中国情感计算大会
2021

交互情境下的多模态情感识别

金琴

AI·M³，中国人民大学信息学院

2021.07.11

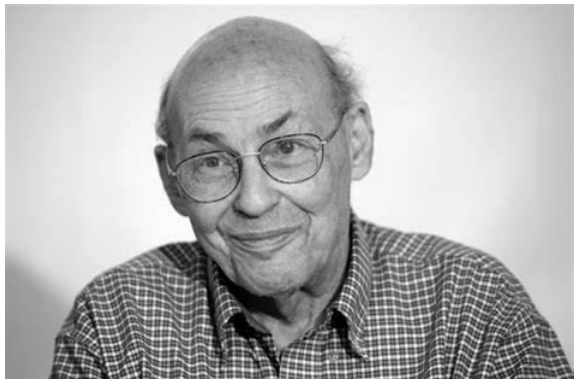
Affective Computing

➢Herbert A. Simon

《Motivational and emotional controls of cognition》

1) human thinking always takes place in, and contributes to, a cumulative process of growth and development;
2) human thinking begins in an intimate association with emotions and feelings which is never entirely lost;
3) almost all human activity, including thinking, serves not one but a multiplicity of motives at the same time.

➢Marvin Minsky



《The Society of Mind》

The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without any emotions

Affective Computing

- **Affective Computing** is computing that relates to, arises from, or deliberately influences emotion or other affective phenomena (Picard, MIT Press 1997).

Affective Computing

- **Affective Computing** is computing that relates to, arises from, or deliberately influences emotion or other affective phenomena (Picard, MIT Press 1997).

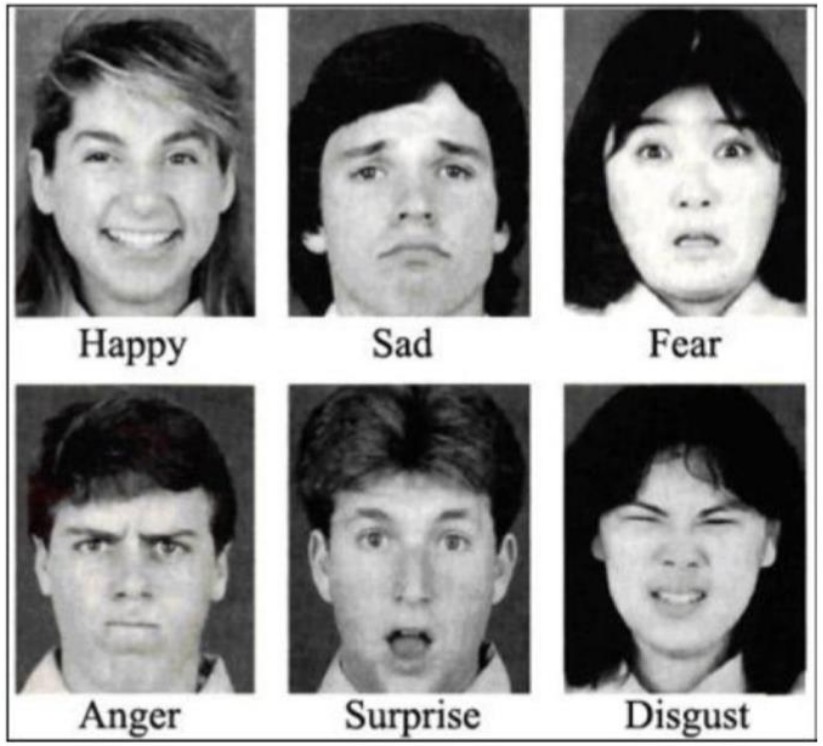
- **Develop new technologies and theories that advance basic understanding of affect and its role in human experience.**

- **Wide range of applications in Human Computer Interactions, Psychological Health Care, Education …**

Emotion Recognition

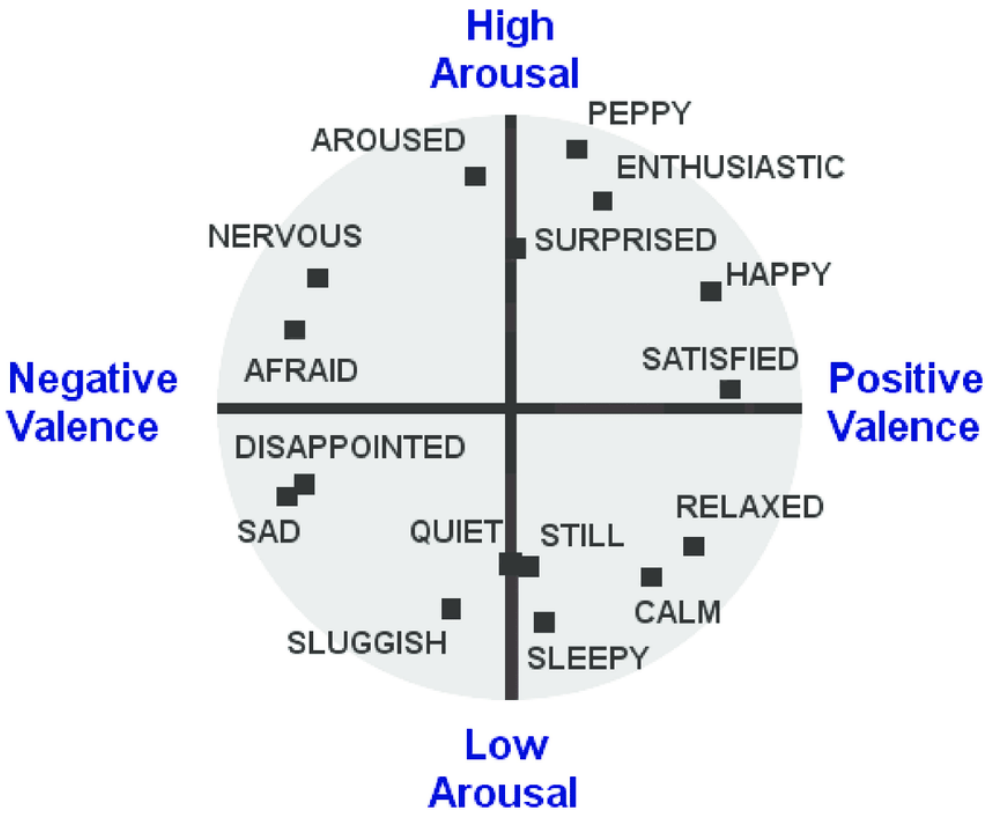
Discrete Emotion Model:

Fear, Angry, Disgust, Sad, Happy, Surprise



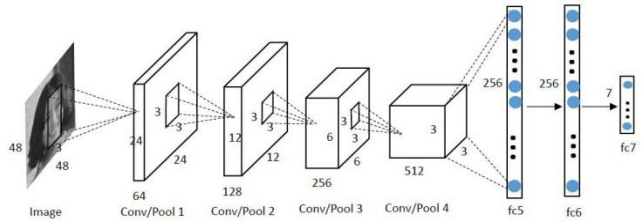
Dimensional Emotion Model:

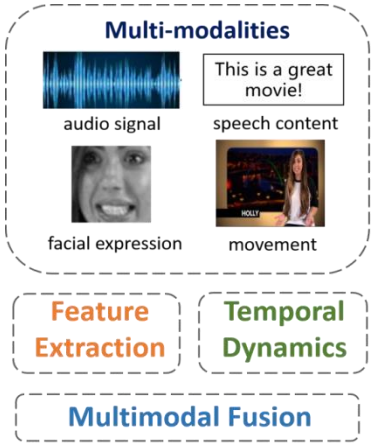
Arousal, Valence, Dominant



Multimodal Emotion Recognition

- Speech
- Text
- Facial
- Gesture/Body Language
- Bio-signals



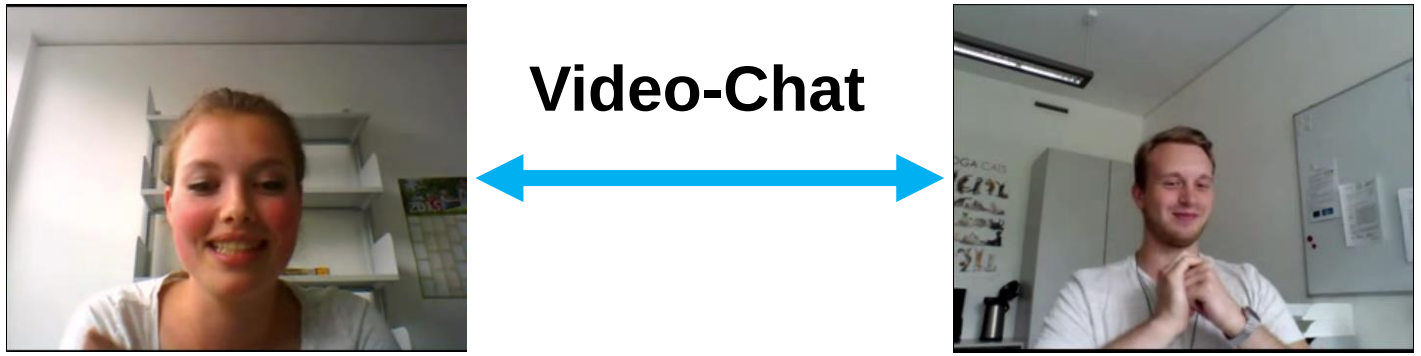


Multimodal Emotion Recognition in Conversation

- More and more multimodal Human-Computer Interaction (HCI) devices and applications emerge in our lives.

- **Emotion detection and expression ability** of intelligent systems are very important in natural human-computer interactions.

Multimodal Emotion Recognition in Conversation



Human-to-human dyadic conversation scenario

- **Conversation Context**
- **Multimodal Fusion**
- **Data Scarcity**
- **Culture Robustness**

Multimodal Conversation Context

Oh, yeah? Have you gotten letters yet?
Um, where? hang out?

Oh my Gosh, that's so cool.

U.S.C..

But big packet, big packet is nice.

So you're going to be right here in Los Angeles; that is so cool.

So can we hang out?

Oh, yay. Thank you. No but that's **an awesome school**.

I'm looking forward to it.

Hey, yes, you know that's - that's what I'm coming for.

Okay.
They have a lot of good schools, right?

Well, you know I'm leaning towards like communicational.

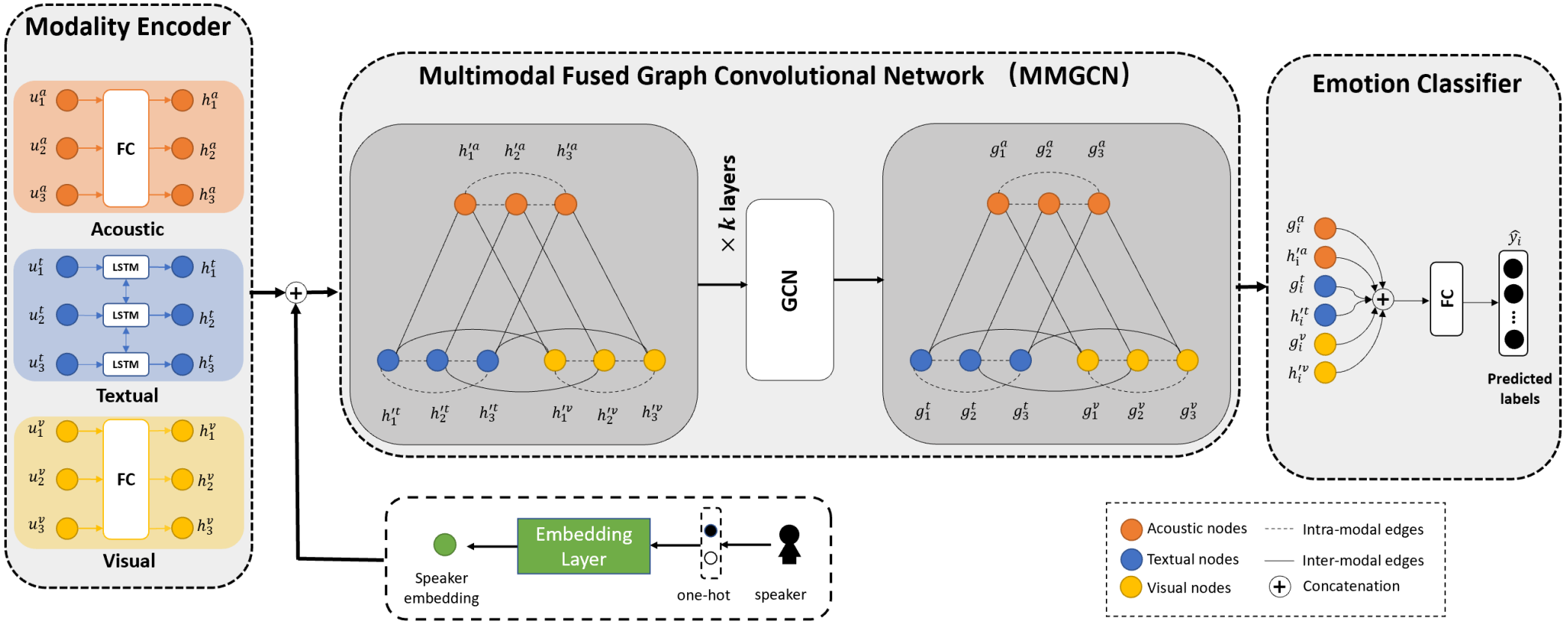
Yeah,
I mean it's just **a really good school**.

