Towards Multimodal Sarcasm Detection

Speaker：李 想

Sarcasm

- *Maybe it's a good thing we came here. It's like a lesson in what not to do.*

Multimodal

Chandler :

Oh my god! You almost gave me a heart attack!

- **Text** : suggests fear or anger.
- **Audio** : animated tone
- **Video** : smirk, no sign of anxiety

MUStARD

- **Source**: TV shows
  - Friends, The Golden Girls, Sarcasmaholics Anonymous, The Big Bang Theory

- **Annotation:**
  - 345 videos labeled as sarcastic
  - 6,020 videos labeled as non-sarcastic

- **Dataset:**
  - 690 samples, balanced
  - video, audio, text, speaker identifier

*ACL 2019, Towards Multimodal Sarcasm Detection (An *Obviously* Perfect Paper)

Sentiment and Emotion help Sarcasm? A Multi-task Learning Framework for Multi-Modal Sarcasm, Sentiment and Emotion Analysis

ACL 2020

Motivation

- Hypothesize that sarcasm is closely related to sentiment and emotion
- MUStARD: only use SVM as baseline
- Contextual Inter-modal Attention for Multi-modal Sentiment Analysis *(EMNLP 2018)*
- Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis (*NAACL-HLT 2019*)

Overview

Manually annotate MUStARD with sentiment and emotion labels

A multi-modal conversational scenario

A multi-task deep learning framework

Solve all these three problems simultaneously

Sentiment and Emotion Annotation

- **Sentiment:**

  Explicit & implicit:

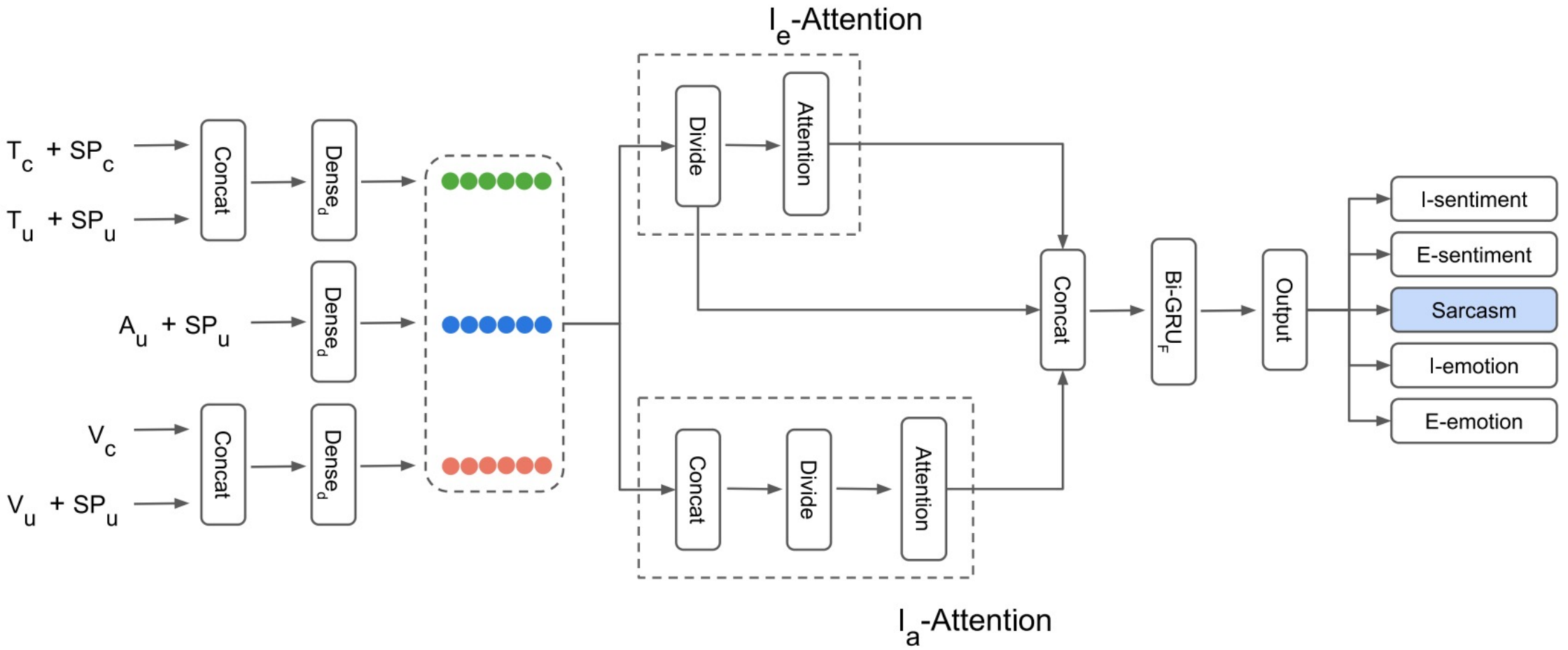
  positive, negative, neutral

- **Emotion**:

