applied sciences

MDPI

Article

Multi-Hypergraph Neural Networks for Emotion Recognition in Multi-Party Conversations

Haojie Xu [1,2], Cheng Zheng [1], Zhuoer Zhao [1,2] and Xiao Sun [2,3,*]

[1] AHU-IAI AI Joint Laboratory, Anhui University, Hefei 230039, China
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230088, China
[3] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China
* Correspondence: sunx@iai.ustc.edu.cn

Abstract: Emotion recognition in multi-party conversations (ERMC) is becoming increasingly popular as an emerging research topic in natural language processing. Recently, many approaches have been devoted to exploiting inter-dependency and self-dependency among participants. However, these approaches remain inadequate in terms of inter-dependency due to the fact that the effects among speakers are not individually captured. In this paper, we design two hypergraphs to deal with inter-dependency and self-dependency, respectively. To this end, we design a multi-hypergraph neural network for ERMC. In particular, we combine average aggregation and attention aggregation to generate hyperedge features, which can allow utterance information to be better utilized. The experimental results show that our method outperforms multiple baselines, indicating that further exploitation of inter-dependency is of great value for ERMC. In addition, we also achieved good results on the emotional shift issue.

Keywords: emotional shift; emotion recognition in conversations; emotion recognition in multi-party conversations

Citation: Xu, H.; Zheng, C.; Zhao, Z.; Sun, X. Multi-Hypergraph Neural Networks for Emotion Recognition in Multi-Party Conversations. *Appl. Sci.* **2023**, *13*, 1660. https://doi.org/10.3390/app13031660

Academic Editors: Ya Li, Kai Yu and Yan Song

Received: 21 December 2022
Revised: 16 January 2023
Accepted: 17 January 2023
Published: 28 January 2023

1. Introduction

Emotion recognition in conversations (ERC) has attracted more and more attention because of the prevalence of dialogue behavior in various fields. The primary purpose of ERC is to recognize the emotion of each utterance in the dialogue. The recognized emotion can be used for opinion mining on social media, such as Facebook and Instagram, building conversational assistants, and conducting medical psychoanalysis [1–4]. However, ERC, especially emotion recognition in multi-party conversations (ERMC), often exhibits more difficulties than traditional text sentiment analysis due to the emotional dynamics of dialogue [4]. There are two kinds of emotional dependencies among the participants in a dialogue—inter-dependency and self-dependency. Self-dependency is the influence of what the speaker says on the current utterance. Inter-dependency is the influence of what others say on what the current speaker says. Therefore, identifying the emotion of an utterance in a multi-party dialogue depends not only on the utterance itself and its context, but also on the speaker's self-dependence and the inter-dependency [5,6].

Existing work on emotion recognition in conversations can be roughly divided into two categories: that based on recurrent neural networks and that based on graph neural networks. Some recent works based on recurrent neural networks [3,7–13] began to focus on conversational context modeling and speaker-specific modeling, and some works [14] have even carried out multi-task learning for speaker-specific modeling on this basis. They tried to deal with speaker-dependent influences through speaker-specific modeling and conversational context modeling, but they could use other speakers' utterances to influence the current utterance well. Meanwhile, some works based on graph neural networks [5,6,15–18] have used relational graph convolutional networks (RGCNs) [19] to

distinguish among different speaker dependencies, and some have even used conversational discourse structure [6] or commonsense knowledge [18] to extend relationships among utterances. These models are intended to establish more perfect utterance relations and then aggregate according to the relations to form the influence of the surrounding utterances on the current utterance. However, the performance of such models is affected by the type and quantity of inter-utterance relations. Moreover, an emotional change in a speaker may be caused by the joint influence of multiple utterances of multiple speakers. This influence may also be caused by the interactions of utterances in different relationships. So, inter-dependency is more complex than self-dependency. We believe that it is necessary to build a graph network alone to model inter-dependency, especially for multi-dialogue, and this can allow better identification of the problem of emotional shifts between consecutive utterances of the same speaker.

Conventional graph neural networks focus on pairwise relationships between objects in the constructed graphs. In many real-world scenarios, however, relationships among objects are not dyadic (pairwise), but rather triadic, tetradic, or higher. Squeezing the high-order relations into pairwise ones leads to information loss and impedes expressiveness [20]. So, we used a hypergraph neural network [21] to deal with two kinds of speaker dependencies instead of using a conventional graph neural network. According to the hypergraph [21] structure, we know that a hyperedge may contain multiple utterances, and an utterance may belong to multiple hyperedges. We let each utterance generate a hyperedge. The nodes on the hyperedge are the corresponding current utterance and the specific context of the current utterance. Hypergraph neural networks [21] can use the structure of a hypergraph to deal with the influences of multiple utterances from multiple speakers on an utterance, that is, they use multiple surrounding utterances to produce an influence on the current utterance. By performing a node–edge–node transformation, the underlying data relationship can be better represented [21], and more complex and high-level relationships can be established among nodes [22]. Previous work has shown that speaker-specific information is very important for ERMC [3,6]. Therefore, the way of using hypergraphs for speaker-specific modeling of ERMC is a very important issue. Second, the current utterance may be influenced by utterances from different speakers. Therefore, the way of using hypergraphs for non-speaker-specific modeling of ERMC is also a very important issue.

In this paper, we construct two hypergraphs for speaker-specific and non-speaker-specific modeling, respectively. The hyperedges in the two hypergraphs are different. The hypergraph for speaker-specific modeling, where the nodes on the hyperedge are from the speaker of the current utterance, mainly deals with self-dependency. The hypergraph for non-speaker-specific modeling, where nodes on a hyperedge contain the current utterance and utterances from other speakers, is primarily used to handle inter-dependency. In Figure 1, we construct two kinds of hyperedges for the third utterance. The hyperedge of the green triangle indicates that the node of the hyperedge is from speaker B of the third utterance. The hyperedge of the blue triangle indicates that the nodes of the hyperedge are from speakers other than speaker B. Note that this hyperedge needs to contain the current utterance so that the nodes within the hyperedge have an effect on the current utterance. We use the location information and node features to aggregate to generate hyperedge features. Here, we use the location information to obtain the weight of the average aggregation, use the node features to perform the attention aggregation to obtain the attention weight, and combine the two weights to obtain the hyperedge features. Then, the hyperedge features are used to model the conversational context by using a recurrent neural network. Finally, the hyperedge features are used to aggregate to obtain new node features. The hypergraph convolution of the two hypergraphs can be used to model specific speakers and non-specific speakers so as to deal with inter-dependency and self-dependency among participants.

