Towards Exploiting Sticker for Multimodal Sentiment Analysis in Social Media: A New Dataset and Baseline

Feng Ge **Weizhao Li** **Haopeng Ren** **Yi Cai** *

South China University of Technology

logosg@foxmail.com

se_weizhao.li@mail.scut.edu.cn

renhp_scut@foxmail.com

ycai@scut.edu.cn

COLING-2022

Reported by JiaWei Cheng

Introduction

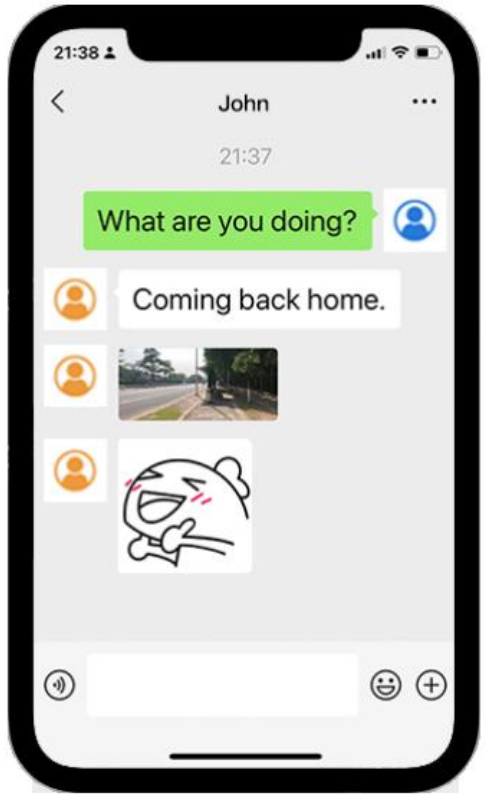


Figure 1: An example of online conversation. The first image is a photo that does not reflect any sentiment. The second is a sticker, expressing the emotional tendency of happiness. The sticker here contributes to supplementing the missing sentiment of the text.

Introduction

First, stickers may be inherently multimodal because they are embedded with texts,while other datasets have only real-world photos

