PEKING UNIVERSITY

Advanced Data & Signal Processing laboratory

Improving the Performance of Automated Audio Captioning via Integrating the Acoustic and Semantic Information

Zhongjie Ye[1], Helin Wang[1], Dongchao Yang[1], Yuexian Zou[1, 2, *]

[1]ADSPLAB, School of ECE, Peking University, Shenzhen, China [2]Peng Cheng Laboratory, Shenzhen, China

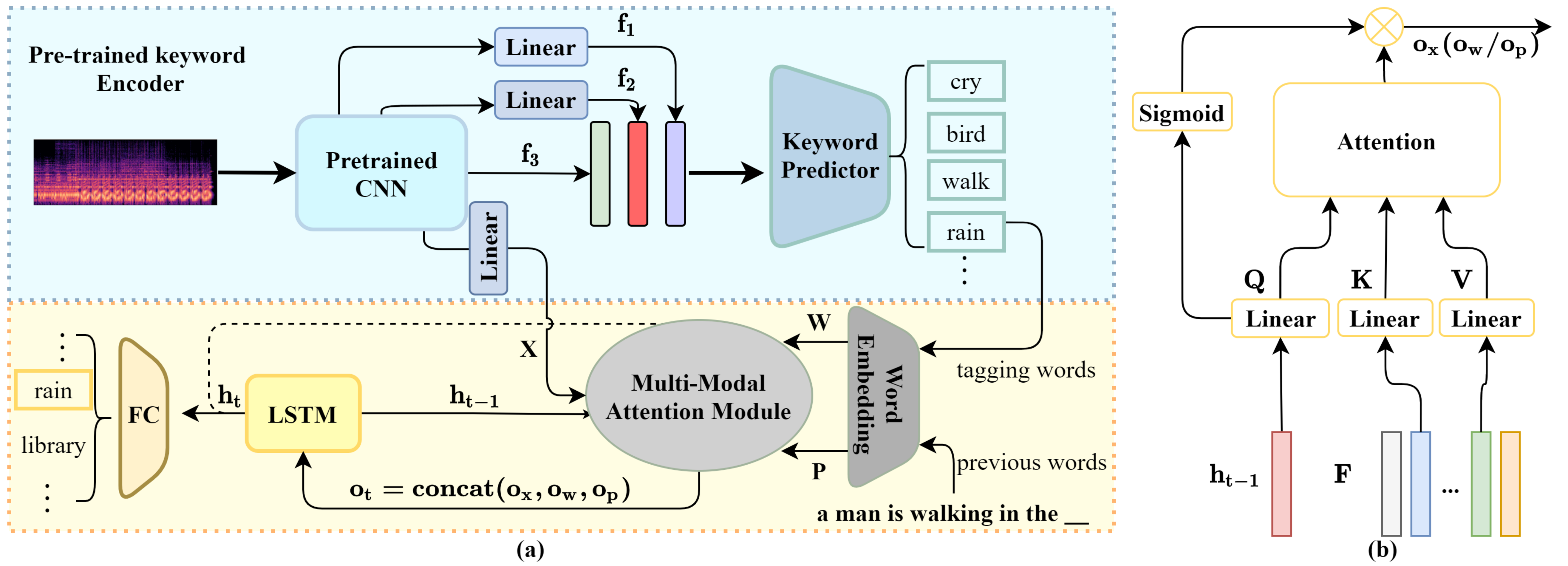
Motivation

Our goal is to improve the performance of automated audio captioning by enhancing the ability of the encoder to recognize audio concepts and utilizing both acoustic and semantic information in the decoder.

- **Acoustic information** is obtained from the encoder.
- **Semantic information** contains (1) tagging words that are audio concepts recognized from the encoder and (2) previously predicted words that contain all the generated words before the current time.

Proposed Method

Key idea: we build a pre-trained encoder to enhance the ability of the encoder to recognize audio concepts and a multi-modal attention module to utilize both acoustic and semantic information.



Pre-trained keyword encoder:

- Using PANNs to initialize the parameters of the encoder.
- A feature hierarchy structure to combine multi-level features:

$$\hat{y} = \sigma(Linear(concat(f_1, f_2, f_3)))$$

- Optimized by minimizing the binary cross-entropy loss:

$$\mathcal{L}_{bce}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N} y(i) log\ \hat{y}(i)$$

Multi-modal attention module:

- The addictive attention with a gated linear unit is applied to acoustic or semantic information shown in (b).
- We add the outputs of acoustic and semantic attention module with the predicted word of the last time step to get the current hidden state:

$$h_t = LSTM(h_{t-1}, Add(o_x, o_w, o_p, \mathbf{Emb}(w_{t-1})))$$

- We use hidden state h_t to predict the word in the current time.