The main contributions of this work are summarized as follows:

- We construct hypergraphs for two different dependencies among participants and design a multi-hypergraph neural network for emotion recognition in multi-party

conversations. To the best of our knowledge, this is the first attempt to build graphs for inter-dependency alone.
- We combine average aggregation and attention aggregation to generate hyperedge features that can allow better utilization of the information of utterances.
- We conducted experiments on two public benchmark datasets. The results consistently demonstrate the effectiveness and superiority of the proposed model. In addition, we achieved good results on the emotional shift issue.

1: Dad, we can't believe you're selling the house. [Sadness] -- A

2: Well, it's time for a new family to start their memories here and hopefully their check will clear before they find the crack in the foundation and the asbestos in the ceiling. [Neutral] -- B

3: Let's grab our stuff and get the hell out of here. [Neutral] -- A

4: I'm sorry we can't store your childhood things anymore. [Sadness] -- B

5: Oh, I can't wait to see everything again! All of the memories. [Joy] -- C

6: Well, I don't know what's in the boxes down here, but I do know there are six or seven. [Neutral] -- B

7: I used to love to play restaurant. [Neutral] -- C

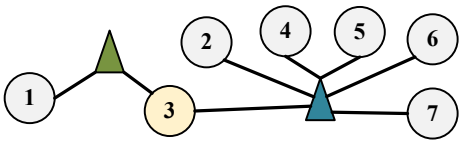

Figure 1. Conversation as a hypergraph. Circles and triangles represent nodes and hyperedges, respectively. A. B, C are participants in the conversation.

2. Related Work

2.1. Emotion Recognition in Conversations

In the following paragraphs, we divide the related works into two categories according to their methods to model a conversation's context. Note that, here, we regard the network by using Transformer [23] without actually building a graph as a model based on a recurrent neural network. DialogXL [24], BERT+MTL [14], and ERMC-DisGCN [6] have been used for some research on emotion recognition in multi-party conversations.

Recurrence-Based Models. ICON [8] separately models each speaker's historical utterances through GRU [25] and uses an additional GRU to model the impacts between speakers. DialogueRNN [3] uses three GRUs to model the speaker, the context given by the preceding utterances, and the emotion behind the preceding utterances, respectively. COSMIC [13], which was built on DialogueRNN, uses commonsense knowledge (CSK) to learn interactions among participating interlocutors. BiERU [10] involved the design of a generalized neural tensor block (GNTB) to generate contextual utterance vectors by taking the context and current utterance as inputs, and then extracting features from the contextual utterance vector by using a two-channel model (LSTM [26] and CNN [27]). EmoCaps [28] introduced the concept of emotion vectors into multi-modal emotion recognition and involved the proposal of a new emotion feature extraction structure, Emoformer. BERT+MTL [14] exploits speaker identification as an auxiliary task to enhance the representation of utterances in conversations. DialogueCRN [12] is used to understand the conversational context from a cognitive perspective and integrates emotional cues through a multi-turn reasoning module for classification. VAE-ERC [29] models the context-aware latent utterance role with a latent variable to overcome the lack of utterance role annotation in ERC datasets. TODKAT [30] proposes a new model in which the transformer model fuses topical knowledge and CSK to predict the emotion label. DialogXL [24] improves XLNet [31] with enhanced memory and dialog-aware self-attention. CoG-BART [32] utilizes supervised contrastive learning in ERC and incorporates response generation as an auxiliary task when certain contextual information is involved.

Graph-Based Models. DialogueGCN [5] used a context window to connect the current utterance with surrounding utterances and treated each dialogue as a graph. RGAT [15]

used BERT [33] to obtain contextual information and proposed relational position encodings to use the location information of utterances. ConGCN [34] regarded both speakers and utterances as graph nodes and the whole dataset as one graph. DAG-ERC [17] involved the design of a directed acyclic graph neural network and provided a method for modeling the information flow between the remote context and local context. SKAIG [18] utilized CSK to enrich edges with knowledge representations and process a graph structure with a graph transformer. MMGCN [35] was the proposal of a new model based on a multimodal fused graph convolutional network. TUCORE-GCN [36] proposed a context-aware graph convolutional network model by focusing on how people understand conversations. ERMC-DisGCN [6] involved the design of a relational convolution to leverage the speaker self-dependency of interlocutors to propagate contextual information and the proposal of an utterance-aware graph neural network.

2.2. Hypergraph Neural Networks

Unlike conventional graph neural networks, hypergraph neural networks no longer focus on only pairwise interactions between nodes. As shown in Figure 1, the relationship between hyperedges and nodes is a many-to-many relationship.

Zhou et al. [37] were the first to introduce hypergraphs in order to represent complicated relationships, and they proved that hypergraph-based learning outperformed graph-based learning on several clustering, embedding, and classification tasks. HGNN [21] was the proposal of a hypergraph neural network framework, and its ability to model complex high-order data dependencies through hypergraph structures was demonstrated. HyperGAT [38] used subject words to construct hypergraphs for text classification. HGC-RNN [22] adopted a recurrent neural network structure to learn temporal dependencies from data sequences and performed hypergraph convolution operations to extract hidden representations of data. HWNN [20] was the proposal of a graph-neural-network-based representation learning framework for heterogeneous hypergraphs, an extension of conventional graphs, which could characterize multiple non-pairwise relations well. SHARE [39] constructed different hyperedges through sliding windows of different sizes and extracted user intent through hypergraph attention for session-based recommender systems. MHGNN [40] used multi-hypergraph neural networks to explore the latent correlations among multiple physiological signals and the relationships among different subjects. HOTL [41] was the proposal of a new online cross-topic ECG emotion recognition method that used online transfer learning based on hypergraphs and effectively handled the online cross-subject scenario in which unknown target ECG data arrived one by one with varying overtime. Shao et al. [42] used hypergraphs of each modality to represent the complex relationships among subjects and used multimodal physiological signals for emotion recognition through an edge-weighted hypergraph neural network.

To the best of our knowledge, there is currently no work on building graph networks by using non-specific speaker contexts alone or on dealing with inter-dependency among speakers by using hypergraphs. In order to better capture the high-order relationships among utterances and model the two speaker dependencies, we treat dialogue as a hypergraph and solve the ERC task by using a hypergraph neural network.