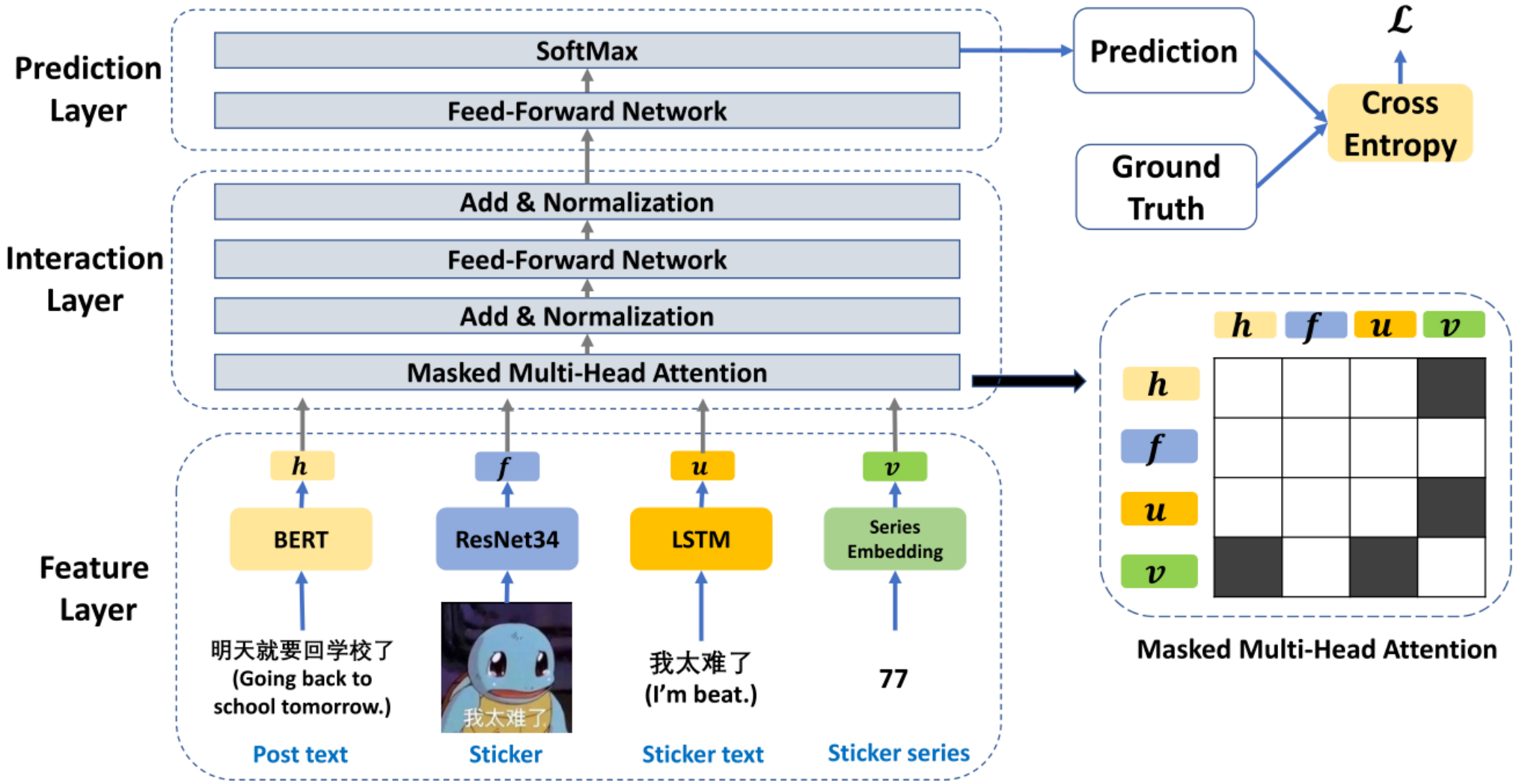
Introduction

Second,stickers are highly different in styles, leading models to learn robust representations for the stickers following various distributions hardly

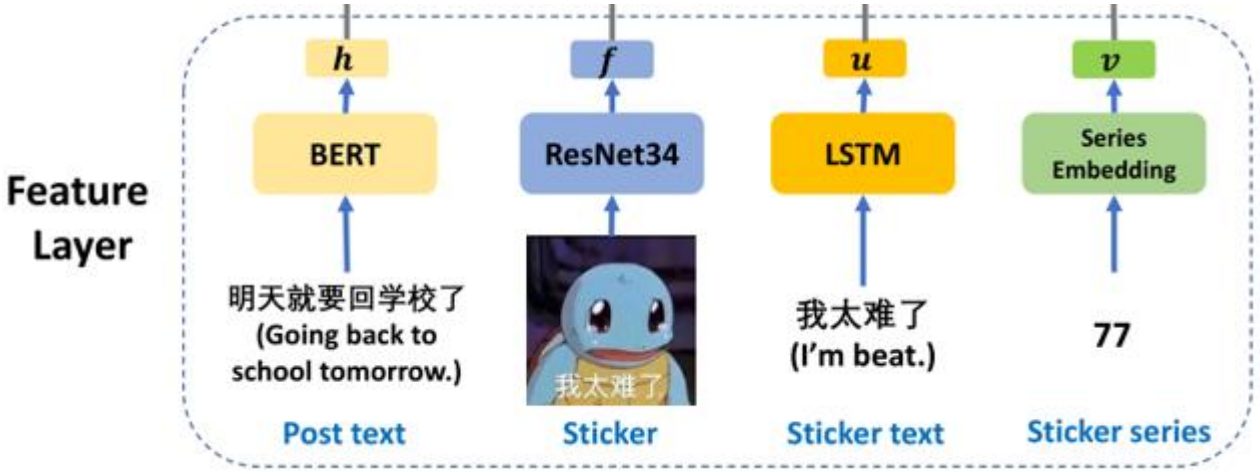
Introduction

Third, the sentiment fusion of text and stickers is complex

Overview


Prediction Layer
SoftMax
Feed-Forward Network
Interaction Layer
Add & Normalization
Feed-Forward Network
Add & Normalization
Masked Multi-Head Attention
Feature Layer
h
f
u
v
BERT
ResNet34
LSTM
Series Embedding
明天就要回学校了
(Going back to school tomorrow.)
我太难了
我太难了
(I'm beat.)
77
Post text
Sticker
Sticker text
Sticker series
Prediction
Ground Truth
Cross Entropy
ℒ
Masked Multi-Head Attention

