論文紹介資料

EMNLP-IJCNLP2019

DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya and Alexander Gelbukh

JAIST 情報科学系 修士1年
林 貴斗(Hayashi Takato)

Introduction

グラフ畳み込みニューラルネット(GCN)を用いた会話中の感情認識手法であるDialogueGCNを提案する.

RNNを用いたの会話中の感情認識手法らは文脈情報を伝播しているものの，話者情報や発話の相対的な位置を考慮していない.

話者情報は，話者間の依存関係をモデル化するために必要であり，これにより話者が他の話者の感情変化をどう誘発するのか理解するのに役立つ．一方で，発話の相対的な位置は，過去の発言が未来の発話にどのように影響を与えるかやその逆について理解するのに役立つ.

ノードを発話，エッジが話者間の依存関係と会話における相対的な位置を表すような有効グラフを作成し，これらをGCNに与えることで，離れた発話間の文脈情報を伝播させる.

Methodology

感情の変化は，話者レベルの文脈である話者間依存性(inter-speaker dependency)と自己依存性(self-dependency)に左右される．

話者間依存性とは，対話者が話者に与える感情的な影響のことである．ただし，すべての対話者が同じように影響を与えるわけではない．各話者は通常，対話者に独自の方法で影響を与える．

自己依存性(感情的慣性)は，会話中に話者が自分自身に与える感情的な影響のことである．話者は，相手が変化を呼び起こさない限り，感情の慣性によって自分の感情状態を維持する傾向がある．

