


Article

Domain Adversarial Network for Cross-Domain Emotion Recognition in Conversation

Hongchao Ma¹, Chunyan Zhang² , Xiabing Zhou^{3,*}, Junyi Chen¹ and Qinglei Zhou¹

¹ School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou 450001, China; ma-hc@foxmail.com (H.M.); junyichen_ch@sina.com (J.C.); ieqlzhou@zzu.edu.cn (Q.Z.)

² State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450001, China; iecyzhang@163.com

³ School of Computer Science and Technology, Soochow University, Suzhou 215006, China

* Correspondence: zhouxiaxing@suda.edu.cn

Abstract: Emotion Recognition in Conversation (ERC) aims to recognize the emotion for each utterance in a conversation automatically. Due to the difficulty of collecting and labeling, this task lacks the dataset corpora available on a large scale. This increases the difficulty of finishing the supervised training required by large-scale neural networks. Introducing the large-scale generative conversational dataset can assist with modeling dialogue. However, the spatial distribution of feature vectors in the source and target domains is inconsistent after introducing the external dataset. To alleviate the problem, we propose a Domain Adversarial Network for Cross-Domain Emotion Recognition in Conversation (DAN-CDERC) model, consisting of domain adversarial and emotion recognition models. The domain adversarial model consists of the encoders, a generator and a domain discriminator. First, the encoders and generator learn contextual features from a large-scale source dataset. The discriminator performs domain adaptation by discriminating the domain to make the feature space of the source and target domain consistent, so as to obtain domain invariant features. Then DAN-CDERC transfers the learned domain invariant dialogue context knowledge from the domain adversarial model to the emotion recognition model to assist in modeling the dialogue context. Due to the use of a domain adversarial network, DAN-CDERC obtains dialogue-level contextual information that is domain invariant, thereby reducing the negative impact of inconsistency in domain space. Empirical studies illustrate that the proposed model outperforms the baseline models on three benchmark emotion recognition datasets.

Keywords: emotion recognition in conversation; domain adversarial network; domain adaptation; transfer learning



Citation: Ma, H.; Zhang, C.; Zhou, X.; Chen, J.; Zhou, Q. Domain Adversarial Network for Cross-Domain Emotion Recognition in Conversation. *Appl. Sci.* **2022**, *12*, 5436. <https://doi.org/10.3390/app12115436>

Academic Editor: Antonio Fernández-Caballero

Received: 14 April 2022

Accepted: 24 May 2022

Published: 27 May 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Emotion plays a significant role in daily life and intelligent dialogue systems. Emotion recognition in conversation (ERC), which is one of the important tasks of Natural Language Processing, has attracted more and more attention in recent years. ERC is to predict the emotion of each utterance in a conversation. ERC is more challenging, considering the sequential information of the conversation and the self-speaker dependencies and inter-speaker dependencies [1].

In the literature, many neural network models have been applied to model dialogue and dependencies, such as recurrent neural networks [2,3], graph-based convolutional neural networks [4,5], and attention mechanisms [6–8]. However, some problems should not be ignored. A vital issue of emotion recognition in conversation is the lack of available labeled data, which is hard to collect and annotate. The emotion of the same statement in different dialogue scenarios is determined according to the context, rather than the same emotion [9]. It is difficult for annotators to figure out the contextual information. So, there are a relatively small number of available datasets, such as in the more commonly used

data IEMOCAP (the IEMOCAP and MELD are two common datasets for ERC, which will be described in detail in Section 4.1) [10], there are only 153 dialogues. It is challenging to finish the supervised training required by large-scale neural networks.

In addition, some datasets contain only two participants in each conversation, while others have multiple participants. Figure 1 shows a conversation from the MELD (the IEMOCAP and MELD are two common datasets for ERC, which will be described in detail in Section 4.1) [11], which comprises five participants. It is challenging to model speakers dependencies. The IEMOCAP dataset has five sessions, and the last session is divided into test data, which means that there are different speakers in the training data and the test data. These factors have made it more challenging to model speakers and emotion dependencies.

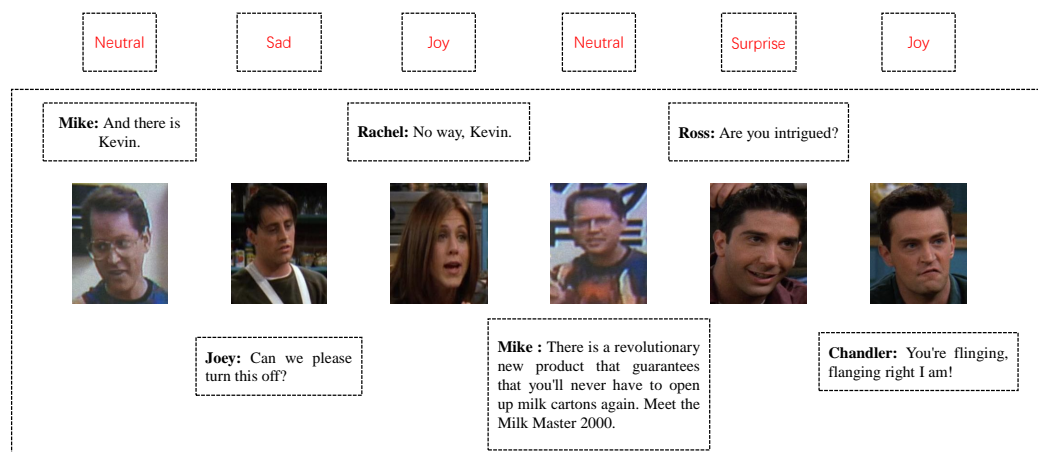


Figure 1. A conversation example with emotion labels from the MELD dataset.

Cross-domain sentiment classification, which aims to transfer knowledge to the target domain from the source domain, is one of the effective ways to alleviate the lack of datasets in the target domain. Many excellent achievements have been made in cross-domain sentiment classification [12–14]. Cross-domain sentiment classification generally requires the source domain data to be labeled. Unlike sentiment classification, ERC lacks large scale datasets and aims to identify the emotion of each utterance in a conversation, rather than a single sentence. An utterance of the speaker is affected by their own and external factors, such as topic, speaker’s personality, argumentation logic, viewpoint, intent, etc. [9]. In turn, the utterance reflects these factors to a certain extent. These factors may lead to improved conversation understanding, including emotion recognition [9]. For the above reasons, aiming at the problem of lack of labeled data, Hazarika et al. [15] pre-trained a whole conversation jointly using a hierarchical generative model and transfer the knowledge to the target domain.

In summary, the generation task can be used to assist the ERC task, since the dialogue generation task and ERC have some similarities. However, the spatial distribution of feature vectors in the source and target domains is inconsistent after introducing external generation datasets.

Inspired by cross-domain sentiment classification, to alleviate the problem of inconsistency in the domain feature space of the source domain dataset and target domain dataset, we propose a Domain Adversarial Network for Cross-Domain Emotion Recognition in Conversation (DAN-CDERC) model, which transfers the knowledge from the domain adversarial model to the emotion recognition model, instead of directly modeling historical information and speaker information. The domain adversarial model consists of the encoders, a generator and a domain discriminator. The encoders are used to learn sequence knowledge from the massive amounts of the generative dataset. The generator generates an utterance in the source domain. The discriminator performs domain adaptation by discriminating the domain, which plays an essential role in reducing the domain inconsistency for domain adaptation. Since both the generative conversational task and the

ERC task also need to model the dialogue context, DAN-CDERC can use the dialogue sequence knowledge from the large-scale dialogue generation dataset to assist the ERC in modeling the dialogue context. Due to the use of the domain adversarial network, our DAN-CDERC avoids the dissimilarity of the domain and vector space during the transferring process. In this paper, we try to achieve the same effect of these models without explicitly modeling speakers.

