Multi-Modal Sarcasm Detection Based on Relationship Dependence of Knowledge Graph

Reported by Danling.Wei

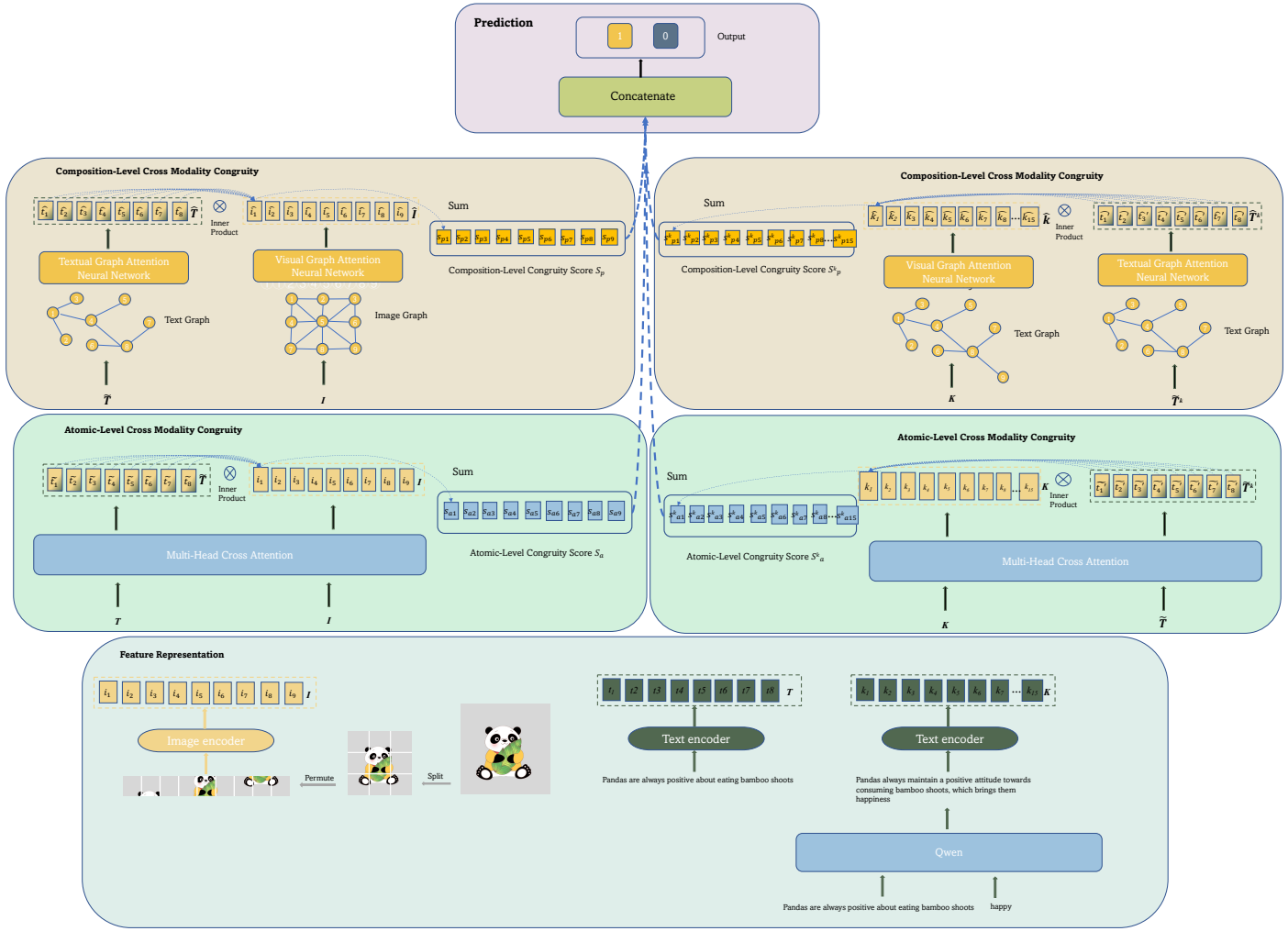
Motivation

Sarcasm
Time to change the channel for some good news

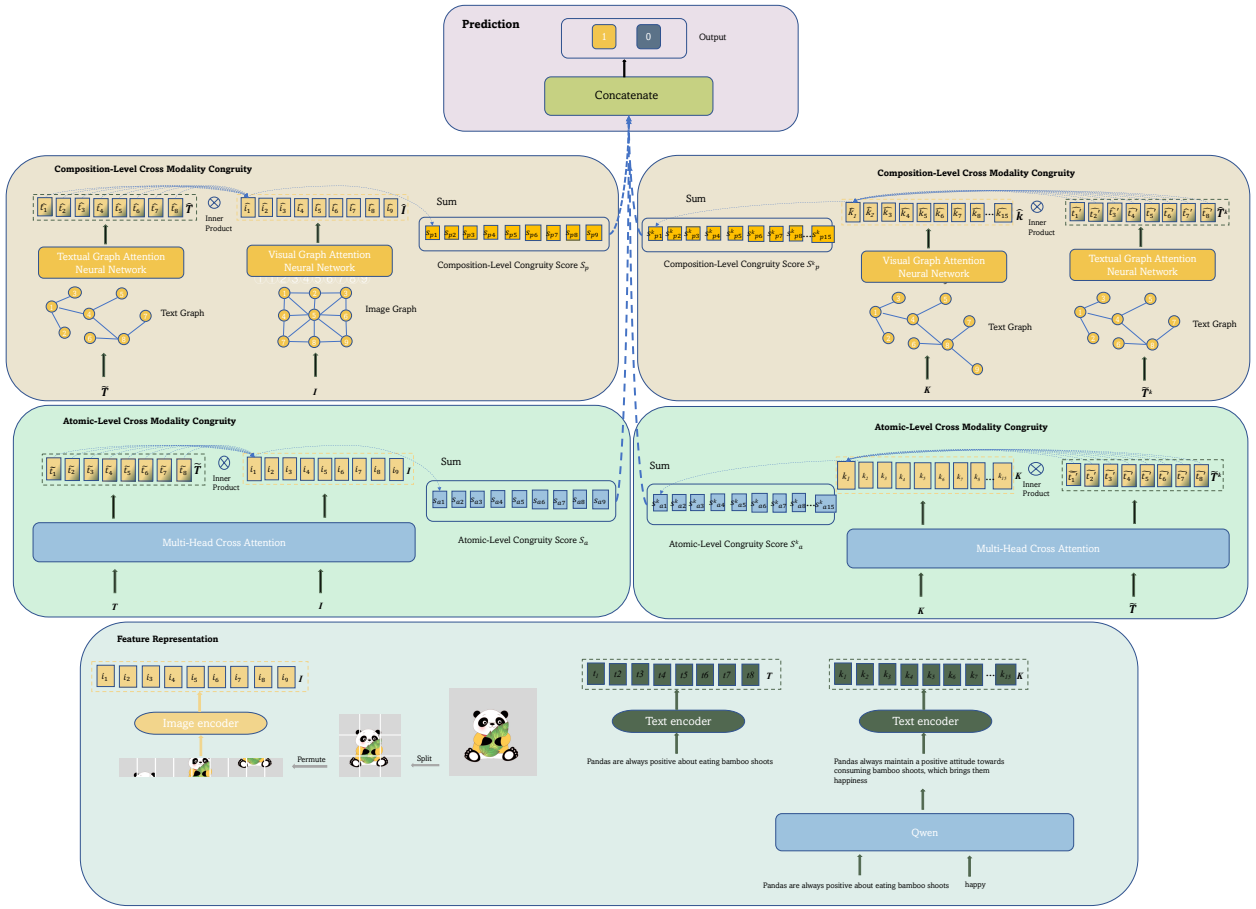
External Knowledge
Image Attributes: green, scene, different, book, signs
ANPs: horizontal text, young driver, clean air, tough competition, evil alien
Image Caption: a map showing the current state of the storm

Figure 1: An example of sarcasm along with the corresponding image and different types of external knowledge extracted from the image. The sarcasm sentence represents the need for some good news. However, the image of the TV program is switched to bad news depicting severe storms (bad weather) which contradicts the sentence.

Overview


Prediction
1
0
Output
Concatenate
Composition-Level Cross Modality Congruity
Inner Product
Sum
Composition-Level Congruity Score S_p
Textual Graph Attention Neural Network
Visual Graph Attention Neural Network
Text Graph
Image Graph
Composition-Level Cross Modality Congruity
Sum
Inner Product
Composition-Level Congruity Score S^k_p
Visual Graph Attention Neural Network
Textual Graph Attention Neural Network
Text Graph
Text Graph
Atomic-Level Cross Modality Congruity
Inner Product
Sum
Multi-Head Cross Attention
Atomic-Level Congruity Score S_a
Atomic-Level Cross Modality Congruity
Sum
Inner Product
Atomic-Level Congruity Score S^k_a
Multi-Head Cross Attention
Feature Representation
Image encoder
Permute
Split
Text encoder
Pandas are always positive about eating bamboo shoots
Text encoder
Pandas always maintain a positive attitude towards consuming bamboo shoots, which brings them happiness
Qwen
Pandas are always positive about eating bamboo shoots
happy