I'm looking forward to getting there.
I mean, God, the campus is cool.

- Long Distance Dependency
- Multimodal Information
- Speaker Information

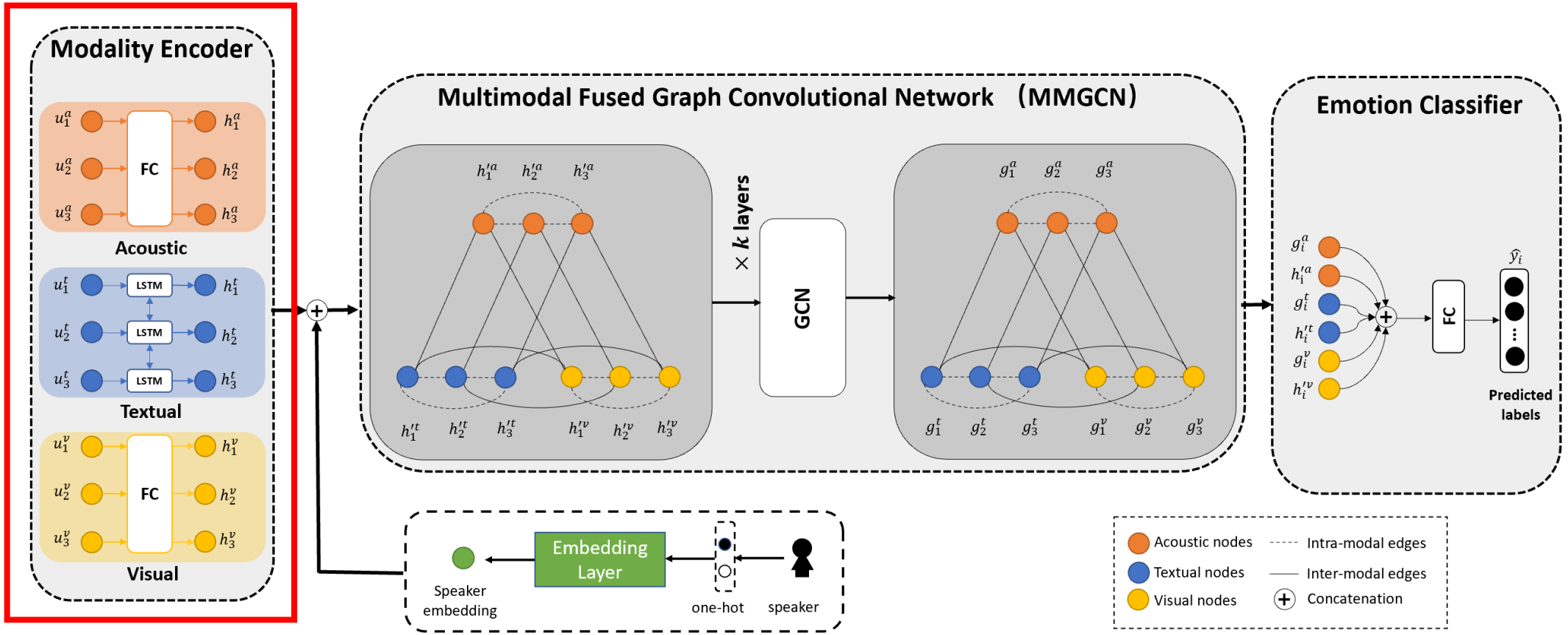
Multimodal Fusion with Graph Convolution Network

Our proposed model **MMGCN**: Dialogues naturally have graph-like structure, therefore we employ GCN with deep layers for emotion modeling in conversation (ACL 2021)



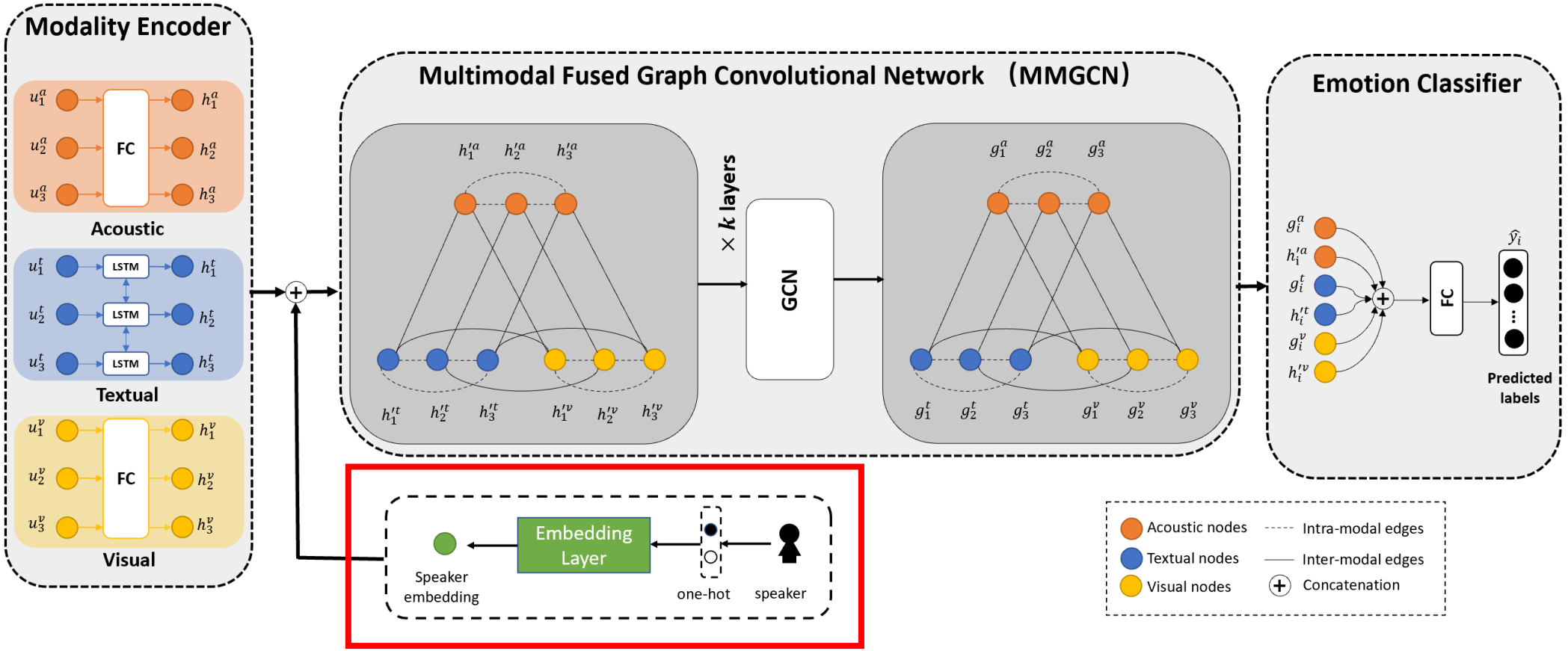
J Hu, Y Liu, J Zhao, Q Jin. **MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation.** ***ACL, 2021.***

Multimodal Fusion with Graph Convolution Network



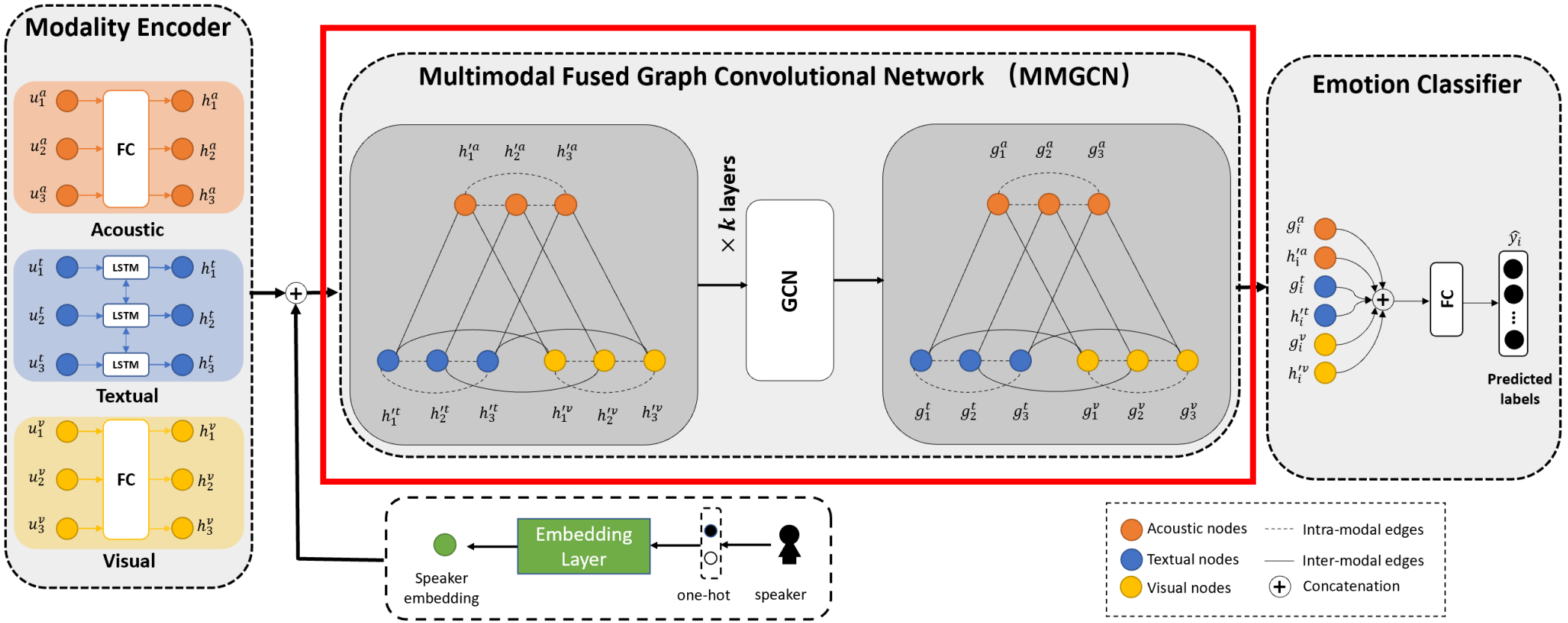
J Hu, Y Liu, J Zhao, Q Jin. MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation. *ACL, 2021.*

Multimodal Fusion with Graph Convolution Network



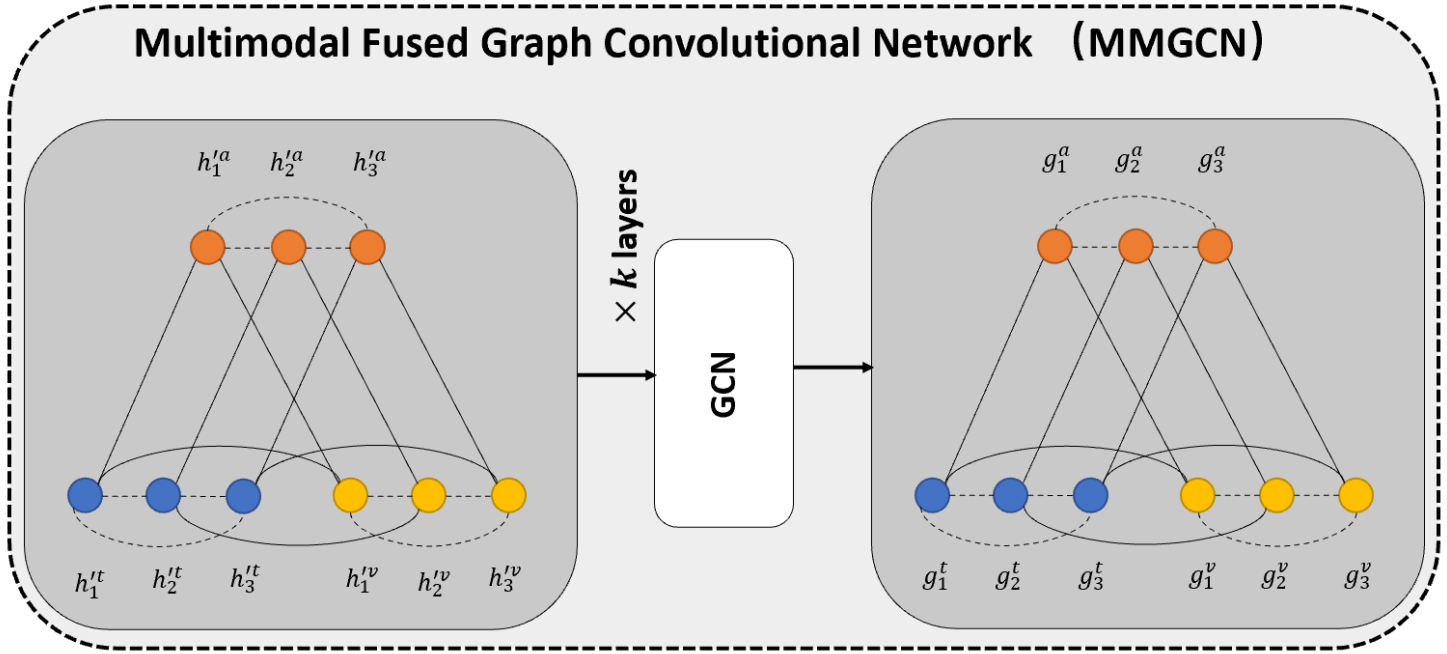
J Hu, Y Liu, J Zhao, Q Jin. MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation. *ACL, 2021.*

