

# Multi-modal Semantic Understanding with Contrastive Cross-modal Feature Alignment

**Ming Zhang**<sup>1,2\*</sup>, **Ke Chang**<sup>1,3\*</sup>, **Yunfang Wu**<sup>1,3†</sup>

<sup>1</sup>National Key Laboratory for Multimedia Information Processing, Peking University

<sup>2</sup>School of Software and Microelectronics, Peking University

<sup>3</sup>School of Computer Science, Peking University

zhangming@stu.pku.edu.cn, {changkegg, wuyf}@pku.edu.cn

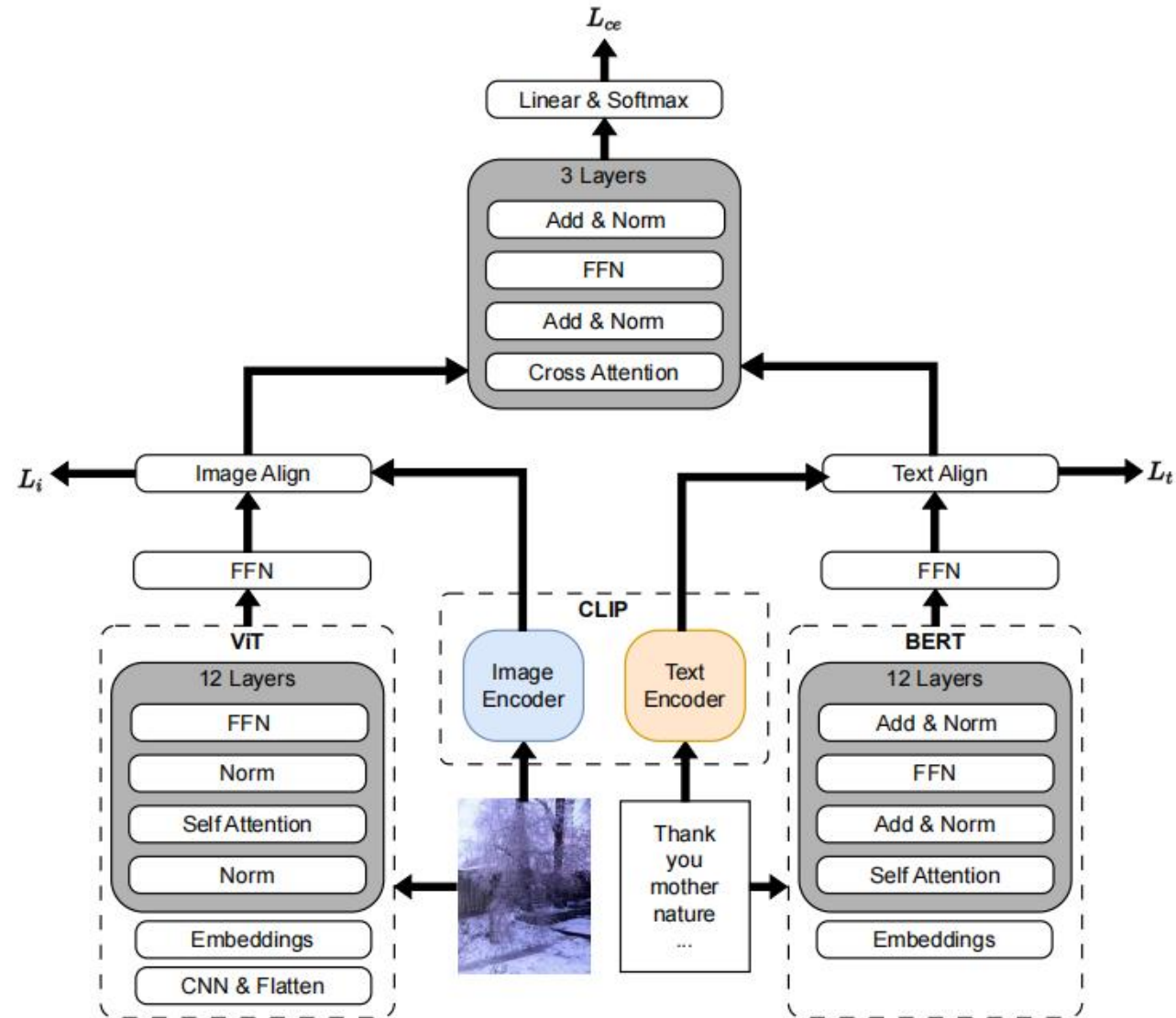
# Motivation

- Multimodal sarcasm detection require both text and image clues.
- Previous methods using dual encoder structure neglect to consider the semantic alignment between different modalities when fusing multi-modal features.