Method



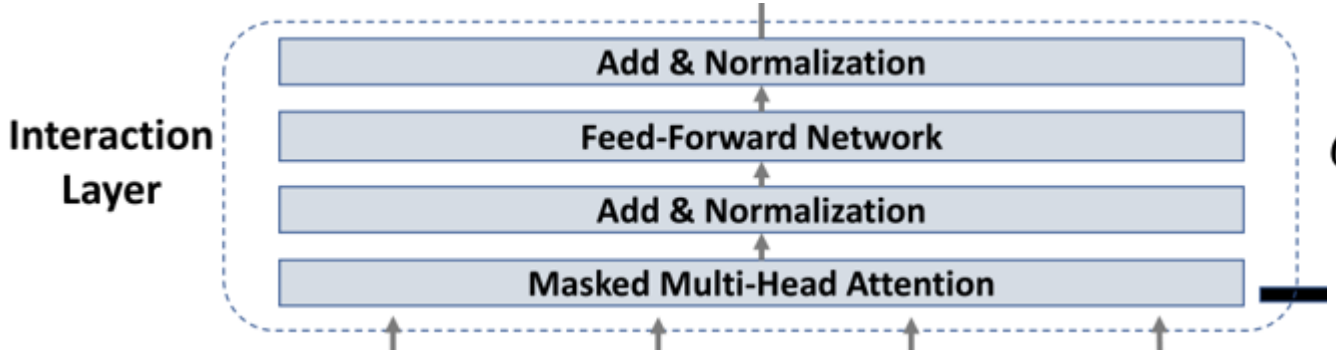
$$h = BERT(X) \quad (1)$$

$$f = ResNet(I) \quad (2)$$

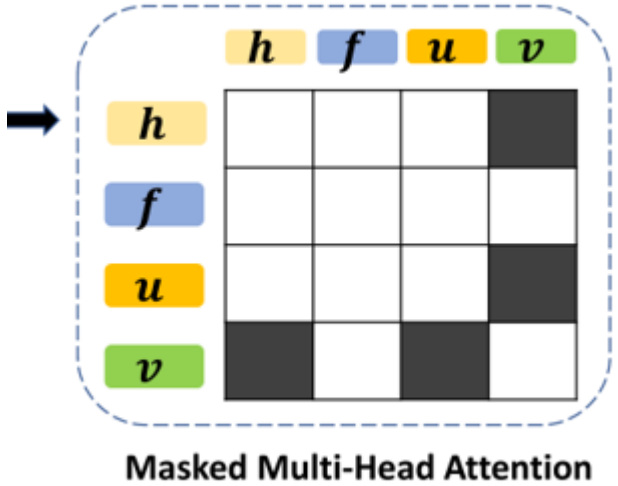
$$u = LSTM(S) \quad (3)$$

$$v = Embedding(E) \quad (4)$$

Method



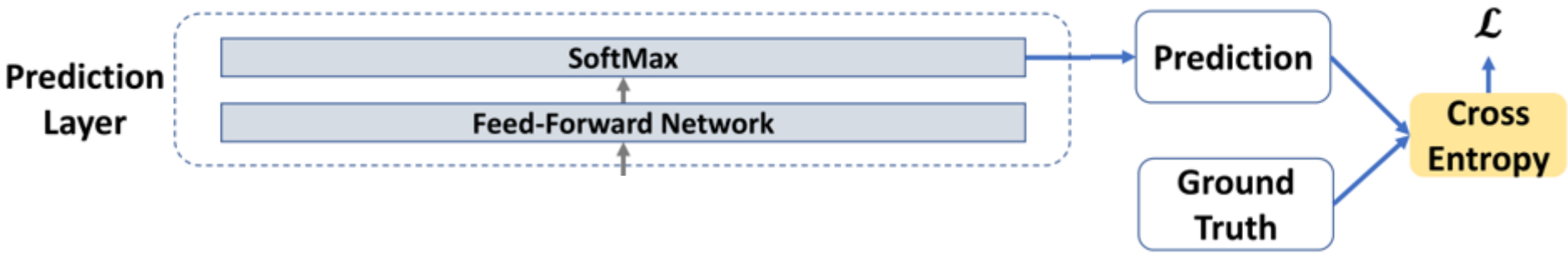
$$Q_i = FN(C_i),\ K_i = FN(C_i),\ V_i = FN(C_i) \tag{5}$$



$$C'_i = \sum_{j=1}^{4} \alpha_{i,j} * V_i \tag{6}$$

$$\alpha_{i,j} = \frac{exp(Q_i * K_j \odot \mathcal{M})}{\sum_{k=1}^{4} exp(Q_i * K_k \odot \mathcal{M})} \tag{7}$$