3. Methodology

3.1. Hypergraph Definition

A hypergraph is defined as $HG = (V, E)$, where $V = \{v_1, v_2, \ldots, v_N\}$ is a node set, and $E = \{HE_1, HE_2, \ldots, HE_N\}$ is a collection of hyperedges. A hyperedge HE_m is a subset of the node set V, that is, the node set belonging to hyperedge HE_m is a subset of V. The structure of a hypergraph HG can also be represented by an incidence matrix A, with entries defined as in Equation (1):

$$A_{ij} = \begin{cases} 0, & v_i \notin HE_j, \\ 1, & v_i \in HE_j \end{cases} \tag{1}$$

We use $X = \{x_1, x_2, \ldots, x_N\}$ to denote the attribute vector of nodes in the hypergraph. So, the hypergraph can also be represented by $HG = (A, X)$. In this paper, we use matrix M to store the relative position weight of an utterance in the hypergraph. The structure of matrix M is similar to that of the incidence matrix A. Each row in M corresponds to a hyperedge, and the non-zero items in each row represent the utterance node in this hyperedge. The size of the non-zero items is related to the positions between nodes in the hyperedge. In the following, we use $HG = (M, X)$ to represent the hypergraph.

Vertices. Each utterance in a conversation is represented as a node $v_i \in V$. Each node v_i is initialized with the utterance embeddings h_i. We update the embedding representations of vertices via hypergraph convolution.

Hyperedge. Since each hyperedge is generated based on a specific current utterance, we need to calculate the influences of other utterances on the current utterance, and these influences will be weakened according to the relative position between the utterances. We set the position weight of the current utterance to 1, and the position weight of the remaining utterances gradually decreases with the relative distance. See Algorithm 1 for the specific process of hypergraph and hyperedge construction.

Algorithm 1 Constructing a Hypergraph

Input: the dialogue $\{h_1, h_2, \ldots, h_N\}$, speaker identity $p(\cdot)$, and context window w.
Output: $SSHG, NSHG$.

$X \leftarrow \{h_1, h_2, \ldots, h_N\}$
$M_{SSHG}, M_{NSHG} \leftarrow \emptyset, \emptyset$
for all $i \in [1, N]$ **do**
  $M^i_{SSHG}, M^i_{NSHG} \leftarrow \{0, 0, \ldots, 0\}, \{0, 0, \ldots, 0\}$ // N zero in total
  w_p, w_f, count $\leftarrow i - w, i + w, 0$ // $w_p, w_f \in [1, N]$
  $M^i_{NSHG}[i] \leftarrow 1$
  for $j = w_p; j <= w_f; j++$ **do**
    if $p(h_i) = p(h_j)$ **then**
      $M^i_{SSHG}[j] \leftarrow 1/(1 + \textbf{abs}(i - j))$
      count++
    else if $p(h_i)! = p(h_j)$ and $count = 0$ **then**
      $M^i_{NSHG}[j] \leftarrow 1/(1 + \textbf{abs}(i - j))$
    end if
  end for
end for
$SSHG \leftarrow (M_{SSHG}, X)$
$NSHG \leftarrow (M_{NSHG}, X)$
return $SSHG, NSHG$

We designed two kinds of hypergraphs—one is speaker-specific hypergraph (SSHG), and the other is a non-speaker-specific hypergraph (NSHG). The hyperedges in the SSHG are speaker-specific hyperedges (SSHEs). We selected some utterances in the context window to add to the SSHEs, and the speaker of these utterances was the same as the speaker of the current utterance. The hyperedges in the NSHG were non-speaker-specific hyperedges (NSHEs). We take the past utterance of the speaker of the current utterance as a selective constraint and selected some utterances in the context window to add to the NSHEs. The speakers of these utterances were different from the speaker of the current utterance.

3.2. Problem Definition

Given the conversation record and the speaker information for each utterance, the ERC task is that of identifying the emotional label of each utterance. More specifically, an input sequence containing N utterances $\{u_1, u_2, \ldots, u_N\}$ is given, and it is annotated with a

sequence of emotion labels $\{y_1, y_2, \ldots, y_N\}$. Each utterance u_i is spoken by $p(u_i)$. The task of ERC aims the prediction of the emotion label y_i for each utterance u_i.

3.3. Model

An overview of our proposed model is shown in Figure 2, which consists of a feature extraction module, a hypergraph convolution layer module, and an emotion classification module. Hyperedges are generated according to the third, fourth, and fifth utterances.

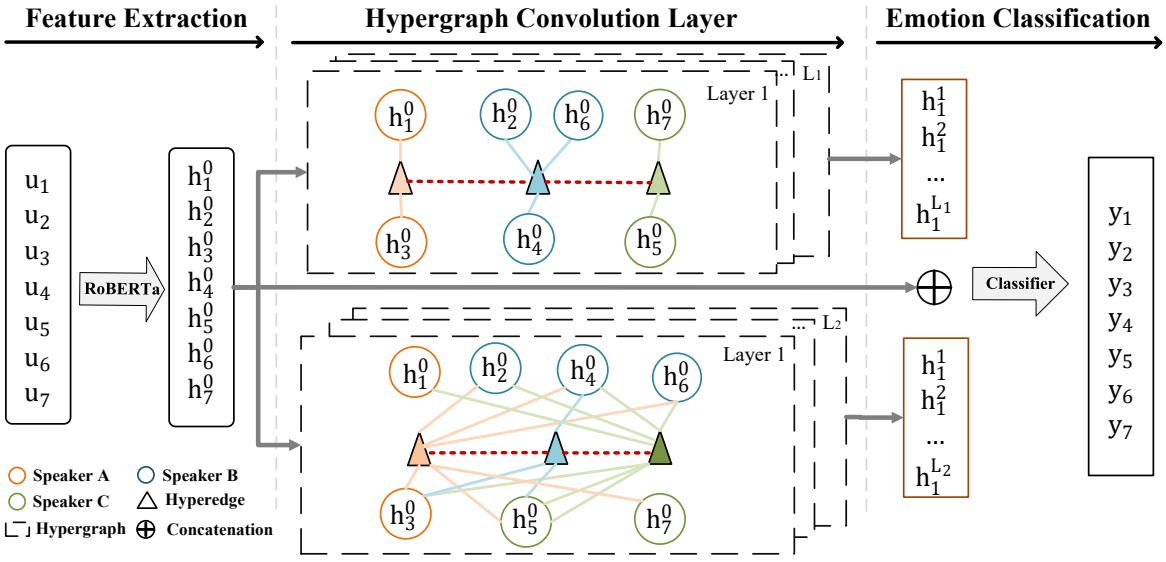

Figure 2. Overview of our proposed model. In the hypergraph convolutional layer module, the red dotted line represents the information transfer between the hyperedges.

3.3.1. Utterance Feature Extraction

Following COSMIC [13], we employed RoBERTa-Large [43] as a feature extractor. The pre-trained model was first fine-tuned on each ERC dataset, and its parameters were then frozen while training our model. More specifically, a special token [CLS] was appended at the beginning of the utterance to create the input sequence for the model. Then, we used the [CLS]'s pooled embedding at the last layer as the feature representation h_i of u_i.