2つの異なる文脈情報スキーム(話者レベルの文脈，順序文脈)を組み合わせることで，会話中の感情変化の理解につながる，より優れた文脈表現を獲得できるのではないか．

Methodology

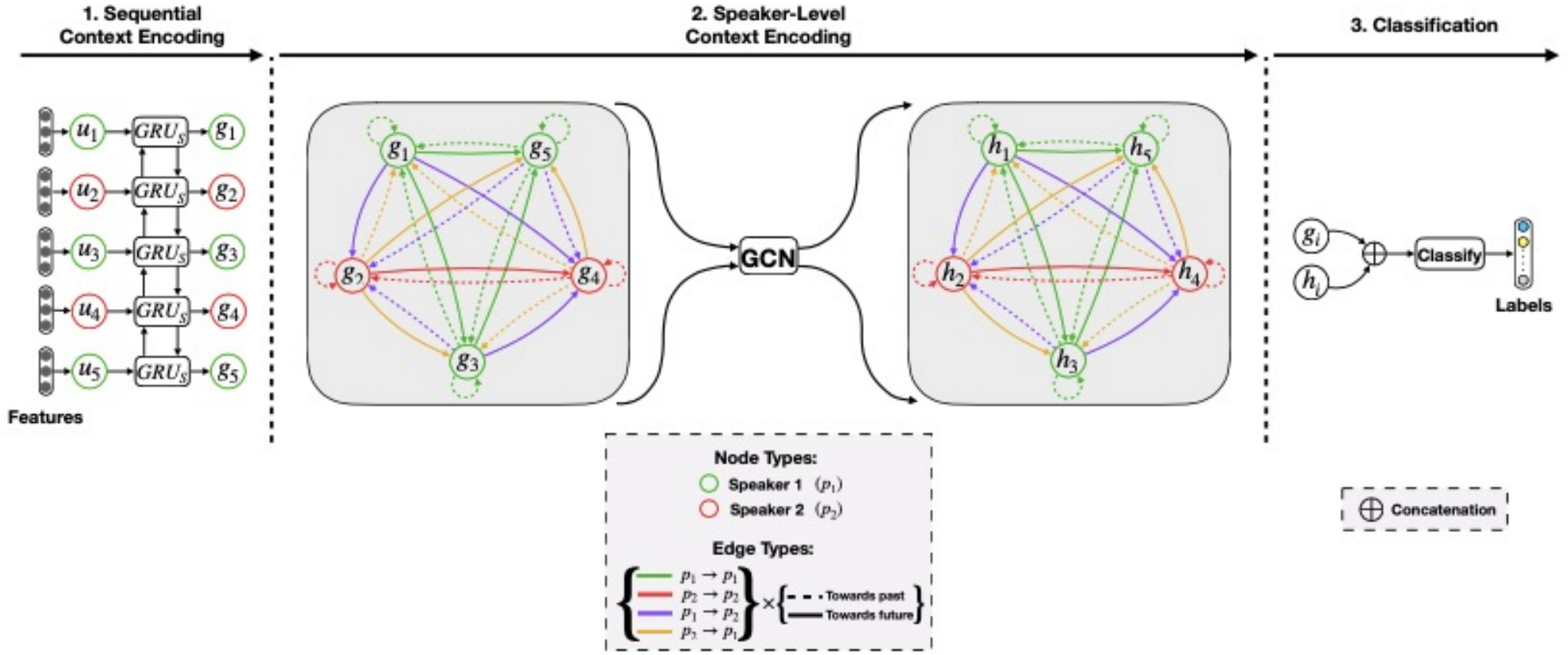
・問題の定義

M人の話者($p_1, p_{2,\cdots}, p_M$)の会話において，ある話者$p_s(u_i)$が発したu_iを構成する発話($u_1, u_{2,\cdots}, u_N$)の感情ラベル(happy, sad, neutral etc.)を予測すること

・文脈から独立したテキストの特徴抽出

CNNを用いてテキストの特徴を抽出する．CNNは1つの畳み込み層とmax pooling層，完全連結層を用いて，テキストの特徴表現を得ることができる．このネットワークは，感情ラベルを用いて発話レベルで学習される．
なお，入力には300次元の事前学習済みモデルである840B GloVeベクトルを用いる．

Methodology-Model



会話における感情認識のためのDialogueGCNの全体像は上図の通りである.
DialogueGCNは1.Sequential Context Encoder(順序文脈)，2.Speaker-Level Context Encoder(話者レベルの文脈)，3.感情分類器の3つのコンポーネントで構成されている

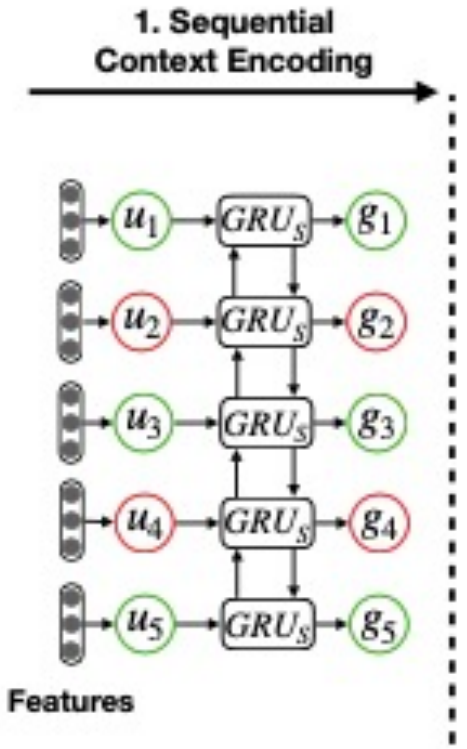
Methodology-Sequential Context Encoder

文脈から独立したテキストの特徴を双方向のGRUに入力することで，
各発話に順序文脈情報を取り込む.

$$g_i = \overleftrightarrow{GRU_S}\left(g_{i(+,-)1}, u_i\right)$$

i: $1,2,\cdots,N$
u_i: 順序文脈を考慮していないテキスト特徴
g_i: 順序文脈を考慮したテキスト特徴

各発話は話者に関係なくベクトル化されるため，この時点では最先端の
DialogueRNNとは対照的に話者情報を考慮していない.