Multimodal Fusion with Graph Convolution Network



J Hu, Y Liu, J Zhao, Q Jin. **MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation.** ***ACL, 2021.***

Multimodal Fusion with Graph Convolution Network



Nodes:

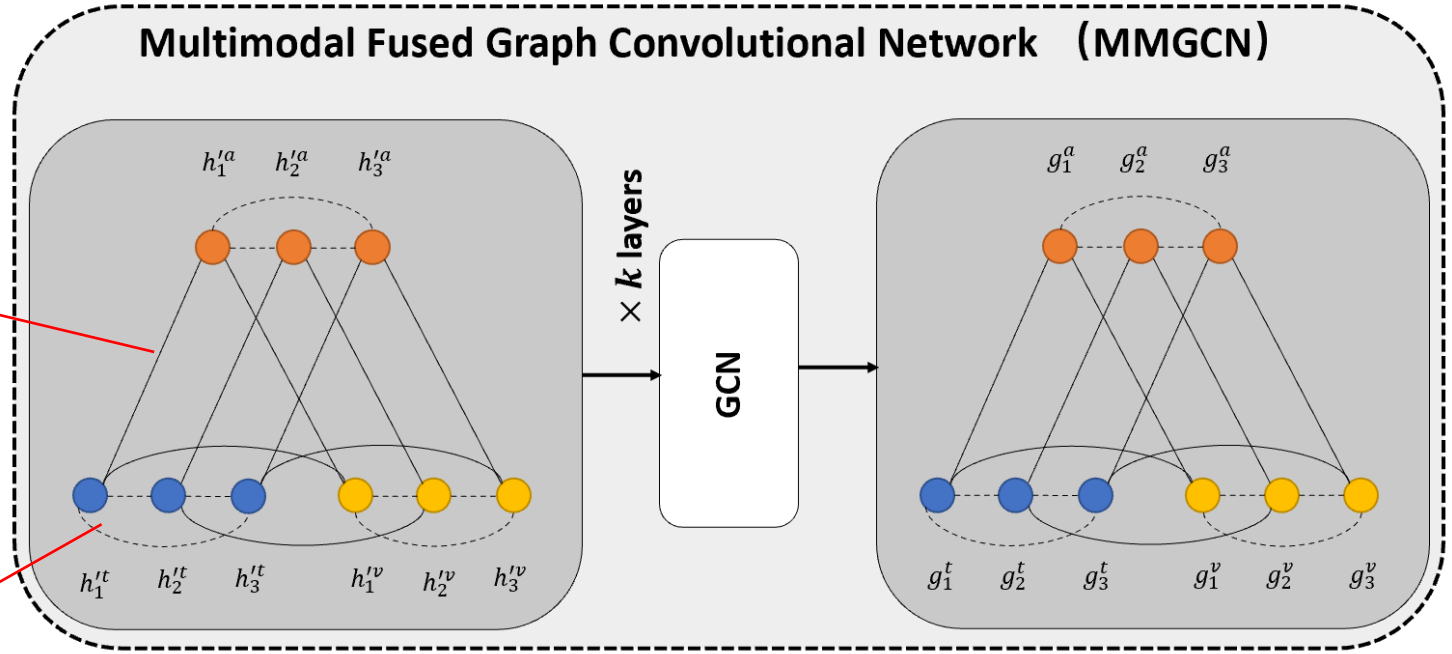
Utterance nodes are initialized with $[h_i^a, S_i]$, $[h_i^v, S_i]$, $[h_i^t, S_i]$.

Edges:

- Any two nodes in the same modality in the same dialogue are connected in the graph; **(Speaker Interaction)**
- Each node is connected with nodes of the same utterance but from different modalities. **(Modality Fusion)**

J Hu, Y Liu, J Zhao, Q Jin. **MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation.** ***ACL, 2021.***

Multimodal Fusion with Graph Convolution Network



Edge Weighting:

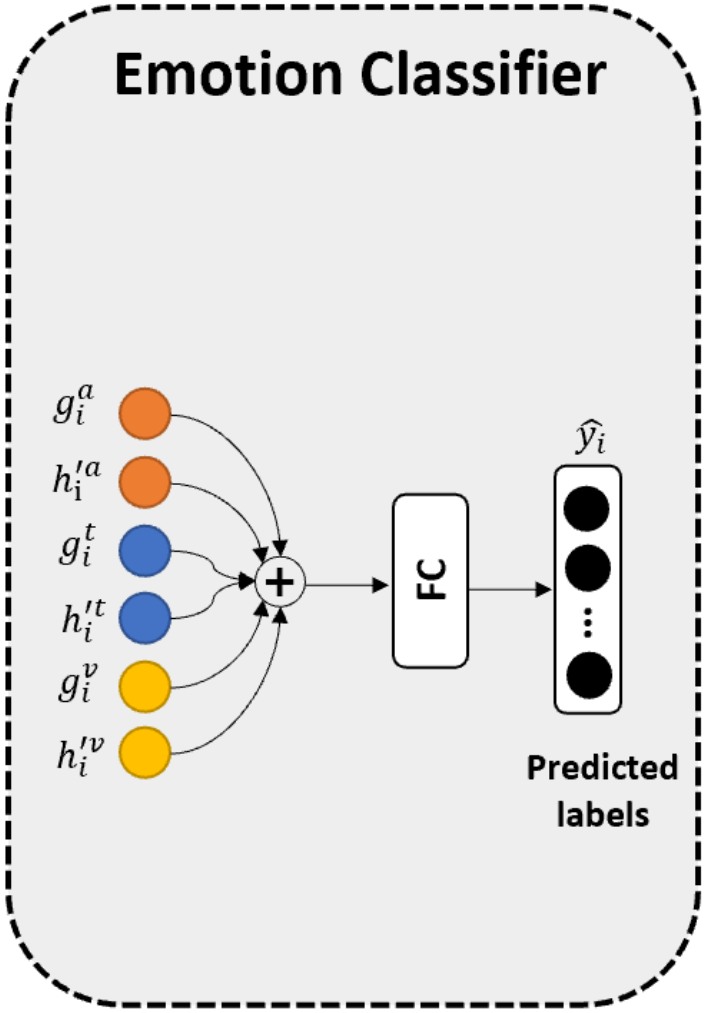
We use the angular similarity to represent the edge weight between two nodes.

$$\mathcal{A}_{ij} = 1 - \frac{arccos(sim(n_i, n_j))}{\pi}$$

$$\mathcal{A}_{ij} = \gamma(1 - \frac{arccos(sim(n_i, n_j))}{\pi})$$

J Hu, Y Liu, J Zhao, Q Jin. **MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation.** ***ACL, 2021.***

Multimodal Fusion with Graph Convolution Network



Emotion Classifier

$$h_i^{'} = [h_i^{'a}, h_i^{'v}, h_i^{'t}]$$

$$g_i = [g_i^a, g_i^v, g_i^t]$$

$$e_i = [h_i^{'}, g_i]$$

We use categorical cross-entropy along with L2-regularization as the loss function during training.

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

J Hu, Y Liu, J Zhao, Q Jin. **MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation.** ***ACL, 2021.***

Experiments

IEMOCAP (2008)

The dataset contains 12 hours of videos of two-way conversations from ten unique speakers, including 7433 utterances and 151 dialogues totally. Each utterance in the dialogue is annotated with six emotion labels.

MELD (2018)

The dataset collects data from TV show Friends, including three modality-aligned conversation data with higher quality, incorporating 13708 utterances, 1433 conversations and 304 different speakers totally.

Dataset	dialogues		utterances	
	train+val	test	train+val	test
IEMOCAP	120	31	5810	1623
MELD	1153	280	11098	2610

J Hu, Y Liu, J Zhao, Q Jin. **MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation.** ***ACL, 2021.***

Experiments

	IEMOCAP							MELD
	Happy	Sad	Neutral	Angry	Excited	Frustrated	**Average(w)**	**Average(w)**
BC-LSTM	34.43	60.87	51.81	56.73	57.95	58.92	54.95	56.80
CMN	30.38	62.41	52.39	59.83	60.25	60.69	56.13	-
ICON	29.91	64.57	57.38	63.04	63.42	60.81	58.54	-
DialogueRNN	39.16	**81.69**	59.77	67.36	72.91	60.27	64.58	57.11
DialogueGCN	**47.1**	80.88	58.71	66.08	70.97	61.21	65.04	58.23
MMGCN	42.34	78.67	**61.73**	**69.00**	**74.33**	**62.32**	**66.22**	**58.65**

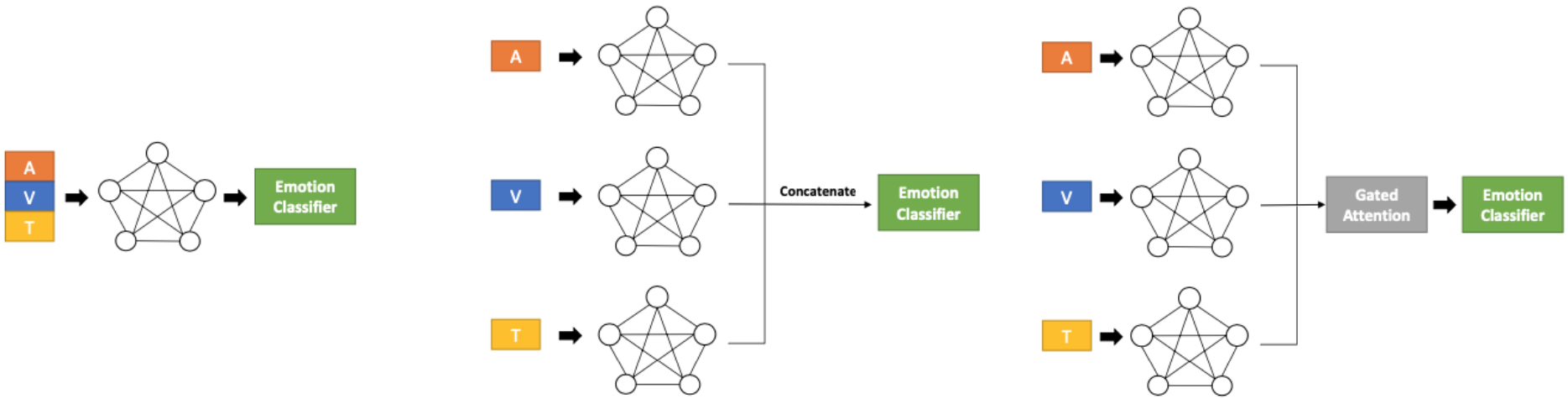
ERC performance (F1-score) of different approaches **under the multimodal setting**, where bold font denotes the best performance. Average(w) means weighted average.

Comparison with other models

Our proposed MMGCN improves the F1-score performance over DialogueGCN under the multimodal setting by absolute **1.18% on IEMOCAP** and **0.42% on MELD** on average.