  Explicit & implicit:

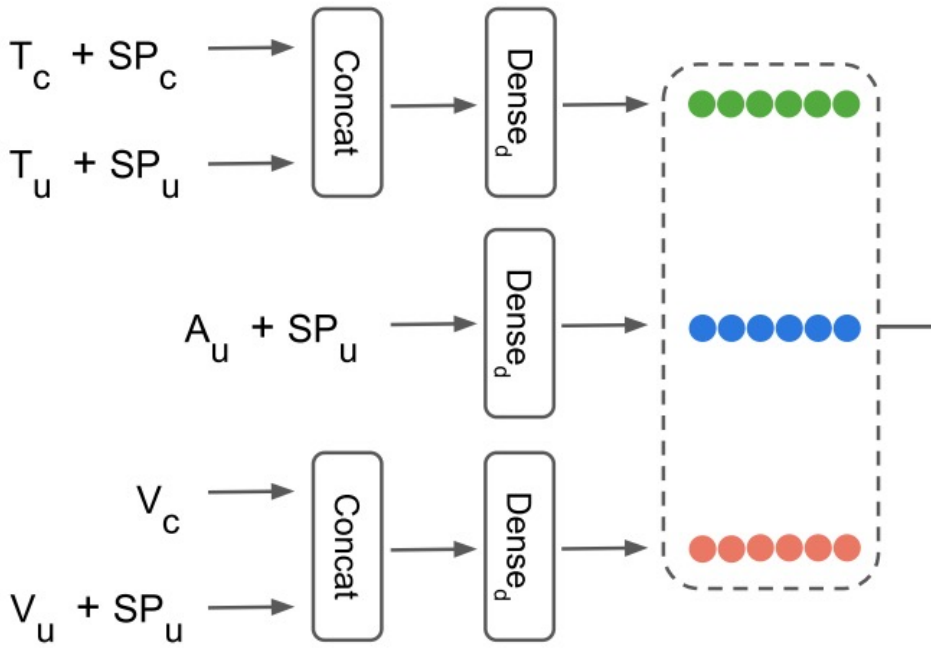
  anger (An), excited (Ex), fear (Fr), sad (Sd), surprised (Sp), frustrated (Fs), happy (Hp), neutral (Neu) and disgust (Dg)

Proposed Methodology



Input Layer Attention Mechanism Output Layer

Input Layer



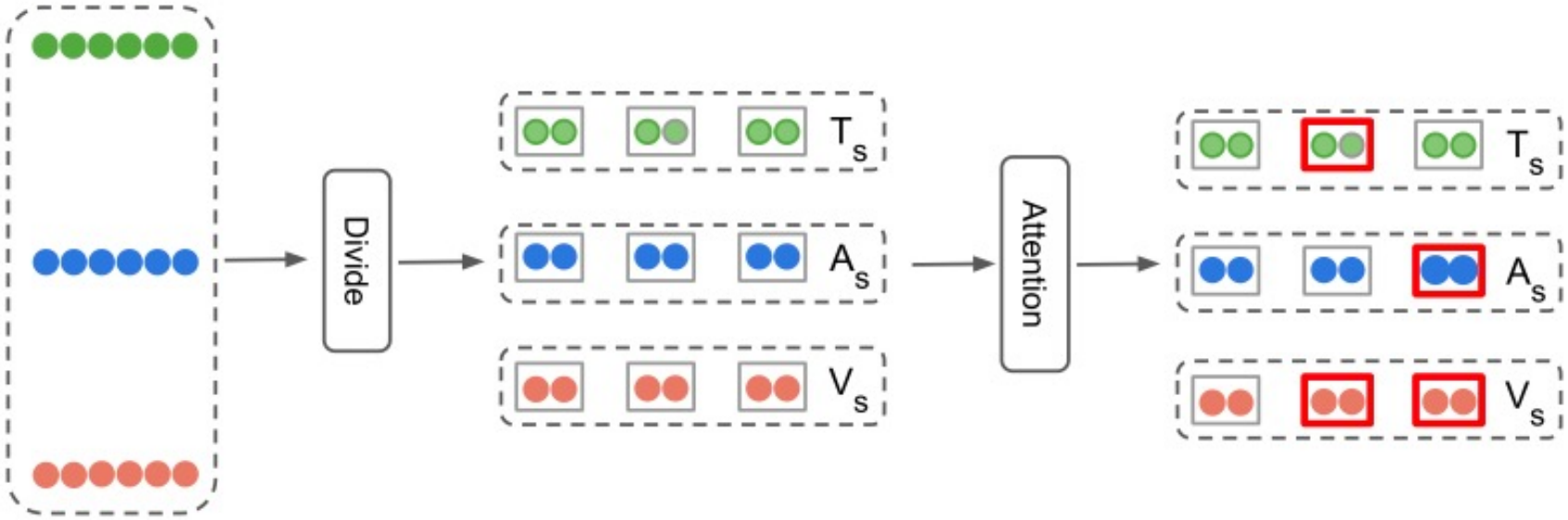
- Text:
  - fastText embedding, 300d
  - BiGRU+Attention
  - Concatenate SP
- Visual:
  - Average of all frames + SP, 2048d
  - Context: average of all sentences, no SP
- Acoustic:
  - Average of all frames + SP, 283d
  - No context information