# Contribution

- We propose a novel CLIP-guided contrastive learning-based architecture for multi-modal semantic alignment
- Our method outperforms baseline models and achieves comparable results with knowledge-enhanced models
- Our method can be easily combined with other knowledge-based models to get higher performance

# Method - CLFA



# Text and Image Encoding

- For given input sentence  $S = [s_1, s_2, \dots, s_n]$ , text representation

$$f_t(S) = \mathbf{BERT}(S)$$

- Input image is resize and diveded to m patches  $P = [P_1, P_2, \dots, P_m]$   
image representation

$$f_i(P) = \mathbf{ViT}(P)$$

# Cross-modal Feature Alignment

- Obtain CLIP representation

$$C_t = [\mathbf{CLIP}_t(T_1), \dots, \mathbf{CLIP}_t(T_B)]$$
$$C_i = [\mathbf{CLIP}_i(I_1), \dots, \mathbf{CLIP}_i(I_B)]$$

- A mapping net to align feature dimension size

$$f'_t = \mathbf{MLP}(f_t)$$
$$f'_i = \mathbf{MLP}(f_i)$$

- Align the CLIP representation and BERT/ViT representation by using contrastive learning

$$\mathcal{L}_{ic} = -\frac{1}{B} \sum_{k=1}^B \log \frac{e^{\mathbf{sim}(F'_{ik}, C_{ik})/\tau}}{\sum_{j=1}^B e^{\mathbf{sim}(F'_{ik}, C_{ij})/\tau}}$$
$$\mathcal{L}_{ci} = -\frac{1}{B} \sum_{k=1}^B \log \frac{e^{\mathbf{sim}(C_{ik}, F'_{ik})/\tau}}{\sum_{j=1}^B e^{\mathbf{sim}(C_{ik}, F'_{ij})/\tau}}$$

# Cross Attention and Classification

- Three cross attention layers to fuse text representation and image representation

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$h = \text{Attn}(W_q F'_t, W_k F'_i, W_v F'_i)$$

- Final objective function

$$\mathcal{L} = \alpha \mathcal{L}_{con} + \mathcal{L}_{ce}$$

# Results

Categories	Models	Acc(%)	P(%)	R(%)	F1(%)	Macro		
						P(%)	R(%)	F1 (%)
Image Based	Resnet*	71.27	63.02	67.36	65.12	70.20	70.61	70.35
	ViT*	72.15	64.72	66.01	65.36	70.97	71.11	71.03
Text Based	BiLSTM*	76.21	71.59	66.74	69.08	75.27	74.61	74.88
	BERT*	79.95	72.2	80.71	76.22	79.18	80.08	79.44
Multi-modal	CLIP*	84.56	<b>84.57</b>	74.87	79.42	84.56	82.92	83.53
	CLIP+Cross Attention*	85.14	80.82	82.17	81.49	84.45	84.64	84.54
	MLP+CNN	81.61	-	-	-	79.52	72.47	75.83
	HFM	83.44	76.57	84.15	80.18	79.40	82.45	80.90
	D&R Net	84.02	77.97	83.42	80.60	-	-	-
	ResBert	86.05	78.63	83.31	80.90	78.87	84.46	82.92
	BERT+ViT*	83.73	78.12	81.54	79.80	82.94	83.41	83.15
<b>Our CLFA*</b>	<b>86.80</b>	<b>81.51</b>	<b>86.44</b>	<b>83.91</b>	<b>86.09</b>	<b>86.74</b>	<b>86.36</b>	
With Knowledge	InCross	86.10	81.38	84.36	82.84	85.39	85.80	85.60
	HKE	87.36	81.84	<b>86.48</b>	84.09	-	-	-
	CMGCN	<b>87.55</b>	<b>83.63</b>	84.69	<b>84.16</b>	<b>87.02</b>	<b>86.97</b>	<b>87.00</b>

# Results

Datasets	Models	Acc(%)	F1(%)
MVSA-Single	BERT+ViT	69.11	68.84
	Our CLFA	73.11	<b>72.45</b>
	RoBERTa+ViT	68.44	68.67
	Our CLFA	<b>73.33</b>	72.01
MVSA-Multiple	BERT+ViT	68.14	67.39
	Our CLFA	<b>69.73</b>	<b>68.31</b>
	RoBERTa+ViT	67.02	65.86
	Our CLFA	69.02	67.26

Table 5: Experimental results on MMSA.

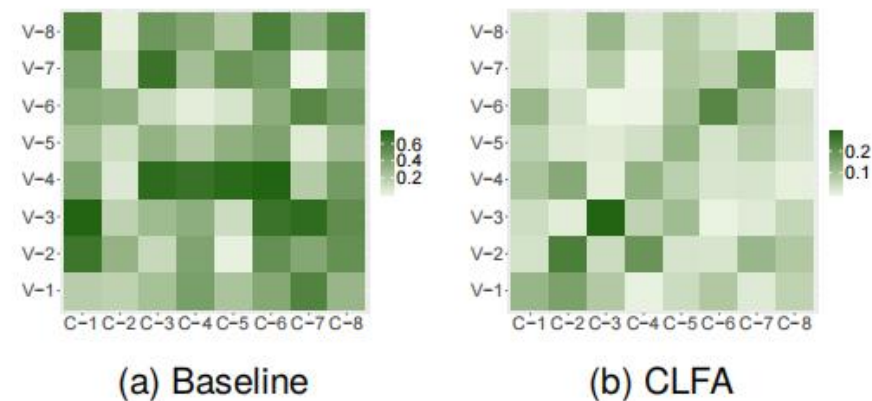


Figure 4: Text-image alignment heatmap on the MMSD data. C indicates the image caption, V indicates the image. The darker is the color, the higher is the similarity.

# Conclusion

- Dual encoder models can learn multi-modal feature alignment by using CLIP as a teacher model
- CLFA gains large improvement on MMSA and MMSSD tasks
- Our method can be combined with other knowledge-enhanced models.

*Thanks*