J Hu, Y Liu, J Zhao, Q Jin. **MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation.** ***ACL, 2021.***

Experiments



(a) early fusion (b) late fusion (c) fusion through gated attention

	IEMOCAP	MELD
$DeepGCN_{early\_fusion}$	64.46	57.94
$DeepGCN_{late\_fusion}$	64.62	58.26
$DeepGCN_{gated\_attention}$	64.45	58.18
$DeepGCN_{MFN}$	62.77	58.21
$DeepGCN_{MulT}$	62.37	57.93
$MMGCN$	**66.22**	**58.65**

Comparison with other fusion methods

MMGCN with the graph-based multimodal fusion outperforms all other compared methods, since MMGCN could let nodes aggregate from both context and different modalities.

J Hu, Y Liu, J Zhao, Q Jin. **MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation.** ***ACL, 2021.***

Experiments

modality	IEMOCAP	MELD
a	54.66	42.63
v	33.86	33.27
t	62.35	57.72
at	65.70	58.02
vt	62.89	57.92
avt	**66.22**	**58.65**

MMGCN under various modality setting

Adding acoustic and visual modalities can bring additional performance improvement over the textual modality.

layers	IEMOCAP	MELD
1	66.12	58.40
2	66.17	58.38
4	**66.22**	**58.65**
8	66.1	58.54
16	66.06	58.38
32	66.1	58.42

MMGCN with different layers

A different number of layers does affect the ERC recognition performance. It proves the efficiency of deep GCN.

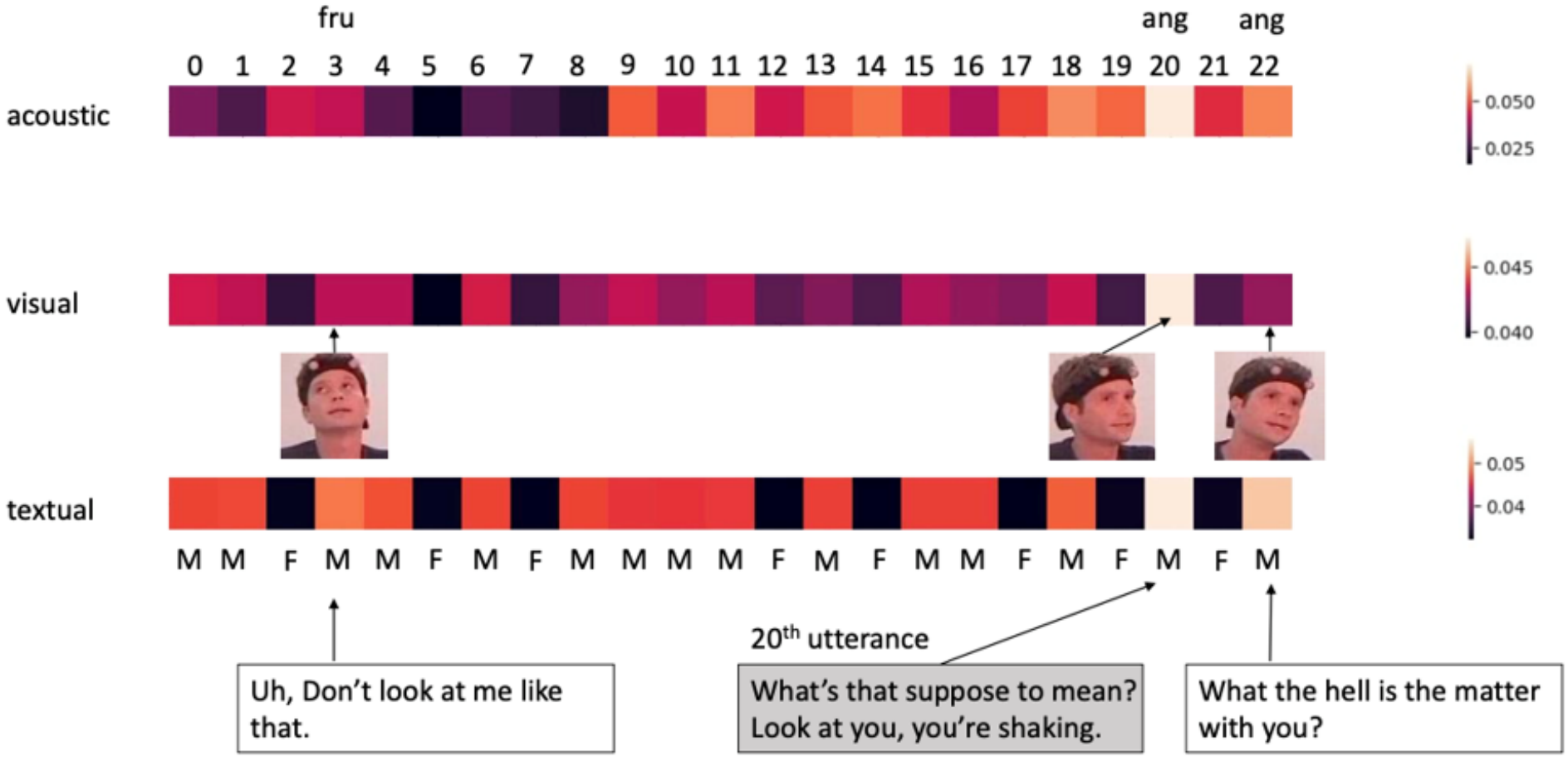
MMGCN	IEMOCAP	MELD
w/ spkr embedding	**66.22**	**58.65**
w/o spkr embedding	65.76	58.38

Impact of Speaker Embedding

The result proves that speaker information can help improve emotion recognition performance.

J Hu, Y Liu, J Zhao, Q Jin. **MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation.** ***ACL, 2021.***

Experiments



visualization of the heatmap of the adjacent matrix for the 20th utterance in a conversation with three modalities. 'M' and 'F' refer to the male and female speakers respectively.

J Hu, Y Liu, J Zhao, Q Jin. **MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation.** ***ACL, 2021.***

Summary

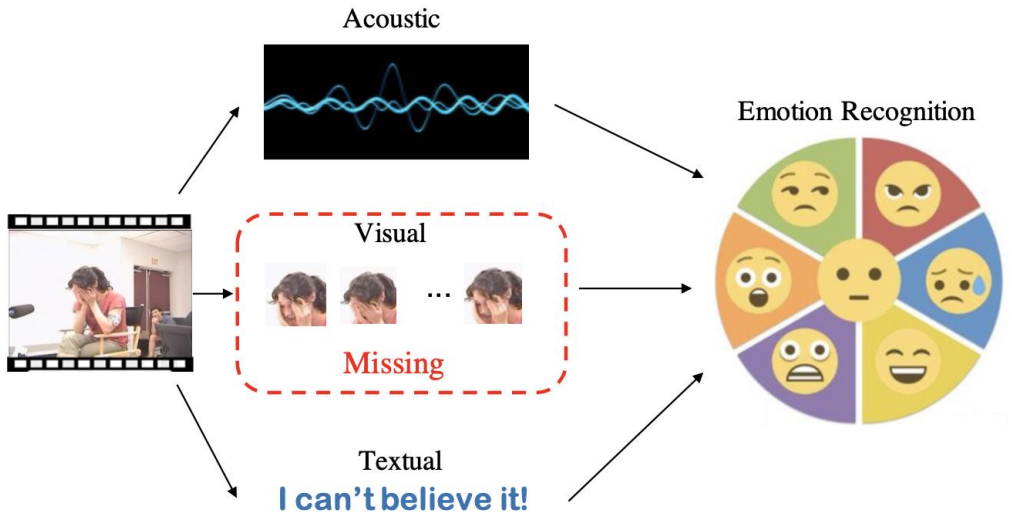
Multimodal Graph Convolution Network (MMGCN)

- Provide a more effective way of utilizing both multimodal and long-distance contextual information.
- Construct a graph that captures not only intra-speaker context dependency but also inter-modality dependency.
- Deepen GCN layers to improve recognition performance.

Missing-modality Problem

Scenarios with missing-modality problem

- Camera is turned off or blocked due to privacy issues; Missing Face Info
- lighting or occlusion issues; Missing Face Info
- ASR errors; Missing text Info
- Silent and using body language as response; Missing audio/text Info



the person's face was obscured by her hands
Missing Face Info

Pitfall: Multimodal emotion recognition models trained on the full-modality data are sensitive to missing-modality scenarios

Missing-modality Problem

- Data Augmentation
  construct missing modalities training samples

- Generative methods
  generate the missing modalities

- Learning multimodal Joint Representation

Missing-modality Problem

- Data Augmentation
  construct missing modalities training samples

- Generative methods
  generate the missing modalities

- Learning multimodal Joint Representation

Our Goal:
A **unified** model deal with both different **uncertain missing-modality** conditions and the **full-modality** condition.

J Zhao, R Li, Q Jin. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. *ACL, 2021.*

Missing Modality Imagination Network

Training Pipeline Overview:

1. Extract Features
2. Extract multimodal representations
3. Forward Imagination
4. Backward Imagination

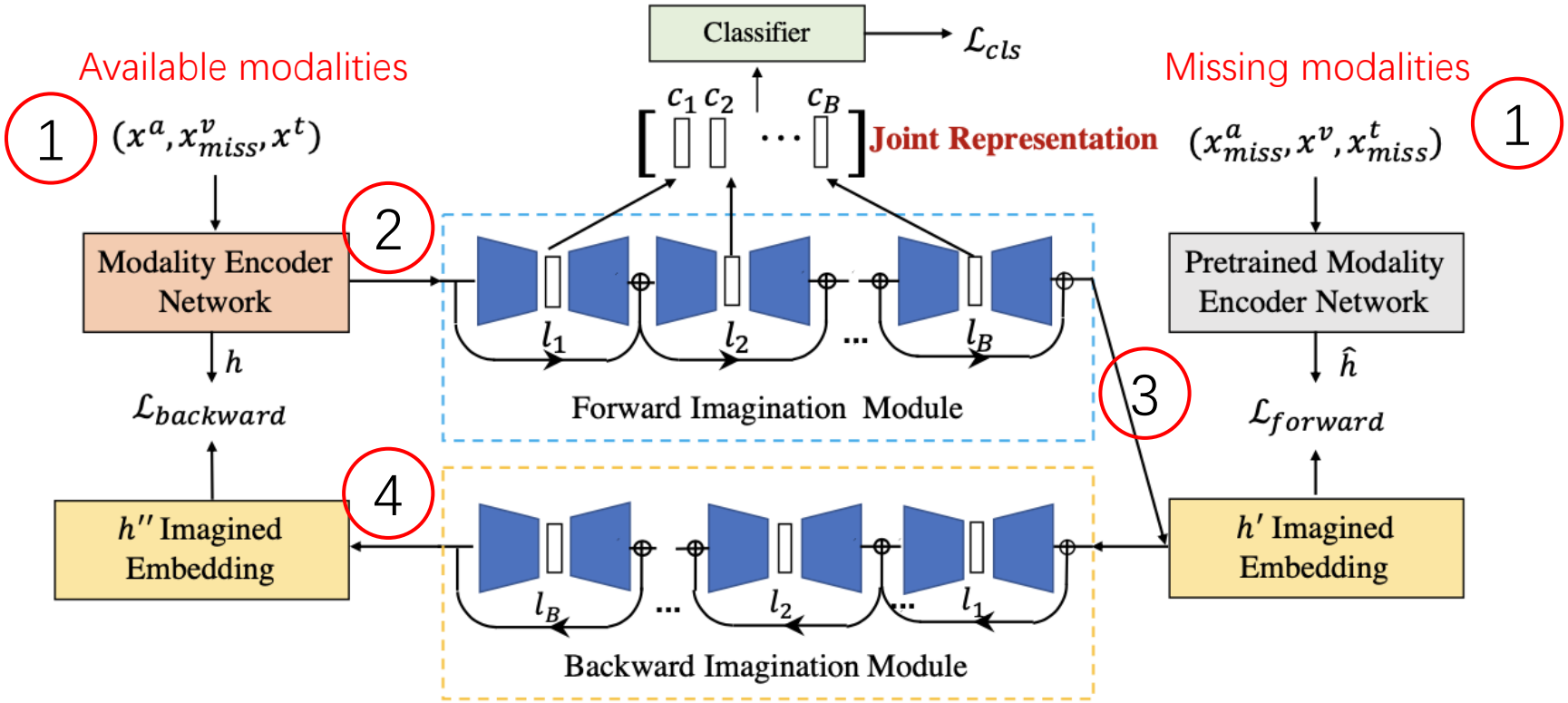
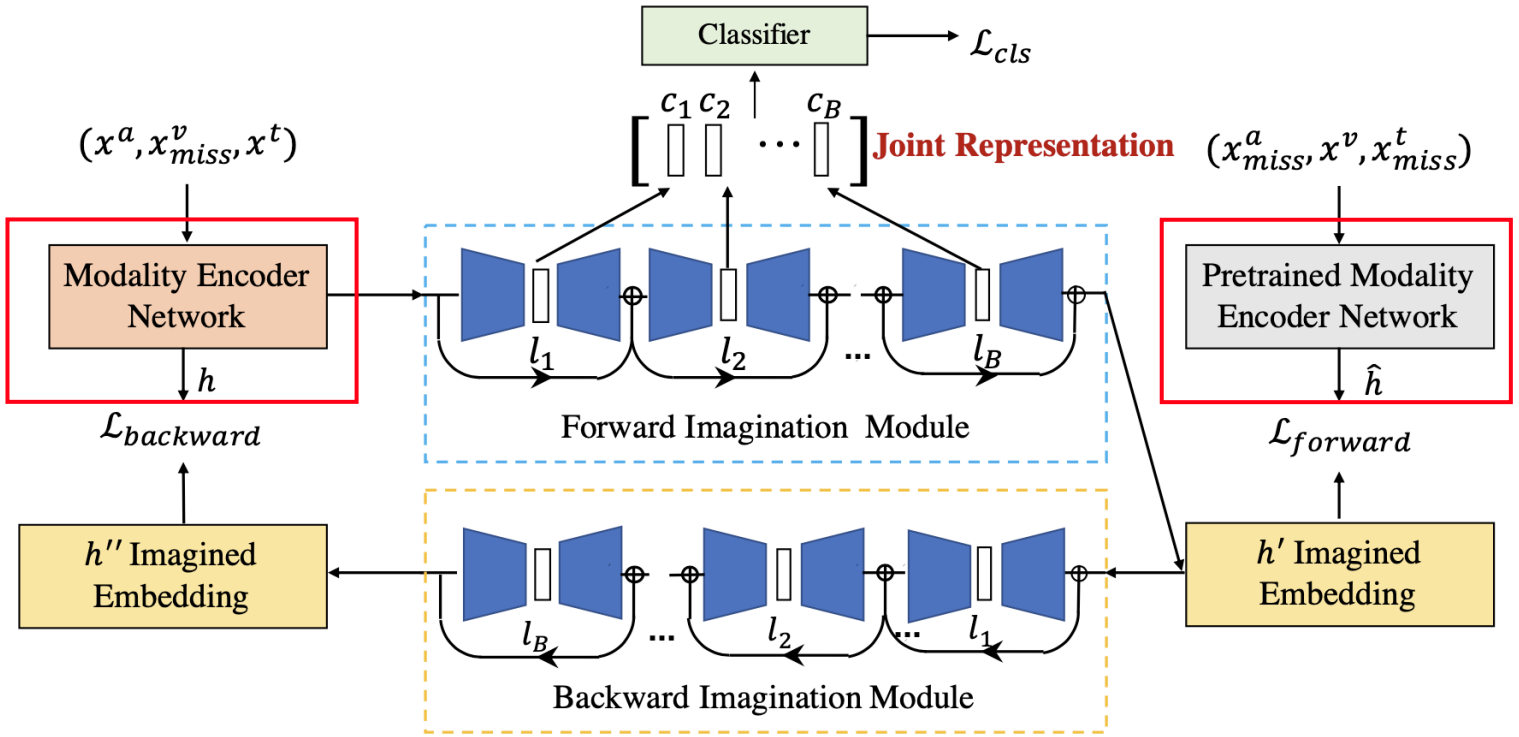


