

# 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.

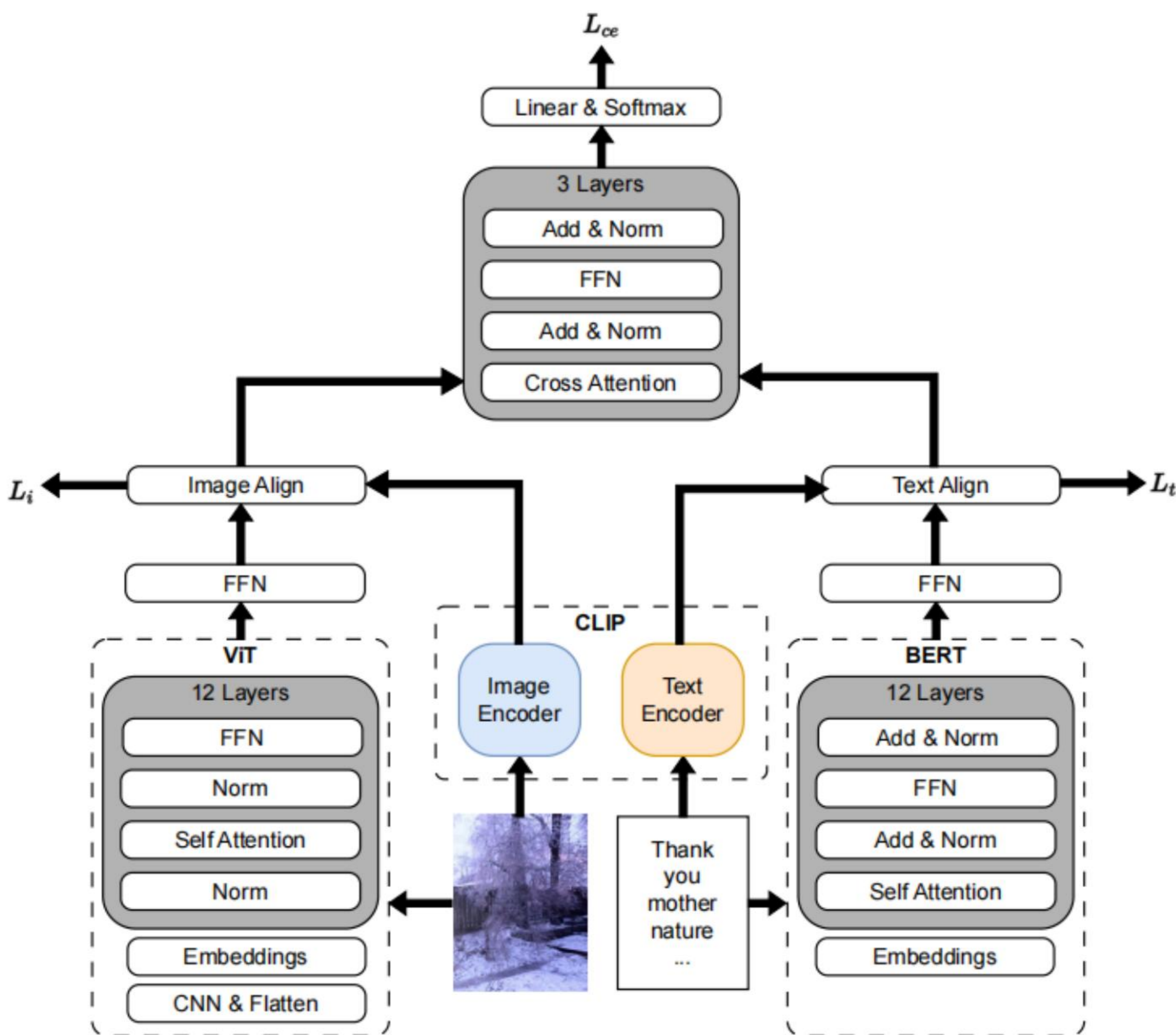
## Introduction

- We propose a novel CLIP-guided contrastive-learning-based architecture to perform multi-modal feature alignment, which projects the features derived from different modalities into a unified deep space.
- On MMSD and MMSA tasks, the experimental results show that our proposed model significantly outperforms several baselines.
- Simple to implement without using task-specific external knowledge, and thus easily migrate to other multi-modal tasks