Method



$$P(y|X, I, S, E) = softmax(FFN(O)) \quad (8)$$

$$O = [C'_1; C'_2; C'_3; C'_4].$$

$$\mathcal{L} = -\frac{1}{|\mathcal{D}|} \sum_{c \in \mathcal{D}} logP(y^{(c)}|X^{(c)}, I^{(c)}, S^{(c)}, E^{(c)}) \quad (9)$$

Experiments

Datasets	Size	# Sti.	# Ano.
MOD	45k	307	0
StickerChat	**340k**	174k	0
CSMSA	28k	16k	1.5k

Table 1: Comparison with other sticker-based datasets. # Sti. represents the number of stickers. # Ano. represents the number of samples with human-annotated multimodal sentiment label.

Experiments

Task	Train	Valid	Test
Easy Task	942	314	314
Hard Task-1	1297	127	146
Hard Task-2	1290	130	150
Hard Task-3	1373	109	88

Table 2: The split statistics of the CSMSA dataset.

Figure 4: An example shows the division of Hard Task-3 and the significant variation of styles between the stickers series.

Experiments

	Methods	Easy Task		Hard Task-1		Hard Task-2		Hard Task-3	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Text-only	BERT	0.6178	0.5764	0.5274	0.4709	0.5467	0.5102	0.5114	0.3617
	BERT-ST	0.6146	0.5609	0.4932	0.4772	0.5333	0.4769	0.5227	0.3814
	RoBERTa	0.6210	0.5941	0.5479	0.4892	0.5133	0.5216	0.5341	0.4194
Image-only	ResNet34	0.6178	0.5701	0.5753	0.5530	0.5342	0.4778	0.4886	0.3484
Multimodal	mBERT	0.5924	0.5576	0.5753	0.5156	0.5600	0.5361	0.5341	0.4037
	MMTF	0.5955	0.5374	0.5479	0.4886	0.5267	0.4876	0.5227	**0.4428**
	SAMSAM	**0.6369**	0.6180	**0.5959**	**0.5669**	**0.5533**	0.5179	**0.5455**	0.4265
Ablation Study	w/o MASK	0.6306	0.6060	0.5685	0.5483	0.5133	0.4929	0.4773	0.3710
	w/o IMG	0.6274	**0.6199**	0.5685	0.5467	0.5200	**0.5251**	0.5114	0.3913
	w/o PT	0.6210	0.5665	0.5411	0.4845	0.5267	0.4924	0.5114	0.3782
	w/o SE	0.6146	0.5594	0.5685	0.5112	0.5467	0.4978	0.5227	0.4133
	w/o ST	0.6242	0.6179	0.5890	0.5664	0.5267	0.5006	0.5117	0.3925

Table 3: Overview of the experimental results. Acc. represents accuracy, and the F1 represents the weighted F1 score. MASK represents the mask mechanism. IMG represents the image feature. PT represents the post text. SE represents the sticker series embedding, and ST represents sticker text.

Experiments

Post Text: *我直接举报了*
(I reported it directly.)

Sticker text: 让人怪不好意思的
(Sorry not sorry.)

× BERT: Negative
√ ResNet: Positive
× MMTF: Negative
√ SAMSAM: Positive
Ground Truth: Positive

Post Text: *你要做什么呀*
(What do you want to do?)

Sticker text: *一个耿直的微笑*
(A straight smile.)

× BERT: Neutral
× ResNet: Neutral
× MMTF: Neutral
√ SAMSAM: Positive
Ground Truth: Positive

Figure 6: Examples of multimodal sentiment analysis produced by different models on CSMSA dataset.

Thanks !