For sentence-level classification, the domain discriminator is used to discriminate sentences. Our discriminator can discriminate each utterance in the conversation that belongs to the source domain or the target domain, rather than the whole conversation. We believe this is important. On the one hand, our model is relatively simple, with only two encoder layers. It is challenging to represent the whole conversation with a vector effectively. On the other hand, using sequential utterances is more conducive to learning and transferring sequence knowledge.

In summary, our contributions are as follows:

- To alleviate the problem of the small scale of the ERC task dataset, we propose Domain Adversarial Network for Cross-Domain Emotion Recognition in Conversation, which not only learns knowledge from large-scale generative conversational datasets, but also utilizes adversarial networks to reduce the difference between source and target domains;
- We use two large-scale generative conversational datasets and three emotion recognition datasets to verify model performance. The empirical studies illustrate the effectiveness of the proposed model, even without modeling information dependencies such as speakers.

The rest of the paper is organized as follows: Section 2 discusses related work; Section 3 provides details of our model; Section 4 shows and interprets the experimental results; Section 5 analyses and discusses the experimental results; and finally, Section 6 concludes the paper.

2. Related Work

Inspired by sentence-level cross-domain sentiment analysis, this paper utilizes large-scale dialogue generative datasets, adversarial networks and transfer learning for Emotion Recognition in Conversation. The related work includes dialogue generation, Emotion Recognition in Conversation, Cross-Domain Sentiment Analysis, Adversarial Network and Transfer Learning.

2.1. Dialogue Generation

Hierarchical Recurrent encoder–decoder (HRED) [16] is a classic generative hierarchical neural network. It has three key components, including the utterance encoder, the context encoder and the decoder. The latent variable hierarchical recurrent encoder–decoder (VHRED) [17] extended HRED. VHRED added a latent variable at the decoder, which is trained by maximizing a variational lower-bound on the log-likelihood. Variational Hierarchical Conversation RNN (VHCR) [18] augmented a global conversational latent variable along with local utterance latent variables to build a hierarchical latent structure with a new regularization technique called utterance drop.

Moreover, ERC is a vital step to endowing the dialogue system with emotional perception. Some researchers are interested in how to make the dialogue system have emotional perception. Zhou et al. [19] proposed novel mechanisms to make the responses more emotional respectively: embedded emotion categories, captured the change of implicit internal emotion states, and used explicit emotion expressions by an external emotion vocabulary. Deeksha et al. [20] employed a multi-task learning framework to predict emotion labels, and used emotion labels to guide the modeling of empathetic conversations. Li et al. [21] proposed a multi-resolution interactive empathetic dialogue model combining coarse-grained dialogue-level and fine-grained token-level emotions, which contains an interactive adversarial learning framework to judge emotional feedback. Xie et al. [22]

combined 32 emotions and eight additional emotion regulation intentions to complete the task of empathic response generation. Ide et al. [23] made the generated responses more emotional by adding emotion recognition tasks.

2.2. Emotion Recognition in Conversation

Unlike the document-level sentiment and emotion classification, neural network models learn the representation of words and documents, and understand the self and inter-speaker dependencies, for sentiment and emotion classification in conversation. Most of the models for ERC are hierarchical network structures, including at least one utterance encoder layer to encode utterances, and one context encoder layer to encode contextual content.

A dialogue, generally composed of multi-turn utterances, happens in a natural sequence, which is suitable for modeling with RNN. So RNN has become a fundamental component for emotion detection in conversation. Poria et al. [2] employed an LSTM-based to model dependencies and relations among the utterances. Majumder et al. [3] used three GRUs to model the speaker, the context and the emotion of the preceding utterances. In addition, the attention mechanism is also an important component. Wei et al. [24] employed GRUs and hierarchical attention to model the self and inter-speaker influences of utterances. Jiang et al. [25] proposed a hierarchical model and introduced a convolutional self-attention network as an utterance encoder layer.

Due to the rising of graph neural network models and the problem of context propagation in the current RNN-based methods, some work RNN-based networks are replaced by graph networks. Ghosal et al. [4] proposed Dialogue Graph Convolutional Network (DialogueGCN) to model self and inter-speaker dependencies. Zhang et al. [5] tried to address context-sensitive dependencies and speaker-sensitive dependencies using a conversational graph-based convolutional neural network in multi-speaker conversation. Sheng et al. [26] introduced a two-stage Summarization and Aggregation Graph Inference Network, which models inference for topic-related emotional phrases and local dependency reasoning over neighboring utterances. Zhang et al. [27] proposed a dual-level graph attention mechanism that augments the semantic information of the utterance and multi-task learning to alleviate the confusion between a few non-neutral utterances and much more neutral ones. Ma et al. [28] used a multi-view network to explore the emotion representation of a query from word-level and utterance-level views. TODKAT [29] used a topic-augmented language model (LM) with an additional layer specialized for topic detection, and combined LM with commonsense statements derived from a knowledge base ATOMIC. SKAIG [30] used commonsense knowledge to enrich the edges of the graph with knowledge representations from the model COMET.

2.3. Cross-Domain Sentiment Analysis

Cross-domain sentiment analysis is one of the areas where a classifier is trained in one source domain and applied to one target domain. Due to different expressions of emotions across several domains, many pivot-based methods [14,31,32] have been proposed to address domain adaptation problems by learning non-pivot words and pivot words. The selection of non-pivot words and pivot words will directly affect the performance of the target domain. Another effective way is adversarial training [13,33–35], which obtains domain-invariant features by deceiving the discriminator.

2.4. Adversarial Network and Transfer Learning

Multi-source transfer learning can also lay a foundation for modeling various aspects of different emotions (e.g., mood, anxiety), where only a limited number of datasets with a small number of data samples are available.

Liang et al. [36] treated emotion recognition and culture recognition as two adversarial tasks for cross-culture emotion recognition to address the problem of generalization across different cultures. Lian et al. [37] and Li et al. [38] treated the speaker characteristics and emotion recognition as two adversarial tasks to reduce the speaker's influence on emotion

recognition. Parthasarathy et al. [39] proposed an Adversarial Autoencoder (AAE) to perform variational inference over the latent factors, including age, gender, emotional state, and content of speech.

Furthermore, some researchers utilize transfer learning for emotion recognition. Gideon et al. [40] investigated that emotion recognition can benefit from using representations originally learned for different paralinguistic and different domains. Felbo et al. [41] used 1246 million tweets to train a pre-training model for emoji recognition. Li et al. [42] utilized a low-level transformer as the utterance encoder layer and a high-level transformer as the context encoder layer. EmotionX-IDEA [43] and PT-Code [44] learn emotional knowledge from BERT. Hazarika et al. [15] pre-trained a whole conversation jointly using a hierarchical generative model and transferred it to the target domain.

Our work strives to tackle a small number of datasets of ERC. Hence we use a large amount of publicly available generative conversational datasets to model conversation, and introduce a domain discrimination task to enhance domain adaptability.

3. Domain Adversarial Network for Emotion Recognition in Conversation

In this paper, there are two domains: a source domain D_s and a target domain D_t . Because the source domain dataset is used to train the generative task, it has no emotional label. For the source domain, given a dialogue containing m utterances $d_s = \{u_1, u_2, \dots, u_m\}$, m is the length of dialogue, we can leverage $\{u_1, u_2, \dots, u_{m-1}\}$ and $\{\hat{u}_2, \hat{u}_3, \dots, \hat{u}_m\}$ to train a generative conversational task. For the target domain, given a dialogue containing n utterances $d_t = \{u_1, u_2, \dots, u_n\}$ and n labels $Y_d = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n\}$, n is the length of dialogue. Our goal is to predict the emotion labels of the d_t .