Figure. 2 Illustration of the Missing Modality Imagination Network (MMIN) framework.

J Zhao, R Li, Q Jin. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. *ACL, 2021.*

Missing Modality Imagination Network

Modality Encoder Network

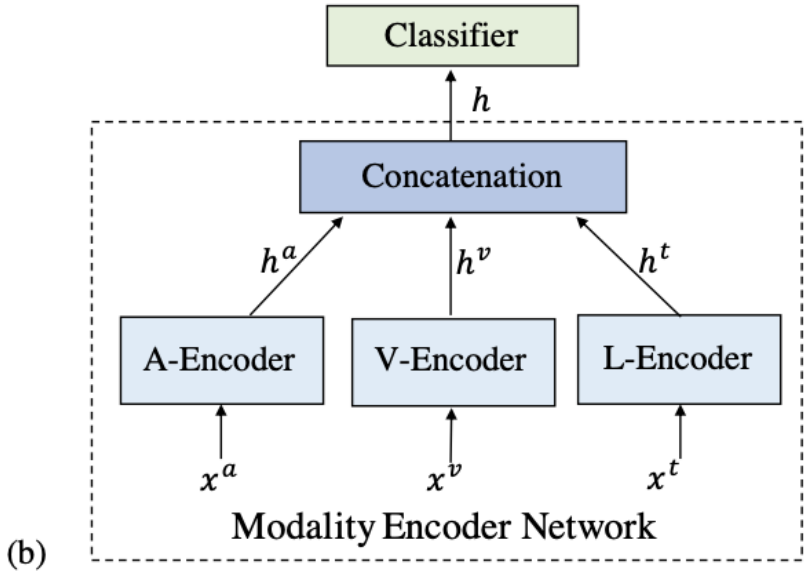


- **Orange Block** is initialized by pretrained-encoder and update during training.
- **Gray Block** is fixed during training and extract modality embeddings as ground-truth for forward imagination.

J Zhao, R Li, Q Jin. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. *ACL, 2021.*

Missing Modality Imagination Network

Modality Encoder Network



Extract the **utterance-level modality-specific embeddings**.

$$h^a = EncA(x^a), \quad h^v = EncV(x^v), \; h^t = EncL(x^t)$$

$$h = concat(h^a, h^v, h^t)$$

J Zhao, R Li, Q Jin. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. *ACL, 2021.*

Missing Modality Imagination Network

Missing Modality Condition Creation

- For 1 full-modality sample, constructing 6 different missing-modality samples.
- Transform the input to the **unified format**.
- Cross-modality **unified triplet format pairs** are used to train the MMIN.

	(available,missing)	unified triplet format pairs
1	$((x^a),(x^v,x^t))$	$((x^a,x^v_{miss},x^t_{miss}),(x^a_{miss},x^v,x^t))$
2	$((x^v),(x^a,x^t))$	$((x^a_{miss},x^v,x^t_{miss}),(x^a,x^v_{miss},x^t))$
3	$((x^t),(x^a,x^v))$	$((x^a_{miss},x^v_{miss},x^t),(x^a,x^v,x^t_{miss}))$
4	$((x^a,x^v),(x^t))$	$((x^a,x^v,x^t_{miss}),(x^a_{miss},x^v_{miss},x^t))$
5	$((x^a,x^t),(x^v))$	$((x^a,x^v_{miss},x^t),(x^a_{miss},x^v,x^t_{miss}))$
6	$((x^v,x^t),(x^a))$	$(x^a_{miss},x^v,x^t),(x^a,x^v_{miss},x^t_{miss}))$

J Zhao, R Li, Q Jin. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. *ACL, 2021.*

Missing Modality Imagination Network

Imagination Module

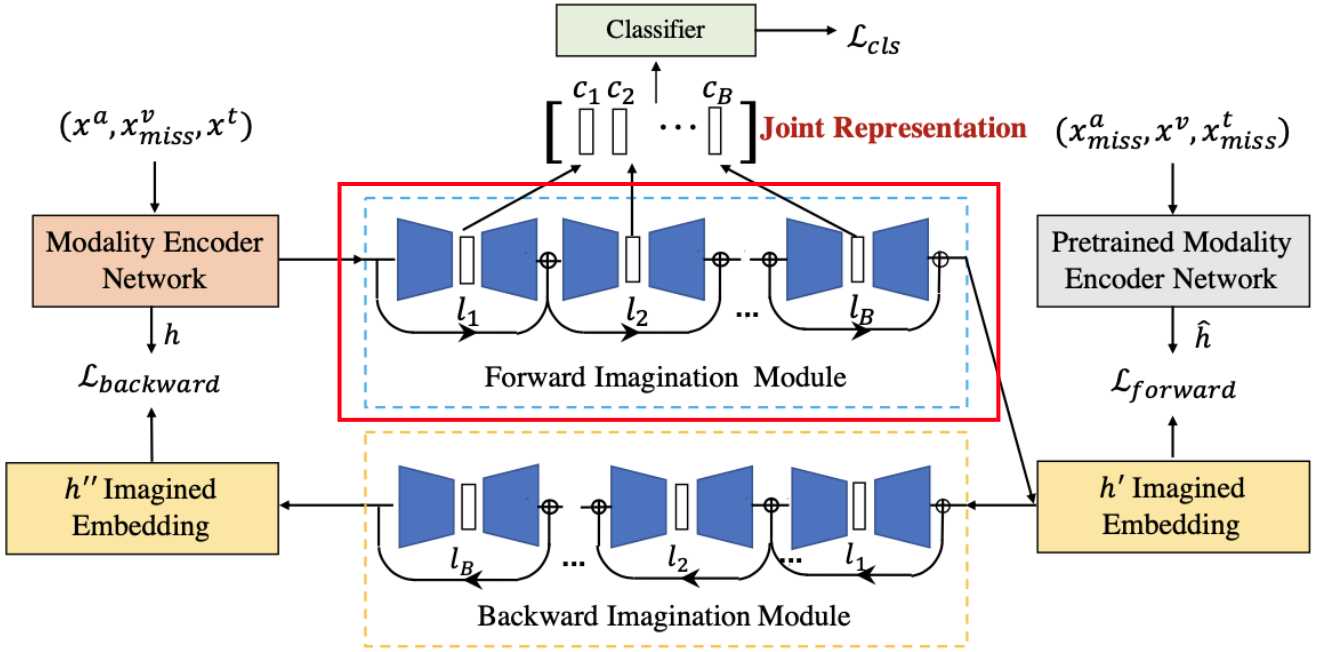
- Learning multimodal joint representations through the cross-modality imagination.

- CRA (Cascade Residual Autoencoder)

$$\begin{cases} \Delta z_k = \phi_k(h), & k = 1 \\ \Delta z_k = \phi_k(h + \sum_{j=1}^{k-1} \Delta z_j), & k > 1 \end{cases}$$

$$h^{'} = imagine_{forward}(h) = h + \sum_{k=1}^{B} \Delta z_k$$

Residual operation can retain more multimodal information.



- Cycle Consistency Learning

J Zhao, R Li, Q Jin. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. *ACL, 2021.*

Missing Modality Imagination Network

Joint Optimization

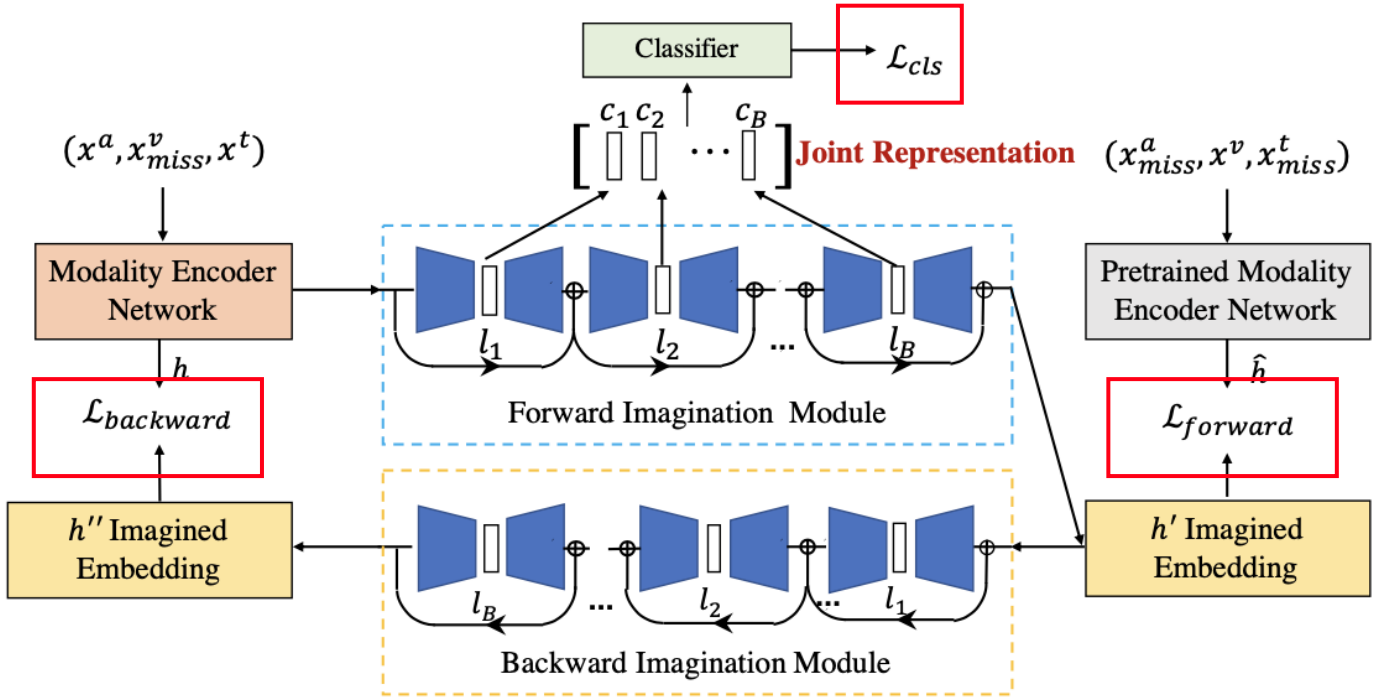
$$\mathcal{L}_{cls} = -\frac{1}{|S|}\sum_{i=1}^{|S|} H(p,q)$$

$$\mathcal{L}_{forward} = \frac{1}{|S|}\sum_{i=1}^{|S|} \left\| \hat{h}_i - h_i^{'} \right\|_2^2$$

$$\mathcal{L}_{backward} = \frac{1}{|S|}\sum_{i=1}^{|S|} \left\| h_i - h_i^{''} \right\|_2^2$$