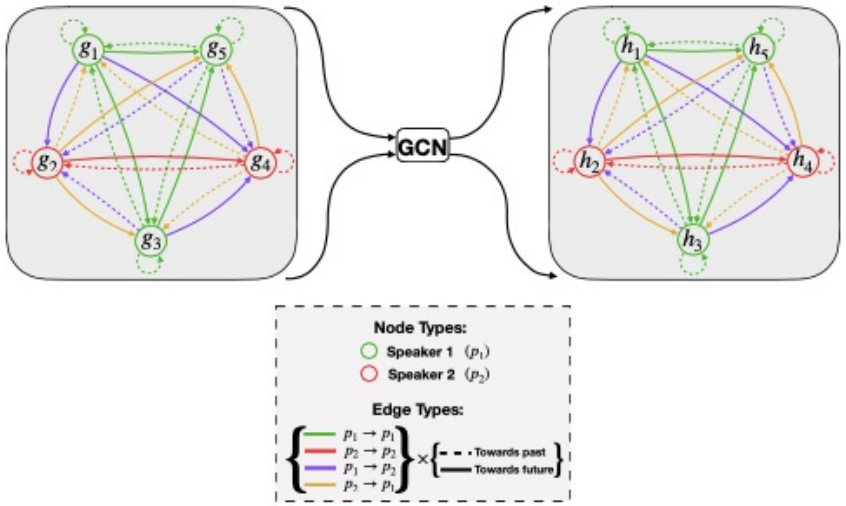


Methodology-Speaker-Level Context Encoder

文脈から話者レベルの文脈情報を取り込むためにGCNを用いたSpeaker-Level Context Encoderを提案する．話者レベルの文脈情報を得るためには，話者間依存性と自己依存性を捉えることが必要である．

各発話から有向グラフを作成し，近隣の情報を集約することによって，話者レベルの文脈情報を取り込む．

N個の発話を持つ会話は有向グラフ$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R}, \mathcal{W})$として表現される．
ここで，頂点(ノード)$v_i \in \mathcal{V}$，エッジ(関係性) $r_{ij} \in \mathcal{E}$，v_iとv_jの関係性の種類r $\in \mathcal{R}$，r_{ij}にかかる重み$\alpha_{ij} \in \mathcal{W}$である．ただし$0 \leq \alpha_{ij} \leq 1$かつi, $j \in [1,2,\cdots,N]$



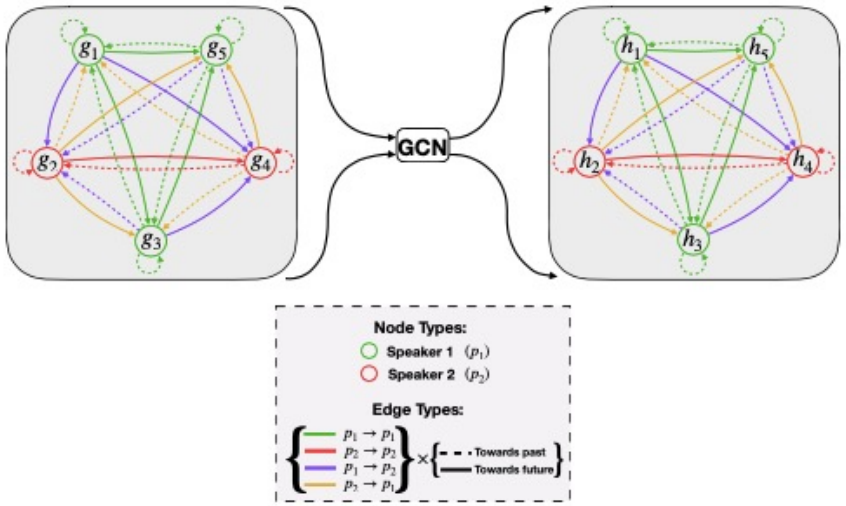
Methodology-Speaker-Level Context Encoder

頂点： $v_i \in \mathcal{V}$

各頂点v_iは順序文脈を考慮したテキスト特徴であるg_iで初期化される．このベクトルをVertex feature(頂点特徴量)と呼ぶ.

エッジ： $\mathcal{E}$

各頂点は直前p個の過去の発話($v_{i-1}, v_{i-2}, \cdots, v_{i-p}$)とf個の未来の発話($v_{i+1}, v_{i+2}, \cdots, v_{i+f}$)，そして自分自身($v_i$)との間にエッジを持つ．本研究では，pとfのウィンドウサイズはともに10個で，有効グラフであるため，2つの頂点は両方向に異なる関係のエッジを持つ.



エッジの重み：

$$\alpha_{ij} = \mathrm{softmax}\left(g_i^T W_e [g_{i-p}, \ldots, g_{i+f}]\right) \quad \text{for } j = i-p, \ldots, i+f \qquad \cdots (1)$$

各頂点に入ってくるエッジの重みの総和は1になる.
つまり，($v_{i-p}, \cdots, v_{i+f}$)からv_iに入ってくる重みの総和は1である.