This study proposes a *Domain Adversarial Network for Cross-Domain Emotion Recognition in Conversation* (DAN-CDERC) model to address emotion recognition with generative conversation. DAN-CDERC contains two key components: the *Domain Adversarial model* and the *Emotion Recognition model*. Figure 2 shows the architecture of the *Domain Adversarial model*, where the input is $\{u_1, u_2, \dots, u_m\}$ for the source domain and $\{u_1, u_2, \dots, u_n\}$ for the source target domain. The output of the generator is the generated response sequence $\{\hat{u}_2, \hat{u}_3, \dots, \hat{u}_m\}$, and the output of the discriminator is the domain labels. Figure 3 shows the architecture of the *Emotion Recognition model*, where the input is $\{u_1, u_2, \dots, u_n\}$ and the output is emotion labels $Y_d = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n\}$.

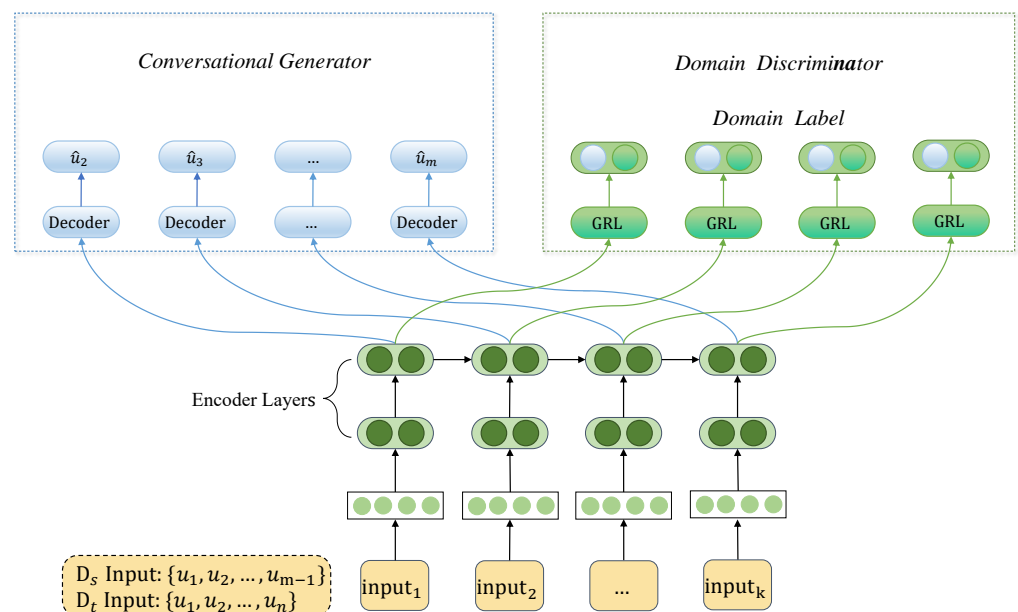


Figure 2. Overall architecture of the Domain Adversarial Network model.

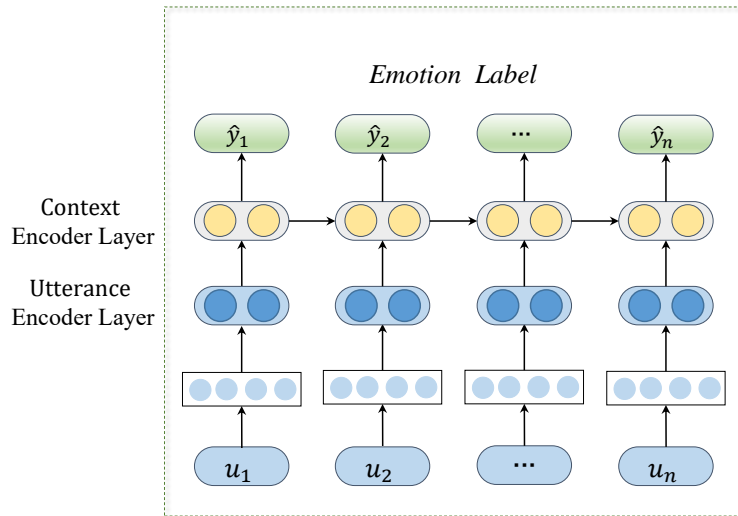


Figure 3. Overall architecture of the Emotion Recognition model.

The *Domain Adversarial model* contains the *Encoder Layers*, the *Generator model* and the *Discriminator*. The *Encoder Layers* and the *Generator*, from Hierarchical Recurrent encoder–decoder (HRED) [16], are used to learn sequence knowledge from the massive amounts of generative dataset. The *Discriminator* performs domain adaptation by discriminating the domain, which plays an essential role in reducing the domain inconsistency for domain adaptation. For the emotion recognition model, we leverage BERT [45] to encode utterances, and LSTM (context encoder) to encode context, which learns context weights from the generative conversational model. First, we leverage $d_s = \{u_1, u_2, \dots, u_{m-1}\}$ and $\{\hat{u}_2, \hat{u}_3, \dots, \hat{u}_m\}$ to train a generative conversational model, and leverage d_s and d_t to train the domain-distinguish task. Then a part of the parameters of the generative model is transferred to the emotion recognition model (target task). We leverage $d_t = \{u_1, u_2, \dots, u_n\}$ and $Y_d = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n\}$ to train the emotion recognition model.

3.1. Domain Adversarial Model

3.1.1. Encoder Layers

The encoder layers include an utterance encoder and a context encoder. A Bidirectional LSTM is used as the utterance encoder, and a unidirectional LSTM is used as the context encoder. Given a dialogue $d_s = \{u_1, u_2, \dots, u_{m-1}\}$, the utterance encoder uses Equation (1) to represent each u_i as a high-dimensional vector h_i . Then, the context encoder uses Equation (2) to learn the sequence knowledge of the context and represent d_s as $\{H_1, H_2, \dots, H_{m-1}\}$.

$$h_i = BiLSTM(u_i) \tag{1}$$

$$H_i = LSTM(h_i). \tag{2}$$

3.1.2. Conversational Generator

The generator is used to decode and generate u^{i+1} one response at a time. In addition, at the decoding stage, the generator generates a new utterance u^{i+1} by computing a distribution over vocabulary V_t for target elements u^{i+1} by projecting the output of the decoder via a linear layer with weights W^o and bias b^o ,

$$p(u^i | u^1, \dots, u^{i-1}; X) = \text{softmax}(W^o H^i + B^o). \tag{3}$$

3.1.3. Domain Discriminator

The role of the domain discriminator is to predict the domain label of the utterance u_i which comes from the target domain or the source domain. The generator and the discriminator are trained in parallel. For each u_i through the encoding stage of Section 3.1.1, an H_i can be obtained. Specifically, before feeding H_i to the domain classification, the H_i goes through the gradient reversal layer (GRL) [33].

During the backpropagation, the role of the GRL is to reverse the gradient. The following equations are the forward propagation and backpropagation when H_i goes through GRL, respectively:

$$Q_\lambda(x) = x \tag{4}$$

$$\frac{\partial Q_\lambda(x)}{\partial x} = -\lambda I. \tag{5}$$

We denote the hidden state H_i through the GRL as \hat{H}_i .

3.2. Emotion Recognition with Transfer Learning

Given a dialogue containing n utterances $d = \{u_1, u_2, \dots, u_n\}$, n is the length of dialogue. Our goal is to predict their labels $\{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n\}$. This model has two components, including an utterance encoder and a context encoder.