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{forward} + \lambda_2 \mathcal{L}_{backward}$$

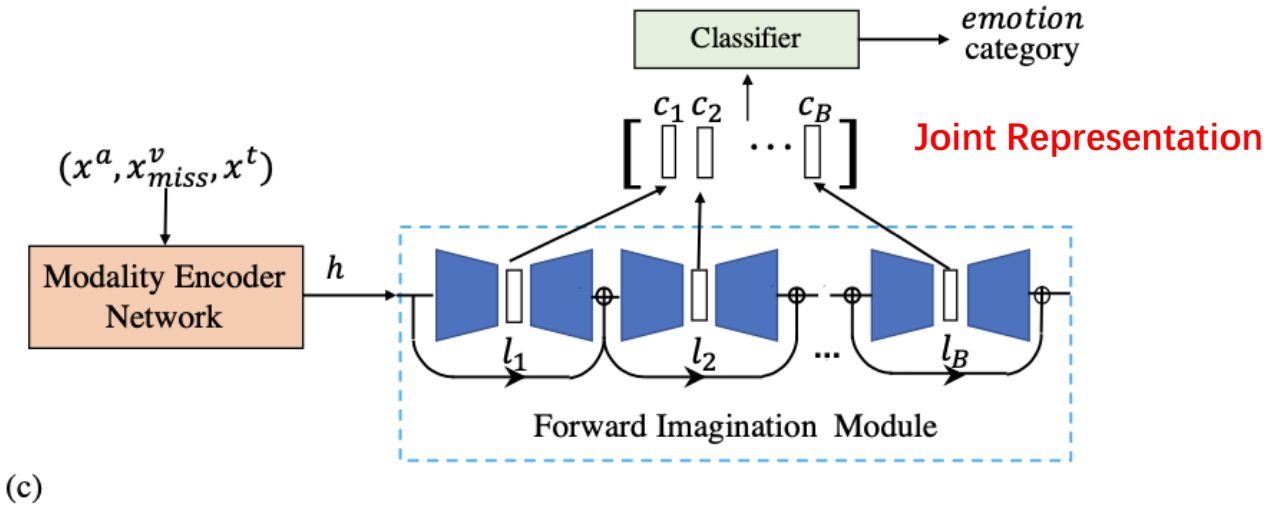
$$(\lambda_1 = 0.1,\ \lambda_2 = 0.1\ )$$



J Zhao, R Li, Q Jin. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. *ACL, 2021.*

Missing Modality Imagination Network

Inference Pipeline



Inference Framework

MMIN can infer under different missing modality conditions.

J Zhao, R Li, Q Jin. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. *ACL, 2021.*

Experiments

Datasets: IEMOCAP and MSP-IMPROV

dataset	Happy	Anger	Sadness	Neutral	Total
IECMOAP	1636	1103	1084	1708	5531
MSP-IMPROV	999	460	627	1733	3819

Table 2: Data Statistics of datasets

- **Full-Modality Training Set** and **Full-Modality Testing Set**

  Original datasets with audio, textual and visual modalities

- **Missing-Modality Training Set**

  Construct 6 different missing-modality samples for each full-modality sample.

- **Missing-Modality Testing Set**

  Similarly, construct 6 different missing-modality testing subsets

J Zhao, R Li, Q Jin. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. *ACL, 2021.*

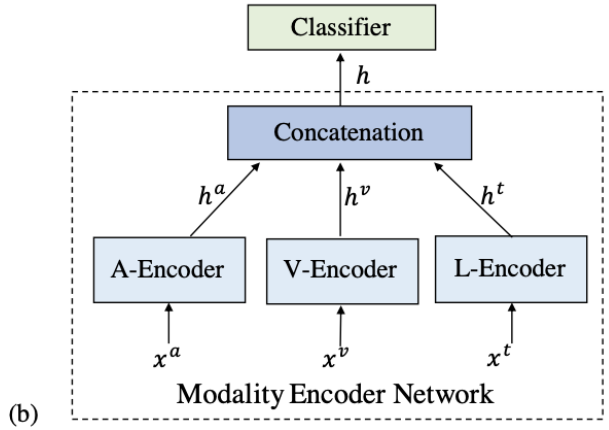
Experiments

Modality Features and Modality Encoders

	Features	Encoder
Acoustic	ComParE frame-level features extracted from OpenSmile toolkit.	1-layer LSTM with 128 hidden units
Visual	Facial expression frame-level features extracted from a pretrained DenseNet model	1-layer LSTM with 128 hidden units
Textual	word embeddings extracted from the pretrained BERT-large model.	Standard TextCNN with (3,4,5) kernel sizes.

	train	test	WA	UA
Our full-modality baseline	$\{a, v, t\}$	$\{a, v, t\}$	**0.7651**	**0.7779**
cLSTM-MMA(Pan et al., 2020)			0.7394	–
SSMM(Liang et al., 2020)			0.7560	0.7450

Table 4: Multimodal Emotion Recognition Results on IEMOCAP under full-modality condition.



(b) Baseline Model

J Zhao, R Li, Q Jin. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. *ACL, 2021.*

Experiments

Uncertain Missing-Modality Results

Dateset	Model	Metric	Testing Condition							
			$\{a\}$	$\{v\}$	$\{t\}$	$\{a,v\}$	$\{a,t\}$	$\{v,t\}$	Average	$\{a,v,t\}$
IEMOCAP	Full-modality baseline	WA(↑)	0.4190	0.4574	0.5646	0.5488	0.7018	0.6217	0.5522	0.7651
		UA(↑)	0.4719	0.3966	0.5549	0.5762	0.7257	0.5971	0.5537	0.7779
	Augmented baseline	WA(↑)	0.5303	0.4864	0.6564	0.6395	0.7251	0.7082	0.6243∗	0.7617
		UA(↑)	0.5440	0.4598	0.6691	0.6434	0.7435	0.7162	0.6293∗	0.7767
	proposed MMIN	WA(↑)	0.5658	0.5252	0.6657	0.6399	0.7294	0.7267	**0.6410∗▲**	0.7650
		UA(↑)	**0.5900**	**0.5160**	**0.6802**	**0.6543**	**0.7514**	**0.7361**	**0.6524∗▲**	**0.7812∗▲**
	MCTN (Pham et al., 2019)	WA(↑)	0.4975	0.4892	0.6242	0.5634	0.6834	0.6784	0.5894∗	–
		UA(↑)	0.5162	0.4573	0.6378	0.5584	0.6946	0.6834	0.5913∗	–

$\{a\}$: Only acoustic modality is available.
Average: Average performance over all 6 missing-modality testing subsets.
Full-modality baseline: Full-modality training set.
Augmented baseline: Full-modality and missing-modality training sets together.
proposed MMIN: MMIN model and missing-modality training set.
MCTN: 6 different MCTN models for different conditions.

J Zhao, R Li, Q Jin. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. *ACL, 2021.*

Experiments

Augmented baseline vs Full-Modality baseline

Dateset	Model	Metric	Testing Condition							
			$\{a\}$	$\{v\}$	$\{t\}$	$\{a,v\}$	$\{a,t\}$	$\{v,t\}$	Average	$\{a,v,t\}$
IEMOCAP	Full-modality baseline	WA(↑)	0.4190	0.4574	0.5646	0.5488	0.7018	0.6217	0.5522	0.7651
		UA(↑)	0.4719	0.3966	0.5549	0.5762	0.7257	0.5971	0.5537	0.7779
	Augmented baseline	WA(↑)	0.5303	0.4864	0.6564	0.6395	0.7251	0.7082	0.6243∗	0.7617
		UA(↑)	0.5440	0.4598	0.6691	0.6434	0.7435	0.7162	0.6293∗	0.7767
	proposed MMIN	WA(↑)	0.5658	0.5252	0.6657	0.6399	0.7294	0.7267	**0.6410∗▲**	0.7650
		UA(↑)	**0.5900**	**0.5160**	**0.6802**	**0.6543**	**0.7514**	**0.7361**	**0.6524∗▲**	**0.7812∗▲**
	MCTN (Pham et al., 2019)	WA(↑)	0.4975	0.4892	0.6242	0.5634	0.6834	0.6784	0.5894∗	–
		UA(↑)	0.5162	0.4573	0.6378	0.5584	0.6946	0.6834	0.5913∗	–

- Full-modality baseline performs very poorly under missing-modality conditions.
- Data augmentation can **significantly improve** over all missing-modality conditions, helps **data mismatch problem**.

J Zhao, R Li, Q Jin. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. *ACL, 2021.*

Experiments

Our MMIN vs Augmented baseline

Dateset	Model	Metric	Testing Condition							
			$\{a\}$	$\{v\}$	$\{t\}$	$\{a,v\}$	$\{a,t\}$	$\{v,t\}$	Average	$\{a,v,t\}$
IEMOCAP	Full-modality baseline	WA(↑)	0.4190	0.4574	0.5646	0.5488	0.7018	0.6217	0.5522	0.7651
		UA(↑)	0.4719	0.3966	0.5549	0.5762	0.7257	0.5971	0.5537	0.7779
	Augmented baseline	WA(↑)	0.5303	0.4864	0.6564	0.6395	0.7251	0.7082	0.6243∗	0.7617
		UA(↑)	0.5440	0.4598	0.6691	0.6434	0.7435	0.7162	0.6293∗	0.7767
	proposed MMIN	WA(↑)	0.5658	0.5252	0.6657	0.6399	0.7294	0.7267	**0.6410∗▲**	0.7650
		UA(↑)	**0.5900**	**0.5160**	**0.6802**	**0.6543**	**0.7514**	**0.7361**	**0.6524∗▲**	**0.7812∗▲**
	MCTN (Pham et al., 2019)	WA(↑)	0.4975	0.4892	0.6242	0.5634	0.6834	0.6784	0.5894∗	–
		UA(↑)	0.5162	0.4573	0.6378	0.5584	0.6946	0.6834	0.5913∗	–

- MMIN performs better under all missing-modality conditions.
- MMIN performs better under full-modality conditions, which trained on the missing-modality training set.
- MMIN learns the joint multimodal representations.

J Zhao, R Li, Q Jin. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. *ACL, 2021.*

Experiments

Our MMIN vs MCTN

Dateset	Model	Metric	Testing Condition							
			$\{a\}$	$\{v\}$	$\{t\}$	$\{a,v\}$	$\{a,t\}$	$\{v,t\}$	Average	$\{a,v,t\}$
IEMOCAP	Full-modality baseline	WA(↑)	0.4190	0.4574	0.5646	0.5488	0.7018	0.6217	0.5522	0.7651
		UA(↑)	0.4719	0.3966	0.5549	0.5762	0.7257	0.5971	0.5537	0.7779
	Augmented baseline	WA(↑)	0.5303	0.4864	0.6564	0.6395	0.7251	0.7082	0.6243∗	0.7617
		UA(↑)	0.5440	0.4598	0.6691	0.6434	0.7435	0.7162	0.6293∗	0.7767
	proposed MMIN	WA(↑)	0.5658	0.5252	0.6657	0.6399	0.7294	0.7267	**0.6410∗▲**	0.7650
		UA(↑)	**0.5900**	**0.5160**	**0.6802**	**0.6543**	**0.7514**	**0.7361**	**0.6524∗▲**	**0.7812∗▲**
	MCTN (Pham et al., 2019)	WA(↑)	0.4975	0.4892	0.6242	0.5634	0.6834	0.6784	0.5894∗	–
		UA(↑)	0.5162	0.4573	0.6378	0.5584	0.6946	0.6834	0.5913∗	–

- MCTN performs better than full-modality baseline under missing-modality conditions.
- MCTN performs worse than our MMIN.