Methodology-Speaker-Level Context Encoder

関係：

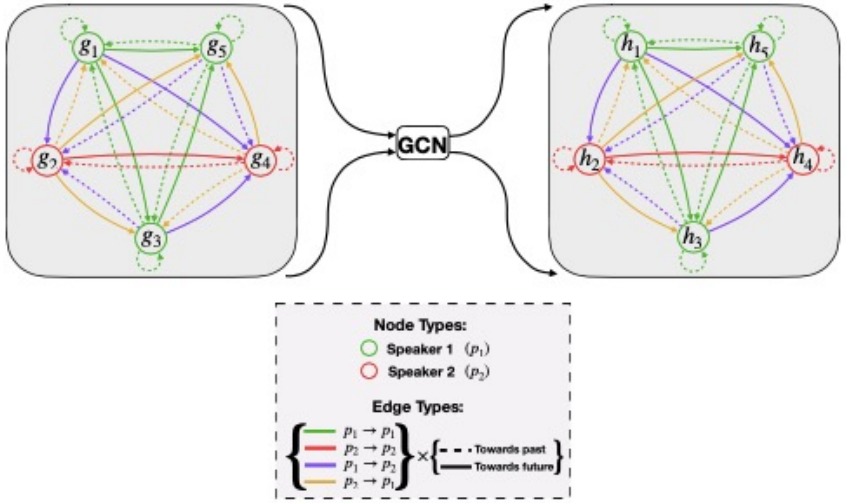
エッジr_{ij} の関係rは話者依存性と時間依存性の２つの観点から設定される.

話者依存性：
関係は発話u_iの話者$p_s(u_i)$と発話u_jの話者$p_s(u_j)$の両方に依存する.
つまり誰の発話から誰の発話へのエッジかということ.

時間依存性：
この関係は発話の相対的な位置に依存する．つまり発話の順序関係.

一つの会話にM人の話者がいる場合，グラフ$\mathcal{G}$には最大でM(u_iの話者)×M (u_jの話者)×2(u_iとu_jの順序)，つまり$2M^2$の関係性の種類を持つ.

例として，2人による会話で，u_1, u_3, u_5がp_1による発話， u_2, u_4がp_2による発話であるとき，関係性の種類は右図のようになる.



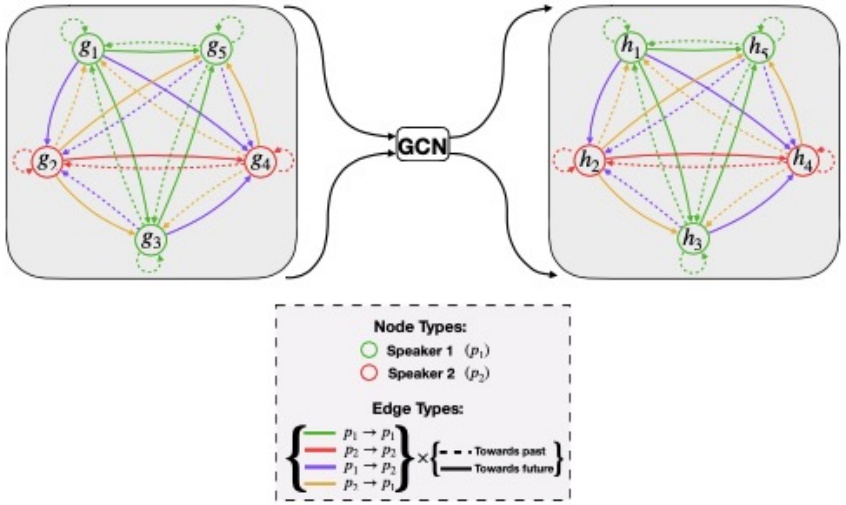
Relation	$p_s(u_i), p_s(u_j)$	$i < j$	(i, j)
1	p_1, p_1	Yes	(1,3), (1,5), (3,5)
2	p_1, p_1	No	(1,1), (3,1), (3,3) (5,1), (5,3), (5,5)
3	p_2, p_2	Yes	(2,4)
4	p_2, p_2	No	(2,2), (4,2), (4, 4)
5	p_1, p_2	Yes	(1,2), (1,4), (3,4)
6	p_1, p_2	No	(3,2), (5,2), (5,4)
7	p_2, p_1	Yes	(2,3), (2,5), (4,5)
8	p_2, p_1	No	(2,1), (4,1), (4,3)

Methodology-Speaker-Level Context Encoder

話者レベルの文脈を考慮していない特徴量ベクトルであるg_iに2段階のグラフ畳み込み処理をすることによって話者レベルの文脈を考慮した特徴量ベクトルに変換する.