3.3.2. Hypergraph Convolution (HGC) Layer

We utilized the two hypergraphs to perform separate hypergraph convolutions, and then obtained different utterance representations. The process of performing hypergraph convolution for each graph can be divided into the following three steps.

Node-to-Edge Aggregation. The first step was the aggregation from the nodes to the hyperedges. Here, we used the position weight m_j^i to calculate the weight α_{ji}^{pos} of the weighted average aggregation. Since some nodes on a hyperedge are informative, but others may not be, we should pay varying attention to the information from these nodes while aggregating them together. We utilized an attention mechanism to model the significance of different nodes. Here, we used a function $S(\cdot, \cdot)$ to calculate the attention weights α_{ji}^{ATT}. Function $S(\cdot, \cdot)$ was derived from the scaled dot-product attention formula [23]. Then, the

obtained weight α_{ji}^{pos}, attention weight α_{ji}^{ATT}, and node information h_i^{l-1} were aggregated to obtain the hyperedge feature f_j^l. The specific process is shown in Equations (2)–(5).

$$\alpha_{ji}^{pos} = \frac{m_j^i}{\sum_{k|v_k \in HE_j} m_j^k} \tag{2}$$

$$\alpha_{ji}^{att} = \frac{S(W_1 h_i^{l-1}, t^l)}{\sum_{f|v_f \in HE_j} S(W_1 h_f^{l-1}, t^l)} \tag{3}$$

$$f_j^l = \sigma(\sum_{v_i \in HE_j} \alpha_{ji}^{pos} \alpha_{ji}^{att} W_2 h_i^{l-1}) \tag{4}$$

$$S(a, b) = \frac{S(a^T b)}{\sqrt{D}} \tag{5}$$

where HE_j is the j-th hyperedge resulting from the j-th utterance. m_j^i is stored in the association matrix M, which represents the size of the position weight of the i-th node in the j-th hyperedge. h_i^{l-1} represents the features of the utterance node. t^l represents a trainable node-level context vector for the l-th HGC layer. W_1 and W_2 is a trainable parameter matrix. D is the dimension size.

Edge-to-Edge Aggregation. The second step was to transfer information between hyperedges. In order to make the current utterance have a better interaction with the context, we used the hyperedge generated by each utterance to model the conversation context. We used BiLSTM to complete the information transfer. The specific process is shown in Equation (6).

$$q_j^l, hidden_j = \overleftrightarrow{LSTM}^c(f_j^l, hidden_{j-1}) \tag{6}$$

where $hidden_j$ is the j-th hidden state of the LSTM, and q_j^l represents the hyperedge feature obtained after the information is passed by the hyperedge.

Edge-to-Node Aggregation. To update the feature for a node, we needed to aggregate the information from all of its connected hyperedges. We also used $S(\cdot, \cdot)$ to calculate the similarity between the node and hyperedge features. The specific process is shown in Equations (7) and (8).

$$h_i^l = \sigma(\sum_{HE_j \in E_i} \beta_{ij} W_3 q_j^l) \tag{7}$$

$$\beta_{ij} = \frac{S(W_4 q_j^l, W_1 h_i^{l-1})}{\sum_{HE_p \in E_i} S(W_4 q_p^l, W_5 h_i^{l-1})} \tag{8}$$

where E_i is the set of hyperedges containing the i-th node. W_3, W_4, and W_5 denote trainable parameters, and $S(\cdot, \cdot)$ is the same as in Equation (5).

3.4. Classifier

We concatenated the hidden states of the two hypergraphs in all HGC layers and passed them through a feedforward neural network to obtain the predicted emotion. The specific process is shown in Equations (9)–(11).

$$H_i^{HG} = \|_{l=1}^{L_{HG}} (h_{HG})_i^l \tag{9}$$

$$P_i = softmax(W_{smax}[h_i^0 : H_i^{SSHG} : H_i^{NSHG}] + b_{smax}) \tag{10}$$

$$\hat{y}_i = argmax_k(P_i[k]) \tag{11}$$

where H_i^{HG} represents the result of the hypergraph convolution performed on the hypergraph, HG can be SSHG and NSHG, and L_{HG} is the number of layers for the hypergraph convolution of the corresponding hypergraph.

For the training of ERMC-MHGNN, we employed the standard cross-entropy loss as an objective function. The specific function is shown in Equation (12).

$$\mathcal{L}(\theta) = -\sum_{i=1}^{C}\sum_{t=1}^{N_i} \mathrm{Log} P_{i,t}[y_{i,t}] \tag{12}$$

where C is the number of training conversations, N_i is the number of utterances in the i-th conversation, $y_{i,t}$ is the ground-truth label, and θ is the collection of trainable parameters of ERMC-MHGNN.

4. Experimental Setting

4.1. Datasets

We evaluated our model on two ERC datasets. Their statistics are shown in Table 1. They were all multimodal datasets, but our task mainly focused on textual modality for conducting our experiments.

MELD [44] was derived from the Friends TV series. The utterances were annotated with one of seven labels, namely, neutral, joy, surprise, sadness, anger, disgust, and fear. The dataset consisted of multi-party conversations and involved too many plot backgrounds. Non-neutral emotions accounted for 53%.

EmoryNLP [45] was also collected from the Friends TV scripts, but differed from MELD in the choice of scenes and emotion labels. The emotion labels included neutral, sad, mad, scared, powerful, peaceful, and joyful. Non-neutral emotions accounted for 73%.

Table 1. The statistics of the datasets . avg_utt denotes the average number of utterances.

	MELD	**EmoryNLP**
#Dial.	1432	897
Train	1038	713
Dev.	114	99
Test	280	85
#Utt.	13,708	12,606
Train	9989	9934
Dev.	1109	1344
Test	2610	1328
avg_utt	9.57	14.05
Classes	7	7
Metrics	Weighted-average F1	Weighted-average F1

4.2. Compared Methods

For a comprehensive evaluation of our proposed ERMC-MHGNN, we compared it with the following baseline methods.

Recurrence-Based Models: DialogueRNN [3], COSMIC [13], DialogueCRN [12], TODKAT [30], DialogXL [24], VAE-ERC [29], DialogueRNN-RoBERTa [13], CoG-BART [32], and EmpCaps [28].

Graph-Based Models: DialogueGCN [5], RGAT [15], RGAT-RoBERTa [17], DialogueGCN-RoBERTa [17], SKAIG [18], ERMC-DisGCN [6], MMGCN [35], TUCORE-GCN [36], and DAG-ERC [17].

