Reactive Multi-Stage Feature Fusion for Multimodal Dialogue Modeling

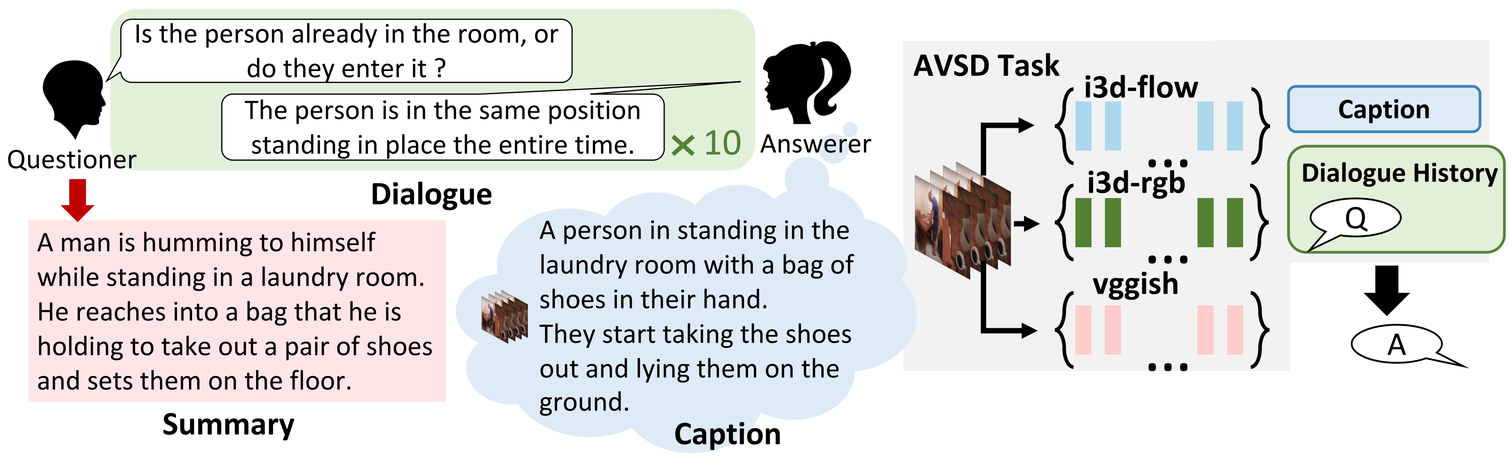
National Taiwan University

Yi-Ting Yeh Tzu-Chuan Lin Hsiao-Hua Cheng Yi-Hsuan Deng Shang-Yu Su Yun-Nung (Vivian) Chen

https://github.com/MiuLab/DSTC7

Summary

➢ **Task: Audio Visual Scene-Aware Dialogue (AVSD)**



➢ **Motivation**

- Single large attention module is not capable to model the complicated relation of different modalities
- To make different modalities fully interact with each other, we need to integrate them in multiple stages.

➢ **Approach**

- Multi-stage fusion encourage the model to thoroughly integrate information from different modalities
- Fuse features from different modalities and stages by cross modality fusion