J Zhao, R Li, Q Jin. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. *ACL, 2021.*

Experiments

Uncertain Missing-Modality Results on MSP

MSP-IMPROV	Full-modality baseline	F1(↑)	0.2824	0.3295	0.4576	0.4721	0.5655	0.5368	0.4543	0.6523
	Augmented baseline	F1(↑)	0.4278	0.4185	0.5544	0.5396	0.6038	0.6295	**0.5455**∗	**0.6663**∗
	proposed MMIN	F1(↑)	**0.4647**	**0.4471**	**0.5573**	**0.5740**	**0.6188**	**0.6411**	**0.5649**∗▲	**0.6855**∗▲
	MCTN (Pham et al., 2019)	F1(↑)	0.3285	0.3810	0.5050	0.4683	0.5611	0.5886	0.4721∗	–

Similar trends on MSP-IMPROV, which demonstrates the good **generalization ability** of MMIN across different datasets.

J Zhao, R Li, Q Jin. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. *ACL, 2021.*

Experiments

Ablation Study

Model	Metric	Testing Condition							
		$\{a\}$	$\{v\}$	$\{t\}$	$\{a,v\}$	$\{a,t\}$	$\{v,t\}$	Average	$\{a,v,t\}$
MMIN-AE	WA(↑)	0.5404	0.5025	0.6588	0.6115	0.7203	0.7125	0.6244	0.7619
	UA(↑)	0.5625	0.4836	0.6689	0.6246	0.7374	0.7187	0.6368	0.7677
MMIN-NoCycle	WA(↑)	0.5503	0.5116	0.6577	0.6239	0.7185	0.7202	0.6304	0.7498
	UA(↑)	0.5821	0.5006	0.6705	0.6454	0.7438	0.7301	0.6454	0.7709
MMIN	WA(↑)	0.5658	0.5252	0.6657	0.6399	0.7294	0.7267	**0.6410**	**0.7650**
	UA(↑)	**0.5900**	**0.5160**	**0.6802**	**0.6543**	**0.7514**	**0.7361**	**0.6524**	**0.7812**

- Different Imagination Module, Default CRA vs AutoEncoder

  Standard auto-encoder performs worse than CRA.

- W/O Cycle Consistence Learning, Default Cycle vs NoCycle

  Only forward-imagination performs worse than cycle consistence learning.

J Zhao, R Li, Q Jin. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. *ACL, 2021.*

Experiments

Joint representation learning ability

	train	test	Baseline	Augmented	MMIN
ComparE	a	a	0.5760	0.5440	**0.5900**
Denseface	v	v	0.5064	0.4598	**0.5160**
Bert	t	t	0.6873	0.6691	0.6802
ComparE+Denseface	a, v	a, v	0.6380	0.6434	**0.6543**
ComparE+Bert	a, t	a, t	0.7533	0.7435	0.7514
Bert+Denseface	v, t	v, t	0.7177	0.7162	**0.7361**
ComparE+Bert+Denseface	a, v, t	a, v, t	0.7779	0.7767	**0.7812**

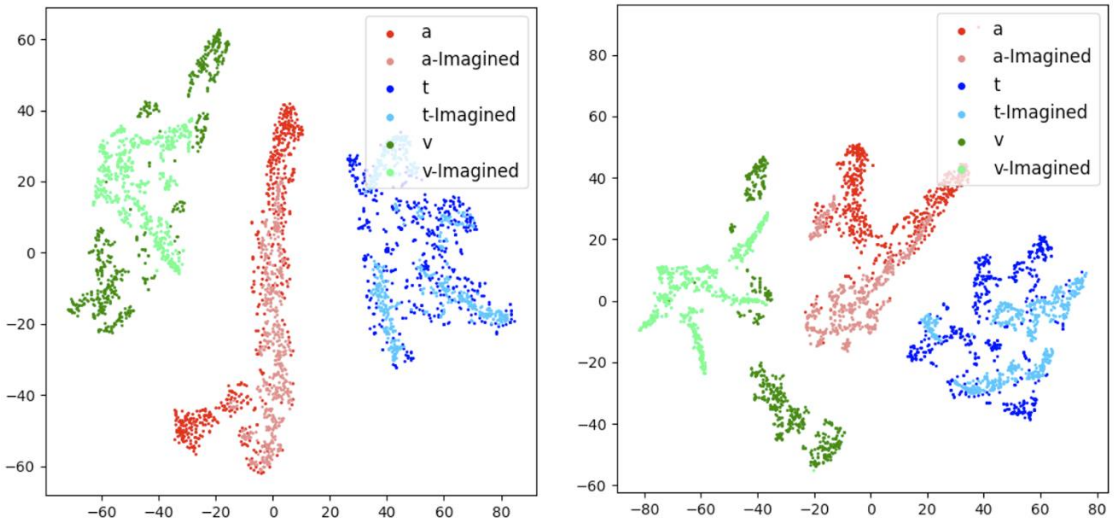
Baseline denotes the results individually trained on partial modalities samples. (**7 Different Models**)

- MMIN can beat almost the partial-modalities baseline models
- MMIN can learn the joint multimodal representation.

J Zhao, R Li, Q Jin. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. *ACL, 2021.*

Experiments

Imagination Ability



(a) A Female Speaker (b) A Male Speaker

Figure 3. Visualization (T-SNE) the distributions of the ground-truth and imagined multimodal embeddings.

The distributions are very similar, which demonstrates the imagination ability of MMIN.

J Zhao, R Li, Q Jin. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. *ACL, 2021.*

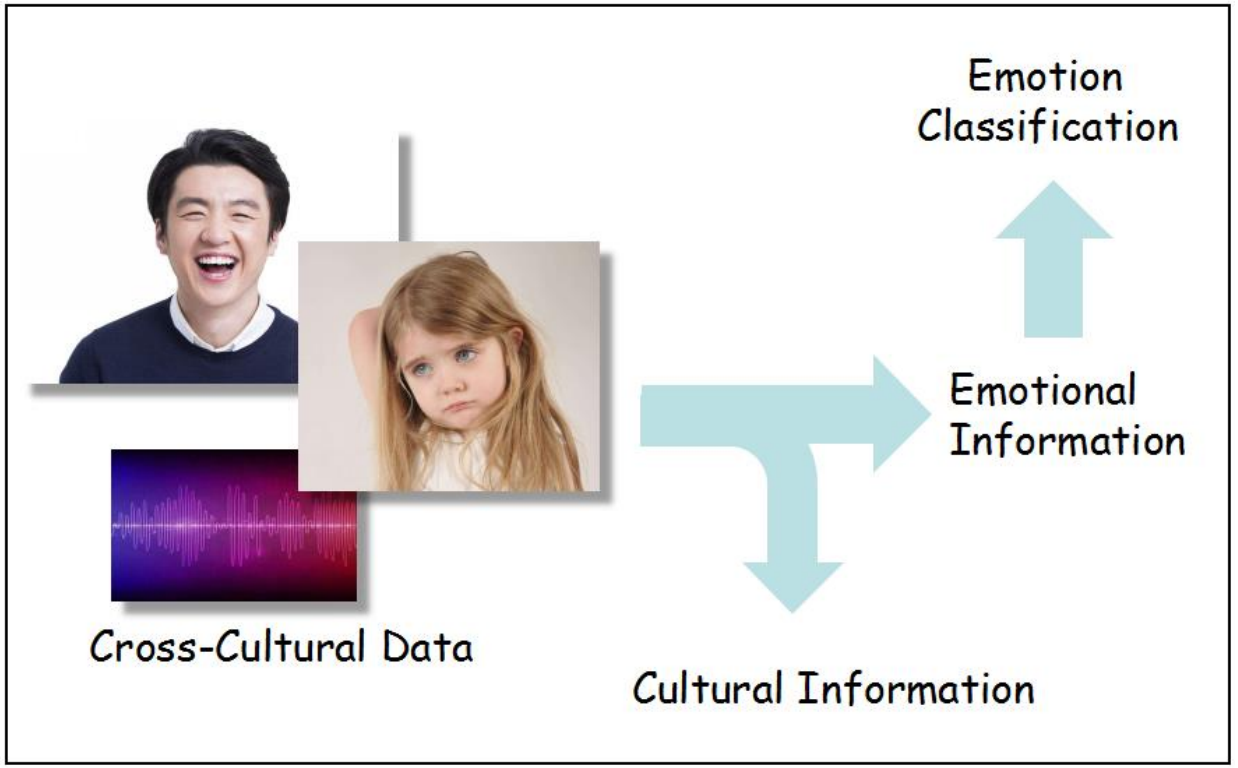
Summary

Missing Modality Imagination Network (MMIN)

- A **unified model**, Missing Modality Imagination Network (MMIN), to improve the robustness of emotion recognition systems under **uncertain missing-modality scenarios**.
- **Cross-modality imagination** based on paired cross-modality data and adopt CRA and Cycle Consistency Learning to learn the multimodal joint representations.
- MMIN can improve the emotion recognition performance under **both the uncertain missing-modality and the full-modality conditions.**

Multimodality Cross-Culture Emotion Recognition

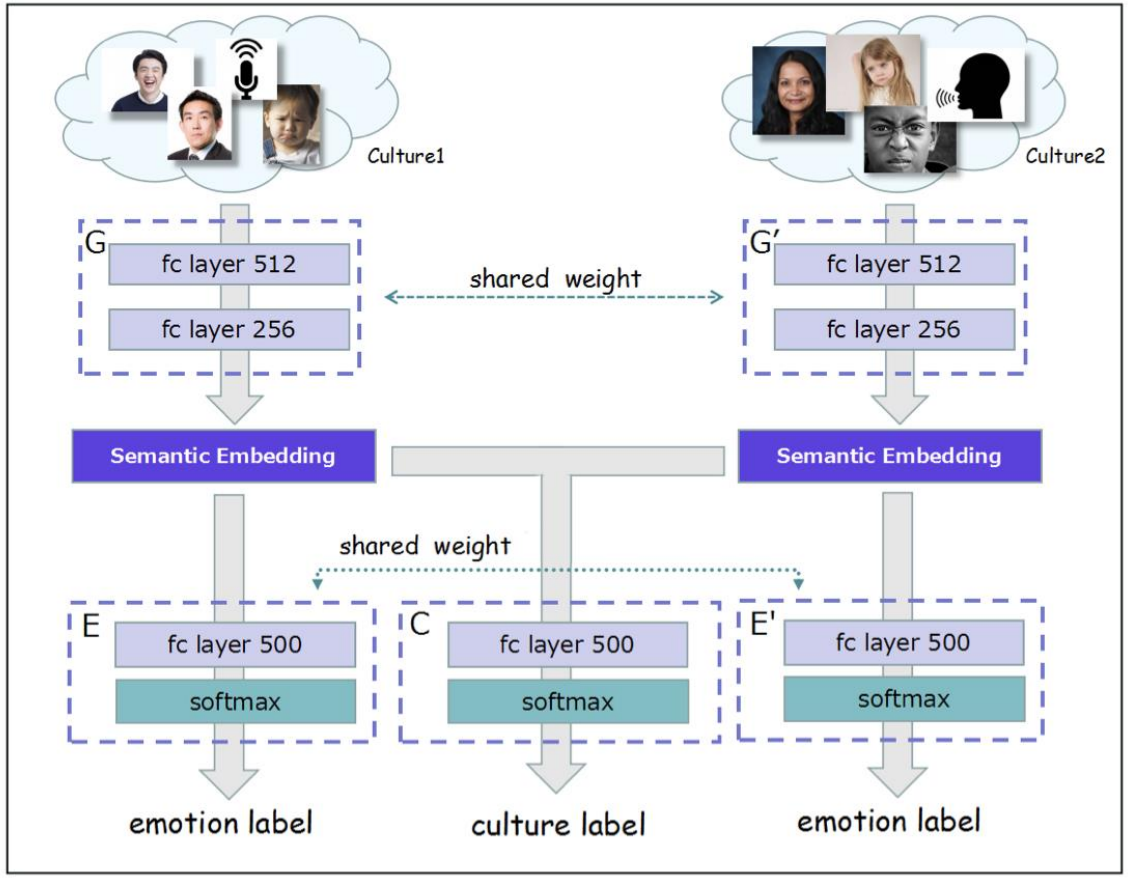
- Cross-culture discrepancy
- Expensive to collect data
  - lacking large-scale multi-cultural labeled dataset



Multimodality Cross-Culture Emotion Recognition

Method:

- Adversarial learning
- Cultural recognition:
  Discriminate different culture
- Emotion recognition:
  Learning emotion salient information