3.2.1. Utterance Encoder

BERT is a classic pre-training model and has achieved good results on many NLP tasks. Consequently, BERT [45] is used to encode utterances. We choose the BERT-based uncased pre-trained model as our utterance encoder. Through BERT, we can obtain representations of utterances, $H_h = \{h_1, h_2, \dots, h_n\}$

$$h_i = \text{BERT}(u_i). \tag{6}$$

3.2.2. Context Encoder and Transfer Learning

The context encoder of classification is the same as the encoder of the generative conversational model. The parameters of the context encoder of the generative model are used for initialization. The input of the context encoder is $H_h = \{h_1, h_2, \dots, h_n\}$, and the output $H_H = \{H_1, H_2, \dots, H_n\}$ can be obtained by the following formulas:

$$r_t = \sigma(W_{ir}h_t + b_{ir} + W_{hr}H_{(t-1)} + b_{hr}) \tag{7}$$

$$z_t = \sigma(W_{iz}h_t + b_{iz} + W_{hz}H_{(t-1)} + b_{hz}) \tag{8}$$

$$n_t = \tanh(W_{in}h_t + b_{in} + r_t * (W_{hn}H_{(t-1)} + b_{hn})) \tag{9}$$

$$H_t = (1 - z_t) * n_t + z_t * H_{(t-1)} \tag{10}$$

$$H_t = \tanh(W^p H_t + b^p). \tag{11}$$

We transfer $\{W_{hr}, W_{hz}, W_{hn}, b_{hr}, b_{hz}, b_{hn}, W^p, b^p\}$ of the adversarial generative model to the context encoder of classification. Then H_t is used as inputs to a softmax output layer:

$$P_p = \text{softmax}(W_p H_t + B_p). \tag{12}$$

Here, W_p and B_p are model parameters, and P_p is used to predict emotion.

3.3. Model Training

3.3.1. Conversational Generator

The goal is to maximize the output X_H probability given the input original X_O . Therefore, we optimize the negative log-likelihood loss function:

$$Loss_{gen} = -\frac{1}{|\tau|} \sum_{(X_O, X_H) \in \tau} \log p(X_H | X_O; \theta), \quad (13)$$

where θ is the model parameters, and (X_O, X_H) is a pair (original utterance-new utterance) in training set τ , then:

$$\begin{aligned} \log p(X_H | X_O; \theta) = \\ \sum_{i=1}^n \log p(x_H^i | x_H^1, x_H^2, \dots, x_H^{i-1}, X_O; \theta), \end{aligned} \quad (14)$$

where $p(x_H^i | x_H^1, x_H^2, \dots, x_H^{i-1}, X_O; \theta)$ is calculated by the decoder.

3.3.2. Domain Discriminator and Joint Learning

We feed \hat{H}_i through the GRL to the domain discriminator as:

$$d = \text{softmax}(W_d \hat{H}_i + b_d). \quad (15)$$

Our training objective is to minimize the cross-entropy loss over a set of training examples:

$$Loss_{domain} = -\frac{1}{N_s + N_t} \sum_i^{N_s + N_t} \sum_j^K \hat{d}^i(j) \log d^i(j). \quad (16)$$

We jointly train the conversational generator and the domain discriminator, and the final loss is the sum of the loss of the two tasks:

$$Loss_{total} = Loss_{gen} + \beta Loss_{domain}. \quad (17)$$

3.3.3. Emotion Recognition in Conversation

Given a dialogue d including n utterances and the pre-defined emotion y_i of u_i , our training objective is to minimize the cross-entropy loss over a set of training examples, with a ℓ_2 -regularization term,

$$\mathcal{J}(\theta_y) = -\sum_{i=1}^N \sum_{j=1}^K y_i \log \hat{y}_i + \frac{\lambda}{2} \|\theta_y\|^2, \quad (18)$$

where \hat{y}_i is the predicted label, and θ_y is the set of model parameters.

4. Experiments

4.1. Experimental Settings

4.1.1. Data

We choose Cornell Movie Dialog Corpus [46] and Ubuntu Dialog Corpus [47] as the source domain datasets, and call the Cornell Movie Dialog Corpus, Cornell, and the Ubuntu Dialog Corpus, Ubuntu. In all experiments, we carry out IEMOCAP [10], MELD [11] and DailyDialog [48] to evaluate the performance of our model. Table 1 and Figure 4 show the statistics of the datasets.

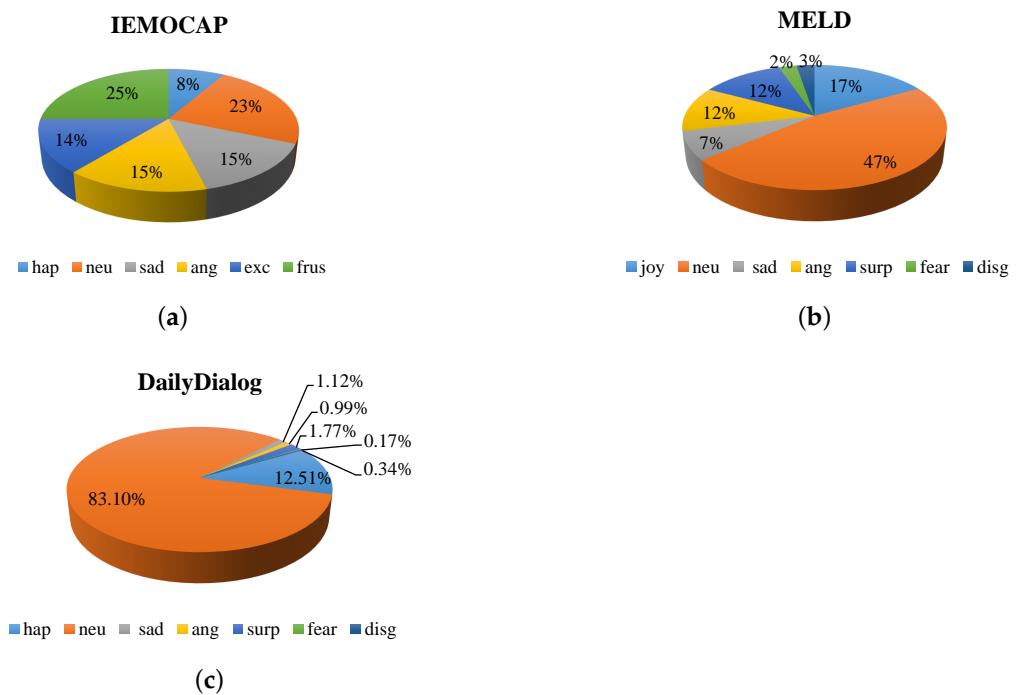


Figure 4. Distribution of emotions. hap: happy; neu: neutral or no emotion; ang: angry; exc: excited; frus: frustrated; surp: surprise; disg: disgust. (a) Distribution of emotions for IEMOCAP. (b) Distribution of emotions for MELD. (c) Distribution of emotions for DailyDialog.

Table 1. Training, validation and test data distribution in the datasets.