4.3. Implementation Details

We conducted the experiments on Windows10 while using an NVIDIA GeForce GTX 1650 GPU with 4 GB of memory. We used PyTorch 1.7.0 and the CUDA toolkit 11.0. We adopted AdamW [46] as the optimizer. Table 2 provides the hyperparameter settings. For

the feature dimension, the utterance feature dimension extracted by the RoBERTA extractor was 1024, and after the linear layer, the utterance feature dimension became 100.

Table 2. Hyperparameter settings.

#	Batch size	Dropout	Lr	Window	SSHG	NSHG
MELD	32	0.1	0.001	1	1	1
EmoryNLP	16	0.4	0.0009	4	4	6

5. Results and Discussions

5.1. Overall Performance

Table 3 shows the performance of different models on the MELD and EmoryNLP test sets. We can see that our model outperformed all baselines, which demonstrated the effectiveness of our proposed model. At the same time, we found that the models using CSK on MELD generally performed better, while our model achieved good results without relying on external knowledge. In this paper, we focused on modeling the two kinds of dependencies among speakers by building multiple hypergraphs, so we did not incorporate external knowledge. On the EmoryNLP dataset, we found that the models using large-scale pre-trained models to extract features had better results. For example, both DAG-ERC and TUCORE-GCN used RoBERTa as feature extractors. These models could achieve over 39% on EmoryNLP. Our model also used RoBERTa as a feature extractor and achieved relatively better results by separately modeling the two speaker dependencies. Compared with DialogXL, BERT+MTL, and ERMC-DisGCN, our model had at least a 2% improvement in the two datasets. This also showed that our model could identify the emotions of utterances better than the previous models studied on the multi-party conversation dataset.

Table 3. Overall performance on the two datasets. '-' signifies that no results were reported for the given dataset. 'CSK' stands for a model that introduced commonsense knowledge, '$\surd$' and '$\times$' represent using 'CSK' and not using 'CSK' respectively. '*' represents the results of the model in the text-only modality.

Model	CSK	MELD	EmoryNLP
RoBERTa	$\times$	62.88	37.78
DialogueRNN	$\times$	57.03	-
+RoBERTa	$\times$	63.61	37.44
DialogueCRN	$\times$	58.39	-
VAE-ERC	$\times$	65.34	-
DialogXL	$\times$	62.4	34.73
BERT+MTL	$\times$	61.90	35.92
CoG-BART	$\times$	64.81	39.04
COSMIC	$\surd$	65.21	38.11
TODKAT	$\surd$	65.47	38.69
EmoCaps *	$\times$	63.51	-
DialogueGCN	$\times$	58.10	-
+RoBERTa	$\times$	63.02	38.10
RGAT	$\times$	60.91	34.42
+RoBERTa	$\times$	62.80	37.89
TUCORE-GCN	$\times$	62.47	36.01
+RoBERTa	$\times$	65.36	39.24
DAG-ERC	$\times$	63.65	39.02
ERMC-DisGCN	$\times$	64.22	36.38
SKAIG	$\surd$	65.18	38.88
MMGCN *	$\times$	57.72	-
ERMC-MHGNN	$\times$	**66.4**	**40.1**

5.2. Ablation Study

To investigate the impacts of the various modules in the model, we evaluated our model by separately removing two weights in the node-to-edge aggregation process in the hypergraph convolution. In addition, we conducted experiments on the hypergraph convolution with a single hypergraph. The results are shown in Table 4.

As shown in Table 4, we can see that, after removing the weights α^{att}, there was a relatively large drop in performance on both datasets. Through the attention function, the surrounding utterances could be given different weights so that the current utterance could better receive information from other utterances. Therefore, the use of attention weights α^{att} was beneficial for the aggregation of the node information. When we removed the α^{pos} weights, both datasets also had relatively large drops. The distance between utterances may affect the interaction between two utterances. Appropriately reducing the influence of surrounding utterances according to the relative distance can also cause the model to better aggregate node features to a certain extent.

Table 4. Results of the ablation study. '↓' represents the reduced performance compared with the 'Full model'.

Method	MELD	EmoryNLP
Full model	66.4	40.1
w/o α^{pos}	65.61 (↓0.79)	39.15 (↓0.95)
w/o α^{att}	65.64 (↓0.76)	39.05 (↓1.05)
w/o $SSHG$	65.3 (↓1.1)	38.93 (↓1.17)
w/o $NSHG$	65.19 (↓1.21)	38.9 (↓1.2)

When we used one hypergraph and removed the other hypergraphs, we only performed the hypergraph convolution of one hypergraph. From the results in Table 4, we can see that the performance of the model was degraded regardless of which hypergraph was removed. Here, the model also had a relatively large performance drop after removing the NSHG, which also showed that the method of modeling for non-specific speakers was feasible. In multi-party dialogues, the influences of utterances from other speakers should be considered in a targeted manner.

5.3. Effects of the Depth of the GNN and Window Sizes

We explored the relationship between model performance and the depth of ERMC-MHGNN. According to Figure 3, the best values of $\{L^{SSHG}, L^{NSHG}\}$ were {1, 1} and {4, 6} on the MELD and EmoryNLP datasets, and a 66.4% weighted-average F1 and 40.1% weighted-average F1, respectively, were obtained. Note that the convolution on the EmoryNLP dataset required more NSHG layers. This may have been related to the number of labels in the conversation and the length of the conversation. The proportion of each label in the EmoryNLP dataset was more balanced than that in the MELD dataset; the proportion of emotional shift was relatively larger, and the conversation length was also larger. Therefore, the EmoryNLP dataset needed more NSHG layers for convolution to deal with inter-dependency. The proportion of neutral labels in the MELD dataset was relatively large, and the conversation length was relatively small. Therefore, the MELD dataset did not require too many convolution layers. In general, the use of two types of hypergraphs was beneficial for understanding contextual cues and speaker dependencies and for enhancing the recognition ability of the model.