## Method

- Obtain text and image representations by using BERT and ViT
- Use CLIP and contrastive learning to Align text and image features
- Cross attention for inter-modal feature fusion

Overview of our proposed model



### Text & Image Encoding

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

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

### Cross-modal Feature Alignment

- Obtain CLIP representation

$$C_t = [\text{CLIP}_t(T_1), \dots, \text{CLIP}_t(T_B)]$$

$$C_i = [\text{CLIP}_i(I_1), \dots, \text{CLIP}_i(I_B)]$$

- Mapping net to align feature dimension

$$f'_t = \text{MLP}(f_t)$$

$$f'_i = \text{MLP}(f_i)$$

- Align the BERT and ViT Features

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

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

### Loss functions

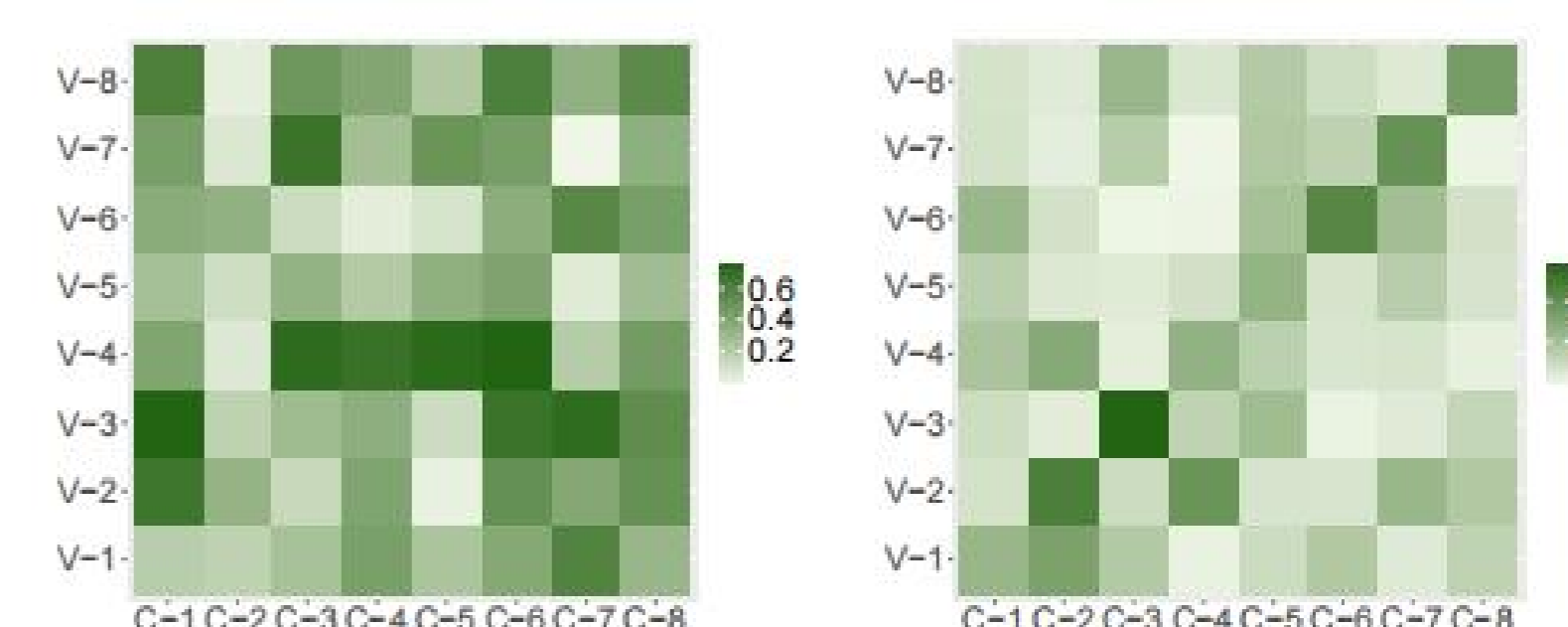
$$\mathcal{L}_{con} = \frac{1}{2}\mathcal{L}_i + \frac{1}{2}\mathcal{L}_t$$

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

Results on MMSD

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>

Visual Analysis



(a) Baseline

(b) CLFA

Results on MMSA

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

## Conclusions

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