Method



$$\mathbf{head}_i = \text{softmax}\left(\frac{(\mathbf{T}\mathbf{W}_q^i)^\top}{\sqrt{d/h}}(\mathbf{I}\mathbf{W}_k^i)\right)(\mathbf{I}\mathbf{W}_v^i), \quad (1)$$

$$\tilde{\mathbf{T}} = \text{norm}(\mathbf{T} + \text{MLP}([\mathbf{head}_1 \,\|\, \mathbf{head}_2 \,\|\, \ldots \,\|\, \mathbf{head}_h])), \quad (2)$$

$$\mathbf{Q}_a = \frac{1}{\sqrt{d}}(\tilde{\mathbf{T}}\mathbf{I}^\top)$$

$$\mathbf{s}_a = \text{softmax}(\tilde{\mathbf{T}}\mathbf{W}_a + \mathbf{b}_a)^\top \mathbf{Q}_a, \quad (3)$$

$$\alpha_{i,j}^l = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{v}_l^\top[\mathbf{\Theta}_l\mathbf{t}_i^l \| \mathbf{\Theta}_l\mathbf{t}_j^l]\right)\right)}{\sum_k \exp\left(\text{LeakyReLU}\left(\mathbf{v}_l^\top[\mathbf{\Theta}_l\mathbf{t}_i^l \| \mathbf{\Theta}_l\mathbf{t}_k^l]\right)\right)}, \quad (4)$$

$$\mathbf{t}_i^{l+1} = \alpha_{i,i}^l \mathbf{\Theta}_l \mathbf{t}_i^l + \sum_{j \in \mathcal{N}(i)} \alpha_{i,j} \mathbf{\Theta}_l \mathbf{t}_j^l, \quad (5)$$

$$\mathbf{c} = \text{softmax}(\mathbf{T}\mathbf{W}_c + \mathbf{b}_c)^\top \tilde{\mathbf{T}}, \quad (6)$$

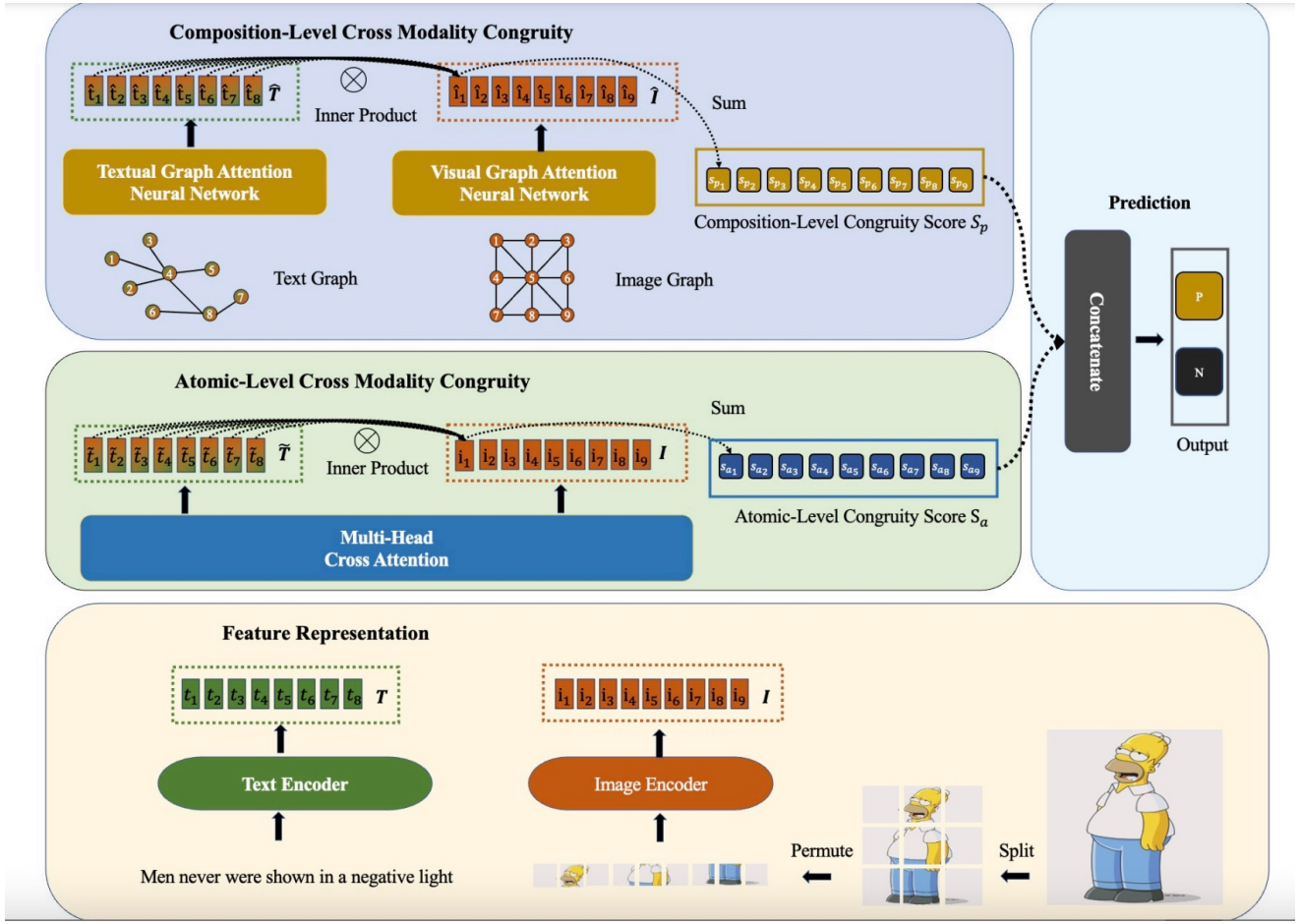
$$\mathbf{Q}_p = \frac{1}{\sqrt{d}}([\hat{\mathbf{T}} \,\|\, \mathbf{c}]\hat{\mathbf{I}}^\top)$$