まずはじめに，各頂点特徴量の局所的な近隣情報(過去10+未来10+自身1の発話)を集約することにより，頂点v_iに対して新しい特徴量ベクトル$h_i^{(1)}$が生成される.

$$h_i^{(1)} = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{j\in N_i^r}\frac{\alpha_{ij}}{c_{i,r}}W_r^{(1)}g_j + \alpha_{ii}W_0^{(1)}g_i\right) \quad \text{for } i = 1, 2, \ldots, N \qquad \cdots(2)$$



α_{ij}, α_{ii}: edge weight
N_i^r: the neighboring indices of vertex i under relation r ∈ $\mathcal{R}$
(iと関係性rをもつ頂点インデックスの集合)
$c_{i,r}$: a problem specific normalization constant
σ: an activation function such as ReLU
$W_r^{(1)}$, $W_0^{(1)}$: learnable parameters of the transformation

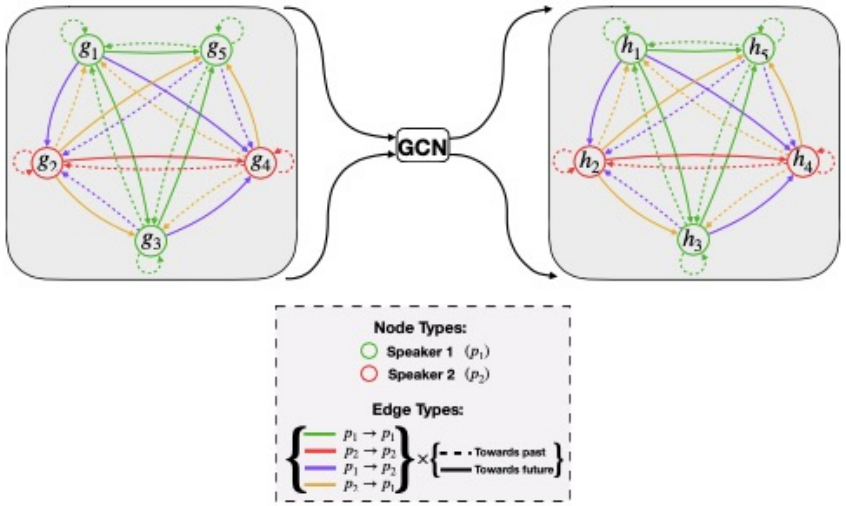
Methodology-Speaker-Level Context Encoder

次に(2)の出力に対して，局所的な近隣情報を再度集約して，新しい特徴量ベクトル$h_i^{(2)}$を生成する．

$$h_i^{(2)} = \sigma\left(\sum_{j \in N_i^r} W^{(2)} h_j^{(1)} + W_0^{(2)} h_i^{(1)}\right) \quad \text{for } i = 1, 2, \ldots, N \qquad \cdots (3)$$

σ: an activation function such as ReLU

$W_r^{(2)}$, $W_0^{(2)}$: learnable parameters of the transformation



(2)と(3)の変換を重ねることによって，グラフ内の各発話の近隣話者情報が効果的に蓄積される．また自己接続により自己依存的な特徴も蓄積される．

Methodology-Emotion Classifier

順序文脈だけを考慮したg_iと話者レベルの文脈を考慮した$h_i^{(2)}$を結合して，
attentionメカニズムを適用して最終的な発話表現が得られる.

$$h_i = \left[g_i, h_i^{(2)}\right] \quad \cdots (4)$$

$$\beta_i = \mathrm{softmax}\left(h_i^T W_\beta [h_1, h_2 \ldots, h_N]\right) \quad \cdots (5)$$