Dataset		Dialogues	Utterances	
Source(S)	Cornell(C)	train	66,477	244,030
		val	8310	30,436
		test	8310	30,247
	Ubuntu(U)	train	898,142	6,893,060
		val	18,920	135,747
		test	19,560	139,775
Target(T)	IEMOCAP(I)	train/val	120	5810
		test	31	1623
	MELD(M)	train	1038	9989
		val	114	1109
		test	280	2610
	DailyDialog(D)	train	11,118	87,170
val		1000	8069	
test		1000	7740	

- Cornell [46] is a conversational dataset of movie scripts collected from various sites. It contains more than 80,000 dialogues, 300,000 sentences, and 9000 characters in 617 movies.
- Ubuntu [47] is dyadic conversations extracted from the Ubuntu chat logs, which are used to receive technical support for various Ubuntu-related problems. It contains more than one million dialogues and 7 million sentences.
- IEMOCAP [10] is a multimodal dataset, and we only use textual data. Emotion recognition of multimodal data is beyond the scope of this paper. It contains 151 dialogues and 7433 utterances. Each conversation consists of multi-turn utterances, and each

utterance is annotated with one of the following emotions: angry, happy, sad, neutral, excited, and frustrated.

- MELD [11] is a multimodal emotion classification dataset that contains textual, acoustic and visual information. MELD is extended from the EmotionLines dataset [49]. It contains 1432 dialogues and 13,708 utterances from the Friends TV series. Each conversation consists of multi-turn utterances, and each utterance is annotated with one of the following emotions: angry, joy, neutral, disgust, sad, surprise and fear.
- DailyDialog [48] is a daily conversation dataset that reflects our daily ways of communication. It contains 13,118 multi-turn dialogues, and the speaker turns are roughly eight. Each utterance has a label of the following emotions: angry, happy, sad, surprise, fear, disgust and neutral (no_emotion).

As displayed in Figure 4, it can be seen that the distribution of labels is relatively balanced in IEMOCAP. Unlike IEMOCAP and MELD, the no_emotion labels account for 83.1% of DailyDialog. It is unbalanced, so the no_emotion will not be evaluated.

4.1.2. Setting

Table 2 shows the hyper-parameters of the model. For the baseline models, we use the hyper-parameters provided in the original papers or the same hyper-parameters as our setting. We employ AdaGrad [50] to optimize the classification model parameters. We utilize F1-score to measure the classification performance for each category, and the average F1-score and accuracy measure the overall performance.

Table 2. Setting of hyper-parameters.

Parameters	IEMOCAP	MELD	DailyDialog
Embedding Size	300	300	300
Utterance Encoder Hidden Size	768	768	768
Context Encoder Hidden Size	256	256	256
Dropout	0.1–0.5	0.1–0.5	0.1–0.5
Learning Rate	0.0001	0.0001	0.0001
Batch Size	4	16	16

4.2. Experimental Results

Table 3 presents the results of using different source domains for three target datasets, where “×” means that only transfer learning is used without the adversarial network [15]; “√” means that the model we proposed uses the adversarial network.

Table 3. F1-score of transfer between different domains. Weighted F1 metrics are used to evaluate classification performance.

Source→Target	Adversarial Network	
	×	√
C→I	59.25	64.40
U→I	58.71	63.94
C→M	57.89	59.44
U→M	57.51	59.23
C→D	48.00	55.20
U→D	47.10	54.60

As can be seen from Table 3, on the three target data, our DAN-CDERC has achieved a significant performance improvement compared with the method without the adversarial network, which is higher by around 5%, 2%, and 7%, respectively. It verifies the effectiveness of our DAN-CDERC model, which can build a good bridge between the

source domain and the target domain. Moreover, it demonstrates the importance of domain adaptation in the transfer between different domains for ERC.

For different source domains, our method shows a primary trend for the three target datasets, where Cornell as the source domain is better than Ubuntu as the source domain, which is higher by 0.46%, 0.21%, and 0.6%, respectively. For transfer learning, in general, the larger amount of source dataset, the better the experimental performance of the target dataset. As shown in Table 1, it is clear that the scale of Ubuntu is an order of magnitude larger than Cornell. To explore this reason, we analyze the characteristics of these datasets. Cornell is composed of movie scripts from multiple websites, and Ubuntu mainly consists of various technical Ubuntu-related problems. For the target domain datasets, the IEMOCAP comes from the drama script, the MELD comes from the movie script, and DailyDialog is the daily dialogue. In terms of content, the similarity between the three target datasets and Cornell is greater than that of Ubuntu. This explains why when Ubuntu is used as the source domain, although the data scale is large, the effect is not as good as Cornell as the source domain dataset.

It can be observed that our model has apparent effects on the IEMOCAP (5%) and the DailyDialog (7%) compared with the method without the adversarial network. Still, it has little impact on the MELD (2%). This may be due to the fact that the size of IEMOCAP is relatively tiny, and DailyDialog is relatively unbalanced. The knowledge brought by domain migration can compensate for these deficiencies and improve performance. However, MELD is relatively large and balanced, and the transfer of different domains does not contribute much to performance improvement.

5. Analysis and Discussion

In this section, we give some analysis and discussion.

5.1. Comparison with Baselines

We compare our model with various baseline approaches for emotion recognition in conversation.

- **bc-LSTM** is a basic model which employs BiLSTM to capture contextual content from the surrounding utterances without distinguishing different speakers;
- **CMN [51]** is the Conversational Memory Network, which models utterance context from dialogue history using two GRUs for speakers. Then, utterance representation is obtained by feeding the current utterance as the query to two memory networks for different speakers;
- **ICON [6]** uses GRU to model the self and inter-speaker sentiment influences and employs a memory network to store contextual summaries for classification. In our implementation, we only use the uni-modal classification;
- **DialogueRNN [3]** employs three GRUs (global GRU, party GRU, and speaker GRU) to model the speaker, the context and the emotion of the preceding utterances;
- **DialogueGCN [4]** uses a graph convolutional neural network to model self and inter-speaker dependencies. It represents each utterance as a node and models the dependencies between the speakers of those utterances by leveraging the edges between a pair of nodes/utterances.
- **DAN_{Cornell}** means that the source domain is Cornell.
- **DAN_{Ubuntu}** means that the source domain is Ubuntu.

IEMOCAP: Table 4 presents the results of our proposed DAN-CDERC model and strong baselines. DAN_{Cornell} and DAN_{Ubuntu} achieve an average F1-score of 64.40%, 63.94% and an accuracy of 65.07% and 64.61%, respectively. To our surprise, the F1-score of DAN_{Cornell} outperforms DialogueGCN (when the learning rate is 0.000085, DAN_{Cornell} achieves an F1-score of 64.68%, which is a 0.5% improvement over DialogueGCN). Although DAN_{Ubuntu} does not perform as well as DialogueGCN, the difference is small, and it improves 1.19% over DialogueRNN. This is because, as mentioned in Section 4.2 above, the similarity between IEMOCAP and Cornell is small, and we mainly use adversarial

networks to reduce the difference between source and target domains, without resorting to complex modeling inter and self-party dependency. For individual labels, our method also achieves a good performance.

Table 4. Comparison with different methods on IEMOCAP.

Methods	IEMOCAP							
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Average (w)	
	F1						Acc.	F1
bc-LSTM	34.43	60.87	51.81	56.73	57.95	58.92	55.21	54.95
CMN	30.38	62.41	52.39	59.83	60.25	60.69	56.56	56.13
ICON	29.91	64.57	57.38	63.04	63.42	60.81	59.09	58.54
DialogueRNN	33.18	78.8	59.21	65.28	71.86	58.91	63.4	62.75
DialogueGCN	42.75	84.54	63.54	64.19	63.08	66.99	65.25	64.18
DAN _{Cornell}	48.69	74.21	64.08	62.14	71.24	60.00	65.07	64.40
DAN _{Ubuntu}	47.58	72.19	62.98	63.45	69.30	61.82	64.61	63.94

MELD: Table 5 presents the results of our proposed model and strong baselines for MELD. DAN_{Cornell} and DAN_{Ubuntu} achieve F1-score of 59.44% and 59.23%, which are 1.34% and 1.13% better than DialogueGCN. DialogueGCN is only 1.06% better than DialogueRNN.