$$\mathbf{s}_p = \text{softmax}([\hat{\mathbf{T}} \,\|\, \mathbf{c}]\mathbf{W}_p + \mathbf{b}_p)^\top \mathbf{Q}_p, \quad (7)$$

$$\mathbf{p}_v = \text{softmax}(\mathbf{I}\mathbf{W}_v + \mathbf{b}_v), \quad (9)$$

$$\mathbf{y}' = \text{softmax}(\mathbf{W}_y[\mathbf{p}_v \odot \mathbf{s}_a \,\|\, \mathbf{p}_v \odot \mathbf{s}_p] + \mathbf{b}_y), \quad (10)$$

Overview


Composition-Level Cross Modality Congruity
Inner Product
Sum
Textual Graph Attention Neural Network
Visual Graph Attention Neural Network
Composition-Level Congruity Score S_p
Text Graph
Image Graph
Prediction
Concatenate
P
N
Output
Atomic-Level Cross Modality Congruity
Sum
Inner Product
Multi-Head Cross Attention
Atomic-Level Congruity Score S_a
Feature Representation
Text Encoder
Image Encoder
Men never were shown in a negative light
Permute
Split

Experiments

Table 3.2 Comparison results for sarcasm detection (%).

模型		准确率	精确率	召回率	F1
文本	TextCNN	78.06	73.30	77.20	73.54
	Bi-LSTM	80.20	75.66	75.40	76.21
	BERT	81.75	76.62	81.82	78.20
图像	ResNet	64.01	53.90	71.23	60.79
	ViT	68.32	54.91	71.05	62.23
多模态	HFM	83.44	76.57	84.15	80.18
	InCrossMGs	85.57	81.26	84.32	82.14
	CMGCN	85.63	-	-	82.65
	HKE	85.92	81.40	84.93	83.13
	RDKG (Qwen)	**86.68**	**82.81**	**85.04**	**83.91**
	RDKG (MiniGPT4)	86.39	82.49	84.62	83.55

Table 2: Comparison results for sarcasm detection. † indicates ResNet backbone and ‡ indicates ViT backbone.

Model		Acc(%)	P(%)	R(%)	F1(%)
Text	TextCNN	80.03	74.29	76.39	75.32
	Bi-LSTM	81.90	76.66	78.42	77.53
	SMSD	80.90	76.46	75.18	75.82
	BERT	83.85	78.72	82.27	80.22
Image	Image	64.76	54.41	70.80	61.53
	ViT	67.83	57.93	70.07	63.43
Multi-Modal	HFM†	83.44	76.57	84.15	80.18
	D&R Net†	84.02	77.97	83.42	80.60
	Att-BERT†	86.05	80.87	85.08	82.92
	InCrossMGs‡	86.10	81.38	84.36	82.84
	CMGCN‡	86.54	–	–	82.73
	Ours†	87.02	**82.97**	84.90	83.92
	Ours‡	**87.36**	81.84	**86.48**	**84.09**

Table 3: Results of different knowledge types.

Knowledge Type	Acc(%)	F1(%)
w/o external knowledge	87.36	84.09
Image Attributes	86.43	83.30
ANPs	86.35	83.54
Image Captions	**88.26**	**84.84**
Image Captions (w/o image)	86.60	83.28

Thanks!