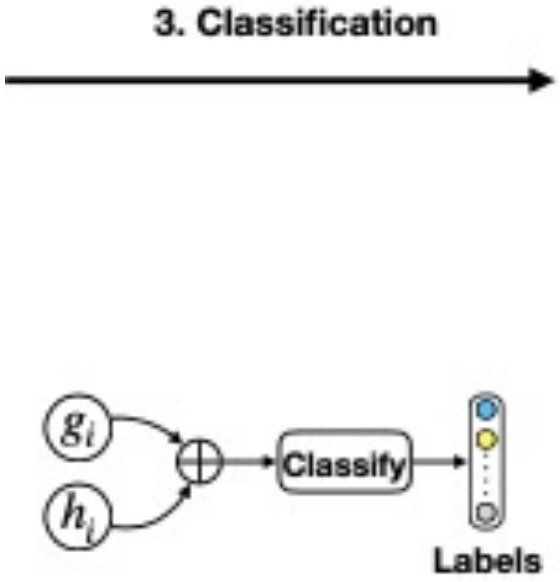
$$\tilde{h}_i = \beta_i [h_1, h_2, \ldots, h_N]^T \quad \cdots (6)$$

最後に，発話を完全連結ネットワークを使って分類する.

$$l_i = \mathrm{ReLU}\left(W_l \tilde{h}_i + b_l\right) \quad \cdots (7)$$

$$\mathcal{P}_i = \mathrm{softmax}(W_{s\max} l_i + b_{smax}) \quad \cdots (8)$$

$$\hat{y}_i = \underset{k}{\mathrm{argmax}}(\mathcal{P}_i[k]) \quad \cdots (9)$$



Methodology-Training Setup

損失関数としてL2正則化を適用した交差エントロピーを計算する.

$$L = -\frac{1}{\sum_{s=1}^{N} c(s)} \sum_{i=1}^{N} \sum_{j=1}^{c(i)} \log \mathcal{P}_{i,j}[y_{i,j}] + \lambda \|\theta\|_2 \qquad \cdots (10)$$

N: the number of samples/dialogues
$C(i)$: the number of utterances in sample i
$P_{i,j}$: the probability distribution of emotional labels for utterance j of dialogue i
$y_{i,j}$: the expected class label of utterance j of dialogue i
λ: the L2-regularizer weight
θ: the set of all trainable parameters

ネットワークの学習にはAdamをハイパーパラメーターの設定にはGrid Searchを用いた.

Experiment Setting-Datasets

Dataset	# dialogues			# utterances		
	train	val	test	train	val	test
IEMOCAP	120		31	5810		1623
AVEC	63		32	4368		1430
MELD	1039	114	280	9989	1109	2610

本研究では**IEMOCAP**(Busso et al., 2008), **AVEC**(Schuller et al., 2012), **MELD**(Poria et al, 2019)の３つのデータセットを使用してDialogueGCNを評価する．これらはテキスト，視覚，音響情報を含むマルチモーダルデータセットであるが本研究ではテキスト情報のみを用いる．

IEMOCAP: 一対一の対話データセットで各発話に6つの感情ラベルのいずれかが付与されている．

AVEC: 一対一の対話データセットで各発話に4つの実数値の感情属性がアノテーションされている．　（valance[-1.1], arousal[-1,1], expectancy[-1,1], power[-1,∞]）