➢Dense:

- Fully-connected layer → feature vector (length d)

Attention Mechanism -- I_e-Attention

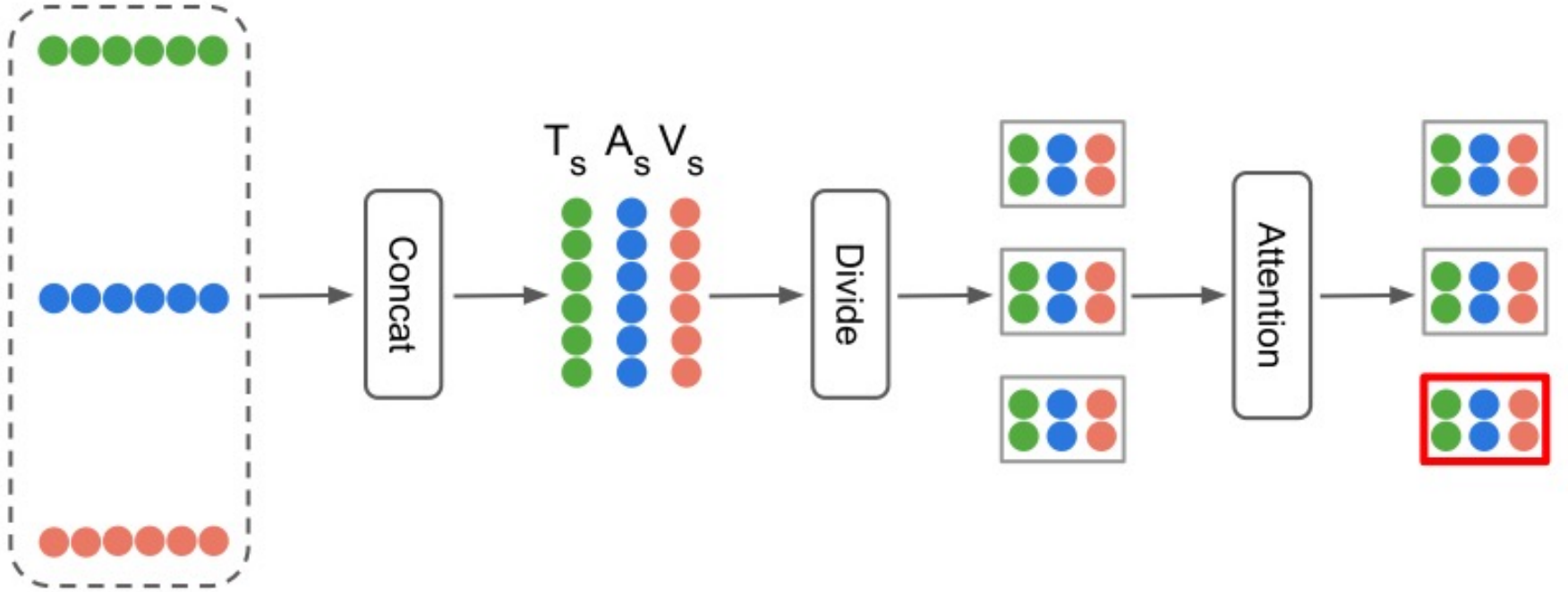
Learn the relationship between the feature vector of a segment of an utterance in one modality and feature vector of the another segment of the same utterance in another modality through this mechanism