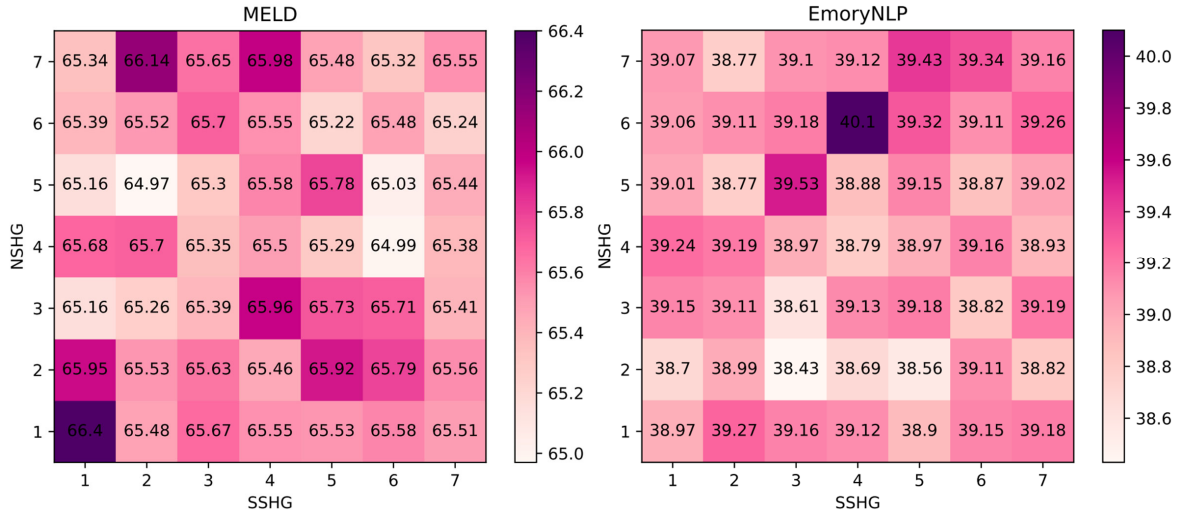

Figure 3. Effect of the depth of the GNN. We report the weighted F1 score on the MELD and EmoryNLP datasets. The darker the color, the better the performance.

We also experimented with both datasets by increasing the window size of the past and future. The experimental results are shown in Figure 4. In the figure, we can see that the window size of the context had a relatively small effect on the two datasets, but the context window sizes for obtaining relatively good results for the different datasets were not the same. In the MELD dataset, there was a relatively certain number of conversations with less than three utterances, while in the EmoryNLP dataset, the length of the conversations was generally greater than five utterances. Therefore, the MELD dataset had better results when the window size of the past and future was 1, while the EmoryNLP dataset required a relatively large context window.

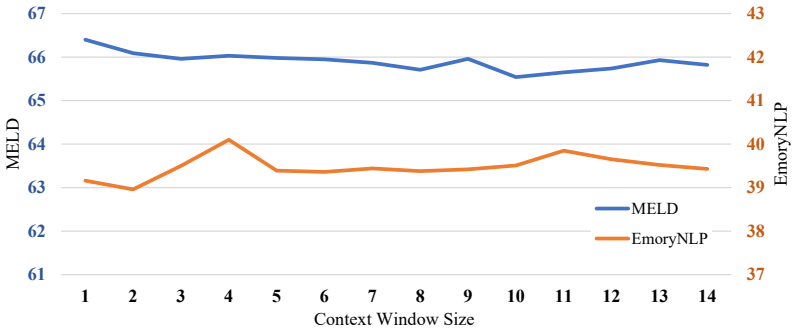

Figure 4. Effects of window sizes.

5.4. Error Analysis

We analyzed our predicted emotion labels and found that misclassifications often occurred in similar emotion classes, such as *scared* vs. *mad* and *joyful* vs. *peaceful*. In addition, some *non-neutral* labels were predicted as *neutral* labels on datasets where *neutral* labels were the majority. Figure 5 shows the confusion matrix obtained by our model on the EmoryNLP dataset.

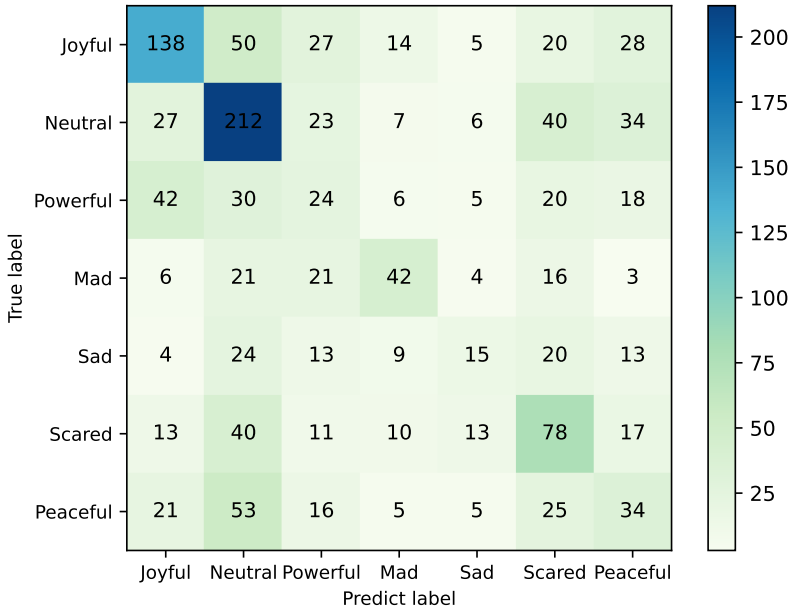

Figure 5. Heatmap of the confusion matrix of ERMC-MHGNN on the EmoryNLP dataset.

We also studied the emotional shift issue, which means that the emotions of two consecutive utterances from the same speaker were different. Since DialogXL did not provide the corresponding emotional shift prediction accuracy on the MELD and EmoryNLP datasets, we reproduced it. The weighted-average F1 of DialogXL on the MELD and EmoryNLP datasets was 62.67% and 35.0%, respectively, which are both higher than the results in the paper. The emotional shift prediction accuracy of DialogXL on the MELD and EmoryNLP datasets is listed in Table 5. It can be seen in Table 5 that, compared with the other two models, our model greatly improved the accuracy in identifying emotional shifts in these two multi-party dialogue datasets. However, improving the accuracy of identifying emotional shifts can easily reduce the accuracy of identifying without emotional shifts. Compared with the other models, we were able to improve the accuracy of recognizing emotional shifts while keeping the accuracy of recognition without emotional shifts at a high level.

Table 5. Test accuracy of ERMC-MHGNN and the partial baseline models on samples with an emotional shift and without one. '()' indicates the number of samples.

#	MELD		EmoryNLP	
	Shift (1003)	**w/o Shift** (861)	**Shift (673)**	**w/o Shift (361)**
DialogXL	57.33	71.43	33.88	43.77
DAG-ERC	59.02	69.45	37.29	42.10
ERMC-MHGNN	62.01	72.36	38.93	41.83

5.5. Case Study

For a comprehensive understanding of our proposed method, we allowed its performance to be visualized through a case study, which was selected from the EmoryNLP test dataset. As illustrated in Figure 6, our model and the baseline model both made mistakes when predicting the emotions of utterance (2) and utterance (3). By checking the video corresponding to this conversation, we found that Emily had already started checking in for boarding. Ross was afraid that Emily could not see him coming to find her, so Ross cried out to Emily anxiously. In this dialogue, if we only consider the text features, we will lack some emotional information, so we cannot accurately predict the corresponding

emotions. In our future work, we will use multimodal features to make up for the lack of emotional information in the text features.