➢ **Results**

- Significantly outperforms the baseline in CIDEr and ROUGE-L

Proposed Model

➢ **Multi-Stage Fusion**

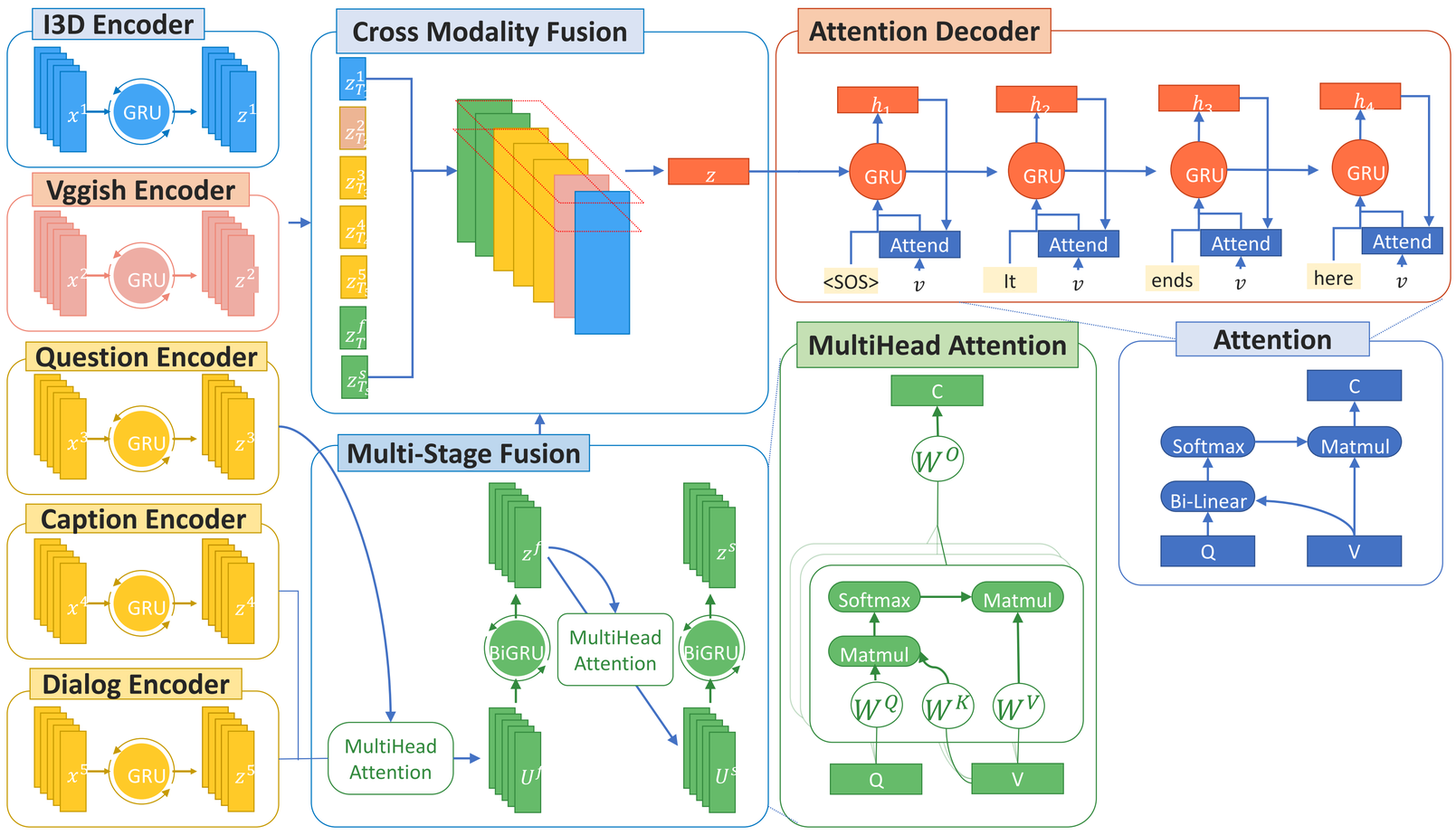
- Apply the **multi-head attention** to fuse encoded question into caption and dialogue
- Perform **self-attention** to help model gather long term information in the dialogue.

➢ **Cross Modality Fusion**

- Insert **1x1 convolution** to help different feature channels interact with each other
- **Weighted sum** fuse all modalities into single vector

➢ **Attention Decoder**

- Calculate the attention of a query and values by **low-rank bilinear method**
- Concatenate the context vector with the next step input as the attention-enhanced input



Experiments & Results

➢ **Dataset:** Official AVSD Dataset

- 7659, 1787, 1710 dialogues for train, dev, and test sets respectively

➢ **Result**

- Encoding
  - The proposed multi-stage fusion boosts the CIDEr score
  - 1x1 convolution fusion enables interactions between different modalities, and hence improves performance
- Decoding
  - Attention decoder mainly improves BLEU scores
- Proposed Fusion Model: multi-stage fusion + 1x1 convolution fusion + attention decoder
  - Outperform baseline by a large margin in CIDEr, improve METEOR and ROUGE-L, and achieve comparable results in BLEU-4

Model Encoder	Model Decoder	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Naïve Copy		0.124	0.077	0.049	0.111	0.235	0.637
Released (Hori et al. 2018)		**0.172**	**0.118**	0.085	0.115	0.292	0.790
Simple Fusion	Simple	0.157	0.112	0.084	0.120	0.305	0.994
Multi-stage	Simple	0.157	0.113	0.086	0.119	0.308	1.009
1x1 Convolution	Simple	0.162	0.117	0.088	0.122	0.310	1.013
Simple Fusion	Attention	0.162	0.115	0.086	0.119	0.308	0.977
Multi-Stage + 1x1 Conv	Attention	0.163	**0.118**	**0.090**	**0.122**	**0.315**	**1.059**

Video Frame	Type	Sentences
	Caption	a person in the entryway is working on something on their phone. they start throwing some clothes at another person who is watching them oddly.
	Question	are they laughing in the video?
	Ground Truth	they are not laughing out loud but are smiling and appear maybe to be flirting a bit.
	Released Baseline	no, they are both talking to each other at the end of the video.
	Basic Fusion Model	no they are not talking.
	Multi-Stage Fusion Model	no, they are not laughing.

Conclusion

- Proposes an intuitive and effective visual dialogue model based on an encoder-decoder design
- Develop a set of modules to fuse multimodal features and perform context-aware decoding
- Significantly improves CIDEr score compared to the baseline
- Outperforms baseline in human evaluation, which shows the superiority of multi-stage model

r07922064@ntu.edu.tw b04902013@ntu.edu.tw
b04705003@ntu.edu.tw f05921117@ntu.edu.tw
b03902024@ntu.edu.tw y.v.chen@ieee.org

Code Available:
https://github.com/MiuLab/DSTC7