I_e-Attention mechanism: Inter-segment Inter-modal Attention

Attention Mechanism -- I_a-Attention

For a specific segment of any particular utterance, to establish the relationship between the feature vectors obtained from the different modalities

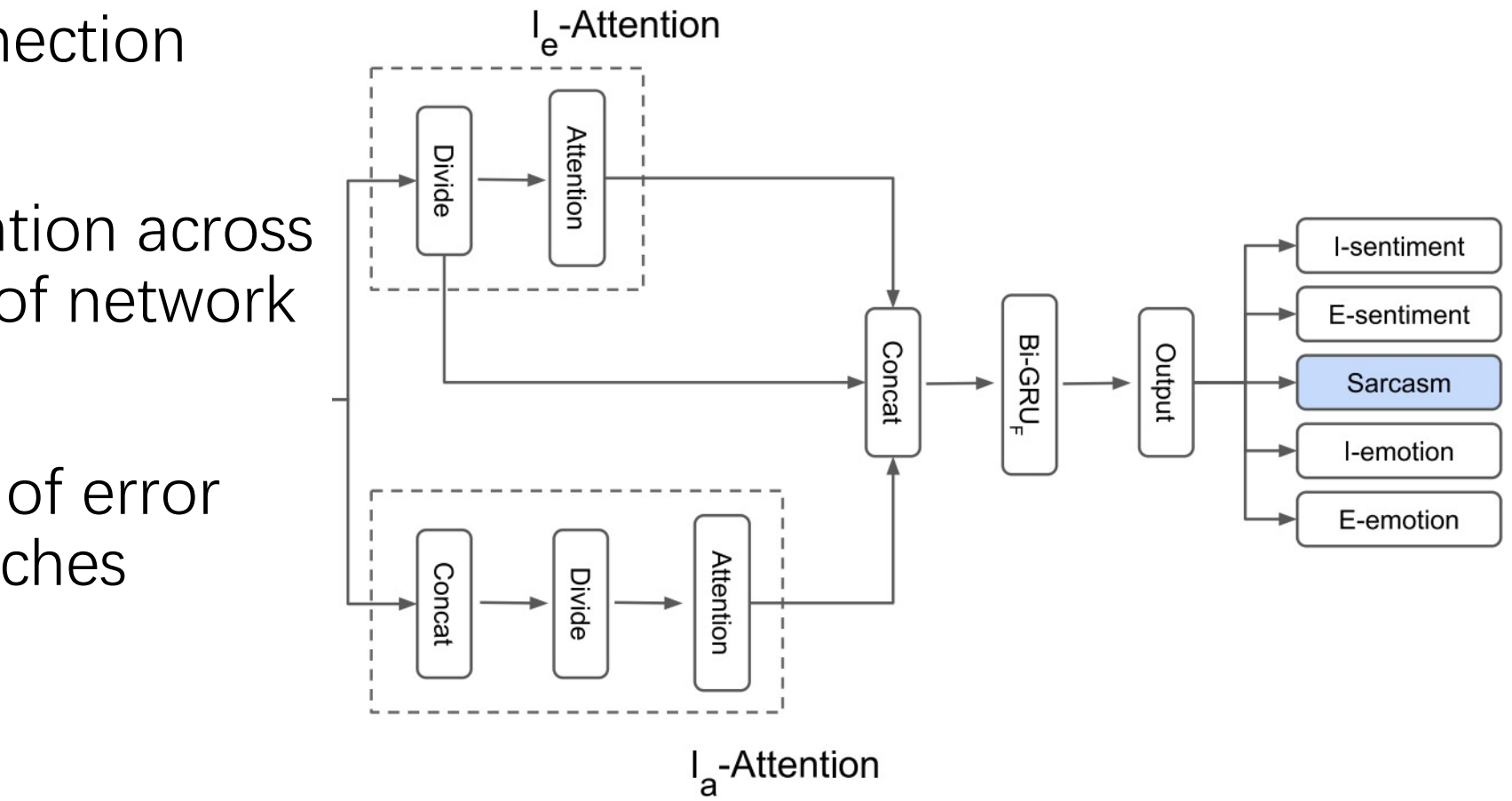


I_a-Attention mechanism: Intra-segment Inter-modal Attention

Output Layer

- Residual skip connection
- Shared representation across the five branches of network
- Receive gradients of error from the five branches