Figure 6. Results of the case study on the EmoryNLP dataset, where three participants in a conversation are provided, along with their dependent historical utterances. We use green and red to highlight the right and wrong predictions.

When the conversation went to utterance (8) and utterance (9), we found that Emily's communication with Ross was interrupted, and Emily was urged to board. In addition, in combination with the semantics of utterance (9), Emily could not carry the big candy bar that Ross gave her. So, Emily's mood changed. We found that the baseline model directly predicted the emotion of utterance (9) within the emotion of utterance (8). The baseline model did not combine the context of this conversation well and did not fully consider multiple historical contexts. Our model processed the utterances of three speakers through hypergraphs and captured Emily's emotional change from 'neutral' to 'angry' in the current context by using two types of speaker dependence.

6. Conclusions

In this paper, two different hypergraphs were constructed for two speaker dependencies for the first time, and a multi-hypergraph neural network—namely, ERMC-MHGNN—was designed for multi-party conversation emotion recognition to better handle speaker dependencies. The experimental results show that ERMC-MHGNN has good performance. Furthermore, through comprehensive evaluation and ablation studies, we can confirm the advantages of ERMC-MHGNN and the impacts of its modules on performance. Several conclusions can be drawn from the experimental results. First, our approach to non-speaker-specific modeling of utterances from other speakers is feasible. Second, combining average aggregation with attention aggregation can allow better hyperedge features to be obtained. Finally, although the model achieved certain results on the emotional shift issue, the ability of the model to recognize similar emotions still needs to be enhanced.

In the future, we plan to build a hierarchical hypergraph neural network based on the existing hypergraph network to deal with interactions within a single modality and interactions among multiple modalities. We believe that hierarchical hypergraph neural networks can not only handle high-order relationships between utterances, but can also alleviate the deficiencies of single-mode features.

Author Contributions: Conceptualization, C.Z., X.S., H.X. and Z.Z.; methodology C.Z., X.S. and H.X.; software, H.X.; validation, H.X.; formal analysis, X.S.; investigation, C.Z.; resources, H.X.; data curation, H.X.; writing—original draft preparation, H.X. and Z.Z.; writing—review and editing, X.S. and C.Z.; visualization, H.X.; supervision, X.S. and C.Z.; project administration, H.X.; funding acquisition, X.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the General Programmer of the National Natural Science Foundation of China (61976078) and the Major Project of Anhui Province under Grant No. 202203a05020011.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Chatterjee, A.; Narahari, K.N.; Joshi, M.; Agrawal, P. SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019 ; Association for Computational Linguistics: Toronto, ON, Canada, 2019; pp. 39–48. [CrossRef]
2. Rashkin, H.; Smith, E.M.; Li, M.; Boureau, Y.L. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; Association for Computational Linguistics: Toronto, ON, Canada, 2019; pp. 5370–5381. [CrossRef]
3. Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; Cambria, E. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'19/IAAI'19/EAAI'19), Honolulu, HI, USA, 27 January–1 February 2019; AAAI Press: Palo Alto, CA, USA, 2019. [CrossRef]
4. 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]
5. 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 ; Association for Computational Linguistics: Toronto, ON, Canada, 2019; pp. 154–164. [CrossRef]
6. Sun, Y.; Yu, N.; Fu, G. A Discourse-Aware Graph Neural Network for Emotion Recognition in Multi-Party Conversation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, 16–20 November 2021; Association for Computational Linguistics: Toronto, ON, Canada, 2021; pp. 2949–2958. [CrossRef]
7. 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, Volume 1 (Long Papers), New Orleans, Louisiana, 1–6 June 2018; Association for Computational Linguistics: Toronto, ON, Canada, 2018; pp. 2122–2132. [CrossRef]
8. 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; Association for Computational Linguistics: Toronto, ON, Canada, 2018; pp. 2594–2604. [CrossRef]
9. Jiao, W.; Lyu, M.R.; King, I. Real-Time Emotion Recognition via Attention Gated Hierarchical Memory Network. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, the Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, 7–12 February 2020; AAAI Press: Palo Alto, CA, USA, 2020; pp. 8002–8009.
10. Li, W.; Shao, W.; Ji, S.; Cambria, E. BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis. *Neurocomputing* **2022**, *467*, 73–82. [CrossRef]
11. Wang, Y.; Zhang, J.; Ma, J.; Wang, S.; Xiao, J. Contextualized Emotion Recognition in Conversation as Sequence Tagging. In Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Virtual, 1–3 July 2020; Association for Computational Linguistics: Toronto, ON, Canada, 2020; pp. 186–195.
12. Hu, D.; Wei, L.; Huai, X. DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, 1–6 August 2021; Association for Computational Linguistics: Toronto, ON, Canada, 2021; pp. 7042–7052. [CrossRef]