Experiments and Results

- We evaluate our proposed model on the Clotho v2 dataset, which contains 3,839 training, 1,045 validation, and 1,045 testing audio clips.
- Training periods: cross-entropy and CIDEr-D optimization.
- B1, B4, RG, ME, CD, SP, and SD denote BLEU-1, BLEU-4, METEOR, ROUGE-L, CIDEr-D, SPICE, and SPIDEr.

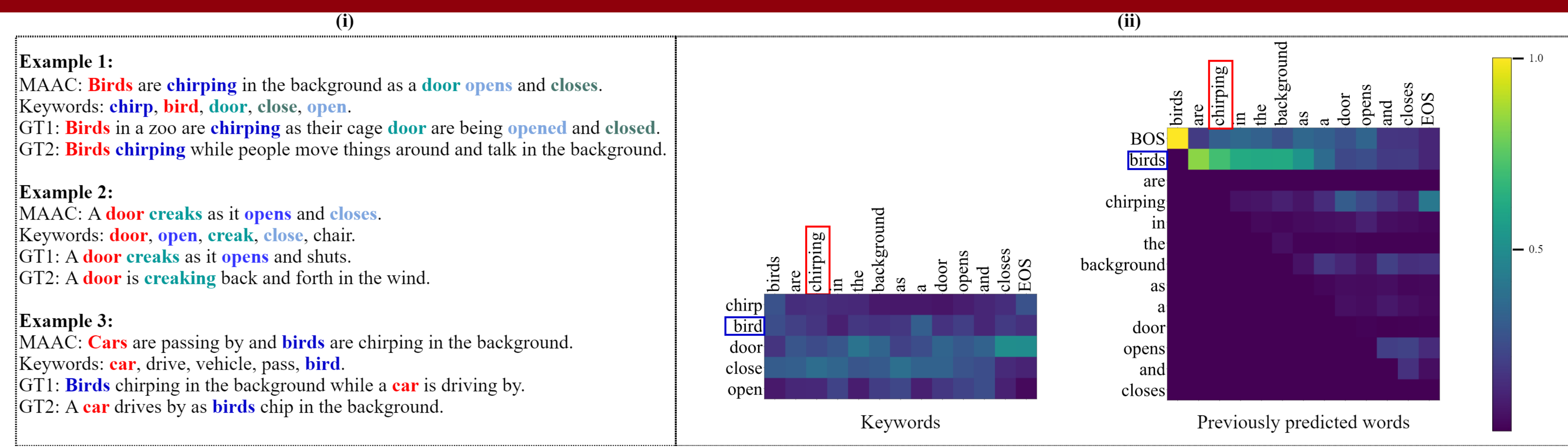
	Cross-entropy							CIDEr-D optimization						
Model	B1	B4	RG	ME	CD	SP	SD	B1	B4	RG	ME	CD	SP	SD
Baseline [10]	37.8	1.7	26.3	7.8	7.5	2.8	5.1	-	-	-	-	-	-	-
TAM [5]	48.9	10.7	32.5	14.8	25.2	9.1	17.2	-	-	-	-	-	-	-
TM [4]	53.4	15.1	35.6	16.0	34.6	10.8	22.7	-	-	-	-	-	-	-
UNIS's model [11]	-	-	-	-	-	-	-	62.5	17.8	40.1	17.6	42.8	12.6	27.7
SJTU's model [12]	56.5	15.5	37.4	17.4	39.9	11.9	25.9	64.0	16.3	40.4	17.8	44.9	12.3	28.6
MAAC (Ours)	57.7	17.4	37.7	17.4	41.9	11.9	26.9	**64.8**	**18.1**	**40.8**	**19.0**	**49.1**	**13.1**	**31.1**

Model	B4	CD	SD
Base	16.5	40.6	26.4
+ Previously predicted words	48.9	10.7	32.5
+ Keywords	-	-	-
+ Both (w/o sharing SAM)	16.8	41.1	26.7
proposed MAAC (Ours)	**17.4**	**41.9**	**26.9**

Results:

- Our proposed model achieves the highest score on all metrics both in the cross-entropy and CIDEr-D optimization stages.
- The CIDEr-D score of the proposed model improves from 41.9 to 49.1 after further optimizing CIDEr-D.
- Pre-trained keyword encoder and multi-modal attention module could improve the performance of the automated audio captioning.

Visualization



Analysis:

- The pre-trained keyword encoder can almost recognize the main concepts *i.e.* keywords and the keywords may appear in different states in the ground-truth captions and the predicted sentences.
- Attention maps of semantic information indicate that keywords and previously predicted words are concerned to generate the current word.