Table 5. Comparison with different methods on MELD.

Methods	MELD								
	Neutral	Surprise	Fear	Sad	Joy	Disgust	Anger	Average (w)	
	F1						Acc.	F1	
bc-LSTM	76.23	45.10	0.00	16.11	53.17	0.00	41.06	59.43	56.44
bc-LSTM+Att	76.20	47.74	0.00	22.16	51.86	0.00	38.76	59.23	56.69
DialogueRNN	75.25	49.56	3.08	22.11	52.19	0.00	42.58	58.47	57.04
DialogueGCN	-	-	-	-	-	-	-	-	58.10
DAN _{Cornell}	77.58	54.29	0.00	25.93	57.00	0.00	40.97	57.56	59.44
DAN _{Ubuntu}	77.18	52.45	0.00	29.14	57.97	0.00	39.23	57.50	59.23

The MELD is a multi-party conversations dataset, and there are more than 300 speakers in the dataset. Normally, there are several participants in each conversation, for example, in Figure 1 where there are five participants. Additionally, we also observe that many speakers in a conversation do not utter alternately, but one speaker may utter several utterances continuously. Hence, it is not easy for models such as DialogueGCN to model the speaker's information successfully.

DailyDialog: Table 6 presents the results of DialogueRNN, DialogueGCN, DAN_{Cornell} and DAN_{Ubuntu} (since the DailyDialog is seriously unbalanced, we add two additional evaluation metrics, Micro F1 and Macro F1). Table 6 clearly shows that DialogueGCN performs poorly. Performance improvement is difficult due to the imbalance of the DailyDialog dataset compared with DialogueRNN, but DAN_{Cornell} and DAN_{Ubuntu} still achieve a 1.42% and 0.82% improvement on Weighted F1-score, respectively. Besides, our proposed DAN-CDERC model outperforms baseline models in terms of Micro F1 and Macro F1. To explain this gap in performance, it is essential to understand the distribution of emotions for DailyDialog. From Figure 4, most of the utterances are emotionless, and it may not be possible to model the speaker's information successfully by using DialogueRNN and DialogueGCN, compared with IEMOCAP.

Table 6. Comparison with different methods on DailyDialog. Micro F1, Macro F1 and Weighted F1 metrics are used to evaluate classification performance.

Methods	DailyDialog						Micro F1	Macro F1	Average (w) F1
	Happy	Disgust	Surprise	Angry	Fear	Sad			
bc-LSTM	55.99	25.17	53.54	22.78	0.00	14.81	50.37	28.72	48.78
DialogueRNN	61.75	0.00	33.06	37.34	0.00	29.44	55.65	26.95	53.78
DialogueGCN	57.67	0.00	8.89	5.19	0.00	1.68	50.91	12.24	42.69
DAN _{Cornell}	61.52	29.38	48.65	41.79	20.00	36.62	56.01	39.66	55.20
DAN _{Ubuntu}	60.95	27.63	48.15	41.41	19.05	37.33	55.61	39.09	54.60

As shown in Table 7, we make statistics on the average length of the dialogue, and the average length of the utterance in the IEMOCAP, MELD and DailyDialog. The average dialogue length of the IEMOCAP dataset is around 50 utterances, while the MELD is around 10 and the DailyDialog is around eight. Moreover, the IEMOCAP has five sessions and two participants in each conversation. The MELD has multiple participants, more than 300. Although there are only two participants in each conversation in the DailyDialog dataset, they are collected from dialogues in different scenarios. That is to say, there is no relationship between the conversations. The IEMOCAP, which more easily models inter-dependencies and self-dependencies than the MELD (short conversations and many participants) and DailyDialog (short conversations and a weak correlation between conversations), have long conversations, few participants, and a strong correlation between conversations. This is why our model does not perform as well on the IEMOCAP as on MELD and DailyDialog compared with the other best models. Those models are more conducive to establishing dependencies, while our model lacks this ability.

Experiments show that our model is effective on three datasets. In addition, since the proposed model does not model the speakers' information, it is effective not only for dyadic conversations, but also for multi-party conversations.

Table 7. Statistics of the datasets.

Dataset	Avg. Dialogue	Avg. Utterance
IEMOCAP	49.23	15.81
MELD	9.57	8.07
DailyDialog	7.81	14.08

5.2. Effectiveness of the Utterance Encoder Layer

We try to replace BERT with LSTM as the classification utterance encoder. The encoder parameters of the utterance of the domain adversarial model are transferred to the encoder of the classification model, and the results are shown in Table 8.

Table 8. Impact of the Different Utterance Encoder Layer. Weighted F1 metrics are used to evaluate classification performance.

Source	Utterance Encoder	IEMOCAP	MELD	DailyDialog
Cornell	LSTM	61.24	58.51	53.08
	BERT	64.40	59.44	55.20
Ubuntu	LSTM	62.83	58.15	53.05
	BERT	63.94	59.23	54.60

When we replace the encoder, results demonstrate that BERT provides better representations of utterances than LSTM. When Cornell is the source domain, the gap is 3.16% on

IEMOCAP and 2.12% on DailyDialog. The other gaps are all around 1%. However, the effect of LSTM as an utterance encoder also exceeds the performance of using BERT as the utterance encoder without the adversarial network. This indicates that a suitable utterance encoder and domain adversarial network can jointly promote performance improvement.

Moreover, we try to employ the utterance encoder parameters of the domain adversarial network to initialize the utterance encoder parameters of the emotion recognition model. However, we find that this method is not helpful for performance improvement. The possible reason for this phenomenon is that the representations of utterances differ between generative and emotion recognition tasks in different domains.

5.3. Source Domain Size

We compare the results of different sizes of source domain dataset, comprising 0%, 10%, 20%, 50% and 100% available in the source domain (The source domain: Cornell; The target domain: IEMOCAP). Figure 5 presents the results from the IEMOCAP dataset. A primary trend can be seen from the figure, which is that as the size of the source domain dataset increases, the classification performance in the target domain gets better. Compared with the method without transfer learning (0% source domain data), the use of only 10% source domain dataset can also significantly improve, with an increase of 2.15%. This shows the effectiveness of our method. This method, based on the adversarial network, has indeed improved the improvement by learning some inherent sequence knowledge instead of just the increasing scale of the dataset.

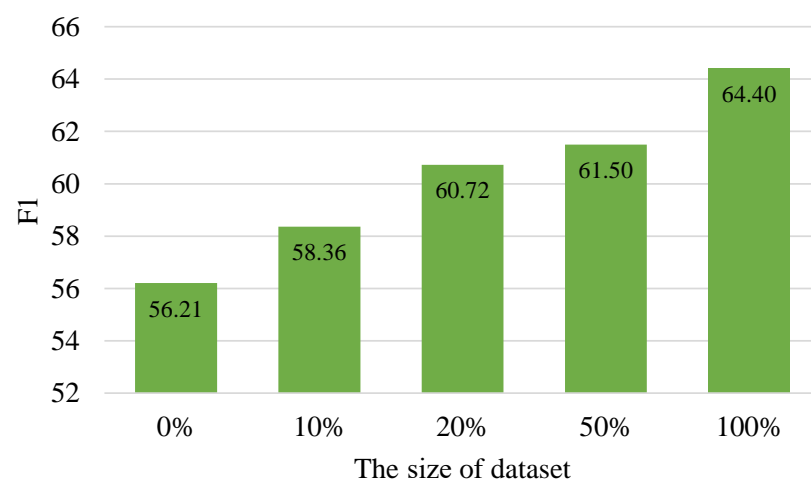


Figure 5. The results of different sizes of source domain data. (The source domain: Cornell; The target domain: IEMOCAP).