  → Adjust weights



Experiments

Labels			T + V			T + A			A + V			T + A + V		
			P	R	F1	P	R	F1	P	R	F1	P	R	F1
Speaker Dependent	*STL*	***Sar***	71.52	70.61	69.32	64.20	64.20	63.88	71.90	71.01	70.64	72.08	71.62	72.01
	MTL	***Sar + Sent***	69.65	69.42	69.33	64.09	60.72	58.21	72.20	71.45	71.18	72.52	71.73	72.07
		Sar + Emo	71.76	70.86	70.54	**65.76**	**65.65**	**65.60**	72.60	71.59	71.25	72.76	71.88	72.11
		Sar + Sent + Emo	**72.76**	**71.88**	**71.61**	62.23	61.15	59.61	**72.73**	**71.88**	**71.81**	**73.40**	**72.75**	**72.57**
Speaker Independent	*STL*	***Sar***	60.11	60.18	60.16	58.23	57.69	57.91	60.44	60.96	60.52	65.98	65.45	65.60
	MTL	***Sar + Sent***	62.74	62.92	62.81	59.25	59.55	52.89	61.60	60.95	61.14	66.97	63.76	63.68
		Sar + Emo	65.11	65.16	65.13	**59.59**	59.55	**59.58**	63.19	63.76	62.91	66.35	65.44	65.63
		Sar + Sent + Emo	**65.48**	**65.48**	**65.67**	59.13	**59.98**	50.27	**65.59**	**63.76**	**63.90**	**69.53**	**66.01**	**65.90**

STL vs. MTL on Sarcasm Classification: *without context without speaker* information

Experiments

Speaker Dependent					
Implicit Sentiment			*Explicit Sentiment*		
P	R	F1	P	R	F1
49.27	57.39	49.12	48.32	52.46	48.11
Speaker Independent					
P	R	F1	P	R	F1
47.05	49.15	40.99	47.73	50.0	45.24

Speaker Dependent					
Implicit ~~Sentiment~~			*Explicit ~~Sentiment~~*		
P	R	F1	P	R	F1
80.66	88.51	83.57	85.01	88.90	85.12
Speaker Independent					
P	R	F1	P	R	F1
81.77	88.29	83.88	83.64	88.35	84.37

Emotion

Results for Single-task experiments for Sentiment/Emotion analysis (T+A+V).

Experiments

Setups		*Speaker Dependent*			*Speaker Independent*		
Context	Speaker	P	R	F1	P	R	F1
✗	✗	73.40	72.75	72.57	69.53	66.01	65.90
✗	✓	**77.09**	**76.67**	**76.57**	74.69	74.43	74.51
✓	✗	72.34	71.88	71.74	71.51	71.35	70.46
✓	✓	76.07	75.79	75.72	**74.88**	**75.01**	**74.72**

Setup	*Speaker Dependent*			*Speaker Independent*		
	P	R	F1	P	R	F1
W/o Attention	71.53	69.71	69.02	60.53	61.23	60.44
Proposed	**73.40**	**72.75**	**72.57**	**69.53**	**66.01**	**65.90**

Comparative Analysis

Setup	*Model*	**T + V**			**T + A**			**A + V**			**T + A + V**		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Speaker Dependent	***Baseline***	72.0	71.6	71.6	**66.6**	**66.2**	**66.2**	66.2	65.7	65.7	71.9	71.4	71.5
	Proposed Model	**72.8**	**71.9**	**71.6**	62.2	61.2	59.6	**72.7**	**71.9**	**71.8**	**73.4**	**72.8**	**72.6**
	T-test	-	-	-	-	-	-	-	-	-	0.0023	0.0098	0.0056
Speaker Independent	***Baseline***	62.2	61.5	61.7	**64.7**	**62.9**	**63.1**	64.1	61.8	61.9	64.3	62.6	62.8
	Proposed Model	**65.5**	**65.5**	**65.7**	59.1	60.0	50.3	**65.6**	**63.8**	**63.9**	**69.5**	**66.0**	**65.9**
	T-test	-	-	-	-	-	-	-	-	-	0.0002	0.0006	0.0012

Conclusion

- Proposed an effective deep learning-based multi-task model to simultaneously solve all the three problems
- Extend MUStARD dataset with sentiment and emotion
- Achieves better performance for sarcasm detection
- The dataset is not big enough for a complex framework to learn from

Thanks

Q & A