MELD: TVシリーズ「Friends」から抽出された多人数会話データセットで各発話に7つの感情ラベルのいずれかが付与されている．

Results and Discussion-Comparison with State of Art and Baseline

Methods	IEMOCAP													
	Happy		Sad		Neutral		Angry		Excited		Frustrated		**Average(w)**	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
CNN	27.77	29.86	57.14	53.83	34.33	40.14	61.17	52.44	46.15	50.09	62.99	55.75	48.92	48.18
Memnet	25.72	33.53	55.53	61.77	58.12	52.84	59.32	55.39	51.50	58.30	67.20	59.00	55.72	55.10
bc-LSTM	29.17	34.43	57.14	60.87	54.17	51.81	57.06	56.73	51.17	57.95	67.19	58.92	55.21	54.95
bc-LSTM+Att	30.56	35.63	56.73	62.90	57.55	53.00	59.41	59.24	52.84	58.85	65.88	59.41	56.32	56.19
CMN	25.00	30.38	55.92	62.41	52.86	52.39	61.76	59.83	55.52	60.25	71.13	60.69	56.56	56.13
ICON	22.22	29.91	58.78	64.57	62.76	57.38	64.71	63.04	58.86	63.42	67.19	60.81	59.09	58.54
DialogueRNN	25.69	33.18	75.10	78.80	58.59	59.21	64.71	**65.28**	80.27	**71.86**	61.15	58.91	63.40	62.75
DialogueGCN	40.62	**42.75**	89.14	**84.54**	61.92	**63.54**	67.53	64.19	65.46	63.08	64.18	**66.99**	65.25	**64.18**

Methods	AVEC				MELD
	Valence	Arousal	Expectancy	Power	
CNN	0.545	0.542	0.605	8.71	55.02
Memnet	0.202	0.211	0.216	8.97	-
bc-LSTM	0.194	0.212	0.201	8.90	56.44
bc-LSTM+Att	0.189	0.213	0.190	8.67	56.70
CMN	0.192	0.213	0.195	8.74	-
ICON	0.180	0.190	0.180	8.45	-
DialogueRNN	0.168	0.165	0.175	7.90	57.03
DialogueGCN	**0.157**	**0.161**	**0.168**	**7.68**	**58.10**

提案するDialogueGCNの性能を最先端のDialogueRNNおよびベースラインモデルと比較した結果を上図に示す．すべての結果は5回の実験の平均値である．
DialogueGCNはすべてのデータセットにおいてベースラインモデルよりも優れていた．

Results and Discussion-Comparison with State of Art and Baseline

Methods	IEMOCAP													
	Happy		Sad		Neutral		Angry		Excited		Frustrated		**Average(w)**	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
CNN	27.77	29.86	57.14	53.83	34.33	40.14	61.17	52.44	46.15	50.09	62.99	55.75	48.92	48.18
Memnet	25.72	33.53	55.53	61.77	58.12	52.84	59.32	55.39	51.50	58.30	67.20	59.00	55.72	55.10
bc-LSTM	29.17	34.43	57.14	60.87	54.17	51.81	57.06	56.73	51.17	57.95	67.19	58.92	55.21	54.95
bc-LSTM+Att	30.56	35.63	56.73	62.90	57.55	53.00	59.41	59.24	52.84	58.85	65.88	59.41	56.32	56.19
CMN	25.00	30.38	55.92	62.41	52.86	52.39	61.76	59.83	55.52	60.25	71.13	60.69	56.56	56.13
ICON	22.22	29.91	58.78	64.57	62.76	57.38	64.71	63.04	58.86	63.42	67.19	60.81	59.09	58.54
DialogueRNN	25.69	33.18	75.10	78.80	58.59	59.21	64.71	**65.28**	80.27	**71.86**	61.15	58.91	63.40	62.75
DialogueGCN	40.62	**42.75**	89.14	**84.54**	61.92	**63.54**	67.53	64.19	65.46	63.08	64.18	**66.99**	65.25	**64.18**

Methods	AVEC				MELD
	Valence	Arousal	Expectancy	Power	
CNN	0.545	0.542	0.605	8.71	55.02
Memnet	0.202	0.211	0.216	8.97	-
bc-LSTM	0.194	0.212	0.201	8.90	56.44
bc-LSTM+Att	0.189	0.213	0.190	8.67	56.70
CMN	0.192	0.213	0.195	8.74	-
ICON	0.180	0.190	0.180	8.45	-
DialogueRNN	0.168	0.165	0.175	7.90	57.03
DialogueGCN	**0.157**	**0.161**	**0.168**	**7.68**	**58.10**

・**IEMOCAPとAVEC**

一対一の対話データセットであるIEMOCAPとAVECではともにDialogueGCNがベースラインを上回った．このパフォーマンスの差は，DialogueRNNとDialogueGCNはどちらも話者レベルの文脈を考慮しているのに対して，その他のベースラインモデルはこれを考慮していないことが要因だと思われる．