13. Ghosal, D.; Majumder, N.; Gelbukh, A.; Mihalcea, R.; Poria, S. COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 16–20 November 2020; Association for Computational Linguistics: Toronto, ON, Canada, 2020; pp. 2470–2481. [CrossRef]
14. Li, J.; Zhang, M.; Ji, D.; Liu, Y. Multi-Task Learning with Auxiliary Speaker Identification for Conversational Emotion Recognition. *arXiv* **2020**, arXiv:2003.01478.
15. Ishiwatari, T.; Yasuda, Y.; Miyazaki, T.; Goto, J. Relation-aware Graph Attention Networks with Relational Position Encodings for Emotion Recognition in Conversations. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; Association for Computational Linguistics: Toronto, ON, Canada, 2020; pp. 7360–7370. [CrossRef]
16. Liang, C.; Yang, C.; Xu, J.; Huang, J.; Wang, Y.; Dong, Y. S+PAGE: A Speaker and Position-Aware Graph Neural Network Model for Emotion Recognition in Conversation. *arXiv* **2021**, arXiv:2112.12389.
17. Shen, W.; Wu, S.; Yang, Y.; Quan, X. Directed Acyclic Graph Network for Conversational Emotion Recognition. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, 1–6 August 2021; Association for Computational Linguistics: Toronto, ON, Canada, 2021; pp. 1551–1560. [CrossRef]
18. 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; Association for Computational Linguistics: Toronto, ON, Canada, 2021; pp. 1204–1214. [CrossRef]
19. Schlichtkrull, M.S.; Kipf, T.N.; Bloem, P.; van den Berg, R.; Titov, I.; Welling, M. Modeling Relational Data with Graph Convolutional Networks. In *The Semantic Web, Proceedings of the 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, 3–7 June 2018, Proceedings*; Gangemi, A., Navigli, R., Vidal, M., Hitzler, P., Troncy, R., Hollink, L., Tordai, A., Alam, M., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2018; Volume 10843, pp. 593–607. [CrossRef]
20. Sun, X.; Yin, H.; Liu, B.; Chen, H.; Cao, J.; Shao, Y.; Hung, N.Q.V. Heterogeneous Hypergraph Embedding for Graph Classification. In Proceedings of the Fourteenth ACM International Conference on Web Search and Data Mining, WSDM '21, Virtual Event, Israel, 8–12 March 2021; Lewin-Eytan, L., Carmel, D., Yom-Tov, E., Agichtein, E., Gabrilovich, E., Eds.; ACM: Boston, MA, USA, 2021; pp. 725–733. [CrossRef]
21. Feng, Y.; You, H.; Zhang, Z.; Ji, R.; Gao, Y. Hypergraph Neural Networks. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, the Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, the Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, HI, USA, 27 January–1 February 2019; AAAI Press: Palo Alto, CA, USA, 2019; pp. 3558–3565. [CrossRef]
22. Yi, J.; Park, J. Hypergraph Convolutional Recurrent Neural Network. In Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD'20, Virtual Event, CA, USA, 23–27 August 2020; Gupta, R., Liu, Y., Tang, J., Prakash, B.A., Eds.; ACM: Boston, MA, USA, 2020; pp. 3366–3376. [CrossRef]
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R., Eds.; 2017; pp. 5998–6008.
24. Shen, W.; Chen, J.; Quan, X.; Xie, Z. DialogXL: All-in-One XLNet for Multi-Party Conversation Emotion Recognition. In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, 2–9 February 2021; AAAI Press: Palo Alto, CA, USA, 2021; pp. 13789–13797.
25. Chung, J.; Gülçehre, Ç.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**, arXiv:1412.3555.
26. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
27. Kim, Y. Convolutional Neural Networks for Sentence Classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Association for Computational Linguistics: Toronto, ON, Canada, 2014; pp. 1746–1751. [CrossRef]
28. Li, Z.; Tang, F.; Zhao, M.; Zhu, Y. EmoCaps: Emotion Capsule based Model for Conversational Emotion Recognition. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, 22–27 May 2022; Muresan, S., Nakov, P., Villavicencio, A., Eds.; Association for Computational Linguistics: Toronto, ON, Canada, 2022; pp. 1610–1618. [CrossRef]
29. Ong, D.; Su, J.; Chen, B.; Luu, A.T.; Narendranath, A.; Li, Y.; Sun, S.; Lin, Y.; Wang, H. Is Discourse Role Important for Emotion Recognition in Conversation? In Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, 22 February–1 March 2022; AAAI Press: Palo Alto, CA, USA, 2022; pp. 11121–11129. [CrossRef]

30. 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 (Volume 1: Long Papers), Online, 1–6 August 2021; Association for Computational Linguistics: Online, 2021; pp. 1571–1582. [CrossRef]
31. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Curran Associates Inc.: Red Hook, NY, USA, 2019.
32. Li, S.; Yan, H.; Qiu, X. Contrast and Generation Make BART a Good Dialogue Emotion Recognizer. In Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022, Virtual Event, 22 February–1 March 2022; AAAI Press: Palo Alto, CA, USA, 2022; pp. 11002–11010.
33. 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, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019 ; Association for Computational Linguistics: Toronto, ON, Canada, 2019; pp. 4171–4186. [CrossRef]
34. 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, IJCAI 2019, Macao, China, 10–16 August 2019; Kraus, S., Ed.; ijcai.org, 2019; pp. 5415–5421. [CrossRef]
35. Hu, J.; Liu, Y.; Zhao, J.; Jin, Q. MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, 1–6 August 2021; Association for Computational Linguistics: Toronto, ON, Canada, 2021; pp. 5666–5675. [CrossRef]
36. Lee, B.; Choi, Y.S. Graph Based Network with Contextualized Representations of Turns in Dialogue. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, 7–11 November 2021; Association for Computational Linguistics: Toronto, ON, Canada, 2021; pp. 443–455. [CrossRef]
37. Zhou, D.; Huang, J.; Schölkopf, B. Learning with Hypergraphs: Clustering, Classification, and Embedding. In Proceedings of the Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 4–7 December 2006; Schölkopf, B., Platt, J.C.,Hofmann, T., Eds.; MIT Press: Cambridge, MA, USA, 2006; pp. 1601–1608.
38. Ding, K.; Wang, J.; Li, J.; Li, D.; Liu, H. Be More with Less: Hypergraph Attention Networks for Inductive Text Classification. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–18 November 2020; Association for Computational Linguistics: Toronto, ON, Canada, 2020; pp. 4927–4936. [CrossRef]
39. Wang, J.; Ding, K.; Zhu, Z.; Caverlee, J. Session-based Recommendation with Hypergraph Attention Networks. In Proceedings of the 2021 SIAM International Conference on Data Mining, SDM 2021, Virtual Event, 29 April–1 May 2021; Demeniconi, C., Davidson, I., Eds.; SIAM: Philadelphia, PA, USA, 2021; pp. 82–90. [CrossRef]
40. Zhu, J.; Zhao, X.; Hu, H.; Gao, Y. Emotion Recognition from Physiological Signals using Multi-Hypergraph Neural Networks. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; pp. 610–615. [CrossRef]
41. Ye, Y.; Pan, T.; Meng, Q.; Li, J.; Lu, L. Online ECG Emotion Recognition for Unknown Subjects via Hypergraph-Based Transfer Learning. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, Vienna, Austria, 23–29 July 2022; Raedt, L.D., Ed.; 2022; pp. 3666–3672, Main Track. [CrossRef]
42. Shao, J.; Zhu, J.; Wei, Y.; Feng, Y.; Zhao, X. Emotion Recognition by Edge-Weighted Hypergraph Neural Network. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 2144–2148. [CrossRef]
43. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
44. 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; Association for Computational Linguistics: Toronto, ON, Canada, 2019; pp. 527–536. [CrossRef]
45. Zahiri, S.M.; Choi, J.D. Emotion Detection on TV Show Transcripts with Sequence-Based Convolutional Neural Networks. In Proceedings of the Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; AAAI Press: Palo Alto, CA, USA, 2018; Volume WS-18 , pp. 44–52.
46. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019.