5.4. Comparison of Time and Number of Parameters

As shown in Table 9, we count the time required and the number of parameters for different methods per epoch on IEMOCAP in the inference stage.

Table 9. The time required and the number of parameters for different methods on IEMOCAP. (The second is used as a unit of time here).

Methods	Time	Parameters
DialogueRNN	18.16 s	3.30 m
DialogueGCN	3.97 s	2.78 m
DAN-CDERC	7.76 s	115.57 m

Since the proposed DAN-CDERC model uses the pre-trained model, it has the most parameters, at 115.57 million. The DAN-CDERC model takes 7.76 s per epoch, which is

between DialogueRNN and DialogueGCN. DialogueGCN used the 300 dimensional pre-trained 840B GloVe vectors [52]. Due to a significant amount of time required to process the pre-trained 840B GloVe vectors, the total time used by DialogueGCN is much more than our DAN model. So the proposed DAN-CDERC model takes the least total time. In addition, our model is simple, with only two layers (a BERT layer and a unidirectional LSTM layer). DialogueGCN has three layers (a CNN layer, a Bidirectional LSTM layer, and a GCN layer); DialogueRNN has a CNN, three GRUs, and an attention layer. They also need to model various information.

5.5. Case Studies

Table 10 presents an example from IEMOCAP. This dialogue is carried out in a pessimistic atmosphere and alternates between emotions. Due to the recognition error of U_F^{45} and the alternation of emotion in the dialogue, DialogueGCN does not perform well in the next several utterances. Paying too much attention to the previous utterances and speaker may cause this phenomenon.

We analyze predicted labels for the IEMOCAP dataset. In the confusion matrix, we find that our model is mainly misclassified in two cases. One is to mistake “Sad” and “Frustrated” as “Neutral”, and the other is to mistake “Neutral” as “Frustrated”. As can be seen from Figure 4a, “Natural” and “Frustrated” account for a large proportion, and the above results may be caused by the imbalance of emotional labels distribution. The recognition of these kinds of emotions depends on contextual emotions, which is a shortcoming of our model compared to DialogueGCN and DialogueRNN.

Table 10. Our method is compared with DialogueGCN for an example from IEMOCAP.

Turn	Utterance	Emotion	DialogueGCN	DAN _{Cornell}
T _F ⁴⁵	What the hell is this?	ang	fru	ang
T _M ⁴⁶	I'll get out. I'll get married and live some place else. Maybe, maybe New York?	fru	hap	ang
T _F ⁴⁷	Are you crazy?	ang	fru	ang
T _M ⁴⁸	Wait a minute. Tell me this. Do you mean to say that you would leave the business?	fru	fru	fru
T _F ⁴⁹	The business? The business? It doesn't inspire me?	ang	fru	ang

6. Conclusions

Given the lack of large-scale publicly available datasets, transfer learning is an effective way to alleviate this problem. We present a *Domain Adversarial Network for Cross-Domain Emotion Recognition in Conversation* (DAN-CDERC) model, consisting of two parts, namely the domain adversarial model and the emotion recognition model. The domain adversarial network employs a conversational dataset to train the generative task and a source domain and target domain dataset to train the domain discriminator for domain adaptation simultaneously. The emotion recognition model receives the transferred sequence knowledge and recognizes the emotions. When Cornell is the source dataset, the DAN-CDERC achieves an F1 of 64.40%, 59.44% and 55.20% on three datasets, all outperforming the baselines, without resorting to complex modeling inter and self-party dependency. In addition, the data scale of the source domain will have an impact on emotion recognition, but for different source domain datasets, domain similarity is more important than the data scale. Since DAN-CDERC does not model speakers' information, it is effective not only for dyadic conversations, but also for multi-party conversations. Our method proves the feasibility of using conversational datasets and domain adaptation for ERC.

Although this paper attempts to solve the problem of domain adaptation for ERC, there is inconsistency in the domain space between different tasks and different datasets, which has not been fully considered. In addition, due to the introduction of large-scale data and the use of adversarial networks in this paper, the training time of the model on the source task is long, which is also one of the shortcomings of adversarial transfer networks.

In the future, using more and faster adaptation strategies to solve the task of ERC is worthy of continuous and in-depth research.

Author Contributions: Conceptualization, H.M. and X.Z.; methodology, H.M. and X.Z.; software and experiments, H.M.; resources, Q.Z.; writing—original draft preparation, H.M.; draft revision, H.M., C.Z., X.Z., J.C. and Q.Z.; supervision, Q.Z.; project administration, Q.Z.; funding acquisition, X.Z. and Q.Z. All authors read and agreed to the published version of the manuscript.

Funding: This work was funded by National Natural Science Foundation of China Grant No. 61972133 and National Natural Science Foundation of China Grant No. 62176174.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The IEMOCAP dataset used in this study can be found at https://sail.usc.edu/iemocap/iemocap_release.htm, accessed on 12 March 2021. The MELD dataset used in this study can be found at <https://affective-meld.github.io/>, accessed on 12 March 2021. The DailyDialog dataset used in this study can be found at <http://yanran.li/dailydialog>, accessed on 12 March 2021.

Acknowledgments: The authors would like to appreciate the National Natural Science Foundation of China for supporting this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Morris, M.W.; Keltner, D. How emotions work: The social functions of emotional expression in negotiations. *Res. Organ. Behav.* **2000**, *22*, 1–50. [[CrossRef](#)]
- Poria, S.; Cambria, E.; Hazarika, D.; Majumder, N.; Zadeh, A.; Morency, L.P. Context-Dependent Sentiment Analysis in User-Generated Videos. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 873–883.
- Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; Cambria, E. Dialoguerrn: An attentive rnn for emotion detection in conversations. *Proc. Proc. Aaai Conf. Artif. Intell.* **2019**, *33*, 6818–6825. [[CrossRef](#)]
- Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; Gelbukh, A. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 154–164.
- Zhang, D.; Wu, L.; Sun, C.; Li, S.; Zhu, Q.; Zhou, G. Modeling both Context- and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019; pp. 5415–5421.
- Hazarika, D.; Poria, S.; Mihalcea, R.; Cambria, E.; Zimmermann, R. ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 2594–2604.
- Aguilar, G.; Rozgic, V.; Wang, W.; Wang, C. Multimodal and Multi-view Models for Emotion Recognition. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 991–1002.
- Gu, Y.; Yang, K.; Fu, S.; Chen, S.; Li, X.; Marsic, I. Hybrid Attention based Multimodal Network for Spoken Language Classification. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 2379–2390.
- Poria, S.; Majumder, N.; Mihalcea, R.; Hovy, E. Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances. *IEEE Access* **2019**, *7*, 100943–100953. [[CrossRef](#)]
- Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [[CrossRef](#)]
- Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 527–536.
- Li, Z.; Zhang, Y.; Wei, Y.; Wu, Y.; Yang, Q. End-to-End Adversarial Memory Network for Cross-domain Sentiment Classification. In Proceedings of the IJCAI, Melbourne, Australia, 19–25 August 2017.
- Du, C.; Sun, H.; Wang, J.; Qi, Q.; Liao, J. Adversarial and domain-aware bert for cross-domain sentiment analysis. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 4019–4028.