また， DialogueRNNとDialogueGCNの性能差は話者レベルの文脈エンコーダーの性質の違いに起因すると考えられる． IEMOCAPとAVECでは70発話以上の対話が多く含まれている． DialogueRNNではこのような長い対話の場合，長期的な情報伝播の問題があるため，話者レベルの文脈のエンコーディングがう上手くいかない．一方で， DialogueGCNでは近隣の畳み込みをすることでこの問題を解決している．

Results and Discussion-Comparison with State of Art and Baseline

Methods	IEMOCAP													
	Happy		Sad		Neutral		Angry		Excited		Frustrated		**Average(w)**	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
CNN	27.77	29.86	57.14	53.83	34.33	40.14	61.17	52.44	46.15	50.09	62.99	55.75	48.92	48.18
Memnet	25.72	33.53	55.53	61.77	58.12	52.84	59.32	55.39	51.50	58.30	67.20	59.00	55.72	55.10
bc-LSTM	29.17	34.43	57.14	60.87	54.17	51.81	57.06	56.73	51.17	57.95	67.19	58.92	55.21	54.95
bc-LSTM+Att	30.56	35.63	56.73	62.90	57.55	53.00	59.41	59.24	52.84	58.85	65.88	59.41	56.32	56.19
CMN	25.00	30.38	55.92	62.41	52.86	52.39	61.76	59.83	55.52	60.25	71.13	60.69	56.56	56.13
ICON	22.22	29.91	58.78	64.57	62.76	57.38	64.71	63.04	58.86	63.42	67.19	60.81	59.09	58.54
DialogueRNN	25.69	33.18	75.10	78.80	58.59	59.21	64.71	**65.28**	80.27	**71.86**	61.15	58.91	63.40	62.75
DialogueGCN	40.62	**42.75**	89.14	**84.54**	61.92	**63.54**	67.53	64.19	65.46	63.08	64.18	**66.99**	65.25	**64.18**

Methods	AVEC				MELD
	Valence	Arousal	Expectancy	Power	
CNN	0.545	0.542	0.605	8.71	55.02
Memnet	0.202	0.211	0.216	8.97	-
bc-LSTM	0.194	0.212	0.201	8.90	56.44
bc-LSTM+Att	0.189	0.213	0.190	8.67	56.70
CMN	0.192	0.213	0.195	8.74	-
ICON	0.180	0.190	0.180	8.45	-
DialogueRNN	0.168	0.165	0.175	7.90	57.03
DialogueGCN	**0.157**	**0.161**	**0.168**	**7.68**	**58.10**

・**MELD**

MELDに含まれる多人数会話では平均10発話しか含まれないうえに5人以上の話者がいるため，それぞれの話者の発話は少数かつ短い．そのため，話者間依存性や自己依存性のモデル化が難しく他のベースラインと比較してもそれほど大きな改善は見られなかった．

・ウィンドウサイズの効果

上図の結果は過去と未来のウィンドウサイズを(10,10)に設定して実験したものである．これを(8,8)，(4,4)，(0,0)と次第に小さくしていくとF1スコアも62%, 59%, 56%と次第に低下していく．計算上の制約から実験できなかったがウィンドウサイズを(10,10)から大きくすれば性能は更に向上すると思われる．

Results and Discussion-Ablation Study

Sequential Encoder	Speaker-Level Encoder	F1
✓	✓	64.18
✓	✗	55.30
✗	✓	56.71
✗	✗	36.75

Speaker Dependency Edges	Temporal Dependency Edges	F1
✓	✓	64.18
✓	✗	62.52
✗	✓	61.03
✗	✗	60.11

・順序文脈エンコーダーと話者レベルの文脈エンコーダー

両方の文脈エンコーダーを除去した場合，F1スコアは36.7%と非常に低くなり，会話型感情認識における文脈モデルの重要性が示された.

・話者依存のエッジと時間依存のエッジ

M人の話者がいる会話には，$2M^2$の異なるエッジ関係がある．まず時間依存性エッジのみを除去し (M^2のエッジ関係)，次に話者依存性エッジのみを除去し(2個のエッジ関係になる)，最後に両方を除去する(グラフ全体で1個のエッジ関係になる)．上の右図の結果からこれらの文脈情報を取り入れることが会話型感情認識における異なる関係のエッジの重要性を示した.