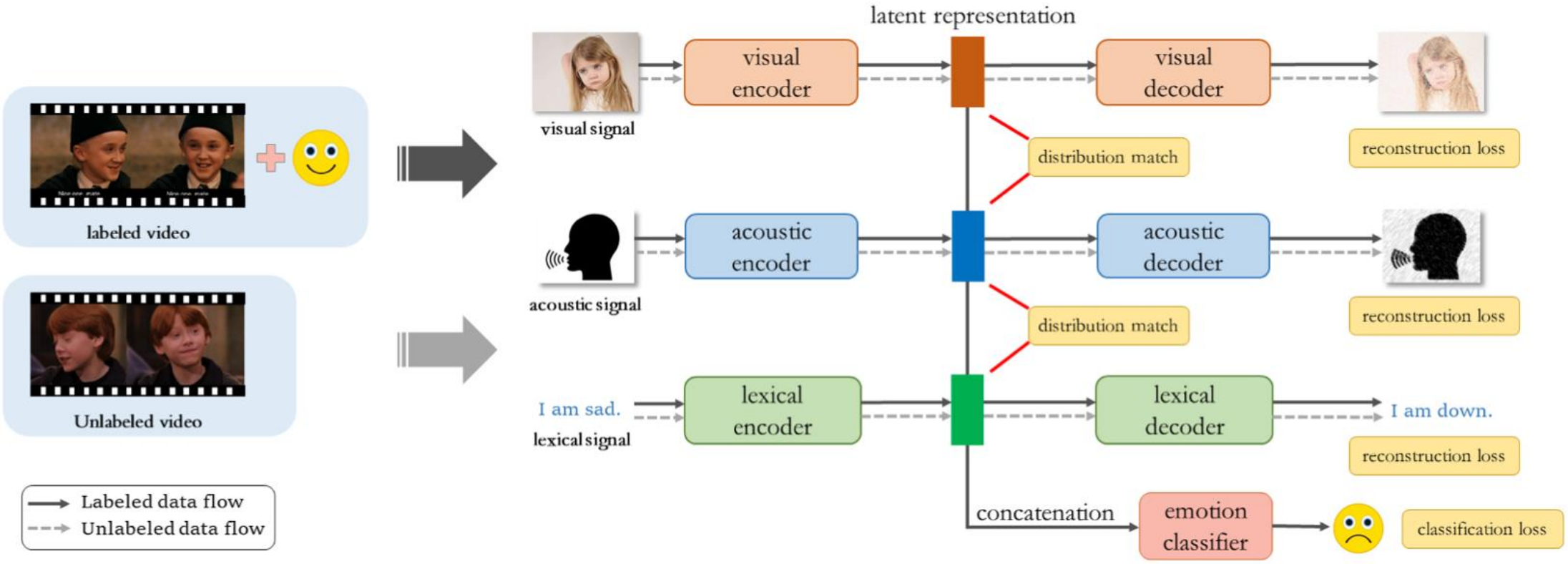
J Liang, S Chen, J Zhao, Q Jin et al. Cross-culture Multimodal Emotion Recognition with Adversarial Learning. ICASSP 2019
J Zhao, R Li, J Liang, Q Jin. Adversarial Domain Adaption for Multi-Cultural Dimensional Emotion Recognition in Dyadic Interactions. AVEC 2019

Data Scarcity Problem

- Hard to collect and hard to annotate
- Current multimodal recognition corpus are small

Dataset	Total duration	Language
IEMOCAP	12h	English
CHEAVD 2.0	7.9h	Chinese
AFEW	< 2.1h	English
Recola	3.8h	French

Emotion Recognition with limited supervised data



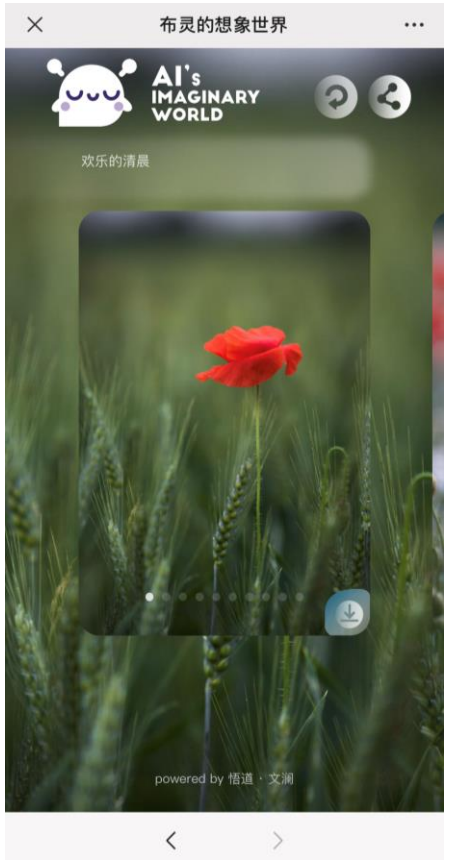
Semi-supervised multimodal emotion recognition

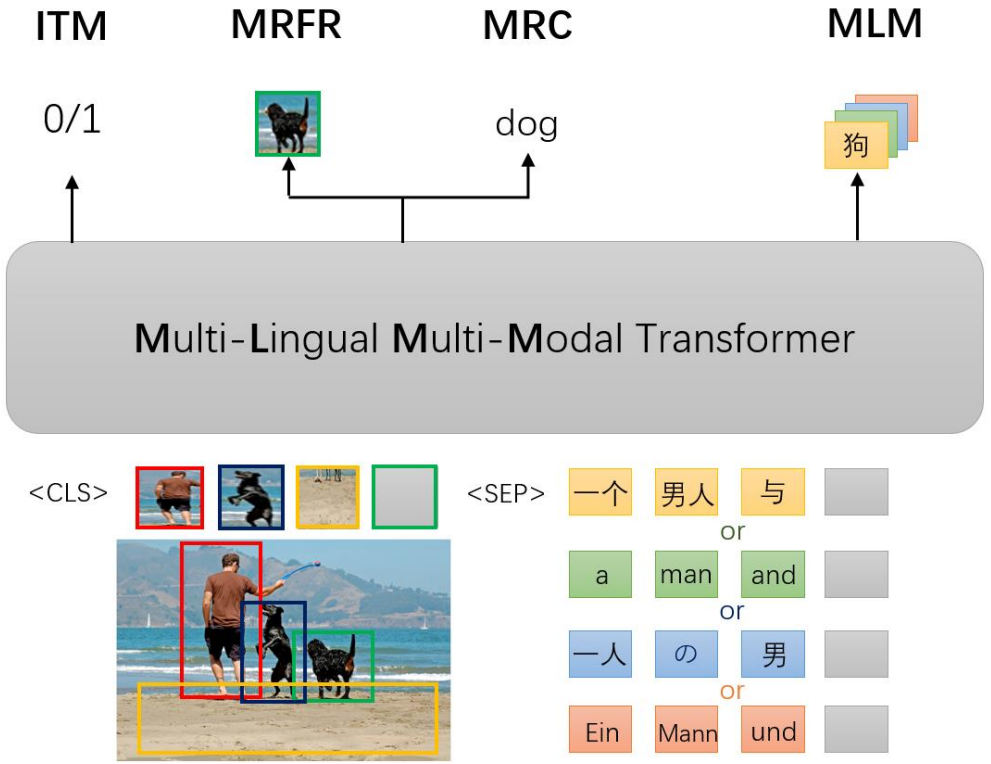
J Liang, R Li, and Q Jin. Semi-supervised Multi-modal Emotion Recognition with Cross-Modal Distribution Matching. ACM MM 2020

Multimodal Pretraining

- **MLMM: Learning Semantic Association cross Multimodalities via Self-supervised Learning Tasks**

悟道-文澜





Emotion Challenge

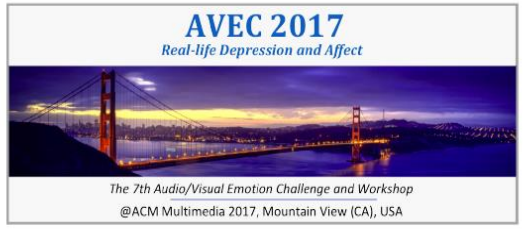
The AAAC (The Association for the Advancement of Affective Computing) is a professional, world-wide association for researchers in Affective Computing, Emotions and Human-Machine Interaction.

Audio-Visual Emotion Challenge (AVEC) aims at comparison of multimedia processing and machine learning methods for automatic audio, visual and audiovisual emotion analysis, with all participants competing under strictly the same conditions.

- Jinming Zhao, Ruichen Li, Jingjun Liang, Shizhe Chen, Qin Jin. **Adversarial Domain Adaption for Multi-Cultural Dimensional Emotion Recognition in Dyadic Interactions.** *ACM Multimedia AVEC 2019.* **(Challenge Winner).**
- Jinming Zhao, Ruichen Li, Shizhe Chen, Qin Jin. **Multi-modal Multi-cultural Dimensional Continues Emotion Recognition in Dyadic Interactions.** ACM Multimedia AVEC 2018. (**Challenge Winner**).
- Shizhe Chen, Qin Jin, Jinming Zhao and Shuai Wang. **Multimodal Multi-task Learning for Dimensional and Continuous Emotion Recognition.** ACM Multimedia AVEC 2017. (**Challenge Winner**).

Emotion Challenge

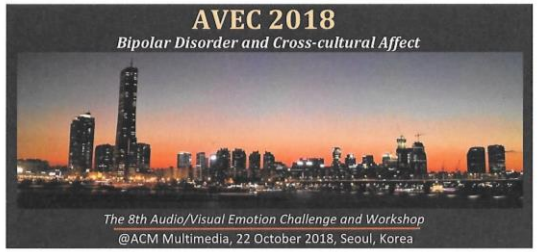
Winner of ACM Multimedia 2017 & 2018 & 2019
Audio Visual Emotion Challenge (AVEC)



Winner of the Affect Sub-Challenge

Multi-task Learning for Dimensional and Continuous Emotion Recognition

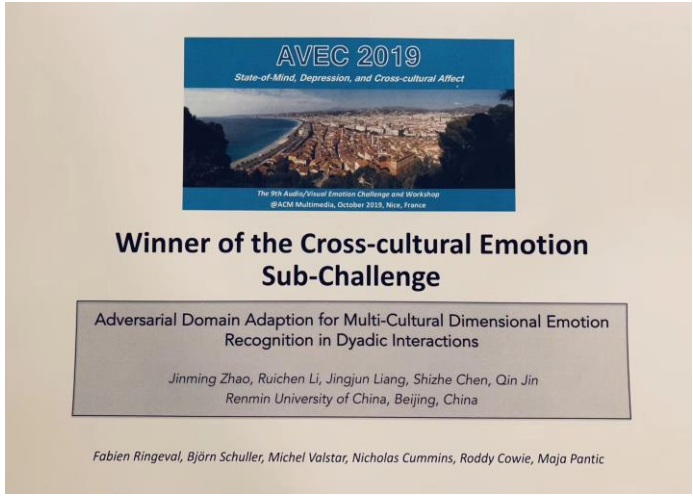
Shizhe Chen, Qin Jin, Jinming Zhao, Shuai Wang



Winner of the Cross-cultural Emotion Sub-Challenge

Multi-modal Multi-cultural Dimensional Continues Emotion Recognition in Dyadic Interactions

Jinming Zhao, Ruichen Li, Shizhe Chen, Qin Jin
Renmin University of China, Haidan, Beijing, China



- Jinming Zhao, Ruichen Li, Jingjun Liang, Shizhe Chen, Qin Jin. **Adversarial Domain Adaption for Multi-Cultural Dimensional Emotion Recognition in Dyadic Interactions.** *ACM Multimedia AVEC 2019.* **(Challenge Winner).**
- Jinming Zhao, Ruichen Li, Shizhe Chen, Qin Jin. **Multi-modal Multi-cultural Dimensional Continues Emotion Recognition in Dyadic Interactions.** ACM Multimedia AVEC 2018. (**Challenge Winner**).
- Shizhe Chen, Qin Jin, Jinming Zhao and Shuai Wang. **Multimodal Multi-task Learning for Dimensional and Continuous Emotion Recognition.** ACM Multimedia AVEC 2017. (**Challenge Winner**).

Multimodal Affective Computing @ RUC AI·M^3

- **Person Affective analysis via multimodalities**
  - Multimodality Fusion
  - Conversation Context Encoding
  - Cross-culture Adaptation
  - Semi-supervised/Self-supervised Learning
- **Media Data Affective Analysis**
  - Image/video Memorability Prediction

- **Affective Computing is an important step in building Artificial Intelligence**
  - Long way to go

Thank You !

qjin@ruc.edu.cn