14. Li, L.; Ye, W.; Long, M.; Tang, Y.; Xu, J.; Wang, J. Simultaneous Learning of Pivots and Representations for Cross-Domain Sentiment Classification. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 8220–8227. [[CrossRef](#)]
15. Hazarika, D.; Poria, S.; Zimmermann, R.; Mihalcea, R. Conversational Transfer Learning for Emotion Recognition. *Inf. Fusion* **2020**, *65*, 1–12. [[CrossRef](#)]
16. Serban, I.V.; Sordoni, A.; Bengio, Y.; Courville, A.; Pineau, J. Building end-to-end dialogue systems using generative hierarchical neural network models. *Proc. AAAI Conf. Artif. Intell.* **2016**, *33*, 3776–3783.
17. Serban, I.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A.; Bengio, Y. A hierarchical latent variable encoder–decoder model for generating dialogues. *Proc. AAAI Conf. Artif. Intell.* **2017**, *31*, 1.
18. Park, Y.; Cho, J.; Kim, G. A Hierarchical Latent Structure for Variational Conversation Modeling. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, New Orleans, LA, USA, 1–6 June 2018; pp. 1792–1801.
19. Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; Liu, B. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. *Proc. AAAI Conf. Artif. Intell.* **2018**, *32*, 1.
20. Varshney, D.; Ekbal, A.; Bhattacharyya, P. Modelling Context Emotions using Multi-task Learning for Emotion Controlled Dialog Generation. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, Online, 19–23 April 2021; pp. 2919–2931.
21. Li, Q.; Chen, H.; Ren, Z.; Ren, P.; Tu, Z.; Chen, Z. EmpDG: Multi-resolution Interactive Empathetic Dialogue Generation. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online), 8–13 December 2020; pp. 4454–4466.
22. Xie Y.; Pu P. Empathetic Dialog Generation with Fine-Grained Intents. In Proceedings of the 25th Conference on Computational Natural Language Learning, Online, 10–11 November 2021; pp. 133–147.
23. Ide T.; Kawahara D. Multi-Task Learning of Generation and Classification for Emotion-Aware Dialogue Response Generation. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics, Online, 6–11 June 2021; pp. 119–125.
24. Wei, J.; Feng, S.; Wang, D.; Zhang, Y.; Li, X. Attentional Neural Network for Emotion Detection in Conversations with Speaker Influence Awareness. In Proceedings of the Natural Language Processing and Chinese Computing, Dunhuang, China, 9–14 October 2019; pp. 287–297.
25. Jiang, T.; Xu, B.; Zhao, T.; Li, S. CAN-GRU: A Hierarchical Model for Emotion Recognition in Dialogue. In Proceedings of the 19th Chinese National Conference on Computational Linguistics, Haikou, China, 30 October–1 November 2020; pp. 1101–1111.
26. Sheng, D.; Wang, D.; Shen, Y.; Zheng, H.; Liu, H. Summarize before Aggregate: A Global-to-local Heterogeneous Graph Inference Network for Conversational Emotion Recognition. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online), 8–13 December 2020; pp. 4153–4163.
27. Zhang, D.; Chen, X.; Xu, S.; Xu, B. Knowledge Aware Emotion Recognition in Textual Conversations via Multi-Task Incremental Transformer. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online), 8–13 December 2020; pp. 4429–4440.
28. Ma, H.; Wang, J.; Lin, H.; Pan, X.; Zhang, Y.; Yang, Z. A multi-view network for real-time emotion recognition in conversations. *Knowl.-Based Syst.* **2022**, *236*, 107751. [[CrossRef](#)]
29. Zhu, L.; Pergola, G.; Gui, L.; Zhou, D.; He, Y. Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Bangkok, Thailand, 1–6 August 2021; pp. 1571–1582.
30. Li, J.; Lin, Z.; Fu, P.; Wang, W. Past, Present, and Future: Conversational Emotion Recognition through Structural Modeling of Psychological Knowledge. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, 16–20 November 2021; pp. 1204–1214.
31. Peng, M.; Zhang, Q.; Jiang, Y.g.; Huang, X. Cross-Domain Sentiment Classification with Target Domain Specific Information. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 2505–2513.
32. Ziser, Y.; Reichart, R. Pivot Based Language Modeling for Improved Neural Domain Adaptation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; pp. 1241–1251.
33. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-Adversarial Training of Neural Networks. *J. Mach. Learn. Res.* **2016**, *17*, 2096–2030.
34. Li, Z.; Li, X.; Wei, Y.; Bing, L.; Zhang, Y.; Yang, Q. Transferable End-to-End Aspect-based Sentiment Analysis with Selective Adversarial Learning. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 3–7 November 2019; pp. 4590–4600.
35. Qu, X.; Zou, Z.; Cheng, Y.; Yang, Y.; Zhou, P. Adversarial Category Alignment Network for Cross-domain Sentiment Classification. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MI, USA, 2–7 June 2019; pp. 2496–2508.

36. Liang, J.; Chen, S.; Zhao, J.; Jin, Q.; Liu, H.; Lu, L. Cross-culture Multimodal Emotion Recognition with Adversarial Learning. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 4000–4004.
37. Lian, Z.; Tao, J.; Liu, B.; Huang, J.; Yang, Z.; Li, R. Context-Dependent Domain Adversarial Neural Network for Multimodal Emotion Recognition. In Proceedings of the Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Shanghai, China, 25–29 October 2020; pp. 394–398.
38. Li, H.; Tu, M.; Huang, J.; Narayanan, S.; Georgiou, P. Speaker-Invariant Affective Representation Learning via Adversarial Training. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7144–7148.
39. Parthasarathy, S.; Rozgic, V.; Sun, M.; Wang, C. Improving Emotion Classification through Variational Inference of Latent Variables. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 7410–7414.
40. Gideon, J.; Khorram, S.; Aldeneh, Z.; Dimitriadis, D.; Provost, E.M. Progressive Neural Networks for Transfer Learning in Emotion Recognition. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 1098–1102.
41. Felbo, B.; Mislove, A.; Søgaard, A.; Rahwan, I.; Lehmann, S. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017; pp. 1615–1625.
42. Li, J.; Ji, D.; Li, F.; Zhang, M.; Liu, Y. HiTrans: A Transformer-Based Context- and Speaker-Sensitive Model for Emotion Detection in Conversations. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online), 8–13 December 2020; pp. 4190–4200.
43. Huang, Y.H.; Lee, S.R.; Ma, M.Y.; Chen, Y.H.; Yu, Y.W.; Chen, Y.S. EmotionX-IDEA: Emotion BERT—An Affectional Model for Conversation. *arXiv* **2019**, arXiv:1908.06264.
44. Jiao, W.; Lyu, M.R.; King, I. PT-CoDE: Pre-trained Context-Dependent Encoder for Utterance-level Emotion Recognition. *arXiv* **2019**, arXiv:1910.08916.
45. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MI, USA, 2–7 June 2019; pp. 4171–4186.
46. Danescu-Niculescu-Mizil, C.; Lee, L. Chameleons in Imagined Conversations: A New Approach to Understanding Coordination of Linguistic Style in Dialogs. In Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics, Portland, OR, USA, 23 June 2011; pp. 76–87.
47. Lowe, R.; Pow, N.; Serban, I.; Pineau, J. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Prague, Czech Republic, 29–31 July 2015; pp. 285–294.
48. Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; Niu, S. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In Proceedings of the Eighth International Joint Conference on Natural Language Processing, Taipei, Taiwan, 27 November–1 December 2017; pp. 986–995.
49. Hsu, C.C.; Chen, S.Y.; Kuo, C.C.; Huang, T.H.; Ku, L.W. EmotionLines: An Emotion Corpus of Multi-Party Conversations. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.
50. Duchi, J.; Hazan, E.; Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **2011**, *12*, 2121–2159.
51. Hazarika, D.; Poria, S.; Zadeh, A.; Cambria, E.; Morency, L.P.; Zimmermann, R. Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; pp. 2122–2132.
52. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 26–28 October 2014; pp. 1532–1543.