 2020. An iterative emotion interaction network for emotion recognition in conversations. In *COLING*.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *AAAI*.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *ACL*.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard H. Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *NIPS*.

Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. *ICCV*.

A Qualitative Analysis

Fig. 4 illustrates the confusion matrix of predicted emotions. We found that negative sentiments such as "sad", "angry" can be easily mispredicted as "frustrated", and vice versa. "Happy" emotions exhibit the worst performance among all of six categories, which is difficult for the model to distinguish from "excited". This is in line with our manually observed prediction results because sometimes it is even not obvious for a human to distinguish the emotions with similar polarities, such as "sad" and "frustrated", "happy" and "excited". Further study on learning sentiments of similar polarity may be a solution to such misunderstanding.

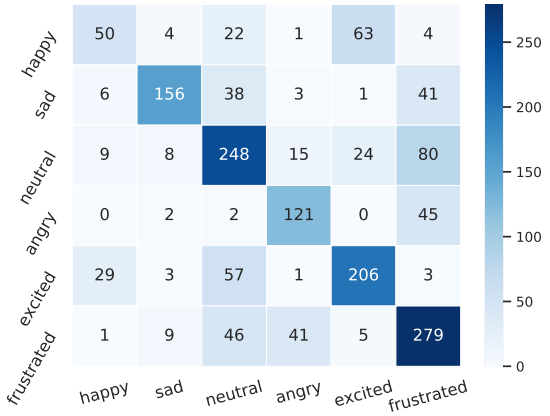


Figure 4: Confusion matrix of emotion predictions.

B Experiments on Adversarial Training

To verify the advantage of our model using adversarial training (AT), we further conduct experiments on different baseline models and report the result in Table 2. It is clear that our model outranks other models in terms of the overall performance, demonstrating the advantage of our model. Also, it is observed that emotion recognition models do not necessarily improve after incorporating the AT method. Specifically, models using single RNNs to simulate the speaker utterances, such as DialogueRNN and DialogueGCN, show the performance drop after adding AT, whereas models using separate RNNs to model different speakers, like ICON and ours, illustrate the advancement. We extrapolate that the emotion dynamics of different speakers may vary, thus the sensitivity of emotion models is affected by the adversarial noise derived from the conversational context. If different RNNs are adopted for various speaker utterance modeling, the noise would greatly rely on the current speaker's utterances despite the noise from noisy dialogue context, which eases the learning process of emotion transition.

C Case Study

Fig. 5 illustrates examples of our case study, which demonstrates that our model can capture the emotion dynamics during the conversation process.

Model	Happy		Sad		Neutral		Angry		Excited		Frustrated		Average	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
ICON +AT	47.93	43.77	84.69	75.28	58.71	60.05	63.49	66.85	72.31	67.26	60.00	65.31	64.98	64.18
DialogueRNN +AT	36.04	31.17	91.10	79.82	53.53	55.35	65.32	65.89	65.65	65.09	55.08	59.22	61.48	60.72
DialogueGCN +AT	60.38	32.49	69.23	76.10	61.31	59.92	53.65	56.91	62.36	74.87	57.79	47.20	60.99	59.38
Ours (w/ AT)	53.12	42.50	85.71	73.07	60.05	62.23	66.48	68.75	69.13	69.01	61.73	66.99	**66.05**	**65.37**

Table 2: Performance of emotion recognition models with Adversarial Training (AT) on the IEMOCAP dataset.

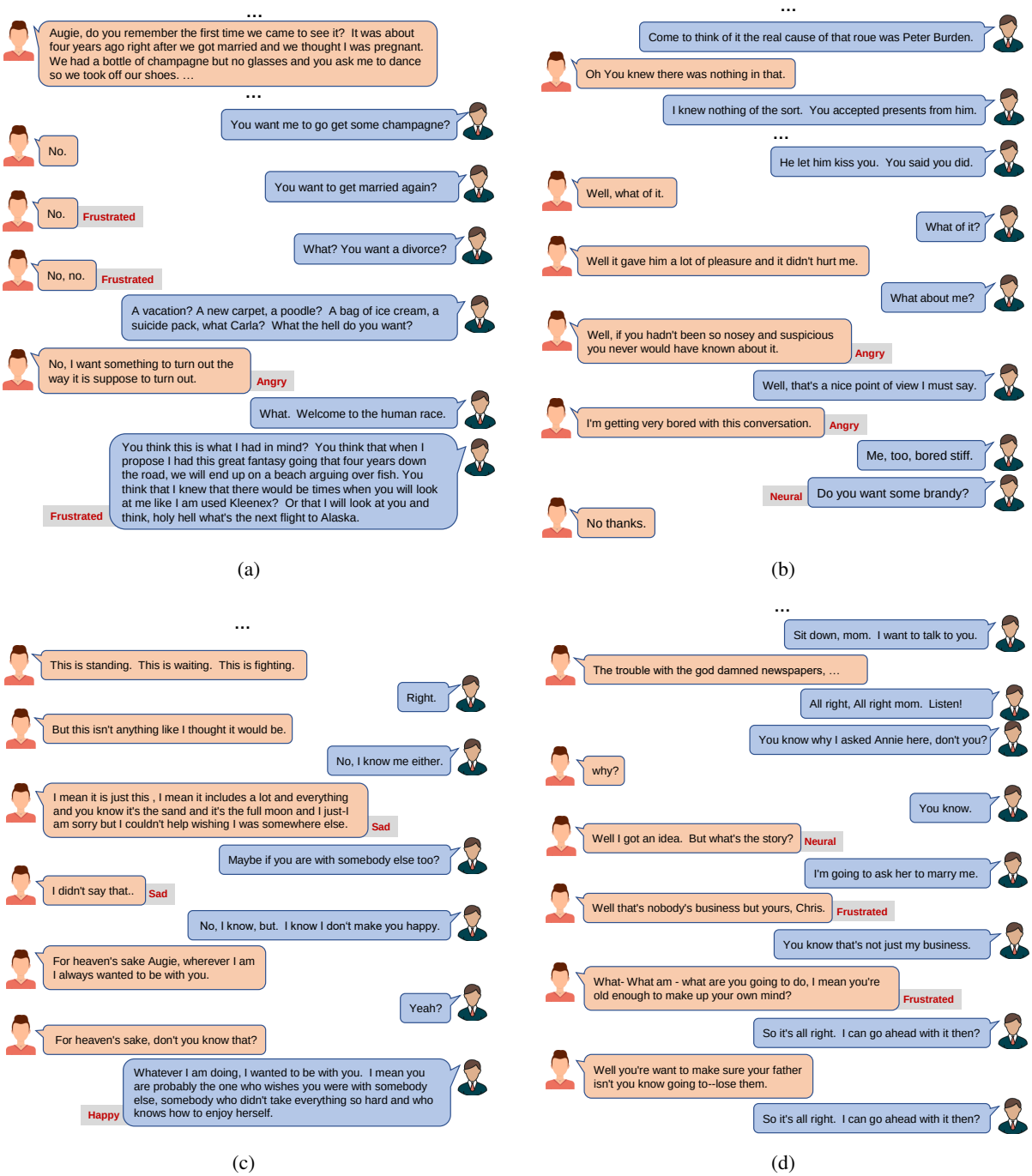


Figure 5: Case study examples.